SMGroup

The Revitalization of Population Genetics: From Coalescence Theory to Phylogeography

Barrandeguy ME[1,2,3]*, Sanabria DJ[3,4] and García MV[1,2,3]

[1]Laboratory of Population and Landscape Genetics, Faculty of Chemical and Natural Sciences National University of Misiones (3300) Posadas, Misiones, Argentina

[2]Institute of Subtropical Biology, Posadas Node (UNaM - CONICET)

[3]National Council of Scientific and Technical Research (CONICET - Argentina)

[4]Laboratorio de Biología Molecular Aplicada, Facultad de Ciencias Exactas, Químicas y Naturales, Universidad Nacional de Misiones

***Corresponding author:** Barrandeguy ME, Laboratory of Population and Landscape Genetics, Faculty of Chemical and Natural Sciences. National University of Misiones. Institute of Subtropical Biology, Posadas Node (UNaM - CONICET), National National Council of Scientific and Technical Research (CONICET - Argentina), (3300) Posadas, Misiones, Argentina, Tel: +54 376 4437023; Fax: +54 376 4425414; Email(s): ebarran@fceqyn.unam.edu.ar and euge_barra@hotmail.com

Published Date: July 26, 2017

ABSTRACT

Complex demographic histories characterize natural populations: their sizes and ranges change over time creating processes that leave signatures on their genetic composition. Coalescence theory formalized a powerful way of using only the sample of alleles such that gene genealogies are modeled backwards in time under virtually any complex demographic history. The main aim of this chapter is to show the advances in how to use genetic data from related populations or species to know about their recent evolutionary history. Rational, quantitative assessment of population genetics models and estimates together with evolutionary scenarios should enable to unravel the complex and subtle relationships between demographic history, natural selection, and genomic diversity. Recent statistical methods in phylogeography have revitalized the population genetics allowing to researchers infer the past evolutionary history of populations and estimate demographic and genetic parameters.

Keywords: Approximate Bayesian Computation; Bayesian Skyline Plot; Coalescence theory; Population Genetics

Abbreviations: ABC: Approximate Bayesian Computation; BSK: Bayesian Skyline or Plot; MCMC: Markov Chain Monte Carlo

INTRODUCTION

Population genetics is the discipline that emerges from the combining of the theory of evolution and the Mendelian laws and it is the central point to many current areas of biological science [1]. An enormous progress in population genetic inference has been made during the last years as a consequence of increase in computer processing speed and available gene sequence data [2]. The current abundance of population genetic data has forced the population genetics to become from a theoretical discipline to a more applied data analysis field. Population genetic data are now commonly used for estimating population sizes, describing the history of divergence and migration among populations. Coalescence theory is particularly useful in these applications because of its focus on the properties of samples taken from a population [3]. Researchers of molecular phylogenetics, phylogeography and population genetics are broadly interested in understand the same point: evolutionary change of genomes and the organisms that host them, and as consequence of their common dependence on molecular sequence data these disciplines could be integrated under the view of molecular evolution [4].

Complex demographic histories characterize natural populations: their sizes and ranges change over time, creating processes that leave signatures on their genetic composition. Molecular data are a good tools for expose the complex demographic and adaptive processes that have acted on natural populations. The widespread availability of different molecular markers and the increased computer power has promoted the development of sophisticated statistical methods that contribute to resolve population complex history [5].

The main aim of this chapter is to show the advances in how to use genetic data from related populations or species to know about their recent evolutionary history. This chapter does not intend to review all methods for phylogeographic inference based on coalescence theory otherwise the purpose is show the state of the discipline and how the population genetics contribute in the study of evolutionary history of the populations in the framework of the coalescence theory driven by the technological and computational advances.

The challenge to the modern population genetics is the development of an estimator capable of jointly inferring the action of both non-neutral and non-equilibrium models simultaneously. This will require at least two components: 1) the ability to identify patterns that distinguish selective from demographic effects, and 2) a computational framework capable of handling whole genomes data, a large number of summary statistics, and the accurate inference of multiple parameters of interest [6].

