COMPARATIVE ANALYSIS OF ERROR CORRECTION IN HIGH-THROUGHPUT SEQUENCES FOR THE HUMAN GUT MICROBIOME

An Undergraduate Research Scholars Thesis

by

NATHAN PURWOSUMARTO

Submitted to the LAUNCH: Undergraduate Research office at
Texas A&M University
in partial fulfillment of requirements for the designation as an

UNDERGRADUATE RESEARCH SCHOLAR

Approved by
Faculty Research Advisor: Dr. Sing-Hoi Sze

May 2023

Major: Computer Science

RESEARCH COMPLIANCE CERTIFICATION

Research activities involving the use of human subjects, vertebrate animals, and/or biohazards must be reviewed and approved by the appropriate Texas A&M University regulatory research committee (i.e., IRB, IACUC, IBC) before the activity can commence. This requirement applies to activities conducted at Texas A&M and to activities conducted at non-Texas A&M facilities or institutions. In both cases, students are responsible for working with the relevant Texas A&M research compliance program to ensure and document that all Texas A&M compliance obligations are met before the study begins.

I, Nathan Purwosumarto, certify that all research compliance requirements related to this Undergraduate Research Scholars thesis have been addressed with my Faculty Research Advisor prior to the collection of any data used in this final thesis submission.

This project did not require approval from the Texas A&M University Research Compliance & Biosafety office.

TABLE OF CONTENTS

Page

ABSTRACT

Comparative Analysis of Error Correction in High-Throughput Sequences for the Human Gut Microbiome

Nathan Purwosumarto
Department of Computer Science & Engineering
Texas A&M University

Faculty Research Advisor: Dr. Sing-Hoi Sze
Department of Computer Science & Engineering
Texas A&M University

With the development of high-throughput sequencing tools over the last few decades, the sequencing of genomic data at a large scale at a relatively low cost has drastically revolutionized the field of bioinformatics. Next-generation sequencing tools, such as the Illumina suite of bridge amplification sequencing technologies, can generate billions of base pair reads per experiment. However, one drawback of these tools is that they produce a lot more errors than early sequencing methods. While error rates may seem to be quite low on paper, they are compounded by the large number of bases sequenced. Since these errors have the potential to confuse analysis and further results within bioinformatics pipelines, many tools have been developed to mitigate this issue. The traditional method is to use clustering & denoising techniques to mitigate the error, but there have been a variety of software that also look at reducing error through correction using alternative methods, such as k-mer analysis. This project looks at the traditional method of error correction of high throughput sequencing using clustering & denoising and seeing if non-

standard error correction models can be included in addition to the traditional pipeline to obtain better results.

As the entire field of high-throughput sequencing is very large, a focus will be placed on error correction in bacterial taxonomic classification. For this project, taxonomic classification for the human gut microbiome will be studied, using the 16S rRNA gene as the target sequence due to its ubiquity and importance in bacterial taxonomic classification. This gene is a highly conserved sequence among most prokaryotes, serving a fundamental role in protein synthesis across bacterial species. Differences within this sequence allow for the analysis of taxonomic composition within bacterial communities, which will be analyzed in the context of the species residing within the human gut microbiome. Existing sequences that have known taxonomic composition for the human gut microbiome will be used with different error correction methods as part of an *in silico* pipeline using the bioinformatics platform QIIME2. This project builds off previous research in the field, studying their methodologies and differences to address the problem of errors arising during sequencing. The human gut microbiome was chosen due to recent studies finding that the diversity of the gut microbiome has been increasingly linked with a variety of overall health conditions. A contrastive approach will be taken to identify the differences between error correction and traditional taxonomic classification methods to determine whether increased taxonomic classification can be obtained with error correction on sequences for the human gut microbiome, focusing on the differences that error correction software can make at the genus and species level.

ACKNOWLEDGEMENTS

Contributors

I would like to thank my faculty advisor, Dr. Sing-Hoi Sze, for their guidance and support throughout the course of this research and for being an amazing professor for Discrete Structures and Computational Techniques for Evolutionary Analysis.

Many thanks also go to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University an uplifting and learning experience. Special thanks also go out to Dr. Jyotikrishna Dass, for being my first research mentor and introducing me to the research process after taking his Computer Architecture course at Texas A&M University. Additional thanks also go out to my friends and faculty I met during my semester abroad as part of the Swansea Semester Exchange, and the staff of Texas A&M Education Abroad & Swansea Go Global for making this study abroad a wonderful undergraduate experience.

The data used for Comparative Analysis of Error Correction in High-Throughput Sequences for the Human Gut Microbiome were provided by the open-source library mockrobiota curated by Bokulich et al. (2016). Portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing, specifically the GRACE supercomputer cluster.

Funding Sources

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

1. INTRODUCTION

1.1 Overview of DNA Sequencing

Since DNA was discovered to be the genetic material that formed the building blocks of life, the constraint for genomic research has been the ability to read, or sequence, DNA on a large scale. As researchers seek to understand how variations in DNA can lead to such a large diversity in observable forms of life, sequencing DNA has been a large building block to the modern-day field of bioinformatics, driven by initiatives such as the Human Genome Project. As DNA is formed by a mere four repeating nucleotide bases: adenine, guanine, cytosine and thymine, the length of unique meaningful DNA sequences becomes quite long. These sequences that form specific coding regions, which when combined form the basis of genes, are of critical importance to individual function within any living organism. These regions also do not have a uniform size, from being a few hundred to even millions of base pairs in length for more complex genes. Because DNA bases are also smaller and more similar to each other than the amino acids that make up protein sequences, being able to read the order of these DNA bases was one of the first challenges in the field of bioinformatics, and an important precursor to many studies within the field.

