

# Bacterial Genome Constraints Due To Environment

Terry Cho, Aidan Pavao

## Preliminary analysis

We have set up the following project Github repository ([Github: Bac Genome Constraint](#)).

To work with quality reference genome, we filtered the NCBI genome database for reference genomes with complete assembly level, year released since 2010, and annotations (both RefSeq and GenBank). We also filtered out genomes that are atypical, or from large multi-isolate projects, or that are metagenome-assembled genomes (MAGs). There were 5,988 reference genomes that passed this criteria (data/metadata/selected\_genomes.txt). The exact search criteria for reproducibility is found here: [Exact NCBI Search Criteria](#) (See Fig S1). We also successfully downloaded the full JGI GOLD (Genomes Online Database: <https://gold.jgi.doe.gov/downloads>) in the shared HMS o2 project directory.

Fig S1. Screenshot of NCBI search criteria

We conducted exploratory analyses to assess data quality and validate downstream modeling approaches for bacterial genome size constraints. The dataset was filtered to exclude obligate intracellular bacteria (e.g., Chlamydia, Rickettsia, Mycoplasma) and genomes smaller than 1.0 Mb, which represent genome decay rather than environmental constraints.

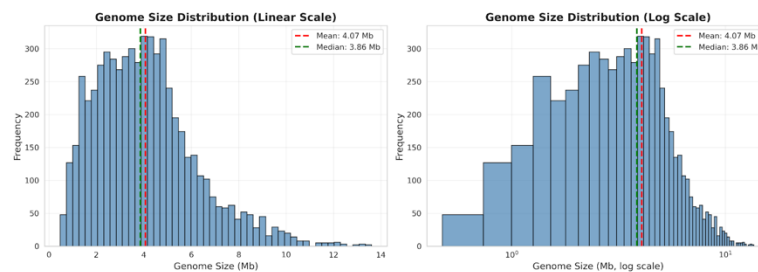


Fig 1. Screenshot of NCBI search criteria

We started out by plotting the genome size distributions using histograms on both linear and log scales to get a sense of what the data looks like (Fig 1). From Fig 1, we were able to visually inspect that genome size distribution is not gaussian nor random, hence we reasoned that there must be biological reason behind such bias around 4Mb and screwed towards smaller genome size (possibly due to favoring smaller genomes)

To obtain ecological context, we integrated data from the GOLD (Genomes Online Database) by matching genomes between NCBI and GOLD databases using taxonomy IDs. This mapping enabled comparisons across broad ecological categories. We were able to obtain 16 ecosystem categories with at least 20 genomes each, ranging from host-associated (mammals, insects, birds) to free-living (terrestrial, aquatic, wastewater) environments; we further reduced this number of ecosystems to top 10 categories based on the completeness of annotations (5 categories were also grouped into the “unclassified” category). (See figure 2).

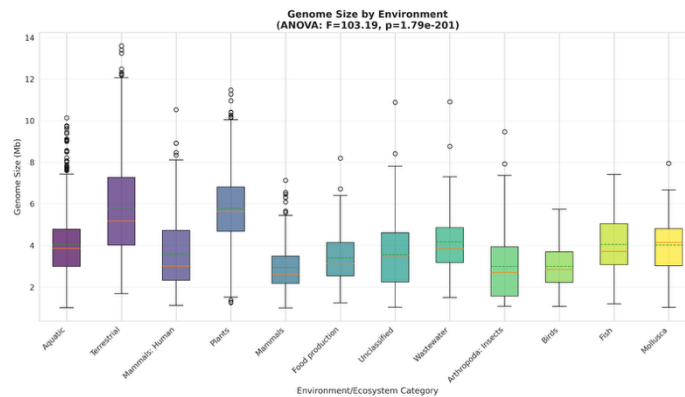


Fig 2. Genome size by environment.

Figure 2 shows significant environment effects (ANOVA  $p=1.79e-201$ ) on genome size variation. Box plots demonstrated clear environment-specific patterns in genome size distributions, supporting convergence of genome sizes within environments despite species diversity.