THE WRIGHT-FISHER MODEL

Natural selection and genetic drift are the most important microevolutionary processes that cause changes of allele frequencies along time. Genetic drift is the change of allele frequencies at random in populations of finite size as consequence of sampling error [3]. Wright-Fisher Model is a theory of random genetic drift with binomial sampling that describes the expected distribution of allele frequencies among subpopulations. Under this model the transition probability, i.e. the probability of the population drifting from a state having i copies to j copies of allele A, is directly obtained from the binomial distribution. The expected outcome of a pure drift process can be obtained from iterations of the Wright-Fisher model [7]. In this model, pq/(2N) describe the variance in allele frequency from one generation of random genetic drift corresponding to the variance of the proportion in a binomial distribution. In this way, a large population will change allele frequency more slowly than a smaller population, because of the sampling variance varies as the reciprocal of population size [7].

Using the Wright-Fisher model under the classic population genetics view, researchers can predict the change in the allele frequency mathematically but they are often more interested in understanding the processes that produced any pattern when they collect data [8]. In this way, using the same model under the coalescence theory researchers will be able to relate the theory to data, so they can use real molecular data to learn more about the populations from which data have been sampled [3].

THE COALESCENCE THEORY

The coalescence theory is based on the Wright-Fisher model but instead of modeling changes of allele frequencies forward in time, this theory considers a sample and the genealogical history of the sample [3]. The coalescence theory describes how population genetic processes determine the shape of the genealogy of sampled gene sequences [2] and allow to describes the relationship between the shape of an intra population genealogy and the demographic history of the sampled population [9].

Looking backward, if we trace the ancestry of alleles in a sample back far enough, all of them will be descended from a single common ancestor [8]. The effective size of the population determine how frequently two (or more) alleles are descended from the same allele in the preceding generation and also the time to find a single common ancestor [8].

While, mathematicians and theoretical population geneticists turned population genetics upside down by introducing the mathematical formulation of coalescence theory Avise and colleagues catalyzed the growth of phylogeography. Coalescence theory formalized a powerful way of using only the sample of alleles such that gene genealogies are modeled backwards in time under virtually any complex demographic history in order to estimate phylogeographic parameters such as historical population sizes, divergence times, and migration rates given the stochastic timing of coalescent events [10].

Simulations as Enhancers of Coalescence Theory

Modern approaches to evolutionary genetics make intensive use of simulation methods motivated by the growth in computational power and data complexity [5].

Coalescent simulations can be used much more broadly than to fit a model to the data. Using simulations is possible to test if simulated data tend to fit the observed data under a particular model. When a model produces simulated data that look similar to the observed data is possible have more faith in it than in a model that does not produce data that look anything like the real data when two models are compared. This basic concept has made coalescent simulations one of the most important computational tools in population genetics [3]. The use of simulations, both as artificial experiments in evolution and inference tools, also has a long tradition in population genetics [5].

PHYLOGEOGRAPHY

The goal of phylogeography is understand the processes that underlie the distribution of genetic variation within and among populations or closely related species [11] by mean of the phylogenetic analysis of organismal data in the context of the geographic distribution of the organism [10]. In the early days, phylogeography was essentially descriptive and not statistical [12]. Nowadays, the statistical rigor of phylogeography has increased because of the advances in coalescence theory which enabled model-based parameter estimation and hypothesis testing [10].

To understand the processes underlying the spatial and temporal dimensions of genetic variation underlies methodological transformations that characterize the phylogeography moving it beyond the original concept of bridging population genetics and phylogenetics [11].

The linkage between geographic and genetic information can certainly bring additional insights on past evolutionary processes such as environmental adaptations, range expansions and migrations [12]. Understanding the evolutionary processes driving patterns of species diversity is critical to knowing the geographic context [13]. While most inferential approaches integrating geography only use information on allele frequencies, coalescent-based approaches integrate molecular information into phylogeographic inferences [12] being coalescence theory the appropriate statistical and methodological framework for testing phylogeographic hypotheses [10]. In this way, phylogeography has become in one of the most integrative fields in evolutionary biology.

