Review I:

The authors propose an approach based on chemostat model analysis to explain phytoplanktonic biodiversity and species successions through different scenario representing various seasons and latitudes. The model combines genomic information to the uptake and growth rate of 3 N and 3 P substrates. Changing environmental conditions are then simulated, in order to simulate phytoplankton community structure.

This work can be seen as an attempt to give more mechanistic explanations to the work of Follows et al., but I don't think it achieves this goal.

As described in the abstract, the main purposes of our model are to first develop a new biogeochemical model that includes phytoplankton genomic information; secondly, to link genetic activation with biogeochemical processes; and finally to study phytoplankton adaptation under the changing environments.

Although the method we use to assemble the phytoplankton communities is similar to that developed by Follows et al. (2007), our overall methodology differs from theirs in that each type of phytoplankton in our model is able to adapt to available nutrient concentrations by using different pathways of nutrient utilization encoded in its genome. A by-product of developing this model is that it does indeed provide a possible mechanistic grounding for the methodology of Follows et al. (2007), but this was not the main aim of the project.

The three principal limit of the paper are 1/ the frightening assemblage of hypotheses, most of them being difficult / impossible to verify, 2/the use of chemostat framework to extrapolate results in the ocean 3/ the lack of systematic simulation exploration of the possible parameter values:

We agree with the reviewer that the model contains some key assumptions. Some of them may be tedious and time-consuming to verify, but we disagree that they are inherently impossible to verify. For example, we know qualitatively that phytoplankton with larger genomes tend to have lower maximum growth rates (Hessen et al., 2009). But the quantitative relationship between genome size and growth rate is not known for some phytoplankton types in our model. Because of this, we run a sensitivity analysis to quantify how our hypothesized relationship will affect the model output. Results from this analysis are shown at the end of this reply.

1- The set of hypotheses which are assumed in order to derive the model strongly limit the breadth of the model. The list of fragile and unverified hypotheses gives the feeling of an abstract exercise.

Equations (5), (6) and (8) would require more experimental support.

The aim of equations (5)-(6) is to provide a simple mathematical representation of nutrient uptake when organisms have the ability to utilize membrane transporters under

low nutrient concentrations. This process of facilitated diffusion is best described used a reaction-diffusion equation which requires knowledge of both diffusion and reaction rates; the latter are known for only a very few, specialized cases. So, to simplify our model, we assume that the chemical reaction rate is much faster than the diffusion rate resulting in a diffusion-limited process.

In order to clarify the above process, we have altered the text on "p. 821, l. 21-22" to read *"Through the further assumption that nutrient elements and specialized enzymes immediately react as soon as they encounter each other at the cell surface (Atkin, 1998; Eigen and Hammes, 2006), the nutrient uptake is limited only by diffusion process and calculated as the nutrient diffusive flux to the surface of a cell of radius R (Jumars et al., 1993; KarpBoss et al., 1996) by*:"

We use equation (5) to describe transport by passive diffusion for NH_4 and PO_4 only when they are abundant. Nutrient transport to the cells is by passive diffusion for all phytoplankton types in our model regardless of genetic combination. Compared to equation (6), Eq (5) indicates that nutrients (NH_4 and PO_4) around the cell surface won't be completely removed. In our model, we use the critical extracellular concentrations ($C_{0,j}$) to quantify the lower limit to the nutrient concentration which still allows for uptake by passive diffusion.

We made this assumption because 1) there is a lack of experimental results that would allow for a more detailed model to be used (as pointed out by the reviewer) and 2) the use of $C_{0,j}$ provides a transition from passive to facilitated diffusion and a convenient way to link environmental changes of nutrient concentrations with genetic activation. We agree with the reviewer that it is important to examine the sensitivity of the model to the values of the constants chosen and so we have run a series of sensitivity analyses to test the influence of different choices of $C_{0,j}$ on the model results (see the end of this document).

Equation (8) serves several purposes. It is well known (e.g. Follows et al., 2007) that this type of simulation can easily create a "super-organism" that can outcompete all others. Equation (8) helps to avoid this situation by reducing the growth rate of organisms with large genomes (i.e. those organisms with greater flexibility in nutrient acquisition etc.). Although a quantitative understanding of how genome size affects cell maintenance costs and growth rate is lacking, there is some evidence of an inverse relationship between growth rate and genome size (e.g. Hessen et al., 2009). Because of this lack of quantitative information we chose to use a very simple form for equation (8) that expresses the core of the idea.

Maximum and minimum cell quota are derived from allometric hypotheses, but the publications which are mentioned do not refer to a nitrogen or phosphorus quota, but to the chla quota, this is rather different.

