Imperial College London

Predicting the Identity and Abundance of Globally-Important Tyrosinase Genes in Bacteria Based on Metagenomics and Machine Learning

Mingji Zhang

CID: 02285741

August 2022

Supervisor: Dr. Samraat Pawar

Words count: 5793

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at Imperial College London
Submitted for the MSc in Computational Methods in Ecology and Evolution

Declaration

I declare that the original data for the publicly available metagenomic dataset comes from the National Council for Biotechnology (NCBI). All sample serial numbers and download links can be found at https://github.com/Mingji0613/MasterProject.git.

I declare that I am responsible for all data processing, analysis and visualisation.

I declare that all third party software and works have been appropriately cited.

My supervisor Samraat Pawar gave me detailed and patient guidance during the project.

Contents

Abstract

Microbial communities play a pivotal role in ecosystem functions, including carbon cycling and climate regulation. Targeting the Tyrosinase (TYR) gene, a key enzyme in phenolic compound metabolism, I delves into the identification of potential bacterial hosts within different aqueous environments. Leveraging metagenomic data and employing advanced machine learning techniques, a comprehensive analysis unveils intricate relationships between TYR gene abundance and microbial hosts across diverse ecosystems. My study's innovative approach encompasses multivariate linear regression and correlation network modeling, shedding light on the ecological and functional significance of TYR genes in distinct environments. Results highlight significant correlation between TYR gene abundance and specific genera such as Gemmiger, Escherichia, Collisella, Streptococcus, Akkermansia, Bifidobacterium, Faecalibacterium, and the yet-to-be-identified k__Bacteria—p__Firmicutes—c__CFGB10477—o__OFGB10477—f__FGB10477—g__GGB9345 bacteria. The Alpha and Beta diversity analyses further uncover compelling trends: certain environments display a prevalence of these TYR gene-associated genera, potentially indicating their ecological significance. Importantly, my research introduces a pioneering predictive model for TYR gene enrichment across diverse environments, exhibiting an accuracy surpassing 0.7. Remarkably, this model excels in the precise identification of strains at the Genus level, a feat unattainable through experimental isolates alone. These findings highlight the potential of machine learning in predicting TYR gene abundance and host biology, which can be targeted to help isolate TYR strains from the vast number of unidentified strains. A promising alternative to expensive molecular diagnostics is provided. Genus-based hierarchical predictions are superior to Family-based hierarchies due to the fact that metagenomic data are more abundant in data volume at the Genus level. While limitations such as dataset representativeness and uncalculated variables remain, my research still advances the understanding of TYR gene dynamics and their ecological impacts in diverse environments, contributing to the fields of environmental microbiology and carbon cycle research.

Keywords: Metagenomics, Tyrosinase genes, Microbial communities, Machine Learning, Abundance prediction, Environmental diversity

1 Introduction

Global warming, as a major challenge facing the world today, has aroused widespread concern and worry (Helbling and Meierrieks, 2023). Its impact not only covers the climate system, but also involves a wide range of fields such as socio-economics and ecology (Sodangi et al., 2011; Grace, 2004; Falkowski et al., 2000; Wani et al., 2023) Against this urgent background, the role of the global carbon cycle has become more and more important, as it carries a huge potential for curbing global warming (Falkowski et al., 2000). The global carbon cycle is a complex ecological process involving the exchange and cycling of carbon between the atmosphere, land and ocean (Grace, 2004; Falkowski et al., 2000). The biosphere converts carbon dioxide into organic matter through photosynthesis, storing it in plants, soil and oceans (Wani et al., 2023). This organic matter gradually releases carbon dioxide during decomposition and biological processes, creating a balanced carbon cycling system (Wani et al., 2023). The soil carbon pool is one of the five important carbon pools of the Earth, absorbing and storing large amounts of carbon dioxide and slowing down the accumulation of greenhouse gases in the atmosphere through its dynamic interaction with the biosphere carbon pool (Wani et al., 2023). Plants in ecosystems absorb large amounts of carbon dioxide through photosynthesis and fix it in their organisms, thus reducing the release of greenhouse gases (Falkowski et al., 2000). At the same time, microorganisms in the ecosystem are also involved in carbon decomposition and transformation, influencing the rate and direction of the carbon cycle (Hawkins et al., 2023; De Mandal et al., 2020). Soil respiration is the main pathway by which carbon dioxide fixed by land plants is returned to the atmosphere(Schlesinger and Andrews, 2000). Therefore, maintaining the integrity and diversity of ecosystems and protecting vegetation and soils are essential for the stabilization of the global carbon cycle (Wani et al., 2023).

Tyrosinases (TYRs) are important phenolic oxidases that are widely involved in the global carbon cycle and play a key role, especially in some ecosystems that exhibit high phenolic compounds (Panis and Rompel, 2022; Panis et al., 2021). TYRs are found in bacteria and eukaryotes and catalyze the oxidation of tyrosine and other phenolic compounds (Panis and Rompel, 2022; Hassan et al., 2023; Panis et al., 2021). TYRs are involved in the degradation and transformation of organic carbon in the environment, facilitating the decomposition and conversion of organic matter into more stable forms, thereby influencing carbon storage and release processes (De Mandal et al., 2020). At the same time TYRs catalyse oxidation reactions involving the conversion of phenolic compounds, which play a crucial role in the decomposition and humification of organic matter (De Mandal et al., 2020). Historically, persistent waterlogging has limited the activities of TYRs (Panis et al., 2021). Peatlands store 30% of global soil carbon, and prolonged summer droughts caused by global warming have facilitated the release of carbon stored as organic compounds from peatland carbon pools by TYRs (Panis et al., 2021). Understanding the role of TYRs is therefore critical to assessing the impact of carbon dioxide in the global carbon cycle (Panis et al., 2021).

However, although the role of TYRs in ecosystems has been recognized to some extent, there is still a knowledge gap in the identification of host environments as well as TYR-positive genera (Panis and Rompel, 2022). With the continuous development of advanced molecular biology and genome sequencing technologies, second-generation gene sequencing techniques provide for the search for potential TYR-positive genera (Slatko et al., 2018; Metzker, 2010). Metagenomic technology provides excellent conditions and

opportunities in identifying hosts of specific genes in the environment (Wooley et al., 2010; Sleator et al., 2008). This technology breaks through the limitations of traditional genetic studies and is no longer limited to studying the genome of a single organism, but is able to simultaneously analyze microbial populations throughout an ecosystem (Metzker, 2010; Wooley et al., 2010; Sleator et al., 2008). This provides new possibilities and depth for revealing the hosts of specific genes in the environment.In metagenomic studies, by collecting environmental samples and extracting the microbial DNA from them, the genetic information of all the microbial communities present in this environment can be obtained (Wooley et al., 2010). Subsequently, a large amount of DNA sequence data is acquired through high-throughput sequencing techniques (Daniel, 2005). These data contain gene sequences from a wide range of microorganisms, which may also cover target genes (Daniel, 2005). Subsequently, through a process of fine-grained data analysis and metagenome assembly, these DNA sequences can be spliced and annotated to determine their possible functions and origins (Daniel, 2005; Sleator et al., 2008). This makes it possible to identify hosts of specific genes in the environment. By comparing databases of known genes, it is possible to determine which groups of microorganisms carry the target genes and thus find their potential hosts (Metzker, 2010).

However, traditional methods for identifying gene-host associations often prove time-consuming and are constrained by prior knowledge of gene functionality (Moradigaravand et al., 2018; Sun et al., 2021). In order to overcome these challenges, recent advancements in machine learning technology have demonstrated the capability to predict specific bacterial hosts of genes even in the absence of prior mechanistic insights (Sun et al., 2021). Notably, a study focused on antibiotic resistance genes has indicated that machine learning models can accurately predict resistance patterns solely based on genomic and epidemiological data, even without detailed understanding of underlying biological mechanisms (Moradigaravand et al., 2018). Machine learning techniques offer an alternative approach for addressing the link between bacterial species and the TYR gene. Multivariate linear regression modeling is a machine learning algorithm that can predict the abundance of TYR genes based on the analysis of bacterial community abundance data obtained from metagenomic studies (Eberly, 2007). By modeling multivariate linear regression, potential TYR hosts in the environment can be identified and bacterial species enriched for TYR genes can be identified (Eberly, 2007).

The main goal of my research is to bridge the existing knowledge gap in identifying TYR-positive genera in different aqueous environments by applying a robust machine learning framework. I used multiple linear regression analysis to train and model, and I utilized 530 publicly available metagenomic datasets to obtain data on bacterial community abundance and relative gene abundance for the TYR gene. Calculated abundances were then meticulously validated on a comprehensive test dataset, a critical step in evaluating their predictive performance. My study predicted the abundance of TYR genes by employing multiple linear regression fitting and discriminated between bacterial species enriched in TYR genes, which indicate potential TYR hosts. In addition, the inclusion of bacterial taxonomic data in the multiple linear regression analysis helped to accurately predict TYR gene abundance in different sample environments. Predictions were performed based on two taxonomic levels, Family and Genus. To better highlight the great potential of utilizing machine learning techniques to effectively identify potential bacterial hosts responsible for TYR activities in the environment and to reveal the complex dynamics and importance of TYR in global carbon cycle and environmental processes by attempting to explore the intricate ecological associations between the TYR genes and

their microbial hosts, I specifically investigated three questions: (1) How well can machine learning techniques be utilized to predict the abundance and TYR-positive genera in host environments? (2) What characterizes the diversity of microbial hosts carrying the TYR gene in different environments? (3) What significant correlations exist between the gene abundance of the TYR gene in different environments and specific bacterial genera?

2 Materials and Methods

2.1 Bioinformatics analysis

A total of 530 sample sequences were randomly selected from the National Council for Biotechnology(NCBI) database, encompassing four primary sequencing environments, namely fecal, urban river, sewage, and large river, as well as two special sequencing environments, glacier lacustrine, and groundwater. The selection aimed to include as many bacterial species as possible, ensuring a comprehensive representation of the diverse microbial communities. Detailed information about all the sample sequence numbers can be found in Supplementary Table 1. The flowchart of the entire project is shown in Figure 1.

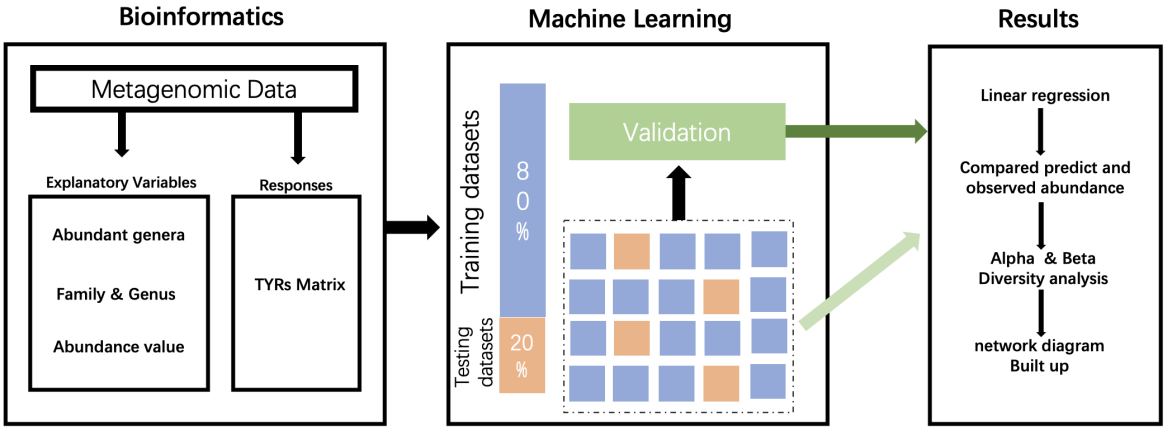


Figure 1: **Flowchart of the project including three periods.**

Prior to analysis, all samples underwent meticulous pre-processing using Metagenomic Phylogenetic Analysis (MetaPhlan) 4.0 (Segata et al., 2012). This preprocessing involved the trimming of macrogenomic sequence reads obtained from public databases, as well as the elimination of low-quality reads and splice sequences (Segata et al., 2012). The trimmed reads were then employed for strain extraction, facilitating the construction of the relative abundance matrix of the bacterial colonies.

Using a query term "tyrosinase," I employed a search of the NCBI database to target the TYR gene sequences, leading to the identification of 74,283 entries (as of May 2023). Subsequently, a reference short sequence library was curated, encompassing all the short sequence entries. The open-source programme DIAMOND (Dual Index Alignment of Next Generation Sequencing Data) was used to carry out comparisons between samples and reference libraries (Buchfink et al., 2015). Whenever a sequence in a sample displayed a precise overlap with a sequence in the database, extending over a span of 25 or

more amino acids, it was considered as a TYR sequence (Kristiansson et al., 2011). In addition, relative abundance values were uniformly normalised to parts per million (ppm) units for ease of comparison and analysis (Yang et al., 2013). This normalisation standardises abundance measurements across samples, allowing for meaningful comparisons of the presence and abundance of tyrosinase genes in different sequencing environments (Yang et al., 2013).

2.2 Development, validation and application of multiple linear regression model

A multivariate linear regression model was developed using the relative abundance matrix and the ST2 gene abundance matrix, which was obtained by rigorous data pre-processing as described in section 2.1. The multiple linear regression model is represented as follows (Eberly, 2007):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

where:

y : The dependent variable (the target variable to be predicted).

$x_1, x_2, \ldots, x_p$: The independent variables (features or predictors).

β_0 : The intercept term, representing the predicted value of y when all independent variables$(x_1, x_2, \ldots, x_p)$ are set to zero.

$\beta_1, \beta_2, \ldots, \beta_p$: The coefficients, indicating the impact of each independent variable on the predicted value of the dependent variable y.

ε : The error term (residual), accounting for the discrepancy between the predicted and actual values of the dependent variable y.
It captures the unexplained variation in the predictions.

The goal of multiple linear regression is to estimate the coefficients $(\beta_0, \beta_1, \ldots, \beta_p)$ that minimize the sum of squared errors, resulting in the best-fitted linear model for predicting y based on the given set of independent variables $(x_1, x_2, \ldots, x_p)$.

When multivariable linear regression is employed for predicting the abundance of the TYR gene within a sample, the equation can be expressed as follows::

$$\text{The prediction abundence } (A) = \sum_s S_i \cdot A_s$$

where:

A : represents the total TYR gene's abundance of the sample,

S_i : represents the relative bacterial abundance of each species in the sample,

A_s : represents the TYR abundance of each species.

The analysis was performed using the Scikit-learn(sklearn) package in R, which incorporates some basic steps such as data normalisation, missing value handling, data partitioning, preprocessing, model adjustment and variable significance assessment (Pedregosa

et al., 2011; Bisong and Bisong, 2019; Kramer and Kramer, 2016).

