Preregistration

My preregistration for a Registered Report

First Author[1], Ernst-August Doelle[1,2]

[1] Wilhelm-Wundt-University

[2] Konstanz Business School

16. September 2022

Main question	Biogeography of nearshore Polychaetes and Bivalves off the Pacific Coast of North America
Hypotheses	I expect their patterns to match the biogeographic breaks recovered by Fenberg 2015 and Blanchette 2008 (insert citations).
Data collection	Two sources of data will inform the analyses: GBIF and OBIS. Records were queried using rgbif and robis, defined by a polygon encompassing the North American Pacific Coast and a selection of invertebrates specified at the family-level.
Inclusion criteria	This project focuses on a subset of families from two classes of marine invertebrates: Bivalves and Polychaetes. Taxa were selected through analysis of a previous eDNA dataset from Calvert Island, BC. These data contained seagrass, kelp, and rocky bottom samples, amplified for COI and filtered to marine invertebrates. I identified generic and species-level

assignments that were unique to seagrass samples, and determined which taxa were most abundant in terms of unique species assignments and unique genetic sequences.

Polychaetes had the most species recovered, no matter how the data was grouped. Bivalves, specifically the species Kurtiella tumida, had the most unique genetic sequences in most samples, but especially in seagrass samples. However, as bivalves and polychaetes are very large classes with incredible levels of diversity, I restricted the taxonomic query to only families found in our environmental DNA dataset. Databases for assigning taxonomic identifications to genetic data are also notoriously incomplete for invertebrates, so selecting only the species identified in the eDNA samples filters to species that were in the database rather than a biologically meaningful subset. Increasing the query to the family level allows for closely related species and genera absent from the database to be recovered in big data occurrence records while still targeting the search towards the taxa that we found to be highly detectable with eDNA.

However, as the taxa were selected using a temperate dataset, I expect some of the analyses to contradict the expected latitudinal diversity gradient. As the search was run at the family, level this allows for subtropical species not present in our temperate dataset to be found, but it does not allow for subtropical and tropical families that may occupy similar functional groups in seagrass habitats but are absent in temperate latitudes.

As our first round of data from PECO will be fish, selecting invertebrates for this project will allow development of scripts without setting specific expectations for the data to come.

Exclusion criteria

Enter your response here.

Quality checks

GBIF records:

85,789 records in raw dataset. 38,423 records in cleaned dataset.

- Remove high levels of coordinate uncertainty
- Restrict data sources, removing fossils and machine observations.
- Retain only Presence data

- Remove individual count == 0
- Remove records without dates and prior to 1944: older records tend to be less reliable and there were many changes in biodiversity recording practices around this time, so WWII works as a decent cutoff.
- Remove records that were only a family-level ID
- Use taxize to update the taxonomy to the WoRMS backbone:
  - retain the species and genus level IDs, retain GBIF original taxonomy and ID for later use
  - get_wormsid_()
  - Use WoRMS aphiaID to pull higher taxonomy
  - Filter out any species that were identified by WoRMS as terrestrial or freshwater only
  - Filter out records with taxonomy not found or synonymized in WoRMS; retain list of rejected GBIF taxa as a 'problem list' to research in the literature later.

OBIS records:

25,914 records in raw dataset, 18,591 records in cleaned dataset

- Remove invalid dates
- Remove fossils and machine observations
- Remove records flagged as 'ON_LAND'
- Standardize taxonRank column
- Remove records that were only a family-level ID
- Rename columns to match GBIF data to merge

Combined dataset:

57.012 records in combined dataset. 44,038 records in cleaned dataset.

- Remove duplicate records: distinct by WoRMS scientific name, year, latitude and longitude. Seasonality and abundance are not important so one occurrence per species per year per coordinate pair is sufficient.

Confirmatory analyses	Maps, clustering.
Data type	Existing data, they are new extractions from OBIS and GBIF so I have never seen the data before.

References