Manuscript outline:

*Topic 1) Exploring global river corridor OM chemistry through the ecological concept of core-satellite species*

*#------------------------------------------------- Please read -----------------------------------------------------#*

Please edit/add suggestions as pleased. Make sure to make edits in **“Suggestion Mode”**. The initial draft is adapted from the potential outline Christof has created using our comments [here](https://docs.google.com/document/d/1k0G3YBb3P1gGnnzCOOUqthC0N7X-BXrxzts2HRTQRe4/edit?usp=sharing). Refer to the same document for other notes and comments that were added during the first paragraph creation process.

#-----------------------------------------------------------------------------------------------------------------------#

# **Summary paragraph**

During the Summer of 2019, the Worldwide Hydrobiogeochemistry Observation Network for Dynamic River Systems (WHONDRS) conducted a crowdsourced sampling campaign and collected surface water and sediment data from 97 river locations. Here, we present the high-resolution dissolved organic matter (DOM) composition measured at ultra-high resolution (FT-ICR-MS) and discuss how riverine DOM assemblages – fingerprints of substrates and metabolic products - are organized across this wide range of riverine settings. To develop a transferable understanding of controls of DOM composition across sites and environmental conditions, we utilize the concept of core and satellite species, indicating coherences and similarities in substrates, dominant inputs or accumulated products. We discuss the patterns in terms of compound chemistry (groups, lability, molecular size) and watershed characteristics (e.g., size, stream order, flow, soil characteristics, geology) for both surface water and sediments. This study shows how community-led efforts can enhance our understanding of how metabolomes are structured across diverse riverine systems and identify underlying factors that influence riverine organic matter composition.

# **Introduction**

* Current challenges in DOM and FT-ICR-MS research
* Previous studies that applied ecological/microbial frameworks in DOM research
* Core and Satellite approach
  + Define Core and Satellite - utility of this distinction → Conceptual diagram of how C-S approach has been used in ecology for reader to better understand its applicability in FT research
  + The potential of core-satellite groupings in FT-ICR-MS data and their respective interpretations (macro/microbiological interpretations vs DOM interpretations) → reference to universal structures (Zark and Dittmar) and island of stability (IOS) as concepts that have shown why this distinction is interesting/important
  + Methodological variety/inconsistency of core-satellite thresholds
* This paper: Investigate the concept of core/satellite
  + Is there a clear distinction of satellite vs. core in riverine DOM?   
    → describe work done here
  + Are there clear differences in properties of satellite vs. core groups?   
    → describe work done here

## **Main objectives:**

* Investigate the occupancy patterns of DOM metabolites in water and sediments across river systems: identify core metabolites, those that are broadly distributed within systems, and satellite metabolites, those which have a sparser spatial distribution.
* Identify the molecular characteristics, and the environmental factors driving the diversity and distribution of core and satellite metabolites.

# **Methods**

* Short version of WHONDRS project introduction, sampling procedure and sampling design (ref: Garayburu-Caruso 2020)
* Data cleaning [this is something Topics 2-6 can then refer to, assuming we all do this consistently]:
  + Handling of replicates: up, mid, downstream of sediments = replicates and surface water triplicates (may be discussed in upcoming WHONDRS meeting organized by leadership team)
  + Rarity cutoff: Keep peaks in present at least 2 samples?
* Analyses (details in Results section):
  + Satellite-Core threshold comparison (see Section 1)
  + Multivariate analyses to explore functional differences in metabolites grouped as core-satellite between sediments and surface water (see Section 2)

**Fig.1:** Map of sample sites (unless this is in the overview paper for the special issue)

# **Results**

## **Section 1: Operational definition of core vs satellite**

### *Question*

* How do different core-satellite definitions affect the outcome of ‘species’ groupings with FT-ICR-MS data?