The reviewer is correct and this was our mistake! The reference cited was from an old version of the manuscript and does not reflect the values that were used in the model. The model used P and N cell quotas obtained from Shuter (1978). We have corrected this in

the manuscript by giving the correct references and citations: the references (p.821, l. 14 & 15) have been adjusted to (Shuter 1978; Blasco et al., 1982).

Moreover, the parameter choice is not very convincing and seems very arbitrary. This is particularly true for the parameters of Table 3. The authors should give a better justification of this choice, and more important, they should assess the influence of a given parameter set on the model results (sensitivity analysis). The way the parameters are chosen is unclear, especially for the parameters which play a key role, such as the critical extracellular concentration.

We thank the reviewer for the suggestion of performing a sensitivity analysis. We have done this for the critical parameters r (the growth reduction constants that appear in equation (8)) and C_0 (the critical extracellular nutrient concentration) and will include a section in the manuscript covering this — the proposed section is included as an appendix to this response. Other parameters in Table 3 are either taken from experiments (such as diffusion coefficients for each nutrient) or widely used in literatures (such as minimum and maximum intracellular N and P content).

2- This is not clear how many simulations have been run to get the results, and if there is a preliminary simulation to reach periodic values, and how many seasons are tested. The results should be stable despite various initialization and parameter choice.

The following paragraph has been added on "p. 826, l. 4" to make the manuscript clearer. *"The initial values of nutrient concentrations are set to zero among all scenarios. The initial values for phytoplankton cell density and cell quota are 10 cells/l and average value of maximum and minimum cell quota respectively. The model run for 5 five years and after a quasi steady-state is established for all variables in the model, the last year's results are presented in the next section."*

Despite not having light controls, the patterns of phytoplankton community diversity in the model scenarios share some similarities with observed biodiversity patterns. This assertion seems rather weak and too vague. It should be supported by statistical analysis of large simulations sets initiated from arbitrary choices of parameters and initial conditions. This property could be claimed only if it turns out to be stable, for randomly chosen initial conditions, parameters and dilution rate (in a given range).

As shown in the sensitivity analysis, the dominant species in experiment I are almost the same as the base when we double the C_0^{NH4} or C_0^{PO4}. The relative percentage difference of phytoplankton biomass is always below 10%. Although nearly half of the dominant species are different from the base in the experiment II, the small, slow-growing phytoplankton types are still the most common species among the dominant species under the settings of Scenario II.

3- The reduction of the model to a CSTR is of course a tremendous assumption which strongly shorten the work scope. Moreover, the choice of the dilution is arbitrary, it is well known however that it can be determinant for the competition outcome.

As we stated in the manuscript, this model is meant to be a proof of concept, and we are currently working to extend the model to more realistic settings such as the Amazon River Plume where we have both biogeochemical and omics data. The current focus is to form a model framework to include genomic information, predict gene expression patterns, and study phytoplankton adaptation. As a result, we have attempted to keep the model as simple as possible.

We agree that the choice of dilution rate will affect the competition outcome. For example, if dilution rate is higher than maximum growth rate, nothing will survive. On the other side, if dilution rate becomes zero, phytoplankton community will be eventually dominated by slow-growing type of small phytoplankton. We chose our dilution rates to represent, in some broad (though admittedly highly simplified) sense, nutrient input to the surface waters from mixing. Although the choice of dilution rate is arbitrary in both scenarios (dilution rates change in Scenario I from 0.1 to 0 and from 0.01 to 0 in Scenario II), a higher dilution rate (0.1) was used in Scenario I in the spring in order to represent stronger mixing from below the mixed layers. Lower dilution rates (0.01) were used in Scenario II in the spring to represent weaker mixing. A zero dilution was used in the middle of the year to represent the ocean surface during the summer stratification period.

4- Other comments: Because of the crucial role of the nitrogen fixing cyanobacteria which is highlighted by the model, a more accurate representation of the underlying mechanism should be included. It should include the (high) energetic cost to activate the nitrogenase.

We agree that the nitrogen fixing cyanobacteria play a crucial role in the model that the activation of the nitrogenase is associated with the high cost. We did not wish to develop a whole cell model (it is impractical to embed such models into large scale biogeochemical models) and so tried to develop a simpler model that encapsulated the main features of the underlying mechanisms without introducing too much detail. To that end, the growth penalties used in equation (8) depend on the genes carried. As can be seen in Table 3, phytoplankton having the "*nif*" gene cluster receives the highest reduction of maximum growth (12). Therefore, the tradeoff associated with the higher energetic costs of N fixation is lower maximum growth rate when cyanobacteria are fixing N in a DIN deplete environment.

Two crucial parameters (light and temperature) are not represented, and they may have a large impact stronger than the nitrogen and phosphorus limitations.