The first major breakthrough in DNA sequencing took place many decades after the discovery of DNA as the source of genetic material, in the form of Sanger's ‘chain-termination’ method developed in 1977; which used radioactive markers on dideoxynucleotides (ddNTPs) that terminated the DNA replication process by lacking the 3’ hydroxyl group required for further extension of the DNA strand (Heather and Chain 2016). This allowed one to amplify the desired DNA sequence and terminate it at each possible base, allowing for the original sequence to be determined through gel electrophoresis.

1.1.1 Next-generation Sequencing

Nowadays, the rise of high throughput sequencing has revolutionized the field of bioinformatics by drastically reducing the time and cost requirements necessary to process large numbers of DNA samples. These technologies include emulsion PCR based sequencing, bridge amplification sequencing, Oxford Nanopore sequencing, and PACBIO sequencing. Next generation sequencing technologies allow for the reading of billions of DNA base pairs, being able to process whole genomes at a time in a single experiment. A long way from the original Sanger sequencing method that worked on sequencing fragments in the magnitude of thousands of base pairs, these technologies have revolutionized studies across various fields, including genomics and transcriptomics. At the base level, the most significant change that these technologies provide is that they allow for the parallelization of the sequencing process, where millions of sequences can be read simultaneously. While the Human Genome Project took billions of dollars and more than a decade to sequence the first human genome, sequencing human genomes nowadays using next-generation sequencing tools can be done within a few days at a fraction of the cost. Next-generation sequencing has also had profound implications for the medical field, as targeted genomic analysis has the potential to find specific mutations that are localized to a single individual.

1.2 Error correction in High-throughput Sequencing

While high-throughput sequencing methods are much more cost and time effective than the earlier generation of sequencing technologies, they are much more prone to error due to the sheer scale of sequencing output. Currently, one of the most used next-generation sequencing tools is the Illumina suite of sequencing platforms, such as the HiSeq 4000, which use bridge amplification sequencing. For these Illumina machines, the estimated error rate on the machines is ∼0.1–1 × 10^-2 per base sequenced (Hussmann et al. 2013). Since the scale afforded by high-throughput sequencing allows for hundreds of thousands of bases to be read in a relatively short amount of time, the accumulation of these errors has the potential to confound the results of an experiment and affect the resulting conclusions.

There have been many different approaches that have attempted to address the problem of errors arising in the results of high-throughput sequencing. Traditional methods used to account for error correction when conducting taxonomic classification are de-noising and clustering techniques (Johnson et al. 2019). However, with the rise of generic error correction libraries in high-throughput sequencing, with Heo et al. (2021) evaluating 17 different error correction tools for Illumina sequencing alone, one must consider the effectiveness of these tools for taxonomic classification. Since there has not been much literature to compare these tools with the traditional methods of denoising and clustering in specific real-world use cases, this project hopes to explore the advantages and disadvantages of including an error correction library as part of the taxonomic classification pipeline in addition to the traditional method of dealing with error in the field of taxonomic classification in the human gut microbiome.

1.3 Importance of the Human Gut Microbiome

The human gut microbiome has been a diverse and extensive research area in the fields of bioinformatics and medicine, and has been shown to be increasingly linked to a plethora of biological functions and overall health. While many instances of bacteria within the human body are commonly seen as pathogens that need to be eliminated by the immune system, the majority of bacteria that live in the gut are non-pathogenic and live in a symbiotic relationship, providing essential functions that include digestion of nutrients, drug metabolism, crowding out and preventing the colonization of pathogenic microorganisms in the gut, supporting intestinal barrier function, and even partial integration with the human immune system to fight off invasive pathogens or other microorganisms (Jandhyala et al. 2015). The diversity of the gut microbiota seems to play a major factor with regards to health conditions as well, as lower bacterial diversity has been observed in people with inflammatory bowel disease, psoriatic arthritis, type 1 diabetes, atopic eczema, coeliac disease, obesity, type 2 diabetes, and arterial stiffness, when compared to healthy controls (Valdes et. al. 2018). Therefore, improvements in the ability to properly identify the specific bacterial communities residing in the human gut microbiome have the potential to revolutionize diagnosis and treatment for a variety of conditions. Currently, taxonomic analysis of the bacterial communities composing the human gut microbiome have been limited to taxonomic features that exist on the genus-level (Yang et al. 2020). At the species level, bacteria within the same genus share a very large amount of identical DNA, making them extremely difficult to distinguish. Furthermore, the discovery of new species that have not been previously categorized is not an uncommon occurrence.

1.4 Taxonomic Classification

Taxonomic classification experiments involve analyzing DNA samples from a target region or sequence to obtain information on the target organisms being sequenced. This can be done to determine the evolutionary relationships between the target organisms being sampled. If DNA samples of known composition are taken from a population of similar organisms, the similarities and differences found between them allow for the creation of taxonomic hierarchies such as phylogenetic trees, which help in determining which organisms are most closely related. This allows researchers to understand the chronology of the DNA changes observed, determining which mutations took place earlier than the others based on the evolutionary relationships in the tree. Taxonomic classification can also be done to determine the origins of unknown genetic sequences by comparing them to existing reference genomes, classifying the unknown sequences to their most probable origin. Additionally, taxonomic classification techniques can also be applied to determining the origins of an unknown DNA sequence that does not have a reference genome by comparing the unknown sequence to DNA sequences from other organisms that show high similarity. This is used in conjunction with de novo sequencing and is particularly useful for organisms that have not been previously sequenced, to determine what other organisms it is related to, especially important in the discovery of new species.

