Spring 5-1-2020

Identification and Analysis of Feather Degrading Bacteria: A Search for Keratinase Genes

Nehal Navali
nehal.navali@uconn.edu

Identification and Analysis of Feather Degrading Bacteria: A Search for Keratinase Genes

Nehal Navali
Undergraduate Honors Thesis
May 2019

Hird Laboratory
Department of Molecular and Cell Biology
Univeristy of Connecticut

Table of Contents

Abstract

Over two million tons of feather waste is generated annually by the poultry industry, the majority of which goes into landfills due to the difficulty of degrading its major component keratin. Although a portion of feather waste is eliminated via incineration or chemical treatment, the use of Feather Degrading Bacteria (FDB) has been proposed as a cheap and eco-friendly alternative. FDBs have been consistently isolated from the feather microbiome of birds and contain genes coding for the specialized protein keratinase which is able to degrade feathers. By doing so, feather waste, which is rich in nutrients, can be repurposed as animal feed or fertilizer. More research into FDBs is needed to determine whether or not this process is viable on a large scale. In this study, I aimed to identify FDBs and conduct a search for keratinase genes within their genomes. Bacterial swabs were collected from Saltmarsh and Seaside Sparrows in the state of Connecticut which were then isolated and tested for feather degrading ability. Six FDBs were then selected, their genomes sequenced and the raw reads assembled. Genomes were taxonomically identified using NCBI BLAST and StrainSeeker and all six bacteria were consistently identified as *Bacillus pumilus*, a known FDB. Primers for keratinase genes were identified using Geneious and it was found that four of the six bacteria contained at least one of the primers. Three of the samples contained multiple primers in the same area suggesting the presence of a keratinase gene, however more research is needed for confirmation.

Introduction

The human body hosts over 100 trillion microbes, both beneficial and harmful (1). It is these communities of microbes which comprise the human microbiome. The definition of the term microbiome has been debated in recent years. Although some definitions refer to only the genomic component of the microbes, in most cases the microbiome is treated as a "micro biome". In this definition, the microbiome refers to the community of microorganisms (including bacteria, archaea, fungi and viruses), the genomes of these microorganisms and the environment which they inhabit. The combined genetic material of all organisms within the microbiome is then referred to as the Metagenome.

Microbiomes are abundant across all parts of the Earth from the soil beneath our feet to the inside of volcanoes and even the buildings we create. One area of microbiome study which has seen an explosive increase in research has been Host-Associated Microbiomes. Many complex organisms contain vast amounts of bacterial cells which in some cases can even outnumber their own eukaryotic cells (2). The trillions of bacteria in the human body contribute an estimated 3.3 million protein coding genes which is over 100 times the number of protein coding genes contained within our eukaryotic cells (2). By studying the bacterial communities within host organisms, it is possible to determine which microorganisms are living within a community, as well as what functions they provide. Using this information, methods can be developed to manipulate the microbiome for benefits to health and conservation.

The human microbiome has been implicated in a wide array of human diseases and health conditions. Gut microbes in particular play a significant role in metabolism (3), the gut-brain axis (3), and immune development (4). Changes in diet and the development of a healthier and more diverse microbial gut community can also improve the symptoms of cardiovascular disease (5),

inflammatory bowel disease (6), obesity (7), type 2 diabetes (8) and a variety of allergies (9). Overall, research into the microbiome has demonstrated how the microbes that inhabit living organisms can have a profound impact upon their health and well-being.

Study of microbes first began in the 1600s when Anton van Leuwenhoek first looked at cells beneath a microscope, however, study of the microbiome proved to be a challenge. Traditional microbial techniques such as bacterial cultures provided only a limited view of the microbiome as the vast majority of microbes are unable to be grown in a laboratory setting (10). In the 1990's, new DNA sequencing technologies were developed which created new ways to study the microbiome in greater detail. This technology made it possible to identify the unculturable majority of bacteria by sequencing short segments of their DNA and matching them to a database of known sequences (10). Common techniques involve sequencing marker genes such as the 16s rRNA gene, or short DNA segments which are unique to the organism from which they are derived. This made it possible to identify a microbe by its genetic data without having to sequence its entire genome (10). These developments led to the initiation of the first phase of the Human Microbiome Project in 2007 with the goals of identifying the microbes in the human microbiome and creating a database for researchers.

Although DNA sequencing greatly enhanced how we identify and study microorganisms, early techniques did not allow for understanding the functions of these microbe. This gave rise to the area of metagenomics which involves not only identifying the various microbes in a microbiome, but also understanding their function to the overall host or environment (11). The hope was that by documenting the role of microbes within the body, it could become possible to use the microbiome to control health and disease. This desire to not only characterize, but also understand the function of the microbiome led to the NIH initiating the second phase of the HMP

with the goal of understanding the role of the microbiome and the microbes within the microbiome on human health and disease (11).

