Dear Dr. Trebs,

we thank you for communicating the updated tasks for the manuscript. One of your comments was related to the presence of colour cells in one of our tables (Tab.5) . We have solved the issue by replacing that table with a figure (Fig. 6). As regards the referees, we acknowledge the fact that Rev #2 has been satisfied by our prior revisions and endorsed the publication of the manuscript. We therefore addressed the remaining issues that concern novel comments from Rev. #3.
We need to add that anonymous Reviewer # 3 has lowered from 'fair' to 'poor' his/her ranking of Scientific Quality Score, in spite of the extensive revision of the manuscript that we had carried out. The revision yielded a 19-page long rebuttal document, which is a rather uncommon case in our experience, and that was moreover in its vast majority actually dedicated to Rev.# 3 issues. This series of additional comments was therefore not expected, also considering that in your own editorial comment of March 16th you had anticipated: *"Your overall responses to the referee comments appear very detailed and complete to me."*
We also noticed that this time the same reviewer has filed the form by checking the option: *"I am not willing to review the revised paper"*. Therefore, our following answers are hereby provided for your attention only. However, we find somewhat uncommon that a person having put forward several novel questions and arguments (and indicating major revisions needed) would disregard hearing whether these have been found appropriate and how authors were responding.

In any event we here below provide the full answers to these novel points.

Suggestions for revision or reasons for rejection (will be published if the paper is accepted for final publication)

The study is of importance to the deciphering of the processes affecting the composition of atmospheric microbiome. It addresses important questions such as the relevance of the time of the year, the occurrence of dust events and the location of the sampling. However, it is my opinion that the small number of samples, compared with the number of examined variables, produces results with low statistical significance, if any.

ANSWER: in first instance we find surprising the Reviewer commenting that the manuscript *"addresses important questions such as the relevance of the time of the year"*. The reason is because in the first round of reviewing the same reviewer insisted on the fact that our original interpretation of 'seasonal' changes (i.e. meant as relative to the spring vs. fall comparison), were instead to be considered as just 'temporal', with no proven correlation with the time of the year. Therefore, we had accordingly dropped the view of cyclic or recurring events, and accepted that of successional phenomena. The concept had been extensively revised throughout the manuscript, including large parts of the discussion, until the change of the original title, in which, as requested by the reviewer, the reference to the time of the year (season) had been removed. This is "why" we do not understand why the Reviewer is instead praising again the original *"relevance of the time of the year"* having asked to dismiss it.
Second, we found surprising that some basic questions related to the experimental design, as number of samples and number of variables, could still arise at this time of the revision. We had thoroughly addressed the concept of sample and that of control in our prior round of revisions and pointed out that we had analyzed <u>all</u> the dust-outbreak events of that year, so that the 'samples' could not be more than those. And that, given the definitions, the controls could not have been 'any' of the quiet days before or after the storms, but just the last day before and the first day after. Therefore, we simply analyzed all samples available and we did it in two opposite locations.
About the "*low statistical significance if any*" sentence, sounding as a disdainful dismissal of the entire research, we do not understand what the reviewer would be referring to. On the contrary, as regards the whole community level, we showed that the September sampling was significantly different from the May

sampling, in terms of ecological indexes (p Value = 0.00584 for species diversity, and pValue = 0.026 for community evenness); while as regards the single species level, we then showed that 76 taxa were found featuring p values < 0.05. From these, upon applying a stringent Bonferroni-adjusted p-value correction, six of them stood above the significance cutoff, and all within minimal false discovery rate values (FDR < 0.005). All of them were cases highly reduced in September in comparison to May. In the most differentially featured cases, we were able to reach a p-value as low as = 0.000019. Moreover, the robustness of these results was tested by running the analyses independently with parametric as well as non-parametric methods, testing both an ANOVA variance analysis and the non-parametric Wilcoxon Rank test for the verification of the ranking. The two tools gave fully coherent results. The distinctness of our rankings of the variables was moreover confirmed (and shown in the manuscript and supplementary materials), by: Cluster Analysis, Principal Component Analysis; Principal Coordinate Analysis, Linear Discriminant Analysis, Bray Curtis Similarity Analysis of all the 66 pairwise community comparisons;

It is suggested that the authors focus their effort in comparing a single, at most two, variables (the location and the time) and re-analyze their results accordingly.

