molecular systems biology

Global biogeography of SAR11 marine bacteria

Mark V. Brown, Federico M. Lauro, Matthew Z. DeMaere, Les Muir, David Wilkins, Torsten Thomas, Martin J. Riddle, Jed A. Fuhrman, Cynthia Andrews-Pfannkoch, Jeffrey Hoffman, Jeffrey B. McQuaid, Andrew Allen, Stephen R. Rintoul, Ricardo Cavicchioli

Corresponding author: Ricardo Cavicchioli, The University of NSW, Sydney

Review timeline:

Submission date:	15 March 2012
Editorial Decision:	23 April 2012
Revision received:	06 June 2012
Editorial Decision:	13 June 2012
Revision received:	15 June 2012
Accepted:	18 June 2012

Transaction Report:

(Note: With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. The original formatting of letters and referee reports may not be reflected in this compilation.)

Editorial Decision 23 April 2012

Thank you again for submitting your work to Molecular Systems Biology. We have now heard back from the three referees who agreed to evaluate your manuscript. As you will see from the reports below, the referees find the topic of your study of potential interest. They raise, however, substantial concerns on your work, which, I am afraid to say, preclude its publication in its present form.

The reviewers provided very detailed reports that list both important concerns, and constructive suggestions for improvements. The editor would like to emphasize that the reviewers felt that some these concerns were sufficient to cast doubt on some of the key claims made in this work, and as such these points will need to be addressed convincingly, and possibly with new evidence and analyses, before this work would be suitable for publication in Molecular Systems Biology. In particular, reviewer #2 felt that some of the claims regarding adaptive changes in gene content remained speculative, and all reviewers had concerns regarding the clarity and/or conclusiveness of the methods used to identify genes experiencing positive selection. In addition, the third reviewer was concerned that geographical sampling issues and upwelling differences could bias the definition of "polar" phylotypes, and felt that this issue required thorough consideration (and ideally inclusion of additional data if available).

If you feel you can satisfactorily deal with these points and those listed by the referees, you may wish to submit a revised version of your manuscript. Please attach a covering letter giving details of the way in which you have handled each of the points raised by the referees. A revised manuscript will be once again subject to review and you probably understand that we can give you no guarantee at this stage that the eventual outcome will be favorable.

PLEASE NOTE As part of the EMBO Publications transparent editorial process initiative (see

http://www.nature.com/msb/journal/v6/n1/full/msb201072.html), Molecular Systems Biology now publishes online a Review Process File with each accepted manuscript. Please be aware that in the event of acceptance, your cover letter/point-by-point document will be included as part of this file, which will be available to the scientific community. Authors may opt out of the transparent process at any stage prior to publication (contact us at msb@embo.org). More information about this initiative is available in our Instructions to Authors.

Sincerely,
Editor - Molecular Systems Biology
msb@embo.org

Referee reports:

Reviewer #1 (Remarks to the Author):

Brown et al use metagenomic analysis to examine the distribution of Pelagibacter phylotypes in surface waters of the global ocean, convincingly linking biogeographic phylotype structuring to a gradient in temperature/latitude. The paper should be of interest to a broad audience, as it provides critical insight into the environmental factors shaping the ecology and evolution of arguably one of the most important organisms on the planet. Furthermore, the work is one of the better recent examples of integrating diverse metagenomic datasets to explore fundamental questions in earth systems biology. Building upon prior research that has focused more targetedly on specific isolates or datasets, the current study, in part by providing key data on polar Pelagibacter strains, should help researchers define the physiological basis for distinct "ecotypes", enabling predictive models of distribution at the global scale, but potentially also at the scale of microhabitat niche (similar to work that has been done on Prochlorococcus).

I enjoyed this paper, both for it subject matter and its technical quality. It is exceptionally well-written, analytically robust and thorough, and makes excellent use of Figures (They are clear, well-annotated, and convincing). Furthermore, the authors are careful not to overstep the bounds of their data - they exercise an appropriate level of caution when inferring major trends and drawing conclusions about drivers of biogeography and the (potential) adaptive value of phylotype genomic differences (e.g., they don't give excessive wait to the modeling predictions, but present the results as interesting food for thought...). More targeted gene-specific or physiological analyses will undoubtedly follow. I have few very (minor) comments.

- Metadata collection. More details could be provided in the SOM regarding metadata collection. E.g., how was chla and temperature measured (notably for the new polar datasets)? Are chla measurements discrete values or some integrated average (over season, vertical gradients)?

- Tree construction. More details could be provided here. I'm not even sure exactly how many nucleotides are being used in the alignments/analysis. Given the importance of delineating clades for this paper, the phylogeny should be as clear and robust as possible. E.g., how are "significant nodes" being defined? Why are there no bootstrap values in Figure 1?

- Signatures of positive selection. Slightly more detail about the threshold for determining selection would be useful, either in the Methods or main text. I think it is appropriate to define "signature of positive selection." Is there a dN/dS threshold being applied?

- The section on "Phylotype analysis" in the Material and Methods is hard to read (lots of information about strains with very similar numerical designations). Some of this info might be better in a table.

- Figure 2 - the country names are distracting.

- Column1, Table S2. Are these accession numbers? Are all of these in CAMERA? Could annotate better. Also, might be useful to include sampling date in the table?

Reviewer #2 (Remarks to the Author):

The paper by Brown et al reports the survey of the diversity and biogeography of SAR11-like bacteria, a very abundant and thus important marine group, in metagenomics datasets covering a large part of the World's oceans. The manuscript is primarily based on the phylogenetic analysis of sequences covering the ITS region of the rRNA operon and secondarily on metagenomics (Sanger and 454) sequences. The authors found that specific phylotypes of SAR11 show strong geographic patterns, driven mostly by temperature and latitude. The main thesis of the manuscript, i.e., understanding what factors limit or promote the dispersion of microbes in the oceans, is an important one for study and the manuscript offers new insights into this issue, albeit some of its findings repeat those of previous studies. The authors could further improve their manuscript by paying more attention to the issues outlined below:

The authors claim in their abstract and introduction that they have identified several genetic elements that differentiate the various SAR11 phylotypes based on metagenomic data. However, the metagenomic part of the paper is very limited; most of the paper is about the distribution of ITS-defined phylotypes and correlation analysis to physicochemical data. The latter represents limited new information since similar analysis, based on several of the exact same samples (e.g., GOS samples, Rusch et al. 2007), was performed previously (albeit comparisons to data from polar sites is unique to this study). The ITS and the 16S rRNA gene levels are also too conserved; thus, important levels of genotypic and phenotypic differentiation most certainly underlie identical ITS sequences. The metagenomic/genomic data that was made available as part of the study could provide higher resolution compared to what was achieved previously but, as indicated above, the information provided is limited and/or the derived conclusions not well-supported by data. For instance, the abstract claims that genomic analysis has revealed signatures of adaptation and positive selection and this comes across as a major finding of the study. However, there is not a single figure or table in the main article with results to back up such a strong conclusion. No details are provided in the main text about how strong the signal of positive selection was, how many genes were detected as positively selected, not even what specific methods were used and what the caveats of the methods are, if any. The manuscript will benefit from focusing on and leveraging more the polar data.

Specific commends:

Line 66: genome streamlining is one hypothesis for the abundance of SAR11, but not directly proven. It should be written as "might be in part" etc.

Line 593: the title of the reference is wrong

Lines 98-99. Is this possibly due to lack of adequate coverage? Only a handful of sequences per sample were obtained, on average, which may not be adequate to detect low-abundance phylotypes in some samples. Perhaps a statistics test will be helpful here to exclude the latter possibility.

Line 110. Here and elsewhere: What is actually analyzed during the multivariate analysis? Was it the number of reads per (predefined) phylotype or the analysis also included the phylogenetic diversity of the sequences recovered (as the latter map on the ITS and 16S rRNA gene trees)?

Lines 147-150. Is it possible that important levels of genotypic differentiation underlie identical, or almost identical, ITS/16S rRNA gene sequences? What is the level of resolution of ITS actually and how is it compared to the whole-genome level? Will be important to at least discuss this somewhere.

Lines 163-173. It is not clear what is compared in this paragraph. What do the eight clusters represent? I was able to understand more after I saw the corresponding figure 2, but the text should stand on its own. It would be useful to refer to the names of the samples like the authors do above. Figure 2 contains more than one cluster with green symbols.

Line 177. How reliable were these assemblies? Was their quality verified somehow? In general, the assembly part of the manuscript requires more attention by the authors and to be accompanied by basic assembly statistics such as coverage, number of contigs, possible mis-assemblies detected, etc.

Line 181. I don't think you can safely draw this conclusion based on the relatedness observed. Organisms related at 86.3% genome average nucleotide identity typically belong to different species and frequently show considerable gene content and genome rearrangement differences.

Lines 189-192: The results described in this paragraph are not very clear. Where the average number of hits refers? Average from the 2 polar genomes compared? Average coverage of the genomes?

Lines 209-213. Hand waving, here and elsewhere. There is no real data to back up the claim that the gene content differences identified play a role in adaptation to polar environments. What about if the genes identified are just the product of mis-assembly? Or that adaptation is conferred at the transcription level of shared genes, not gene content? The authors need to qualify their conclusions better.