NEW APPROACHES IN STATISTICAL PHYLOGEOGRAPHY

Phylogeographic studies are increasingly using simulation-based statistical methods that employ an explicit parameterized coalescent model to estimate parameters as well as test alternative a priori historical hypotheses. In these new statistical phylogeographic

methodologies, the gene genealogy is not the central point of a phylogeographic analysis because of the demographic history is not directly interpreted from the gene genealogy. Instead the gene genealogy is a transition variable for connecting data to demographic parameters under an explicit statistical coalescent model [10]. Detailed estimates of the past, such as population size, growth, subdivision, admixture, patterns of gene flow, and the timing of divergence could be provide by methods that calculate the probability of the data using likelihood-based or Bayesian approaches. These methods become increasingly popular because of their potential to disentangle complex histories (e.g., distinguishing divergence with gene flow from retention of ancestral polymorphism) whereas multiple evolutionary forces have been difficult to distinguish in the past. Such approaches offer extreme flexibility, resolving complex situations that involve combinations of processes (e.g., population divergence, admixture, size change, and gene flow) and any number of populations and samples, while also offering a framework for comparing competing scenarios, estimating parameters and computing bias and precision measures for any given scenario (Figure 1) [11].

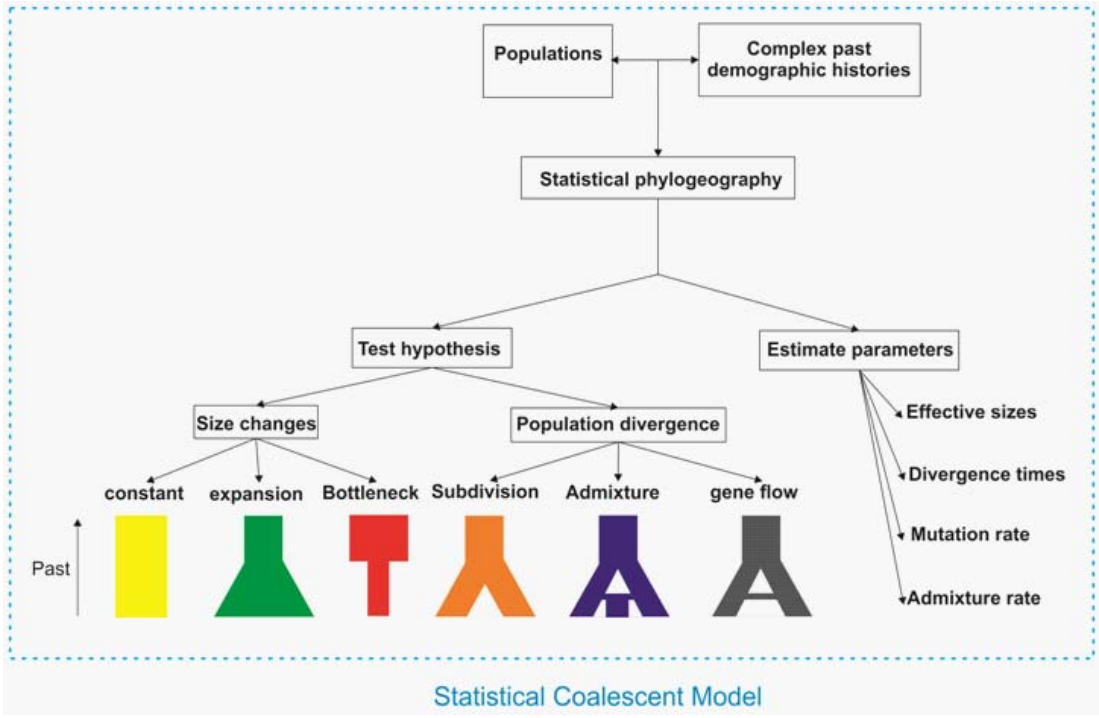

Figure 1: Usefulness of the new methods in statistical phylogeography.

