How to estimate clonality from genetic data: use large samples and consider the biology of the species

Myriam Heuertz

Peer Community In
Evolutionary Biology

How to estimate clonality from genetic data: use large samples and consider the biology of the species

Myriam Heuertz based on reviews by David Macaya-Sanz, Marcela Van Loo and 1 anonymous reviewer

Open Access

A recommendation of:

Solenn Stoeckel, Barbara Porro, Sophie Arnaud-Haond. **The discernible and hidden effects of clonality on the genotypic and genetic states of populations: improving our estimation of clonal rates** *(2019), arXiv, 1902.09365, ver. 4 peer-reviewed and recommended by Peer Community in Evolutionary Biology.* https://arxiv.org/abs/1902.09365v4

Published: 30 July 2019

Submitted: 28 February 2019, Recommended: 12 August 2019

Cite this recommendation as:

Myriam Heuertz (2019) How to estimate clonality from genetic data: use large samples and consider the biology of the species. *Peer Community in Evolutionary Biology, 100078.* 10.24072/pci.evolbiol.100078

Population geneticists frequently use the genetic and genotypic information of a population sample of individuals to make inferences on the reproductive system of a species. The detection of clones, i.e. individuals with the same genotype, can give information on whether there is clonal (vegetative) reproduction in the species. If clonality is detected, population geneticists typically use genotypic richness R, the number of distinct genotypes relative to the sample size, to estimate the rate of clonality c, which can be defined as the proportion of reproductive

events that are clonal. Estimating the rate of clonality based on genotypic richness is however problematic because, to date, there is no analytical, nor simulation-based, characterization of this relationship. Furthermore, the effect of sampling on this relationship has never been critically examined. The paper by Stoeckel, Porro and Arnaud-Haond [1] contributes significantly to the characterization of the relationship between rate of clonality and genetic and genotypic parameters in a population. The authors use an extensive individual-based simulation approach to assess the effects of rate of clonality (fully sexual, fully clonal and a range of intermediate levels of clonality, i.e., partial clonality) on genetic and genotypic parameters, considering variable population size, sample size, and numbers of generations elapsed since population initiation. Based on their simulations, they derive empirical formulae that link for the first time the rate of clonality to the genotypic richness and to the size distribution of clones (genotypic parameters), as well as to the population inbreeding coefficient and to a metric of linkage disequilibrium (genetic parameters). They then use the simulated data to assess the accuracy of their predictions. In a second phase, the authors use a Bayesian supervised learning algorithm to estimate rates of clonality from the simulated data. The authors show that the relationship between rate of clonality and genotypic richness is not linear: genotypic richness decreases slowly with increasing clonality, a large drop in genotypic richness is only seen for rates of clonality ≥ 0.90. Genetic parameters are only sensitive to high rates of clonality. The practical implications of these results are that genotypic and genetic parameters can complement each other for the estimation of rates of clonality, with genotypic parameters most useful throughout most of the range of clonality values and with genetic parameters complementing them meaningfully at higher values. The most meaningful practical result of the paper is the demonstration of sampling bias on the estimation of genotypic richness. Commonly used population sample sizes in population genetics studies ($n \leq 50$) lead to great overestimation of genotypic richness, which consequently leads to a severe underestimation of the rate of clonality in most systems, irrespectively of whether they have reached stationary equilibrium. Only in small populations, these effects are attenuated. Biologists interested in the estimation of the rate of clonality will find this paper highly useful to design their sampling, and to choose their statistics for inference in a meaningful way. This paper also calls for a careful reappraisal of previously published works that infer rates of clonality from genetic data, and highlights the prime importance of complementary information on species life history data for a correct understanding of partial clonality.

