Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Genetics & Development

Fine-tuning of Approximate Bayesian Computation for human population genomics

Niall P Cooke and Shigeki Nakagome

Approximate Bayesian Computation (ABC) is a flexible statistical tool widely applied to addressing a variety of questions regarding the origin and evolution of humans. The significant growth of genomic scale data from diverse geographic populations has facilitated the use of ABC in modelling the complex processes that underlie human demography and local adaptation. However, a fundamental issue still remains in how to efficiently capture patterns of genetic variation with a set of summary statistics in order to achieve better approximation of Bayesian inference. Here, we review recent advances in ABC methodology and its applications for human population genomics, with a particular focus on optimal tuning of ABC approaches for different types of genetic data and different sets of evolutionary parameters.

Address
Trinity Translational Medicine Institute, School of Medicine, Trinity College Dublin, Dublin, Ireland

Corresponding author: Nakagome, Shigeki (nakagoms@tcd.ie)

Current Opinion in Genetics & Development 2018, **53**:60–69

This review comes from a themed issue on **Genetics of human origins**

Edited by **Lluis Quintana-Murci** and **Brenna Henn**

https://doi.org/10.1016/j.gde.2018.06.016

0959-437X/© 2018 Elsevier Ltd. All rights reserved.

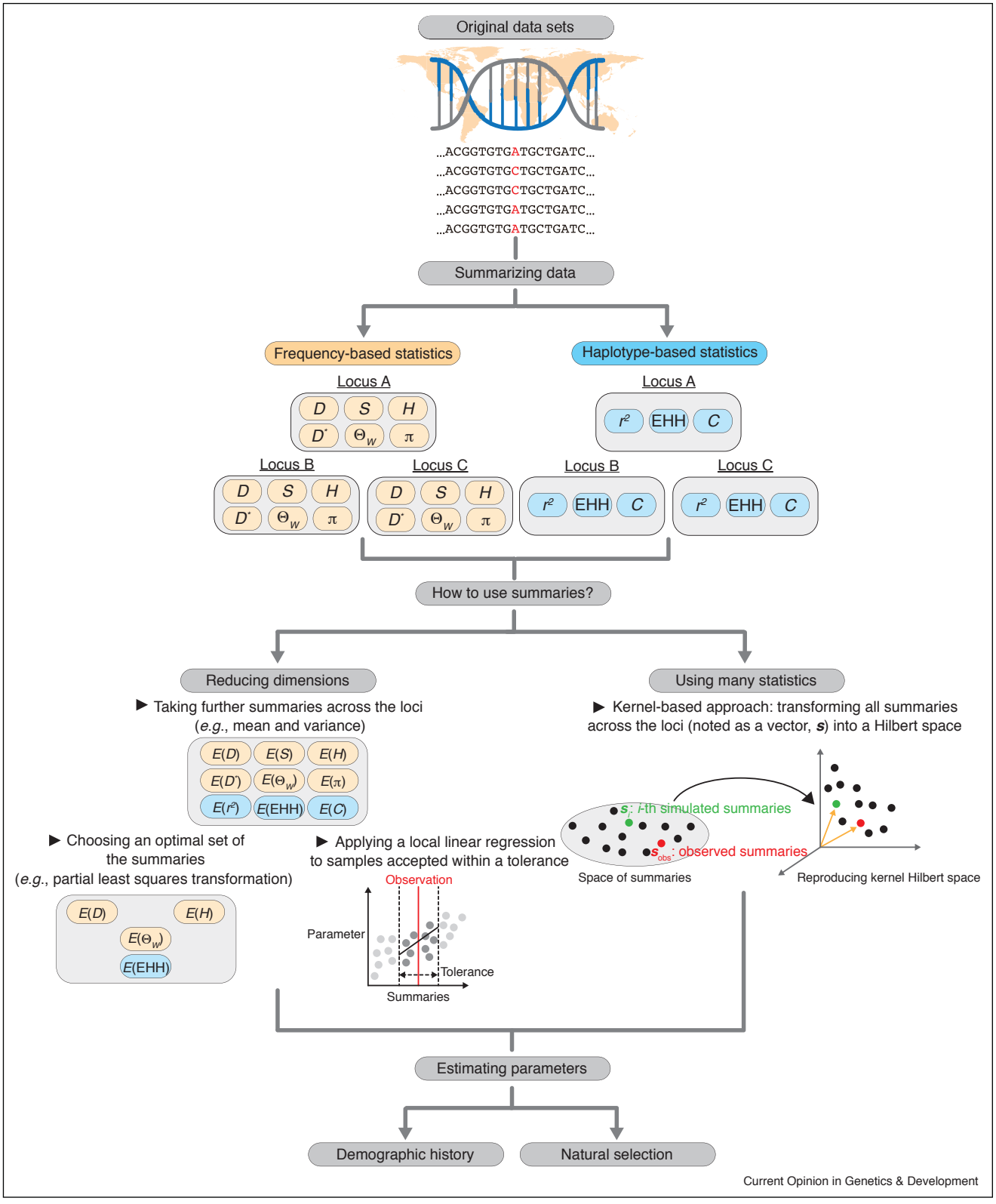
Introduction

Statistical inference on genetic variation provides powerful tools for unravelling the evolutionary history of human populations [1–3]; however the complexity of realistic models underlying the history, as well as of genomic data obtained from diverse geographic populations, makes it challenging to formulate statistical frameworks. One of the leading techniques to address these obstacles is Approximate Bayesian Computation (ABC). ABC works by simulating data and quantifying the fit between observed and simulated data, and therefore there is no need to define and compute the likelihood function (i.e., the probability of the data for parameters of interest) [4,5]. A key feature of ABC-based methods is to capture the characteristics of genetic data (e.g., DNA sequences in a population) through summary statistics (e.g. the number of segregating sites) and to use this simplified information to make analytical approximation of the posterior probability distribution of a parameter of interest. Because it is not easy to find statistics that are sufficiently informative with respect to the parameters in population genetics inference, constructing summary statistics is a crucial step to minimizing the loss of information and achieving better approximations to the posterior estimates given the full