MICB425\_portfolio

Jawairia Atif, 39139134

version April 20, 2018

Table of Contents

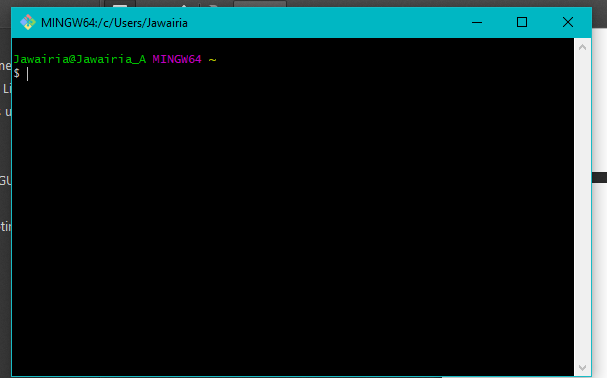
# Module 01

## Module 01 portfolio check

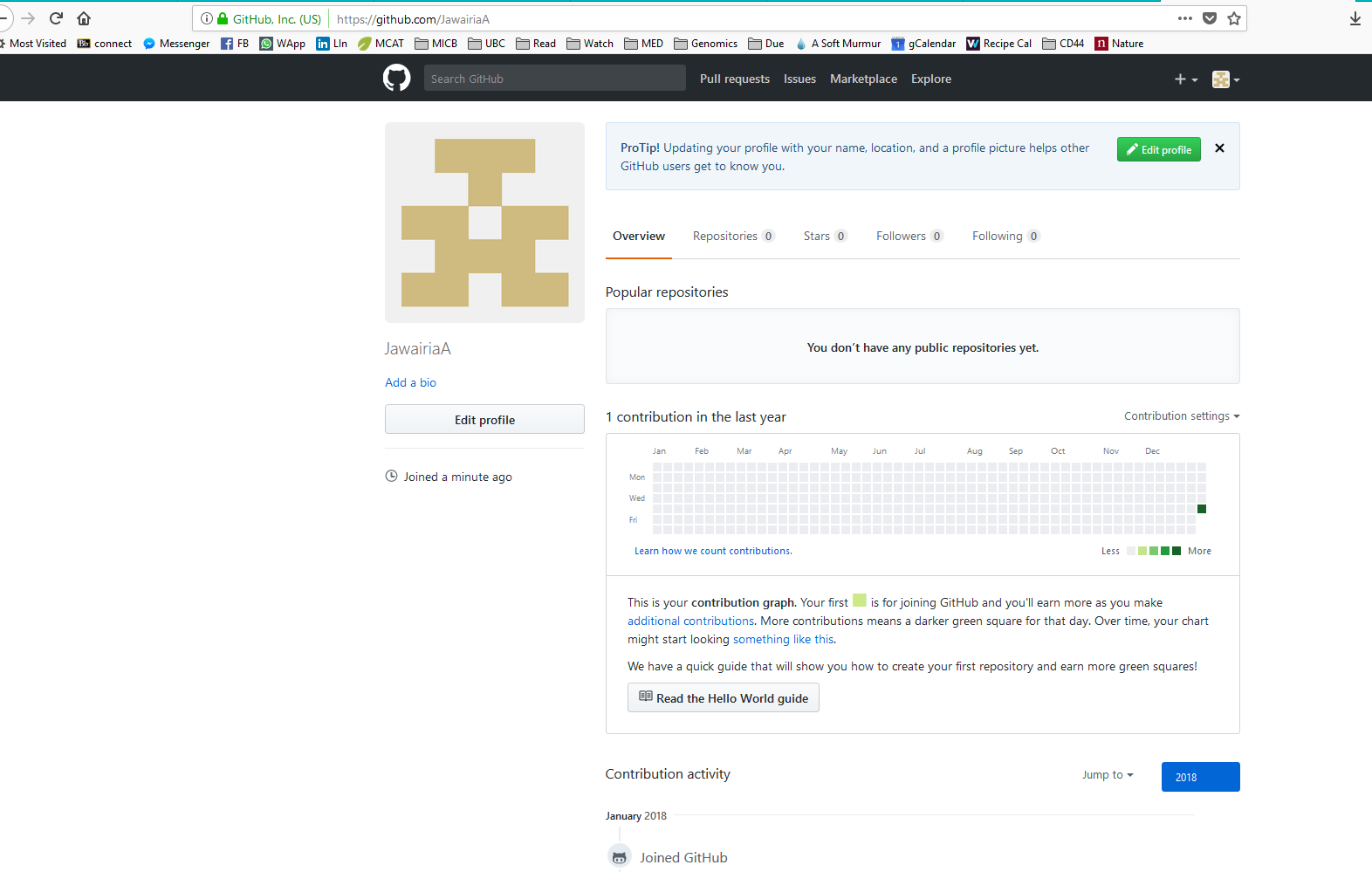
* Installation check
  + Completion status:
  + Comments:
* Portfolio repo setup
  + Completion status:
  + Comments:
* RMarkdown Pretty PDF Challenge
  + Completion status:
  + Comments:
* Evidence worksheet\_01
  + Completion status:
  + Comments:
* Evidence worksheet\_02
  + Completion status:
  + Comments:
* Evidence worksheet\_03
  + Completion status:
  + Comments:
* Problem Set\_01
  + Completion status:
  + Comments:
* Problem Set\_02
  + Completion status:
  + Comments:
* Writing assessment\_01
  + Completion status:
  + Comments:
* Additional Readings
  + Completion status:
  + Comments

## Data science Friday

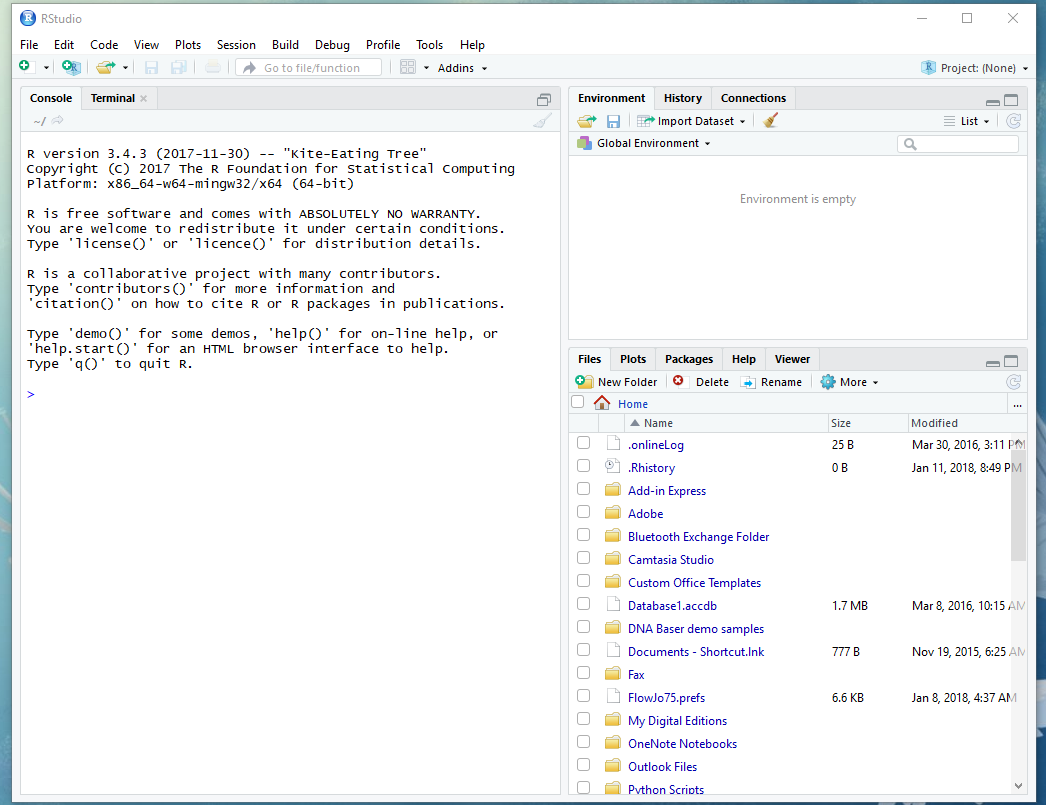
### Installation check



Screenshot of GIT console



Screenshot of GitHub home page



Screenshot of Rstudio

### Portfolio repo setup

**Creating the direcrory** cd ~/documents  
mkdir MICB425\_portfolio  
touch ID.txt

**Initializing repo on Github** git init git add .  
git commit -m“first commit”  
git remote add origin <https://remote_repository_URL> git remote -v  
git push -u origin master

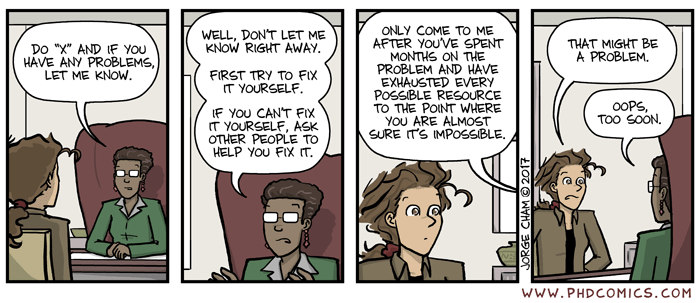
**Pushing to Github**

git add . git commit -m “comment” git status git push

### RMarkdown pretty PDF challenge

# R Markdown PDF Challenge

The following assignment is an exercise for the reproduction of this .html document using the RStudio and RMarkdown tools we’ve shown you in class. Hopefully by the end of this, you won’t feel at all the way this poor PhD student does. We’re here to help, and when it comes to R, the internet is a really valuable resource. This open-source program has all kinds of tutorials online.



<http://phdcomics.com/> Comic posted 1-17-2018

## Challenge Goals

The goal of this R Markdown html challenge is to give you an opportunity to play with a bunch of different RMarkdown formatting. Consider it a chance to flex your RMarkdown muscles. Your goal is to write your own RMarkdown that rebuilds this html document as close to the original as possible. So, yes, this means you get to copy my irreverant tone exactly in your own Markdowns. It’s a little window into my psyche. Enjoy =)

**hint: go to the** [**PhD Comics website**](http://phdcomics.com/) **to see if you can find the image above**  
*If you can’t find that exact image, just find a comparable image from the PhD Comics website and include it in your markdown*

### Here’s a header!

Let’s be honest, this header is a little arbitrary. But show me that you can reproduce headers with different levels please. This is a level 3 header, for your reference (you can most easily tell this from the table of contents)

#### Another header, now with maths

Perhaps you’re already really confused by the whole markdown thing. Maybe you’re so confused that you’ve forgotton how to add. Never fear! ~~A calculator~~ R is here:

1231521+12341556280987

## [1] 1.234156e+13

### Table Time

Or maybe, after you’ve added those numbers, you feel like it’s about time for a table!  
I’m going to leave all the guts of the coding here so you can see how libraries (R packages) are loaded into R (more on that later). It’s not terribly pretty, but it hints at how R works and how you will use it in the future. The summary function used below is a nice data exploration function that you may use in the future.

library(knitr)

## Warning: package 'knitr' was built under R version 3.4.4

kable(summary(cars),caption="I made this table with kable in the knitr package library")

I made this table with kable in the knitr package library

|  |  |  |
| --- | --- | --- |
|  | speed | dist |
|  | Min. : 4.0 | Min. : 2.00 |
|  | 1st Qu.:12.0 | 1st Qu.: 26.00 |
|  | Median :15.0 | Median : 36.00 |
|  | Mean :15.4 | Mean : 42.98 |
|  | 3rd Qu.:19.0 | 3rd Qu.: 56.00 |
|  | Max. :25.0 | Max. :120.00 |

And now you’ve almost finished your first RMarkdown! Feeling excited? We are! In fact, we’re so excited that maybe we need a big finale eh? Here’s ours! Include a fun gif of your choice!