Linear regression analyses were conducted to determine whether there was a statistically significant relationship between the sample variables (Maulud and Abdulazeez, 2020). This analytical procedure is crucial as it ensures the robustness and interpretability of the predictive model (Bisong and Bisong, 2019). By employing linear regression, I aimed to explore potential dependencies and correlations between the variables, thereby laying a solid foundation for subsequent modelling and prediction tasks. The rigour of the linear regression analysis contributes to the reliability and validity of the findings, leading to meaningful interpretations and in-depth analyses of the potential relationships between the variables under investigation. In addition, 95% confidence intervals were incorporated to account for sample uncertainty and to provide a measure of confidence in the correlation (Ci and Rule, 1987). This correlation validation process is essential to determine the statistical significance of TYR gene abundance predictions and to gain insight into the relationships between other genes or environmental factors and TYR gene expression, thereby enhancing our understanding of TYR gene function.

The development of the model involved several steps. Firstly, I randomly divided the samples into an 80% training subset and a 20% validation subset using the sklearn package to ensure the balance of samples in both subsets (Kramer and Kramer, 2016). Next, we utilized the training function from the sklearn package to generate multiple linear regression models separately based on the Family and Genus dimensions (Bisong and Bisong, 2019). This approach allowed us to make predictions at both the Family and Genus taxonomic levels.

Predictions based on the Family level exhibit a higher degree of aggregation, encompassing a wider range of different bacterial species, thus capturing common features across different bacterial genera. This broad aggregation aids in identifying overall trends and patterns in the data, especially when the sample size is small or the number of bacterial species is large, resulting in more reliable results. Furthermore, due to the relatively high Family classification, the prediction results are more generalized, facilitating an understanding of the overall relevance of the bacterial community to TYR genes.

On the other hand, predictions based on the Genus level are more refined and detailed, enabling a more accurate differentiation of different bacterial genera. This fine-grained prediction approach is instrumental in capturing subtle differences and ecological associations between different bacterial genera. In cases of larger sample sizes or fewer bacterial species, Genus-level predictions provide more precise information, unveiling specific relationships between bacterial genera and TYR genes.

To ensure the model's reliability and generalizability, I recorded the intercepts and coefficients for each fitted model on the validation set. Subsequently, we ranked all sample regression coefficients and grouped them into high and low correlation categories based on the absolute value of the coefficients. Random cross-validation was conducted to ensure predictive accuracy.

In addition, as an integral part of the model evaluation process, I compared the validation set gene abundance obtained through model fitting to the observed gene abundance. This rigorous evaluation process allows for a quantitative comparison between the predicted gene abundance and the actual gene abundance in the validation set, thereby determining

the accuracy and reliability of the model predictions.

2.3 Statistical analysis and data visualization

In order to visualise the predictions, the Matplotlib package (Barrett et al., 2005) for python was used to present the results of the linear regression. The scatterplot visualises the correlation between the variables. A direct comparison between predicted and actual values is shown in a line graph. The line graph description provides a visual assessment of the accuracy of the predictions. Sample's phylum level histograms were plotted by Origin software (Edwards, 2002) and ggplot (Wickham, 2006) to show the number or abundance of bacterial species under different phylums. The distribution of the major phyla of bacteria can be visualised. The Alpha diversity and Beta diversity analysis plots used to assess the diversity of bacterial species and differences in species composition within the samples were plotted by ggplot in R.The Alpha diversity was calculated using Shannon Diversity formula. The formula was calculated as follow (Nolan and Callahan, 2006):

$$\text{Shannon Diversity Index } (H') = -\sum_{i=1}^{S} p_i \cdot \ln(p_i)$$

where:

H' : represents the Shannon diversity index,
S : represents the total number of species in the community,
p_i : represents the proportion or relative abundance of the i-th species in the community.

The Beta diversity was calculated using the Bray-Curtis algorithm. The formula was calculated as follows (Beals, 1984):

$$\text{Bray-Curtis distance (BC)} = \frac{\sum |a_i - b_i|}{\sum (a_i + b_i)}$$

where:

a_i : represents the abundance or frequency of the i-th species in sample A,
b_i : represents the abundance or frequency of the i-th species in sample B.

Key strains derived from differential expression calculations and Pearson correlation calculations were performed with TYR genes. The following formula was used (Cohen et al., 2009):

$$\text{Pearson correlation coefficient (r)} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

where:

$$X_i : \text{represents the i-th observation of variable X,}$$
$$Y_i : \text{represents the i-th observation of variable Y,}$$
$$\bar{X} : \text{represents the mean of all observations of variable X,}$$
$$\bar{Y} : \text{represents the mean of all observations of variable Y.}$$

To determine differentially expressed genes, a significance test is performed in the differential expression analysis to assess whether the observed differences are due to random factors. The p-value is obtained through the t-test. The calculation formula for the p-value is as follows (Ho et al., 2019):

$$t = \frac{\bar{D}}{s_D/\sqrt{n}}$$

where:

$$t : \text{t-statistic,}$$
$$\bar{D} : \text{mean of the differences between paired samples,}$$
$$s_D : \text{standard deviation of the differences between paired samples,}$$
$$n : \text{number of paired samples.}$$

A smaller p-value indicates that the observed differences are likely to be real rather than due to chance variation (Ho et al., 2019). The significance level is set to 0.05, meaning that when the p-value is less than 0.05, the differences are considered significant (Kim, 2015). If the p-value is smaller than the predefined significance level, the null hypothesis can be rejected, indicating that the differences are significant (Ho et al., 2019).

The Log2 Fold Change (Log2FC) is a measure used to quantify the degree of differential expression (Erhard, 2018). It represents the fold change in gene expression levels between two groups of samples. The formula for Log2FC was calculated as follows (Erhard, 2018):

$$\text{Log2FC} = \log_2 \left(\frac{\text{Expression in Treatment Group}}{\text{Expression in Control Group}} \right)$$

where:

$$\text{Log2FC} : \text{Log2 Fold Change value,}$$
$$\log_2 : \text{logarithm base 2,}$$
$$\frac{\text{Expression in Treatment Group}}{\text{Expression in Control Group}} : \text{ratio of gene expression in the treatment group to gene expression in the control group.}$$

A positive Log2FC indicates that the gene's expression level is higher in the first group of samples compared to the second group, while a negative Log2FC indicates that the gene's expression level is higher in the second group.The calculation was completed to get the correlation coefficient between them. The strains that were significantly correlated with the TYR gene were screened out and these strains were constructed into an association

network diagram using Gephi software (Bastian et al., 2009). The association network diagram visualises the strains significantly associated with the TYR gene and the association between them. This helps to understand the interactions between TYR genes and specific strains, and may reveal the potential role of TYR genes in regulating the abundance or function of these strains. In addition, the association network diagrams may also provide clues for further functional analyses and biological interpretations to help us better understand the associative relationships between TYR genes and gut microbial communities.

3 Result

A total of 530 metagenomic datasets were selected based on the defined criteria as outlined in Table S1. In May 2023, a comprehensive collection of 530 metagenomic datasets was procured from the NCBI database in the form of FASTQ files. Within these datasets, 125 datasets corresponded to fecal environments, 168 datasets to sewage environments, 176 datasets to urban river environments, 16 datasets to glacier lake environments, 9 datasets to groundwater environments, and 36 datasets to large river environments. The DNA read counts for individual samples spanned a range from 515,744 to 111,343,794, with a mean value of 26,972,889. Across this diverse compilation of datasets, taxonomic classification encompassed 10 phyla, 45 classes, 51 orders, 66 families, 153 genera, and 11,709 species.

From Figure 2, it can be observed that the TYR gene-containing microbial communities exhibit a diverse array of microbial phyla, and significant variations in relative abundances among these phyla are evident. Among the identified phyla, several show relatively higher relative abundances and count values, indicating their prominent presence and richness in the samples. Specifically, Proteobacteria, Bacteroidetes, and Firmicutes emerge as the most abundant phyla, with relative abundances of 530, 523, and 502, respectively, and correspondingly high count values, further emphasizing their significance within the TYR gene-containing microbial communities. These phyla are known to be widely distributed in soil and other sample sources, suggesting their potential roles in various environmental contexts.

Moreover, Figure 2 highlights certain microbial phyla that exhibit higher relative abundances and count values in specific samples, such as Actinobacteria, Synergistetes, and Tenericutes. This suggests that these phyla might be sensitive to specific ecological conditions or possess competitive advantages in these particular environments, leading to their relatively higher abundance in the corresponding samples.

In contrast, some phyla, including Candidatus Melainabacteria, Verrucomicrobia, and Lentisphaerae, display relatively lower abundances and count values in certain samples. This indicates that these phyla might not be as common in TYR gene-containing microbial communities or exhibit lower relative abundances in these specific environments. This could be attributed to their lesser adaptability to specific environmental conditions or potential competition from other microbial communities present in these samples.

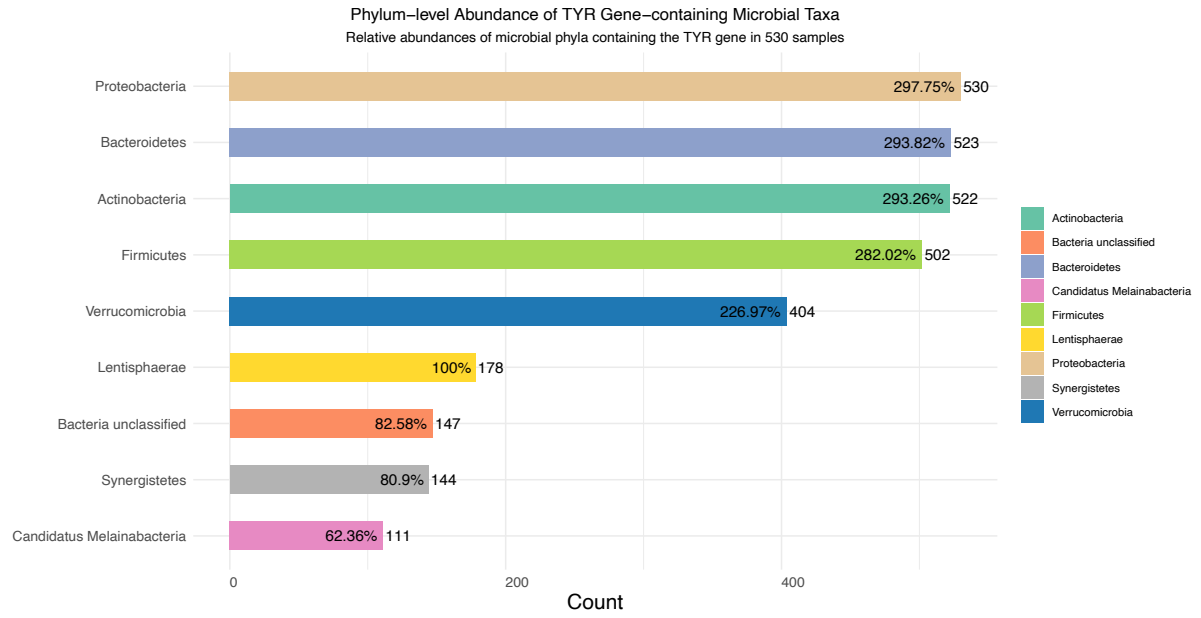


Figure 2: **Phylum-level Bar Chart. Phylum-level bar chart depicting the relative abundances of microbial taxa containing the target gene TYR in 530 samples. Horizontal coordinates (X-axis): the count value of each microbial phylum in the sample or community. Vertical coordinate (Y-axis): different microbial phyla (Phylum). The number of clades was calculated as a percentage relative to the abundance of Lentisphaerae clades. Proteobacteria had the highest TYR enrichment. Candidatus Melainabacteria showed the lowest TYR enrichment.**

The analysis provides insights into the diversity and relative abundance distribution of microbial phyla within TYR gene-containing microbial communities.Taking the phylum Lentisphaerae as the reference, percentage calculations were conducted for the remaining phyla. Four phyla, surpassing Lentisphaerae in abundance, demonstrated significant disparities at a categorical level. These results contribute valuable information for understanding the ecological characteristics and potential functional roles of microbial communities associated with the TYR gene, and offer valuable data for further investigations into microbial biodiversity and environmental microbiology.

Correlation analyses performed to investigate the relationship between bacterial population abundance and TYR gene abundance in the samples showed that scatter plots and best-fit straight lines showed a clear linear trend. This analysis is depicted in Figures 3 and 4, showing scatter plots and corresponding best-fit straight lines at the family and genus levels, respectively.

The slope of the best-fit straight line in the figure indicates the strength and direction of the linear association between the relative abundance of bacterial populations and TYR gene abundance. A positive slope suggests a positive correlation, indicating that as the relative abundance of bacterial populations increases, the TYR gene abundance also tends to increase. Conversely, a negative slope would indicate a negative correlation, implying that as the relative abundance of bacterial populations increases, the TYR gene abundance tends to decrease.