Lines 214-219. More results about this section need to be shown in the main text; the reader should not be referred to the supplementary material for one of the two main lines of the paper (based on the abstract at least).

Lines 232-238. Dn/Ds analysis to detect positive selection has several limitations (see for instance Rocha et al., 2006 and Schmid and Yang 2009), none of which seems to be taken into account or at least discussed by the authors here.

Line 321. What do you mean? Do you mean that P1a.3 is the same as S1a? Please revise.

Lines 347-361. This section is just hand waving. I would recommend delete it completely as no data is available to back it up. For instance, why warmer phylotypes will invade polar regions as the latter become warmer and not polar phylotypes adapt to live at higher temperatures? Data is needed before one can make educated predictions about what may happen during these scenarios.

Although the text is well-written in general, with only a few grammatical errors or inaccuracies, the figures are not of the same quality. Figure keys, graph labels, and text should be added to the figures to make them easier to follow and stand on their own.

Reviewer #3 (Remarks to the Author):

This research analyzes a global genomic dataset to characterize and propose a temperature-driven biogeography of SAR11 ecotypes. Overall the authors were very clear in conveying the scientific motivations and study design. The analyses performed were generally thorough and appropriate, and provided compelling results. Although the main body of the manuscript was well written the figures could more clearly illustrate some of the points being referred to in the text (see specific comments below). In addition, two main issues need to be addressed regarding this manuscript.
First, while the data sufficiently supports the conclusion that SAR11 phylotypes biogeography is linked to temperature, from my perspective it seems difficult to eliminate confounding factors that could also explain the observed trends. For example, although the "polar" biome dataset very extensive, it is highly skewed. That is, 80% of the samples were collected from Antarctic waters and it appears that all of these show a consistent SAR11 community signature (i.e., dominance by P1a.1 and sub-dominance by P2.2). Conversely, the remaining 20% of "polar biome" samples were taken from North Atlantic waters and do not appear to consistently have the same trend in community composition. Admittedly, this is difficult to explicitly determine without sample labels on figure 1. Thus, it could be equally concluded that P1a.1 and P2.2 are not "polar" phylotypes on a global sense, but occur regionally in Antarctic waters. Additionally, it is difficult to de-convolute the role of upwelling in the distribution of the phylotypes commonly found in colder waters. Given that SAR11 populations have been shown in previous studies to vary according to seasonal mixing and depth, it would seem important to address these confounding factors. The discussion touches on this issue but does not clearly indicate that the phylotypes found in the coldest waters are also not areas of active upwelling. It would seem appropriate to include an upwelling variable, if possible, in the dbRDA and DISTLM analyses to account for any community composition variability captured by this factor. Ideally, the appropriate samples and variable measurements to test these alternatives would be available. However, given the scope of this research it is understandable that this may not be possible. As such, appropriate considerations should be added to the discussion section if these concerns are valid.

The second issue is that the materials and methods section was superficial and needs more detail to allow readers to properly evaluate the methods applied. Specific notes on proposed additions/changes to this section are provided below.

Minor comments:

1) Line 52: The term "robust" does not seem appropriate here. Perhaps "measureable" instead?

2) Line 84: Add "latitude" after temperature since these factors could not be de-convoluted.

3) Line 90: There seem to be only 862 ITS sequences represented by the trees of figure 1 and figure S1, please account for this discrepancy. Also, sequences from the "NA" clusters seem to be mislabeled as the group of sequences with the prefix "NA1" contains 12 sequences, but is shown in figure 1 to have nine members and vice versa.

4) Figure 1: It may be useful to highlight the "biomes" being referred to in the text in some way. Perhaps, simply dividing the background by lines, or colors according to temperature divisions? Also, the differentiation of colors for P2.1 and P2.2 in the phylogenetic tree is difficult for me to determine, although the bar graph is fine. Maybe a stronger contrast of colors could be used for these groups.

5) Line 104: Change this sentence to note that at ~ 20 {degree sign}C P1a.2 begins to be the dominant P1a phylotype, since its appearance occurs at much lower temperatures.

6) Line 111: Why was longitude not included as a test variable in the DISTLM and dbRDA models? I don't believe that its lack of significance should be ruled out a priori, if that was the case. I also wonder if date of sampling and/or a categorical variable to designate upwelling state would be useful to test, as mentioned above. Along these lines, is there a reason why the deep Mediterranean samples were not included in this analysis?

7) Figure 2: It would be beneficial to include supp. figure S2 into this figure to illustrate how the dbRDA model is partitioning the variance.

8) Lines 178 - 184: Please further justify the selection of P3.2 as a polar representative. I understand that its assembly from Ace Lake is the driving rationale, but this phylotype is largely uncharacterized and based on the ITS analysis P3.2 is found in both temperate and polar regions, a conclusion which seems based on very small sample sizes. It is understandable that the authors would want to include this genome in the comparative analysis, but it seems a bit suspect to rely on this genome as one of the two representative of the polar phylotypes. Perhaps it would be more convincing if the bias analyses performed were reported for P3.2 explicitly. For example, Lines 190-192 suggest a bias in blast hits between polar metagenomes and the two mosaic assemblies, but the individual hit rate between ACE_P1a.1 and ACE_P3.2 are not reported. Additionally, on Line 206 the percent of hits for unique ACE_P3.2 genes to polar metagenomes are not reported.

9) Line 192: What is the statistical test used to denote this bias?

10) Figure 3: It would be useful for the reader if the regions of low recruitment referred to in the text (Lines 193 - 196) were highlighted or otherwise noted. Also, labeling the genomes with their names in the figure would be more helpful than referring to the legend for them. Finally, please indicate on the figure or in the legend the threshold level of % identity used to denote matching regions in the ACT plots.

11) Line 199: I may be confused but how many genes that are specific to the polar representatives were shared between phylotypes? (I.e., are any of the 344 and 444 unique genes for ACE_P1a.1 and ACE_P3.2 shared between the two genomes? Or are they completely unique with respect to all genomes, not just the tropical genomes?)

12) Line 206: The methodology to assign significant biases is unclear, please elaborate on this more in the Materials and Methods section.

13) Line 270 - 271: Since the effects of upwelling, and thus greater nutrient concentrations, cannot be ruled out at this point, please note that selection of transporters could also be a result in differences in substrate concentration and associated KM values.

14) Line 282 - 290: Most genes demonstrating positive selection seem to be specifically linked to cold adaptation (Lines 237 - 238). Therefore, the impact of this finding does not seem to extend to the other differences in environment mentioned beyond temperature (e.g., salinity, mineral concentrations, nutrient inputs).

15) Lines 298 - 300: If the data supporting this conclusion is figure 2 alone, it does not seem justified. For example, using a similar logic it could be said that Monterey Bay waters (Mont_Bay_1) influence the community in the North Atlantic (Gilbert_Apr_day).

16) Lines 304 - 307: Along with the previous comment above, I'm not sure much can be said about these samples given the results presented.

17) Line 317: To avoid confusion, please consistently refer to the subgroups as SIa/SIb or S1a/S1b.

18) Line 323: I am assuming due to technological differences it can not be determined if there was a consistent phylotype of the S1b subgroup associated with these mixing/upwelling events across sites?

Materials and methods questions/comments:

19) Is the sequencing technology the same for all metagenomes? If not, please add this information to one of the supplemental tables (see comment 23 below).

20) Line 384: No reference or website is given for the ITS database.

21) Lines 378 - 396: The section describing ITS identification and how the abundance analysis was performed needs much greater detail. It is assumed that the analysis was performed with reads and that some sort of clustering was performed (similar to pyrotag studies?). If this is true, how are the 2,983 markers clustered into 865 ITS sequences? Is the discrepancy between these numbers due to redundancy? Are clusters generated by grouping perfect matches together with the longest read representing the consensus sequence to remove any redundancy? Is abundance then determined by simply summing the number of reads in each metagenome assigned to each of the 865 ITS sequences and dividing by the total? Once abundances of each phylotype were assigned by sample, were the assignment distributions relatively even or could specific samples be classified as outliers based on low total numbers? Or how many samples of the 127 were represented by only one phylotype with one read?

22) Line 389: Is temperature considered a continuous variable in the dbRDA or categorical by biome? Also, what is the "chlorophyll" variable a measure of (total Chl, ChlA, ChlB)?

23) Line 393: How many samples did not have all associated environmental variables and were then estimated? It would be useful to include all of these measured and estimated variables in a supplemental table along with the sequencing technology information for each metagenome.

24) Line 395 & 397: How were variables normalized? The text mentioned that a square-root transformation was used on the relative abundances (as opposed to the raw count data). Was there a reason an arcsin transformation was not used instead? I am of the impression that this type of transformation is more appropriate for percentage data (Quinn and Keough, 2002, pg. 66).

25) Lines 403 - 408: It may be useful to state within the text the phylogenetic diversity captured by each of the phylogenetic levels you describe (i.e., clade, subgroup, and phylotype). For a more universal point of reference, it may be best to use the 16S tree from figure S1 to calculate these values from.

26) Line 478: How were scaffolds ordered? (Alignments to known genomes?)

27) Line 493: Was depth measured at each base position of the genome and averaged? Was the variation accounted for on the same scale?

28) Line 500: I'm a bit confused about how paralogs are being identified. Can you clarify what is the query and database in each case?