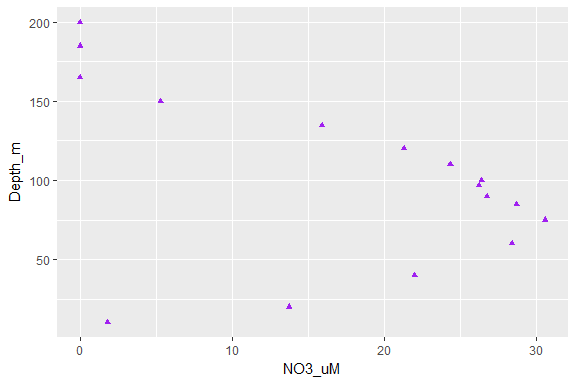
meow

### Data science assignment 4

# install.packages("tidyverse")  
library(tidyverse)  
  
metadata <- read.table(file="c:/Users/Jawairia/Documents/MICB425\_portfolio/Saanich.metadata.txt", header=TRUE, row.names=1, sep="\t", na.strings="NAN")  
otu= read.table(file="c:/Users/Jawairia/Documents/MICB425\_portfolio/saanich.otu.txt", header=TRUE, row.names=1, sep="\t", na.strings="NAN")  
# source("https://bioconductor.org/biocLite.R")  
# biocLite("phyloseq")  
  
library(phyloseq)  
  
load("phyloseq\_object.RData")

#### Exercise 1

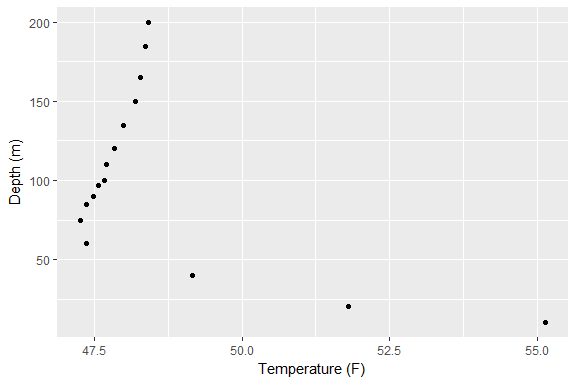
#Figure 1: Plotting NO3\_uM against depth  
ggplot(metadata, aes(x=NO3\_uM, y=Depth\_m)) +   
 geom\_point(shape=17, color="purple")



#### Exercise 2

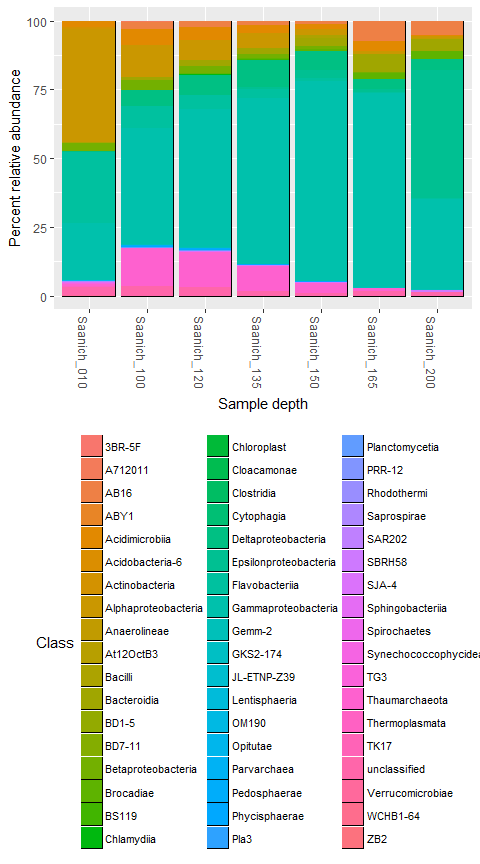
#Figure 2:   
metadata %>%   
 mutate(Temperature\_F = Temperature\_C\*1.8 +32) %>%   
 ggplot() + geom\_point(aes(x=Temperature\_F, y=Depth\_m))+  
 labs(x= "Temperature (F)", y="Depth (m)")

## Warning: package 'bindrcpp' was built under R version 3.4.4



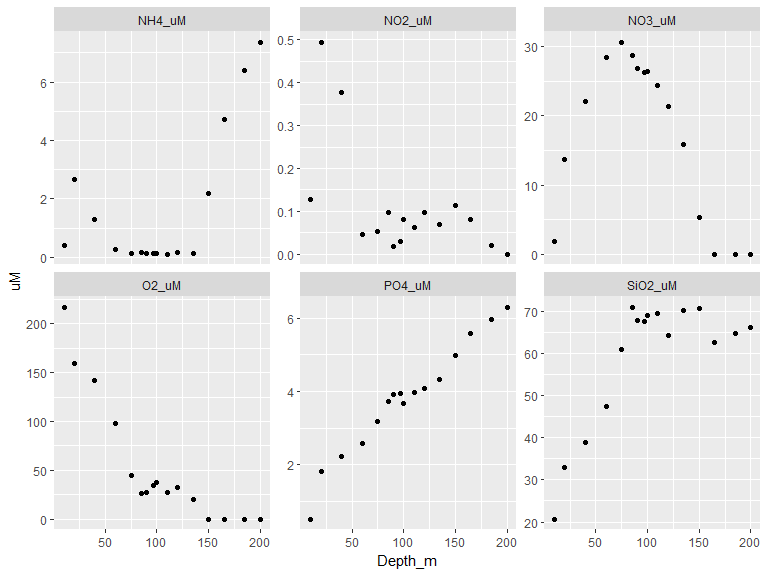
#### Exercise 3

#Figure 3:  
physeq\_percent = transform\_sample\_counts(physeq, function(x) 100 \* x/sum(x))   
 plot\_bar(physeq\_percent, fill="Class") +   
 geom\_bar(aes(fill=Class), stat="identity") + labs(x="Sample depth", y="Percent relative abundance")+  
 theme(legend.position = "bottom" , legend.text=element\_text(size=8), plot.title = element\_text(size=12, face="bold", hjust=0.5), axis.text= element\_text(size=9))+  
 guides(fill=guide\_legend(ncol=3,bycol=TRUE))



#### Exercise 4

#Figure 4:   
library(tidyverse)  
metadata %>%   
 select(Depth\_m, O2\_uM, PO4\_uM, SiO2\_uM, NO3\_uM, NH4\_uM, NO2\_uM) %>%  
 gather("Nutrients","uM", O2\_uM:NO2\_uM)%>%  
 ggplot() + geom\_point(aes(x=Depth\_m, y=uM)) +  
 facet\_wrap(~Nutrients, scales="free\_y") +  
 theme(legend.position="none")



## Origins and Earth Systems

### Evidence worksheet 01

Paper: [Whitman *et al* 1998](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC33863/)

#### Learning objectives

Describe the numerical abundance of microbial life in relation to ecology and biogeochemistry of Earth systems.

#### General questions

#### What were the main questions being asked?

The main questions were:

* To determine the number of prokaryotes in different habitats and how much they contribute to the total number of microbes on Earth.
* What are the turnover rates of the microbes in these environments and which habitats are the most productive?
* What is the total amount of cellular carbon and other nutrients (N, P) produced by these prokaryotes on earth?
* Estimate prokaryotic diversity in different habitats.

#### What were the primary methodological approaches used?

* Pooling data from studies in which there is sampling of prokaryotes from different habitats, and quantification of cells in these samples and extrapolation, estimations, assumptions, mathematical formulas to calculate cell numbers, nutrient contents etc.:
* Soil:

estimating prokaryotic abundance in differant habitats

+ ocean below 200m  
 -average cellular density and multiplied it with the estimated amount of marine and fresh water quantities  
   
+ different soils  
 - estimations cellular density from direct cell counts from different soils in past field studies.  
 - calculate number of microbes by looking at perviously estimated amounts of soil on earth  
   
+ terrestrial subsurface in various dephts  
 - estimated prokaryotic numbers from ground water based on values of several sites, and multiplied this using the estimated value of ground water on Earth.   
 - looked at the average porosity of Earth's soil and used known values of space occupied by prokaryotes in these pores  
   
+ Animals and other habitats  
 -human: skin surface area multiplied by cell density of prokaryotes on the skin   
 - insects: using termites as a model: multiple number of microbes in insect by total number of insect on the planet.  
 - leaves: leaf area x assumed dense population of microbes   
 - air: pre-calculated data used

#### Summarize the main results or findings.

3 habitats contribute to prokaryotic abundance on Earth. + Open ocean: 1.2x 1029 cells + Soil: 2.6x 1029 cells + Subsurfaces: terrestrial and marine: 0.25-2.5x 1030 cells

Other important habitas: + Animals, leaves, air, groundwater Further important habitats but with minor contributions to total cell number:

Total number of prokaryotes on earth: + ~4-6x 1030 cells

Total amount of Carbon in prokaryotes: 350-550 Pg which is roughly 60-100% of the total Carbon of plants

Total prokaryotic nutrients (N,P)= 10x nutrient content of plants. (N: 85-130 Pg, P: 9-14 Pg)

Turnover rates for the different habitats:  
+ sOil: 2.5 years + Ocean above 200m: 6-25 days  
+ Ocean below 200m: 300 days  
+ Terrestrial subsurface: 1-2x 103 years

cellular productivity: # of cells produced per a unit of time: + Ocean above 200m: highest productivity: 8.2\*1029 cells/year + Total Earth: 1.7x 1030 cells/year

The higher the cellular productivity: the more mutation events occur: leads to more diversity. This leads to the emergence of new cycles and pathways.

Authors mention that the number of prokaryotic species may be underestimated due to inadequate sampling methods.

#### Do new questions arise from the results?

* Subsurface prokaryote turnover rate: very high: is this an over-estimation: need to investigate this habitat further.
* Since the 1980s tools to study microbial diversity have evolved: do they give us better estimates for the above measurements?
* In the lsat 40 years, has the microbial abundance and diversity changed?
* How does the abundance of the prokaryotes in these environments play a role in the total metabolic capcity of these microbes in affecting the global nutrient cycles?

#### Were there any specific challenges or advantages in understanding the paper (*e.g.* did the authors provide sufficient background information to understand experimental logic, were methods explained adequately, were any specific assumptions made, were conclusions justified based on the evidence, were the figures or tables useful and easy to understand)?

It was hard to understand where some of the numbers, calculations and assumptions came from because it was not very well explained. A discussion on the precision of the data presented in this paper would have been cool.

### Problem set 01

Paper: [Whitman *et al* 1998](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC33863/)

#### Learning objectives:

Describe the numerical abundance of microbial life in relation to the ecology and biogeochemistry of Earth systems.

#### Specific questions:

#### What are the primary prokaryotic habitats on Earth and how do they vary with respect to their capacity to support life? Provide a breakdown of total cell abundance for each primary habitat from the tables provided in the text.

Open ocean: Total 1.2x 1029 cells

* Top 200m: 3.6x 1028
* Below 200m (+10 cm of marine sediment): 8.2x 1028

Soil: 2.5x 1029 cells

Subsurfaces (oceanic + terrestrial): ~3.8x 1030 cells

#### What is the estimated prokaryotic cell abundance in the upper 200 m of the ocean and what fraction of this biomass is represented by marine cyanobacterium including Prochlorococcus? What is the significance of this ratio with respect to carbon cycling in the ocean and the atmospheric composition of the Earth?

Upper 200m of the ocean contains a total 3.6 x 10^28 cells:

->2.9x 1027 autotrophs (e.g, cyanobacteria like Prochlorococcus): 8% of total: responsible for carbon assimilation to sustain C requirements of heterotrophic (comprising 92% of total cells) cells. This menans that the carbon in the ocean is cycled within the ocean at higher rates than the carbon from the atmoshpere is fixed. This essentially results in a loss of carbon from the ocean into the atmosphere.

#### What is the difference between an autotroph, heterotroph, and a lithotroph based on information provided in the text?

* Autotroph: Fix inorganic carbon (CO\_2) -> biomass.
* Heterotroph: use organic carbon to build biomass
* Lithotroph: inorganic substances as electron donor like NH3

#### Based on information provided in the text and your knowledge of geography what is the deepest habitat capable of supporting prokaryotic life? What is the primary limiting factor at this depth?

4km below the surface (4 km below terrestrial surface or marine sediments) is the deepest that life has been measured. The main limiting factor for life at this depth is temperature: \* temperature is about 125 deg celcius: limit for prokaryotic life. \* temperature drops at 22 degrees per km of sruface: Marian’s trench, at 10.9 km deep, is the deepest part of the ocean that can support like: another 4 km below the marine sediments of Marian’s trench would also contain life.

#### Based on information provided in the text your knowledge of geography what is the highest habitat capable of supporting prokaryotic life? What is the primary limiting factor at this height?

Spores can survivee at altitudes up to 77km above the surface of the earth. But metabollically active cells can only survive at roughly 20 km above the surface of the earth (so highest would be: Mt. everest: 8.8km + 20 km on top). Limiting factor: cold temperatures, ionizing radiationm lack of nutrients and no pressure.

#### Based on estimates of prokaryotic habitat limitation, what is the vertical distance of the Earth’s biosphere measured in km?