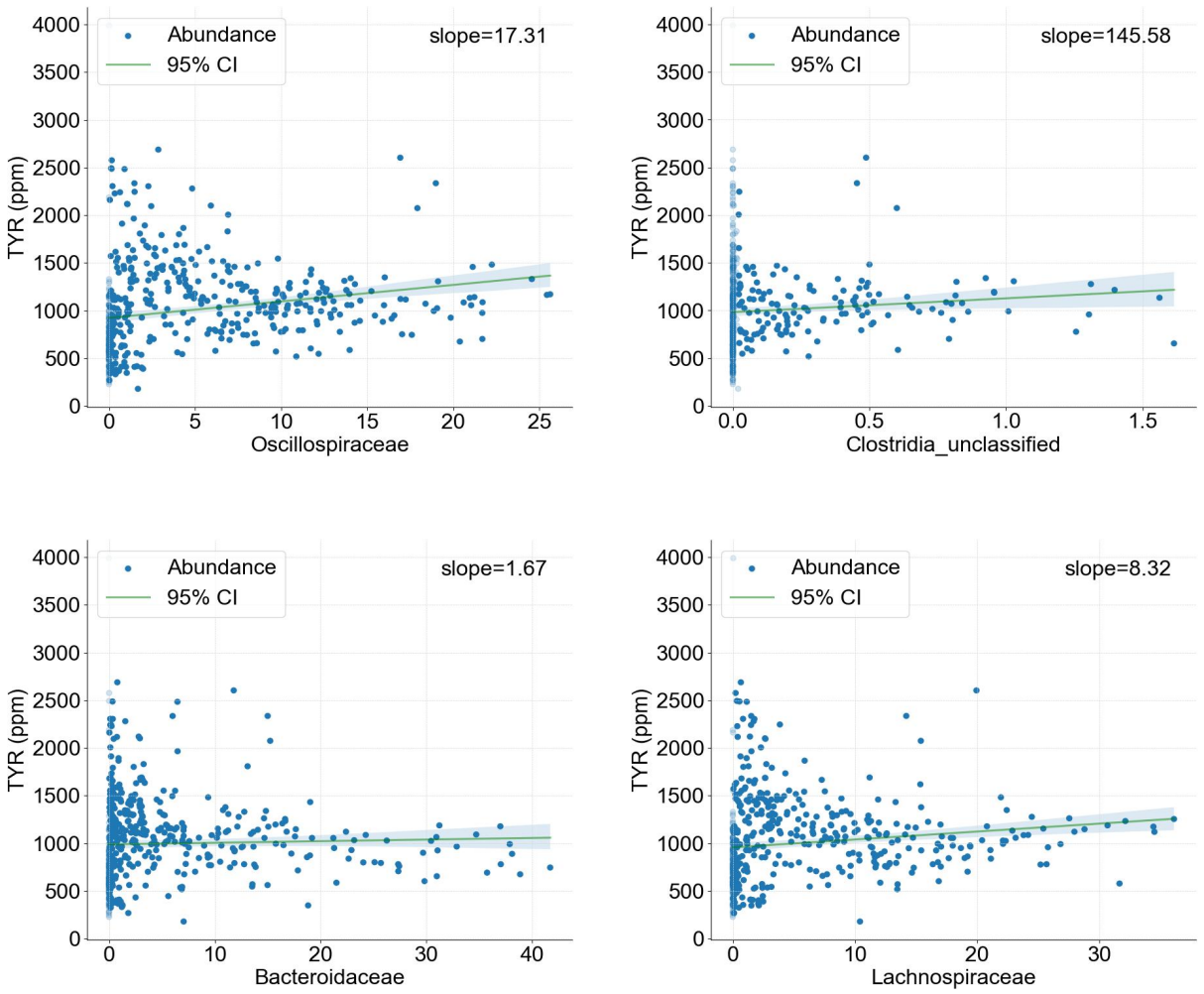


Figure 3: **Linear regression relative population abundance and gene abundance based on Family. Only the 4 positive correlation scatter plots from the test data are shown.(Oscillospiraceae & Clostridia_unclassified & Bacteroidaceae & Bacteroidaceae)**

The presence of confidence bands in the figure is of great significance as well. These bands provide an estimation of the uncertainty associated with the predicted values and establish confidence intervals for the predicted values. The width of the confidence bands illustrates the range within which the predicted values are expected to fall with a certain level of confidence. This information aids in assessing the reliability and precision of the predictions made based on the correlation analysis. These plots provide a detailed visual representation of the correlations between bacterial population abundance and TYR gene abundance for various strains, supplementing the analysis conducted. The correlation coefficients of all fitted straight lines were recorded and ranked, and samples with slopes smaller or equal to less than 0 were specifically flagged and the weights were reduced by adjusting the parameters accordingly in the subsequent model building phase. Enrichment of samples with relative abundance less than 0.02 was observed in all scatter plots, with clustering occurring at a transparency of 80%.

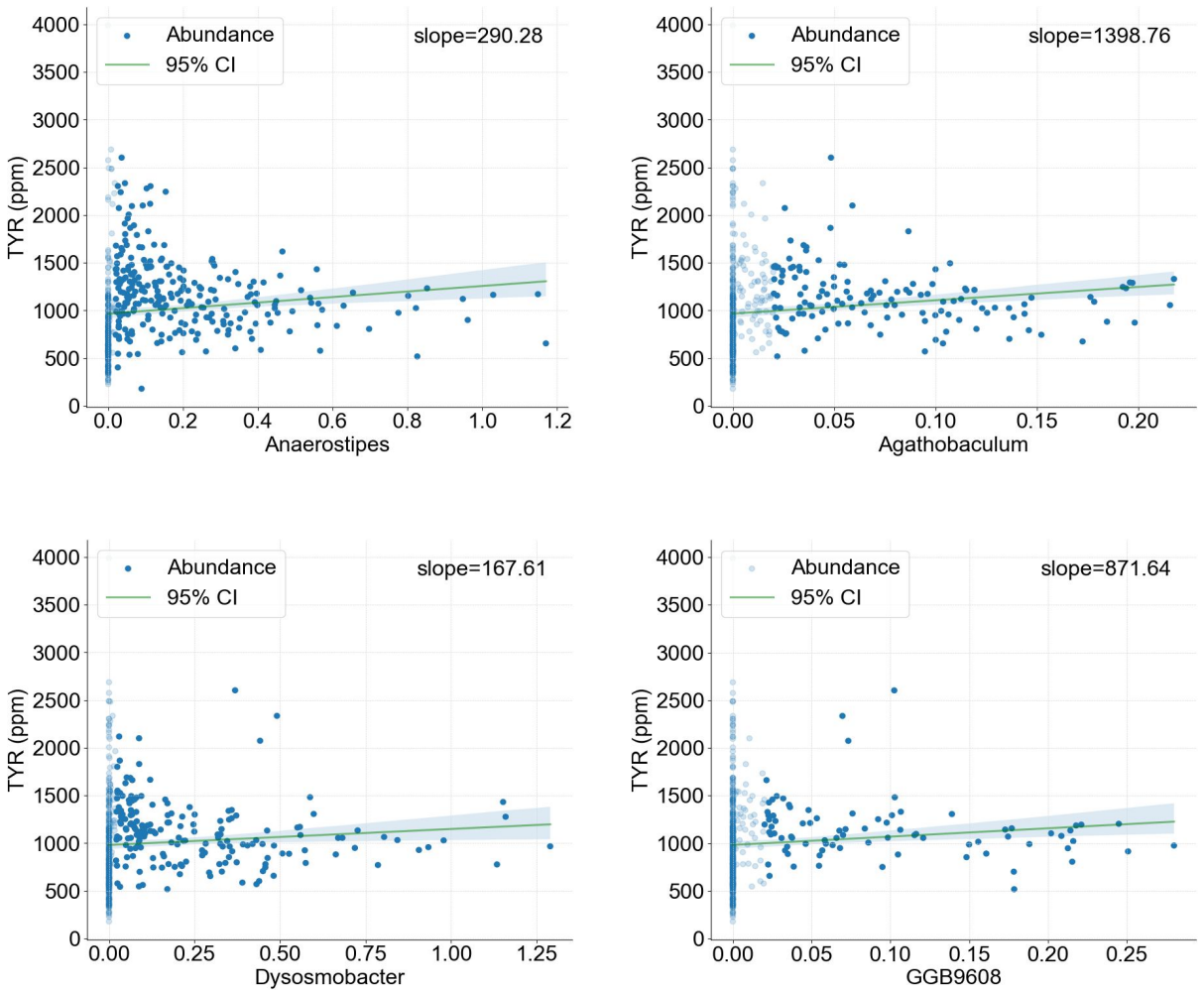


Figure 4: **Linear regression relative population abundance and gene abundance based on Genus. The horizontal coordinate is the relative abundance of the population in the sample. The vertical coordinate is the abundance of the TYR gene in the sample.Transparency of points with relative abundance less than 0.02 will be increased to 80%.Only the 4 positive correlation scatter plots from the test data are shown.(Anaerostipes & Agathobaculum & Dysosmobacter & GGB9608)**

Figures 5&6 show the results of model predictions based on both Family and Genus classes. The horizontal axis represents the test number ID of the random test subset, while the vertical axis represents the value of TYR gene abundance in the sample (PPM). The blue line corresponds to the actual observed values of the test dataset, while the red line represents the predicted values generated by the model for the same test dataset.

By visually comparing the blue and red lines, we can assess how well the model fits the test data. When the red line is closely aligned with the blue line, it indicates that the model's predictions are very close to the actual observations, suggesting that the model fits the test data well. Conversely, if there is a significant deviation between the red and blue lines, it indicates that the model's predictions deviate from the actual observations, suggesting that the model may not adequately account for changes in the test data.

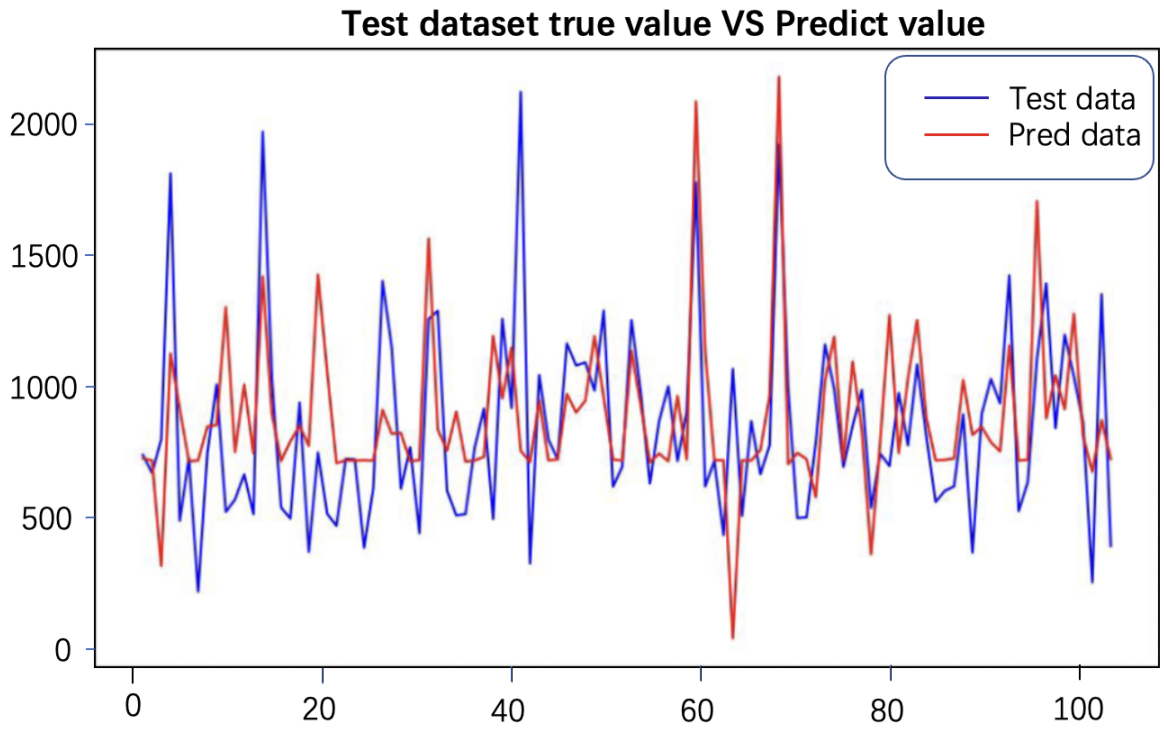


Figure 5: **Comparison of predicted abundance and observed values based on family level. The horizontal coordinate is the test set ID and the vertical coordinate is the predicted and actual value of TYR gene abundance.**

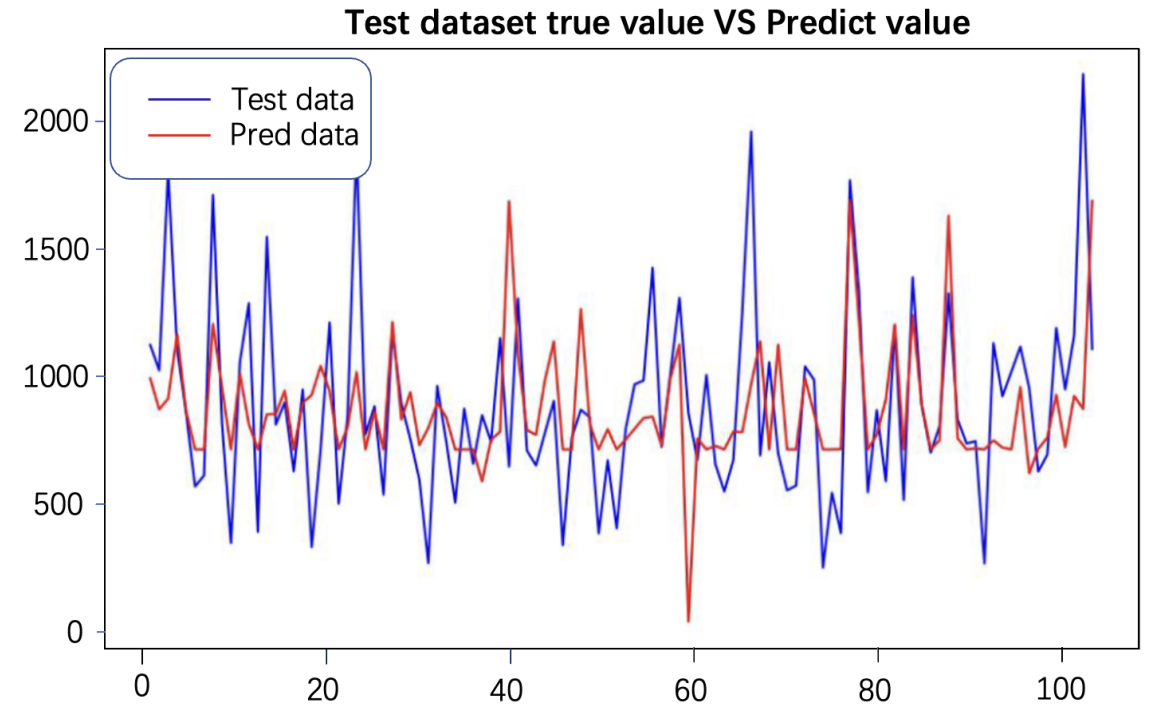


Figure 6: **Comparison of predicted abundance and observed values based on genus level. The horizontal coordinate is the test set ID and the vertical coordinate is the predicted and actual value of TYR gene abundance.**

In Figures 5 and 6, the red and blue lines for both levels (family and genus) show a significant overlap, indicating that the predictive model is capable of capturing the underlying relationship between bacterial population distribution and TYR gene abundance with a

considerable degree of accuracy. The majority of the residuals fall within the range of plus or minus 500, suggesting that the predicted values closely align with the observed values for most of the samples. The average prediction error being at 25% for an average abundance of 2000 ppm further proves the model's ability to accurately predict TYR gene abundance based on the bacterial population distribution.

The model's good fit to the test data indicates that the predictions are reliable and accurate for a substantial portion of the dataset. However, there are some interesting observations to be made from the data. The red line, which represents the predictions at the family level, exhibits a tendency to level off compared to the blue line (genus level). This indicates that there may be certain factors or complexities that the model fails to account for, particularly in extreme low peak sections.

The influence of the relative abundance of strains and the degree of TYR gene enrichment on model fitting becomes evident. Samples with higher relative abundance of strains show a relatively higher predictive fit, whereas lower relative abundance results in predicted data being biased high due to parameter corrections. This suggests that the model's performance may be influenced by the specific composition of bacterial strains in the test samples.

Moreover, the presence of outliers in the predicted data for certain test samples suggests the existence of unexplained sources of variation or factors not adequately addressed by the predictive model. These biases could arise from unmeasured variables, measurement errors, or other unknown factors that impact the relationship between bacterial population distribution and TYR gene abundance.