1.4.1 The 16S rRNA Sequence as a Taxonomic Marker

Taxonomic classification in the human gut microbiome involves analyzing a sample from a bacterial community found in the human gut to determine its bacterial composition. The DNA sequences taken will be of unknown composition, but the resulting reads will be compared to a library of the common bacteria residing in the human gut microbiome to determine the taxonomic composition of the sample. Each unique DNA sequence found in the sample is called an amplicon sequence variant (ASV) and is labeled with a predicted organism based on the reference genomes of the common human gut bacteria. The target DNA region most commonly used for taxonomic composition of the human gut microbiome is the 16S rRNA gene. Since the 16S rRNA sequence codes for the main component of the small ribosomal subunit in prokaryotes, it is highly conserved across species, as ribosomes are necessary to produce all proteins within an organism. Therefore, the 16S rRNA sequence has been commonly used in the study of bacterial phylogeny and taxonomy, and its critical function has not changed over time, which allows for a comparison of minute differences as a result of evolutionary mutation (Janda and Abbott 2007). However, since this gene serves the same function across different species, comparative analysis of the whole sequence between different organisms will result in a very high level of similarity. Thus, specific hypervariable regions within the 16S gene that have a higher prevalence of differences between species are more often used to distinguish the composition of bacterial communities. One of these hypervariable subregions, V4, is only a couple of hundred base pairs long and has been demonstrated to be able to differentiate phylogenetic relationships closest to those based on the full-length sequence, which is why it is used for many analyses of the human gut microbiome (Yang et al. 2016).

2. RELATED WORKS

2.1 Evaluation of Error Correction

The paper *Comprehensive assessment of error correction methods for high-throughput sequencing data* published by Heo et al. (2021) describes challenges with error correction for various types of high-throughput sequencing data and presents a solution to measure the performance of error-correction algorithms through a software package. The first challenge described is the difference in error correction between DNA and RNA reads, mentioning how non-uniform expression levels and alternative splicing make RNA reads a lot more susceptible to variation and thus may require a different approach. The second challenge lies with differences in the underlying process of data collection. Because of the variety of tools available for high throughput sequencing, there are differences in what errors are produced based on the sequencing technology itself. For example, the dominant form of errors created in reads taken from Illumina machines is substitutions, while the dominant form of errors from PacBio and Oxford Nanopore machines are insertions and deletions. As such, error correction methods would perform differently for each of the sequencing technologies, and some error correction algorithms were developed only for a certain type of sequencing machine. The third challenge and the one most directly addressed in the paper is the lack of tools and metrics available to test the accuracy of error correction methods against one another, even though there have been many papers proposing various algorithms for error correction. The paper attributes this issue to the difficulty in determining whether an error correction method generated new errors during the error correction process, making it unclear whether an error was fixed or a new one was generated.

To address these problems, Heo et al. created a software package called SPECTACLE to evaluate the performance of 23 different error correction tools, which can be applied to both DNA and RNA reads and is invariant to what sequencing technology was used to obtain the initial reads. This software works by applying each error correction method to reads taken from a mix of simulated and real data that have a known reference sequence and calculates the accuracy of error correction by each method by comparing the results to said reference sequence. The results showed that while some error correction technologies were benchmarked to be better than others in the experimental setup, each tool performed differently based on factors such as read coverage and how repetitive the sequenced genome was. However, this software package would be extremely useful in the creation of new error correction algorithms, as it provides a baseline on which to test new methods of error correction against preexisting algorithms.

2.1.1 Benchmarking based on unique molecular identifier sequencing

Another study conducted by Mitchell et al. (2020) also looks at benchmarking error correction methods by using a special type of sequencing, based on the unique molecular identifier (UMI) method to act as the error-free template. This protocol involves adding short random sequences to every molecule of DNA to serve as a molecular barcode, allowing for the identification of duplicate sequences during the amplification process. These markers allow for the correction of errors that arise by the consensus method, by comparing all the sequences with the same UMI barcode after amplification. The same input data is then sequenced normally, and error correction methods are applied to the resulting reads. The accuracy of each error correction method is computed by taking the error-corrected sequencing reads compared to the UMI sequencing reads. While UMI sequencing is not completely error free, it provides much higher accuracy than normal sequencing methods. For the study, the authors only used UMI clusters that had above 80% consensus on all nucleotides in the target segment and disregarded the rest, guaranteeing that the remaining UMI sequencing results are able to be used as the error-free template.

The error correction methods tested by Mitchell et al. were more limited in scope than the ones tested by Heo et al., as the study was limited to Illumina error correction tools. Between these tools, the effect of k-mer size on the accuracy of the error correction method was determined individually for each input dataset, and the best value was used. This study also goes further in-depth into explaining the differences between each error correction tool when the coverage size of the original datasets was modified. The general trend was that increased k-mer size and coverage depth improved the accuracy of error correction, regardless of the specific algorithm used. This makes logical sense as larger k-mers and increased coverage of the target sequence allow for error correction tools to have more information on the sequence as a whole and make more accurate decisions throughout the error correction process. Their paper also used viral sequencing data as part of the benchmark, in addition to the more common human and bacteria sequencing data. The results of the paper showed large variability across different types of datasets for each error correction method, with no single method significantly performing better than the others. However, the results also showed that error correction methods are able to achieve performance comparable to the UMI barcoding sequence data in the right conditions. Their conclusion from the study was that error correction tools may replace UMI sequencing in the future, as UMI sequencing increases the size of the DNA being sequenced and is more expensive than error correction software.

