# Block Course Projects

DALL E

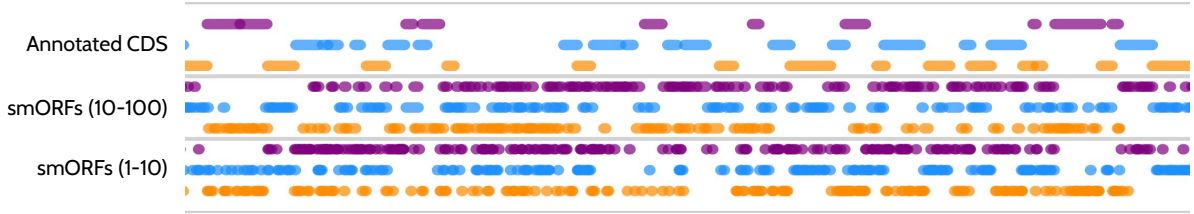
Projects 1 - 3

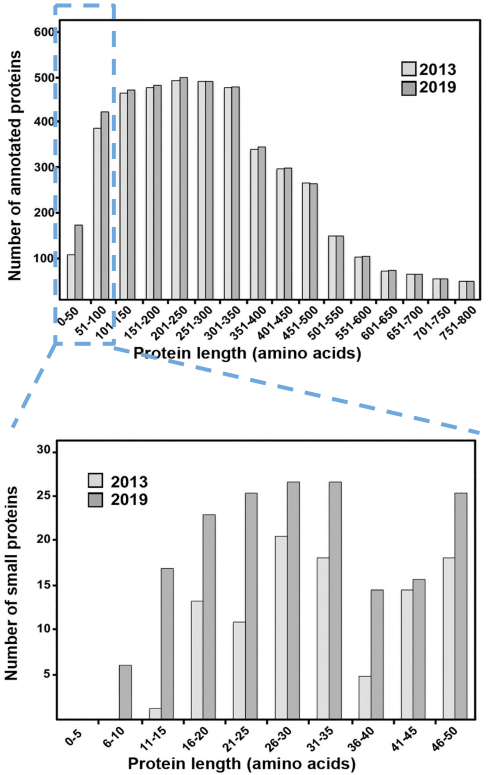
# Biological determinants on smORFs and AMPs rates (also within BGCs) in the ocean

–

Project descriptions

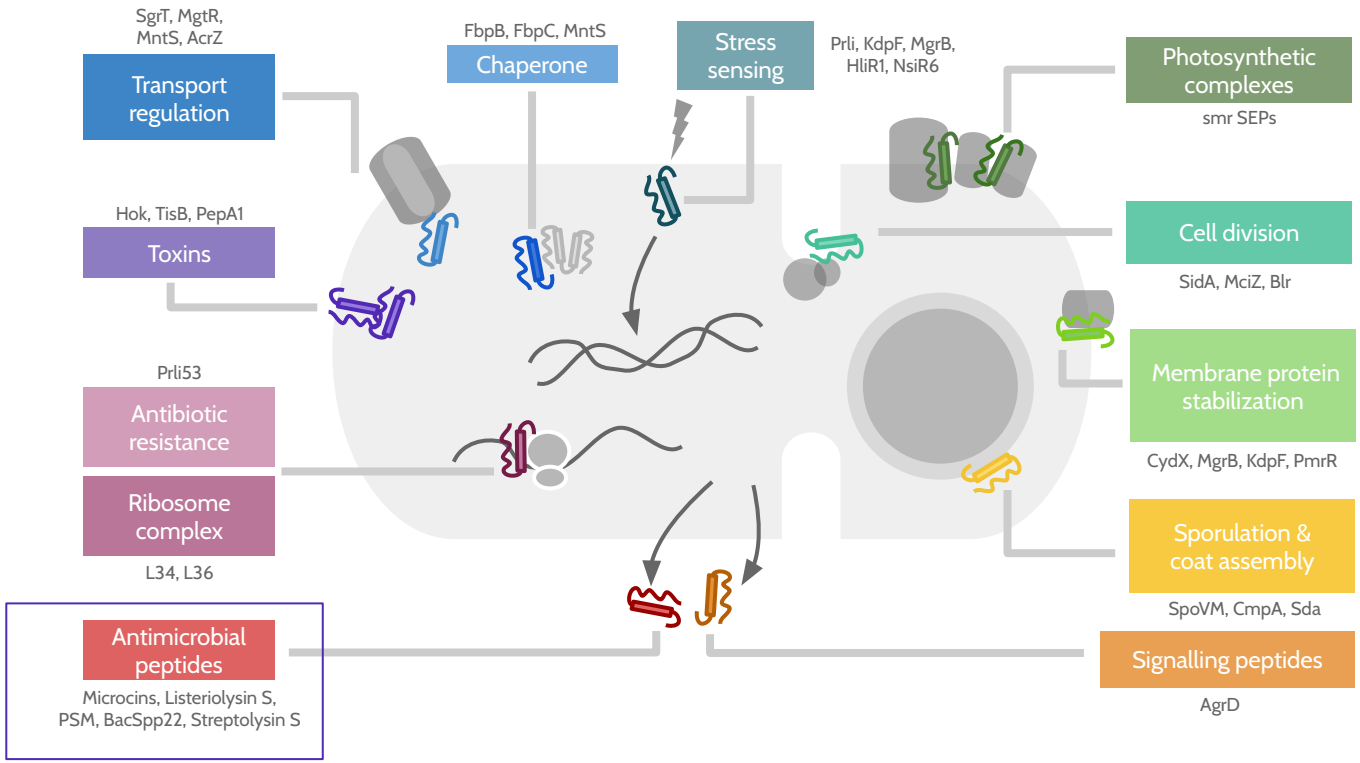
# smORFs are a pool for protein and function discovery



