Advances in Computer Simulation of Genome Evolution: Toward More Realistic Evolutionary Genomics Analysis by Approximate Bayesian Computation

Miguel Arenas

Abstract NGS technologies present a fast and cheap generation of genomic data. Nevertheless, ancestral genome inference is not so straightforward due to complex evolutionary processes acting on this material such as inversions, translocations, and other genome rearrangements that, in addition to their implicit complexity, can co-occur and confound ancestral inferences. Recently, models of genome evolution that accommodate such complex genomic events are emerging. This letter explores these novel evolutionary models and proposes their incorporation into robust statistical approaches based on computer simulations, such as approximate Bayesian computation, that may produce a more realistic evolutionary analysis of genomic data. Advantages and pitfalls in using these analytical methods are discussed. Potential applications of these ancestral genomic inferences are also pointed out.

Keywords Population genetics · Computer simulations · Approximate Bayesian computation · Genome evolution · Molecular evolution

Next generation DNA sequencing techniques have lead to fast and cheap generation of sequence genomes. Moving forward though, these genomes should be properly analyzed to test molecular evolutionary hypotheses of interest (Smith 2013). Common analyses of genome data involve the estimation of genetic statistics such as genetic diversity and genetic differentiation (e.g., Abecasis et al. 2012). However, if we are interested in ancestral genome inferences, these are not so straightforward because the need to jointly consider a variety of genomic events of structural variation such as duplications, insertions, deletions, inversions, and translocations of genomic regions, or other phenomena such as gene–gene interactions. These processes determine the evolutionary history of genomes and despite the fact that they have been widely studied (e.g., Chain and Feulner 2014), they are still very challenging to implement jointly in current analytical methods. For example, it is known that the computation of a likelihood function based on a relatively complex model of evolution can be intractable, thus restricting the use of likelihood-based inference to simple evolutionary scenarios and models (e.g., Marjoram et al. 2003; Wegmann et al. 2009). Therefore, in order to deal with complex evolutionary models, statistical approaches based on computer simulations such as approximate Bayesian computation (ABC) (e.g., Beaumont 2010; Sunnaker et al. 2013), that avoid the need for a likelihood function, are being established. These methods provide promising alternative analytical strategies and can generate very accurate inferences because of their joint consideration of different evolutionary processes. For example, they have already outperformed approximate maximum likelihood methods based on more approximate models (Lopes et al. 2014; Arenas et al. 2015). Nevertheless, ABC approaches definitely require extensive computer simulation with evolutionary frameworks that must be able to model the evolutionary process in a way that is as realistic as possible.

This letter provides an overview for the application of computer simulation of complex genome evolution to evolutionary genomics. Recent advances in the modeling of complex genome evolution and its implementation in
state-of-the-art computer simulators are described first. Next, a methodology based on the ABC approach to properly perform ancestral genomic inferences accounting for complex genome evolution is proposed. Advantages and limitations are discussed, and lastly, a variety of potential applications are suggested.

In the last few years, some sophisticated computer simulators of genome evolution have emerged. One of them is the evolutionary framework ALF (Dalquen et al. 2012), which can simulate genome evolution accounting for gene duplication and loss, gene fusion and fission, lateral gene transfer (LGT), genome rearrangement, and speciation under a birth-death process (e.g., Dalquen et al. 2013). In addition, this framework can simulate genome evolution under a heterogeneous substitution process wherein each genomic region can evolve under a particular substitution model. SGWE (Arenas and Posada 2014) is another sophisticated simulator of genome evolution that implements homogeneous and heterogeneous (hot-spots and cold-spots) recombination, along with different gene histories into a species tree taking into account complex demographics under a coalescent-based approach. In a similar way to ALF, SGWE allows for heterogeneous substitution along the genome. Another interesting contribution was recently presented by Peischl et al. (2013). These authors presented the modeling of large chromosomal inversions in large populations through a sequential coalescent approach with recombination. For example, this simulation tool can be useful to explore the effects of polymorphic inversions on patterns of recombination.

Of course there are complex processes acting on genome evolution that have not yet been modeled, such as complex gene–gene interactions, coevolution, and heterogeneous selective pressure along the genome (e.g., Makino and McLysaght 2008; Blanc et al. 2010).