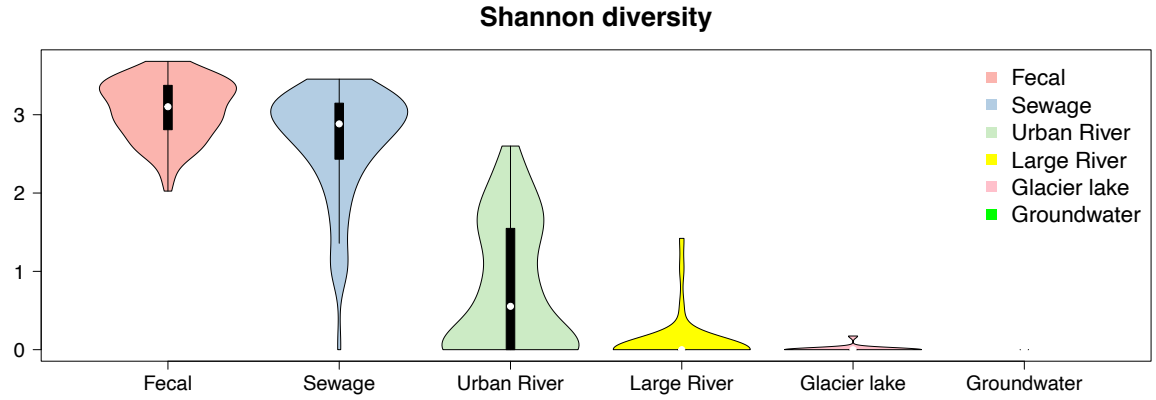


Figure 7: **Alpha Diversity Plot. The horizontal coordinate is the sample category. The vertical coordinate is the shannon diversity value.**

Figure 7 presents an alpha diversity plot based on the Shannon diversity index, revealing intriguing patterns of microbial diversity among bacterial hosts carrying the TYR gene across different environmental categories. The vertical axis represents the Shannon diversity value, reflecting the richness and evenness of species in each sample.

Notably, the Shannon diversity values are relatively high in feces and wastewater environments, approaching 3, indicating rich and evenly distributed microbial communities within TYR gene-carrying hosts. Conversely, the Shannon diversity value in urban river

environments is moderate, around 0.5, suggesting lower species richness and uneven distribution of microbial taxa in hosts carrying the TYR gene. Meanwhile, the Shannon diversity values for the other three environmental categories are relatively low, around 0, indicating limited species diversity and uneven distribution of microbial taxa within TYR gene-carrying hosts.

Furthermore, the alpha diversity plots for each category exhibit non-square shapes and uneven distributions. Feces and wastewater environments are characterized by an inverted triangular pattern, highlighting the diversity and evenness of microbial communities within TYR gene-carrying hosts. In contrast, the plots for other environments show an upward triangular pattern, indicating lower species richness and less uniform distribution of microbial taxa in hosts carrying the TYR gene. This suggests significant differences in microbial community structure among different environments, and TYR gene-carrying microbial hosts in each environmental category possess unique compositions and ecological characteristics.

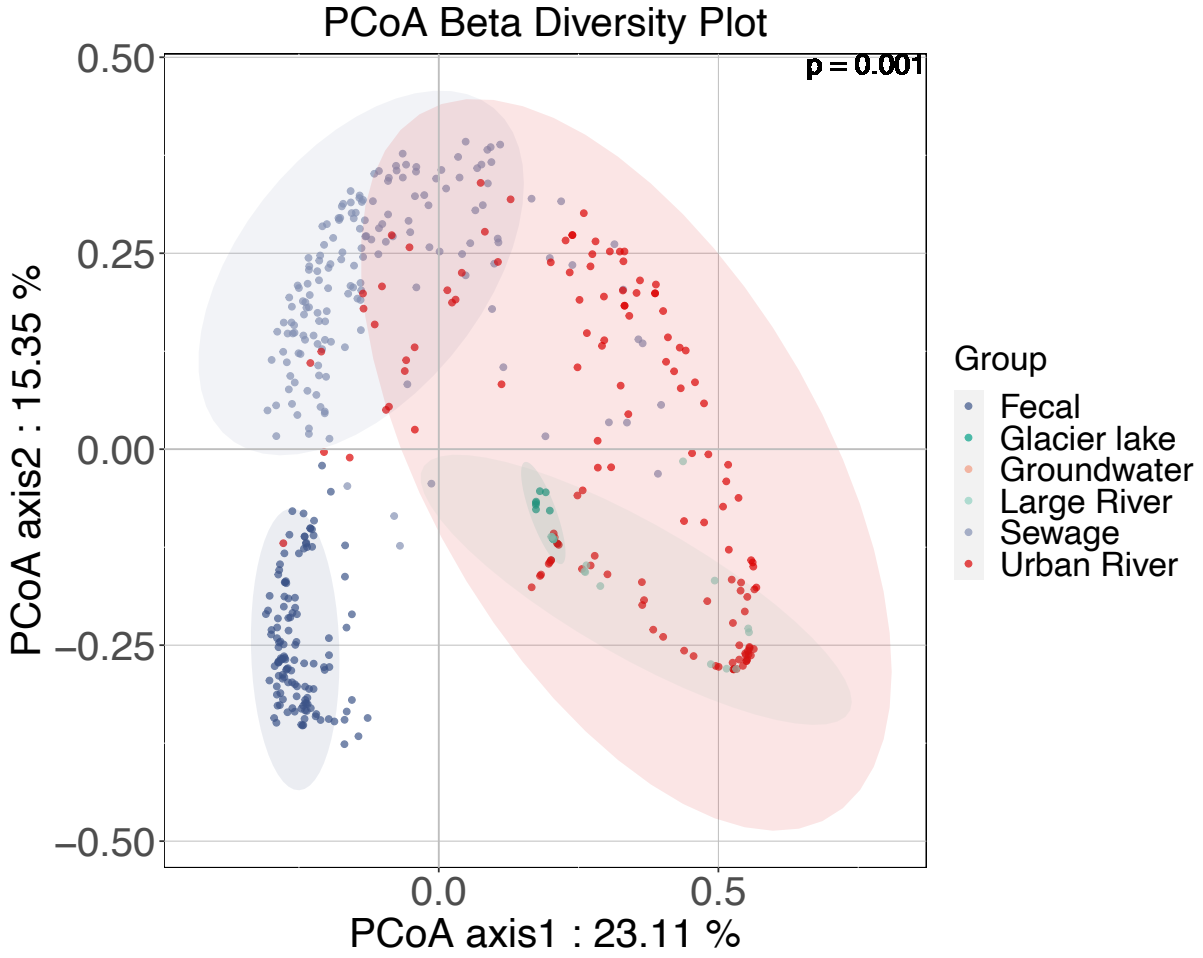


Figure 8: **Beta Diversity Plot. The horizontal and vertical coordinates represent the coordinate values of the samples on the different principal axes.**

Figure 8 Beta diversity plots based on Bray-Curtis distances and visualised by principal coordinate analysis (PCoA) effectively reveal the overlapping patterns of TYR bacterial hosts in different environments, providing strong evidence of similarities and differences in bacterial community structure between samples. The ellipses in the plots represent confidence intervals that visualise the variability and dispersion of the sample groups in

PCoA space. Samples within the same ellipse have similar bacterial community structures, while samples from different ellipses have more pronounced differences. Overlap of ellipses indicates partial similarity in bacterial composition, whereas non-overlap indicates significant differences between groups of samples.

Samples were significantly separated along the first principal axis (p=0.001), indicating significant overall differences in microbial composition across environments. This finding suggests that unique environmental conditions may create unique bacterial communities associated with TYR genes.

Notably, the figure shows that glacial lake, groundwater, and large river samples were completely included in the urban river clustering. This observation suggests a high degree of similarity in TYR bacterial hosts in these environments.

In addition, there was about 40% overlap between the wastewater samples and the urban river samples, suggesting some similarity in their bacterial community structure. This commonality may be due to common anthropogenic influences or similar ecological factors. In contrast, fecal samples were significantly different from other environmental samples. This apparent difference highlights the unique microbial characteristics associated with faecal environments, which may be influenced by host-specific factors. In addition, due to the limited number of groundwater samples, only coordinate points were plotted and elliptical intervals were not calculated to ensure the accuracy of the analyses.

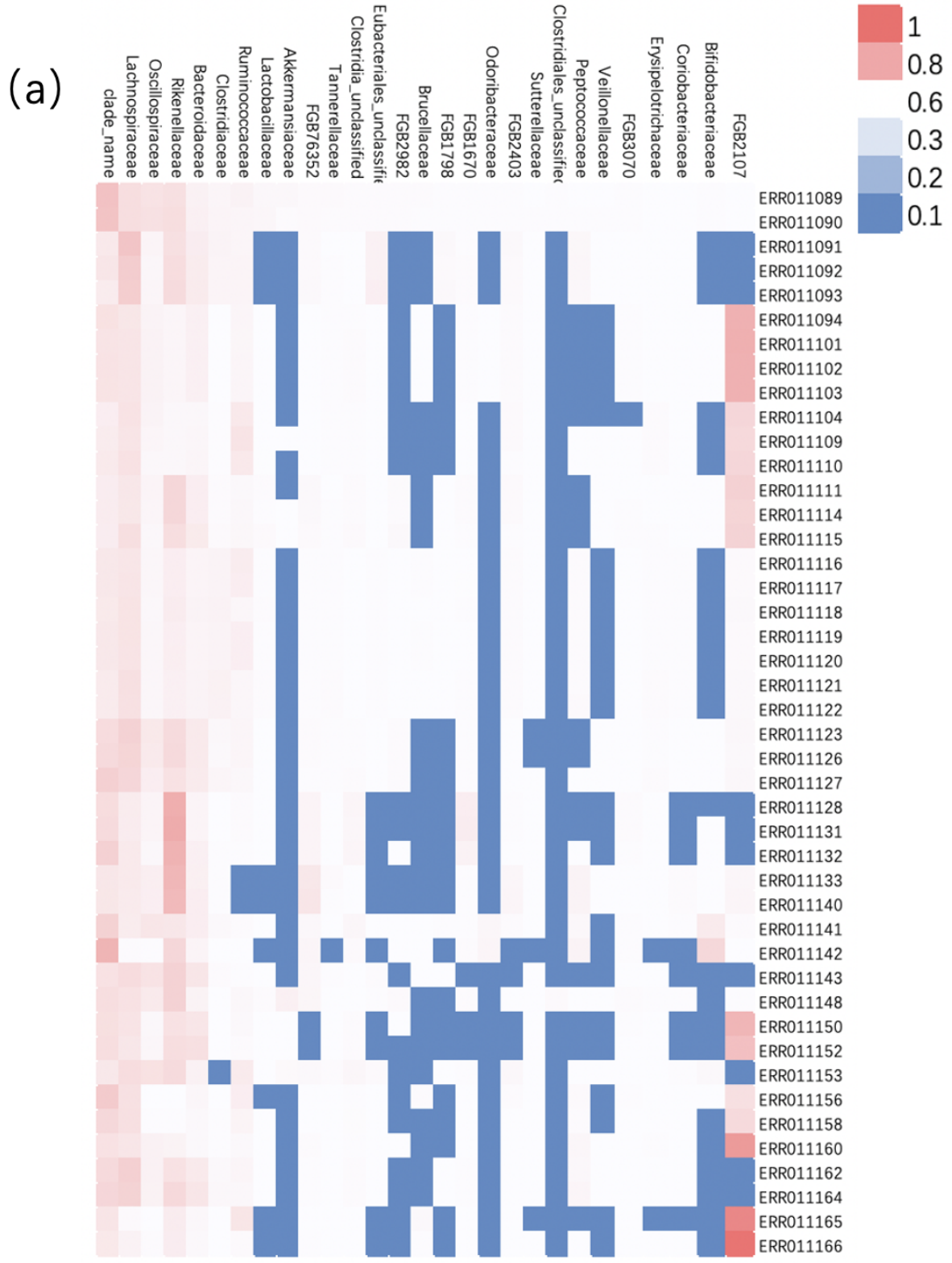
The overlapping patterns of TYR bacterial hosts in different environments were effectively demonstrated by PCoA's beta diversity plot based on Bray-Curtis distances, revealing similarities and differences in bacterial community structure. This analysis helps to understand how environmental conditions affect the microbial communities associated with TYR genes and provides important clues to the ecological significance of TYR gene-carrying bacterial hosts in different environments.

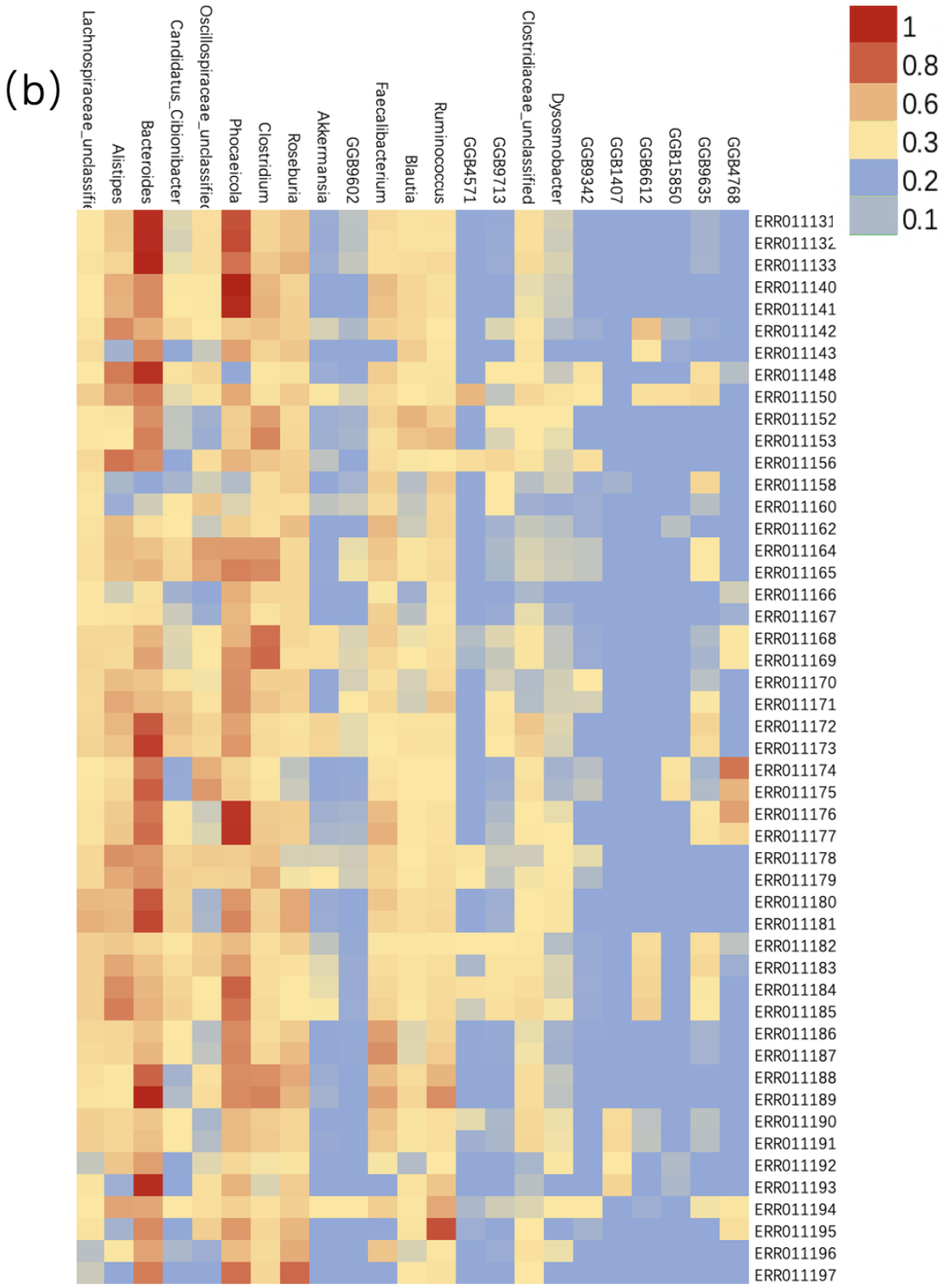
I conducted differential expression analysis to identify key predicted bacterial genera and performed Pearson correlation analysis with the TYR gene. The results of the correlation analysis are presented in Figure 9, displaying the strength of correlations between TYR gene abundance and bacterial taxa in six different sample environments. Among the 153 genera analyzed, 113 genera with significant correlations (correlation coefficient >0.2) were included in the visualization.