We agree that crucial environmental parameters like light and T may have a strong impact on phytoplankton community and should be considered in a fully predictive model. However, the goal of this paper is to introduce the model framework as a proof of concept and to show that the model structure captures certain features seen in natural communities; in particular, that communities can adapt to changing environmental conditions Therefore, we consider our study as an initial attempt to build an ecological model that integrates microbial genomics and predicts gene expression patterns. It is far

from being a mature model, but the initial results are sufficiently encouraging for us that we are developing a more predictive version of the model.

The genomic side in the model is very rough and simply consists in assuming some cell capability, and it links the number of these pathways to a maximal growth rate value. At the end, there is no real contribution of any genomic knowledge, and I have the feeling that the results obtained by this approach are less clear than for Follows approach, despite more fragile underlying hypotheses. Finally, the gain of the approach is not clear.

The aim of this manuscript was to present a modeling framework that allowed for a link to be made between genomic information, biogeochemistry and community structure. We did not intend the manuscript to contribute to genomic knowledge per se, but instead to provide a framework for synthesizing genomic, biogeochemical and community structure information. We agree that the representation of genetic information in the model is very simple indeed. However, as the reviewer has alluded to, whilst there is abundant qualitative genomic information, there is little quantitative information (though that is changing). As the reviewer appreciates, this hinders any modeling effort. So, the question arises of how to incorporate the available genomic information into a model with biogeochemistry? To do this we have adapted the approach taken by Mick Follows and colleagues and extended it by developing a means to allow the organisms to adapt to changing environmental conditions by giving them suites of pathways that can be used under different conditions. In doing so we can track the changes in activation of these pathways and thereby provide a link to the -omics information available. For example, we can track the relative changes in activation of a certain pathway as environmental conditions change and compare that with quantitative changes seen in the –omics information. This link between the –omics information, biogeochemical information and community structure is what we gain by this approach. However, in order to do this, we have had to make simplifying assumptions. The initial "proof of concept" version of the model gives results that are generally satisfactory. The next phase is to incorporate this model into a larger framework and make comparisons with specific observations.

The authors should work hard to address these points and propose a more convincing revision.

Minor comment:

Equation (2) should involve μ, instead of u, this seems to be simply a typo.

This has been corrected.

The meaning of the parameters is unclear and unconsistant. What the authors denote μmax is not the maximum growth rate. The maximum growth rate depends on the ratio Qmin/Qmax. Change μ'max into \bar \mu to avoid any confusion. But then, several relationships along the paper should be revisited.

μ_{max}: in the modified version of our manuscripts, we will use μ_{max} to denote "maximum

potential growth rate" that only depends on the cell size, and call μ'_{max} the maximum growth rate (pp 823, ll 10).

Appendix:

We propose to run a series of sensitivity analysis and will add the following parts in the method and result sections of the manuscripts.

2.8 Sensitivity Analysis:

Values for two crucial model parameters are unknown. These are the growth reduction factor (r) and the critical extracellular nutrient concentration (c_0) that controls the activation of the functional gene clusters. We examined the effect that the chosen values of these parameters have on the model results using a sensitivity analysis. The baseline model run used for this was scenario II and the 8 parameters were doubled individually (Exp I.1, c_0^{NH4} doubles; Exp I.2, c_0^{PO4} doubles; Exp II.1, r_{amt} doubles; Exp II.2, r_{nr} doubles; Exp II.3, r_{nif} doubles; Exp II.4, r_{pst} doubles; Exp II.5, r_{pho} doubles; Exp II.6, r_{phn} doubles) and the resulting total biomass and dominant species were compared with the baseline model run.

3.4 Results of sensitivity analysis

In general the total biomass predicted by the model was insensitive to the values for C_0, with maximum relative differences of about 6% from the baseline simulation (Figure A.1). These differences were seen when C_0^{NH4} was varied, and the model showed no discernable difference from the baseline simulation when C_0^{PO4} was varied. Differences in biomass were significantly greater when varying the growth reduction parameters (Figure A.1). Differences from the baseline simulation were reasonably small (< 15%) when r_{amt} was doubled, and slightly larger relative differences (< 25%) were seen when r_{cop} was doubled. Variations in the remaining growth reduction parameters were larger (< 35%)

The different choices of parameters (C_0 and r) also change the dominant species in the model results. In experiment I, increasing C_0^{NH4} and C_0^{PO4} does not significantly change the pattern of dominant species. Compared to the baseline simulation, the dominant species are generally small, slow-growing phytoplankton that contains P related gene clusters (Figure A.2). In experiment II we increased the growth reduction constants and this led to a change in which species dominated. The pattern of this change is especially obvious in experiment II.2, 3, 6; species that were dominant in the baseline simulation had lower relative abundances because of the increased penalty on growth. For example, phytoplankton types that possess *nr*, *nif,* and *phn* gene cluster are commonly found in the baseline simulation, but only one species having the *nr* gene is dominant in experiment II.2, one having *nif* gene in experiment II.3, and one having *phn* gene in experiment II.6 (Figure A.2).