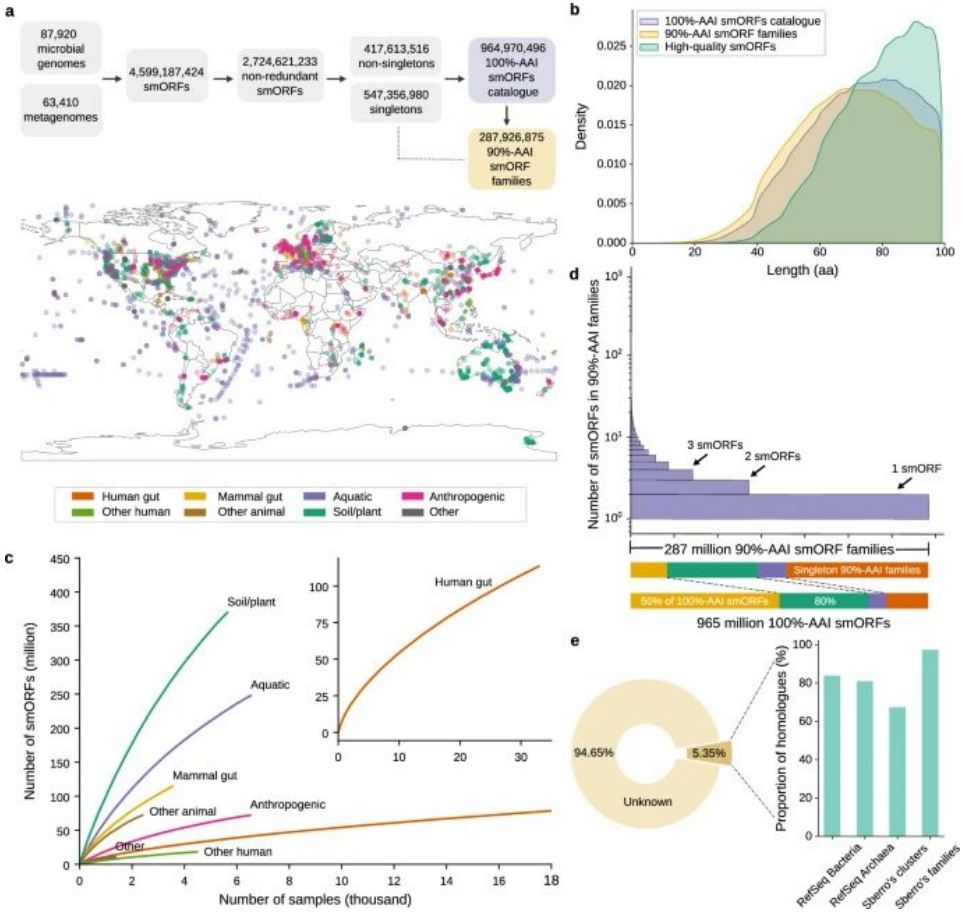

Total number of new SEPs doubled number in *E. coli.* Adapted from Hemm MR, *et al.* (EcoSal Plus 2020)