2.2 Taxonomy of the Human Gut Microbiome

The paper *Exploring the universal healthy human gut microbiota around the world* published by Piquer-Estaban et al. (2022) explores the history of taxonomic studies on the human gut microbiome, focusing on the differences between microbiomes throughout the world and determining a core taxonomy of bacterial species that roughly universal. Looking at research published in the field regarding the human gut microbiome, the authors found that this field is relatively new, with almost all studies conducted within the two few decades, starting with the Human Microbiome Project in 2007. They emphasize that the human microbiome has been found to be linked to host health for a variety of health conditions, making it a prime target of recent studies. However, one problem highlighted within the study was that most reference genomes for the human gut are mainly limited to western microbiomes, due in part to the higher levels of economic development in the western world. By comparing a large number of existing data from individual microbiome studies around the world, Piquer-Estaban et al. were able to determine a universal core of 20 bacterial genera. This taxonomy also accounts for the most abundant genera found within the gut microbiome itself, independent towards lifestyle and geographical differences, thus showing that they are crucial to a variety of biological functions within the human body. This is not to say that these differences are not important within the taxonomic composition of the gut, as they account for many of the other non-universal bacterial genera, but to highlight the existence of how some genera may be fundamental as part of co-evolution with the history of the human species.

3. METHODS

3.1 Bacterial Communities Used

Data for this project is taken from the open source *mockrobiota* library, a publicly available resource used for microbiome bioinformatics benchmarking using artificially constructed communities. This library was used because it consists of many different datasets curated together into one public repository with clear labels, allowing for specific communities with the desired target sequences to be obtained in one centralized repository. Each community within this library is also clearly labeled with their taxonomic composition, allowing for a comparison to be made between expected results and actual results. One advantage of *mockrobiota* is that the communities included in the library were based off real-world biological observations. By basing them off real-world observations, these mock communities represent real communities far more accurately and are a better benchmark than synthetically generated data.

Within the *mockrobiota* library, there were three groups of mock communities chosen as input data to be used for the taxonomic classification pipelines. The first group consisted of mock communities 13-15, which contained even amounts of purified genomic DNA from 21 bacteria strains generated by Kozich et al. (2013). The second group consisted of mock communities 20 and 22, which contained even amounts of purified genomic DNA from 20 bacterial strains generated by Gohl et al. (2016). The third group consisted of mock communities 21 and 23, also generated by Gohl et al. using the same bacterial strains from the second group, but with an uneven composition. These three groups of mock communities were chosen because they are the most recent within the *mockrobiota* library to emulate bacterial sequencing data using the 16S gene as the target gene and the V4 region as the target subfragment. These communities were also based on observations from the Human Microbiome Project, and resembled Illumina reads taken from biological samples regarding bacteria commonly found within the human gut. Thus, taxonomic classification of this data would closely resemble a real-world scenario of obtaining sequence data from the human gut microbiome.

3.2 High Level Overview

To be able to compare the effectiveness of error correction, two parallel bioinformatics pipelines were created. The first uses the traditional method of analyzing sequence data using denoising and clustering techniques, while the second is identical to the first save an error correction step on the raw sequence data taken from *mockrobiota*. The bulk of the pipeline was implemented using the QIIME2 software package, an open-source microbiome bioinformatics platform for sequence analysis. Some output file postprocessing was done using Kaiser Galaxy, a web-based data analysis platform. This workflow and the tools described above were run on the TAMU HPRC Grace supercomputer cluster, and the result of the pipeline was run through a Python script to calculate the taxonomic accuracy. Since the composition of the input sequences are labeled, accuracy computations in the final stage of the two analysis pipelines will provide a benchmark to analyze the effectiveness of error correction methods, to determine whether external error correction software made a difference in our classification results.

3.3 Importing Data

The first step involves taking the raw sequence data from *mockrobiota*, which comes in FASTQ format. This format includes the sequences obtained from each sample similar to FASTA format, but an additional character is included per base to encode the quality scores from the sequencing run. These quality scores are a critical component when conducting error correction, as they account for the confidence that bases are correctly determined. Typically, reverse reads have lower quality scores than forward reads to the nature of Illumina bridge amplification sequencing, as they are generated after the forward reads. Bases at the start or end of an individual sequence read tend to also have a lower quality score than those in the middle of the sequence. For the control pipeline, we take the sequence data as-is without further modification and import it directly into the QIIME 2 platform for further processing, using a manifest file to import into a QIIME 2 artifact with the datatype being paired-end sequences with quality. For the error correction pipeline, we run the Blue V2 error correction software, which is an error correction tool based on k-mer consensus and context (Greenfield et al. 2014). The executable binaries for Blue, and the preprocessing tool Tessel, which Blue relies on to generate the k-mer tiles for the sequence data, were downloaded and run on the raw sequence data. Since the Blue software was already created with an emphasis on bacterial genomes, the default parameters were used, with a k-mer length of 25. Running Tessel followed by Blue on the FASTQ files generated a set of error corrected FASTQ files, which were then also imported into QIIME 2 using the same procedure.

