The paper addresses the topic of the relationships between the rate of clonality (*c*) and some genotypic and genetic parameters often used to assess it and their empirical estimations basing on sampling. It is a very necessary study, as it is highlighted by the finding of somewhat counterintuitive results.

Probably, the parameter more generally used to assess clonality is genotypic richness (denoted by *R*) and this work shows how not only this parameter is not linearly dependent of *c*, but the relationship slightly depends on the population size. This is important, because to actually have a strong decline in the genotypic richness in a population, *c* has to be rather high.

I have found the paper very informative and an eye-opener. In general, I recommend its endorsement but I have some comments that I believe should be addressed in advance.

Some major concerns:

Why did you use different mutation rates for somatic mutations and for sexual events? Can you introduce your rationale in the text? If it is related with germinal cell lines, I am not how this approach could be extrapolated to plants. Have you made a sensitivity analysis to evaluate the impact on your results of choosing different mutation rates?

Is selection modeled in any way in the simulations? My feeling is that it is not and it may be an important factor driving the demographics of PC populations. A little of discussion about how selection could affect your inferences would increase the completeness of the paper (though I know is not the focus of the paper). Related to this, in the simulations, how were the individuals to be reproduced clonally selected? Same for the sexual events? Were they selected completely randomly?

The main result of the study, or at least the most stressed one in the discussion, is the effect of sampling size in the estimation of the values of the genotypic parameters. I think it would be interesting to see some more figures depicting this phenomenon from different perspectives. At least to me, it is not very intuitive, and some more figures would help to understand the mechanics of this bias. Maybe focusing in the case of the sample size of 10,000 out of 100,000. Some errors bars could provide more insights of the magnitude of this issue.

Minor comments:

Format:

You may want to set clearer demarcations of paragraphs, such as first line indentation or separations between paragraphs.

Use consistent format for your parameter symbols across the text (i.e. r_d). Also use consistently either Pareto β or β Pareto (I prefer the first one).

Place a space before and after the equal, the minor and the major signs.

Other:

P3L18: consider changing “based on” by “by”

P3L18: e.g. instead of i.e.

P4L4: Consider removing “but”

P4L6: “Disabling the information” sounds weird to me. Please, consider rephrasing.

P4L9: the value of c… Maybe: the extent of c….

P7L16: Not sure if the word “during” is appropriate here: In clonal reproduction events,…

P9L6: It could be interesting to report the statistic that reflect the goodness of fit of the population empirical distribution with the theoretical power-law inverse cumulative distribution, if you have those data saved.

P14L14: You mention that there are two inflection points in the curve, but I only can see one around $c = 0.5$

P15: I would add some other reference to Figure 1 in the text.

Figure 1: Include a note to stress that the x axes are not linear.

Figure 2: The meaning of the vertical lines should be stated. In addition, I find this figure difficult to interpret. Most lines overlap, and the plots are small. You may consider removing some *c* lines and maybe spliting the figure in several pages, removing some unnecessary labels and reducing subplot margins to increase the plotting area. Also, it would be nicer to have the Y axes lined up. Not sure if the text header matches with the plot either. It talks about 10000 generations, but in the plot the X axes finish at 500. It is difficult also to appreciate the relationship between the main text (P17) and the figure.

Figure 3: Consider placing it in an edge (top or bottom) of the page to avoid splitting the main text.

P19L7: Looking at Figure 3, and not considering the subsampling bias, I feel that the parameter R is enough to evaluate *c*, being the other ones informative but redundant.

Figure 4: Consider here as well the issues raised for Figure 2.

Figure 1 and 5: At which generation were those values drawn?

P21 last lines, P22 first lines: Consider rephrasing.

P22L13: consider rephrasing content within parentheses.

P33L3: Capital letters

Disclaimers: Since English is a second language for me, I have some limitations when assessing the quality of the writing. My impression is that the overall quality is more than adequate for a scientific publication, but I may have overlooked some errors.

In addition, I have no experience in the use of machine learning, so I can only scrutinize that analysis in a shallow manner.