# smORFs are a pool for protein and function discovery

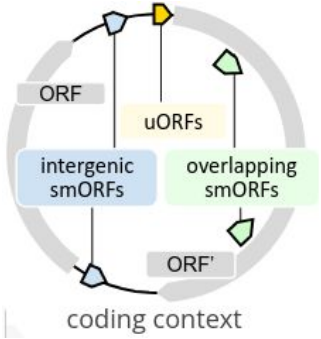



From 'Development of computational and experimental tools for the identification of small proteins in bacterial genomes' - S Miravet-Verde (2021) https://www.tdx.cat/handle/10803/671772

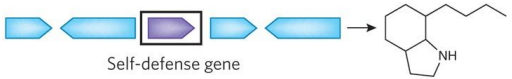
# smORFs are a pool for protein and function discovery




?








https://www.nature.com/articles/s41467-024-51894-6

Projects 1 - 3

# Biological determinants on smORFs and AMPs rates (also within BGCs) in the ocean

–

## General Introduction

---

All **files** and README.md describing tables provided in:

`/nfs/teaching/551-1119-00L-2024/masterdata_smorfs`

# P1. Genomic determinants of smORFomes diversity

**Rationale:** number of smORFs dramatically surpasses the number of ORFs in a genome but a large percentage of them are not going to encode for a protein (i.e., SEP). These smORFs represent an evolutionary pool for new functions as they might acquire mutations that will allow their expression or even become part of larger genes. Additionally, only intergenic smORFs have been studied so far, but certain bacteria can have overlapping expression of genes. Thus, understanding which factors determine the number of smORFs an organism will present (also including partial and overlapping smORFs) can provide insight into functional adaptation and its potential to encode for novel genes.

## Goals:

1. Work with the counting of context types intergenic, partial and overlapping smORFs in a subset of ~90k genomes from OMD2.
2. Explore the ratios of each smORF context type. Is intergenic always the dominant class?
3. Merge the OMD2 genome collection information and metadata to explore the relationship between different factors with the smORF rates estimated:
    - 3.1. What is the impact of completeness and contamination in the numbers of smORFs we can identity in a genome?
    - 3.2. At genomic level, you will explore the impact of genome size (do smORF count scale linearly with genome size?) and/or GC content (remember the last exercise in the hands-on!).
    - 3.3. At sampling conditions level, do environment (ocean vs. host-associated), location, temperature or depth relate with less/more smORFs in genomes from bacteria living in those sites.
    - 3.4. Who are the bacteria taxa presenting surprisingly low or high smORF numbers?
4. Is the average aa length covered by smORFs constant or it varies by any of the previous factors? (i.e., is there a bias in how long smORF tend to be in a genome?)

# P2. Genomic determinants of AMPs diversity

**Rationale:** Antimicrobial peptides are thought to be of relevance in combating the current antibiotic resistance crisis due to their lower resistance induction and their synthesis availability. So far, only intergenic AMPs have been studied so far and it is still unclear if AMPs encoded overlapping larger genes exist. The search for new AMPs could benefit from understanding which factors determine the number of AMPs an organism will present (also including partial and overlapping) can help in identifying potential 'AMP super producers' and characterizing the factors that will make a microbe to produce more AMPs.

## Goals:

1. Work with the counting of context types intergenic, partial and overlapping AMPs in a subset of ~90k genomes from OMD2.
2. Explore the ratios of each AMP context type. How many partial or full overlapping AMPs can be identified?
    2.1. Do these ratios scale linearly with the number of possible smORFs (more smORFs → more AMPs)? Or a specific selection is observed?
3. Merge the OMD2 genome collection information and metadata to explore the relationship between different factors with the AMPs rates estimated:
    3.1. What is the impact of completeness and contamination in the numbers of AMPs we can identity in a genome?
    3.2. At genomic level, you will explore the impact of genome size (do AMP count scale linearly with genome size?) and/or GC content (remember the last exercise in the hands-on!).
    3.3. At sampling conditions level, do environment (ocean vs. host-associated), location, temperature or depth relate with less/more AMP in genomes from bacteria living in those sites.
    3.4. Who are the bacteria taxa presenting surprisingly low or high AMP numbers? Any super producer?
4. Are organisms presenting more AMPs more abundant in the ocean? (i.e., they compete better in the environment thus be present in higher levels in the community)

Note there are 2 types of AMPs and they can be hemolytic or not, these are different categories to take into account!