However, make decisions about the appropriate approach to use for phylogeographic analyses is not easy and the knowledge of the biological system under study is essential. Depending on the specifics of each species history and the type of data collected, the decision could be make for one study keeping in mind that what might be an appropriate method for one study might not be for another [11].

MODEL-BASED INFERENCE IN PHYLOGEOGRAPHY

Models are frequent in biology because of they help to describe the observations and also for a given model parameters it is possible to simulate artificial data sets in order to propose hypothesis [14]. Any one competing model is a useful approximation that should show the essential features of a demographic history and it is the most useful if it is robust to violation of model assumptions [10]. Commonly model-based statistical approach, calculate summary statistics from simulated data sets under each model to obtain a distribution of the summary statistic under each respective model. In this way, the probabilities of both models are evaluated with respect to the summary statistic calculated from the empirical data [10].

Most of these complex statistical methods are based on the concept of likelihood (i.e. a function that describes the probability of the data given a parameter) what are based on classical population genetics or coalescence theory using Markov chain Monte Carlo **(MCMC)** techniques.

Approximate Bayesian Computation (ABC)

The difficult for computing the likelihood function restrict their use to simple evolutionary scenarios and molecular models. The development of new methods that approximate the likelihood has been stimulated by the large amounts of data generated by recently developed, high-throughput DNA sequencing technologies and the increasing computational power. One elegant solution is the recent approach Approximate Bayesian Computation **(ABC)** what is a statistical method used to infer parameters and choose between models in the complicated scenarios avoiding the need of a likelihood function [14]. ABC avoids exact likelihood calculations by using summary statistics and simulations. Summary statistics are parameters calculated from the data to represent the maximum amount of information in the simplest possible form [5].

ABC has its roots in the rejection algorithm (i.e. a simple technique to generate samples from a probability distribution) [5]. Also, ABC is based on the idea that models that have a high posterior probability will produce summary statistics closed to those estimated from the observed data [15]. The basic rejection algorithm consists of simulating large numbers of datasets under a hypothesized evolutionary scenario. The parameters of the scenario are sampled from a probability distribution. The data generated by simulation are then reduced to summary statistics, and the sampled parameters are accepted or rejected on the basis of the distance between the simulated and the observed summary statistics. The sub-sample of accepted values contains the fitted parameter values, and allows us to evaluate uncertainty on parameters given the observed statistics [5].

However, the ABC has some objections: i) inference is limited to a finite set of phylogeographical scenarios and ii) arbitrary choice of summary statistics. As a consequence of these, conclusions from ABC could be influenced by subjective inclusion of evolutionary scenarios and implicit model assumptions that are not expected by the modeler [5]. The universe of potential historical

scenarios is huge but only a limited subset of these possibilities is considered. Finally, it is the creativity of the individual researcher, not simply their computational prowess, which determines the insights gained from a phylogeographic analysis [11].

Also, there are two fundamental steps in the modeling and inference process: Compare models and evaluate their goodness-of-fit. Model comparison is often based on a decision theoretical framework and its goal is to choose models receiving high posterior support. On the other hand, the goal of model checking is to understand the ways in which models do not fit the data [5].

Briefly, ABC implementation can be summarized in three main steps: formulating the model, fitting the model to data, and improving the model by checking its fit and comparing it with other models. These steps are strongly inter-dependent and should be considered as a unified approach, with a possibility of cycling through the three stages [16, 5].

Cornuet et al. [17, 16] developed the coalescent based software DIYABC (available from http://www1.montpellier.inra.fr/CBGP/diyabc/) in which a user-friendly interface helps non-expert users to perform historical inferences using ABC. DIYABC is software to make inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data; it allows considering complex population histories including any combination of population divergence events, admixture events and changes in past population size [18]. The most widespread application of ABC is make inferences about demographic history and local adaptation, but applications outside evolutionary biology have already appeared for example in epidemiology and in systems biology [5].