3.4 Analysis Pipeline

After the sequence data is imported into QIIME 2, we run the DADA2 plugin. This is used to generate the amplicon sequence variants (ASVs) from the sequence data, which are the inferred unique DNA sequences from the result of high-throughput sequencing. The DADA2 plugin denoises the sequences and creates the ASV feature table, assigning a unique id to each of these features. Another QIIME 2 tool is then used to run this feature table through a pretrained Naïve-Bayes classifier on the SILVA Database for 99% OTUs from the 515F/806R (V4) subregion, to determine the taxonomic predictions of these ASVs. QIIME 2 has pretrained

Naïve-Bayes classifier for both the Greengenes and SILVA reference databases trained on either full-length sequences or V4 subregion sequences. This project looks at the V4 subregion and thus uses the corresponding classifier, and the SILVA model was used because it is a newer reference database. The Naïve-Bayes classifier itself works by implementing a machine learning algorithm based on probabilistic calculations based on Bayes theorem. In taxonomic classification, the classifier determines the probability of a sequence belonging to a predicted organism by looking at the observed evidence, or unique features, within each sequence. This probability of a sequence being from a certain species is calculated based on the previous training examples used when the model was created. Since this project uses pretrained models, no computational overhead was spent on training the predictive classifiers. In addition, the results of the taxonomic classifiers can be visualized in a taxonomic bar plot if determining the bacterial composition of the sample, but this project only looks at the predictions themselves and whether they were correct.

The QIIME 2 portion of the analysis pipeline used for this project was run with the resources provided by Texas A&M High Performance Research Computing (TAMU HPRC), on the Grace supercomputer cluster. The QIIME 2 module is part of the software packages available on this cluster, and a command line interface was used to run each command of the pipeline. To streamline this process, batch files were created that allowed for multiple commands to be queued and run in succession using the integrated SLURM job scheduling system. Construction of these job files was done after each step of the pipeline was manually run to make sure each step of the pipeline had the desired input and output format, and that each command run resulted in the desired functionality. Each step within the pipeline was given a maximum time limit of a couple hours of processing time on the cluster, but most tools did not require more than an hour of computation time.

3.5 Computation of Taxonomic Accuracy

The results of the taxonomic prediction from the Naïve-Bayes classifier are then compared with the known composition of the *mockrobiota* community, the true taxonomy of the sampled data. This was done by creating a Python script using the Pandas library, with the taxonomic predictions loaded in as a Pandas dataframe. Some data pre-processing was used to convert the QIIME 2 artifacts and true taxonomy FASTA file from their file formats into a tabular format for the Python script. This step was done using Kaiser Galaxy, a web-based data analysis tool also provided by TAMU HPRC.

This project looks at the genus and species level predictions, since identifying the specific bacteria found in the human gut microbiome is the target of these experiments. Using the results obtained from the classifier in the pipeline, the Python script finds where the classifier made a prediction on the genus or species level. When the classifier did not have a high confidence, the prediction would be left blank, indicating that certain ASVs were not able to be identified at that taxonomic level, which were then labelled unknown. When the classifier made a prediction, it did so at a very high level of confidence, so these predictions are assumed to be correct if they match with an identified genus or species in the true taxonomic classification. If a prediction did not match an identified genus or species in the true taxonomic classification, then it was labelled as incorrect. The percentage of sequences identified was then computed by taking the number of correct and incorrect predictions divided by the total number of ASVs, while accuracy was computed by the number of correct predictions divided by the numbers of correct and incorrect predictions.

4. RESULTS

4.1 Pipeline Comparisons

Tables showing the final results of the two pipelines regarding taxonomic accuracy, with and without the error correction software on the raw sequence data are displayed below. The tables are broken down into genus level and species level, which can be looked at independently. For example, Table 1 shows the summary results of the two pipelines for mock community 13.

Table 1: Control and error correction pipelines on mockrobiota community 13

Mockrobiota 13 (Mock-13)	Control	Blue Error Correction
ASV Sequences Generated:	64	76
Unknown Genus:	1 (98.44% identified)	3 (96.05% identified)
Genus Correct:	57	67
Genus Incorrect:	6	6
Genus Accuracy:	0.90476	0.91781
Unknown Species:	39 (39.06% of sequences identified)	44 (42.11% of sequences identified)
Species Correct:	12	12
Species Incorrect:	13	20
Species Accuracy:	0.48	0.375

The results obtained for each community were highly variable at the species level, with the worst results to species level accuracy found with mock community 15, with -40%

percentage change after the error correction step. The best results to species level accuracy were found with mock community 20, going from 0% correct predictions for the control and 85.71% correct predictions after the error correction step. These summary statistics are shown in Table 2 and Table 3, respectively. The full set of summary statistics for each community are not shown for sake of brevity, but their overall results are displayed in the subsequent section.

Table 2: Control and error correction pipelines on mockrobiota community 15

Mockrobiota 15 (Mock-15)	Control	Blue Error-Correction
ASV Sequences Generated	53	75
Unknown Genus:	2 (96.23% of sequences identified)	2 (97.33% of sequences identified)
Genus Correct:	45	69
Genus Incorrect:	6	4
Genus Accuracy:	0.88235	0.94521
Unknown Species:	35 (33.96% of sequences identified)	45 (40.0% of sequences identified)
Species Correct:	10	10
Species Incorrect:	8	20
Species Accuracy:	0.55556	0.33333

Table 3: Control and error correction pipelines on mockrobiota community 20

Mockrobiota 20 (Mock-20)	Control	Blue Error Correction
ASV Sequences Generated	7	20

Unknown Genus:	1 (85.71% of sequences identified)	0 (100% of sequences identified)
Genus Correct:	6	20
Genus Incorrect:	0	0
Genus Accuracy:	0.88235	0.94521
Unknown Species:	6 (14.29% of sequences identified)	13 (35% of sequences identified)
Species Correct:	0	6
Species Incorrect:	1	1
Species Accuracy:	0	0.8571

4.2 Overall Results

The impact that error correction had on genus level accuracy is shown in Table 4. From the table it can be determined that error correction showed a slight improvement to accuracy, with only one dataset where accuracy decreased after error correction applied. The average improvement across all datasets was a ~2.1% increase for genus level accuracy.