In Figure 9, each bacterial genus is represented as a node. The size of the node indicates the strength of the correlation between the genus and the TYR gene, with larger nodes indicating stronger correlations and smaller nodes indicating weaker correlations. The thickness of the lines connecting the nodes to the center represents the strength of the correlation, with thicker lines indicating stronger correlations and thinner lines indicating weaker correlations. The color of the lines is determined by the ranking of the correlation strength. The top 1.77% of correlations are displayed in purple, representing two genera, Gemmiger and Escherichia, which have the strongest correlations with the TYR gene. The next 1.77-5.31% of correlations are shown in red, encompassing six genera: Collisella, Streptococcus, Akkermansia, Bifidobacterium, Faecalibacterium, and an unidentified genus classified as k__Bacteria—p__Firmicutes—c__CFGB10477—o__OFGB10477—f__FGB10477—g__GGB9345. Fourteen genera with correlations ranging from 5.31% to 12.39% are represented by orange lines. The remaining genera with weaker correlations are shown with

light orange lines at 40% transparency.

These significant correlations indicate that these bacterial taxa may play crucial roles in the regulation and function of the TYR gene in different sample environments. The visualization of these correlations in Figure 9 provides valuable insights into the potential interactions between microbial communities and the TYR gene, contributing to a better understanding of the ecological and functional implications of the TYR gene in diverse environments.




(b)
Lachnospiraceae_unclassifie
Alistipes
Bacteroides
Candidatus_Cibionibacter
Oscillospiraceae_unclassifiec
Phocaeicola
Clostridium
Roseburia
Akkermansia
GGB9602
Faecalibacterium
Blautia
Ruminococcus
GGB4571
GGB9713
Clostridiaceae_unclassified
Dysosmobacter
GGB9342
GGB1407
GGB6612
GGB15850
GGB9635
GGB4768
ERR011131
ERR011132
ERR011133
ERR011140
ERR011141
ERR011142
ERR011143
ERR011148
ERR011150
ERR011152
ERR011153
ERR011156
ERR011158
ERR011160
ERR011162
ERR011164
ERR011165
ERR011166
ERR011167
ERR011168
ERR011169
ERR011170
ERR011171
ERR011172
ERR011173
ERR011174
ERR011175
ERR011176
ERR011177
ERR011178
ERR011179
ERR011180
ERR011181
ERR011182
ERR011183
ERR011184
ERR011185
ERR011186
ERR011187
ERR011188
ERR011189
ERR011190
ERR011191
ERR011192
ERR011193
ERR011194
ERR011195
ERR011196
ERR011197
1
0.8
0.6
0.3
0.2
0.1

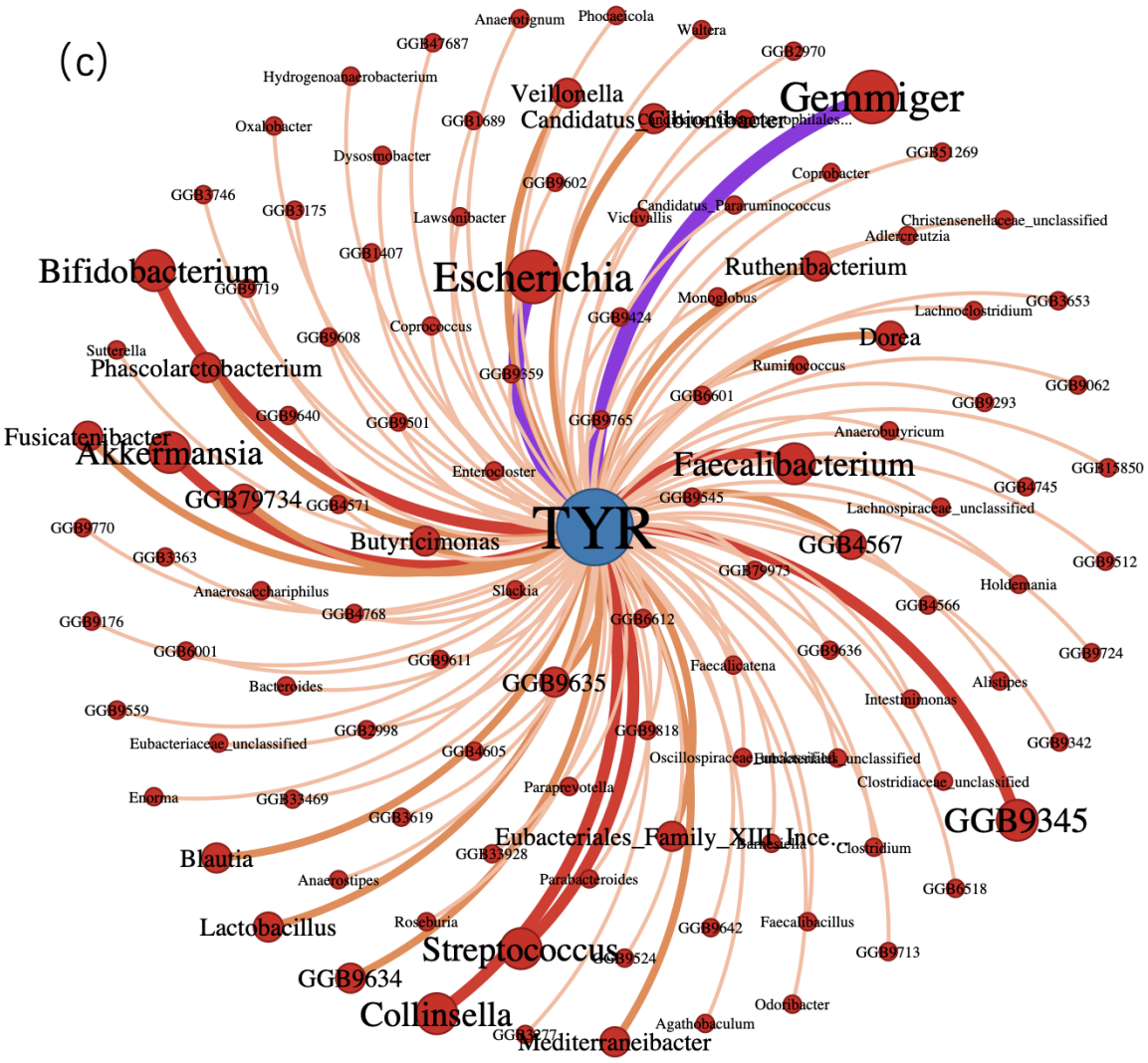


Figure 9: **(a) Schematic of TYR gene heat based on FAMILY level. The first 44 numbered IDs and 28 strains are shown. (b) Schematic of TYR gene heat based on genus level. The first 49 numbered IDs and 23 strains are shown. (c)Network plot of association of significantly related species with TYR genes. The thickness of the connecting line between the node and the centre and the size of the node represent the strength of the correlation. The colour shade of the connecting line indicates the order of significance of the correlation between the node and the central TYR gene.**

4 Discussion

In this study, I aimed to bridge the knowledge gap in identifying bacterial hosts of the TYR gene in different environments by applying a robust machine learning framework. I utilized multivariate linear regression modeling to predict the abundance of TYR genes based on the analysis of bacterial community abundance data obtained from Metagenomic studies. Our findings demonstrate the great potential of using machine learning techniques to effectively identify potential bacterial hosts responsible for TYR activity in the environment.

My findings demonstrate that (1) the abundance and hosts of the TYR gene in microbial

communities can be predicted using machine learning techniques and to a certain extent can be useful for subsequent targeted identification of incompletely recognised strains (2) the abundance of the TYR gene in microbial hosts in different environments is comparable to that of Gemmiger, Escherichia, Collisella, Streptococcus, Akkermansia, Bifidobacterium, Faecalibacterium and the not fully recognised k__Bacteria—p__Firmicutes—c__CFGB10477—o__OFGB10477—f__FGB10477—g__GGB9345 bacteria are extremely correlated (3) The diversity of microbial hosts carrying TYR genes in different environments reflects the variability. Fecal and wastewater environments had abundant and evenly distributed host communities. Urban rivers were not particularly characterised in terms of colony numbers and diversity. Groundwater and glacial lake environments still need further study.

I successfully developed a multivariate linear regression model using the relative abundance matrix of bacterial colonies and the gene abundance matrix to predict TYR gene abundance. The model's good fit to the test data indicates that the predictions are reliable and accurate for a substantial portion of the dataset. However, I observed some outliers in the predicted data, suggesting the existence of unexplained sources of variation or factors not adequately addressed by the predictive model. Further research is needed to identify and account for these factors to improve the model's performance.Even so, my current research continues to demonstrate that it is feasible to use machine learning techniques to help predict the abundance and hosts of TYR genes in microbial communities. To a certain extent, expensive upfront full molecular diagnostic identifications can be circumvented (Váradi et al., 2017), opening up new possibilities for subsequent purification and identification of bacteria.

The correlation analysis between TYR gene abundance and bacterial taxa provided valuable insights into the potential interactions between microbial communities and the TYR gene in different environments. Among the 153 genera analyzed, 113 genera showed significant correlations with the TYR gene, indicating their potential roles in the regulation and function of the TYR gene. This information helps to better understand the ecological and functional implications of the TYR gene in diverse environments.

The alpha diversity and beta diversity analysis revealed intriguing patterns of microbial diversity among bacterial hosts carrying the TYR gene across different environmental categories (Socolar et al., 2016). Feces and wastewater environments exhibited high Shannon diversity values, indicating rich and evenly distributed microbial communities within TYR gene-carrying hosts (Thukral, 2017). On the other hand, urban river environments showed moderate Shannon diversity values, suggesting lower species richness and uneven distribution of microbial taxa in hosts carrying the TYR gene. These findings emphasize the differences in microbial community structure among different environments and the unique compositions and ecological characteristics of TYR gene-carrying microbial hosts in each category.

The association network diagram constructed using significant correlations between TYR gene abundance and bacterial taxa provided a visual representation of the relationships between TYR genes and specific strains. The strength and direction of the correlations were represented by the size and thickness of the nodes and connecting lines, respectively. This network diagram helps to identify key predicted bacterial genera that are significantly associated with the TYR gene and contributes to a better understanding of the potential microbial hosts responsible for TYR activity.In particular, the incompletely identified

bacterium k__Bacteria—p__Firmicutes—c__CFGB10477—o__OFGB10477—f__FGB10477—g__GGB9345 was successfully identified. This suggests that the use of a sufficiently rich metagenome can circumvent to some extent the time cost required for experimental validation of identifications, and improve the efficiency and accuracy of identifications by predictively probing high-probability strains obtained by screening.

Overall, my research has provided valuable information to the field of environmental microbiology and carbon cycle research. Using machine learning techniques, I have made significant progress in identifying potential bacterial hosts of TYR genes in sampled environments, and through machine learning algorithms, I have calculated enriched strains of TYR genes and possible hosts for subsequent qualitative validation. This in part contributes to the subsequent understanding of the complex dynamics of TYR in the global carbon cycle and environmental processes. Identifying as many TYR hosts as possible can greatly assist in subsequent conditional and quantitative studies of the role of TYR.

However, there are certain limitations to my research. The dataset used for model training and validation was based on publicly available Metagenomic datasets that may not fully represent the diversity of bacterial communities in all environments. Future studies should consider using larger and more comprehensive datasets to improve the accuracy and generalisation of predictive models. In addition, further experiments are needed to validate the predictive correlation between bacterial taxa and TYR gene abundance. Moreover, the model in this study discarded many influencing factors and variables in the real environment. For example, sampling time, sampling temperature, sample size, geographic location elevation and many other scientific factors. Subsequently, the model can be optimised by adding subject variables as needed, or changing variables to target variables for specific identification based on this algorithm.

5 Conclusions

In summary, my study takes the perspective of the importance of bacteria in ecosystems in reducing greenhouse gas emissions and mitigating climate change, using the application of multiple linear regression and machine learning techniques in order to be able to identify potential bacterial hosts of TYR genes in different environments. This provides valuable assistance for future studies on the ecological and functional significance of TYR in different environments. It lays the foundation for future studies on microbial diversity of ecosystems, environmental microbiology and carbon cycle processes.

6 Data Availability Statement

The data and code underlying this study are available from https://github.com/Mingji0613/MasterProject.git.

7 Acknowledgements

I extend my heartfelt appreciation to my esteemed supervisor, Samraat Pawar, whose unwavering care, meticulous attention, and boundless patience have been instrumental throughout the course of my project. My profound gratitude also goes to the dedicated researchers whose invaluable efforts in collecting the essential data have lent remarkable support to this endeavor.

References

Barrett, P., Hunter, J., Miller, J. T., Hsu, J.-C., and Greenfield, P. (2005). matplotlib–a portable python plotting package. In *Astronomical data analysis software and systems XIV*, volume 347, page 91.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the international AAAI conference on web and social media*, volume 3, pages 361–362.

Beals, E. W. (1984). Bray-curtis ordination: an effective strategy for analysis of multivariate ecological data. In *Advances in ecological research*, volume 14, pages 1–55. Elsevier.