29) Line 501 - 507: Can more details on the resampling method discussed here be provided? I am assuming it is to test the biases present in the distribution of hits to the subdivided read sets according to water temperature, but it is a little unclear how this is being tested. Also, do repeat regions strongly influence the analysis or are hits widely distributed among the reads and across the whole protein query sequences?

30) Line 515: Change "proteins that" to "protein pairs that", if this is what is meant.

1st Revision - authors' response 06 June 2012

We provide a revision of our manuscript MSB-12-3650 entitled, "**Global Biogeography of SAR11 Marine Bacteria**", authored by Mark V. Brown, Federico M. Lauro, Matthew Z. DeMaere, Les Muir, David Wilkins, Torsten Thomas, Martin Riddle, Jed A. Fuhrman, Cynthia Andrews-Pfannkoch, Jeffrey M. Hoffman, Jeffrey B. McQuaid, Andrew Allen, Stephen R. Rintoul, Ricardo Cavicchioli.

We appreciate the comments of the reviewers which were largely positive and constructive. We have carefully and thoroughly revised the manuscript. We have provided the manuscript as a Word doc that includes a statement about author contributions and a thumbnail jpg image, and below provided bullet point main findings and standfirst text, and detailed the changes made in response to reviewers comments. The License to Publish has also been faxed through. A "compare" version of the main document has also been provided to assist in identifying changes to the original submission. The figures have been provided zipped as high resolution pdf files.

Main findings:

- By generating 37 new Antarctic metagenomes and analyzing the internal transcribed spacer (ITS) regions of the SAR11 clade in a total of 128 surface marine metagenomes we identified phylotype distributions that strongly correlated with temperature and latitude.
- By assembling SAR11 genomes from Antarctic metagenome data we identified specific genes, biases in gene functions and signatures of positive selection in the genomes of the polar SAR11 – genomic signatures of adaptive radiation.
- Our data demonstrates the importance of adaptive radiation in the organism's ability to proliferate throughout the world's oceans, and describes genomic traits characteristic of different phylotypes in specific marine biomes.
- The study has important predictive implications for the responses of a microbial group that plays a fundamental role in nutrient cycling in the microbial loop, and the findings may now facilitate the development of oceanographic models that predict the effects of ocean temperature on the distribution and function of dominant marine heterotrophic bacteria.

Standfirst text:
The global biogeography of the dominant marine heterotrophic bacterial clade, SAR11, was examined. It has evolved through adaptive radiation in response to environmental parameters, particularly temperature, by selecting for specific phylotypes at different latitudes that differ in gene content.

Reviewer #1 (Remarks to the Author):
Brown et al use metagenomic analysis to examine the distribution of Pelagibacter phylotypes in surface waters of the global ocean, convincingly linking biogeographic phylotype structuring to a gradient in temperature/latitude. The paper should be of interest to a broad audience, as it provides critical insight into the environmental factors

shaping the ecology and evolution of arguably one of the most important organisms on the planet. Furthermore, the work is one of the better recent examples of integrating diverse metagenomic datasets to explore fundamental questions in earth systems biology. Building upon prior research that has focused more targetedly on specific isolates or datasets, the current study, in part by providing key data on polar Pelagibacter strains, should help researchers define the physiological basis for distinct "ecotypes", enabling predictive models of distribution at the global scale, but potentially also at the scale of microhabitat niche (similar to work that has been done on Prochlorococcus).

I enjoyed this paper, both for it subject matter and its technical quality. It is exceptionally well-written, analytically robust and thorough, and makes excellent use of Figures (They are clear, well-annotated, and convincing). Furthermore, the authors are careful not to overstep the bounds of their data - they exercise an appropriate level of caution when inferring major trends and drawing conclusions about drivers of biogeography and the (potential) adaptive value of phylotype genomic differences (e.g., they don't give excessive wait to the modeling predictions, but present the results as interesting food for thought...). More targeted gene-specific or physiological analyses will undoubtedly follow. I have few very (minor) comments.

- Metadata collection. More details could be provided in the SOM regarding metadata collection. E.g., how was chla and temperature measured (notably for the new polar datasets)? Are chla measurements discrete values or some integrated average (over season, vertical gradients)?

RESPONSE: The following statement has been added to the Materials and methods.

> "Surface temperature, salinity and chlorophyll measurements, along with depth of water column were obtained using the underway line aboard the Aurora Australis."

Chlorophyll data for all other datasets were taken from the relevant metadata fields in the MG-RAST database or from the relevant manuscripts.

- Tree construction. More details could be provided here. I'm not even sure exactly how many nucleotides are being used in the alignments/analysis. Given the importance of delineating clades for this paper, the phylogeny should be as clear and robust as possible. E.g., how are "significant nodes" being defined? Why are there no bootstrap values in Figure 1?

RESPONSE: We have added details concerning the tree construction and phylotype designation processes to both the Materials and methods and figure legends to make this clearer for the reader. Significant nodes were identified using data from the literature along with the identification of new phylotypes based on bootstrap support and conserved motifs. Bootstrap values for the clades are provided in Supplementary Figure S1. This supplementary figure is exhaustive and was included to provide readers with such details.

- Signatures of positive selection. Slightly more detail about the threshold for determining selection would be useful, either in the Methods or main text. I think it is appropriate to define "signature of positive selection." Is there a dN/dS threshold being applied?

RESPONSE: A statement specifying that only protein pairs with dN/dS>1 over a 60 amino acid window were considered under positive selection has been included in the Materials and methods. This is generally accepted as a fairly robust signature for diversifying selection (see for example Smith, N.H., Maynard Smith, J., Spratt, B.G., 1995. Sequence evolution of the porB gene of Neisseria gonorrhoeae and Neisseria

meningitidis: evidence of positive Darwinian selection. Mol. Biol. Evol. 12, 363–370). During the revision of the manuscript we noted that the wrong column for dN/dS ratios had been pasted into Supplementary Table S4 for the comparison IMCC9063 vs. HIMB114 and perhaps this was a source of confusion. We now state:

> "Orthologous protein pairs were considered under positive selection only if dN/dS ratio was >1 over at least one window of 60 amino acids."

- The section on "Phylotype analysis" in the Material and Methods is hard to read (lots of information about strains with very similar numerical designations). Some of this info might be better in a table.

RESPONSE: Based on this comment from reviewer 1 and comments from reviewer 3 we have added a new Table 1 which outlines the clade, subgroup, phylotype hierarchy along with other pertinent information including providing the phylogenetic diversity captured at each level, based on full-length 16S rRNA gene sequences used in Supplementary Figure S1.

- Figure 2 - the country names are distracting.

RESPONSE: We would prefer to keep the figure the way it is because removing country names requires changing to a satellite image which results in a dark blue coloring for the ocean and distracting oceanic features. The overall effect was that the colored stars representing the sample locations (the main features) were far less distinct. As a result we have not changed the figure.

- Column1, Table S2. Are these accession numbers? Are all of these in CAMERA? Could annotate better. Also, might be useful to include sampling date in the table?

RESPONSE: We obtained all publically available marine metagenome datasets from the MG-RAST depository and have added the MG-RAST accession numbers and the date of sampling to Supplementary Table S2. All these datasets should also be available in CAMERA and the new Antarctic datasets will be available in CAMERA upon release.

Reviewer #2 (Remarks to the Author):

The paper by Brown et al reports the survey of the diversity and biogeography of SAR11-like bacteria, a very abundant and thus important marine group, in metagenomics datasets covering a large part of the World's oceans. The manuscript is primarily based on the phylogenetic analysis of sequences covering the ITS region of the rRNA operon and secondarily on metagenomics (Sanger and 454) sequences. The authors found that specific phylotypes of SAR11 show strong geographic patterns, driven mostly by temperature and latitude. The main thesis of the manuscript, i.e., understanding what factors limit or promote the dispersion of microbes in the oceans, is an important one for study and the manuscript offers new insights into this issue, albeit some of its findings repeat those of previous studies. The authors could further improve their manuscript by paying more attention to the issues outlined below:

The authors claim in their abstract and introduction that they have identified several genetic elements that differentiate the various SAR11 phylotypes based on metagenomic data. However, the metagenomic part of the paper is very limited; most of the paper is about the distribution of ITS-defined phylotypes and correlation analysis to physicochemical data. The latter represents limited new information since similar analysis, based on several of the exact same samples (e.g., GOS samples, Rusch et al. 2007), was performed previously (albeit comparisons to data from polar sites is unique to

this study). The ITS and the 16S rRNA gene levels are also too conserved; thus, important levels of genotypic and phenotypic differentiation most certainly underlie identical ITS sequences. The metagenomic/genomic data that was made available as part of the study could provide higher resolution compared to what was achieved previously but, as indicated above, the information provided is limited and/or the derived conclusions not well-supported by data.

For instance, the abstract claims that genomic analysis has revealed signatures of adaptation and positive selection and this comes across as a major finding of the study. However, there is not a single figure or table in the main article with results to back up such a strong conclusion. No details are provided in the main text about how strong the signal of positive selection was, how many genes were detected as positively selected, not even what specific methods were used and what the caveats of the methods are, if any. The manuscript will benefit from focusing on and leveraging more the polar data.

