Comparative Study of Machine Learning Techniques for Population Genetics

Muhammad Arslan Amin, Muhammad Kashif Hanif*, Muhammad Umer Sarwar, Mohsin Abbas, Muhammad Haroon Jilani, Usman Nasir, Muhammad Bilal Sarwar, and Hafiz Muhammad Talha

Department of Computer Science, Government College University, Faisalabad, Pakista

Abstract
As the size of population genetic data increases, researchers face difficulties in understanding this huge amount of data. In order to work with complex data, computational methods are being developed to work precisely with population genetic data. Various kinds of computational techniques have been developed to analyze population genetic data. Machine learning is a significant area that has considerable potentials for population genetics. Machine learning aims to implement computer algorithms that learn with experience to help humans in the analysis of complex and large data sets. Machine learning is still in its infancy for various problems, especially in the area of evolutionary and population genetics. This study presents machine learning applications in order to investigate the genetic data of population including different concepts that are relevant to population genetics

Key words:
Machine Learning; Computer Algorithms; Genetic Data; Population Genetics; Evolutionary Genetics.

1. Introduction

Population genetics is a part of evolutionary biology and sub-field of genetics that deals with genetics differences within and between populations. One of the significant purpose of evolutionary biology is to understand the development of phenotype diversity that can occur by differences at the deoxyribonucleic acid (DNA) level. Furthermore, the technology can help us to achieve this information and analyze different individuals. The field of population genetics seeks to understand the variations of DNA at the population level. There exists different models to understand the recent evolutionary changes that give rise to observe genetic variations. DNA sequence of an individual is built in the form of character strings A (Adenine), C (Cytosine), G (Guanine), T (Thymine). These strings may affect due to various factors including natural selection, population size changes, mutation, migration, recombination, and population splits. These factors are important to model with the goal for obtaining a correct information about the forces that can cause genetic variation Due to rapidly increasing growth of population genetic data, complexity for understanding data is increased. Computational methodologies are being developed to work efficiently with population genetic data. Machine learning is considered to be an important and significant computational area having considerable potential for population genetics. In this area, different computer algorithms are developed which can enhance the performance with experiences. Machine learning techniques have been implemented in a vast variety of concepts within population genetics. The objective is to discuss prediction of population genetic data by using appropriate machine learning models. This work presents different prediction techniques of population genetic data obtained by applying the machine learning [1], [2].

The rest of the paper is organized in different sections. In the next section, introduction to population genetics is presented. Section 3 describes different machine learning techniques. Section 4 discusses different applications of machine learning in population genetics. The outcomes are concluded in section 5.

2. Population Genetics

Population is considered to be a vital point for finding analysis about a specific group of individuals. The researchers have defined population in various ways. Population is depicted by various analyst, physicians, and biologists in their own ways. In this study, main focus is the field of population genetics, where population can be defined as a group of individuals (Plants, Humans and Animals etc.) or species within which breeding occurs. In population genetics, population plays a vital role for inference of different statistical predictions among genetic data of specific species [3].

Genetic variation is a change that exist in the genes of individuals in a population. There are various factors involved that causes variation in the individuals of population. These variations can occur due to changes in the sequences of DNA, chromosomal variations, variation in proteins, and function of proteins etc. Organisms within a population vary in their particular attributes. For example, we can easily notice the phenotype variations (eye color, facial features, and height) in humans. Genetic variation makes individuals unique in the form of hair colors, height and facial changes. Genetic variations occur due to the insertion, deletion or rearranging of genes that can influence multiple genes or large chromosomal areas. The genetic

Manuscript received June 5, 2019
Manuscript revised June 20, 2019

variations within or between population are studied in population genetics [4].
The population genetics is rapidly growing field and an active area for researchers. It can be considered as a part of evolutionary biology where changes are made with the passage of time in the genetic composition of population. It is a challenging task to find that how variations will occur in the population. The individuals of a population carries a set of genomes called as genetic composition. These composition of genes can vary due to death or moving an individual within or outside of population. Population genetics can be defined in different ways. Population genetics is a field of biology that studies the variation among genetic compositions of population which are results of different evolutionary forces. These genetic variations are occurred by different factors including migration, mutation, natural selection, recombination, and genetic drift [4], [5].