Bayesian Skyline Plots (BSK)

The 'skyline plot' framework, introduced by Pybus et al. [19], is an intuitive visual framework to explore the demographic history of sampled DNA sequences [20] and enables the estimation of historical patterns of population size from a genealogy without the need for a priori restrictions on possible demographic scenarios. A number of methodological extensions have subsequently been described, giving rise to a small family of skyline plot methods. One of these extensions is the Bayesian skyline plot **(BSK)**, which has become in a powerful method for estimating ancestral population dynamics from a sample of molecular sequences [21].

The BSK model uses standard MCMC sampling procedures to estimate a posterior distribution of effective population size through time directly from a sample of gene sequences, given any specified nucleotide-substitution model. A relaxed molecular clock can be employed to accommodate rate variation among lineages. Unlike previous methods, the Bayesian skyline plot includes credibility intervals for the estimated effective population size at every point in time, back to the most recent common ancestor of the gene sequences [2].

In brief, the BSK model has the following characteristics: (i) employs a piecewise-constant model in which the population size is constant in each interval and changes instantaneously between successive intervals; (ii) allows multiple coalescent intervals to be grouped, thereby

reducing the noise associated with short coalescent interval and (iii) requires that the number of groups be chosen a priori [21].

The Bayesian skyline plot is implemented in the phylogenetic software BEAST (available from http://evolve.zoo.ox.ac.uk/beast/) [22]. The Bayesian framework allows the estimation of genealogy and demographic history jointly. Consequently, estimates of model parameters are integrated over phylogenetic and coalescent uncertainty. For this reason, the Bayesian skyline plot can be a useful method even when a specific population parameter is of primary interest and the demographic history is a parameter to be integrated out [23]. Estimation of effective population size, as well as its rate of change through time, can provide useful information about the evolutionary and demographic history of a population [20].

Approximate Bayesian Computation Skyline Plots

The use of the skyline plot in the ABC framework was first proposed in Burgarella et al. [24]. Navascués et al. [25] provide an implementation of the skyline plot within the ABC framework and provide an assessment of its performance for microsatellite data. The flexibility of ABC allows extend it to other type of data, to models with multiple populations, or to other parameters that could change through time. The incorporation of microsatellites to the software BEAST [26] allowed make skyline plot inference for this type of data [25]. The ABC skyline approach is implemented in a suite of R scripts [27] named DIYABCskylineplot [28]. For each simulation the number of population size changes is sampled using the prior probabilities [25].

REMARKABLE CONCLUSIONS

The patterns of DNA sequence variation that exist among present day individuals have record of population history. However, both comprehensive and accurate data and methods to interpret the complicated interplay of forces are required for reading the stories of population history written in DNA sequence variation [15].

Getting useful information about evolutionary processes from population genetic data is hard, and requires appreciation of the mathematical and conceptual of population genetics theory. Rational, quantitative assessment of population genetics models and estimates together with evolutionary scenarios from the spatial distribution of genetic data are unavoidable. It should enable us unravel the complex and subtle relationships between demographic history, natural selection, and genomic diversity [12].

Nowadays, there is a revolution of technological advances that make available to researchers a large amount of molecular data and high computational power for its usefulness in population genetics analyses. This context represents a challenge because of researchers must develop adequate statistical methods. In this way, recent statistical methods in phylogeography such as ABC or BSK have revitalized the population genetics allowing to researchers infer the past evolutionary history of populations and estimate demographic and genetic parameters and evaluate the confidence of that estimations with strength statistical rigor.

ACKNOWLEDGEMENT

M. E. Barrandeguy, D.J. Sanabria and M. V. García wishes to thank to Consejo Nacional de Investigaciones Científicas y Técnicas **(CONICET)**. This work has been funded by grant: PICT 2014 N°2352 from Agencia Nacional de Promoción Científica y Tecnológica **(AGENCIA)** to M.E. Barrandeguy.