# P3. Exploring the relation between smORFs and AMPs with BGCs

**Rationale:** Biosynthetic gene clusters can produce compounds that improve the fitness or survival of the organism expressing it. BGCs have been extensively studied but smORFs (and consequently the potential encoded SEPs) have never been integrated in those analysis. Functional characterization of SEPs is extremely challenging but in bacteria it is common to find genes with similar functions close to each other. Some of these, could be in fact AMPs making BGCs even more biotechnologically/biomedically relevant but this analysis has not been addressed neither. Thus, exploring the relationship between smORFs (also including partial and overlapping) and BGCs could highlight or provide hints on potential functions these putative genes could perform.

## Goals:

1. Work with the counting of context types intergenic, partial and overlapping smORFs in relation to BGCs in a subset of ~90k genomes from OMD2.
2. Define a working metric to represent properly the probability to find a smORF within a BGC.
    2.1. Be aware of the biases: number of BGCs, number of smORFs…
    2.2. Do ratios between intergenic, partial or full overlapping compare?
    2.3. Are these ratios scale in the same way than the number of smORFs per genome? (do we find more or less smORFs in BGCs compared to the whole genome).
3. Reutilize the previous analysis to run the same comparisons with AMPs. Are AMPs enriched in BGCs or the probability is comparable to find AMPs at whole genome level.
4. We can explore AMPs in BGCs in relation to the product synthesized.
    4.1. Is the probability to find AMPs in a BGC the same independently of the product that BGC synthesizes? Or there are specific BGC types predicted to contain more AMPs?

Note: you can merge the OMD2 genome collection information, metadata to explore the relationship between different factors with the smORFs/AMPs/BGCs rate estimated (such as taxa, genome metrics, and so on).

Projects 4 - 6

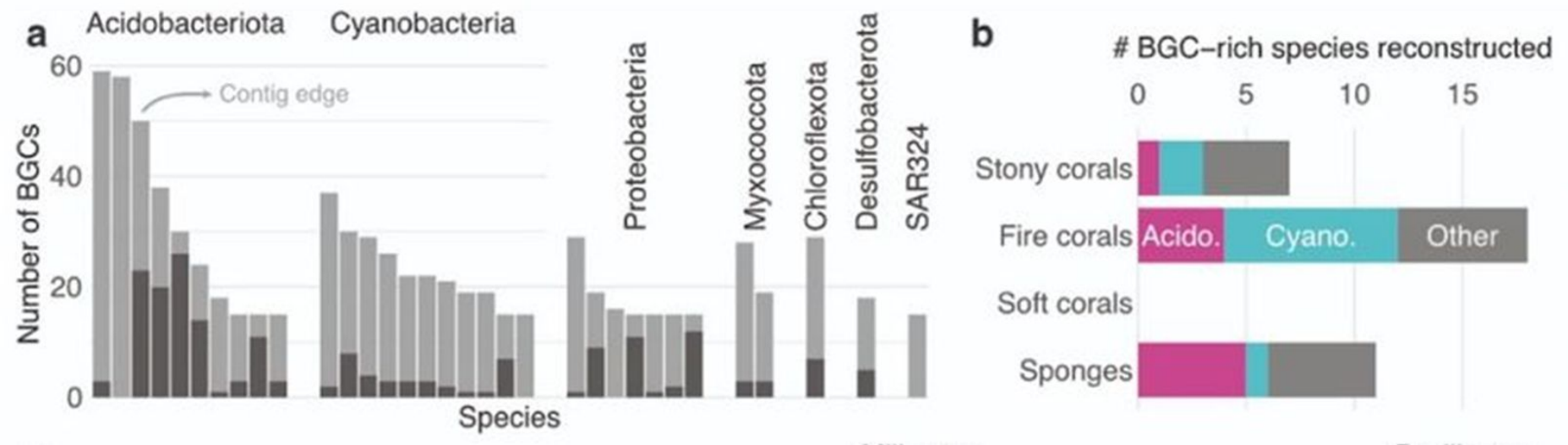
# The Ecology of Marine Acidobacteriota

–

General Introduction

# Coral-associated Acidobacteriota are superproducers for natural products

**(enrichment of Biosynthetic Gene Clusters – BGCs)**




Wiederkehr, Paoli and Richter *et al*., *in-revision;* https://doi.org/10.1101/2024.08.18.608444