3. Machine Learning

Over the last few decades, the methods of machine learning have revolutionized in many fields like image classification [6], [7], speech recognition [8], [9], natural language processing [10], [11], and bioinformatics [12]–[15]. However, machine learning is not yet widely applied in the field of evolutionary and population genetics.
Machine learning is a field which enables the computers to use past experiences and generate accurate predictions. Machine learning techniques can automatically build computational models of complex relationships by computing the available data. The process of automatically building models is called training and the data which is being utilized for training is called training data. In machine learning, usually number of computational models are trained for a certain problem. Unfortunately, there are no such rules for a particular model or algorithm selection. There are different factors to check the performance of a model such as the size and complexity of data, time for training and available memory etc. Some problems may necessarily require training of more than one model and algorithm to choose the specific one. In machine learning algorithms, features (properties of a problem) are used as input basis on which we would like to predict results [16].
There are various kinds of machine learning techniques that have been developed for efficiently performing analysis on data and make successful predictions. These techniques are briefly discussed in following subsection.

3.1 Classification of Machine Learning

Machine learning can be classified as supervised and unsupervised learning. Supervised learning learn to predict output when given an input vector or instances are given with known labels. Supervised machine learning utilizes classification technique and classify the data accordingly. Supervised machine learning works comparatively faster than other techniques. Supervised machine learning provides accurate results when input a new data without having a priori knowledge about it. Therefore, supervised machine learning is considered more sophisticated learning. A problem that is more common in supervised learning is known as over-fitting that occurs when data performs well on the data utilized for training and ineffectively with the new data. The examples of supervised learning are Naive Bayes, Support Vector Machines (SVM), linear and logistics regression etc [17].
Unsupervised machine learning creates an internal representation of input or instances with unknown labels. Unsupervised learning utilizes clustering techniques and data with the similar observations are kept in the same cluster. After the completion of clustering process, the data within the clusters are labeled by own because there is no known labels in the method of unsupervised machine learning. Clustering is mostly used for statistical data. There are various types of clustering like k-mean clustering, principle component analysis etc [18].

3.2 Artificial Neural Network

An artificial neural network (ANN) was initially motivated by the functionality of human brain where large number of neurons are interconnected to process information. ANN plays a vital role in machine learning to analyze data. In machine learning, neural network consists of multiple artificial neurons. An ANN contains an input layer, one or more hidden layer and one output layer to process information. Each layer of neural network connects through a numeric weights. Artificial neural network receives data in an input layer which is then transferred to the successive layers.
The step by step processing of information executed in ANN is shown in Figure 1 where panel (A) shows the inputs are passed through hidden layer of neural network to the output layer. Panel (B) shows that the neurons in ANN computes the weighted sum of inputs and an activation function will be applied to generate output. Panel C shows the local and global optimum for weights [12].
There are various algorithms to train ANN. Backpropagation is an algorithm of ANN which is most used due to programming ease and has a power to manipulate large amount of data. The working purpose of back-propagation algorithm is to learn neural network in efficient way. Back-propagation algorithm consists of test and train phases. The back-propagation algorithm have forward and backward propagation (Figure 1).
The first part of back-propagation algorithm is feed-forward pass which present inputs to the network and propagate forward to produce the output. The second part is backward propagation which computes the gradient descent and

weights are updated according to error correction rule. The iteration in back propagation algorithm is repeated until the problem obtains minimum error [12], [19], [20].

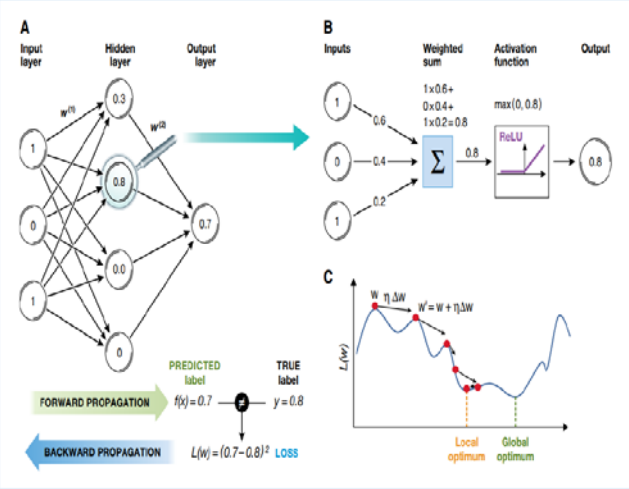

Fig. 1 Work Flow of Artificial Neural Network [12].