Bisong, E. and Bisong, E. (2019). Introduction to scikit-learn. *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, pages 215–229.

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using diamond. *Nature methods*, 12(1):59–60.

Ci, B. and Rule, R.-O. (1987). Confidence intervals. *Lancet*, 1(8531):494–7.

Cohen, I., Huang, Y., Chen, J., Benesty, J., Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

Daniel, R. (2005). The metagenomics of soil. *Nature reviews microbiology*, 3(6):470–478.

De Mandal, S., Laskar, F., Panda, A. K., and Mishra, R. (2020). Microbial diversity and functional potential in wetland ecosystems. In *Recent Advancements in Microbial Diversity*, pages 289–314. Elsevier.

Eberly, L. E. (2007). Multiple linear regression. *Topics in Biostatistics*, pages 165–187.

Edwards, P. M. (2002). Origin 7.0: scientific graphing and data analysis software. *Journal of chemical information and computer sciences*, 42(5):1270–1271.

Erhard, F. (2018). Estimating pseudocounts and fold changes for digital expression measurements. *Bioinformatics*, 34(23):4054–4063.

Falkowski, P., Scholes, R., Boyle, E., Canadell, J., Canfield, D., Elser, J., Gruber, N., Hibbard, K., Högberg, P., Linder, S., et al. (2000). The global carbon cycle: a test of our knowledge of earth as a system. *science*, 290(5490):291–296.

Grace, J. (2004). Understanding and managing the global carbon cycle. *Journal of Ecology*, 92(2):189–202.

Hassan, M., Shahzadi, S., and Kloczkowski, A. (2023). Tyrosinase inhibitors naturally present in plants and synthetic modifications of these natural products as anti-melanogenic agents: a review. *Molecules*, 28(1):378.

Hawkins, H.-J., Cargill, R. I., Van Nuland, M. E., Hagen, S. C., Field, K. J., Sheldrake, M., Soudzilovskaia, N. A., and Kiers, E. T. (2023). Mycorrhizal mycelium as a global carbon pool. *Current Biology*, 33(11):R560–R573.

Helbling, M. and Meierrieks, D. (2023). Global warming and urbanization. *Journal of Population Economics*, 36(3):1187–1223.

Ho, J., Tumkaya, T., Aryal, S., Choi, H., and Claridge-Chang, A. (2019). Moving beyond p values: data analysis with estimation graphics. *Nature methods*, 16(7):565–566.

Kim, T. K. (2015). T test as a parametric statistic. *Korean journal of anesthesiology*, 68(6):540–546.

Kramer, O. and Kramer, O. (2016). Scikit-learn. *Machine learning for evolution strategies*, pages 45–53.

Kristiansson, E., Fick, J., Janzon, A., Grabic, R., Rutgersson, C., Weijdegård, B., Söderström, H., and Larsson, D. J. (2011). Pyrosequencing of antibiotic-contaminated river sediments reveals high levels of resistance and gene transfer elements. *PloS one*, 6(2):e17038.

Maulud, D. and Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, 1(4):140–147.

Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature reviews genetics*, 11(1):31–46.

Moradigaravand, D., Palm, M., Farewell, A., Mustonen, V., Warringer, J., and Parts, L. (2018). Prediction of antibiotic resistance in escherichia coli from large-scale pan-genome data. *PLoS computational biology*, 14(12):e1006258.

Nolan, K. A. and Callahan, J. E. (2006). Beachcomber biology: The shannon-weiner species diversity index. In *Proc. workshop able*, volume 27, pages 334–338.

Panis, F., Krachler, R. F., Krachler, R., and Rompel, A. (2021). Expression, purification, and characterization of a well-adapted tyrosinase from peatlands identified by partial community analysis. *Environmental Science & Technology*, 55(16):11445–11454.

Panis, F. and Rompel, A. (2022). The novel role of tyrosinase enzymes in the storage of globally significant amounts of carbon in wetland ecosystems. *Environmental Science & Technology*, 56(17):11952–11968.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Schlesinger, W. H. and Andrews, J. A. (2000). Soil respiration and the global carbon cycle. *Biogeochemistry*, 48:7–20.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*, 9(8):811–814.

Slatko, B. E., Gardner, A. F., and Ausubel, F. M. (2018). Overview of next-generation sequencing technologies. *Current protocols in molecular biology*, 122(1):e59.

Sleator, R. D., Shortall, C., and Hill, C. (2008). Metagenomics. *Letters in applied microbiology*, 47(5):361–366.

Socolar, J. B., Gilroy, J. J., Kunin, W. E., and Edwards, D. P. (2016). How should beta-diversity inform biodiversity conservation? *Trends in ecology & evolution*, 31(1):67–80.

Sodangi, I. A., Izge, A., and Maina, Y. (2011). Climate change: causes and effects on african agriculture.

Sun, Y., Clarke, B., Clarke, J., and Li, X. (2021). Predicting antibiotic resistance gene abundance in activated sludge using shotgun metagenomics and machine learning. *Water Research*, 202:117384.

Thukral, A. K. (2017). A review on measurement of alpha diversity in biology. *Agricultural Research Journal*, 54(1).

Váradi, L., Luo, J. L., Hibbs, D. E., Perry, J. D., Anderson, R. J., Orenga, S., and Groundwater, P. W. (2017). Methods for the detection and identification of pathogenic bacteria: past, present, and future. *Chemical Society Reviews*, 46(16):4818–4832.

Wani, O. A., Kumar, S. S., Hussain, N., Wani, A. I. A., Subhash, B., Parvej, A., Rashid, M., Popescu, S. M., and Mansoor, S. (2023). Multi-scale processes influencing global carbon storage and land-carbon-climate nexus: A critical review. *Pedosphere*, 33(2):250–267.

Wickham, H. (2006). An introduction to ggplot: An implementation of the grammar of graphics in r. *Statistics*, pages 1–8.

Wooley, J. C., Godzik, A., and Friedberg, I. (2010). A primer on metagenomics. *PLoS computational biology*, 6(2):e1000667.

Yang, Y., Li, B., Ju, F., and Zhang, T. (2013). Exploring variation of antibiotic resistance genes in activated sludge over a four-year period through a metagenomic approach. *Environmental science & technology*, 47(18):10197–10205.

Supplementary Information

Supplementary Table 1 Environment information and reads of 530 Samples

Sample ID	Reads	Environment
ERR011089	4189308	Fecal
ERR011090	11165183	Fecal
ERR011091	7927536	Fecal
ERR011092	4477312	Fecal
ERR011093	10856019	Fecal
ERR011094	9857500	Fecal
ERR011101	14814589	Fecal
ERR011102	15627318	Fecal
ERR011103	15916662	Fecal
ERR011104	15103477	Fecal
ERR011109	4876142	Fecal
ERR011110	11283995	Fecal
ERR011111	13062929	Fecal
ERR011114	4956118	Fecal
ERR011115	11088635	Fecal
ERR011116	13201615	Fecal
ERR011117	5074119	Fecal
ERR011118	10687883	Fecal
ERR011119	13150141	Fecal
ERR011120	15824363	Fecal
ERR011121	16560727	Fecal
ERR011122	15771599	Fecal
ERR011123	15956567	Fecal
ERR011126	4778196	Fecal
ERR011127	10029059	Fecal
ERR011128	13340573	Fecal
ERR011131	4840022	Fecal
ERR011132	9033892	Fecal
ERR011133	13321829	Fecal
ERR011140	10094227	Fecal
ERR4682452	43195714	Sewage
ERR4682453	41840131	Sewage
ERR4682454	35106658	Sewage
ERR4682455	38106623	Sewage
ERR4682456	30104711	Sewage
ERR4682459	35820665	Sewage
ERR4682461	14690822	Sewage
ERR4682462	31029992	Sewage
ERR4682463	39817018	Sewage
ERR4682464	31922765	Sewage
ERR4682466	35987243	Sewage
ERR4682467	46414839	Sewage
ERR4682773	47140348	Sewage
ERR4682774	52235960	Sewage
ERR4682775	53725060	Sewage
ERR4682778	34807224	Sewage
ERR4682780	52127399	Sewage
ERR4682781	37873741	Sewage
ERR4682782	36402766	Sewage
ERR4682783	43977696	Sewage
ERR4682785	43354717	Sewage
ERR4682788	51751124	Sewage
ERR4682789	38324022	Sewage
ERR4682790	39934216	Sewage
ERR4682791	50731053	Sewage
ERR4682792	46586141	Sewage
ERR4682793	42960416	Sewage
ERR4682794	92911992	Sewage
ERR4682797	43416998	Sewage
ERR4682798	56336851	Sewage
SRR8942237	90030803	Large River
SRR8942238	73016059	Large River
SRR8942239	72476722	Large River
SRR8942240	68191296	Large River
SRR8942241	71661632	Large River
SRR8942242	49217494	Large River
SRR8942243	67310254	Large River
SRR8942244	66736832	Large River
SRR8942245	72021766	Large River
SRR8942246	78824062	Large River
SRR8942247	61342709	Large River
SRR8942248	59507691	Large River
SRR8942249	55019460	Large River
SRR8942250	58798464	Large River
SRR8942251	60002675	Large River
SRR8942253	81751105	Large River
SRR8942254	58273488	Large River
SRR8942255	73073394	Large River
SRR8942256	62477120	Large River
SRR8942257	50590507	Large River
SRR8942258	46362573	Large River

SRR8942259	78567507	Large River
SRR8942260	76019378	Large River
SRR8942261	73717827	Large River
SRR8942262	72600958	Large River
SRR8942263	76519203	Large River
SRR8942264	86345482	Large River
SRR8942265	56764625	Large River
SRR8942266	65550706	Large River
SRR8942298	60897027	Large River
SRR8942301	63127691	Large River
SRR8942302	73581852	Large River
SRR8942303	61642275	Large River
SRR8942304	54884052	Large River
SRR8942307	69030168	Large River
SRR8942308	65309625	Large River
SRR14120361	28123961	Urban River
SRR14120362	26977361	Urban River
SRR14120363	32361347	Urban River
SRR14120364	32348999	Urban River
SRR14120365	31525348	Urban River
SRR14120366	29405837	Urban River
SRR14120367	30188915	Urban River
SRR14120368	31117061	Urban River
SRR14120369	26684409	Urban River
SRR12113387	36428361	Groundwater
SRR12113388	32120602	Groundwater
SRR12113389	35027139	Groundwater
SRR12113391	38352667	Groundwater
SRR12113392	56449759	Groundwater
SRR12113393	11510055	Groundwater
SRR12113394	33033546	Groundwater
SRR12113395	12105425	Groundwater
SRR12113396	10440637	Groundwater
SRR7343998	80887417	glacier lake
SRR7343999	23029835	glacier lake
SRR7344000	15192890	glacier lake
SRR7344001	35195281	glacier lake
SRR7344002	22332493	glacier lake
SRR7344003	24482762	glacier lake
SRR8670866	45556983	glacier lake
SRR8670867	39576709	glacier lake
SRR8670868	33471465	glacier lake
SRR8670869	36247242	glacier lake
SRR8670870	49956242	glacier lake
SRR8670871	45970673	glacier lake
SRR8670872	35425143	glacier lake
SRR8670873	36119367	glacier lake
SRR8670874	36628784	glacier lake
SRR8670875	38135666	glacier lake
SRR15237466	21978742	Urban River
SRR15237465	26730431	Urban River
SRR15237464	22870201	Urban River
SRR15237463	25830999	Urban River
SRR15237462	21455861	Urban River
SRR15237461	20188764	Urban River
SRR15237458	23442192	Urban River
SRR15237457	32027408	Urban River
SRR15237456	20932467	Urban River
SRR15237455	20779148	Urban River
SRR15237454	21582379	Urban River
SRR15237453	22372965	Urban River
SRR15237452	35297029	Urban River
SRR15237451	23878460	Urban River
SRR15237450	31012402	Urban River
SRR15237449	21987836	Urban River
SRR15237448	23907356	Urban River
SRR15237447	23742279	Urban River
SRR15237446	23485666	Urban River
SRR15237445	21878569	Urban River
SRR15237444	21094813	Urban River
SRR15237443	22994405	Urban River
SRR13225485	19070893	Urban River
SRR13225484	19545039	Urban River
SRR13225483	19026779	Urban River
SRR13225482	19248525	Urban River
SRR13225481	18471666	Urban River
SRR13225480	19097424	Urban River
SRR13225479	19265987	Urban River
SRR13225478	19091237	Urban River
SRR13225477	19089055	Urban River
SRR13225476	19249365	Urban River
SRR13225475	19794906	Urban River
SRR13225474	19658358	Urban River
SRR13225473	19516498	Urban River
SRR13225472	19695845	Urban River
SRR13225471	19216442	Urban River
SRR13225470	19148905	Urban River
SRR13225469	19774976	Urban River
SRR13225468	19689976	Urban River
SRR13225467	19156541	Urban River
SRR13225466	19788080	Urban River
SRR13225465	19409533	Urban River

SRR13225464	19603801	Urban River
SRR13225463	19310533	Urban River
SRR13225462	19216930	Urban River
SRR13225461	19122145	Urban River
SRR13225460	19780875	Urban River
SRR13225459	19239676	Urban River
SRR13225458	19244944	Urban River
SRR13225457	19768172	Urban River
SRR13225456	19093462	Urban River
SRR13225455	19316938	Urban River
SRR13225454	19683436	Urban River
SRR13225453	19230844	Urban River
SRR13225452	19201071	Urban River
SRR13225451	17410226	Urban River
SRR13225450	19342600	Urban River
SRR13225449	19261611	Urban River
SRR13225448	19287940	Urban River
SRR13225447	19298669	Urban River
SRR10571256	19258841	Urban River
SRR10571255	19608037	Urban River
SRR10571254	19335373	Urban River
SRR10571253	19550669	Urban River
SRR10571252	19410066	Urban River
SRR10571251	19361143	Urban River
SRR10571250	19554814	Urban River
SRR10571249	19614219	Urban River
SRR10571248	19604401	Urban River
SRR10571246	19656927	Urban River
SRR10571245	19570052	Urban River
SRR10571244	19477593	Urban River
SRR10571243	19345584	Urban River
SRR10571242	19589156	Urban River
SRR10571241	19614955	Urban River
SRR10571240	19392131	Urban River
SRR10571239	19457493	Urban River
SRR10571238	19333330	Urban River
SRR10571237	19537247	Urban River
SRR10571214	19375842	Urban River
SRR10571213	19731310	Urban River
SRR10571212	19187926	Urban River
SRR10571211	19181343	Urban River
SRR10571210	19786590	Urban River
SRR10571209	18818562	Urban River
SRR10571208	19724797	Urban River
SRR10571207	19627740	Urban River
SRR10571206	19793194	Urban River
SRR10571205	19595114	Urban River
SRR10132701	19335451	Urban River
SRR10132699	19562974	Urban River
SRR10132698	19552540	Urban River
SRR10132697	19194841	Urban River
SRR10132696	19274144	Urban River
SRR10132695	19115086	Urban River
SRR10132694	19009488	Urban River
SRR10132693	19134666	Urban River
SRR10132692	19134257	Urban River
SRR10132691	19134335	Urban River
SRR10132690	19018269	Urban River
SRR10132688	18905970	Urban River
SRR10132687	19311354	Urban River
SRR10132686	18933521	Urban River
SRR10132685	18637046	Urban River
SRR10132684	18895087	Urban River
SRR10132683	19057657	Urban River
SRR10132682	19212543	Urban River
SRR10132681	16772860	Urban River
SRR10132680	19375842	Urban River
SRR10132679	19731310	Urban River
SRR10132677	19187926	Urban River
SRR10132676	19181343	Urban River
SRR10132675	19786590	Urban River
SRR10132674	18818562	Urban River
SRR10132673	19724797	Urban River
SRR10132672	19095892	Urban River
SRR10132671	15249669	Urban River
SRR10132670	19644101	Urban River
SRR10132669	19162033	Urban River
SRR10132668	19094262	Urban River
SRR10132666	19164253	Urban River
SRR10132665	19441895	Urban River
SRR10132664	19450471	Urban River
SRR10132663	19563563	Urban River
SRR10132662	18177286	Urban River
SRR10132661	19196701	Urban River
SRR10132660	19594864	Urban River
SRR10132659	17467906	Urban River
SRR10132658	19632188	Urban River
SRR10132657	19184720	Urban River
SRR10132655	19259609	Urban River
SRR10132654	19310054	Urban River
SRR10132653	19219441	Urban River
SRR10132652	19447206	Urban River