Table 4: Genus-level accuracy statistics for all datasets

Dataset	Control	Blue Error Correction	Percentage Change
Mock-13	0.90476	0.91781	1.442371458
Mock-14	0.84286	0.90667	7.570652303
Mock-15	0.88235	0.94521	7.124157081
Mock-20	1	1	0
Mock-21	1	1	0

Mock-22	0.9667	0.9429	-2.46198407
Mock-23	0.9091	0.9333	2.66197338
Averaged Results	0.9293957143	0.9494128571	2.153780413

The impact that error correction had on species level accuracy is shown in Table 5. These results are more inconclusive than the genus level accuracy, as for all datasets in the first group, Mock-13 to Mock-15, species level accuracy decreased after error correction. However, for the other two groups, species level accuracy increased, most prevalent for the Mock-20 and Mock-21 communities where the control pipeline failed to predict any species correctly, but the error correction pipeline managed to achieve a reasonable 83-85% accuracy for species predictions in these datasets.

Table 5: Species-level accuracy statistics for all datasets

Dataset	Control	Blue Error Correction	Percentage Change
Mock-13	0.48	0.375	-21.875
Mock-14	0.42308	0.36667	-13.33317576
Mock-15	0.55556	0.33333	-40.00107999
Mock-20	0	0.8571	N/A
Mock-21	0	0.8333	N/A
Mock-22	0.5714	0.6666	16.66083304
Mock-23	0.25	0.2857	14.28
Averaged Results	0.32572	0.5311	63.05415694

The most promising results that error correction made was to the number of unique ASV's generated, shown in Table 6. Having more ASVs allows for the better identification of unique DNA sequences in the sample, which is an important step for taxonomic classification when determining the different bacterial species within the sample. Every dataset tested generated more ASVs after error correction, and an average percentage increase of 38% was found across all datasets.

Table 6: ASVs generated for all datasets

Dataset	Control	Blue Error Correction	Percentage Change
Mock-13	64	76	18.75
Mock-14	71	77	8.450704225
Mock-15	53	75	41.50943396
Mock-20	7	20	185.7142857
Mock-21	5	20	300
Mock-22	30	41	36.66666667
Mock-23	22	39	77.27272727
Averaged Results	36	49.71428571	38.0952381

4.3 Data Patterns

The results obtained from these experiments seem promising, as the error correction on raw sequence data resulted in more unique ASVs being generated during the DADA2 noise correction and feature table creation step. An increase in ASVs found in this step show that the error correction helped to distinguish more defining features from the sequence data, which has the potential to better identify taxonomic variety. For example, in the Mock-13 community, the DADA2 step only managed to be able to generate 64 unique ASVs in the control pipeline, while it generated 76 ASVs in the error correction pipeline, a significant improvement of 18.75%. In addition, the accuracy of the genus-level predictions showed a slight improvement for all datasets tested, showing that error correction at this level of taxonomic classification was able to improve the genus-level predictions by a slight margin.

However, the results show that this pipeline did not improve the taxonomic classification at the species level as the results were highly variable, with the results for the first group of mock communities showing a decrease in accuracy after error correction. For this group of communities, the classifier did not see a benefit in having the sequences error-corrected for the sequence level taxonomic classification. Looking at the actual outputs and predictions of the classifier it seems that since there are more ASVs after the error correction step, the classifier is making the same classification mistake more times, which explains why accuracy decreased. For example, if there are two ASVs predicted to be *Staphylococcus carnosus* instead of *Staphylococcus aureus* in the control pipeline, there could be three or four of them in the error correction pipeline, and thus the classifier can be seen to be making more mistakes.

4.4 False Negatives and Recall

For the previous results, accuracy was computed based on the correct predictions made by the classifier, which can be seen as true positives, divided by the total number of predictions made by the classifier, where the incorrect predictions made by the classifier can be seen as false positives. To better understand the result of error correction we can determine a false negative metric, which we take as the number of unique genera or species in the true taxonomic composition of each sample that are not predicted as a result of the ASVs generated for each community. For example, if *Bacillus cereus* exists in the true taxonomic composition of a mock community, but the classifier never made a prediction for *Bacillus cereus* for any of the ASVs in that run, then it is considered a false negative. These numbers are shown below in Table 7, where it can be seen that error correction made a slight improvement on both the genus and species level.

Table 7: False negatives for all datasets

Dataset	Control (Genus)	Control (Species)	Blue Error Correction (Genus)	Blue Error Correction (Species)
Mock-13	0	12	1	11
Mock-14	0	13	1	11
Mock-15	1	12	1	12
Mock-20	14	20	2	13
Mock-21	17	20	5	14
Mock-22	1	17	1	16
Mock-23	6	19	5	18
Averaged Results	5.571428571	16.14285714	2.285714286	13.57142857

Additionally, we can calculate the recall for each dataset based on the number of true positives divided by the total sum of true positives and false negatives. The recall statistic for the genus level is shown in Table 8, which showed a large improvement on communities 20, 21, and 23, but a small decline on communities 13 and 14.