About having considered three variables instead of less (preferably one, as recommended by the reviewer #3). We need to point out three facts; 1) aim (and title) of this paper is "Determining the hierarchical order by which the variables of sampling period, dust outbreaks occurrence, and sampling location, can shape the airborne bacterial communities in the Mediterranean basin". It should be clear that if one wants to rank items into an order, they need to be more than one. 2) What the reviewer is asking about the analysis is indeed what was done. Results were analyzed independently one variable at a time. We compared which variable yielded the highest shifts in pairwise community distances. We did not analyze these data with general linear nor regression models in which one would test interactions between variables. 3) In submitted reports we were used to hear reviewers asking to do more analyses, and we found strange to hear a complaint about having done too many.

PDF attachment comments:

The manuscript describes a study on the parameters affecting airborne microbial community composition, e.g., season, dust intrusion, geographic proximity to the dust source. These are important questions in the study of aerobiology, especially in the Mediterranean basin that is prone to increasing frequency of Saharan dust intrusions.

My specific remarks are:

L. 20: "…two opposite localities…" – should be "locations".

ANSWER: Correction done.

L. 26: "(118 vs. 65)" – should be moved towards the end of the sentence or removed altogether, at its current location it is unclear what it refers to.

ANSWER: The word orders was moved.

L. 67-68: "Fungal taxa were also analyzed along with their and relationships…" – please correct this sentence.

ANSWER: Correction done.

L. 130-144: This added section describes the samples and their conditions fully and clearly.

ANSWER: We acknowledge this positive comment.

L. 143-144: The term "flanking" is somewhat misleading here. The selected sampling events, on a chronological axis are: dust event, clear day, clear day, dust event. There are no samples before the first dust event or after the last one.

ANSWER: Correction done, the word was removed.

L. 147: There is a stray "." . Please remove.

ANSWER: Correction done.

L. 247: "collection during dust outbreaks under winds from Africa" – the phrasing "under winds" is not correct. Please rephrase.

ANSWER: Correction done. (…brought by winds from Africa)

L. 248: "collection upon under opposite…" – please correct.

ANSWER: Correction done.

L. 252: "add visual aids to each chart attributes" – please remove.

ANSWER: Correction done.

L.257: "their maxima were seen…" – does this refer to all core taxa mentioned, or only to Actinomycetales? Please clarify.

ANSWER: It referred to all three core taxa. We clarified that in the text.

L. 273-274: abrupt line break, please correct.

ANSWER: Correction done.

Table 1: Please add standard deviation to mean values.

ANSWER: values have been added.

Table 2: According to the data presented in this table, the mean variance between the two halves of a single dust event was higher in Cagliari than in Sassari, this is opposed to the number of orders that showed a variance that surpassed the chosen threshold (Table 3). How do the authors interpret each observation?

ANSWER: As will be clear from the answer to the next query, there is no expectedly proportional correlation between the order and number of OTUs that it could encompass, as some orders could have undergone active diversification and others could be at the opposite and be represented by as little as a single OTU. In this Table we used the OTU unit as we wanted to catch the slightest changes possible in the collected air. In Table 3 we wanted instead to assess the extent of the diversity spectrum in its broadest systematics span.

Table 3: Why show the richness in orders? How is it better than showing the number of OTUs?

ANSWER: The OTUs are not reflecting the completeness of bacterial lineages and ecosystem function varieties, as they are simply defined by a clustering cutoff of 97% nucleotide identity. Thus, theoretically, even all different OTUs of a project could even fall within the same single genus. The OTUs richness of a community is more reflective of the micro-evolutionary history of the place than its overall taxonomical latitude. For this reason, a relatively high rank as the order was deemed more apt to convey a true picture of the existing differences among populations.

L. 377-392: ANOVA is an inadequate statistical test to determine differential abundance. I suggest following the approach presented in Gloor et al., 2017 (Front. Microbiol.), which the authors are familiar with (L. 467), yet for some reason chose not to follow.