References

[1] Stoeckel, S., Porro, B., and Arnaud-Haond, S. (2019). The discernible and hidden effects of clonality on the genotypic and genetic states of populations: improving our estimation of clonal rates. ArXiv:1902.09365 [q-Bio] v4 peer-reviewed and recommended by Peer Community in Evolutionary Biology. Retrieved from http://arxiv.org/abs/1902.09365v4

Revision round #2

2019-07-21

Dear authors, Thank you for sending your revised manuscript to PCI for evaluation. I have now gone through your reply to editor (recommender) and reviewers and I have inspected the manuscript. I am overall satisfied with your replies and revisions. Thank you for adding the box with definitions, for investigating the effect of rate of clonality on genetic and genotypic parameters in early generations of your simulations, i.e., when a stationary equilibrium is not reached yet, and for the improvement of figures. At this stage I have no further comments on the science you present. However, before I can formally recommend your paper, I have some concerns related to the presentation of your work that should be addressed. My decision is thus “Revise”.

1. English language

In my first inspection of the manuscript I have not commented on the use of English language, although the reviewers highlighted some places where the language was not clear. In my reading of the revised version, I found that that problems with the use of English language persist, especially in the introduction where the quality of language is below that of the rest of the manuscript. Sentences are often too long or wordy and unprecise, with some problems in syntax or grammar that negatively affect their comprehension. Please revise the language of your paper carefully, preferably using a professional language editing service. I give some examples, mostly in the the introduction where I found the problems most striking. I did not comment on the language throughout the manuscript: P3L5-6: the link between the “dynamics and evolution” of PC and “ecosystems and human health” is not clear for the reader, it relies on a non-explicit shortcut. Please revise. As I understand it, it’s the evolutionary trajectories of PC species that compose ecosystems (or the human microbiome) that can have effects on ecosystem functioning (or human health). P3: “PC is rather well

represented in engineers…" this refers to people who do engineering. Revise. P3: "(PC) has been shown to increase in challenging environments": this is not clear. PC is a mode of reproduction, which a species has, or does not have. Thus, PC is present or absent, it cannot increase. Instead, rates of clonality, or the proportion of species with PC, can increase. Please revise. Box1 "Clonality" is the property of being clonal. The definition you are giving is the definition of clonal, not of clonality. Revise (e.g., "an individual is defined to be clonal if it produces…"). Partial clonality: "through selfing and outcrossing". I think it would be more appropriate to say "through selfing or outcrossing" (the latter expression is quoted on top of P4). P4: revise to "the ability of a given genotype to persist"; "three main knowledge gaps"; "is not obviously inferred from classical…". Do not use "partial clonals" but partially clonal species. P5: "We still face difficulties…., preventing access" Wordy sentence. It is not clear what exactly prevents what. Revise. P5: "Indirect reconstruction" of what? This paragraph is on tracking clonal spread /determination of clonal identity, which is needed to estimate the rate of clonality. Revise. P5-P6: I suggest breaking this sentence into two parts: "… genotypic (clonal) richness. Genotypic richness is often assumed to…". P6L6: at lower rates (of clonality). P6L14: Revise to "to conclude on a neglible…" P6L16: "are overlooked in terms of clonality": a shortcut is made, revise (e.g., "are not interpreted in relationship to clonality"). P6 bottom: Say precisely which families of parameters. P10: mention "Pareto". "Pareto" appears in Results without explicit connection to Materials and Methods. P12. Formulas, or formulae.

2. Consistent and explicit use of terminology

Please use your terminology consistently and explicitly, especially parameter names. E.g., P16, title, what is meant precisely with "Evolution of genotypic states": you are describing genotypic richness and distribution of clonal size at equilibrium under an increasing rate of clonality. First sentence, if you refer to both R and Pareto β, please say so; avoid mixing the use of "parameter β", "β-values", "Pareto β". Figure legends: please make sure that all parameters are properly identified and named, and that links with parameters in the figures are made. Readers often have a fast glance at figures first, so this is important. Figure 1, the link between "Rate of clonality" in the Figure and "values of c" in the legend is not made. Please improve the legend. Figure 3. Subscripts appear cut off in the figure.