Table 8: Genus level recall statistics for all datasets

Dataset	Control	Blue Error Correction	Percentage Change
Mock-13	1	0.9853	-1.47
Mock-14	1	0.9855	-1.45
Mock-15	0.9783	0.9857	0.7564141879
Mock-20	0.3	0.9091	203.0333333
Mock-21	0.15	0.8	433.3333333
Mock-22	0.9667	0.9706	0.4034343643
Mock-23	0.7692	0.8485	10.30941238
Averaged Results	0.7377428571	0.9263857143	25.57027226

The recall for the species level is shown in Table 9. Interestingly, the recall for the species level improved after error correction for all datasets and had a larger average improvement than the recall statistic on the genus level.

Table 9: Species level recall statistics for all datasets

Dataset	Control	Blue Error Correction	Percentage Change
Mock-13	0.5	0.5217	4.34
Mock-14	0.4583	0.5	9.098843552
Mock-15	0.4545	0.4545	0
Mock-20	0	0.3158	N/A
Mock-21	0	0.2632	N/A
Mock-22	0.1905	0.2	4.98687664

Mock-23	0.05	0.1	100
Averaged Results	0.2361857143	0.3364571429	42.45448497

With the recall statistic we can calculate the resulting F1 scores, since the accuracy measurements earlier are equivalent to precision in the absence of a true negative class. The F1 scores for the genus and species level are shown in Table 10 and Table 11.

Table 10: Genus level F1 score for all datasets

Dataset	Control	Blue Error Correction	Percentage Change
Mock-13	0.94999895	0.9503583009	0.03782645102
Mock-14	0.9147303648	0.9444429253	3.248231578
Mock-15	0.9278510252	0.9650304748	4.007049468
Mock-20	0.4615384615	0.952385941	106.3502872
Mock-21	0.2608695652	0.8888888889	240.7407407
Mock-22	0.9667	0.9565495061	-1.050014881
Mock-23	0.8333190967	0.8888820855	6.667672569
Averaged Results	0.7592867805	0.9352197318	23.17081712

Table 11: Species level F1 score for all datasets

Dataset	Control	Blue Error Correction	Percentage Change
Mock-13	0.4897959184	0.4363499498	-10.91188525
Mock-14	0.4399863033	0.423079142	-3.842656275
Mock-15	0.4999742986	0.3845968927	-23.07666737

Mock-20	0	0.4615434905	N/A
Mock-21	0	0.4000447971	N/A
Mock-22	0.2857374984	0.3076852066	7.681073825
Mock-23	0.08333333333	0.1481462276	77.77547317
Averaged Results	0.256975336	0.3659208152	42.39530567

5. CONCLUSION

5.1 Role of Error Correction

In conclusion, the results of this project show that error correction software can be used on sequence data to better identify unique ASVs from that data in the context of taxonomic classification for the human gut microbiome for the samples tested. Each mock community was found to have a higher number of ASVs generated after running DADA 2 on the error-corrected sequences than compared to the control sequences. Genus level prediction accuracy showed some slight improvements, but sequence level prediction accuracy dropped because of the classifier model still not being able to distinguish between bacterial species of the same genus, a problem that still pervades the field of human gut taxonomic identification. The recall scores for species and genus level also showed some improvement after error correction, showing that error correction can generate less false negatives, predicting some genera and species that went unpredicted in the control. The F1 score combines the accuracy and recall statistics, supporting the claim that error correction made a difference at the genus level, as error correction improved this metric on all datasets, but is inconclusive at the species level due to the high level of variability.

The contrastive approach used in this paper also shows that error correction tools can be readily utilized directly with the traditional methods of taxonomic classification. Integrating the error correction step into the analysis pipeline showed that such a step could be used to improve results with little additional overhead. Potentially, such error correction software can also be made into a plugin to better synchronize with bioinformatics platforms such as QIIME 2. In this way error correction to process raw sequence data, as a preprocessing step before the generation of feature tables and taxonomic prediction, can be a valuable step in sequence analyses.

5.2 Future Work/Improvements

Testing different classifiers with the conjunction of error correction software in taxonomic classification seems to be a promising direction for future work. Since the pretrained Naïve-Bayes classifier as part of the QIIME 2 feature classifier tool is based on the entirety of the SILVA bacterial database, training one to be fine-tuned on the specific species one expects to find in the human gut microbiome would increase the accuracy of the species level identification, as certain bacteria species within the same genus are more likely to inhabit the gut microbiome. Additionally, a different prediction architecture could be used entirely, trying different machine learning models that may be more accurate and not solely based on probabilistic calculation using Bayes Rule. Some other supervised learning models used in classification include the k nearest neighbour classifier or linear regression-based classifiers. While Naïve-Bayes classifiers are commonly used, and quite accurate when individual features are independent, this assumption may not hold true at the species level, where similar species within the same genus have an overlap of defining features.

Furthermore, different error correction software could be tested on the raw sequence data, or additional pre-processing steps included before importing the sequences into QIIME 2. Another option that could be tried is instead of using DADA 2 to generate the ASV feature table, the Deblur plugin could also be tried, which also generates an ASV feature table using a different algorithm for quality control. However, these changes to the analysis pipeline would not be as impactful as making improvements to the classifier model, as that is what needed to be improved the most.