Only two isolates of marine Acidobacteriota yet available

(compared to ~ 65 terrestrial Acidobacteriota isolates)

→ We lack understanding about their ecological role in the ocean

→ We require new isolation strategies

# *Acanthopleura japonica → Acanthopleuribacter pedis*

http://animal.memozee.com/view.php?tid=5&did=5542

# *Porites lutea → Sulfidibacter corallicola*

https://en.wikipedia.org/wiki/Porites_lutea

# Utilizing a coral-reef model system to study the ecophysiology of Acidobacteriota

+ Sea anemone (*Aiptasia*)
+ Algae (*Symbiodinium*)
+ Bacteria

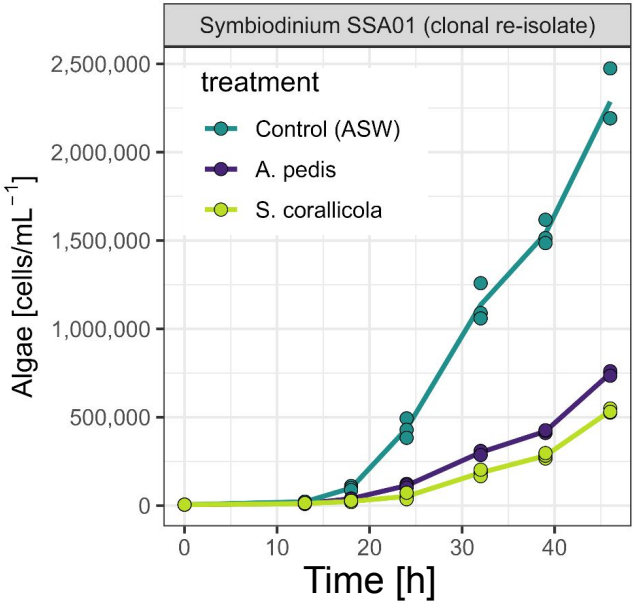
# *Acanthopleuribacter pedis* kills anemones

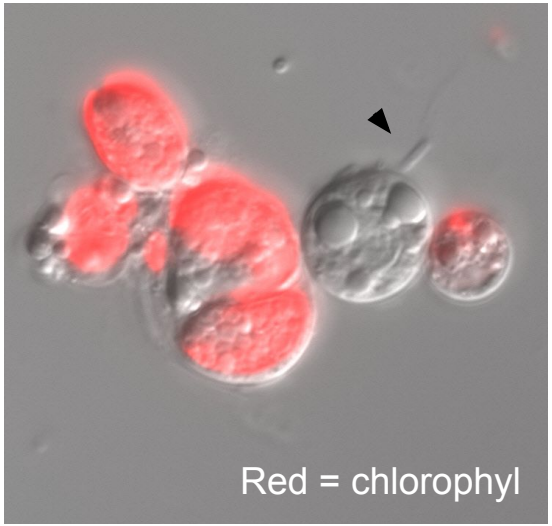
Control

+ A. pedis (rep. 1)

+ A. pedis (rep. 2)

# Algae grow slower in the presence of Acidobacteriota




Algae-Acidobacteriota co-culture

Projects 4 - 6

# The Ecology of Marine Acidobacteriota

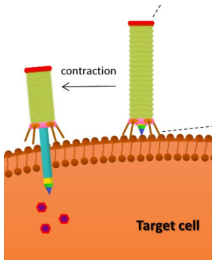
–

Project descriptions

# P4. Extracellular contractile injection systems (eCIS) in Acidobacteriota

- **Rationale:** Extracellular contractile injection systems (eCIS) could expand the weapon arsenal of Acidobacteriota. Preliminary analyses revealed traces of eCIS (i.e. phag tail fibres) in BGC-superproducing *Acanthopleuribacteraceae* (family of Acidobacteriota), hower, a systematic profiling awaits to be conducted.