The development of High Throughput Sequencing (HTS) has greatly enhanced our ability to quickly sequence large quantities of DNA and has facilitated Whole Genome Sequencing (WGS). Instead of sequencing small markers and comparing them to known databases, WGS compares much larger segments of the genome allowing for more accurate identifications (12). Additionally, having the ability to assemble the complete genomes of bacteria has enabled researchers to more closely study microorganisms and create annotations with functions for the various parts of the genome (12). This in turn has enhanced our understanding of the microbiome and the microbes which inhabit it.

Currently, there are several methods for whole genome sequencing. Most projects, however, use shotgun sequencing: a method involving breaking the genome into a collection of small DNA fragments which are sequenced individually. Programs have been developed to assemble this collection of small reads through a process known as *de Novo* assembly. *De Novo* assembly algorithms are particularly useful because they assemble the genome using overlapping reads from within the sample thus eliminating the need for a reference genome. Many different tools exist for *de Novo* assembly such as ABySS (13), Soap De Novo (14) and Ray (15), however the program SPAdes is known to be the best for assembly of short bacterial genomes (16). SPAdes uses a de Bruijn algorithm and is capable of handling multiple types of sequence reads such as Illumina, Ion Torrent and PacBio (16).

The de Bruijn algorithm revolves around *k*-mers, or short DNA segments of length *k*. Each sequenced segment is broken down into *k*-mers of a specified length and overlapping sequences are placed next to one another creating a continuous sequence that will become the

assembly. An important aspect of this method of assembly is choosing the correct *k*-mer length. Choosing a value for *k* which is too high can cause a loss of coverage over the entire genome due to how specific the overlap must be. Conversely, choosing a *k*-mer value which is too low will increase the coverage, but will also increase the amount of errors in the assembly due to a decrease in overlap specificity. As a result, SPAdes uses a variety of *k*-mer values. By selecting multiple values between 20 and 80 bp in length, the program is able to produce the longest error-free contiguous sequences possible (16).

A number of software programs have been developed for the identification of genomic data. One of the most accessible and commonly used programs is NCBI BLAST (Basic Local Alignment Search Tool) (17). BLAST is an application which allows users to compare primary biological sequence data, such as nucleotide sequences, amino acid sequences, or marker genes such as the 16s rRNA gene, against a vast database of sequences to identify those which are similar to the query sequence (17).

Although any section of a genome can be "BLASTed" and identified, the 16s rRNA gene has become increasingly popular as a means to identify organisms. The 16s rRNA strand of RNA makes up part of the ribosomal scaffold and is vital for proper protein synthesis in the human body. Despite its importance, the 16s rRNA gene contains both highly conserved regions as well as nine hypervariable regions (V1-V9) ranging from 30-100 bp long (18). These regions allow the 16s rRNA sequence to model the evolutionary relationship between organisms as more similar hypervariable regions indicate a closer evolutionary relationship. A number of universal primers have been discovered which are common between many organisms. The most common primers are the 8F and 27F primers which cover the hypervariable regions V1-V3 and can be used to identify unknown genomes (19).

Alternative identification programs have also been developed to streamline the identification of bacteria. The BLAST-Like Alignment Tool (BLAT) is very similar to BLAST, however it significantly speeds up the search process at the expense of some sensitivity allowing for more matches (20). A newer program called StrainSeeker (21) is a command-line program which identifies raw genomic data without the need for a full assembly. Although the program only works on small bacterial genomes, it provides a quicker and less memory-intensive method of identification by cutting out the assembly step albeit at the potential cost of some accuracy (21).

The human skin microbiome is known to contain a diverse group of microbes which provide health and immune boosting benefits to the host (22). However, relatively little is currently known about the feather microbiome and how these microbial communities contribute to the health of birds. Birds play a vital role in ecosystems across the world. Feathers are exceptionally important to the anatomy of birds since not only do they allow birds to fly, but they also provide birds with insulation, camouflage from predators, and waterproofing (23, 24). Many of these functions mirror the functions of human skin, further making the feather microbiome an intriguing topic of study.