ANSWER: in first instance, ANOVA is a suitable tool to assess differential abundance. This is also testified by the practice of similar countless studies, as well as of being among the standard tools included into the Calypso webtool suite for microbial communities sequencing data comparisons (Zakrzewski, et al., Bioinformatics 33, 782–783,2017), which has, to date, used and cited by 292 articles indexed in Web of Science.
Moreover, since the only concern with the use of ANOVA regards its well-known basic requirements of distribution normality and variance equality, in our manuscript we had also provided the results of a parallel analysis of the very same data, independently done with the non-parametric alternative to ANOVA, the Wilcoxon Rank Test. That means, that we verified each result with a procedure that does not involve parameters as means and standard deviation and, as stated in each of the versions of the manuscript, we obtained matching results in terms of significantly different taxa. This was already clearly written in the same lines that the reviewer has signaled. Quoting our manuscript: "In order to determine which bacterial taxa were mostly accompanying/causing those changes in a statistically significant manner, and to rank their individual importance in this phenomenon, we run an analysis of the differentially featured taxa, testing both an ANOVA variance analysis and a non-parametric Wilcoxon Rank test verification of the ranking. The two tools gave coherent scores and the results of the ANOVA output are shown in Supplementary Table S1."
As regards the mention to the approach described by Gloor et al. 2017, the reviewer has assumed that we have not taken that into consideration, but instead we had. In addition to the square root transformation, we had verified also the Aitchison's centered log-ratio transformation (CLR), and compared the results of each transformation method. Notwithstanding some slight changes in the shape of the ordination plots, observed phenomena and ensuing trends that we point out and describe in this report were found to be exactly the same. The caveats signaled by Gloor et al., 2017 imply that some communities may suffer from compositional constraints, and some communities could be minimally affected. It depends on the community structure, and does not imply that every author should only construct graphs based on centered log ratio from now on. The reasons for which, once verified the output equivalences with CLR, we opted for showing the former transformation procedure (TSS) is that the combination of total sum scaling with square root transformation is the renowned 'Hellinger transformation'. This transformation has been praised as a preferable choice in ecological community comparisons, (Legendre and Legendre, 1998, Numerical ecology, 2nd English edn. Elsevier, Amsterdam), as it offers the best trade-off between linearity and resolution in comparison to chi-square metrics and other approaches. It is also recognized as more balanced for the weight given to rare species. For this reason it is also the first recommended choice of data transformation in the Calypso webtool suite (Zakrzewski, et al., Bioinformatics 33, 782–783,2017), which, as mentioned above, has to date used and cited by 292 articles indexed in Web of Science. Therefore, once we ruled out the risk that the compositional nature of the datasets could affect the results (by testing CLR as well), we opted for the Hellinger method. We need to remark that this explanation had been already given to a different Reviewer (Rev. #2) in our prior 'Answers to reviewers' document.

L. 424-430: The authors' claim, that an air sample is a built-in average of XXX liters of changing atmosphere, is unreasonable. As the authors themselves state, there is a high day-to-day variability in atmospheric microbiome, yet the hour-by-hour changes are gradual, as the authors themselves must have noticed when analyzing dust events on a 12h basis. When comparing samples that were obtained by 24h of sampling, it is still important to obtain sufficient replicates of the same scale, for statistical significance. Had one chosen to sample 10 liter of sea water, would it be acceptable to obtain a single sample to represent each condition, claiming that it is an average of 10000 ul of diverse microcosms?

ANSWER: No, it is the opposite. And exactly by the same argument that the reviewer points out: since there is a continuous hour-by-hour change, sampling for short times instead of sampling for 24 hours, would not yield replicates but different samples representing successional frames of the evolving new condition, which would not be legitimate to treat as replicates and to compare. Only if one could afford the expense of 24 Skypost Tecora motorized devices, it could be possible to filter simultaneously and in parallel, air for one hour in the same location and obtain truly legitimate replicates. Besides, the variable that we considered here is the dust outbreak, which, as such, exists in two possible statuses: ON or OFF. Therefore the 24 h sampling is carried out completely in the presence of the dust event and indeed not only accounts for the possible variation that takes place in it, but definitely seeks to incorporate as much as possible of that to be more adequately representative of that condition. The fact that variation occurs is also demonstrated by our first-12h/second-12h variation. The concept we are relying on, is exactly the same that is used in soil analyses. In these cases, a plot is analyzed by tracing a cross and pooling five subsamples collected at the center and at each of the four corners transect to obtain a single sample that incorporate as much as possible the spatial variability. And even the example of 10 liters of water chosen by the reviewer is not different, since if one is not sequencing 10 liters but small aliquots, the possibilities are 1) using some of these aliquots to analyze them and consider them as replicates; or 2) Filter all of those 10 liters and sequence the resulting concentrated community. In the first case if one takes for example 10 replicates of 1 ml each one would have analyzed 1 /10000th of the sampled environment. In the second case, filtering all as we did with the air, and resuspending the filtered cells in 1 ml, one would analyze the whole community. Contrary to the reviewer affirmation, our single sample is not "an average of 10000 ul of diverse microcosms". In fact it is not an average at all, as it is the totality. (By the way, in 10 liters there are not 10000 ul but 10000000 ul.)