3.3 Deep Learning

Deep learning is considered most powerful method of machine learning which has been applied to various problems in different fields. Deep learning revolutionized many fields like speech recognition, text recognition, image recognition, and computational biology etc. The use of deep learning method by the researchers in the field of computational biology is increased to some extent but still this work is in its infancy. Deep learning helps to make efficient predictions against complex data models. Deep learning is considered to be most commonly used neural network. Deep learning is utilized for large data sets and not appropriate for the small data sets. The data set that are being utilized for deep learning requires a lot of time for training and testing. After testing and training, it results an accurate predictions. Like other neural network, deep learning neural network has phases of feed forward propagation and back-propagation. These are the basic phases for neural network to make prediction accurate [12], [21].

Figure 2 shows the framework of deep learning neural network where first layer is an input layer, the next two layers are hidden or processing layers and the last one is output layer to estimate the parameters of interest. The number of neurons in each layer are (Input layer) L1=4, (Hidden layers) L2, L3=3, and (Output layer) L4=2. Each layer have a (+1) node represented as a bias term.

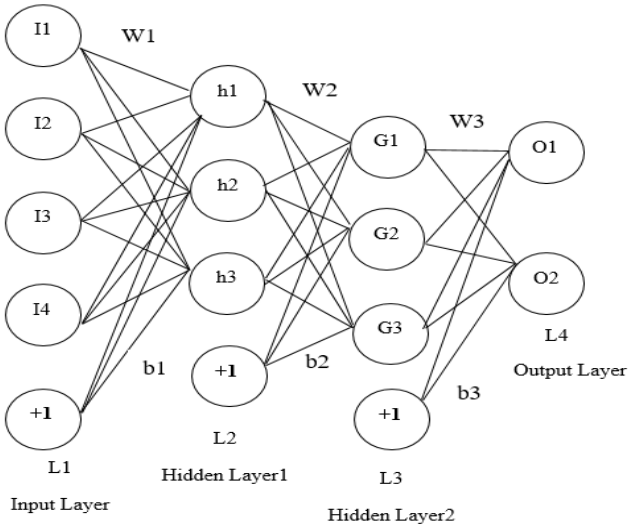

Fig. 2 Framework of deep learning neural network [22].

3.4 Convolutional Neural Network

Convolution neural network (CNN) plays an important role in various fields like image detection, face recognition, handwriting recognition and various machine learning problems etc. CNN consists of layers that can be 2 or 3 dimensional. The layers of CNN are used to input data and get results after performing operations on these inputs [23]. CNN layers are similar to regular neural networks because they also contains fully connected layers which means that all neurons are links with the previous neurons. The layer in CNN are divided into two layers known as pooling layer and convolution layer (Figure 3).

The convolution layer is also known as detection layer because it can be used to detect features. The convolution layer contains a set of filters/kernels. The images are convolved with a filter and the results are obtained as a set of feature maps. Convolution layer also performs forward and back-propagation in their own way. The second type of layer is a pooling layer which reduces the size of parameter/representation and build in invariance to little transformations. Most commonly used type of pooling layer is max pool which compute maximum value of nodes. Similarly min pool computes minimum value of nodes. Furthermore, pooling layer also computes forward and back propagation in a specific way.

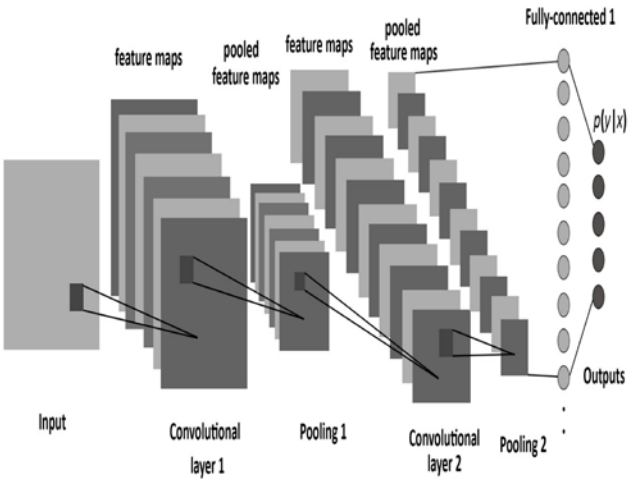

Fig. 3 Architecture of Convolutional Neural Network [24].