RESPONSE: The extensive changes we have made in response to all reviewers comments have provided greater depth and reasoning about the conclusions we draw – in particular the relationship between subgroup, phylotype and genome, and genomic signatures of adaptation and selection.

Specific commends:

Line 66: genome streamlining is one hypothesis for the abundance of SAR11, but not directly proven. It should be written as "might be in part" etc.

RESPONSE: This change has been made.

Line 593: the title of the reference is wrong

RESPONSE: The title has been corrected.

Lines 98-99. Is this possibly due to lack of adequate coverage? Only a handful of sequences per sample were obtained, on average, which may not be adequate to detect low-abundance phylotypes in some samples. Perhaps a statistics test will be helpful here to exclude the latter possibility.

RESPONSE: We agree with the reviewer that analysis of these metagenomic datasets will not provide adequate coverage to detect “rare” organisms. Hence we were careful to phrase this sentence stating that “no phylotype is abundant everywhere” rather than saying that no phylotype could possibly be present everywhere. To ensure this is clear we modified the sentence to state:

> “Analysis of the 2983 resultant ITS markers indicates that no phylotype is abundant everywhere in surface waters (Figure 1), although our ITS analysis cannot determine if a low abundance phylotype has a consistent level of abundance globally.”

Line 110. Here and elsewhere: What is actually analyzed during the multivariate analysis? Was it the number of reads per (predefined) phylotype or the analysis also included the phylogenetic diversity of the sequences recovered (as the latter map on the ITS and 16S rRNA gene trees)?

RESPONSE: The multivariate dataset is the relative number of reads of each predefined phylotype observed in each of the metagenomic samples. This has been clarified at the

first mention in the Results, in the Materials and methods and the Figure legend for Figure 2 has been enhanced.

Lines 147-150. Is it possible that important levels of genotypic differentiation underlie identical, or almost identical, ITS/16S rRNA gene sequences? What is the level of resolution of ITS actually and how is it compared to the whole-genome level? Will be important to at least discuss this somewhere.

RESPONSE: We agree that it is possible that there may be differences within genomes of organisms with closely related 16S/ITS gene sequences. We have added a section to the Introduction describing what is known in terms of genomic diversity in relation to ITS sequence diversity.

Lines 163-173. It is not clear what is compared in this paragraph. What do the eight clusters represent? I was able to understand more after I saw the corresponding figure 2, but the text should stand on its own. It would be useful to refer to the names of the samples like the authors do above. Figure 2 contains more than one cluster with green symbols.

RESPONSE: We have re-written part of this section to clarify the analysis, and added a section to the Figure 2 legend to clarify what each cluster represents.

Line 177. How reliable were these assemblies? Was their quality verified somehow? In general, the assembly part of the manuscript requires more attention by the authors and to be accompanied by basic assembly statistics such as coverage, number of contigs, possible mis-assemblies detected, etc.

RESPONSE: Assemblies were manually validated. A statement to clarify this has been added to the Materials and methods section “Genome assembly and annotation”.

> “Assemblies were manually inspected and validated using AMOS 3.1.0 (Phillippy *et al*, 2008) and Hawkeye 2.0 (Schatz *et al*, 2007). The mosaic draft genome of ACE_P1a.1 (approximately 1.26 Mbp) contained 79 contigs (N50 = 44.3 kbp) that were assembled into 12 scaffolds (N50 = 150.2 kbp) with 87.8% of the total base-pairs in contigs larger than 10 kbp). The mosaic draft genome of ACE_P3.2 (approximately 1.27 Mbp) contained 129 contigs (N50 = 15.2 kbp) that were assembled into 8 scaffolds (N50 = 329.6 kbp) with 73.1% of total base-pairs in contigs larger than 10 kbp.”

Line 181. I don't think you can safely draw this conclusion based on the relatedness observed. Organisms related at 86.3% genome average nucleotide identity typically belong to different species and frequently show considerable gene content and genome rearrangement differences.

RESPONSE: We have removed this conclusion as per reviewer suggestion and rephrased the paragraph. The genome assembly statistics that we have added should provide sufficient information to enable the reader to establish the reliability of the assemblies. Moreover, we avoid issues to do with subjective definitions about sequence identity (i.e. regarding “species”). Furthermore, our data are consistent with the recent report regarding sequence divergence of the SAR11 clade (Viklund *et al*, 2012).

Lines 189-192: The results described in this paragraph are not very clear. Where the average number of hits refers? Average from the 2 polar genomes compared? Average coverage of the genomes?

RESPONSE: This section refered to coverage. To also address comments from reviewer 3 (e.g. point 8) we have edited this section. The text now states.

> "The proportion of the normalized reads from the polar metagenomes that recruited to each genome representative was 60% for ACE_P3.2 vs 40% for HIMB114, and 65% for ACE_P1a.1 vs 35% for HTCC7211. The proportion of the normalized reads from the tropical metagenomes that recruited to each genome representative was 44% for ACE_P3.2 vs 56% for HIMB114, and 45% for ACE_P1a.1 vs 55% for HTCC7211. Similar trends were observed for temperate metagenome data (data not shown). All polar versus tropical/temperate differences were statistically significant (two tailed t-test $P < 0.005$)."

Lines 209-213. Hand waving, here and elsewhere. There is no real data to back up the claim that the gene content differences identified play a role in adaptation to polar environments. What about if the genes identified are just the product of mis-assembly? Or that adaptation is conferred at the transcription level of shared genes, not gene content? The authors need to qualify their conclusions better.

RESPONSE: Adaptation at the level of gene regulation is likely to be part of the story but this is not mutually exclusive with adaptation at the genomic level. The various levels of "adaptability" of a cell, whether it be epigenetic, post-translational modifications, transcriptional/translational regulation and how all these parts interact in a cellular system, are relevant to any complete consideration. However, the base-line is the genomic complement and this is the relevant part we focus on in this metagenome-based study. The question about potential mis-assembly can be addressed with reference to the sequence of strain IMCC9063. We have identified 96 genes under positive selection by comparing ACE_P3.2 and HIMB114. Amongst these 42 (>40%) were also identified as being positively selected when comparing the sequence of IMCC9063 to HIMB114. While it could be argued that only partial overlap exists between the 2 analyses, no overlap at all would be expected if the results were a product of mis-assembly. Moreover, similar classes of genes were detected suggesting redundancy and alternative pathways for functional adaptation. Because the genome of IMCC9063 is from a cultured isolate, rather than invoking potential synteny in the mis-assemblies of the respective genomes, the simplest explanation is that the differences reflect true functional adaptations. A further confirmation of this rationale comes from repeating a similar analysis on the genomes of ACE_P1a.1 vs. HTCC1062, which belong to the same group. If the observed differences were due to sequencing errors a similar number of positive hits would be expected. However, the number of genes detected in this case was much lower (26). Moreover, as discussed in the work by Rocha et al. (Rocha EP, Smith JM, Hurst LD, Holden MT, Cooper JE, Smith NH, Feil EJ. (2006) Comparisons of dN/dS are time dependent for closely related bacterial genomes. J Theor Biol. 239:226-35.), because of the high level of similarity between these 2 genomes, many of the genes under positive selection in this last comparison could be carrying slightly deleterious mutations not yet purged from the population amongst fewer positively selected adaptive mutations. For these reasons, the cause and nature of the differences observed can be debated but our data support that their existence is real.

Lines 214-219. More results about this section need to be shown in the main text; the reader should not be referred to the supplementary material for one of the two main lines of the paper (based on the abstract at least).

RESPONSE: We are not certain what reviewer 2 would like us to do with this section. It seems the reviewer argues strongly against the use of dN/dS in the context of detecting positive selection, yet requests more details be reported. We purposely avoided

overstepping the boundaries of our data in this section, and focused just on the stronger and broader results to avoid possible pitfalls related to the analysis (see responses below and above).

Lines 232-238. Dn/Ds analysis to detect positive selection has several limitations (see for instance Rocha et al., 2006 and Schmid and Yang 2009), none of which seems to be taken into account or at least discussed by the authors here.

RESPONSE: We agree that a dN/dS analysis can have pitfalls. However, a lengthy discussion on this topic is beyond the scope of this manuscript and is not warranted. For example, we agree that genome-wide detection of positive selection by this method is sensitive to sequencing errors, but, as noted above, this is not likely to have occurred in our analyses. It has also been suggested through comparative genomics and computer simulations (Rocha et al. 2006) that dN/dS comparisons between closely related strains might be skewed because mildly deleterious non-synonymous substitutions would not have had enough time to be purged from the population. However, these signatures are expected to plateau in a taxon-dependent fashion and can be detected by computing an average dN/dS ratio. In our case the average dN/dS ratio over all the pairwise comparisons of the available SAR11 genomes was in the range 0.1-0.3, and was fairly consistent regardless of the inferred time of divergence amongst the phylotypes. This is indicative of purifying selection being in place. It should also be noted that this slightly high average dN/dS ratio is in accordance with the recent report of the absence of a *mutLS* system from all the sequenced SAR11 strains (Viklund et al., 2012), as this has the potential to lead to increased evolutionary rates and up to 1000 fold nucleotide substitution rates. We have cited this study in the Introduction and in the section on "Genomic signatures of adaptive radiation".

Line 321. What do you mean? Do you mean that P1a.3 is the same as S1a? Please revise.