The incorporation of these complex models of genome evolution in ABC techniques is not straightforward because different genomic regions may have evolved under different evolutionary histories (e.g., Larkin et al. 2009) and may present a region-specific best-fit evolutionary model (Arbiza et al. 2011). Thus, the first stage of an ABC methodology oriented to analyze genomic data accounting for complex genome evolution may involve the identification of those genomic regions that could have evolved under a specific evolutionary history and evolutionary process (Fig. 1). A variety of comparative genomics tools have been designed to perform this analysis (e.g., Baudet et al. 2010; Skovgaard et al. 2011). In the second stage, summary statistics should be designed and computed at the local level (for each genomic region) and at the global level (whole genome) (see Fig. 1). Additional summary statistics could provide information about genetic differences among genomic regions (e.g., pairwise genetic differentiation) that can be relevant to describe genomic heterogeneity. Similarly, computer simulations might be also performed with different prior distributions to local and global parameters. This aspect was already devised by Arenas and Posada (2014) with the implementation of region-specific prior distributions (i.e., each genomic region can evolve under a particular model whose parameters are based on a particular set of prior distributions) and genome-specific prior distributions (e.g., priors for the global substitution and recombination rates) in their evolutionary framework SGWE (see above). Actually, these authors found that the consideration of local processes (i.e., variable codon frequencies across codon positions) can dramatically affect evolutionary estimates such as the ratio of non-synonymous to synonymous substitution rates (Arenas and Posada 2014) and therefore, the consideration of region-specific evolutionary scenarios can be fundamental for the evolutionary analysis. Next, as any other ABC method, the same summary statistics used to extract the genetic information from the real data must be used to extract the information of the simulated data (Fig. 1). Then, posterior distributions can be computed to evaluate candidate models (e.g., models ranging from low to high genomic heterogeneity or models with/without a particular genome rearrangement) and to estimate the evolutionary parameters of interest (Fig. 1) while accounting for complex genome evolution.

However, the application of ABC techniques to analyze complex genome evolution may present severe technical limitations. One of them is that the complexity of genome evolution—with a large number of evolutionary processes and parameter space to be explored—may require a huge amount of computer simulations. Marjoram and Tavaré (2006) proposed to capture the essential features of the evolutionary process and simplify the models accordingly in order to eliminate unnecessary parameterization. In this concern, efforts can be required to develop models of complex genome evolution that remain workable. Storing and accessing genomes are additional current problems when dealing with genomic data (Kahn 2011). Hopefully these technical limitations will be solved with the evolution of informatics frameworks and available storage.

The consideration of complex genome evolution in the evolutionary analysis can be applied to a variety purposes. For example, it can be used to identify local evolutionary processes such as selection regimens (which can vary among genomic regions), to allow ortholog prediction in the presence of events such as LGT, or to improve current methods of gene/species tree reconciliation (e.g., Dalquen et al. 2012). Moreover, genomic events can play a significant role in molecular adaptation and speciation phenomena (e.g., Lawrence 1999; Barrick et al. 2009), and they have also been associated with genetic diseases (Weischenfeldt et al. 2013).

With the advent of NGS technologies, the quantity and complexity of molecular data increase. This leads to a strong demand for robust analytical frameworks of genome
Fig. 1 Illustrative example of an ABC methodology oriented to the evolutionary analysis of genomic data. Real data consist of a set of genome sequences (upper rectangle with lines inside). Five genomic regions are shown with different colors. The upper section involves the direct analysis of the real data with the detection of genomic regions and the computation of the summary statistics for each genomic region (SSRg) and for the whole genome (SSRG). The middle section involves the application of the prior distributions to get values for the parameters of every genomic region (PrDg) and whole genome (PrDG). Next, computer simulations can be performed according to the prior distributions and then, summary statistics can be computed for every simulated data set. The lower section illustrates the computation of the posterior distributions PoDg for every parameter of interest (pi) (corresponding prior distributions are shown in gray). Note that some parameters could be used for simulating data but not to be estimated (nuisance parameters) and therefore p and pi can be different.

evolution that account for complex genomic events. On the other hand, advances and applications of the ABC approach have been amazing in recent years as a consequence of the analysis of complex problems and the emergence of more sophisticated methods. Important advances in combining ABC strategies and complex genome evolution are long awaited and are likely to have a significant impact on ancestral genome inferences.

Acknowledgments I thank the Editor for his detailed comments. This work was supported by the Spanish Government through the “Juan de la Cierva” fellowship JCI-2011-10452.

Conflict of interest The author declares that there is no conflict of interests.

Compliance with Ethical Standards This study does not involve research with humans and/or animals and it follows all the ethical standards.

References


Smith DR (2013) Death of the genome paper. Front Genet 4:72