Reviewer's report

Title: Analysis of HCV quasispecies dynamic under selective pressure of combined therapy

Version: 2 Date: 2 August 2012

Reviewer: Yury Khudyakov

Reviewer's report:

The authors analyze hepatitis C Virus NS5A sequences in 11 followed-up patients under interferon therapy. Although the sample size is small (divided into three different response groups), a good deal of experimental work and analysis was performed, producing 690 NS5A sequences. However, I must say that the data analysis is rather simple, favoring lengthy description over identification of statistically significant trends. The major finding of the article is that samples from SVR have lower diversity than samples from NR and ETR. This result is commendable but it has already been found in several articles (for instance in references 19 to 22, and 30 to 35). I believe that the results need to be more compact and that a little more analysis is needed. Also the abstract needs to be reworked.

Major comments:

- Abstract page 2. You say: “This study provides evidence that homogeneity of quasispecies composition, low diversity and less complexity of the NS5A region pre-therapy are associated with viral clearance”.

A common theme of this article is to analyze the diversity in different ways: homogeneity of quasispecies, diversity and complexity. In reality, these are all slightly different measures of the same thing, the sequence heterogeneity of the population. I think that nucleotide diversity (their mean genetic distance) is the best of the three and much of the lengthy description can be shortened and made compact. Here it also makes it looks as if three significant things were found, they are all highly correlated measures.

- Abstract page 2. You say “However, distinctive clustering of pre/post-therapy sequences was observed, suggesting that an evolutionary process occurred during the time course examined”.

Here and also in results and discussion, it is suggested that finding two clusters separating pre and post therapy indicates an evolutionary process. However, the results were not clear in the distinction of these clusters. Even if they were clearly distinct, you should test if therapy caused a significant change in the population over time, but this is not systematically analyzed in the paper. Also, to say that “an evolutionary process occurred” is too vague, evolution is happening all the time.
Abstract page 2. You say “This could explain the initial diversified composition of quasispecies at baseline, followed by an increase in the frequency of a predominant quasispecies in ‘after treatment’ samples of non-responders and end-of-treatment responders, probably because it offers some advantage for the virus”. Also found in discussion, page 14.

The changes in diversity over time were not systematically addressed; the authors just described the % of unique sequences over time in each patient, but all behaved in different ways. The authors state that some variants rise in frequency, but this was not the case for most patients and it was not presented in a quantitative way.

Abstract page 2. You say “These results suggest that quasispecies diversity of the NS5A region could be important for understating the mechanism underlying treatment failure in patients infected with chronic hepatitis C”.

As I stated before, this is already well known, it should be written as saying that this analysis confirms the hypothesis that a relationship exists between NS5A heterogeneity and response to therapy.

Results, page 6. “NS5A quasispecies composition is dynamic over time”

I think that the section on quasispecies composition is too wordy; I don’t think it is necessary to describe the disparate behavior of each patient. Suffice to say that there is not an apparent pattern. As a display, figure 3 is fine, but in reality it would have been more informative to plot and discuss the changes over time of the average genetic distance, which is identical to nucleotide diversity. The percentages of different quasispecies is a very raw estimate, the mean genetic diversity of the population includes the number of unique sequences, their frequencies and their distances, giving a much better estimate of population diversity.

Results, page 7. You say “NS5A quasispecies experience genetic evolution over time”. Also found in pages 12 and 14.

This is a very vague statement, repeated in other places of the article. Genetic evolution of a population over time is a continuous and obvious process. What is of interest is to ascertain the nature and extent of this evolution, for instance if there is drift, or demographic changes, or bottlenecks, or positive and negative selection.

Results, page 7. You say: “The topology of clusters was 80% sustained by the bootstrap value for PO3 sequences”.

I think this paragraph is too descriptive and wordy; it can be put into a table form, so each patient doesn’t need to be described individually. Also, you could measure the distance between every sample and the first “before treatment”. This could be done in several ways, I recommend the Fst implemented in the software Arlequin, which takes into account the distances among sequences and
their frequencies. Then you could test if this distance is significantly different among response groups.

The Arlequin software also allows calculating Fst per site, comparing the frequencies of variants at each amino acid position (or nucleotide), so specific sites can be found with significant changes between before and after therapy. Perhaps the number of significantly different sites is different among the three groups.

- Discussion, page 12. You say “This may explain the initial diversified composition of quasispecies at baseline for some ETR and NR patients”

You could test for significance of this assertion, by means of a paired t-test, comparing for each group the mean genetic distance of two time-points. In the cover letter, the authors explain the differences between their previous article and the newly submitted one, the main difference being the follow-up in many patients. However, the submitted paper doesn’t make much use of the follow-up samples. We don’t know if in average the diversity goes up or down, amino acid variants rose in frequency or if individual codons were selected.

- Discussion, page 12. You say “The purifying selection was relaxed in some cases (clades selected for patients P11 and P47), and allows non-predominant and less advantageous variants to survive at a lower frequency”.

This is speculation, maybe the negative selection is not as strong as in other patients, but how do you know that some of this variants are less advantageous? The population is still under negative selection and these variants survived therapy. Assertions about the fitness of the variants are better studied with a mathematical model.

Minor comments:

- Discussion, page 12. You say: “As with ETR patients, the HCV RNA is undetectable by the end of therapy, suggesting that the selective pressure of therapy acts on quasispecies diversity during therapy until HCV becomes undetectable in samples”.

Please rephrase, the word therapy is used too many times.

- Discussion, page 12: You say: “In order to investigate whether quasispecies in samples collected after treatment or predominant quasispecies identified over the time course were under differential selective pressure, the set of sequences of each NR and ETR patient underwent selective pressure analysis”.

Please rephrase, this is redundant. In order to do X, we did X.

- Methods, page 15. You say: “Samples from the SVR group and before treatment samples from most patients (P05, P35, P40, P44, P03, P42 and P37) enrolled in this study were analyzed in a previous work and were used in this study for comparative analyzes of the evolutionary dynamic of quasispecies [35].
You must cite the relevant work (ref 50), not ref 35.

- You say “The dataset of 690 entire NS5A sequences was submitted to genetic analysis”, “All sequences were subjected to Phred-Phrap programs [55-57]” and later, “All sequences of complete NS5A generated in this study were subjected to the program”.

Please rephrase, the current form “the data was submitted to” is awkward, it should be “the data was analyzed with… or by means of a…

- The software LOCQSPEC is described at too much depth, it was published already and in this paper it was used just to extract the unique sequences, which can be done with several tools, including excel.

- Author’s contribution, you say “CB and CMAC analyzed the data e critically revised the manuscript”.

Please change “e” to “and”.

**Level of interest**: An article whose findings are important to those with closely related research interests

**Quality of written English**: Acceptable

**Statistical review**: No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests**: 

'I declare that I have no competing interests'