RESPONSE: We have revised this sentence to convey the meaning more clearly.

> "Our analysis shows P1a.3 is the dominant phylotype within the S1a subgroup in samples taken across a wide range of warm oligotrophic waters (Figure 1) indicating P1a.3 is likely the organism observed by Carlson *et al*, (2009) and other studies (Wilhelm *et al*, 2007)."

Lines 347-361. This section is just hand waving. I would recommend delete it completely as no data is available to back it up. For instance, why warmer phylotypes will invade polar regions as the latter become warmer and not polar phylotypes adapt to live at higher temperatures? Data is needed before one can make educated predictions about what may happen during these scenarios.

RESPONSE: We argue that this speculation is warranted as a natural extension of our analyses and provides an interesting dimension to prompt discussion and future hypothesis driven experimentation. Given that our study indicates that the SAR11 clade has evolved to proliferate across the temperature gradient that presently exists in global oceanic waters through the generation of distinct phylotypes (including genomic changes), the most parsimonious means of these communities responding to long term temperature increase would be for the adapted warmer water organisms to invade by outcompeting those that are adapted to cold waters. The converse that the cold water phylotypes adapt (through mutation or physiological adaptation) and resist the warmer water phylotypes from invading is a far more complex and involved scenario and therefore less likely.

Furthermore, in the Antarctic polar region, the Antarctic Circumpolar Current and associated Polar Front (the northern edge of the Antarctic Circumpolar Current) form a

biogeographic zone with a microbial community that is distinct from the zone north of the Polar Front (Chiba S, Ishimaru T, Hosie G, Fukuchi M (2001) Spatio-temporal variability of zooplankton community structure off east Antarctica (90 to 160°E). *Marine Ecology Progress Series* 216, 95-108; Esper O, Zonneveld KA (2002) Distribution of organic-walled dinoflagellate cysts in surface sediments of the Southern Ocean (eastern Atlantic sector) between the Subtropical Front and the Weddell Gyre. *Marine Micropaleontology* 46, 177-208; Hunt BPV, Pakhomov EA, McQuaid CD (2001) Short-term variation and long-term changes in the oceanographic environment and zooplankton community in the vicinity of a sub-Antarctic archipelago. *Marine Biology* 138, 369-381; Ward P, Whitehouse M, Brandon M, Shreeve R, Woodd-Walker R (2003) Mesozooplankton community structure across the Antarctic Circumpolar Current to the north of South Georgia: Southern Ocean. *Marine Biology* 143, 121-130; Abell GCJ, Bowman JP (2005) Ecological and biogeographic relationships of class Flavobacteria in the Southern Ocean. *FEMS Microbiology Ecology* 51, 265-277; Giebel H-A, Brinkhoff T, Zwisler W, Selje N, Simon M (2009) Distribution of Roseobacter RCA and SAR11 lineages and distinct bacterial communities from the subtropics to the Southern Ocean. *Environmental Microbiology* 11, 2164-217; Selje N, Simon M, Brinkhoff T (2004) A newly discovered Roseobacter cluster in temperate and polar oceans. *Nature* 427, 445-448; Weber TS, Deutsch C (2010) Ocean nutrient ratios governed by plankton biogeography. *Nature* 467, 550-554).

If the Polar Front shifts southward, which it is predicted to do as a result of global warming (Biastoch A, Boning CW, Schwarzkopf FU, Lutjeharms JRE (2009) Increase in Agulhas leakage due to poleward shift of Southern Hemisphere westerlies. *Nature* 462, 495-498; Boning CW, Dispert A, Visbeck M, Rintoul SR, Schwarzkopf FU (2008) The response of the Antarctic Circumpolar Current to recent climate change. *Nature Geoscience* 1, 864-869; Fyfe JC, Saenko OA (2005) Human-Induced Change in the Antarctic Circumpolar Current. *Journal of Climate* 18, 3068-3073.), the size of the Antarctic Circumpolar Current will shrink and along with it the cold dwelling communities. In this scenario, as the Polar Front heads south, the warmer waters will replace cold waters and bring with it the indigenous communities. While there is scope to extend the discussion we presented and include this type of information, we have kept this section brief and provocative, retaining focus on the biogeography and providing scope for follow up studies in this arena.

Although the text is well-written in general, with only a few grammatical errors or inaccuracies, the figures are not of the same quality. Figure keys, graph labels, and text should be added to the figures to make them easier to follow and stand on their own.

RESPONSE: All the colors in the figures have now been produced using a common color palette so that they tie together throughout the paper. Addressing comments of reviewer 3 as well (e.g. point 4), for Figure 1 we precisely matched the colors of the ITS tree with the colors in the bar graph. Given the clear distinctions of the phylotype groupings in the ITS tree, and the separation of phylotypes into 4 separate bar graphs, the figure should provide a clear and accurate visual aid for readers. We experimented with distinguishing biomes but in our view it only complicated the bar graph and did not facilitate ease of understanding, and therefore we made no further changes to the figure. In the Figure 2 legend (now panel E), we clarified which sequences were present in all 8 features on the plot. Figure 2 now includes what was Figure S2 (the panels from Figure S2 now precede the 2 panels that were in Figure 2), and collectively they describe the analyses of variance of the phylotypes and geolocation of the samples. For Figure 3 the genomes have been labeled on the figure, arrows included to highlight regions of low recruitment, and text has been added to the legend describing examples of how the color of the recruitment plots can be interpreted. As Figure S4 was referred to in the text a number of times and this figure provides a visual aid regarding positive selection, we moved this into the main text as Figure 5. Collectively Figures 4 and 5 effectively describe genomic signatures of adaptive radiation.

Reviewer #3 (Remarks to the Author):

This research analyzes a global genomic dataset to characterize and propose a temperature-driven biogeography of SAR11 ecotypes. Overall the authors were very clear in conveying the scientific motivations and study design. The analyses performed were generally thorough and appropriate, and provided compelling results. Although the main body of the manuscript was well written the figures could more clearly illustrate some of the points being referred to in the text (see specific comments below). In addition, two main issues need to be addressed regarding this manuscript.

First, while the data sufficiently supports the conclusion that SAR11 phylotypes biogeography is linked to temperature, from my perspective it seems difficult to eliminate confounding factors that could also explain the observed trends. For example, although the "polar" biome dataset very extensive, it is highly skewed. That is, 80% of the samples were collected from Antarctic waters and it appears that all of these show a consistent SAR11 community signature (i.e., dominance by P1a.1 and sub-dominance by P2.2). Conversely, the remaining 20% of "polar biome" samples were taken from North Atlantic waters and do not appear to consistently have the same trend in community composition. Admittedly, this is difficult to explicitly determine without sample labels on figure 1. Thus, it could be equally concluded that P1a.1 and P2.2 are not "polar" phylotypes on a global sense, but occur regionally in Antarctic waters.

RESPONSE: The number of sampling sites in the Arctic vs Antarctic is less relevant than the number of ITS sequences in our database that define the phylotypes and that have come from the Arctic and Antarctic. ITS sequences were previously obtained from studies describing Arctic samples, and the capacity of our present study to include the Antarctic derives from our research contributing 37 polar metagenomes and extracting ITS sequences from that shotgun sequence data. Given that 49.7% of the sequences in our ITS database are sourced from waters located in the Arctic or Sub-Arctic, while only 3% are sourced from Antarctic waters (Supplementary Table S1), by comparing Antarctic ITS sequences from our metagenomic samples to a majority of Arctic sequences in the database, the analysis is truly bi-polar. We have included the following statement in the Results:

> "Although the majority of polar metagenomic datasets originate from the Antarctic, it should be noted that 49.7% of the ITS sequences in our database are from samples taken in the Arctic while only 3% are derived from Antarctic samples (with the rest from temperate and tropical regions - Supplementary Table S1). Thus, by identifying phylotypes in Antarctic waters that have been defined using Arctic samples, the distributions of these phylotypes can be considered bi-polar."

Additionally, it is difficult to de-convolute the role of upwelling in the distribution of the phylotypes commonly found in colder waters. Given that SAR11 populations have been shown in previous studies to vary according to seasonal mixing and depth, it would seem important to address these confounding factors. The discussion touches on this issue but does not clearly indicate that the phylotypes found in the coldest waters are also not areas of active upwelling. It would seem appropriate to include an upwelling variable, if possible, in the dbRDA and DISTLM analyses to account for any community composition variability captured by this factor. Ideally, the appropriate samples and variable measurements to test these alternatives would be available. However, given the scope of this research it is understandable that this may not be possible. As such, appropriate considerations should be added to the discussion section if these concerns are valid.

RESPONSE: Regarding whether the affects of upwelling or mixing might confound the interpretation of the distribution of phylotypes, given the observed distributions of wind-driven upwelling/downwelling (ie Ekman pumping) and of mixed layer depth, we think this is unlikely. The large-scale pattern of wind-driven vertical velocity consists of upward motion in cold high latitude regions and in the warm tropics, and downward motion in the subtropical gyres. If the phylotype distributions were strongly linked to the pattern of Ekman pumping, we would expect a weak relationship with temperature (as both the warmest and coldest waters upwell, and waters of intermediate temperature downwell). Other factors, like eddies and convergent/divergent surface currents, can drive upwelling and downwelling, but there is no data with which to assess their global distribution.