4. Machine Learning for Population Genetics

During past decade, researchers have developed various machine learning tools and techniques in order to analyze population genetic data. The techniques developed for population genetics are still not efficient. This work discusses previously developed techniques to predict population genetic data using machine learning. Machine learning technique are able to predict a number of problems arise in population genetics very successfully and accurately.

4.1 Analyzing Population Genetic Data using Hidden Markov Model and Principal Component Analysis

Hidden Markov model (HMM) has been employed in the field of evolutionary and population genetics. HMMs was used to reveal contrast among evolutionary rates with chromosome [25], [26]. HMMs for population genetics have been used to identify the selective sweeps and genomic regions [27] – [30]. . Pei et al. have addressed three distinct problems for the inference of population genetics with the aim of developing an accurate and fast algorithm. This method estimates maximum likelihood based on HMM. Moreover, species delimitation problem was addressed using an efficient SVM classification based method known as CLADES (a classification based delimitation of species). CLADES can work well when existing techniques have difficulties in species delimitation

[31]. Further, they designed a machine learning method to resolve the inference problem of demographic histories under isolation with migration model. Regression trees were trained for the estimation of parameters using boosting [32].

genomes. SweeD software quickly identifies sweeps when compared with SweepFinder [40].

Principle component analysis (PCA) can be used to find relationship of unknown relatedness among individuals in the context of population genetics. PCA utilizes high dimensionality inputs and then (after performing computations) obtain summaries with reduced dimensionality. Novembre et al. utilized PCA for the analysis of individuals from Europe to find the relationship among them. There exists multiple unsupervised machine learning application that have been applied to population genetics beyond PCA [33].

There are number of problems in population genetics requiring efficient analysis on the data set. Supervised ML can be used to make accurate predictions in data sets that cannot be adequately modeled with a reasonable number of parameters. Recently, a supervised machine learning technique has been developed in order to find purifying selection from human genome data. Schrider and Kern have implemented supervised machine learning based SVM classifier to analyze population genetic data. This technique helps to find the earlier known human specific loss or gains of function. This supervised machine learning method is helpful to perform complex tasks and very effective and accurate [34].

4.2 Analyzing Population Genetic Data using Hidden Markov Model and Principal Component Analysis

One of the main purpose of population genetics is detection of selective sweeps. The classical population genetic approaches for discovering sweeps formulates test statistics [35]– [38]. In recent years, there exists various machine learning methods for population genetics to improve the inferential power and efficiently finding selective sweeps. Machine learning algorithms can classify genome into neutral vs selective regions. Most of these techniques are based on SVM [39]– [41] or performance boosting techniques [42], [43].

Pavlidis et al. used SVM to combine statistics of site frequency spectrum (SFS) and linkage disequilibrium (LD) [39]. In order to obtain these statistics (SFS and LD), they have used SweepFinder and statistic [36]. The focus was the genetic patterns of variation with chromosome recombination which utilizes a large number of contrast among selective and neutral regions. They have used SVM classifier to distinguish between selective or neutral regions. The combination of SFS and LD had great power to identify selective sweeps [39]. Afterwards, Pavlidis et al. have developed an efficient software sweep detector (SweeD) for sophisticated detection of sweeps on the basis of likelihood. This software provides quick identification of selective sweeps in the whole genome data. SweeD is an efficient, expandable and well-structured implementation of SweepFinder that performs analysis on large number of

Moreover, Ronen et al. have proposed a technique to make contrast among neutral and selected regions. For this

purpose, they have introduced a SVM classifier called SFselect. The separation between selective versus neutral regions is based upon feature vector of SFS. These feature are used in SVM (i.e., SFselect) to obtain efficient discrimination [41]. Lin et al. have designed a boosting method for accurately identifying the selective sweeps using feature vectors [42]. Pybus et al. described a powerful framework for selective sweeps based on supervised machine learning. They have developed a hierarchical boosting classifier to identify selective sweeps and classify them in genomic regions [43].