Mariana trench: ~ -10.9 km deep, + - 4 km further below marine sediment Mt. everest: 8.8km + 20km on top vertical distance of Earth’s biosphere: 22 + 8.8 +10.95 +4= 45.7km

#### How was annual cellular production of prokaryotes described in Table 7 column four determined? (Provide an example of the calculation)

Annual cellular production \* account for population size and growth rate \* Annual cellular production = Population size (number of cells) x Turnover rate ( in years) \* Using Subsurface ocean as an example: + # cells = 3.6x 1028 cells  
+ Turnover time: 16 days + Turnover rate = 365 (days/year)/16 ( turnover/days)= 22.81 turnovers per year + annual cellular production: 3.6x 1028 Cells x 22.81 turnovers per year = 8.2x 1029 cells per year \* Soil:  
+ # cell = 2.6x 1029 cells + turnover time: 900 days + Turnover rate = 0.4 turnovers per year (= 365/900) + annual cellular production:2.6x 1029 x 0.4= 1029 cells per year

#### What is the relationship between carbon content, carbon assimilation efficiency and turnover rates in the upper 200m of the ocean? Why does this vary with depth in the ocean and between terrestrial and marine habitats?

* the net productivity in the ocean is: 51 Pg C/year
* Total amount of marine C in ocean: 0.7-2.9 Pg
* calculating the upper limit (2.9 Pg C)
  + Carbon in one cell: 5-20 fg C/cell
  + 20 fg C/cell=20x0-30 petagreams C
  + 3.6\*1028 cells x 20x 10-30 pg/cell= 0.72 Pg of C in marine prokaryotes
  + Carbon efficieny while 20% (a factor 5 needed), Authors used 4 in the paper for some reason:
  + 4 x 0.72 Pg = 2.88Pg/year
* 85% of net productivity in the ocean is consumed in the upper 200 m: 51 Pg C/year X 0.85 = ~43 Pg C/year
  + Turnovers per year in upper 200 m of ocean: 43 pg/yr / 2.88 pg/yr (turnover rate/total amounf of carbon in the ocean)= 14.9 Turnovers per year
  + So: (14.9 turnovers/year) x (1 year/365 days) = one turnover every 24.5 days. or using the lower limit of C in the ocean: 61 turnovers per eyara ax.

The relationship between carbon assimilation efficiency adn turnover rate with depth varies in different environments varies because the primary producer fraction of prokaryotes reach different depths in different habitats.

#### How were the frequency numbers for four simultaneous mutations in shared genes determined for marine heterotrophs and marine autotrophs given an average mutation rate of 4 x 10-7 per DNA replication? (Provide an example of the calculation with units. Hint: cell and generation cancel out)

* average mutation rate per DNA replication: 4 x 10-7 (mutations/gerneration)
* for 4 simultaneous mutations: (4 x 10-7)4 = 2.56 x 10-26 mutations per gene per generation  
  \*16 turnovers/day: 22.5turnovers/year
* 3.6 x 1028cells (total number of marine cells) x 22.8 turnovers per year = 8.2 x 1029 cells produced per year
* (8.2 x 1029 cells per year) x (2.56 x 10-26 mutations per generation) = 2.1 x 104 times per year 4 simulataneous mutations happen
  + 1/[(2.1 x 104 events/ year) \* (1 year/365 days) \* (1 day /24 hours)]= 2.5 events/hour: or in other words
  + four simoultanous mutations happen every 0.4 hours every 24 minutes.

#### Given the large population size and high mutation rate of prokaryotic cells, what are the implications with respect to genetic diversity and adaptive potential? Are point mutations the only way in which microbial genomes diversify and adapt?

Due to high mutation rates and a large population size, the adaptive potential might be much higher than previously thought. The number of prokaryotic species might be much higher than previously estimated. The high diversity implies that prokaryotic cells have the potential to evolve and adapt relatively quick. High mutation rates might allow cells to adapt quickly to changing environments and new metabolic cycles can emerge that might have a great impact on the global nutrient cycle.

Beyond point mutations, microbes can diversify with bigger earrangements in the genome like inversions and deletions. Furthermore, genes can be adopted into the cells from DNA rom neighbouring cells or the environment in a process known as horizontal gene transfer.

#### What relationships can be inferred between prokaryotic abundance, diversity, and metabolic potential based on the information provided in the text?

prokaryotic abundance correlates with magnitude of meetabolic potential.

prokaryotic diversity correlates with diversity of metabolic characteristics.

In conclusion: microbes have a large potential to participate in and alter the earth’s biogeochemical cycles.

/pagebreak

### Evidence worksheet 02: Life and the Evolution of Earth’s Atmosphere

Paper 1: [Nisbet et al. 2001](https://www.nature.com/articles/35059210) Paper 2:

[Kasting et al. 2002](http://science.sciencemag.org/content/296/5570/1066)

#### Learning objectives:

Comment on the emergence of microbial life and the evolution of Earth systems

#### Indicate the key events in the evolution of Earth systems at each approximate moment in the time series.

**Hadean:**

* 4.6 Ga:
  + Formation of solar system:
  + Earth: Atmosphere comprises of: water vapor, methane, ammonia, hydrogen, nitrogen.
* 4.5 Ga:
  + Moon formed: gave Earth its spin, tilt, and day/ night cyles and seasons.
* 4.5 Ga - 4.1 Ga:
  + High levels of CO\_2: increased temperature despite the sun being weak
* 4.4 Ga:
  + Oldest zircoin forms
  + Oceans, and atmosphere forms
* 4.4- 4.1 Ga:
  + meteorite impacts
* 4.3-3.8 Ga: Heavy bombardment of Earth
* 4.1 Ga:
  + First evidence of life in zirocns and from preserved Carbon isotopes
* 4.0 Ga: Plate subduction
  + Oldest rock: acasta gneiss in Canada and evidence of plate subduction.
  + CO\_2 converted to carbonate: and lcts less as greenhouse gas: temperatures become glacial.

**Archaean:**

* 3.8 Ga:
  + Oldest sedimentary rocks show evidence of life and methanogenesis from carbon isotopes.
* 3.5 Ga:
  + Evidence for photosynthesis: in fossils, and stromatolites.
  + Rubisco: enzyme for carbon assimilation detected: cyanobacteria photosynthesize: oxygen being produced.
* 3 Ga:
  + first Glaciation event: Oxygen production leads to less methanogen production.- less greenhouse effect
* 2.7 Ga:
  + great oxydation event-> glaciation
  + Fossils contain molecules that suggest presence of cyanobacteria (hydrocarbon biomarkers) but also eukaryotes (steranes)
* 2.2 Ga:
  + red rock beds obserevd: evidence of oxidation: oxygens level increased sharply.

**Proterozoic:**

* 1.8 Ga:
  + Evidence of eukaryotes.
  + Changing carbon cycle
* 1 Ga:
  + First major ice age has a large impact on carbon cycle

**Phanerozoic:**

* 540 Ma:
  + Cambrian explosion: expansion of multicellular life
  + Emergence of animals.
* 400 Ma:
  + Land plants observed
  + Strong increased oxygenation of atmosphere
  + Denovian extinction:
* 250 Ma:
  + Permian extinction (95% of species disappear)
  + gigantism of organisms, rise of dinosaurs
  + Pangea forms, dry, harsh climate
* 66 Ma:
  + Cretaceous-palogene extinction event
  + Dramatic global warming
  + Dominating forest
* 200,000 years ago:
  + *Homo sapiens* appears

#### Dominant physical/ chemical characteristics of Earth systems

**Hadean** + High CO\_2 levels (in the atmosphere) keep the earth warm: 500 deg Celcius + Loss of hydrogen to space + weak sun + Earth mostly molten rock + Rock vapour atmosphere + heavy meteror bombardment + Evidence of plate subduction near the end of the period

**Archean**  
+ Atmosphere failled with methane: keeps the earth warm: methanogenesis + Earth still anoxic at the beginning. + Earth starts cooling + Rock formation + As photosynthesis evolved, some O2 was present

**Proterozoic** + Oxygen levels in atmosphere rising + oxygen reacts with atmospheric methan to produce CO\_2 + atmospheric oxygen level rise resulting in glaciation events. + oxic air

**Phanerozoic**

* Formation of pangea and subsequent breaking apart of pangea
* Ice agaes
* Modern atmosphere with increased oxygen levels.
* Emergence of complex global nutrient cycles

### Problem set 2: Microbial Engines

[Falkowski et al. 2008](https://www.ncbi.nlm.nih.gov/pubmed/18497287)

#### What are the primary geophysical and biogeochemical processes that create and sustain conditions for life on Earth? How do abiotic versus biotic processes vary with respect to mattern and energy transformation and how are they interconnected?

Geophysical processes: Plate tectonics and atmospheric photochemical processes. These two processes allow for the interaction of molecules and their cyclation.

Abiotic geochemical reactions are based on acid/base chemistry; proton transfers

Biogeochemical processes: Redox reactions driven mostly by microbes. -> Formation of two half-cells leads to linked cycles.

Biotic processes are based on redox reactions; electron transfers

Biogeochemical processes depend on resupply of C, S and P by tectonic cycles in geological time-scales. Abiotic cycles can supply biogeochemical reactions with new molecules carrying electrons which are required for the redox reactions. Therfore, abiotic acd/base reactions are interconnected with biotic redox reactions. Both reaction types support the other type with metabolites and energy required to sustain their cycles. This connection also leads to feedback on the microbial evolution, changing their metabolism and eventually the global redox state.

#### Why is Earth’s redox state considered an emergent property?

The Earth’s redox state depends on microbial metabolism which in turn adapts to the properties of geochemical cycles and therefore is subject to constant change. Further, a large part of the Earth’s redox state is determined by photosynthesis, which is a process independent of already available energy stored in metabolites. Thus, many different processes and cycles are nested in a complex manner and together lead to the Earth’s redox state as an emergent property.

#### How do reversible electron transfer reactions give rise to element and nutrient cycles at

different ecological scales? What strategies do microbes use to overcome thermodynamic barriers to reversible electron flow?

Reversible metabolic pathways can be directly related and catalyzed by microbes from similar species. On another scale, the reversible pathways may be catalyzed in a more global manner, by very diverse microbes. An expample for the first case is the formation of methane by methanogenic archaea, when the hydrogen pressure is high enough. In low hidrogen tension however, the reaction gets inversed by oxidation of methane to CO2. This reverse pathway is catalyzed by Archeae closely related to the methanogens. The second model can be represented by the global nitrogen cycle as an example. In this cycle, redox reactions are spatially and temporally separated and catalyzed by many different microbes. Atmospheric nitrogen is fixed by transformation of N2 to NH4+ by the oxygen sensitive enzyme nitrogenase. NH4+ is then oxidized in several steps to nitrite and finally nitrate in presence of oxygen. This nitrification is performed by several only distantly related bacteria and archaea. Again another set of microbes then uses the nitrate and nitrite as electron acceptors in anoxic conditions to generate energy. These bacteria thereby close this diverse, multipsecies cycle by the formation of N2.

Thermodynamic barriers to reversible electron flow can be overcome by coupling the unfavourable reaction to an energy yielding reaction such as catabolism of organic compounds. Else, reactions can be made thermodynamically favourable by changing the redox couples in a manner that again leads to a positive redox potential of the wanted reaction. Different microbes use the end products of other microbes as their substrate. In photosynthesis for example, CO2 is used as an electron acceptor to generate reduced organic carbon. In this reaction, H2O, which is used as electron donor, gets oxidized to O2 as end product. Different organisms can then ireverse these reactions by using the organic carbon as electron donor and O2 as electron acceptor.

#### Using information provided in the text, describe how the nitrogen cycle partitions between

different redox “niches” and microbial groups. Is there a relationship between the nitrogen cycle and climate change?

In this cycle, redox reactions are spatially and temporally separated and catalyzed by many different microbes. Atmospheric nitrogen is fixed by transformation of N2 to NH4+ by the oxygen sensitive enzyme nitrogenase in bacteria like rhizobia. NH4+ is then oxidized in several steps to nitrite and finally nitrate in presence of oxygen. This nitrification is performed by several only distantly related nitrifying bacteria and archaea. Again another set of microbes then uses the nitrate and nitrite as electron acceptors in anoxic conditions to generate energy. These bacteria thereby close this diverse, multipsecies cycle by the formation of N2.