- **Goals:**




1) Compare both PFAM annotation methods (eggNOG vs HMMER), using the ~7k Acidos as test dataset.
2) Adopt strategy to identify eCIS in bacteria based on PFAMs (see Geller *et al*., 2021, Chen et al., 2019 and eCIS DB).
3) identify eCIS superproducing taxa in OMDv2.
4) Are Acidobacteriota in the upper, middle, or lower range of eCIS production (in OMDv2)?
5) Correlation of Acidobacteriota eCIS with Biogeography (e.g. terrestrial vs marine; water vs host?).
6) Correlation of Acidobacteriota eCIS abundance with BGC abundance?
7) How often are Acidobacteriota eCIS within or next to BGCs?
8) Search for potential cargo of Acidobacteriota eCIS by qualitative analysis of enclosed or neighboring genes (PFAMs). This could be e.g. cell-lysis functions (holins, lysins, spanins).

# P5. Complex sugar degradation in Acidobacteriota

- **Rationale:** Some terrestrial Acidobacteriota or known to be potent degraders of polysaccharides (complex sugars), which is a function that may be attributed with a host-associated lifestyle or pathogenicity; Kielak et al., 2016, Wang et al., 2022). Profiling the complex sugar degradation potential in marine Acidobacteriota will shed light into their lifestyle, and could be harnessed as a novel isolation strategy.

- **Goals:**
  1) Compare both CAZy annotation methods (eggNOG vs dbcan), using the ~7k Acidos as test dataset.
  2) Compare the complex sugar degradation potential in ~7k Acidos to other bacteria (e.g. on phylum or mOTU level) in OMDv2, using eggNOG CAZys.
  3) Explore the dbcan CAZy, CGC counts, and substrate profiles between Acidobacteriota ranks (e.g. families, genera, species, mOTUs…).
  4) Explore the dbcan CAZy, CGC counts, and substrate profiles between Acidobacteriota from different habitats, with specific focus on marine vs terrestrial, and host-associated (e.g. coral, sponge) vs. free living.
  5) Is there a correlation of CGC counts and BGC counts in Acidobacteriota?
  6) Do some specific CAZy susbtrates correlate with specific BGC products in Acidobacteriota?

# P6. Eukaryote-like proteins (ELPs) in Acidobacteriota

- **Rationale:** Eukaryote-like proteins (ELPs) were recently shown to be enriched in coral-associated bacteria (Sweet *et al*., 2021), and they could thus serve as markers for bacteria-host interactions. A preliminary screen of the OMD database indicated that ELPs are enriched in Acidobacteriota, which awaits to be substantiated.

- **Goals:**
  1) Compare both PFAM annotation methods (eggNOG vs HMMR), using the ~7k Acidos as test dataset.
  2) Compare ELPs in Acidos with other bacteria (on phylum, family and mOTUs level; eggNOG PFAMS in OMDv2) – who are the super-interactors?
  3) Identify the Acidos with highest ELP abundance.
  4) Compare ELP abundance in terrestrial vs marine Acidos.
  5) Compare ELP abundance in host-associated (e.g. sponge or coral) vs free-living (marine) Acidos.
  7) Correlation of ELPs with BGCs?
  8) How often are ELPs encoded within or next to BGCs?
  9) Search for ELP effectors by profiling the neighboring genes (PFAMs, KEGG KOs, CAZys).

# Instructions

# General instructions

- We recommend to work in pairs but you can decide working by yourself
- Plan tasks and split responsibilities with your partner
  - Sharing is caring! Not only with your team colleague but also with everyone else :)
  - Keep your project clean and explanatory (more recommendations next Tuesday)
- To organize teams and preferences:
  - PROJECTS_SHARED_DRIVE
  - This also includes some useful folders:
    - A folder per project where you can store:
      - Preparation documents
      - Presentations updates (every Friday)
    - Examples: progress doc, presentation and report examples
    - Final presentation and reports folder
      - Please upload in each folder
      - Report has to be sent by email to Martin, Sam and Shini, more info in the future.