Customize and get the most out of your reduced-representation sequencing experiment with the new simulation software RADinitio

Choquet, M.

Published in: Molecular Ecology Resources
DOI: 10.1111/1755-0998.13218

Available online: 30 Jun 2020

Title:

Customize and get the most out of your reduced-representation sequencing experiment with the new simulation software *RADinitio*

Author:
Marvin Choquet*
*Faculty of Biosciences and Aquaculture, Nord University, Bodø, Norway

ORCID:
Marvin Choquet https://orcid.org/0000-0001-6719-2332

Correspondence:
Dr Marvin Choquet
Faculty of Biosciences and Aquaculture, Nord University
Campus Mørkved, 8049 Bodø, Nordland, Norway
marvin.choquet@nord.no

Whole genome sequencing is still often a difficult, costly and time-consuming task. The emergence of various genome reduced-representation sequencing (RRS) protocols such as restriction site-associated DNA sequencing (RADseq) has facilitated the access to genome-wide information, without the need for whole genome sequencing. Reaching the full potential of RRS protocols though, requires adjustments and tailoring to the species under investigation. To that end, simulation software have been developed to guide researchers in the customization of their RADseq experiment, but the extent to which these tools mimic the behavior of a protocol in generating sequencing data is limited. In this current issue of *Molecular Ecology Resources*, Rivera-Colón et al. (2020) introduce *RADinitio*, a new software for simulating RADseq data designed to perform simulations at the highest level of representativeness. By taking into account the effects of library preparation and sequencing parameters on the resulting sequences, *RADinitio* allows the precise identification of the sources of failure when designing a RADseq experiment. This new software represents a considerable advance in RADseq data simulation and will likely lead to increased success in RADseq experiments.

Keywords: Genotyping-by-sequencing, RADseq, population genomics, next-generation sequencing, non-model species

Over the last decade, the rapid development of genome reduced-representation sequencing (RRS) protocols has revolutionized the fields of molecular ecology, evolutionary genetics and conservation genetics. With diverse procedures, RRS protocols aim to reduce the complexity of a genome by sampling a fraction of it, sufficient to address various biological questions, and much easier to sequence and analyze compared to a whole genome. These protocols offer a solution for exploring population genomics at a reasonable cost in model organisms and in non-model organisms often left understudied due to their genome complexity.

Restriction site-associated DNA sequencing (RADseq) is currently one of the most popular approaches and consists in using restriction enzyme(s) to digest a genome, followed by the sequencing of restriction sites flanking regions. The appeal of this method relies on its applicability to theoretically any type of organism, with or without prior genomic resources associated (Davey & Blaxter, 2010). A variety of RADseq protocols have been described (Andrews et al, 2016) and in spite of reviews assessing pros and cons related to the use of each method (e.g. Andrews & Luikart, 2014), figuring out the most fitting protocol for a new experiment may still be challenging. Standardized protocols exist though, such as the ezRAD (Toonen et al, 2013), and those are often perceived as attractive due to their reported experimental ease of implementation. However, the relevance of data yielded by any protocol, even standardized, to answer a biological question cannot be ensured, except through prior prospective data simulation.

The article by Rivera-Colón et al. (2020) in the current issue of *Molecular Ecology Resources* introduces the new simulation software *RADinitio.* The function of *RADinitio* is to simulate behaviors of different RADseq protocols and parameters of library preparation on a particular species in generating sequencing data. Thereby the user can make informed decisions on how to tailor a specific RADseq experiment. The software *RADinitio* can be downloaded from https://pypi.org/project/radinitio/ and is used via a command-line interface. Its development emanated from the authors' view that two sources of error often impede researchers from reaching the full potential of success with their RADseq experiment. The first challenge relies in the selection of a protocol (e.g. single-digest or double-digest RADseq) that may be suboptimal for a specific organism. The second challenge lies in the processes of library preparation (i.e. quality of starting template, choice of restriction enzyme) and sequencing (i.e. coverage given). Other simulation software were developed in the past, but *RADinitio* represents an important step forward in RADseq data simulations due to its capacity to generate variants actually relevant for population genetic studies and due to the inclusion of parameters of library preparation and sequencing in its simulations.