Schrider and Kern designed a technique capable of efficient discrimination between soft and hard sweeps on standing variation. This technique is based on supervised machine learning called soft or hard inference through classification (S/HIC). This helps to find soft and hard sweep [44]. In S/HIC, Kern and Schrider used Extra-Tree classifier for the detection of hard and soft sweeps. Moreover, they have represented an update in S/HIC that appeared to be both exact and strong.

4.3 Inference of Recombination Rate and Demography

Machine learning techniques can be utilized to infer recombination rates and demography in population genetics. Approximate Bayesian computation (ABC) techniques are considered popular for the inference of demographic history [46]. Aeschbacher et al. have introduced an approach for choosing best statistical summaries in ABC. Their work is based on boosting in order to select summary statistics. Further, they have compared different types of boosting technique to select highly accurate approach [47]. Moreover, Jiang et al. described how to build informative and low dimensional summary statistics automatically with the help of ABC. They have trained statistical summaries through a deep neural network. The trained data has been utilized for parametric prediction [48].

Furthermore, supervised machine learning has been utilized for the selection of demographic models. Pudlo et al. have developed an efficient technique based on random forests (RF) to perform selection among complicated models executed by ABC algorithm [49]. Smith et al. have designed a method to overcome the curse of dimensionality and issues related to summarization and simulations of large scale SNP data sets. They utilized an RF classifier for demographic model selection to address the dimensionality issue and apply binning strategy for the construction of multidimensional site frequency spectrum (mSFS) to circumvent the issue of summarization and simulation of large scale SNP data set. They have compared different demographic models. The results indicates that the combination of binning strategy for mSFS with RF strategy for demographic model selection can preforms well for phylogeographic model selection [50]. Schrider et al. have developed an Extra-Tree classifier for locus-specific demography model selection to recognize regions with gene flow among a couple of correlated species [51]. Moreover, Beichman et al. have discussed various efficient statistical approaches to make inference of demographic histories and how these approaches are applied to different non-model organisms. They provided an overview of the theory and logics regarding every approach and recommendations for researchers to use genomic data for the inference of demographic history [52].

Additionally, supervised machine learning has also been utilized for analyzing recombination rates and its patterns in the genome. Adrian et al. have developed a method to specify recombination rates. They have trained a random forest classifier to differentiate between the classes of recombination rates in drosophila melanogaster [53]. Lin et al. have introduced a regression based method for the estimation of recombination rates. For large data, this method has significantly less computational power than other techniques. They have utilized the combination of boosting and regression for the best model selection [54]. Gao et al. have introduced a machine learning based software FastEPRR for the estimation of recombination rates within large data sets [55]. FastEPRR has similar accuracy and significantly more efficiency than the recently developed method LDhat [56]. Chan et al. have built a method for population genetic data, i.e., likelihood free inference that does not depends on the handmade summary statistics. They have composed a group of neural networks to learn the exchangeable features of genetic data and applied genotypic data to a group of neural networks. Further, this method was applied on the complex issue of recombination hotspot testing. Moreover, they have developed techniques of machine learning to build a powerful scientific simulator to empower the already developed likelihood-free inference techniques [57].

Simultaneously estimation of demography and selection has been done by applying deep learning neural network. Sheehan and Song computed composite estimation of demography and selection. They have introduced an algorithm likelihood free inference which is implemented through deep learning neural network. They have selected samples of drosophila melanogaster from Zambia for analysis and finds regions of genomes under selection. These regions are later transferred into summary statistics to estimate the demography and selected regions of genome. Their main contribution is the inference of demography and selection by applying neural network [22].

References

[1] S. Sheehan, “Scalable algorithms for population genomic inference”, Ph.D. dissertation, UC Berkeley, 2015

[2] D. R. Schrider and A. D. Kern, “Machine learning for population genetics: A new paradigm”, bioRxiv, p. 206482, 2017