Avoiding adequate sampling design by suggesting that the samples themselves represent averages of smaller increments is unacceptable in aerobiology as it is in any environmental microbiome study.

ANSWER: No, besides the fact that in aerobiology, sampling through a continuous time lapse as we did is the standard practice, as just answered above. We did not do any averaging here as we analyzed the whole sampled amount.

Each environmental sample is somewhat different than its replicate, and some environmental conditions are difficult to replicate, such are dust events. This should encourage deeper and more extensive sampling, and not the opposite.

ANSWER: There is an open and naïve contradiction in this sentence: "Each environmental sample is somewhat different than its replicate". If a sample is acknowledged in the first place, as 'already known to be different from another sample', that can never be taken as 'replicate'. Replicates are legitimately assumed as samples which obey to the assumption of being comparable as belonging to the same experimental condition. If one already knows that an incremental gradient of variability among replicates is predetermined, the legitimate choice is to pool them into a single sample that represents the persisting overall condition, in our case the 'ON' state of the dust discharge.

The authors could make the effort to sample more times on clear days, which are not as rare, and their study would only benefit from this choice.

ANSWER: We had actually already addressed this in the prior round of revisions and we state that again here:
Since it is already known from literature that, as microbial community composition is concerned, even in the absence of dust outbreaks, the ambient state of the atmosphere is not stable either, our goal was not to compare an hypothetical *status quo* with an altered one. The meaning of the 'control' here was to catch 1) the first possible timeframe after the stopping of a northbound dusty wind outbreak (it occurred in May) or 2) the latest possible timeframe of a situation before the onset of a dust-carrying change of wind regime (it occurred in September). Thence, in the latter event the control is not intended as a situation of calm that could represent a period of unknown length, but rather the time-zero sample of the dust event itself. While for the former case in May, the control is symmetrically designed as the quiet after the storm.

In the lack of more samples, it is statistically irresponsible to compare 3 different variables.

ANSWER: Asa already commented above. The reviewer is apparently missing the point that we are comparing the variables independently one at a time and we are not testing their interaction in a linear model or any other way.

L. 467: As the authors cite the problem well defined by Gloor et al. (2017), why not implement the suggested solution to their work? The cited paper gives adequate tools to overcome the compositional nature of sequencing data, yet the authors chose to ignore it altogether

ANSWER: We just paste-repeat the answer given above:
As regards the mention to the approach described by Gloor et al. 2017, the reviewer has assumed that we have not taken that into consideration, but instead we had. In addition to the square root transformation, we had verified also the Aitchison's centered log-ratio transformation (CLR), and compared the results of each transformation method. Notwithstanding some slight changes in the shape of the ordination plots, observed phenomena and ensuing trends that we point out and describe in this report were found to be exactly the same. The caveats signaled by Gloor et al., 2017 imply that some communities may suffer from compositional constraints, and some communities could be minimally affected. It depends on the community structure, and does not imply that every author should only construct graphs based on centered log ratio from now on. The reasons for which, once verified the output equivalences with CLR, we opted for showing the former transformation procedure (TSS) is that the combination of total sum scaling with square root transformation is the renowned 'Hellinger transformation'. This transformation has been praised as a preferable choice in ecological community comparisons, (Legendre and Legendre, 1998, Numerical ecology, 2nd English edn. Elsevier, Amsterdam), as it offers the best trade-off between linearity and resolution in comparison to chi-square metrics and other approaches. It is also recognized as more balanced for the weight given to rare species. For this reason it is also the first recommended choice of data transformation in the Calypso webtool suite (Zakrzewski, et al., Bioinformatics 33, 782–783,2017), which, as mentioned above, has to date used and cited by 292 articles indexed in Web of Science. Therefore, once we ruled out the risk that the compositional nature of the datasets could affect the results (by testing CLR as well), we opted for the Hellinger method. We need to remark that this explanation had been already given to a different Reviewer (Rev. #2) in our prior 'Answers to reviewers' document.

Counting on having clarified all the pending issues we thank you for your kind attention and cooperation.

Andrea Squartini