Additional requirements of the managing board:

As indicated in the 'How does it work?' section and in the code of conduct, please make sure that: -Data are available to readers, either in the text or through an open data repository such as Zenodo (free), Dryad or some other institutional repository. Data must be reusable, thus metadata or accompanying text must carefully describe the data. -Details on quantitative analyses (e.g., data treatment and statistical scripts in R, bioinformatic pipeline scripts, etc.) and details concerning simulations (scripts, codes) are available to readers in the text, as appendices, or through an open data repository, such as Zenodo, Dryad or some other institutional repository. The scripts or codes must be carefully described so that they can be reused. -Details on experimental procedures are available to readers in the text or as appendices. -Authors have no financial conflict of interest relating to the article. The article must contain a "Conflict of interest disclosure" paragraph before the reference section containing this sentence: "The authors of this preprint declare that they have no financial conflict of interest with the content of this article." If appropriate, this disclosure may be completed by a sentence indicating that some of the authors are PCI recommenders: "XXX is one of the PCI XXX recommenders."

Preprint DOI: https://arxiv.org/abs/1902.09365v3

Author's reply:

Dear PCI Evol Biol Editors and Reviewers, Please find our third version of our manuscript taking into account for editor suggestions on English language and consistent and explicit use of terminology. We attached two files: a certificate of language editing and a version with all track changes made visible. We are thankful (again) to Myriam Heuertz for her helpful and explicit proposals. The manuscript was edited by AJE editing services (please find the attached certificate), we also corrected all explicated suggestions. We harmonized all terminologies (as well checked by the AJE editing services), added an explicit description of all parameters in figure legends to be readable by population geneticists without having to move into the text, and modified figure texts to be visible. We hope the current (and greatly enhanced by Myriam Heuertz, Marcela Van Loo and David Macaya-Sanz) version of the manuscript will fit PCI Evol Biol standard. Thank you for your consideration, and we look forward to hearing from

you at your earliest convenience. Sincerely, Dr. Solenn Stoeckel, Barbara Porro and Dr. Sophie Arnaud-Haond

*new version on Arxiv wil appear on Mon, 05 Aug 2019 18:00 UTC but all corrections making the new version are readable on the tracked changes document. Arxiv version will be this tracked changes document with all changes accepted.

Download author's reply (PDF file)

Revision round #1

2019-05-13
Dear authors,

We have now received three reviewer reports for your manuscript. Two of the reviewers found that the paper represents a significant contribution for evolutionary biologists interested in clonally reproducing organisms; the third reviewer was unable to assess the relevance of the paper. The reviewers made a series of suggestions which I invite you to take into account before your paper can be reconsidered for recommendation by PCI. My decision on this version of your manuscript is thus “revise”.

Two reviewers pointed out the need for a clearer definition of research concepts and a clearer framing of research questions: please make sure all concepts are defined, including in the abstract of the manuscript. Clonality should be defined as a form of reproduction/multiplication at the first use of the term. The definition of the rate of clonal multiplication rests on the concepts of reference time frame, reference population and reference individual or assessment unit. Defining those concepts across (partially) clonal organisms is not straightforward but the topic requires to be addressed in the paper, to clarify and justify the definitions of these concepts used in the simulation approach you develop. Please see the reviewer reports for comments and suggestions on how to improve the manuscript. Please also consider my additional comments below.

p.10 Can you be more specific about/ justify the “known shapes of curves” used to assess the relationship between c and genotypic descriptors?

p.14: The relationship between R and c is not illustrated in Figure 3; it is in Figure 1. The dependence of beta on population size for a given level of clonality is not illustrated. Can you improve this?

p. 15, bottom: you give the example of multi-locus genotypes present in small strictly clonal populations. I assume you are giving the value at equilibrium. Please link this result to the evolution of R over time (Fig S2). It is not clear how long it takes for the main MLG to establish, i.e., what is the impact of drift, and what is the impact of somatic mutation in this pattern. This example illustrates that R, which reflects MLGs, is not necessarily a very good statistic to reflect the diversity of a population with a high level of clonality. In this case, R appears to represent a higher estimate of diversity compared to the (more intuitive) number of clonal lineages present in the population because R confounds the number of MLGs and their origin (MLLs) (taking this logic to the extreme, the more markers you genotype, the higher will be R because the absolute number of somatic mutations will increase). Now that your genetic data allow sorting MLGs into MLLs, would it be beneficial to pull apart the roles of drift and of somatic mutation in the diversity pattern? I would assume that the role of drift is especially marked in small populations, whereas somatic mutation has the same effect for any population size (and the longer the time, the more of them accumulate). The number of MLLs represents information that is not much exploited in this manuscript, and it would be interesting to assess its usefulness in the context of realistic sampling.