Humans strongly affect the global nitrogen cycle. Extensive usage of synthetic nitrogen fertilizers and fossil fuel processing lead to strong increases of reactive nitrogen in the atmosphere. These nitrogen species affect the abundance of the greenhouse gases CO2, CH4, O3 and N2O and therefore contribute to global warming.

#### What is the relationship between microbial diversity and metabolic diversity and how does this

relate to the discovery of new protein families from microbial community genomes?

The huge amount of microbial cells and their fast turnover leads to an enourmous genetic diversity through mutations. A further mechanism strongly promoting microbial diversity and evolution is horziontal gene flow. In different envrionments, different selective pressures exist, which require specialized metabolic pathways. Therefore, the high genetic diversity and many different strong selective pressures lead to a huge metabolic diversity of microbes. This huge diversity can be observed by the fact that to date, the number of new protein families discovered still rises linearly with the number of newly sequenced genomes. This means, the microbial and metabolic diversity are so high, that the total number of genes and protein families are still unknown and can only be estimated very imprecisely

#### On what basis do the authors consider microbes the guardians of metabolism?

The core genes responsible for most of the environment-specific metabolic pathways are spread in microbes all over the world. In addition, essential genes for general houskeeping pathways are highly conserved among global microbes.

Different environments favour the evolution and survival of the best adapted microbes. This means, less adapted microbes and their specialized pathways go extinct in a certain environment. However, thanks to the global distribution of the core gene set, the specialized metabolic pathways of this extinct strain are very likely to survive in another strain in a different environment. In addition, the survival of the essential genes is even more expected, as these genes are conserved throughout most of the microbes on earth. Thus, despite changing environments and the selective pressures they cause, the genes responsible for different metabolic pathways always survive in some micrboes which therefore can be considered as the guradians of metabolism.

### Evidence Worksheet 03

[Waters et al. 2016](http://science.sciencemag.org/content/351/6269/aad2622)

#### Evaluate human impacts on the ecology and biogeochemistry of Earth systems

#### What were the main questions being asked?

* Whether humans changed the Earth system strong enough that the stratighrapic signature is altered in a way that the current epoch can be considered as distinct from the Holocene.
* To determine the time-point this human made stratigraphical signal became recognizable in a significant manner.

In general: To rieview anthropogenic markers of changes in different systems ( biochemical cycles, sediment composition, sea-level, climate, biotic systems)

#### What were the primary methodological approaches used?

The paper is a review. -> Results collected from other studies. Measured concentrations of different molecules, isotope frequencies, temperatures etc from different places (soil, ocean, glacier etc).  
-> Other measurements also taken to reconstruct/ extrapolate/ estimate the corresponding values of past times in Earth’s history.  
-> Comparison of values from present and past times to infer whether humans have caused a significant change in stratighrapic signature.

#### Summarize the main results or findings.

Anthropogenic deposits: (great acceleration at ~1950 CE)

* Products of mining, waste desposal, construction, urbanization.
  + Great expansion of new minerals: new geological materials with long term persistance (new “rocks”)
    - Aluminium, concrete, organic polymers: “technofossils”, provide stratigraphic resolution in time scales of years to decades.
  + Combustion of fossil fuels lead to global distribution of airborn particles: black carbon, inorganic ash spheres, sperical carbonaceous particles. These particles are long time persistent stratigraphic markers.

Modification of sedimentary processes:

* Transformation of >50% of Earth’s land surface: landfills, urban structure, mine tailings, deforestation, cultivated soils, sediment retention (dams, leading to reduced flux, subsided deltas)
* Ocean: coastal reclamation works, sediment reworking, sand extraction, rising sea level, eutrophication, coral bleaching
* Subsurface: mineral extraction, waste storage

Geochemical signatures in sediments and ice:

* Elevated concentrations of polyaromatic hydrocarbons, polychlorinated biphenyls, pesticide residues, lead.
* Fertilizer usage: Doubled concentrations of nitrogen and phosphorus in soil. Influx to lakes led to oxygen deficiency, increased animal mortality.
* Decrease of 15N in lakes, ice sheet.
* Increase in nitrate: Values higher than any for the previous 100’000 years.-> distinct from Holocene background level.
* Industrial metals: cadmium, chromium, copper, mercury, nickel, lead, zinc.

Radiogenic signatures:

* Fallout from nuclear weapons testing: Most widespread, globally synchronous anthropogenic signal.  
  -> Start of Anthropocene may be defined by detonation of the trinity atomic device at alamogordo, 1945.
* Increased 14C, 239P ( 239P may be best radioisotope for marking the start of the Anthropocene because of long half-life and low solubility)

Carbon cycle:

* Atmospheric CO2: >400 ppm, exceeding Holocene levels since 1850 CE.
* 13C levels decrease of > 0.2% because burning of fossil fuels leads to increase of 12C (organic carbon has increased 12C because lighter isotope reacts faster in biochemical reactions (photosynthesis)
  + Permanent signal, stored in tree rings, lime-stones, fossils.
* Increase of methane to 1700 ppb, 900ppb higher than highest value in past 800’000 years.

Climate and sea-level change

* Given orbital trend, earth should be cooling ( as it was since 8200 B.P. untill 1800CE)
* Emission of greenhouse gases lead to climate warming.
* Average temperature increase of 06.-0.9 degrees from 1906 to 2005, exceeding natural variability.
* Average global sea levels are higher than highest levels of the past 115’000 years.
* Rise of 3.2 mm per year from 1993 to 2010
* Climate and sea-level changes are not as strong as other stratiraphic changes, but are likely to exceed the envelope of quaternary system baseline conditions.
* Change in planetary energy balance: radiative forcing increased by 2.29 Wm-2 compared to 1750 CE. 8because of burning of fossil fuels)

Biotic change:

* Extinction rates since 1500 CE are far above mean per-million-year background rates.
* species abundances and assamblages strongly altered; transglobal species invasions, agriculture, fishing.

Conclusion:

* stratigraphic signatures are either novel or outside the range of variation of the holocene, supporting the formalization of the Anthropocene as stratigraphic epoch.
* Dating of begining of Anthropocene proposed to lay between 1945 and 1964 CE

#### Do new questions arise from the results?

* Should the Anthropocene be formalized as epoch or left as informal time term?
* How to define the Anthropocene?  
  -> By GSSA (calendar age) or GSSP ( reference point in a stratal section)  
  -> Define start point of the Anthorpocene (1945-1964)?
* How are the stratigraphical changes going to proceed in the future? -> Make projections of climate, sea-level, biodiversity etc to future

#### Were there any specific challenges or advantages in understanding the paper (e.g. did the authors provide sufficient background information to understand experimental logic, were methods explained adequately, were any specific assumptions made, were conclusions justified based on the evidence, were the figures or tables useful and easy to understand)?

The Review mostly showed very well, on what data/assumptions/estimations their results are based. However, sometimes it was not clear, how the data were optained, i.e. which methods lead to the obtained results.  
Further, it is not always clear, how exact the obtained values for concentrations/ temperatures etc are. Especially for the estimations of the values for earlier points in time, some more information about error rates would have been useful.

### Module 1 essay

“Microbial life can easily live without us; we, however, cannot survive without the global catalysis and environmental transformations it provides.”

Ever since the emergence of humans, we have been interacting with microbes. We live in symbiosis with microbes in many different aspects, which rises the important question whether we could survive without them. Microbial life emerged about 4 billion years (1) before the first humans developed, which prooves that microbes can easily live without us. In contrast, confirming that humans cannot survive without the global catalysis provided by microbes, is much more challenging. Several aspects need to be considered in order to answer the question whether humans rely on the presence of microbes. Firstly, after their emergence, the microbes rapidly spread all over the earth and rose to incredible numbers of individuals. Secondly, fast reproduction and constant turnover of the cells leads to massive requirements of nutrients and thereby promotes the global turnover of these nutrients. The fast reproduction of microbes further generates an enormous genetic diversity through mutations, which allows the microbes to constantly adapt present pathways or even generate new metabolic processes. Through this extremely high abundance and diversity, microbes developed a massive potential not only to actively participate in, but also to significantly alter important biogeochemical cycles, which can be shown by several examples covered in this assay. Lastly, this potential for global catalysis leads to important, microbially driven environmental transformations on Earth, creating an atmosphere inhabitable by humans. This not only allowed for the emergence of more complex organisms and the human species, but will also be important for human survival in the future.

Microbes have a massive potential to impact global cycles because of their sheer number of cells present. They account for by far the biggest part of the number of organisms alive on earth and arguably also the greatest part of nutrients stored in living beings. The number of prokaryotes living on earth could be obtained by cell counts of probes sampled from several different habitats. Subsequent projection and extrapolation for habitats not available for sampling led to an estimation of 4-6\*1030 cells present (2). Considering average values of nutrient contents per cell, leads to the conclusion that microbes totally contain 60-100% of the amount of carbon stored in plants. The fraction of nitrogen and phosphorus stored in microbes is even higher, with about ten times more of these elements stored in microbes than in plants (2). These numbers show, that through their enormous abundance and capacity, microbes technically have the potential to provide important global catalysis.

In addition to the huge number of microbial cells, their fast turnover rates generate an even higher potential of the microbes to significantly catalyze and transform global cycles. An average turnover rate of 22 turnovers per year leads to about 8.2\*1029 heterotrophic cells produced every year, just in the upper 200 meters of the ocean (2). Therefore, the vast amounts of cells continuously being produced further explains their impact on global cycles as their constant metabolism leads to massive requirements and turnover of nutrients. The constant turnover of prokaryotes additionally offers the opportunity to generate an enormous genetic diversity through billions of mutation events (2). The large genetic variation leading to continuous adaptation through evolution enables the bacteria to constantly generate new metabolic pathways and cycles. In summary, the fast microbial reproduction accounts for high nutrient requirement, global circulation of these nutrients and generation of metabolic diversity through evolving cells. This supports the conclusion, that microbes have the potential to significantly transform the environment and even catalyze global cycles in such a powerful way that humans become dependent on their presence on Earth.

The enormous abundance, turnover and diversity provides a huge potential for the microbes to actively contribute to the composition of Earth’s properties. These contributions thereby are of such importance that humans would not be able to survive without microbes providing them, which can be shown by several examples. Firstly, marine microorganisms are responsible for the generation of nearly all of the oxygen present in the atmosphere (3). The oxygen produced by plants only contributes to a small fraction of atmospheric oxygen because most of the O2 produced by plants is used up again by their own respiratory processes and upon decay of dead plant material (3). Therefore, the oxygen produced by microbes is substantial for human respiration and hence, also existence. Second, Earth’s redox state depends mostly on microbial life and therefore is an emergent property of their existence (4). This means, the global fluxes of some of the most important elements (H, C, N, O, S) are controlled in large parts by redox reactions which are catalyzed by prokaryotes. Thus, microbial photosynthesis not only provides humans with breathable air, but also drives the Earth’s oxidation in general, and finally, supplies heterotrophic organisms with reduced carbon. Further, not only photosynthesis, but many other microbial processes contribute to important global nutrient cycles. One example for a cycle largely controlled by microbes is the nitrogen cycle (4). Microbes catalyze all the steps present in the global nitrogen cycle and thereby control the oxidation state in which the nitrogen species are present in Earth’s atmosphere, soil and oceans. Many other prokaryotes and eukaryotes, that cannot process atmospheric nitrogen, rely on the supply of these nitrogen species provided by nitrogen-fixing microbes. Also humans rely on fixed nitrogen, which initially is provided by microbes and wanders through the food chain until eventually taken up as part of the human nourishment. All of these examples show, that the global nutrient cycles, which are in large parts controlled by microbes, are of such importance, that human existence relies on their catalysis provided by microbes.

The potential of microbes to catalyze global processes does not only affect nutrient cycles, but also the global environment and climate. Microorganisms can significantly change the climate by altering the atmospheric composition (1, 3, 4). The production of several gases like methane and nitrous oxide strongly influences the global climate through the greenhouse effect (1, 3). The great importance of this effect for the survival of organisms can be shown not only for present days, but also for early stages in time. In the distant past, about 3.5 billion of years ago, microbial methane production might have contributed to a global shield, warming up the planet (1). At this time, the sun was much weaker than today and therefore, without a greenhouse gas like methane, Earth would have been completely frozen over. By keeping the earth from freezing, this methane shield might have allowed for fast reproduction, leading to evolution and the emergence of more complex life-forms. About 500 million years later, microbes again significantly transformed the environment. The emergence of photosynthetic cyanobacteria led to a strong increase of the oxygen level in the atmosphere (1). This resulted in decreased viability of methanogenic organisms and therefore decreased methane concentrations, which in turn reduced the greenhouse effect. The consequence was a global glaciation, again significantly changing the composition of organisms capable of living on Earth (1). In summary, in the absence of microbes, the global environment with its properties as present now, could not have been created. In a world depleted of the global catalysis provided by microbes, Earth’s environment would have been very unfavorable for the emergence of complex life forms as they are present today. As the past times show, a stable environment, which can be provided and maintained by microbes, will also be necessary for humans to live in future times. If humans wanted to survive without microbes, they would have to artificially control all the cycles currently run by microbes. This, however, is most likely not possible. Humans will not be able to replace all the microbially driven cycles in an efficient way. We are not able to generate machines catalyzing processes as efficient as microbes do after millions of years of evolution. Further, as microbial diversity is unimaginably high, the range of the processes they catalyze can always adapt quickly to changing conditions. Humans will not be able to adapt their engineered machines fast enough in order to provide sufficient flexibility to changing demands.