REFERENCES

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., Huttley, G. A., & Gregory Caporaso, J. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome, 6(1), 90. https://doi.org/10.1186/s40168-018-0470-z

Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., Mills, D. A., & Caporaso, J. G. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nature Methods, 10(1), 57–59. https://doi.org/10.1038/nmeth.2276

Bokulich, N. A., Rideout, J. R., Mercurio, W. G., Shiffer, A., Wolfe, B., Maurice, C. F., Dutton, R. J., Turnbaugh, P. J., Knight, R., & Caporaso, J. G. (2016). mockrobiota: a Public Resource for Microbiome Bioinformatics Benchmarking. MSystems, 1(5). https://doi.org/10.1128/mSystems.00062-16

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., … Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology, 37(8), 852–857. https://doi.org/10.1038/s41587-019-0209-9

Edgar, R. C. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. BioRxiv, 081257.

Gohl, D. M., Vangay, P., Garbe, J., MacLean, A., Hauge, A., Becker, A., Gould, T. J., Clayton, J. B., Johnson, T. J., Hunter, R., Knights, D., & Beckman, K. B. (2016). Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. Nature Biotechnology, 34(9), 942–949. https://doi.org/10.1038/nbt.3601

Greenfield, P., Duesing, K., Papanicolaou, A., & Bauer, D. C. (2014). Blue: correcting sequencing errors using consensus and context. Bioinformatics, 30(19), 2723–2732. https://doi.org/10.1093/bioinformatics/btu368

Hawkins, J. A., Jones Jr, S. K., Finkelstein, I. J., & Press, W. H. (2018). Indel-correcting DNA barcodes for high-throughput sequencing. Proceedings of the National Academy of Sciences, 115(27), E6217-E6226.

Heo, Y., Manikandan, G., Ramachandran, A., & Chen, D. (2020). Comprehensive assessment of error correction methods for high-throughput sequencing data. arXiv preprint arXiv:2007.05121.

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics, 107(1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Janda, J. M., & Abbott, S. L. (2007). 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls. Journal of Clinical Microbiology, 45(9), 2761–2764. https://doi.org/10.1128/JCM.01228-07

Jandhyala, S. M., Talukdar, R., Subramanyam, C., Vuyyuru, H., Sasikala, M., & Nageshwar Reddy, D. (2015). Role of the normal gut microbiota. World journal of gastroenterology, 21(29), 8787–8803. https://doi.org/10.3748/wjg.v21.i29.8787

Johnson, J. S., Spakowicz, D. J., Hong, B. Y., Petersen, L. M., Demkowicz, P., Chen, L., ... & Weinstock, G. M. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nature communications, 10(1), 1-11.

Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K., & Schloss, P. D. (2013). Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. Applied and Environmental Microbiology, 79(17), 5112–5120. https://doi.org/10.1128/AEM.01043-13

Lou, D. I., Hussmann, J. A., McBee, R. M., Acevedo, A., Andino, R., Press, W. H., & Sawyer, S. L. (2013). High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. Proceedings of the National Academy of Sciences, 110(49), 19872-19877.

Medvedev, P., Scott, E., Kakaradov, B., & Pevzner, P. (2011). Error correction of high-throughput sequencing datasets with non-uniform coverage. Bioinformatics (Oxford, England), 27(13), i137–i141. https://doi.org/10.1093/bioinformatics/btr208

Mitchell, K., Brito, J. J., Mandric, I., Wu, Q., Knyazev, S., Chang, S., Martin, L. S., Karlsberg, A., Gerasimov, E., Littman, R., Hill, B. L., Wu, N. C., Yang, H. T., Hsieh, K., Chen, L., Littman, E., Shabani, T., Enik, G., Yao, D., … Mangul, S. (2020). Benchmarking of computational error-correction methods for next-generation sequencing data. Genome Biology, 21(1), 71. https://doi.org/10.1186/s13059-020-01988-3

Piquer-Esteban, S., Ruiz-Ruiz, S., Arnau, V., Diaz, W., & Moya, A. (2022). Exploring the universal healthy human gut microbiota around the World. Computational and Structural Biotechnology Journal, 20, 421–433. https://doi.org/10.1016/j.csbj.2021.12.035

Robeson, M. S., O’Rourke, D. R., Kaehler, B. D., Ziemski, M., Dillon, M. R., Foster, J. T., & Bokulich, N. A. (2021). RESCRIPt: Reproducible sequence taxonomy reference database management. PLOS Computational Biology, 17(11), e1009581. https://doi.org/10.1371/journal.pcbi.1009581

Shi, H., Schmidt, B., Liu, W., & Müller-Wittig, W. (2010). A parallel algorithm for error correction in high-throughput short-read data on CUDA-enabled graphics hardware. Journal of computational biology : a journal of computational molecular cell biology, 17(4), 603–615. https://doi.org/10.1089/cmb.2009.0062

Valdes, A. M., Walter, J., Segal, E., & Spector, T. D. (2018). Role of the gut microbiota in nutrition and health. BMJ, k2179. https://doi.org/10.1136/bmj.k2179

Yang, B., Wang, Y., & Qian, P. Y. (2016). Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. BMC bioinformatics, 17, 135. https://doi.org/10.1186/s12859-016-0992-y

Yang, J., Pu, J., Lu, S., Bai, X., Wu, Y., Jin, D., Cheng, Y., Zhang, G., Zhu, W., Luo, X., Rosselló-Móra, R., & Xu, J. (2020). Species-Level Analysis of Human Gut Microbiota With Metataxonomics. Frontiers in Microbiology, 11. https://doi.org/10.3389/fmicb.2020.02029