Figure 2: please verify the number of generations in the figure (500) vs. the legend (10,000).

P. 17 and Figure 2. You state that equilibrium is reached in tens to hundreds of generations, but the evolution of parameter values through time is not illustrated for the early time frame. I think it might be insightful to zoom into what happens in the first tens/hundreds of generations; this would be pertinent for some organisms such as trees that display clonal reproduction.

In line with the prior reflection: you assess subsampling effects on the estimates of genotypic and genetic indices when your simulations have reached equilibrium. How realistic are such conditions for a real life population and the real life situation of the population geneticist sampling the population? I understand you are interested in equilibrium conditions to derive the relationships between c and indicators, but the choice of equilibrium conditions for assessing subsampling effects should be at least discussed.

As a last comment: the discussion is long and you might be able to reduce its length without losing much information: see also reviewer reports.

Additional requirements of the managing board: As indicated in the 'How does it work?' section and in the code of conduct, please make sure that: -Data are available to readers, either in the text or through an open data repository such as Zenodo (free), Dryad (to pay) or some other institutional repository. Data must be reusable, thus metadata or accompanying text must carefully describe the data. -Details on quantitative analyses (e.g., data treatment and statistical scripts in R, bioinformatic pipeline scripts, etc.) and details concerning simulations (scripts, codes) are available to readers in the text, as appendices, or through an open data repository, such as Zenodo, Dryad or some other institutional repository. The scripts or codes must be carefully described so that they can be reused. -Details on experimental procedures are available to readers in the text or as appendices. -Authors have no financial conflict of interest relating to the article. The article must contain a "Conflict of interest disclosure" paragraph before the reference section containing this sentence: "The authors of this preprint declare that they have no financial conflict of interest with the content of this article." If appropriate, this disclosure may be completed by a sentence indicating that some of the authors are PCI recommenders: "XXX is one of the PCI XXX recommenders."

Preprint DOI: https://arxiv.org/abs/1902.09365v1

Reviewed by anonymous reviewer, 2019-04-10 16:06

I am unable to say whether this paper has merit. However, I can say that it would be easier to review than it was if the matters it discusses were defined precisely. A first example is "clonality" itself. Authors should be sensitive to the fact that in another context, "clonality" is the collision probability associated with pairs of rearrangements in the adaptive human immune system. After you say what clonality is, why, intuitively, do only high values of {\bf \it c} influence genetic description of {\bf \it R} and {\bf \beta}? More generally, this reviewer's task could have been helped by statement of a precise mechanism by which observations are generated. Otherwise, conclusions are qualitative at best. Here are some items of concern. There is discussion of a machine learning approach to a 12-class problem in classification. What, exactly, were the 12 classes? What was the methodology for "machine learning?" One infers that "neural nets" were used to pick features and also to do classification, but these are only guesses.

What was the larger list of features from which the neural net (if that's what was employed) picked its features? To what extent does the "power law" really apply?

Reviewed by David Macaya-Sanz, 2019-04-23 06:15

Download the review (PDF file)

Reviewed by Marcela Van Loo, 2019-04-30 19:48

Download the review (PDF file)

Author's reply:

Dear PCI Evol Biol Editors and Reviewers, Please find our new version of our manuscript taking into account for editor and reviewers suggestions, and our reply to recommender reviews. We are thankful to editor and reviewers for their comments and proposals that improved the clarity and readability of our manuscript and its messages. We did modifications to address most of the points reported by reviewers. We changed Figure 2 and 4, and supplementary figures so those ones better plot distributions, using violin plots with varying y-axis scaling to better picture identifiable (and un-) signals. We also reduced the discussion part of one page, and, due to clarity pictured by violin plot, reduced the number of figures from 5 to 4. Please find the detailed list of actions and when relevant comments/answer in the “Author’s Reply” file. Thank you for your consideration, and we look forward to hearing from you at your earliest convenience. Sincerely, Dr. Solenn Stoeckel, Barbara Porro and Dr. Sophie Arnaud-Haond

Download author's reply (PDF file)