### *Paragraphs + analyses ideas*

* Different thresholds evaluated in this study:

What is the difference resulting from different approaches to separate core vs. satellite species? Examples of thresholding approaches include:

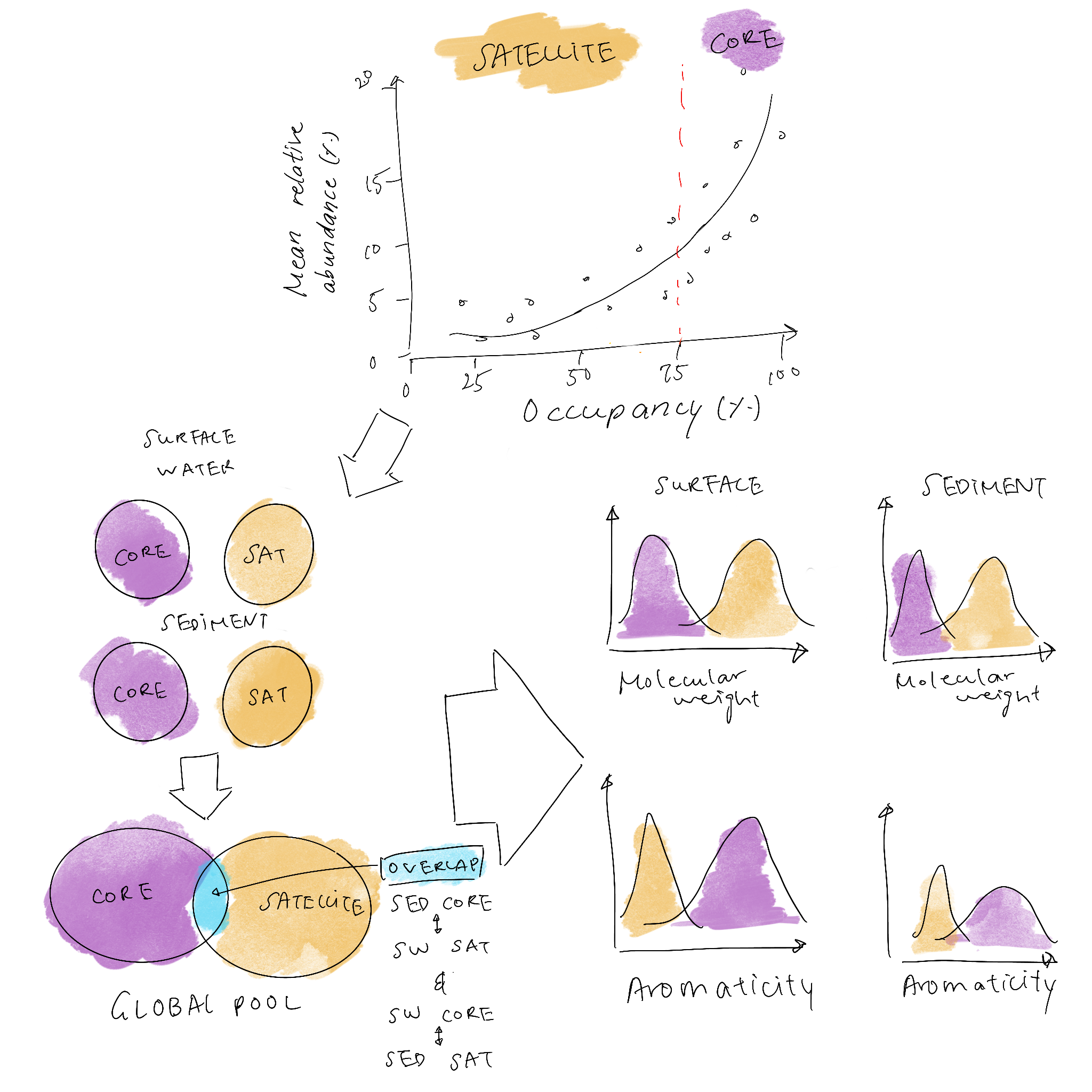
* + Literature threshold: Presence/absence persistence threshold (%), or universal presence vs any molecular formula (MF) that is absent somewhere
  + Emergent threshold: Let the data guide us.
    - exponential/log-normal curve and take 2nd derivative to find moment of maximum acceleration to find thresholds
    - Jenks Natural Breaks (clustering)
    - K-means clustering (clustering)
    - Machine learning? (Yuehan and Malcolm will check with AI colleagues)
  + (Maybe: Network approach)
* Are there pronounced differences between environments? The separation between core and satellite species may be quite different between water and sediment sample, and they may differ between broadly defined geographic regions.
  + Separate/Classify core and satellite for sediment and surface water separately
    - identify classification overlaps
    - How does the separation between core and satellite change/differ between sediment and water?
  + Investigate differences between geographic regions (see also Sengupta et al. In review). Do core/satellite compositions and the separation between them differ for
    - Global vs regional core?
    - US west vs east?
    - Distinction based on vegetation/biomes?