References

1. Ewens WJ. Mathematical Population Genetics: Lecture Notes Cornell University. 2006.
2. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent iInference of past population dynamics from molecular sequences. Molecular Biology and Evolution. Molecular Biology and Evolution. 2005; 22: 1185-1192.
3. Nielsen R, M Slatkin. An introduction to Population Genetics. Theory and Applications. Sinauer Associates, Sunderland USA. 2013; 287.
4. Cutter AD. Integrating phylogenetics, phylogeography and population genetics through genomes and evolutionary theory. Molecular Phylogenetics and Evolution. 2013; 1172-1185.
5. Csilléry K, Blum MGB, Gaggiotti OE, Francois O. Approximate Bayesian Computation (ABC) in practice. Trends in Ecology and Evolution. 2010; 25: 410-418.
6. Crisci JL, Poh YP, Bean A, Simkin A, Jensen JD. Recent progress in polymorphism-based population genetic inference. Journal of Heredity. 2012; 103: 287-296.
7. Hartl DL, Clark AG. Principles of population genetics. Sinauer Associates, Inc Publishers, Sunderland. 2007.
8. Holsinger K. Lecture Notes in Population Genetics. Creative Commons License. Stanford, California 94305, USA. 2010.
9. Kingman JFC. The coalescent. Stochastic Processes and their Applications. 1982; 13: 235-248.
10. Hickerson MJ, Carstens BC, Cavender-Bares J, Crandall KA, Graham CH, et al. Phylogeography's past, present, and future: 10 years after Avise, 2000. Molecular Phylogenetics and Evolution. 2010; 54: 291-301.
11. Knowles LL. Statistical Phylogeography. Annu. Rev. Ecol. Evol. Syst. 2009; 40: 593-612.
12. Beaumont MA, Nielsen R, Robert C, Hey J, Gaggiotti O, et al. In defence of model-based inference in phylogeography. Molecular Ecology. 2010; 19: 436-446.
13. Knowles L and Carstens BC. Estimating a geographically explict model of population divergence. Evolution. 2007; 61: 477-493.
14. Beaumont MA. Approximate Bayesian Computation in Evolution and Ecology. Annual Review of Ecology, Evolution, and Systematics. 2010; 41: 379-406.
15. Schraiber JG, Akey JM. Methods and models for unravelling human evolutionary history. Nature Reviews Genetics. 2015; 16: 727-740.
16. Cornuet JM, Ravigné V, Estoup A. Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1.0). BMC Bioinformatics. 2010; 11: 401.
17. Cornuet JM, Santos F, Beaumont MA, Robert CP, Marin JM, et al. Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. Bioinformatics. 2008; 24: 2713-2719.
18. Cornuet JM, Pudlo P, Veyssier J, Dehne-Garcia A, Gautier M, et al. DIYABC v2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. Bioinformatics Applications Note. 2014; 30: 1187-1189.
19. Pybus OG, Rambaut A, Harvey PH. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics. 2000; 155: 1429-1437.
20. Strimmer K, Pybus OG. Exploring the Demographic History of DNA Sequences Using the Generalized Skyline Plot. Mol. Biol. Evol. 2001; 18: 2298-2305.
21. Ho SY, Shapiro B. Skyline-plot methods for estimating demographic historyfrom nucleotide sequences. Molecular Ecology Resources. 2011; 11: 423-434.
22. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology. 2007; 7: 214.

23. Burgarella C, Navascués M, Zabal-Aguirre M, Berganzo E, Riba M, et al. Recent population decline and selection shape diversity of taxol-related genes. Molecular Ecology. 2012; 21: 3006-3021.

24. Navascués M, Leblois R, Burgarella C. Demographic inference through approximate-Bayesian-computation skyline plots. 2017.

25. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS Biology. 2006; 4: e88.

26. Wu CH, Drummond AJ. Joint inference of microsatellite mutation models, population history and genealogies using transdimensional Markov Chain Monte Carlo. Genetics. 2011; 188: 151-164.

27. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2017.

28. Navascués M. DIYABCskylineplot v1.0.1: A suite of R scripts to perform a Approximate Bayesian Computation Skyline Plot. 2017.