[3] N. Krieger, “Who and what is a “population”? historical debates, current controversies, and implications for understanding “population health” and rectifying health inequities”, The Milbank Quarterly, vol. 90(4), pp. 634–681, 2012
[4] M. De Vicente, C. Lopez, and T. Fulton, “Genetic diversity analysis with molecular marker data: learning module”, International Plant Genetic Resources Institute (IPGRI), 2004
[5] S. Choudhuri, Bioinformatics for beginners: genes, genomes, molecular evolution, databases and analytical tools. Elsevier, 2014
[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks”, in Advances in neural information processing systems, pp. 1097–1105, 2012
[7] R. McCoppin and M. Rizki, “Deep learning for image classification”, in Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR V, vol. 9079. International Society for Optics and Photonics, p. 90790T, 2014
[8] G. Hinton et al, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups”, IEEE Signal processing magazine, vol. 29, no. 6, pp. 82–97, 2012
[9] T. Afouras et al, “Deep audio-visual speech recognition”, arXiv preprint arXiv 1809.02108, 2018
[10] F. Sebastiani, “Machine learning in automated text categorization”, ACM computing surveys (CSUR), vol. 34(1), pp. 1–47, 2002
[11] T. Young et al, “Recent trends in deep learning based natural language processing”, ieee Computational intelligenCe magazine, vol. 13(3), pp. 55–75, 2018
[12] C. Angermueller et al, “Deep learning ¨ for computational biology”, Molecular systems biology, vol. 12(7), p. 878, 2016
[13] E. Byvatov and G. Schneider, “Support vector machine applications in bioinformatics”, Applied bioinformatics, vol. 2(2), pp. 67–77, 2003
[14] M. W. Libbrecht and W. S. Noble, “Machine learning applications in genetics and genomics”, Nature Reviews Genetics, vol. 16(6), p. 321, 2015
[15] S. Min, B. Lee, and S. Yoon, “Deep learning in bioinformatics,” Briefings in bioinformatics, vol. 18(5), pp. 851–869, 2017
[16] Y. Baştanlar, and M. Özuysal, “Introduction to machine learning”, in miRNomics: MicroRNA Biology and Computational Analysis. Springer, pp. 105–128, 2014
[17] S. B. Kotsiantis et al, “Supervised machine learning: A review of classification techniques”, Emerging artificial intelligence applications in computer engineering, vol. 160, pp. 3–24, 2007
[18] Z. Ghahramani, “Unsupervised learning”, in Advanced lectures on machine learning. Springer, pp. 72–112, 2004
[19] R. C. S. Nakano et al, “Implementation of an artificial neural network in recognizing in-flight quadrotor images”, in TENCON 2015-2015 IEEE Region 10 Conference. IEEE, pp. 1–5, 2015
[20] J. C. Chaudhari, “Design of artificial back propagation neural network for drug pattern recognition”, International Journal on Computer Science and Engineering (IJCSE), pp. 1–6, 2010
[21] W. Jones et al, “Computational biology: deep learning”, Emerging Topics in Life Sciences, vol. 1(3), pp. 257– 274, 2017
[22] S. Sheehan and Y. S. Song, “Deep learning for population genetic inference,” PLoS computational biology, vol. 12, no. 3, p. e1004845, 2016
[23] K. O’Shea and R. Nash, “An introduction to convolutional neural networks”, arXiv preprint arXiv:1511.08458, 2015
[24] S. Albelwi and A. Mahmood, “A framework for designing the architectures of deep convolutional neural networks”, Entropy, vol. 19(6), p. 242, 2017
[25] J. Felsenstein and G. A. Churchill, “A hidden markov model approach to variation among sites in rate of evolution”, Molecular biology and evolution, vol. 13(1), pp. 93–104, 1996
[26] A. Siepel et al, “Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes”, Genome research, vol. 15(8), pp. 1034–1050, 2005
[27] S. Boitard et al, “Detecting selective sweeps: a new approach based on hidden markov models”, Genetics, 2009
[28] S. Boitard et al, “Detecting selective sweeps from pooled next-generation sequencing samples”, Molecular biology and evolution, vol. 29(9), pp. 2177– 2186, 2012
[29] S. Boitard et al, “Pool-hmm: a python program for estimating the allele frequency spectrum and detecting selective sweeps from next generation sequencing of pooled samples”, Molecular ecology resources, vol. 13(2), pp. 337–340, 2013
[30] A. D. Kern and D. Haussler, “A population genetic hidden markov model for detecting genomic regions under selection”, Molecular biology and evolution, vol. 27(7), pp. 1673–1685, 2010
[31] J. Pei et al, “Clades: A classification based machine learning method for species delimitation from population genetic data”, Molecular ecology resources, 2018
[32] J. Pei, “Methods and algorithms for inference problems in population genetics”, Methods, vol. 7, pp. 9–2018, 2018
[33] J. Novembre et al, “Genes mirror geography within europe”, Nature, vol. 456, no. 7218, p. 98, 2008
[34] D. R. Schrider and A. D. Kern, “Inferring selective constraint from population genomic data suggests recent regulatory turnover in the human brain”, Genome biology and evolution, vol. 7(12), pp. 3511– 3528, 2015
[35] J. C. Fay and C.-I. Wu, “Hitchhiking under positive darwinian selection”, Genetics, vol. 155(3), pp. 1405–1413, 2000
[36] Y. Kim and R. Nielsen, “Linkage disequilibrium as a signature of selective sweeps”, Genetics, vol. 167(3), pp. 1513–1524, 2004
[37] F. Tajima, “Statistical method for testing the neutral mutation hypothesis by dna polymorphism”, Genetics, vol. 123(3), pp. 585–595, 1989
[38] B. F. Voight et al, “A map of recent positive selection in the human genome”, PLoS biology, vol. 4(3), no. 3, p. e72, 2006
[39] P. Pavlidis et al, “Searching for footprints of positive selection in whole-genome snp data from non-equilibrium populations”, Genetics, 2010
[40] P. Pavlidis et al, “Sweed: ´ likelihood-based detection of selective sweeps in thousands of genomes”, Molecular biology and evolution, vol. 30(9), pp. 2224– 2234, 2013
[41] R. Ronen et al, “Learning natural selection from the site frequency spectrum”, Genetics, pp. genetics– 113, 2013