Figure A1: Sensitivity analysis of phytoplankton biomass.

(Relative Difference (%) = $\frac{(Experiment-Base)}{Base} \times 100\%$)

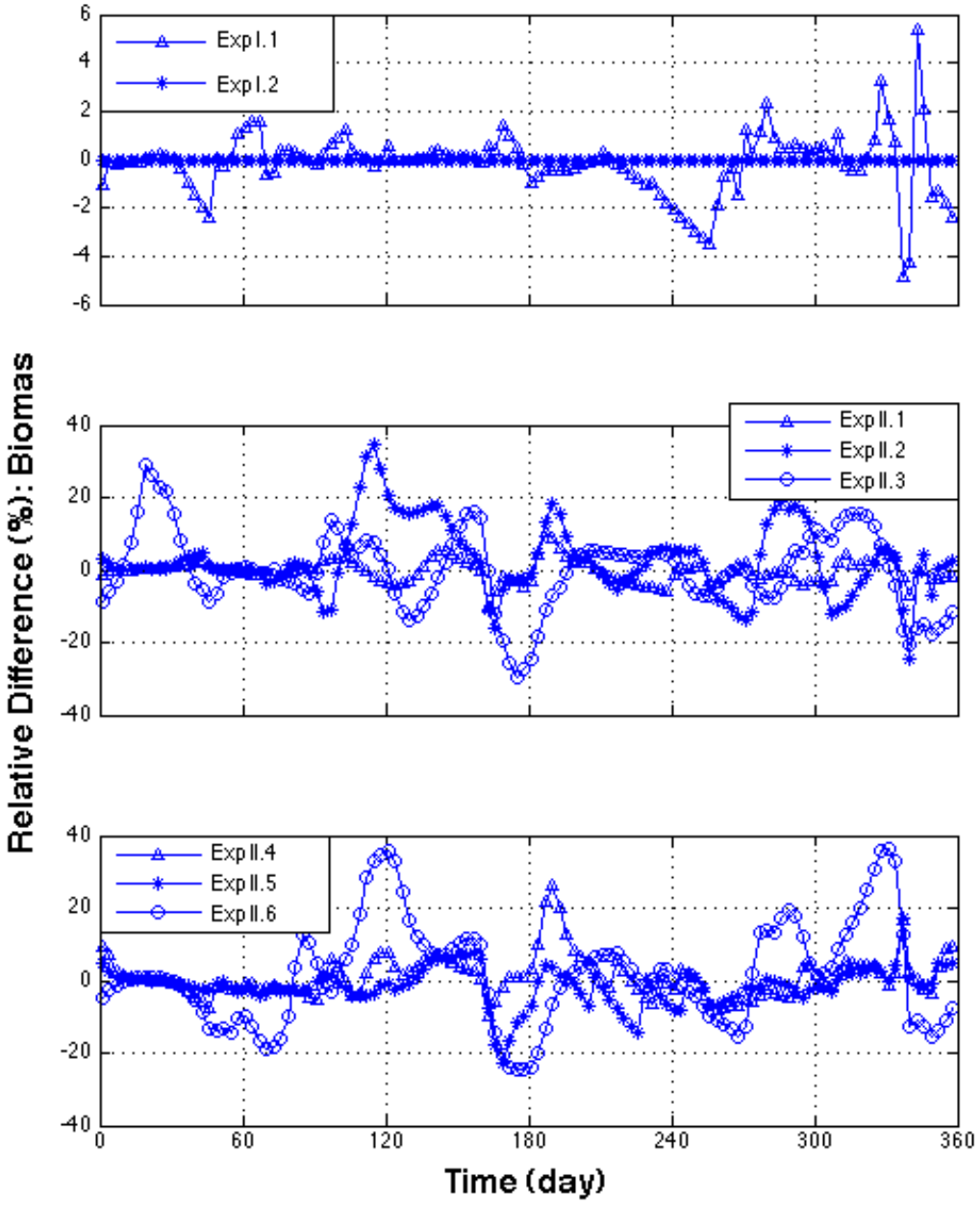

Figure A2: Sensitivity analysis of temporal variations of genetic combinations from dominant species (Note: dominant species on each day is represented by its combination of six genes on the y-axis according to the order of *amt, nr, nif, pst, pho, phn* from the bottom to top of y-axis; white bar stands for the presence of a particular gene.)