In summary, the enormous number of prokaryotic cells present, their fast turnover and the high genetic diversity offer an unimaginable potential for the microbes to significantly contribute to the presence and properties of biogeochemical cycles. Microbes provide a global environment-composition which allowed for the emergence of humans and will also in the future be necessary for our persistence. From the beginning of our existence, we have been living in symbiosis with microbes. We do not only interact with microbes living in and on our bodies, but also with the global biogeochemical cycles they are a significant part of. Artificially replacing all of the processes provided by microbes will not be possible in a sufficient way. Therefore, existence of human life without the presence of microbes is most likely not possible as our survival strongly depends on stable global cycles, kept intact by microbes. Microbes have been the guardians of global metabolism for billions of years (5). It is very unlikely that humans could further exist without microbes present, continuously guarding global metabolic processes. A new question that arises is however, how likely it is, that the human race eventually manages to destroy this essential function of the microbes as guardians of global metabolism?

#### References

1. Nisbet, E. G., and N. H. Sleep. 2001. ‘The habitat and nature of early life’, Nature, 409: 1083.
2. Whitman, W. B., D. C. Coleman, and W. J. Wiebe. 1998. ‘Prokaryotes: the unseen majority’, Proc Natl Acad Sci U S A, 95: 6578-83.
3. Kasting, J. F., and J. L. Siefert. 2002. ‘Life and the evolution of Earth’s atmosphere’, Science, 296: 1066-8.
4. Falkowski, P. G., T. Fenchel, and E. F. Delong. 2008. ‘The microbial engines that drive Earth’s biogeochemical cycles’, Science, 320: 1034-9.
5. Waters, Colin N., Jan Zalasiewicz, Colin Summerhayes, Anthony D. Barnosky, Clément Poirier, Agnieszka Galuszka,  
   Alejandro Cearreta, Matt Edgeworth, Erle C. Ellis, Michael Ellis, Catherine Jeandel, Reinhold Leinfelder, J. R. McNeill, Daniel deB. Richter, Will Steffen, James Syvitski, Davor Vidas, Michael Wagreich, Mark Williams, An Zhisheng, Jacques Grinevald, Eric Odada, Naomi Oreskes, and Alexander P. Wolfe. 2016. ‘The Anthropocene is functionally and stratigraphically distinct from the Holocene’, Science, 351.

### Module 01 references

Whitman WB, Coleman DC, and Wiebe WJ. 1998. Prokaryotes: The unseen majority. Proc Natl Acad Sci USA. 95(12):6578-6583.  
[Whitman et al. 1998](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC33863/)

Falkowski, P. G., T. Fenchel, and E. F. Delong. 2008. ‘The microbial engines that drive Earth’s biogeochemical cycles’, Science, 320: 1034-9.  
[Falkowski et al. 2008](https://www.ncbi.nlm.nih.gov/pubmed/18497287)

Kasting, J. F., and J. L. Siefert. 2002. ‘Life and the evolution of Earth’s atmosphere’, Science, 296: 1066-8.  
[Kasting et al. 2002](http://science.sciencemag.org/content/296/5570/1066)

Nisbet, E. G., and N. H. Sleep. 2001. ‘The habitat and nature of early life’, Nature, 409: 1083.  
[Nisbet et al. 2001](https://www.nature.com/articles/35059210)

Waters, Colin N., Jan Zalasiewicz, Colin Summerhayes, Anthony D. Barnosky, Clément Poirier, Agnieszka Galuszka,  
Alejandro Cearreta, Matt Edgeworth, Erle C. Ellis, Michael Ellis, Catherine Jeandel, Reinhold Leinfelder, J. R. McNeill, Daniel deB. Richter, Will Steffen, James Syvitski, Davor Vidas, Michael Wagreich, Mark Williams, An Zhisheng, Jacques Grinevald, Eric Odada, Naomi Oreskes, and Alexander P. Wolfe. 2016. ‘The Anthropocene is functionally and stratigraphically distinct from the Holocene’, Science, 351.  
[Waters et al. 2016](http://science.sciencemag.org/content/351/6269/aad2622)

[The Washington Post, 12-01-03, ’Spaceship Earth: A new view of envrionmentalism](https://www.washingtonpost.com/national/health-science/spacesh/)

[Canfield et al. 2010](http://science.sciencemag.org/content/330/6001/192)

A. Leopold, 1949, ‘The Land Ethic’, from ‘A Sand County Almanac’

[Rockstrom et al. 2009](https://www.nature.com/articles/461472a)

D. P. Schrag, 2012, ‘Geobiology of the Anthropocene’

Zehnder, 1988, ‘Biology of anaerobic microorganisms’, chapter 1

# Module 2

## Module 02 portfolio check

* Evidence worksheet\_03
  + Completion status:
  + Comments:
* Problem Set\_03
  + Completion status:
  + Comments:
* Supergraphic
  + Completion status:
  + Comments:
* Writting assesment\_02
  + Completion status: N/A
  + Comments: cancelled
* Additional Readings
  + Completion status:
  + Comments:

### Problem set 03

[Wooley et al. 2010](http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000667) [Madsen et al. 2005](https://www.ncbi.nlm.nih.gov/pubmed/15864265)

#### How many prokaryotic divisions have been described and how many have no cultured

representatives (microbial dark matter)?

More than 40 major bacterial divisons and 12 divisions in archeae have been described (1,2). The total number of microbial species on earth is estimated to be up to 1012 (3). Up to 99.8% of microbial species are not culturable (4).

1. [Pace et al. 1997](https://www.ncbi.nlm.nih.gov/pubmed/9115194)
2. [DeLong et al. 2001](https://www.ncbi.nlm.nih.gov/pubmed/12116647)
3. [Locey et al. 2016](http://www.pnas.org/content/113/21/5970)
4. [Streit et al 2004](https://www.ncbi.nlm.nih.gov/pubmed/15451504)

In-class solution:  
89 bacterial phyla, 20 archaeal phyla  
->But could be up to 1500 bacterial phyla  
->26 of 52 major bacterial phyla have been cultivated.(But only few examples for some phyla.-> still most species uncultivated)

#### How many metagenome sequencing projects are currently available in the public domain

and what types of environments are they sourced from?

The Gold database counts 1424 Metagenomic studies, with 51513 sequencing projects  
[Gold database](https://gold.jgi.doe.gov/studies?Study.Metagenomic+Study=Yes&Study.Is+Public=Yes)

The main environment types are:  
-Marine  
-Soil  
-Sediment  
-Gut of humans/ animals -Especially environments where it is hard to cultivate communities in the lab.

#### What types of on-line resources are available for warehousing and/or analyzing

environmental sequence information (provide names, URLS and applications)?

Types of resources cover metagenomics Assembly, Binning, Annotation, Analysation. Some examples are:

Shtogun metagenomics:

* Assembly:  
  [EULER](https://omictools.com/euler-sr-tool)
* Binnig:
* MEGAN, CARMA, Phymm: For toxonomy-dependent binning  
  [MEGAN](http://ab.inf.uni-tuebingen.de/data/software/megan5/download/welcome.html)  
  [CARMA](https://omictools.com/carma-tool)  
  [Phymm](https://omictools.com/phymm-phymmbl-tool)  
  (MEGAN: Also for comparing OTUs of different samples)
* TETRA: Taxonomy-independent binning  
  [TETRA](http://www.megx.net/tetra)
* Annotation:  
  KEGG: Metabolic pathways, combining genes, proteins, metabolites [KEGG](http://www.genome.jp/kegg/)
* Analysis:  
  [Megan 5](http://ab.inf.uni-tuebingen.de/data/software/megan5/download/welcome.html)
* MG-RAST: Gene calling and annotation by sequence similarity. - Analysis and processing database.  
  [MG\_RAST](https://metagenomics.anl.gov/)

Marker Gene metagenomics:

* Standalone software:  
  [OTUbase](https://www.bioconductor.org/packages/release/bioc/html/OTUbase.html)
* Analysis:  
  [SILVA](https://www.arb-silva.de/)
* Denoising:  
  [AmpliconNoise](http://nebc.nox.ac.uk/bioinformatics/docs/Ampliconnoise.html)
* Databases:  
  [Ribosomal Database Project (RDP)](https://rdp.cme.msu.edu/)
* Some more online sources:
* BLAST: Finding homologs in DNA/ protein sequences  
  [BLAST](https://blast.ncbi.nlm.nih.gov/Blast.cgi)
* RAMMCAP: Clusters translated ORFs by sequence similarity, can also be used to compare metagenomes.  
  [RAMMCAP](http://weizhong-lab.ucsd.edu/rammcap/rammcap_20130510.html)
* Pfam: Protein family database, uses Hidden markov model  
  [Pfam](http://pfam.xfam.org/)
* UniFrac, MetaStats: To compare metagenomes based on phylogenetic data  
  [UniFrac](http://bmf.colorado.edu/unifrac)  
  [MetaStats](http://www.cbcb.umd.edu/software/metastats)
* ShotgunFunctionalizeR: Metagenome analysation in R  
  [ShotgunFunctionalizeR](http://shotgun.zool.gu.se)

sources: [Oulas et al. 2015](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4426941/) and Wooley et al. 2010

#### What is the difference between phylogenetic and functional gene anchors and how can they be used in metagenome analysis?

Phylogenetic gene anchors:

* Are used to determine the different OTUs present in a sample and to order them in a tree. 16s RNA is mostly used as phylogenetic anchor gene.
* Vertical gene transfer
* Ideally single-copy genes
* Taxonomic

Functional gene anchors:

* Show the genes for important metabolic functions that are present in a certain environment. These can be used to determine the specific metabolic potential of a community in its respective environment.
* More horizontal gene transfer, therefore not so usefull for phylogeny

[Krause et al. 2008](https://www.ncbi.nlm.nih.gov/pubmed/18285365)

#### What is metagenomic sequence binning? What types of algorithmic approaches are used to produce sequence bins? What are some risks and opportunities associated with using sequence bins for metabolic reconstruction of uncultivated microorganisms?

Binning: To build longer sequences from overlapping fragments and assign them to OTUs. -> Group fragments together that come from the same genome.  
Approaches:

* Taxonomy-dependent: Supervised
  + Sequences get assigned to homologs in reference sequence databases, often using BLAST
  + OTUs can be assigned to a last common ancestor.-> Phylogenetic trees are established. (MEGAN/ CARMA/ Phymm algorithm)
* Good method if sequences are similar to referencce sequence.
* Composition-dependent: Unsupervised
  + No reference sequences needed. -> Good for unknown sequences without homologs
  + Assignment based on GC content or short nucleotide sequences like tetranucleotides (TETRA algorithm).
  + Other method: Based on codon-usage: Cann be combined with TETRA to improve binning.
* Assambling algorithm: Represent each read as a vertex and detect overlapping vertices. Find correct assembly: NP- Hamiltonian path problem.-> gets too time expensive when amount of fragments is too high, as produced by second gen sequencing of metagenome. Solution: use k-mer words instead of the whole reads. -> Assembling can now be done in linear-time with Eulerian path.
* Other method: ORFome assembly: First find putative open reading frames and then assemble only these regions. - Problem: non-coding regions or undetected ORFs get lost.(Wooley et al, 2010)

Risks:

* Composition-based binning has high missclassification rates, especially when many close related OTUs present in the sample.
* Assembling sequences from different OTUs together. This generates interspecies chimeras which don’t represent the actual metabolic potential of an exisisting OTU.
* Not DNA from all organisms of the habitat present, some are more abundant.
* Not the whole genomes covered in the samples.
* ORFs often not detected because fragments only contain parital ORFs
* Some sequences cannot be assigned:  
  Long repeat sequences, sequences that have no homologs in database yet
* Cloning bias: Some sequences do not get incorporated in cloning vector because they are toxic for the host

Opportunities:

* To infer the genomic/metabolic potential of whole communities.
* Fast and efficient process, algorithms.
* Algorithms are continously improving.