SRR10132651	18758085	Urban River
SRR10132650	19355414	Urban River
SRR10132649	19371015	Urban River
SRR10132648	19141441	Urban River
SRR10132647	19170104	Urban River
SRR10132646	19080554	Urban River
SRR10132644	19228917	Urban River
SRR10132643	17494404	Urban River
SRR10132642	19215848	Urban River
SRR10132641	19132414	Urban River
SRR10132640	19312839	Urban River
SRR10132639	19627740	Urban River
SRR10132638	19793194	Urban River
SRR10132637	19595114	Urban River
SRR10132636	19748565	Urban River
SRR10132635	19258841	Urban River
SRR10132632	19608037	Urban River
SRR10132631	19335373	Urban River
SRR10132630	19588666	Urban River
SRR10132629	19273304	Urban River
SRR10132628	18854913	Urban River
SRR10132627	19614892	Urban River
SRR10132626	19188306	Urban River
SRR10132625	19230619	Urban River
SRR10132624	19815447	Urban River
SRR10132623	19127139	Urban River
SRR10132621	19165660	Urban River
SRR10132620	19742511	Urban River
SRR10132619	19738872	Urban River
SRR10132618	19083808	Urban River
SRR10132617	19105059	Urban River
SRR10132616	19298903	Urban River
ERR011141	12803061	Fecal
ERR011142	8958733	Fecal
ERR011143	4168325	Fecal
ERR011148	10710932	Fecal
ERR011150	11437582	Fecal
ERR011152	10893200	Fecal
ERR011153	7847511	Fecal
ERR011156	10317483	Fecal
ERR011158	10126724	Fecal
ERR011160	10768394	Fecal
ERR011162	9133455	Fecal
ERR011164	9129898	Fecal
ERR011165	13366786	Fecal
ERR011166	4806843	Fecal
ERR011167	13893002	Fecal
ERR011168	10922014	Fecal
ERR011169	13705983	Fecal
ERR011170	10483020	Fecal
ERR011171	13367401	Fecal
ERR011172	8643128	Fecal
ERR011173	12372278	Fecal
ERR011174	9567382	Fecal
ERR011175	13162165	Fecal
ERR011176	8838893	Fecal
ERR011177	12518664	Fecal
ERR011178	8679313	Fecal
ERR011179	13444486	Fecal
ERR011180	9032917	Fecal
ERR011181	13151792	Fecal
ERR011182	8802750	Fecal
ERR011183	12900414	Fecal
ERR011184	11610294	Fecal
ERR011185	13058081	Fecal
ERR011186	9171391	Fecal
ERR011187	13259545	Fecal
ERR011188	11544188	Fecal
ERR011189	13170508	Fecal
ERR011190	6052183	Fecal
ERR011191	11844870	Fecal
ERR011192	10769772	Fecal
ERR011193	2694802	Fecal
ERR011194	10563584	Fecal
ERR011195	2920218	Fecal
ERR011196	10935278	Fecal
ERR011197	4391268	Fecal
ERR011198	11556341	Fecal
ERR011199	12338464	Fecal
ERR011200	10151349	Fecal
ERR011201	2825973	Fecal
ERR011202	10592713	Fecal
ERR011203	3693580	Fecal
ERR011204	8624377	Fecal
ERR011205	13016035	Fecal
ERR011206	8493115	Fecal
ERR011207	12636094	Fecal
ERR011208	10018598	Fecal
ERR011209	13535415	Fecal
ERR011210	11324515	Fecal
ERR011211	13242905	Fecal
ERR011212	8046008	Fecal

ERR011213	13477358	Fecal
ERR011214	11595122	Fecal
ERR011215	13311136	Fecal
ERR011216	10012573	Fecal
ERR011217	10494611	Fecal
ERR011218	8422281	Fecal
ERR011219	12132695	Fecal
ERR011220	11046354	Fecal
ERR011221	11175648	Fecal
ERR011222	9350300	Fecal
ERR011223	8905846	Fecal
ERR011224	11837789	Fecal
ERR011225	11140970	Fecal
ERR011226	11167478	Fecal
ERR011227	11196413	Fecal
ERR011228	11217122	Fecal
ERR011229	11069860	Fecal
ERR011230	10943896	Fecal
ERR011231	7112559	Fecal
ERR011232	11600193	Fecal
ERR011233	10641529	Fecal
ERR011234	11097318	Fecal
ERR011235	10834690	Fecal
ERR011236	10853791	Fecal
ERR011237	11959007	Fecal
ERR011238	12043015	Fecal
ERR011239	11320404	Fecal
ERR011240	12636427	Fecal
ERR011241	11194636	Fecal
ERR011242	11922800	Fecal
ERR011243	12098158	Fecal
ERR011244	12539905	Fecal
ERR011245	11027848	Fecal
ERR011246	12586444	Fecal
ERR011247	11172112	Fecal
ERR4682857	111343794	Sewage
ERR4682347	69184522	Sewage
ERR4682900	62816644	Sewage
ERR4682866	62137616	Sewage
ERR4682349	60409715	Sewage
ERR4682345	60117785	Sewage
ERR4682804	58948286	Sewage
ERR4682335	58298080	Sewage
ERR4682358	57909414	Sewage
ERR4682359	57432027	Sewage
ERR4682355	57304731	Sewage
ERR4682353	57095621	Sewage
ERR4682334	56671203	Sewage
ERR4682339	55781548	Sewage
ERR4682836	55234812	Sewage
ERR4682805	54945758	Sewage
ERR4682427	54745223	Sewage
ERR4682839	53335414	Sewage
ERR4682346	52997161	Sewage
ERR4682354	52109453	Sewage
ERR4682835	51776040	Sewage
ERR4682401	51711865	Sewage
ERR4682397	51394099	Sewage
ERR4682409	49997715	Sewage
ERR4682820	49990869	Sewage
ERR4682343	49685848	Sewage
ERR4682801	49599987	Sewage
ERR3562852	49002440	Sewage
ERR4682832	48664807	Sewage
ERR4682387	48015736	Sewage
ERR4682851	47465761	Sewage
ERR4682411	47276063	Sewage
ERR4682407	46941683	Sewage
ERR4682421	46795743	Sewage
ERR4682418	46535033	Sewage
ERR4682352	46152331	Sewage
ERR4682869	46122903	Sewage
ERR4682825	45197410	Sewage
ERR4682852	45000450	Sewage
ERR4682389	44951034	Sewage
ERR3562861	44493451	Sewage
ERR4682399	44269067	Sewage
ERR4682875	43793467	Sewage
ERR4682356	43523026	Sewage
ERR4682840	43343824	Sewage
ERR4682847	43314151	Sewage
ERR4682386	43092208	Sewage
ERR4682833	42886674	Sewage
ERR4682378	42846025	Sewage
ERR3562837	42800891	Sewage
ERR4682420	42636846	Sewage
ERR4682432	42103536	Sewage
ERR4682819	41885523	Sewage
ERR4682429	41854509	Sewage
ERR4682362	41852142	Sewage
ERR4682372	41718898	Sewage
ERR4682895	41579390	Sewage

ERR4682867	41343706	Sewage
ERR4682844	41226720	Sewage
ERR4682373	41023432	Sewage
ERR4682812	40813794	Sewage
ERR4682426	40730992	Sewage
ERR4682366	40356751	Sewage
ERR4682430	40295015	Sewage
ERR4682408	40134589	Sewage
ERR4682388	39617152	Sewage
ERR4682850	39321072	Sewage
ERR4682374	39035215	Sewage
ERR3562866	38958998	Sewage
ERR4682811	38932695	Sewage
ERR4682404	38765953	Sewage
ERR4682896	38670754	Sewage
ERR4682874	38516422	Sewage
ERR4682849	38342705	Sewage
ERR4682344	38250141	Sewage
ERR4682414	38109092	Sewage
ERR4682818	38075884	Sewage
ERR4682391	37918380	Sewage
ERR4682406	37845363	Sewage
ERR4682423	36498124	Sewage
ERR4682845	36337401	Sewage
ERR4682380	36205127	Sewage
ERR3562849	36195360	Sewage
ERR4682415	36032883	Sewage
ERR4682443	36022960	Sewage
ERR3562867	35827486	Sewage
ERR4682428	35757114	Sewage
ERR4682422	35574745	Sewage
ERR4682361	35514206	Sewage
ERR4682431	35409996	Sewage
ERR4682417	35147507	Sewage
ERR4682450	35032307	Sewage
ERR4682826	34647359	Sewage
ERR4682897	34428523	Sewage
ERR3562865	34199832	Sewage
ERR4682889	33892180	Sewage
ERR3562870	33521075	Sewage
ERR3562869	32796585	Sewage
ERR3562850	30515644	Sewage
ERR3562853	30206345	Sewage
ERR3562868	29926797	Sewage
ERR3562864	27997532	Sewage
ERR4682398	27624799	Sewage
ERR3562836	26035123	Sewage
ERR3562863	25039649	Sewage
ERR3562854	24599973	Sewage
ERR3562862	23158276	Sewage
ERR3562843	19158198	Sewage
ERR3562851	16841347	Sewage
ERR3562848	15443092	Sewage
ERR3562846	13603243	Sewage
ERR3562847	13316798	Sewage
ERR3562845	13292182	Sewage
ERR3562844	12472118	Sewage
ERR3562839	4223695	Sewage
ERR3562838	3773399	Sewage
ERR3562860	2426487	Sewage
ERR3562842	2388949	Sewage
ERR3562876	1712526	Sewage
ERR3562891	1603589	Sewage
ERR3562888	1471517	Sewage
ERR3562875	1469807	Sewage
ERR3562895	1295213	Sewage
ERR3562856	984302	Sewage
ERR3562857	894620	Sewage
ERR3562874	839740	Sewage
ERR3562859	816202	Sewage
ERR3562873	696063	Sewage
ERR3562893	694210	Sewage
ERR3562855	666328	Sewage
ERR3562841	658881	Sewage
ERR3562872	657883	Sewage
ERR3562871	643261	Sewage
ERR3562897	617838	Sewage
ERR3562840	616149	Sewage
ERR3562885	561563	Sewage
ERR3562882	559283	Sewage
ERR3562858	515744	Sewage

Supplementary Table 2 TYR Abundence value of 530 Samples

SAMPLE ID	tyr(PPM)
ERR011141	900.3315691
ERR011142	861.7289967
ERR011143	1232.869318
ERR011148	872.3797332
ERR011150	1432.033449

ERR011152	792.4209599	**ERR011204**	1311.978825
ERR011153	570.2445017	**ERR011205**	1203.745995
ERR011156	1338.989364	**ERR011206**	1007.28649
ERR011158	993.7073431	**ERR011207**	1058.238408
ERR011160	1224.788023	**ERR011208**	1032.37998
ERR011162	776.5954943	**ERR011209**	1055.157895
ERR011164	925.3115424	**ERR011210**	701.9285153
ERR011165	995.2280227	**ERR011211**	975.9188033
ERR011166	179.3276793	**ERR011212**	799.0297797
ERR011167	699.7767653	**ERR011213**	706.8150894
ERR011168	1377.95099	**ERR011214**	1055.961291
ERR011169	1196.411815	**ERR011215**	949.9564876
ERR011170	780.309491	**ERR011216**	1057.370568
ERR011171	964.7350296	**ERR011217**	1348.025191
ERR011172	586.4774882	**ERR011218**	986.9060412
ERR011173	793.1441566	**ERR011219**	991.2059934
ERR011174	546.439977	**ERR011220**	674.7927868
ERR011175	1121.396062	**ERR011221**	745.3706488
ERR011176	602.3378719	**ERR011222**	1164.133771
ERR011177	925.0188359	**ERR011223**	1170.803987
ERR011178	1306.439807	**ERR011224**	743.8889137
ERR011179	1133.773355	**ERR011225**	843.4633609
ERR011180	928.6036836	**ERR011226**	955.8111509
ERR011181	1028.453005	**ERR011227**	865.0091775
ERR011182	517.9063361	**ERR011228**	1215.998186
ERR011183	914.8543605	**ERR011229**	1086.55394
ERR011184	854.1558035	**ERR011230**	1022.213661
ERR011185	891.6317796	**ERR011231**	806.3202006
ERR011186	1119.459415	**ERR011232**	760.3321772
ERR011187	1174.022186	**ERR011233**	780.6209051
ERR011188	1120.99699	**ERR011234**	1065.122221
ERR011189	1024.409992	**ERR011235**	888.9963626
ERR011190	844.9843635	**ERR011236**	1482.154945
ERR011191	975.1056787	**ERR011237**	1084.872682
ERR011192	770.5827013	**ERR011238**	1153.780843
ERR011193	801.5431189	**ERR011239**	1261.792424
ERR011194	882.0869887	**ERR011240**	1147.47626
ERR011195	1071.495347	**ERR011241**	1103.921557
ERR011196	746.7574213	**ERR011242**	947.5962022
ERR011197	838.4821878	**ERR011243**	780.3667302
ERR011198	1031.468351	**ERR011244**	991.2355795
ERR011199	1132.393789	**ERR011245**	1177.745649
ERR011200	751.4272241	**ERR011246**	1187.38859
ERR011201	1087.059218	**ERR011247**	967.0508137
ERR011202	887.4025002	**ERR3562836**	1629.375824
ERR011203	1254.338609	**ERR3562837**	1444.923191