Similarly, the depth of the surface mixed layer and its seasonal variation is not a strong function of latitude or surface temperature. Deepest winter mixed layers are observed on the equatorward side of strong mid-latitude currents (the so-called mode waters), while relatively shallow mixed layers are observed in the tropics and polar latitudes. There is also a strong contrast in winter mixed layer depth between the North Pacific and North Atlantic, despite similar surface temperatures (e.g. Tomczak and Godfrey, 1994). Summer mixed layer depths are relatively constant from the tropics to high latitude while summer sea surface temperature is a strong function of latitude. The observed strong association between phylotypes and temperature/latitude argues against a strong dependence on the depth of the surface mixed layer.

We have added to the Discussion words to this effect concerning upwelling in the context of the global patterns of SAR11 distribution.

The second issue is that the materials and methods section was superficial and needs more detail to allow readers to properly evaluate the methods applied.

RESPONSE: We have added a range of details and restructured some sections in the Materials and methods to provide more details.

Specific notes on proposed additions/changes to this section are provided below.

Minor comments:

1) Line 52: The term "robust" does not seem appropriate here. Perhaps "measureable" instead?

RESPONSE: This suggestion has been included.

2) Line 84: Add "latitude" after temperature since these factors could not be de-convoluted.

RESPONSE: This change has been made.

3) Line 90: There seem to be only 862 ITS sequences represented by the trees of figure 1 and figure S1, please account for this discrepancy. Also, sequences from the "NA" clusters seem to be mislabeled as the group of sequences with the prefix "NA1" contains 12 sequences, but is shown in figure 1 to have nine members and vice versa.

RESPONSE: We thank the reviewer for noting this and have corrected the figure.

4) Figure 1: It may be useful to highlight the "biomes" being referred to in the text in some way. Perhaps, simply dividing the background by lines, or colors according to temperature divisions? Also, the differentiation of colors for P2.1 and P2.2 in the phylogenetic tree is difficult for me to determine, although the bar graph is fine. Maybe a stronger contrast of colors could be used for these groups.

RESPONSE: We precisely matched the colors of the ITS tree with the colors in the bar graph. Given the clear distinctions of the phylotype groupings in the ITS tree, and the separation of phylotypes into 4 separate bar graphs, the figure should provide a clear and accurate visual aid for readers. We experimented with distinguishing biomes but in our view it only complicated the bar graph and did not facilitate ease of understanding, and therefore made no further changes to the figure.

5) Line 104: Change this sentence to note that at ~ 20 {degree sign}C P1a.2 begins to be the dominant P1a phylotype, since its appearance occurs at much lower temperatures.

RESPONSE: We replaced "….appears…" with "…dominates…" to clarify this point.

6) Line 111: Why was longitude not included as a test variable in the DISTLM and dbRDA models? I don't believe that its lack of significance should be ruled out a priori, if that was the case. I also wonder if date of sampling and/or a categorical variable to designate upwelling state would be useful to test, as mentioned above. Along these lines, is there a reason why the deep Mediterranean samples were not included in this analysis?

RESPONSE: We included longitude in our DISTLM analysis as the reviewer requested however its utility in describing patterns of phylotypes was weak and not significant (pseudo F= 5.7, p=0.007). We also included time of sampling as a variable and this was in fact a relatively weak but significant descriptor (pseudo=29, p=0.001). However, given that temperature and latitude both explain SAR11 phylotype distribution ~5 times better (pseudo ~169 and 167) and the fact that major sampling expeditions which have focused on the different biomes have been temporally separate (e.g. GOS sampling and Antarctic sampling separated by 3 years) we did not include the temporal variable in our discussion. To test what effect temporal succession might have would require repeat sampling/analyses of at least tropical and polar waters spanning a several year period. While this type of analysis, and in effect monitoring is warranted, it is well beyond the scope of this study. In the text we included the statement:

> "We also analysed datasets against longitude but these results were not significant and were removed from the model."

We could not determine an upwelling variable for the 128 samples but as described above, we have included a paragraph in the Discussion concerning upwelling and its likely effect on the global patterns we observe.

We have restricted our analysis to surface water samples only as there is sufficient data to perform robust statistical analyses, whereas there is presently insufficient data for deep samples.

7) Figure 2: It would be beneficial to include supp. figure S2 into this figure to illustrate how the dbRDA model is partitioning the variance.

RESPONSE: We agree and have incorporated these panels into Figure 2.

8) Lines 178 - 184: Please further justify the selection of P3.2 as a polar representative. I understand that its assembly from Ace Lake is the driving rationale, but this phylotype is largely uncharacterized and based on the ITS analysis P3.2 is found in both temperate and polar regions, a conclusion which seems based on very small sample sizes. It is understandable that the authors would want to include this genome in the comparative analysis, but it seems a bit suspect to rely on this genome as one of the two representative of the polar phylotypes. Perhaps it would be more convincing if the bias analyses performed were reported for P3.2 explicitly. For example, Lines 190-192 suggest a bias in blast hits between polar metagenomes and the two mosaic assemblies, but the individual hit rate between ACE_P1a.1 and ACE_P3.2 are not reported. Additionally, on Line 206 the percent of hits for unique ACE_P3.2 genes to polar metagenomes are not reported.

RESPONSE: We have greatly extended description about ANI and PCD comparisons. ACE_P3.2 has high identity to the Arctic isolate IMCC9063. ACE_P3.2 is therefore a good bi-polar representative. We have also included all the data for the genome

recruitments so the proportion of hits to the individual genomes can be considered. We state:

> "ACE_P1a.1 has equivalent genome length and average nucleotide identity (ANI) of 83.27% and percentage conserved DNA (PCD) of 6.16% (Figure 3; Goris *et al*, 2007) when compared to HTCC1062 (Giovannoni *et al*, 2005) and ANI of 76.12% and PCD of 0.41% when compared to HTCC7211. The length of the ACE_P3.2 and HIMB114 genomes were also equivalent (Figure 3) with ANI of 69.53% and PCD of 0.257%. The ANI and PCD values for ACE_P1a.1 compared to HTCC1062 are similar to that determined for *Shewanella* species, all of which had >70% ANI and >94% 16S rRNA gene sequence identity (Goris *et al*, 2007). The ANI and PCD values for ACE_P1a.1 compared to HTCC7211, and ACE_P3.2 compared to HIMB114 are similar to the values calculated for *Pseudomonas* species, which had PCD values ranging from 13.1 to 0.0001% (Goris *et al*, 2007). The Arctic strain, IMCC9063 (Oh *et al*, 2011) also has equivalent genome length, and has an ANI with ACE_P3.2 of 94.17% and PCD of 74.05 % which would be considered similar to a DNA-DNA hybridization of approximately 70% (Goris *et al*, 2007). Taken together these results support a strong correlation between ITS identity and genome similarity and are also consistent with the extensive sequence divergence recently reported for members of the SAR11 clade (Viklund *et al*, 2012)."

> "The proportion of the normalized reads from the polar metagenomes that recruited to each genome representative was 60% for ACE_P3.2 vs 40% for HIMB114, and 65% for ACE_P1a.1 vs 35% for HTCC7211. The proportion of the normalized reads from the tropical metagenomes that recruited to each genome representative was 44% for ACE_P3.2 vs 56% for HIMB114, and 45% for ACE_P1a.1 vs 55% for HTCC7211. Similar trends were observed for temperate metagenome data (data not shown). All polar versus tropical/temperate differences were statistically significant (two tailed t-test $P < 0.005$)."

9) Line 192: What is the statistical test used to denote this bias?

RESPONSE: A two tailed t-test was used. This was described in the last sentence of the Materials and methods section, "Genome alignment and recruitment". Also see point 8 above.

10) Figure 3: It would be useful for the reader if the regions of low recruitment referred to in the text (Lines 193 - 196) were highlighted or otherwise noted. Also, labeling the genomes with their names in the figure would be more helpful than referring to the legend for them. Finally, please indicate on the figure or in the legend the threshold level of % identity used to denote matching regions in the ACT plots.

RESPONSE: Arrows were added to the figure to point to two examples of low recruitment. The genomes were labeled. We didn't use a percentage cutoff but a match-length cutoff (which is the standard type of cutoff for ACT). We did clarify that:

> "blastn (run with standard parameters and the -m8 flag) matches of at least 30 nucleotides or longer are displayed."

11) Line 199: I may be confused but how many genes that are specific to the polar representatives were shared between phylotypes? (I.e., are any of the 344 and 444 unique

genes for ACE_P1a.1 and ACE_P3.2 shared between the two genomes? Or are they completely unique with respect to all genomes, not just the tropical genomes?)

RESPONSE: There was essentially no overlap (1 gene) between the gene sets the reviewer highlights. This is not an unexpected result. These phylotypes are not closely related. There are few reports in the literature of individual genes conferring thermal adaptation. Rather, it appears that specific categories of genes are involved in the process. Therefore it is not unexpected to find very little overlap between the genes under positive selection in ACE_P3.2 and ACE_P1a.1. By analogy this would be similar to asking why isn't there extensive gene overlap between uropathogenic *E. coli* CFT073 and enterohemorrhagic *E. coli* EDL933 just because both are pathogenic (see Welch et al., 2002). To be as accurate as possible, we changed "strain-specific" or "unique" to "genome-specific".