Source: Previous genomics lecture at ETH Zürich and Wooley et al, 2010

(In-class solution:  
Binning is the process of grouping sequences that come from a single genome  
Types of algorithms:  
1. Align sequences to database  
2. Group to each other based on DNA characteristics: GC content, Codon Usage

Risks:

* Incomplete coverage of genome sequence
* Contaminaton from different phylogeny)

#### Is there an alternative to metagenomic shotgun sequencing that can be used to access the metabolic potential of uncultivated microorganisms? What are some risks and opportunities associated with this alternative?

Single Cell sequencing: Single cells are separated, the genome amplified using a special polymerase and then sequenced.

Polymerase colonies: Polonies  
The sampled DNA is diluted to separate the single DNA fragments. Single nucleic acids are then amplified by a polymerase. The amplicons can then be sequenced using a next generation sequencing technique like Illumina.

Second Generation sequencing (Illumina):  
DNA fragments get amplified by bridge amplificationo, fragments get sequenced by synthesis, in parallel.

Advantage: More sequences produced than in shotgun-sanger sequencing, cheap per bp, sensitive because of amplification, only ng amounts of DNA needed.  
Problem: The produced sequences are shorter, which makes binning more difficult.

Third generation sequencing:  
PacBio or Oxford Nanopore

* Single DNA molecules get directly real-time sequenced.  
  Advantage: No amplification needed, very long reads, fast, good for difficult genomes, eg with long repeat sequences With third gen sequencing, even unassembled reads can be used to identify the metabolic function. The long single reads can be directly BLASTed to find ORFs  
  Disadvantage: Higher per-base pair costs, less fragments, more sample needed (ug in 3rd gen vs ng in second gen)

Other methods:

* Functional screens (biochemical): Directly asses the metabolic potential of a colony/community for example by testing the ability to use a certain substrate.
* FISH probe: In situ hybridization with 16sRNA oligos or mRNA matching oligos to observe expressed genes, metabolic potential of a community

Source: Previous genomics lecture at ETH Zürich and Wooley et al, 2010

### Evidence worksheet 04

[Martinez et al. 2007](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1838496/)

#### What were the main questions being asked?

* The main goal was to characterize the proteorhodopsin (PR) photosystem structure, function and genetics. (PR are retinal binding membrane proteins that catalyze light activated proton transfer across the cell membrane. PR are found globally in a broad range of marine bacteria and archaea in the photic zone.)
* To identify the minimal genetic requirements to transfer phototrophy to a heterologus host.
  + Whether PR can be functionally expressed in E. coli and if there are other genes (eg for biosynthesis of cofactors like retinal…) required to induce photophosphorylation.

#### What were the primary methodological approaches used?

* Take DNA from the environment, transfer it to E. coli, induce an observable phenotype. -> Program a heterologus host and look for gene expression responsible for the induced phenotype.
* E. coli cells were transformed with a large-insert fosmid DNA library from marine picoplankton.
* The cells were then screened for PR-containing clones which were expected to show an organge phenotype when grown on retinal containing media. A vector copy-control system was used to enhance the assay sensitivity by inducing high copy numbers of the vector and therefore higher expression. (L-arabinose was used to induce single copy vector to replicate to up to 100 copies)
* The isolated vectors which induced the phenotype were sequenced by transposon-sequencing.
* The generated transposon mutants were further analyzed to confirm the functional annotation of the cloned photosystem biosynthesis genes. The transposon mutants did not show the orange pigmentation phenotype anymore and HPLC analysis showed, that retinal production was inhibited. this confirmed that the cloned genes are necessary and sufficient for retinal biosynthesis.
* Light activated PR-catalyzed proton transfer was confirmed by measuring the pH in media containing PR+ or PR- colonies.
* Light induced ATP synthesis was tested by measuring ATP levels with a luciferase assay. As controls, DCCD and CCCP were added to the cells to inhibit the ATPsynthase or generate a H+ permeable membrane respectively

#### Summarize the main results or findings.

* PR-based photosystems can be functionally expressed in E. coli, even without addition of exogenous retinal:  
  Three colonies were found that showed the orange phenotype. These clones exhibited orange pigmentation even in abscence of exogenous retinal, which indicates that the cells not only obtained the PR gene, but also genes for retinal biosynthesis. Sequencing revealed that a operon containing six genes involved in beta-carotene and retinal biosynthesis was located adjacent to the PR gene. This means, the retina biosynthetic pathway is associated with the PR gene and these genes can radily be transfered together.
* In these transformed E.coli colonies, light activates PR catalyzed proton transport through the cell membrane, generating a proton gradient.
* The light activated PR catalyzed proton translocation activates photophosphorylation (ATP synthesis by ATP synthase). This shows that a single DNA transfer of the PR gene and its associated retina biosynthesis genes can result in the acquisition of photrophism in E. coli. The opportunity of horizontal gene transfer of the PR gene and associated retinal operon explains the high abundance of PR photosystems amongst different microbes. Even chemoorganotrophic microbes can easily acquire the ability to perform photophosphorylation by just a simple horizontal gene transfer.
* Another result is that they could show that increasing fosmid copy number could enhance gene expression, which facilitates screens where the observation of phenotypes is necessary.

#### Do new questions arise from the results?

* A direct link between PR and enhanced growth induced by light is not definitely shown yet, but the results in this paper in combination with previous findings strongly support the role of PR catalyzed phototrophy in marine microbes.
* Whether the light induced PR-dependent ATP generation could be used in industrial biotechnology.
* Whether PR has other effects like sensory functions.
* In what conditions PR is most important. -> e.g. in starvation, anaerobic conditions.
* Wheter the PR generated proton gradient is used for other functions like flagellar motility or active transmembrane transport.

#### Were there any specific challenges or advantages in understanding the paper (e.g. did the authors provide sufficient background information to understand experimental logic, were methods explained adequately, were any specific assumptions made, were conclusions justified based on the evidence, were the figures or tables useful and easy to understand)?

The paper was very well written and understandable, with clear explanations of methods, figures, results and conclusions.

### Module 2 References

Wooley JC, Godzik A, Friedberg I (2010) A Primer on Metagenomics. PLoS Comput Biol 6(2): e1000667. <https://doi.org/10.1371/journal.pcbi.1000667>  
[Wooley et al. 2010](http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000667)

Madsen, E.L., Identifying microorganisms responsible for ecologically significant biogeochemical processes. Nature Reviews Microbiology, 2005. 3: p. 439.  
[Madsen et al. 2005](https://www.ncbi.nlm.nih.gov/pubmed/15864265)

Pace, N.R., A molecular view of microbial diversity and the biosphere. Science, 1997. 276(5313): p. 734-40.  
[Pace et al. 1997](https://www.ncbi.nlm.nih.gov/pubmed/9115194)

DeLong, E.F. and N.R. Pace, Environmental diversity of bacteria and archaea. Syst Biol, 2001. 50(4): p. 470-8  
[DeLong et al. 2001](https://www.ncbi.nlm.nih.gov/pubmed/12116647)

Locey, K.J. and J.T. Lennon, Scaling laws predict global microbial diversity. Proceedings of the National Academy of Sciences, 2016. 113(21): p. 5970-5975  
[Locey et al. 2016](http://www.pnas.org/content/113/21/5970)

Streit, W.R. and R.A. Schmitz, Metagenomics–the key to the uncultured microbes. Curr Opin Microbiol, 2004. 7(5): p. 492-8  
[Streit et al 2004](https://www.ncbi.nlm.nih.gov/pubmed/15451504)

Oulas A, Pavloudi C, Polymenakou P, et al. Metagenomics: Tools and Insights for Analyzing Next-Generation Sequencing Data Derived from Biodiversity Studies. Bioinformatics and Biology Insights. 2015;9:75-88. <doi:10.4137/BBI.S12462>.  
[Oulas et al. 2015](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4426941/)

Krause, L., et al., Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Res, 2008. 36(7): p. 2230-9.  
[Krause et al. 2008](https://www.ncbi.nlm.nih.gov/pubmed/18285365)

Martinez, A. et al. “Proteorhodopsin Photosystem Gene Expression Enables Photophosphorylation in a Heterologous Host.” Proceedings of the National Academy of Sciences of the United States of America 104.13 (2007): 5590-5595. PMC. Web. 7 Apr. 2018.  
[Martinez et al. 2007](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1838496/)

# Module 3

## Module 03 portfolio check

* Evidence worksheet\_05
  + Completion status:
  + Comments:
* Problem Set\_04
  + Completion status:
  + Comments:
* Writing assessment\_03
  + Completion status:
  + Comments:
* Additional Readings
  + Completion status:
  + Comments

### Problem set 04

library(kableExtra)

## Warning: package 'kableExtra' was built under R version 3.4.4

library(knitr)  
library(tidyverse)  
library(vegan)

## Warning: package 'vegan' was built under R version 3.4.4

## Loading required package: permute

## Warning: package 'permute' was built under R version 3.4.4

## Loading required package: lattice

## This is vegan 2.5-1

#### Part 1

data1 = data.frame(  
 number = c(1:32),  
 name = c("m&m green", "m&m red", "m&m blue", "m&m yellow", "m&m brown", "m&m orange", "skittle brown", "skittle red", "skittle green", "skittle orange", "skittle yellow", "bear red", "bear pink", "bear green", "bear orange", "bear yellow", "bear white", "m&i pink", "m&i green", "m&i yellow", "m&i orange", "m&i red", "worms red", "balls yellow", "balls green", "balls purple", "balls orange", "balls red", "chocolate kiss", "lego pink", "lego yellow", "lego blue" ),  
 characteristics = c("m&m green", "m&m red", "m&m blue", "m&m yellow", "m&m brown", "m&m orange", "skittle brown", "skittle red", "skittle green", "skittle orange", "skittle yellow", "bear red", "bear pink", "bear green", "bear orange", "bear yellow", "bear white", "m&i pink", "m&i green", "m&i yellow", "m&i orange", "m&i red", "worms red", "balls yellow", "balls green", "balls purple", "balls orange", "balls red", "chocolate kiss", "lego pink", "lego yellow", "lego blue"),  
 occurences = c(28,28,60,44,30,63,39,33,42,35,23,15,16,18,15,19,16,39,36,27,32,40,14,4,5,3,5,7,16,7,5,4)  
)  
  
data1 %>%   
 kable("html") %>%  
 kable\_styling(bootstrap\_options = "striped", font\_size = 10, full\_width = F)

number

name

characteristics

occurences

1

m&m green

m&m green

28

2

m&m red

m&m red

28

3

m&m blue

m&m blue

60

4

m&m yellow

m&m yellow

44

5

m&m brown

m&m brown

30

6

m&m orange

m&m orange

63

7

skittle brown

skittle brown

39

8

skittle red

skittle red

33

9

skittle green

skittle green

42

10

skittle orange

skittle orange

35

11

skittle yellow

skittle yellow

23

12

bear red

bear red

15

13

bear pink

bear pink

16

14

bear green

bear green

18

15

bear orange

bear orange

15

16

bear yellow

bear yellow

19

17

bear white

bear white

16

18

m&i pink

m&i pink

39

19

m&i green

m&i green

36

20

m&i yellow

m&i yellow

27

21

m&i orange

m&i orange

32

22

m&i red

m&i red

40

23

worms red

worms red

14

24

balls yellow

balls yellow

4

25

balls green

balls green

5

26

balls purple

balls purple

3

27

balls orange

balls orange

5

28

balls red

balls red

7

29

chocolate kiss

chocolate kiss

16

30

lego pink

lego pink

7

31

lego yellow

lego yellow

5

32

lego blue

lego blue

4

#### Ask yourself if your collection of microbial cells from seawater represents the actual diversity of microorganisms inhabiting waters along the Line-P transect. Were the majority of different species sampled or were many missed?