### *Figure ideas*

* Conceptual summary of thresholds applied in the study

## **Section 2: Emerging characteristics of core vs satellite MF**

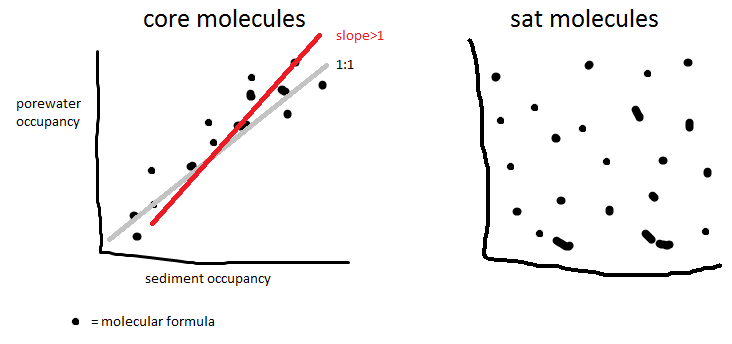
### *Questions*

* Are there statistically meaningful differences between core vs satellite species?
* How do sediment and surface water core-satellite species differ?
  + What are the characteristics of species core in the sediment/satellite in surface water and vice versa

### *Paragraphs + analysis ideas:*

* Alpha diversity (calculate richness for SED and SW, and compare between C-S thresholds)
* Unconstrained multivariate analysis: Beta diversity (do NMDS/PCoA and colour C-S/or polygon C-S, panels for each used threshold, compare similarity among thresholds with Procrustes analysis)
* Constrained multivariate analysis: PCA/NMDS or equivalent of core and satellite with respiration loadings, maybe even consider adding gibbs and lambda loadings as long as there is no overlap with other groups
* Distribution of compound groups: compound type (lipids, sugars, proteins …), molecular size, compound relative abundance, tendency towards more autochthonous vs allochthonous OM?
* (Reactivity: lability, dG, lambda, respiration rates (sediment only))

### *Figure ideas*



## **Section 3: Drivers of core vs satellite MF**

### *Questions*

* What drives/underlies the separation of different MF into core vs satellite?

### *Paragraphs*

* Scales (regions) - global, regional and even local.
* Hydrological/fluvial variables (stream order, catchment area, flow rate, base flow)
* Different proportions of allochthonous vs autochthonous inputs
* Catchment characteristics (soil type, vegetation type, land use)
* Near stream characteristics - light inputs, riparian characteristics (similar to hydrological/fluvial characteristics).
* Hydro-Climatic conditions (can we include climate change here?)
* Autochthonous vs. allochthonous distinction in core vs. satellite? -> Van Krevelen classification, lignin, tannins vs. amino sugars (relative abundance of “-like” compounds)
* Ecological /biogeographical differentiations - biomes, PFTs (plant functional types), etc.

### *Analysis + figure ideas:*

* Constrained multivariate analysis to parse out which drivers are significant
* ANCOVA of specific FT variables (i.e. NOSC) with environmental variables, Satellite-Core and SED-SW as different groupings (different slopes)?

# **Discussion**

* Would different methodologies lead to different results and interpretations?
* What factors determine consistently present molecules?
* Can we relate this to biogeochemical processes in sediments or the surface water?
* Potential use of findings on core/satellite characteristics in models

References

* Garayburu-Caruso VA, Danczak RE, Stegen JC, Renteria L, McCall M, Goldman AE, et al. Using community science to reveal the global chemogeography of river metabolomes. *Metabolites* 2020; **10**: 1–20.
* Sengupta A et al. Global Trait-Based Chemogeography of Organic Matter Thermodynamics. In review