12) Line 206: The methodology to assign significant biases is unclear, please elaborate on this more in the Materials and Methods section.

RESPONSE: We have added a paragraph to the Materials and methods describing the resampling algorithm used to detect statistically significant biases in the number of hits of the phylotype specific proteins.

13) Line 270 - 271: Since the effects of upwelling, and thus greater nutrient concentrations, cannot be ruled out at this point, please note that selection of transporters could also be a result in differences in substrate concentration and associated KM values.

RESPONSE: We note the reviewers comment but suggest that our data are not significantly biased by upwelling – see responses above.

14) Line 282 - 290: Most genes demonstrating positive selection seem to be specifically linked to cold adaptation (Lines 237 - 238). Therefore, the impact of this finding does not seem to extend to the other differences in environment mentioned beyond temperature (e.g., salinity, mineral concentrations, nutrient inputs).

RESPONSE: This is correct. We agree it would not be sensible to extend the analyses beyond temperature.

15) Lines 298 - 300: If the data supporting this conclusion is figure 2 alone, it does not seem justified. For example, using a similar logic it could be said that Monterey Bay waters (Mont_Bay_1) influence the community in the North Atlantic (Gilbert_Apr_day).

RESPONSE: We have removed this sentence.

16) Lines 304 - 307: Along with the previous comment above, I'm not sure much can be said about these samples given the results presented.

RESPONSE: We have also removed this sentence.

17) Line 317: To avoid confusion, please consistently refer to the subgroups as SIa/SIb or S1a/S1b.

RESPONSE: These designations have been standardized.

18) Line 323: I am assuming due to technological differences it can not be determined if there was a consistent phylotype of the S1b subgroup associated with these mixing/upwelling events across sites?

RESPONSE: That is correct – the previous studies used 16S rRNA gene sequencing which does not discriminate between the ITS phylotypes we have identified.

Materials and methods questions/comments:

19) Is the sequencing technology the same for all metagenomes? If not, please add this information to one of the supplemental tables (see comment 23 below).

RESPONSE: This information has been added to Supplementary Table S2.

20) Line 384: No reference or website is given for the ITS database.

RESPONSE: We have clarified this by stating:

> "The annotated database and ITS alignment files are available from the authors upon request."

21) Lines 378 - 396: The section describing ITS identification and how the abundance analysis was performed needs much greater detail.

It is assumed that the analysis was performed with reads and that some sort of clustering was performed (similar to pyrotag studies?). If this is true, how are the 2,983 markers clustered into 865 ITS sequences? Is the discrepancy between these numbers due to redundancy? Are clusters generated by grouping perfect matches together with the longest read representing the consensus sequence to remove any redundancy? Is abundance then determined by simply summing the number of reads in each metagenome assigned to each of the 865 ITS sequences and dividing by the total? Once abundances of each phylotype were assigned by sample, were the assignment distributions relatively even or could specific samples be classified as outliers based on low total numbers? Or how many samples of the 127 were represented by only one phylotype with one read?

RESPONSE: We have re-written the section describing how the data were collected and analysed in order to clarify. See section entitled "Determination of phylotype abundance in metagenomic samples and biogeography".

22) Line 389: Is temperature considered a continuous variable in the dbRDA or categorical by biome? Also, what is the "chlorophyll" variable a measure of (total Chl, ChlA, ChlB)?

RESPONSE: Yes temperature is considered a continuous variable. Chlorophyll is a measure of ChlA.

23) Line 393: How many samples did not have all associated environmental variables and were then estimated? It would be useful to include all of these measured and estimated variables in a supplemental table along with the sequencing technology information for each metagenome.

RESPONSE: This was clarified in the text where we stated,

> "However, all patterns reported were checked for consistency against the subset of 79 metagenomes for which all variables were present."

We have also added a new Supplementary Table S3 that contains the environmental data for each metagenome, including indicating which measures were estimated. The

sequencing technology used for each metagenome has been added to the new Supplementary Table S2.

24) Line 395 & 397: How were variables normalized? The text mentioned that a square-root transformation was used on the relative abundances (as opposed to the raw count data). Was there a reason an arcsin transformation was not used instead? I am of the impression that this type of transformation is more appropriate for percentage data (Quinn and Keough, 2002, pg. 66).

RESPONSE: To normalize the data, the values for each variable have their mean subtracted and divided by their standard deviation. We have added this information to the Materials and methods. The square root transformation is generally appropriate for these types of environmental data.

25) Lines 403 - 408: It may be useful to state within the text the phylogenetic diversity captured by each of the phylogenetic levels you describe (i.e., clade, subgroup, and phylotype). For a more universal point of reference, it may be best to use the 16S tree from figure S1 to calculate these values from.

RESPONSE: Based on comments from both reviewer 1 and reviewer 3 we have added a new Table 1 which outlines the clade, subgroup, phylotype hierarchy along with other pertinent information including providing the phylogenetic diversity captured at each level based on full-length 16S rRNA gene sequences used in Supplementary Figure S1.

26) Line 478: How were scaffolds ordered? (Alignments to known genomes?)

RESPONSE: The scaffolds were orientated according to the closest known reference genome sequence and a statement clarifying this has been added to the Materials and methods section “Genome alignment and recruitment”.

27) Line 493: Was depth measured at each base position of the genome and averaged? Was the variation accounted for on the same scale?

RESPONSE: The recruitment depth was measured at each base position and averaged. The variation was accounted for by choosing metagenomic datasets that were very similar in size as described in the Materials and methods.

28) Line 500: I'm a bit confused about how paralogs are being identified. Can you clarify what is the query and database in each case?

RESPONSE: The initial query and subject databases are the two predicted proteomes of the genomes. During the first run, the reciprocal smallest distance (RSD) algorithm identifies the most probable orthologs based on the reciprocal smallest distance algorithm. After removing the identified orthologs from the query proteome and repeating the RSD analysis, another set of homologs is identified. Because this was not identified in the first round it is assumed to be a set of paralogs. The process is repeated recursively until no new matches can be found.

29) Line 501 - 507: Can more details on the resampling method discussed here be provided? I am assuming it is to test the biases present in the distribution of hits to the subdivided read sets according to water temperature, but it is a little unclear how this is being tested. Also, do repeat regions strongly influence the analysis or are hits widely distributed among the reads and across the whole protein query sequences?

RESPONSE: That is correct: resampling is used to test whether the distribution of the hits of phylotype-specific proteins to different metagenomes is correlated with the temperature of the metagenomes. We have added a description to the Materials and methods section "Detection of homology, COG identification and analysis" describing in detail how the resampling method was implemented. We are unsure what reviewer 3 means by influence of protein repeats as these would be masked during the tblastn searches.

30) Line 515: Change "proteins that" to "protein pairs that", if this is what is meant.

RESPONSE: This change has been made.

We hope you find that the revised manuscript appropriately addresses the reviewers' concerns and is now suitable for publication in *Mol Syst Biol*.

2nd Editorial Decision 13 June 2012

Thank you again for submitting your work to Molecular Systems Biology. We have now heard back from the two referees who agreed to evaluate your revised study. As you will see, the referees felt that the revisions made to this work had satisfied their main concerns, and they are now largely supportive. The last reviewer has an important remaining concern, and we have some minor formatting issues, which we would ask you to carefully address in a final revision of the present work.

The last reviewer is still not convinced that one can reliably conclude that P1a.1 and P2.2 phylotypes are dominant in Arctic waters, given the more limited Arctic sampling. While this reviewer does not seem to feel that this point is sufficient to undermine the value of this work, this point will need to addressed rigorously, possibly with additional statistical analyses, before this work would be appropriate for publication. For example, it may be helpful to present a direct comparison of the distributions in the Arctic vs Antarctic samples. Clearer discussions of the potential caveats imposed by the more limited sampling Arctic sampling also seem needed.

In addition, when preparing your final revision, please address the following format and content issues:

1. The editor has some concerns regarding the presentation of the current figures. Please make sure that all fonts used within the figures are easily readable when the figures are printed to typical single page dimensions (including numbers in scale bars and axis labels). Also, some of the text in Fig. 2 is blurry and should probably be remade in a vector graphics program like Illustrator or Inkscape. Please also correct the overlapping labels in panel D. The labels in panel E are largely unreadable so I would advise removing them entirely, and adding back some larger text labels for the groups defined by the circles and for particularly important samples or outliers. The map image in panel F could also be improved, possibly by using a simpler map that does not include the currently unreadable geographical labels.

2. The Supplementary Information pdf should begin with a Table of Contents listing all supplementary materials, and the supp. figure legends should be placed directly below the appropriate Supp. Figures.

3. If possible, we encourage you to provide machine readable versions of the neighbor joining trees in Fig. S1, ideally in a common community-standard format (e.g. Newick or Nexus)? These supplemental tree files should be listed in the Supp. Information Table of Contents as "separate files".

4. Lastly, the current thumbnail is well made, but we may wish to run an image with this work that better emphasizes the ecological and biogeographic focus. If you have compelling images that you think might be useful please feel free to include them with your resubmission (for example, images of marine bacteria or the ocean environments from where they were sampled). This may help our artists when they are composing the final thumbnail and visual title images. http://mts-msb.nature.com/letters/msb_copyright.pdf

Thank you for submitting this paper to Molecular Systems Biology.