Rivera-Colón et al. (2020) describe *RADinitio* as a three-step pipeline. The user first needs to feed the software with a reference sequence that will be used to simulate a metapopulation. Genetic variants will then be generated via the coalescent simulator *msprime* (Kelleher et al, 2016) following a demographic model that can be defined by the user to match as precisely as possible the study system. In the case when the user does not know the details of the underlying model, *RADinitio* uses reasonable default parameters. The second step consists in digesting *in silico* the reference genome provided by the user with either one or two restriction enzymes, selected by the user, thereby generating a series

of RAD loci across the genome. This set of loci is intersected with the set of genetic variants so as to keep only the variants present within RAD loci. At this step, *RADinitio* reproduces a very famous characteristic of RADseq by taking into account the possibility of mutations at restriction sites (following mutation and recombination rates defined by the user), and subsequently saving only sequences with intact cut-sites, thus simulating the allele dropout effect. Thirdly, paired-end sequences are generated from the extracted pool of RAD alleles. Again, *RADinitio* reflects what happens in reality by including read duplicates to imitate the effect of PCR amplification, following the number of PCR cycles defined by the user, which allows exploring the impact of performing more or less PCR cycles. Sequences are produced according to a sequencing coverage determined by the user and random sequencing errors are added to each individual read pair, mirroring Illumina sequencing patterns of error rates.

The level of detail thought by the authors to make *RADinitio* mimic the effects of several variables (via parameters defined by the user) on data generated by a RADseq experiment brings RADseq simulation software to a new level. By considering allele dropout resulting from natural polymorphism in *RADinitio* simulations on a set of species, Rivera-Colón et al. (2020) were able to characterize for the first time the degree of contribution of library preparation and sequencing parameters to total allele dropout. They found that using a bad template quality and a suboptimal sequencing coverage influence much more the amount of allele dropout compared to the effect of natural polymorphism at enzyme restriction sites. The software *RADinitio* is therefore the first of its kind that will guide users in keeping allele dropout to a minimum.

As mentioned above, the usefulness of *RADinitio* depends on the availability of a reference sequence, which may or may not be available depending on the species under investigation. For a model species where a genome is available, the process is straightforward (Fig. 1). By testing different genomes with different conditions, Rivera-Colón et al. (2020) concluded that *RADinitio* simulations resemble empirical RAD loci. The user can then confidently use the *RADinitio --tally-rad-loci* command to calculate the number of genetic variants expected to be yielded in a RADseq experiment depending on the protocol and parameters chosen. Rivera-Colón et al. (2020) also tested the scenario of a non-model species with no reference genome available, using the example of a salmonid fish. In that situation, the authors concluded that *RADinitio* can still perform informative simulations under the condition that several genomes from a variety of related species of different evolutionary distances are used as input (Fig. 1). There are, however, numerous cases of non-model organisms for which not even one closely related genome is available, making tailoring of a RADseq experiment very difficult. This is particularly striking in the field of marine zooplankton, left largely ignored by genomics, where despite the huge diversity of organisms reported, only a very few genomes are published (Bucklin et al, 2018). One alternative may thus be to generate sequencing data from the species of interest prior to the RADseq experiment in order to feed *RADinitio* with at least some genomic data (Fig. 1). The inconvenient of such practice is that it may be difficult to assess how much data will be necessary for the simulations to be accurate in regard to the actual genome of the target species, and *RADinitio* was not tested for this. Besides, the success of a RADseq experiment cannot be guaranteed if data simulation is not feasible or not accurate, in which case other RRS protocols may need to be considered or developed. A case study performed on the

non-model zooplankton species *Calanus finmarchicus*, known for its large genome, illustrates the potential challenges linked to the absence of a reference genome when trying to simulate RADseq data (although *RADinitio* was not available back then) (Choquet et al, 2019). In that study, a specific RRS protocol relying on target capture had to be developed to achieve generation of variants despite starting with no reference genome, via the sequencing of a draft-transcriptome instead.

To conclude, the new simulation software *RADinitio* is a promising tool that should be used before starting any RADseq experiment in species where a reference genome is available, or at least several closely related genomes. The level of representativeness implemented in *RADinitio* simulations will help users customize their experiment to get the most out of it. This pipeline represents a substantial advance for the field of RRS and particularly for RADseq users.