ERR3562838	1077.013059	**ERR4682334**	1049.968888
ERR3562839	1110.875667	**ERR4682335**	1365.928346
ERR3562840	353.8105231	**ERR4682339**	1458.636465
ERR3562841	701.1888338	**ERR4682343**	785.5355513
ERR3562842	925.5115953	**ERR4682344**	1037.434084
ERR3562843	1451.180325	**ERR4682345**	906.2210126
ERR3562844	1177.025426	**ERR4682346**	852.7626602
ERR3562845	1418.803925	**ERR4682347**	1079.721271
ERR3562846	1685.112881	**ERR4682349**	1212.470544
ERR3562847	1377.433224	**ERR4682352**	577.8473031
ERR3562848	1127.753432	**ERR4682353**	1302.481674
ERR3562849	1327.628735	**ERR4682354**	1115.99713
ERR3562850	1525.80755	**ERR4682355**	755.0336463
ERR3562851	1569.173772	**ERR4682356**	763.8715194
ERR3562852	1300.384226	**ERR4682358**	1208.508171
ERR3562853	1530.075883	**ERR4682359**	1197.990104
ERR3562854	1148.415895	**ERR4682361**	1161.169139
ERR3562855	366.1860225	**ERR4682362**	1469.912818
ERR3562856	409.427188	**ERR4682366**	1211.792297
ERR3562857	388.9919742	**ERR4682372**	1229.34695
ERR3562858	325.7430043	**ERR4682373**	1217.133662
ERR3562859	346.7278933	**ERR4682374**	818.3380058
ERR3562860	852.2609023	**ERR4682378**	1571.300955
ERR3562861	1403.15032	**ERR4682380**	1316.553868
ERR3562862	1141.967563	**ERR4682386**	783.2738578
ERR3562863	1157.20472	**ERR4682387**	1911.352562
ERR3562864	1342.082581	**ERR4682388**	1318.267401
ERR3562865	1102.461556	**ERR4682389**	839.4022705
ERR3562866	1495.675017	**ERR4682391**	1103.211688
ERR3562867	1301.821735	**ERR4682397**	2304.992252
ERR3562868	1458.726104	**ERR4682398**	1252.063409
ERR3562869	1293.122439	**ERR4682399**	1293.905742
ERR3562870	1204.07833	**ERR4682401**	1661.707618
ERR3562871	348.2256813	**ERR4682404**	1134.165333
ERR3562872	343.5261285	**ERR4682406**	1250.721257
ERR3562873	405.135742	**ERR4682407**	1169.66407
ERR3562874	402.5055374	**ERR4682408**	1117.31554
ERR3562875	964.7525151	**ERR4682409**	751.4343405
ERR3562876	1016.042968	**ERR4682411**	2245.406941
ERR3562882	1451.858898	**ERR4682414**	975.2265942
ERR3562885	956.2595826	**ERR4682415**	734.107232
ERR3562888	1065.56703	**ERR4682417**	554.0933529
ERR3562891	865.5584442	**ERR4682418**	524.4650842
ERR3562893	1153.829533	**ERR4682420**	723.5056739
ERR3562895	1033.034721	**ERR4682421**	1544.456725
ERR3562897	715.3978875	**ERR4682422**	1297.634038

ERR4682423	1733.321965	**SRR10132618**	804.766009
ERR4682426	1054.577802	**SRR10132619**	458.8914706
ERR4682427	1402.350667	**SRR10132620**	615.575192
ERR4682428	1278.738547	**SRR10132621**	553.6986464
ERR4682429	1516.156837	**SRR10132623**	635.2230723
ERR4682430	1618.19024	**SRR10132624**	846.4103787
ERR4682431	1385.682167	**SRR10132625**	929.1952589
ERR4682432	2303.98701	**SRR10132626**	1145.645686
ERR4682443	1436.000817	**SRR10132627**	700.7940701
ERR4682450	2240.531861	**SRR10132628**	681.5730203
ERR4682801	1238.629357	**SRR10132629**	824.1970344
ERR4682804	1069.547637	**SRR10132630**	646.2920956
ERR4682805	1552.822331	**SRR10132631**	876.3213412
ERR4682811	2160.0611	**SRR10132632**	789.625193
ERR4682812	2228.021242	**SRR10132635**	762.5069442
ERR4682818	2486.928472	**SRR10132636**	710.4313655
ERR4682819	1329.528582	**SRR10132637**	576.6233358
ERR4682820	1472.72895	**SRR10132638**	693.6222623
ERR4682825	728.3603198	**SRR10132639**	635.2743617
ERR4682826	1494.110994	**SRR10132640**	1073.068543
ERR4682832	2005.227309	**SRR10132641**	1153.539747
ERR4682833	1654.429998	**SRR10132642**	944.9491899
ERR4682835	1892.786702	**SRR10132643**	1155.340874
ERR4682836	2279.757918	**SRR10132644**	1171.828866
ERR4682839	940.6133043	**SRR10132646**	1055.524908
ERR4682840	1550.809176	**SRR10132647**	1158.10535
ERR4682844	1686.285011	**SRR10132648**	736.5171723
ERR4682845	1470.055605	**SRR10132649**	721.9033179
ERR4682847	1665.114018	**SRR10132650**	543.052192
ERR4682849	1278.313567	**SRR10132651**	1633.535619
ERR4682850	873.6282673	**SRR10132652**	1451.108195
ERR4682851	1866.103021	**SRR10132653**	1111.010461
ERR4682852	1107.98892	**SRR10132654**	780.3706815
ERR4682857	1616.174495	**SRR10132655**	630.4904736
ERR4682866	1801.002472	**SRR10132657**	701.0266504
ERR4682867	658.0203526	**SRR10132658**	832.2047446
ERR4682869	3990.121784	**SRR10132659**	894.3831046
ERR4682874	2094.976527	**SRR10132660**	911.7184993
ERR4682875	1306.587578	**SRR10132661**	416.4257181
ERR4682889	1131.116381	**SRR10132662**	561.5249713
ERR4682895	2687.918221	**SRR10132663**	744.0873628
ERR4682896	785.8134858	**SRR10132664**	728.46565
ERR4682897	1117.765058	**SRR10132665**	823.890881
ERR4682900	1374.778952	**SRR10132666**	602.580231
SRR10132616	1060.474784	**SRR10132668**	642.8632853
SRR10132617	781.1543529	**SRR10132669**	576.8698968

SRR10132670	784.4594161	**SRR10571244**	576.5599476
SRR10132671	1329.405904	**SRR10571245**	558.4553378
SRR10132672	1038.390875	**SRR10571246**	565.6530138
SRR10132673	822.1630874	**SRR10571248**	506.7229547
SRR10132674	362.4612763	**SRR10571249**	578.7128205
SRR10132675	481.7909503	**SRR10571250**	658.1499573
SRR10132676	568.6775947	**SRR10571251**	533.0263818
SRR10132677	661.9787881	**SRR10571252**	572.3834221
SRR10132679	779.0156862	**SRR10571253**	523.2046024
SRR10132680	490.0948305	**SRR10571254**	876.3213412
SRR10132681	1029.341448	**SRR10571255**	789.625193
SRR10132682	589.1463717	**SRR10571256**	762.5069442
SRR10132683	570.5843063	**SRR13225447**	972.7613858
SRR10132684	409.2598251	**SRR13225448**	556.3061685
SRR10132685	562.0525914	**SRR13225449**	927.1290963
SRR10132686	754.4291418	**SRR13225450**	1136.455285
SRR10132687	797.6136733	**SRR13225451**	993.0370806
SRR10132688	775.7338026	**SRR13225452**	674.5977868
SRR10132690	432.531478	**SRR13225453**	1000.268111
SRR10132691	1398.63758	**SRR13225454**	914.3220726
SRR10132692	379.0583559	**SRR13225455**	878.7624622
SRR10132693	410.6682604	**SRR13225456**	843.1682007
SRR10132694	353.7181012	**SRR13225457**	480.4187256
SRR10132695	354.9552432	**SRR13225458**	615.6422175
SRR10132696	544.2524451	**SRR13225459**	1208.54426
SRR10132697	1111.704963	**SRR13225460**	992.5243449
SRR10132698	511.8516571	**SRR13225461**	599.3574466
SRR10132699	331.9536181	**SRR13225462**	769.2695972
SRR10132701	378.99297	**SRR13225463**	1215.761367
SRR10571205	576.6233358	**SRR13225464**	614.1155993
SRR10571206	693.6222623	**SRR13225465**	437.0017558
SRR10571207	635.2743617	**SRR13225466**	285.9297112
SRR10571208	822.1630874	**SRR13225467**	2186.668251
SRR10571209	362.4612763	**SRR13225468**	593.8046852
SRR10571210	481.7909503	**SRR13225469**	828.0161756
SRR10571211	568.6775947	**SRR13225470**	675.2344325
SRR10571212	661.9787881	**SRR13225471**	725.9928763
SRR10571213	779.0156862	**SRR13225472**	657.9560308
SRR10571214	490.0948305	**SRR13225473**	1964.952934
SRR10571237	834.5085671	**SRR13225474**	572.0213255
SRR10571238	499.6035344	**SRR13225475**	538.4718675
SRR10571239	556.443731	**SRR13225476**	1177.857036
SRR10571240	419.2422174	**SRR13225477**	1123.156699
SRR10571241	415.0404627	**SRR13225478**	1305.101393
SRR10571242	465.1042648	**SRR13225479**	535.7109397
SRR10571243	492.8773409	**SRR13225480**	268.1513486

SRR13225481	432.2836933	**ERR011120**	1156.760623
SRR13225482	1238.172795	**ERR011121**	1069.940951
SRR13225483	996.12236	**ERR011122**	1195.31317
SRR13225484	1080.427622	**ERR011123**	1186.282739
SRR13225485	2334.657323	**ERR011126**	2335.190938
SRR15237443	730.1776236	**ERR011127**	2074.172662
SRR15237444	930.4182976	**ERR011128**	2603.036616
SRR15237445	899.8303317	**ERR011131**	1091.937185
SRR15237446	995.3305135	**ERR011132**	691.9498263
SRR15237447	1019.868396	**ERR011133**	965.5581077
SRR15237448	963.8874328	**ERR011140**	654.3344032
SRR15237449	934.0164262	**ERR4682452**	1689.889881
SRR15237450	989.313888	**ERR4682453**	2492.822023
SRR15237451	1431.248079	**ERR4682454**	1457.729186
SRR15237452	1511.798628	**ERR4682455**	1540.598336
SRR15237453	1174.676669	**ERR4682456**	1427.351354
SRR15237454	1551.91418	**ERR4682459**	2575.775743
SRR15237455	2483.691824	**ERR4682461**	1067.060781
SRR15237456	1807.479262	**ERR4682462**	1417.854056
SRR15237457	1118.323406	**ERR4682463**	1478.563764
SRR15237458	849.792545	**ERR4682464**	1346.280625
SRR15237461	735.6071922	**ERR4682466**	1033.560698
SRR15237462	959.2250807	**ERR4682467**	1519.794133
SRR15237463	1129.72789	**ERR4682773**	2100.472402
SRR15237464	946.9964868	**ERR4682774**	1494.640857
SRR15237465	1012.291945	**ERR4682775**	1830.244582
SRR15237466	941.7736466	**ERR4682778**	1679.680057
ERR011089	1276.821852	**ERR4682780**	1386.526115
ERR011090	957.9780287	**ERR4682781**	1586.798621
ERR011091	776.9122714	**ERR4682782**	816.4489479
ERR011092	1330.262443	**ERR4682783**	1277.374786
ERR011093	1141.118121	**ERR4682785**	1240.141874
ERR011094	1055.947248	**ERR4682788**	1452.548161
ERR011101	1091.424136	**ERR4682789**	1235.80453
ERR011102	1247.23897	**ERR4682790**	2115.378952
ERR011103	1233.110309	**ERR4682791**	1516.999854
ERR011104	1289.835447	**ERR4682792**	1458.867349
ERR011109	1300.413319	**ERR4682793**	1055.040994
ERR011110	1026.675393	**ERR4682794**	2118.58551
ERR011111	986.9149561	**ERR4682797**	1792.086132
ERR011114	1208.203679	**ERR4682798**	871.0462003
ERR011115	980.4633302	**SRR12113387**	266.5780105
ERR011116	1097.895977	**SRR12113388**	226.0231611
ERR011117	1143.646808	**SRR12113389**	280.2112956
ERR011118	1015.636118	**SRR12113391**	246.9710907
ERR011119	1076.946627	**SRR12113392**	339.6117245

SRR12113393	352.1269012
SRR12113394	321.6427325
SRR12113395	418.1596268
SRR12113396	377.1800514
SRR14120361	591.2751764
SRR14120362	445.9665273
SRR14120363	561.6268074
SRR14120364	424.1244064
SRR14120365	570.6201879
SRR14120366	536.0160297
SRR14120367	566.9630724
SRR14120368	946.7475093
SRR14120369	816.844023
SRR7343998	548.8369099
SRR7343999	912.6856532
SRR7344000	633.6516621
SRR7344001	493.3900087
SRR7344002	642.6062688
SRR7344003	604.0576631
SRR8670866	1058.871699
SRR8670867	752.3616984
SRR8670868	756.7640078
SRR8670869	753.8228702
SRR8670870	921.3062904
SRR8670871	1052.149052
SRR8670872	680.6465114
SRR8670873	681.6287783
SRR8670874	626.1196113
SRR8670875	629.3845766
SRR8942237	655.7977718
SRR8942238	637.4625067
SRR8942239	684.9785508
SRR8942240	780.8181267
SRR8942241	1287.634644
SRR8942242	714.258227
SRR8942243	638.3425622
SRR8942244	637.3691817
SRR8942245	574.4791096
SRR8942246	631.9136408
SRR8942247	544.3189671
SRR8942248	558.8185231
SRR8942249	528.3766871
SRR8942250	481.6622421
SRR8942251	640.7214345
SRR8942253	606.7563246
SRR8942254	510.7811635
SRR8942255	590.5979952
SRR8942256	558.5724822
SRR8942257	766.013276
SRR8942258	866.9708646
SRR8942259	640.7610719
SRR8942260	587.5475593
SRR8942261	588.8670593
SRR8942262	636.7960048
SRR8942263	1139.04741
SRR8942264	825.7872716
SRR8942265	611.8951724
SRR8942266	710.091513
SRR8942298	408.6899021
SRR8942301	635.4897409
SRR8942302	589.5883132
SRR8942303	382.545907
SRR8942304	417.1339244
SRR8942307	525.9584476
SRR8942308	574.3716948