set of genetic data. In this review, we focus on the fine-tuning of ABC approaches for the study of human population genomics. We first present an overview of the recent advances in ABC techniques for overcoming the curse of dimensionality in summary statistics (Figure 1). We then describe its applications in the inference of human demographic history and local adaptation. Finally, we conclude with examples of a potential pitfall in summarizing data and future improvements of ABC.

Recent advances of ABC techniques

The generic principle of ABC is to draw samples of parameters from a posterior distribution through two steps of approximation. The complete set of genetic data observed at a population scale is complex and of high dimension, thus making it infeasible to find simulated data that are identical to the observed data using a basic rejection-sampling algorithm [6,7]. The first step of approximation aims to reduce the dimensionality of genetic data using summary statistics. Although it is theoretically optimal to only use sufficient statistics that are both low dimensional and best represent the information on the data, the availability of such statistics is limited to simple models or particular types of genetic data (e.g., a number of alleles used to estimate a scaled mutation rate under the infinite-allele model in which each mutational event creates a new allele that has not been present in a population [8]). In most cases, it is required to use as many summary statistics as possible to be able to capture much information contained within the data; however, this makes it difficult to find the exact match between simulated and observed summaries. Therefore, the second approximation is introduced by relaxing the condition to accept parameters wherein summaries of simulated data are not the same but rather similar to those of the observed data. The cut-off as to whether or not to reject these parameters is defined with a tolerance that measures the similarity as a distance between the two sets of summaries (e.g., an Euclidean or a Chebyshev distance). The posterior distribution from which accepted parameters are sampled provides the probabilities of different parameter values that produced

Figure 1



Schematic diagram of ABC frameworks for the inference of human evolutionary history. Large scale genomic data (i.e., sequencing or genotyping) become available across diverse human populations. Such original data sets are summarized into two different types of population genetics statistics: frequency-based and haplotype-based statistics. In this example, patterns of genetic variation in three independent loci are captured by

Download English Version:

https://daneshyari.com/en/article/8625622

Download Persian Version:

https://daneshyari.com/article/8625622

Daneshyari.com