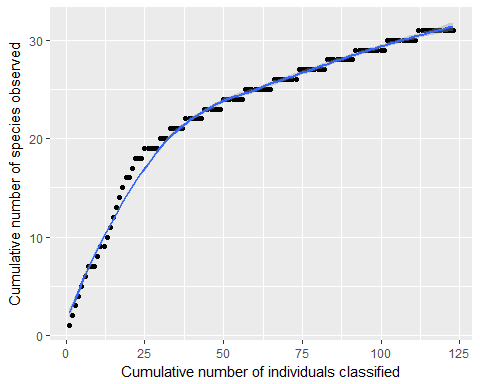
There might be many species missed, as we only have 32 different species

#### Part 2

y = c(1,2,3,4,5,6,7,7,7,8,9,9,10,11,12,13,14,15,16,16,17,18,18,18,19,19,19,19,19,20,20,20,21,21,21,21,21,22,22,22,22,22,22,23,23,23,23,23,23,24,24,24,24,24,24,24,25,25,25,25,25,25,25,25,25,26,26,26,26,26,26,26,26,27,27,27,27,27,27,27,27,27,28,28,28,28,28,28,28,28,28,29,29,29,29,29,29,29,29,29,29,30,30,30,30,30,30,30,30,30,30,31,31,31,31,31,31,31,31,31,31,31,31)  
  
data2 = data.frame(  
   
 y = c(1,2,3,4,5,6,7,7,7,8,9,9,10,11,12,13,14,15,16,16,17,18,18,18,19,19,19,19,19,20,20,20,21,21,21,21,21,22,22,22,22,22,22,23,23,23,23,23,23,24,24,24,24,24,24,24,25,25,25,25,25,25,25,25,25,26,26,26,26,26,26,26,26,27,27,27,27,27,27,27,27,27,28,28,28,28,28,28,28,28,28,29,29,29,29,29,29,29,29,29,29,30,30,30,30,30,30,30,30,30,30,31,31,31,31,31,31,31,31,31,31,31,31),  
 x = c(1: length(y))  
)

ggplot(data2, aes(x=x, y=y)) +  
 geom\_point() +  
 geom\_smooth() +  
 labs(x="Cumulative number of individuals classified", y="Cumulative number of species observed")

## `geom\_smooth()` using method = 'loess'



#### Does the curve flatten out? If so, after how many individual cells have been collected?

The curve does not flatten out completely, however, the slope decreases strongly after 20 observed species

#### What can you conclude from the shape of your collector’s curve as to your depth of sampling?

From the shape of the collectors curve, there would be more species expected to be discovered if the sample was bigger

#### Part 3

sum\_occurence = sum (data1 %>%   
 select(occurences))  
  
spec\_occurence = c()  
for (i in 1:32) {  
 spec\_occurence[i] = data1 %>%  
 filter(number == i) %>%  
 select(occurences)  
}  
spec\_occurence= unlist(spec\_occurence)  
  
  
spec\_occurence

## [1] 28 28 60 44 30 63 39 33 42 35 23 15 16 18 15 19 16 39 36 27 32 40 14  
## [24] 4 5 3 5 7 16 7 5 4

spec\_p =c()  
for (i in 1: 32){  
 spec\_p[i] = spec\_occurence[i] / sum\_occurence   
}  
  
spec\_p2= spec\_p^2  
  
D= 1 / sum(spec\_p2)  
D

## [1] 22.18718

-> The Simpson Reciprocal Index for the total community is 22.187

data\_sample = data.frame(  
 number = c(1:32),  
 name = c("m&m green", "m&m red", "m&m blue", "m&m yellow", "m&m brown", "m&m orange", "skittle brown", "skittle red", "skittle green", "skittle orange", "skittle yellow", "bear red", "bear pink", "bear green", "bear orange", "bear yellow", "bear white", "m&i pink", "m&i green", "m&i yellow", "m&i orange", "m&i red", "worms red", "balls yellow", "balls green", "balls purple", "balls orange", "balls red", "chocolate kiss", "lego pink", "lego yellow", "lego blue" ),  
 characteristics = c("m&m green", "m&m red", "m&m blue", "m&m yellow", "m&m brown", "m&m orange", "skittle brown", "skittle red", "skittle green", "skittle orange", "skittle yellow", "bear red", "bear pink", "bear green", "bear orange", "bear yellow", "bear white", "m&i pink", "m&i green", "m&i yellow", "m&i orange", "m&i red", "worms red", "balls yellow", "balls green", "balls purple", "balls orange", "balls red", "chocolate kiss", "lego pink", "lego yellow", "lego blue"),  
 occurences = c(8,7,6,2,9,1,7,7,5,2,5,2,2,2,1,5,2,5,4,5,11,7,1,6,1,1,2,2,3,1,2,0)  
)  
  
sample3=c(8,7,6,2,9,1,7,7,5,2,5,2,2,2,1,5,2,5,4,5,11,7,1,6,1,1,2,2,3,1,2,0)  
sum3=sum(sample3)  
spec\_p3 =c()  
for (i in 1: 32){  
 spec\_p3[i] = sample3[i] / sum3  
}  
  
spec\_p3= spec\_p3^2  
  
D= 1 / sum(spec\_p3)  
D

## [1] 21.17906

-> The Simpson Reciprocal Index for my sample is 21.179

schao\_tot = 32 + 0  
schao\_tot

## [1] 32

schao\_sample= 31 + (6^2 / (2\*25))  
schao\_sample

## [1] 31.72

chao1 for the total community is 32  
chao1 for the sample is 31.72

#### Part4

data1\_diversity =   
 data1 %>%   
 select(name, occurences) %>%   
 spread(name, occurences)  
  
data1\_diversity

## balls green balls orange balls purple balls red balls yellow bear green  
## 1 5 5 3 7 4 18  
## bear orange bear pink bear red bear white bear yellow chocolate kiss  
## 1 15 16 15 16 19 16  
## lego blue lego pink lego yellow m&i green m&i orange m&i pink m&i red  
## 1 4 7 5 36 32 39 40  
## m&i yellow m&m blue m&m brown m&m green m&m orange m&m red m&m yellow  
## 1 27 60 30 28 63 28 44  
## skittle brown skittle green skittle orange skittle red skittle yellow  
## 1 39 42 35 33 23  
## worms red  
## 1 14

diversity(data1\_diversity, index="invsimpson")

## [1] 22.18718

specpool(data1\_diversity)

## Species chao chao.se jack1 jack1.se jack2 boot boot.se n  
## All 32 32 0 32 0 32 32 0 1

datasample\_diversity =   
 data\_sample %>%   
 select(name, occurences) %>%   
 spread(name, occurences)  
  
diversity(datasample\_diversity, index="invsimpson")

## [1] 21.17906

specpool(datasample\_diversity)

## Species chao chao.se jack1 jack1.se jack2 boot boot.se n  
## All 31 31 0 31 0 31 31 0 1

-> Simpson index for total community: 22.187  
-> Simpson index for my sample: 21.179  
-> Chao1 for total community: 32  
-> Chao1 for my sample: 31

-> Values match previous calculations

#### Part 5

##### How does the measure of diversity depend on the definition of species in your samples?

If we assign more CFUs to the same species, we end up with less different species and therefore less diversity. Or, if different species definitions lead to the same number of species, but with different abundance-proportions amongst the species, the diversity might again be different.

#### Can you think of alternative ways to cluster or bin your data that might change the observed number of species?

We could have sorted the candies just by color, regardless of their shape, brand.  
Another alternative would have been to ignore the color and sort the candies only by the shape, type (e.g. assign all m&m’s to the same species and all skittels to another one..)

#### How might different sequencing technologies influence observed diversity in a sample?

The observed diversity might be overestimated, if a more error-prone sequencing method is applied. More errors in base calling leads to higher diversity in sequences, leading to a higher number of estimated species.  
Second generation sequencing often has higher error rates and produces shorter reads than sanger sequencing. This leads to overestimated diversity, especially when reads with unresolved bases or abnormal read lenghts are removed from the data.[Kunin et al. 2010](https://www.ncbi.nlm.nih.gov/pubmed/19725865)

### Evidence Worksheet 05

[Welch et al. 2002](http://www.pnas.org/content/99/26/17020)

#### What were the main questions being asked?

* How the genomes of different E. coli strains are composed.
* How the different genome compositions and structures can explain pathogenicity of different E. coli strains.
* How the genome structure can be used to understand to evolutionary history of E. coli.
* ->Find differences between the three very similar strains.

#### What were the primary methodological approaches used?

* Genome sequence of uropathogenic E. coli strain CFT073 compared to enterohemorrhagic strain EDL933 and nonpathogenic lab strain MG1655.
* CFT073 was isolated from a woman with acute pyelonephritis and subsequently sequenced by generating whole-genome libraries.
* The genome sequence was analyzed using MAGPIE. Homologs in EDL933 and MG1655 of the potential CFT073 genes were searched using BLAST. Orthology was assumed when matches reached a threshold of 90% identity with at least 90% of the genes covered in an alignment.

#### Summarize the main results or findings.

* The CFT073 genome of 5.2 Mb was successfully sequenced.
* Several features of the genome could be identified:
  + A restriction map was generated to confirm the circular structure of the genome.
  + The Origin and terminus of the CFT073 genome corresponds to those of the MG1655 strain
  + The genome contains no plasmids
  + 5.533 coding genes were detected
  + 247 CFT073-specific islands found with 2004 genes, 60 unique segments for virulence genes
    - The specific islands account for 1.303 Mb, whereas MG1655 contains only 716 kb that are specific for this strain.
* Codon usage of islands was found to be different from the backbone genes, which is an indicator of lateral gene transfer.
  + The backbone codon usage from CFT073 does not differ from the backbones of the other two strains. But the specific islands have different codon usage amongst the three strains.
  + A Mosaic genome structure is generated where newly acquired genes are grouped together.
  + CFT073 misses type III secretion system and plasmid encoded virulence genes, which explains differences in disease potential from EDL933.
* The CFT073 specific virulence genes contain fimbrial adhesins, autotransporters and phase switch recombinases.
* The chromosomal location and order of the pathogenicity genes differ from other uropathogenic strains.
* One CFT073 island contains part of a pathogenicity island from Yersinia pestis, suggesting the introduction of this island in the early evolution of extra intestinal E. coli.
* Many CFT073 islands are needed for the ability to colonyze the urinary tract:
  + Fimbriae, pili for attachment to host cells.
  + Phase switch recombinases that control the expression of fimbriae encoded on the fim operon.
  + Autotransporters that export virulence factors
  + Hemolysin genes that encode cytolytic toxins and their secretion system.
* Numbers of proteins shared between CFT073, DEL933 and MG1655: 2996 (39.2%)
* Numbers of proteins shared between CFT073 and DEL933: 204 (2.6%)
* Numbers of proteins shared between CFT073 and MG1655: 193 (2.5%)
* Numbers of proteins shared between DEL933 and MG1655: 514 (6.7%)
* Proteins specific for CFT073: 1623 (21.2%)
* Proteins specific for EDL933: 1346 (17.6%)
* Proteins specific for MG1655: 585 (7.6%)

Conclusion:

* E. coli genomes show a mosaic structure with conserved backbones and strand specific islands.
* Uropathogenic stains show bigger genetic differences to other strains than previously assumed.
* Uropathogenic or intestinal pathogenic strains acquired specific pathogenicity islands witch made them able to colonyze the host’s organs.
* Each strain contains different specific islands with strong variation in their linkage and chromosomal arrangement, which gives the strains their characteristic traits.

#### Do new questions arise from the results?

* The results suggest that extraintestinal strains might be as diverse as the intestinal strains. This remains to be verified.
* Whether specific gene subsets can be found that can be used to differentiate between uropathogenic strains.
* Whether new genetic regions can be found that are subject to variation due to recombinase activities.
* Whether “black holes” can be found in CFT073: Deletions of genes that would lead to disadvantages in the specific lifestyle of uropathogens.
* The presence of huge amounts of specific islands raises the question, whether species definitions based on only a few phenotypic traits and low-resolution mapping should be reconsidered

#### Were there any specific challenges or advantages in understanding the paper (e.g. did the authors provide sufficient background information to understand experimental logic, were methods explained adequately, were any specific assumptions made, were conclusions justified based on the evidence, were the figures or tables useful and easy to understand)?

In general, the paper was simple to follow, with reasonable assumptions and conclusions. However, the methods how they analyzed and compared the genomes were not explained very detailed. More detailed description of their methods would have been helpful for better understanding of the obtained results.

#### Based on your reading and discussion notes, explain the meaning and content of the following figure derived from the comparative genomic analysis of three E. coli genomes by Welch et al. Remember that CFT073 is a uropathogenic strain and that EDL933 is an enterohemorrhagic strain. Explain how this study relates to your understanding of ecotype diversity. Provide a definition of ecotype in the context of the human body. Explain why certain subsets of genes in CFT073 provide adaptive traits under your ecological model and speculate on their mode of vertical descent or gene transfer.