Yours sincerely,
Editor - Molecular Systems Biology
msb@embo.org

Referee reports:

Reviewer #2 (Remarks to the Author):

The revised article by Brown et al., has addressed, more or less, all my major comments related to the resolution provided by ITS sequence analysis, the quality of metagenomic assemblies, the inference of adaptive (positive) selection in the cold-adapted SAR-11 phylotypes and the comparative genome analysis of selected SAR-11 genomes or assemblies. I have no major issues remaining with the manuscript.

Reviewer #3 (Remarks to the Author):

I would like to thank the authors for their detailed and thoughtful responses and thorough corrections to manuscript. The only outstanding issue of note is that I may not have been clear in my first comment regarding the conclusions drawn about the P1a.1 and P2.2 "polar" phylotypes. The authors reply: "RESPONSE: The number of sampling sites in the Arctic vs Antarctic is less relevant than the number of ITS sequences in our database that define the phylotypes and that have come from the Arctic and Antarctic... by comparing Antarctic ITS sequences from our metagenomic samples to a majority of Arctic sequences in the database, the analysis is truly bi-polar." I agree that the methodology is not in question and that the database is sufficient. However, my comment was directed to one the conclusions of this analysis (i.e., "In polar waters, P1a.1 (HTCC1002, HTCC1062) and P2.2 dominated..."; Line 117). These phylotypes are clearly dominant in Antarctic waters, but does the data support their dominance in Arctic samples as well? Maybe for P1a.1 (although speculating beyond the fjord samples seems troublesome), but I do not see evidence of P2.2's dominance. Again, my conclusions are based on examining figure 1 and attempting to match that data to sample desciptions in the supplemental tables. Thus, I could be wrong in my understanding of the underlying data. If my point is valid, though, I believe the conclusions from this section need to be modified to account for the lack of support. The only other comments I have are with regard to typos found: (1) Figure 5E - genome "IMMCC9063" => "IMCC9063"; Line 696 (comparison doc) - Should this reference Supp. Table S3?; and Line 702 (comparison doc) - "PERMANOVER" => "PERMANOVA+".

2nd Revision - authors' response 15 June 2012

We provide a revision of our manuscript MSB-12-3650R entitled, "**Global Biogeography of SAR11 Marine Bacteria**", authored by Mark V. Brown, Federico M. Lauro, Matthew Z. DeMaere, Les Muir, David Wilkins, Torsten Thomas, Martin Riddle, Jed A. Fuhrman, Cynthia Andrews-Pfannkoch, Jeffrey M. Hoffman, Jeffrey B. McQuaid, Andrew Allen, Stephen R. Rintoul, Ricardo Cavicchioli.

Thank you for the encouraging news in your decision letter June 13. Below are responses to reviewer and editorial comments.

Reviewer #3 (Remarks to the Author):
I would like to thank the authors for their detailed and thoughtful responses and thorough corrections to manuscript.

The only outstanding issue of note is that I may not have been clear in my first comment regarding the conclusions drawn about the P1a.1 and P2.2 "polar" phylotypes. The authors reply: "RESPONSE: The number of sampling sites in the Arctic vs Antarctic is less relevant than the number of ITS sequences in our database that define the phylotypes and that have come from the Arctic and Antarctic... by comparing Antarctic ITS sequences from our metagenomic samples to a majority of Arctic sequences in the database, the analysis is truly bi-polar."

I agree that the methodology is not in question and that the database is sufficient. However, my comment was directed to one the conclusions of this analysis (i.e., "In polar waters, P1a.1 (HTCC1002, HTCC1062) and P2.2 dominated..."; Line 117). These phylotypes are clearly dominant in Antarctic waters, but does the data support their dominance in Arctic samples as well? Maybe for P1a.1 (although speculating beyond the fjord samples seems troublesome), but I do not see evidence of P2.2's dominance. Again, my conclusions are based on examining figure 1 and attempting to match that data to sample desciptions in the supplemental tables. Thus, I could be wrong in my understanding of the underlying data. If my point is valid, though, I believe the conclusions from this section need to be modified to account for the lack of support.

RESPONSE: We agree with the reviewer that there is reason to clarify the statement so that P1a.1 and P2.2 are dealt with separately. Evidence based on clone libraries (sequences forming the database) shows that the majority of SAR11 sequences in the Arctic surface waters belong to the P1a.1 phylotype. This does not extend to the P2.2 phylotype which is only defined by sequences from Antarctic waters. We have modified our statements to say that P1a.1 displays a bi-polar distribution, while P2.2 is currently only described in Antarctic data and four northern hemisphere cool/temperate metagenomes. We now state in the Results:

> "In cold water samples, P1a.1 (HTCC1002, HTCC1062) dominated. Although in lower total abundances, P2.2 was more abundant than P2.1 in waters <5°C and P3.2 was present only in waters <18.2°C. Although the majority of polar metagenomic datasets originate from the Antarctic, it should be noted that 49.7% of the ITS sequences in our database are from samples taken in the Arctic while only 3% are derived from Antarctic samples (with the rest from temperate and tropical regions - Supplementary Table S1). Thus, by identifying P1a.1 phylotypes in Antarctic waters that have been defined using Arctic samples, the distribution of

> this phylotype can be considered bi-polar. Phylotype P2.2 was defined from Antarctic waters (Garcia-Martinez and Rodriguez-Valera F, 2000) but is not represented in Arctic clone libraries. This phylotype is characteristic of Antarctic water metagenome data (Figure 1) and also appears in metagenomes from the English Channel (Aug10pm, 15.8°C), Delaware Bay (GS011, 11°C), Newport Harbor (GS008, 9.4°C) and Nags Head (GS013, 9.3°C). Its distribution in Arctic waters will be able to be determined when metagenome data becomes available."

And in the Discussion:

> "The co-occurrence of different phylotypes from subgroups S1 and S2 in tropical (P1a.3 and P2.1), bi-polar (P1a.1) and Antarctic (P2.2) waters (Figure 1) indicates that these subgroups have undergone adaptive radiation generating phylotypes that have distinct temperature preferences."

The only other comments I have are with regard to typos found: (1) Figure 5E - genome "IMMCC9063" => "IMCC9063"

RESPONSE: This has been changed.

Line 696 (comparison doc) - Should this reference Supp. Table S3?

RESPONSE: We have now cited both Table S3 and S4 as they are both relevant to the statement.

and Line 702 (comparison doc) - "PERMANOVER" => "PERMANOVA+".

RESPONSE: This has been changed.

Editors comments
In addition, when preparing your final revision, please address the following format and content issues:

1. The editor has some concerns regarding the presentation of the current figures. Please make sure that all fonts used within the figures are easily readable when the figures are printed to typical single page dimensions (including numbers in scale bars and axis labels). Also, some of the text in Fig. 2 is blurry and should probably be remade in a vector graphics program like Illustrator or Inkscape. Please also correct the overlapping labels in panel D. The labels in panel E are largely unreadable so I would advise removing them entirely, and adding back some larger text labels for the groups defined by the circles and for particularly important samples or outliers. The map image in panel F could also be improved, possibly by using a simpler map that does not include the currently unreadable geographical labels.

RESPONSE: As much of the artwork as possible in all figures has been changed to vector graphics sources. In addition, the font size, particularly in figures specifically mentioned, was increased for legibility and Helvetica font was used where possible. Overlapping labels in Figure 2D have been fixed to ensure readability and arrowheads added to clarify the presence of colinear vectors. Labels have been removed from Figure 2E and the images and ring redefined for clarity. Some labels (in ocean regions) have been removed from the map image (Figure 2 F) to reduce clutter. Minor changes were made accordingly to the legend of Figure 2. The spelling mistake noted by reviewer 3 in Figure 5 has been corrected. In addition to high quality pdf, we have provided the compiled InDesign figures in case they may be useful.

2. The Supplementary Information pdf should begin with a Table of Contents listing all supplementary materials, and the supp. figure legends should be placed directly below the

appropriate Supp. Figures.

RESPONSE: This has been done.

3. If possible, we encourage you to provide machine readable versions of the neighbor joining trees in Fig. S1, ideally in a common community-standard format (e.g. Newick or Nexus)? These supplemental tree files should be listed in the Supp. Information Table of Contents as "separate files".

RESPONSE: This has been done – three tree files have been provided.

4. Lastly, the current thumbnail is well made, but we may wish to run an image with this work that better emphasizes the ecological and biogeographic focus. If you have compelling images that you think might be useful please feel free to include them with your resubmission (for example, images of marine bacteria or the ocean environments from where they were sampled). This may help our artists when they are composing the final thumbnail and visual title images.

RESPONSE: Thank you for considering ways to visually highlight our work. I have provided two low resolution pdf posters that include a total of 36 images of scenes of the Southern Ocean. If any of these are useful for constructing the thumbnail I can provide the original image files.

Regarding your question about the licence to publish, please note that I received an email June 6 from Jana Christopher indicating that the licence to publish that I had sent had arrived safely, so I presume this is now finalized?

We hope the revised manuscript appropriately addresses everyone's concerns and is now suitable for publication in *Mol Syst Biol*.