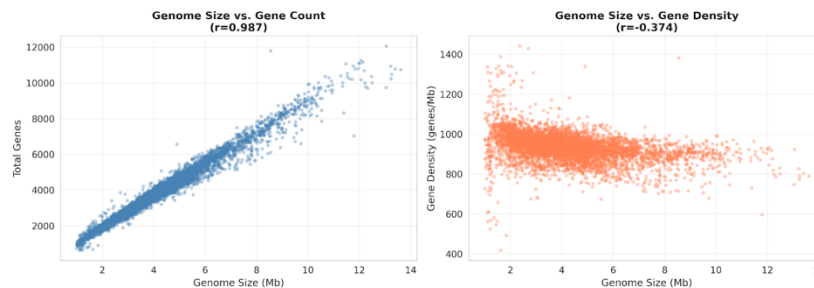


Fig 3. Genome size vs gene density

We wanted to do a sanity check on the overall GC / repeat content compared to total gene count, hence assessed Gene density (genes per Mb) and GC content relationship (Fig 3). Genome size showed strong positive correlation with total gene count, indicating that variation in genome size primarily reflects differences in coding capacity. were also examined.

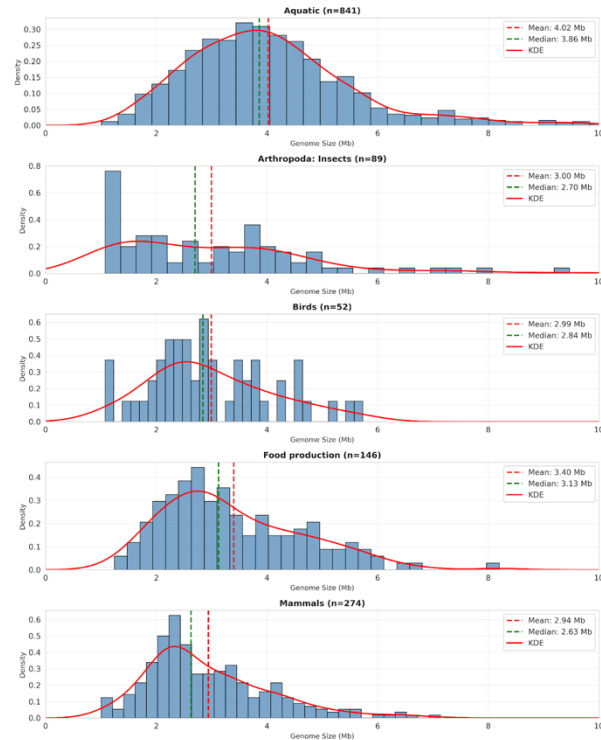


Fig 4. Genome size KDE of selected environment

For environments with at least 50 genomes, histograms with kernel density estimation were generated (Fig. 4) for us to find out whether there are any potential multimodal distributions.

## Planned analysis strategy

Our preliminary analyses shows that we have sufficient dataset robust and appropriately structured for downstream modeling, with clear environmental patterns in genome size that could be used to test hypothesis that bacterial genome size reflects ecological constraints.

Environments will be scored and selected for downstream analysis based on sample size, phylogenetic diversity (multiple phyla and genera), and convergence (coefficient of variation in genome size). This selection process would allow us to focus on environments suitable for detailed modeling, resulting in a curated dataset with complete metadata for downstream analyses.

We will construct phylogenetic trees for each environment and select 10 representative species that are spaced throughout each tree for downstream analyses. Prior to downstream analyses, environments will also be classified by environmental factors like nutrient availability, UV exposure, etc. Using the selected strains, the following analyses will be performed per environment:

1. Analyze how metabolic coding capacity associates with genome size and environment.
  - a. Compare average % coding DNA between environments:
    - i. ANOVA to determine whether the trait is different across environments
    - ii. T-tests between all environment pairs with multiple hypothesis correction to determine which environment pairs are driving the signal

- b. Compare number of metabolic genes, presence of different metabolic pathways, etc. across environments
    - i. ANOVA
    - ii. T-tests between all environment pairs
- 2. Analyze other genomic signatures of nutrient limitation and other stressors (eg. UV exposure) for evidence of association with genome size/environment
  - a. Amino acid usage
  - b. Enrichment analyses of metabolic categories across different environments
  - c. Run KEGG annotation (KofamScan) on protein sequences, map KOs to modules and pathways and create comprehensive feature matrix for downstream modeling
  - d. We will also investigate how transcription factors and mobile elements explains the patterns found in KEGG
- 3. Possible further directions:
  - a. Analyze metabolic pathway content (from KEGG) in metagenomes across environments

We also plan to come up with a good project name (Preferrably some french food name)