An Ecotype is a bacterial species that occupies a specific ecological niche. Different ecotypes can live in the same habitat but do not compete for the same nutrients/ environmental factors and therefore obtain different niches, which allows them to evolve independently from each other. In the human body for example, several different bacteria can live in the gut but rely on different carbon sources/ factors and therefore are different ecotypes. They can also have differences in their exact localization within the gut, eg in the microvilli/ gut lumen/ cryptae.. Different ecotypes can have very identical 16s RNA sequences but differ significantly in genetic islands specific for the niche they occupy.

The figure shows that the different strains have different genetic islands in their genomes. The differences in the islands can be explained by the different niches/ habitats within the human that these two strains occupy. The colonization of the urinary tract requires some different genes than the colonization of the gut. Different genes are needed to colonize the urethra and to metabolize the present nutrients than to colonize and live in the gut. The figure also shows that the specific islands are localized on a shared backbone of the different strains.

The genes that are specific for a distinct niche are mainly inherited vertically to conserve the separation of the different ecotypes. The vertical inheritance of the specific genes (gene islands) conserves the ability of the strains to colonize their specific habitats. If horizontal gene transfer occurs between cells from different ecotypes, a switch in the ecotype of the recipient cell occur, leading to the occupation of a new niche for this ecotype.

### Module 3 essay

Discuss the challenges involved in defining a microbial species and how HGT complicates matters, especially in the context of the evolution and phylogenetic distribution of microbial metabolic pathways. Can you comment on how HGT influences the maintenance of global biogeochemical cycles through time? Finally, do you think it is necessary to have a clear definition of a microbial species? Why or why not?

How to decide for a clear definition of microbial species is an often-discussed issue. The biological species concept cannot be applied for microbes as they do not reproduce sexually. Many alternative concepts exist that use different approaches to define species. Each of these definitions has their own advantages and disadvantages, which makes it difficult to agree on one that fits the needs of all purposes. Some of the most important concepts will be discussed in more detail in this assay. At first, probably the most often applied concept is to assign two individuals to the same operational taxonomic unit (OTU) if their 16s rDNA shows more than 97% identity. A more recently developed method is to define species by amplicon sequence variants (ASV). This method has the potential of even higher resolution than the OTU definition as it defines an exact nucleotide sequence for each species. Another factor that makes species definitions even more challenging is horizontal gene transfer (HGT). This microbial version of “sexual” recombination leads to difficulties in many phylogenetic approaches. Through HGT, even distantly related species can harbor very similar DNA fragments. This allows for a global distribution and maintenance of key metabolic pathways but in turn also leads to a high potential for wrong species-assignments. One method to overcome the problems of HGT is multi-locus sequence analysis (MLSA). MLSA tries to eliminate genes acquired from horizontal transfer by comparing several conserved genes instead of just one, as the OTU and ASV methods do. Regarding all these challenges in species definition, the question arises whether it is even necessary to have a clear definition? In order to find an answer to this question, it is necessary to first obtain a deeper understanding of the present species definitions.

Defining microbial species by a 16s RNA identity of 97% or more has become the standard approach over the last years. This method has been shown to achieve well dissolved phylogenetic assignments in numerous cases. However, with increasing amounts of available data, several problems of this method could be identified. Firstly, as only the 16s rDNA is considered, a lot of variation might be missed. Two species can have completely different sequences in the main parts of their genomes, with just very similar 16s rDNA and thus get clustered together as the same species. Second, one species can have multiple copies of the 16s RNA gene. This means that we have to decide which of the copies we want to use for the species definition. In addition, if the sequence data comes from an unknown environmental sample, one might not even realize that the two 16s rDNA sequences come from the same genome and therefore assign them to two different species. Third, as more genomic data of very closely related species became available, the resolution of 16s rDNA turned out not to be sufficient anymore to separate these species. In these cases, the 16s rDNA shows too high levels of conservation to be sufficient for separation of very similar species. Another challenge that arises from this problem is the decision of how to differentiate between species and strains. This problem again could be the subject of a whole essay.

The more recent amplicon sequence variants (ASV) method has been shown to overcome some of the problems of the 97% rDNA identity OTU definition approach. This method does not rely on an arbitrarily chosen identity threshold, but instead defines species by one exact nucleotide sequence[1]. Therefore, this method provides higher resolution as it distinguishes sequences that even differ by just one nucleotide. However, this method too struggles with some challenges in species definition. In the process of defining one exact sequence to be the true one, many sequences which are assumed to contain sequencing errors have to be discarded. This means, that sequences which are more error-prone get discarded more often and might not be detected as species. In addition, this denoising step is very laborious and therefore requires a lot of capacity if big data sets have to be processed. Another problem is that if a wrong sequence is chosen as a species-defining ASV, the mistake immediately leads to an incorrectly defined species. One challenge the ASV method shares with the OTU method is, that it tries to define species by only one DNA locus. Thus, with ASV again, there is no guarantee that two species actually are the same or even closely related just because they have the same sequence in the observed amplicon.

The past two paragraphs show that the process of defining species can be very challenging. One factor that makes these definitions even more difficult is horizontal gene transfer (HGT). Many bacterial species can exchange some of their genes through HGT even if they are very distantly related. Therefore, individual genes within a single genome can have different evolutionary histories. This leads to varying phylogenetic trees and species assignments, depending on which genes are analyzed. For example, two species can have high similarity in 16s rDNA and other housekeeping genes, but still differ in great parts of their genomes, like specific pathogenicity islands[2]. Else, due to HGT, two species can be identical in most parts of their genomes, with differences in only a few very specific loci [3]. In both cases, 16s rDNA alone would not be sufficient to separate the ecotypes that share the same ribosomal DNA but have clear phenotypic or ecological differences in terms of the niches they occupy. A method to deal with the problem of HGT interfering with the definition of these ecotypes, is multi-locus sequence analysis (MLSA). In MLSA, several conserved housekeeping genes are concatenated and analyzed [4]. This method has two major positive properties. First, the interfering effects of HGT get eliminated, with increasing success, the more different genes are analyzed. Second, using multiple genes for sequence comparison, strongly increases the obtained resolution to distinguish species or ecotypes. Thus, the MLSA method is well worth considering as alternative to the single gene based methods which are more prone to get biased by HGT.

Because of horizontal gene transfer and following selective pressure, the genes for many key metabolic pathways get distributed all over the world in many different species. On one hand, this makes it difficult to generate phylogenetic trees and distinguish species, as discussed in the previous paragraph. But on the other hand, HGT assures the maintenance of global biogeochemical cycles, allowing the microbes to function as guardians of the metabolism [5]. Different environments with strong selective pressures favor the survival of only the best adapted microbes. Bacteria that fail to adapt fast enough to a changing environment go extinct. However, because of HGT, their specialized metabolic pathways often survive in another species in a different environment somewhere else on the planet. Thus, HGT ensures the maintenance of global biogeochemical cycles through time as it enables the distribution of key metabolic genes across species all over the world.

The massive amount of HGT plays an important role in the distribution of metabolic pathways and evolution of microbes. This extensive genomic fluidity strongly increases the challenges in defining microbial species and has led to some scientists questioning the existence of any microbial species [6]. But is it even necessary to have a clear definition of microbial species? Intuitively, one might think that a single, universal definition of microbial species would be very useful. Assigning an individual to a species would be much simpler if just one clear rule with a universal cutoff existed. Also, if two species are compared, one would immediately know the minimal level of genetic difference between the two species. However, despite these obvious and strong advantages of a clear definition, I think that it is important to allow for varying species definitions. An example from Fraser et al [7] illustrates this on three closely related bacterial species: Streptococcus pneumoniae, a human pathogen, S. mitis, a commensal and S. pseudopneumoniae, a species of uncertain status. Of these three species, S. mitis contains the highest genetic diversity of about 5%. This distance is about the same as the average distance between two genomes of S. pneumoniae and S. pseudopneumoniae. Therefore, applying a universal cutoff means that either S. pneumoniae and S. pseudopneumoniae had to be clustered together, or S. mitis would be split into many distinct species. This example clearly shows that one clear species definition with a defined cutoff is not only unnecessary, but even detrimental in many cases. Often, several context dependent aspects have to be considered in order to achieve satisfying clustering of microbes. For the generation of phylogenetic trees, varying phenotypical and ecological properties are arguably as important as genetic similarity. By considering only the 16s rDNA identity, many ecotypes would be missed in the analysis. Thus, it is often important to look at many different metabolic genes to be able to identify differences in ecotypes occupying different niches, despite their great genetic similarity.

In conclusion, the definition of microbial species encounters many challenges. Several approaches exist that try to sort species by a clearly defined threshold of genetic similarity in some single or multiple genes. Each of these methods has their own advantages and disadvantages and none can be clearly considered as the best one. A factor making species definitions even more complicated is HGT. Extensive gene exchanges distribute key genes all across the bacterial domain, leading to the maintenance of biogeochemical cycles by a globally shared core genome. This genome fluidity allows for fast bacterial evolution and adaptation, but in turn causes serious problems in the attempt of defining species. However, a clear species definition is not always necessary after all. Satisfying clustering of microbes often needs context dependent, varying definitions, considering also ecological aspects and many metabolic genes, not just similarity in the 16s rDNA. As a new approach, it might even be worth considering to completely give up on any universal species concepts. Instead, many different methods could be applied simultaneously, leading to several different, even overlapping clusters of taxa. Therefore, for each application, the best clustering method could be chosen to address the particular problem or question in an optimal way.

#### References

1. Callahan, B.J., P.J. McMurdie, and S.P. Holmes, Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J, 2017. 11(12): p. 2639-2643.
2. Welch, R.A., et al., Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proceedings of the National Academy of Sciences, 2002. 99(26): p. 17020-17024.
3. Shapiro, B.J., et al., Population genomics of early events in the ecological differentiation of bacteria. Science, 2012. 336(6077): p. 48-51.
4. Glaeser, S.P. and P. Kampfer, Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. Syst Appl Microbiol, 2015. 38(4): p. 237-45.
5. Falkowski, P.G., T. Fenchel, and E.F. Delong, The microbial engines that drive Earth’s biogeochemical cycles. Science, 2008. 320(5879): p. 1034-9.
6. Riley, M.A. and M. Lizotte-Waniewski, Population Genomics and the Bacterial Species Concept. Methods in molecular biology (Clifton, N.J.), 2009. 532: p. 367-377.
7. Fraser, C., et al., The Bacterial Species Challenge: Making Sense of Genetic and Ecological Diversity. Science, 2009. 323(5915): p. 741-746.

### Module 3 References

Welch, R.A., et al., Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proceedings of the National Academy of Sciences, 2002. 99(26): p. 17020-17024.  
[Welch et al. 2002](http://www.pnas.org/content/99/26/17020)

Callahan, B.J., P.J. McMurdie, and S.P. Holmes, Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J, 2017. 11(12): p. 2639-2643.  
[Callahan et al. 2017](https://www.ncbi.nlm.nih.gov/pubmed/28731476)

Gaudet, A.D., et al., Small-Group Learning in an Upper-Level University Biology Class Enhances Academic Performance and Student Attitudes Toward Group Work. PLOS ONE, 2011. 5(12): p. e15821.  
[Gaudet et al. 2010](http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0015821)

Hallam, S.J., M. Torres-Beltran, and A.K. Hawley, Monitoring microbial responses to ocean deoxygenation in a model oxygen minimum zone. Sci Data, 2017. 4: p. 170158  
[Hallam et al. 2017](https://www.ncbi.nlm.nih.gov/pubmed/29087370)

Hawley, A.K., et al., A compendium of multi-omic sequence information from the Saanich Inlet water column. Sci Data, 2017. 4: p. 170160  
[Hawley et al. 2017](https://www.ncbi.nlm.nih.gov/pubmed/29087368)

Kunin, V., et al., Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ Microbiol, 2010. 12(1): p. 118-23  
[Kunin et al. 2010](https://www.ncbi.nlm.nih.gov/pubmed/19725865)

Sogin, M.L., et al., Microbial diversity in the deep sea and the underexplored “rare biosphere”. Proceedings of the National Academy of Sciences, 2006. 103(32): p. 12115-12120  
[Sogin et al. 2006](http://www.pnas.org/content/103/32/12115)

Torres-Beltrán, M., et al., A compendium of geochemical information from the Saanich Inlet water column. Scientific Data, 2017. 4: p. 170159. [Torres\_Beltràn et al. 2017](https://www.nature.com/articles/sdata2017159)