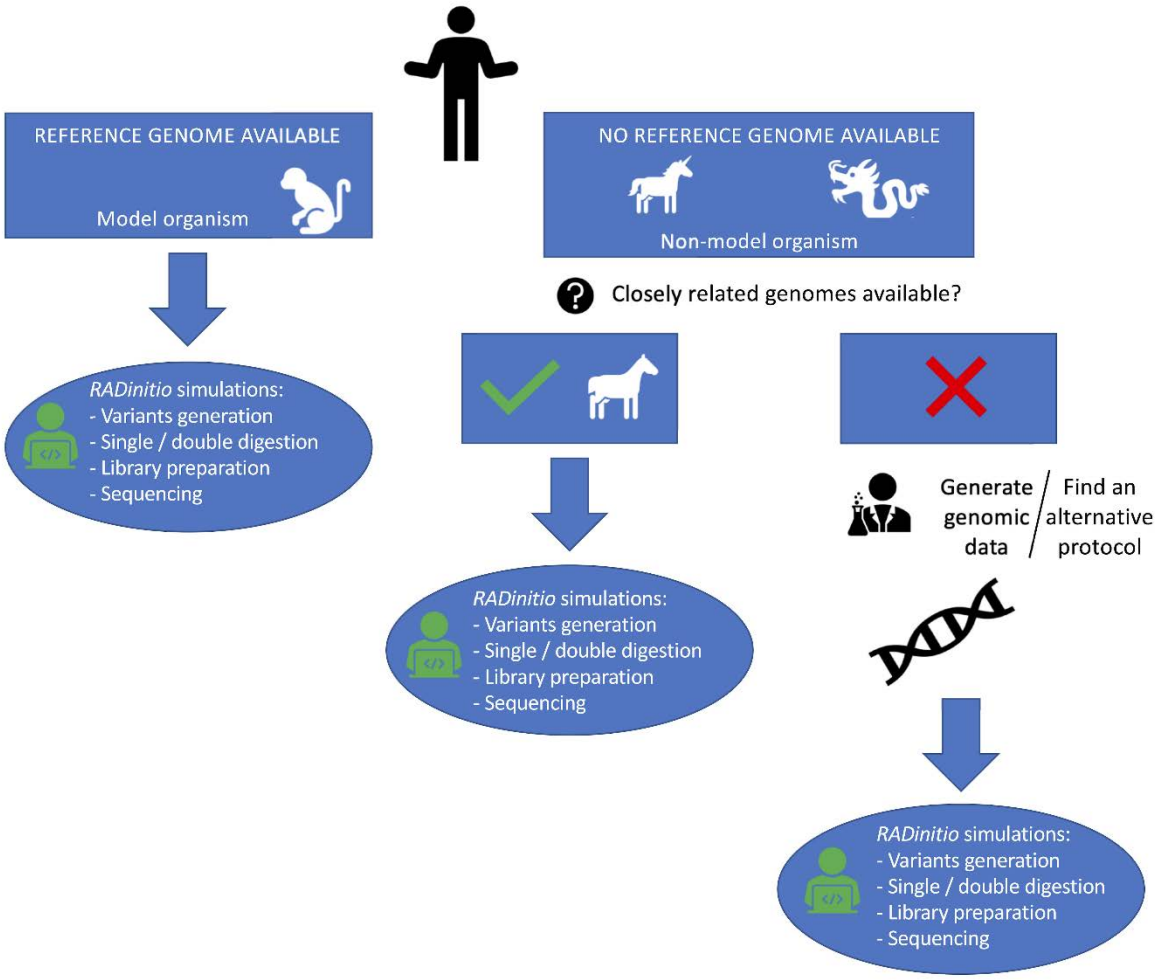

Figure 1: Different scenarios when starting a RADseq experiment

References:

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, **17**, 81.

Andrews KR, Luikart G (2014) Recent novel approaches for population genomics data analysis. *Molecular Ecology*, **23**, 1661-1667.

Bucklin A, DiVito KR, Smolina I *et al.* (2018) Population genomics of marine zooplankton. In: *Population Genomics: Marine Organisms* (ed. Rajora Om P & Oleksiak MF), pp. 61-102. *Springer*.

Choquet M, Smolina I, Dhanasiri AK *et al.* (2019) Towards population genomics in non-model species with large genomes: a case study of the marine zooplankton *Calanus finmarchicus*. *Royal Society Open Science*, **6**, 180608.

Davey JW, Blaxter ML (2010) RADSeq: next-generation population genetics. *Briefings in functional genomics*, **9**, 416-423.

Kelleher J, Etheridge AM, McVean G (2016) Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, **12**.

Toonen RJ, Puritz JB, Forsman ZH *et al.* (2013) ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ*, **1**, e203.