[42] K. Lin et al, “Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics”, Genetics, 2010
[43] M. Pybus et al, “Hierarchical boosting: a machine learning framework to detect and classify hard selective sweeps in human populations”, Bioinformatics, vol. 31(24), pp. 3946–3952, 2015
[44] D. R. Schrider and A. D. Kern, “S/hic: robust identification of soft and hard sweeps using machine learning”, PLoS genetics, vol. 12(3), p. e1005928, 2016
[45] A. D. Kern and D. R. Schrider, “diplos/hic: an updated approach to classifying selective sweeps”, G3: Genes, Genomes, Genetics, pp. g3– 200 262, 2018
[46] M. A. Beaumont, “Approximate bayesian computation in evolution and ecology”, Annual review of ecology, evolution, and systematics, vol. 41, pp. 379–406, 2010
[47] S. Aeschbacher et al, “A novel approach for choosing summary statistics in approximate bayesian computation”, Genetics, pp. genetics–112, 2012
[48] B. Jiang et al “Learning summary statistic for approximate bayesian computation via deep neural network”, arXiv preprint arXiv:1510.02175, 2015
[49] P. Pudlo et al, “Reliable abc model choice via random forests”, Bioinformatics, vol. 32(6), pp. 859–866, 2015
[50] M. L. Smith et al, “Demographic model selection using random forests and the site frequency spectrum”, Molecular ecology, vol. 26(17), pp. 4562–4573, 2017
[51] D. R. Schrider et al, “Supervised machine learning reveals introgressed loci in the genomes of drosophila simulans and d. sechellia”, PLoS genetics, vol. 14(4), p. e1007341, 2018
[52] A. C. Beichman et al, “Using genomic data to infer historic population dynamics of nonmodel organisms”, Annual Review of Ecology, Evolution, and Systematics, no. 0, 2018
[53] A. B. Adrian et al, “Predictive models of recombination rate variation across the drosophila melanogaster genome”, Genome biology and evolution, vol. 8(8), pp. 2597–2612, 2016
[54] K. Lin et al, “A fast estimate for the population recombination rate based on regression”, Genetics, pp. genetics–113, 2013
[55] F. Gao et al, “New software for the fast estimation of population recombination rates (fasteprr) in the genomic era”, G3: Genes, Genomes, Genetics, vol. 6(6), pp. 1563–1571, 2016
[56] G. A. McVean et al, “The fine-scale structure of recombination rate variation in the human genome”, Science, vol. 304(5670), pp. 581–584, 2004
[57] J. Chan et al, “A likelihood-free inference framework for population genetic data using exchangeable neural networks”, arXiv preprint arXiv:1802.06153, 2018