One group of microorganisms which has been consistently isolated from a variety of feather samples including the feathers of wild birds, the feathers of domesticated birds and even feather waste are Feather Degrading Bacteria (FDB). FDBs are microorganisms with the ability to degrade keratin, a fibrous structural protein which makes up over 90% of the composition of feathers (25). Keratin is extremely difficult to break down. It is insoluble in water, weak acids and inorganic solvents, and it is not broken down by common proteolytic enzymes such as trypsin or pepsin (26). A high number of disulfide bonds, hydrogen bonds, salt linkages and

cross linkages means that a specialized group of enzymes, keratinases, are necessary to degrade the protein (26). Keratinases vary greatly in their physical and chemical properties, however most are optimally active between pH 6-9 and temperature 40-60 degrees Celsius (27). Feather degradation is thought to occur in two steps. In the first step, disulfide bonds are broken using disulfide reductase (28). The second step then involves keratinase hydrolyzing the keratin and breaking the remaining bonds (29).

Bacteria of the *Bacillus* genus, specifically *Bacillus licheniformis*, are the most efficient known FDBs and, as a result, have been the target of the most study (30). However, many other bacteria such as *Stenotrophomonas*, *Pseudomonas*, *Brevibacillus*, *Fusarium*, *Geobacillus*, *Chryseobacterium*, *Xanthomonas*, *Nesterenkonia*, and *Serratia* have been shown to degrade feathers as have some species of fungi (30).

Feather waste has long been a by-product of the poultry industry and it is estimated that two million tons of feathers are generated annually (31). Due to poor waste management, the vast majority of feather waste either ends up in landfills or becomes a major pollutant due to how difficult feathers are to degrade (31). This waste often plays host to pathogens such as *Salmonella* and *Vibrio* and can also emit pollutants such as ammonia, nitrous oxide and hydrogen sulfide which can have a negative impact upon human health (32). Despite this, the main component of this waste, keratin, is a good source of peptides, amino acids and minerals such as nitrogen, potassium, calcium, magnesium, iron, manganese, zinc and copper (33). These nutrients alongside the desire to reduce pollution has increased interest in repurposing this feather waste into animal feed or fertilizer.

A portion of this waste is already processed through chemical treatment and steam pressure cooking (34). Although this treatment ultimately converts the waste into animal feed, it

costs a very large amount of energy and ultimately destroys many of the nutrients that would make the process worthwhile (35). As a result, it is not a viable long-term solution for the elimination of poultry feather waste. More recently, FDBs have been proposed as a cheap and eco-friendly alternative to chemical methods. By adding FDBs, which can produce keratinase, to feather waste, it can be degraded into nutrients to feed livestock. Although this theory has a great deal of merit, it is still to be seen whether or not the use of FDBs is practical in real-life application.

As an alternative to livestock feed, it has been suggested that feather waste could also be broken down into a nitrogen-rich fertilizer for plant growth. Many FDBs have been found to combat plant pathogens (36). As a result, introducing FDBs and feather waste into the agriculture industry allows them to act as a feather waste degrader, organic fertilizer, growth promoter and disease protectant all at once which can be hugely beneficial especially when compared with common chemical-based fertilizers and disease protectants (36).

The overall goal of this study was to assemble the genomes of FDBs, identify their species, and attempt to locate keratinase genes within the genome assemblies. Ventral and tail feather samples were collected from Saltmarsh and Seaside Sparrows throughout the state of Connecticut. Bacterial isolates were collected from each of the feathers, cultured, and then tested for keratinolytic activity. Of the bacteria found to be FDBs, six isolates were chosen to be sequenced. The sequences were analyzed for quality and assembled using the *de Novo* assembler SPAdes on the university cluster. The bacterial genomes were then identified using two different programs. NCBI BLAST was used on three separate parts of the assembled genomes: The first 10,000 base pairs of the longest node, the first 10,000 base pairs of the second longest node and the 16s rRNA sequence which was identified using the 8F primer. Secondly, the program

StrainSeeker was used on the raw genomic data to confirm the BLAST identification. After identification, a search was conducted using known keratinase primers in an attempt to determine if the assemblies contain genes for keratinase.

Methods

Sampling and Bacterial Isolation

Six bacterial isolates were selected from a previously established collection of FDBs taken from Saltmarsh and Seaside Sparrow feathers at sampling localities in Hammonnasset State Park in Connecticut, USA. Each isolate had been previously tested for its ability to degrade feathers by Elizabeth Herder, a graduate student in the Hird lab (Unpublished data).

DNA Extraction and Sequencing

DNA from the bacterial isolates was extracted using the QIAamp PowerFecal DNA (Qiagen, Hilden, Germany) kit. Manufacturer instructions were followed to produce a final volume of 100 µl. Sequencing occurred using a MiSeq Illumina machine at the Center for Genome Innovation (CGI) by Dr. Bo Reese.

Quality Control

Raw reads were first tested for quality using the R program DADA2 (37). A Phred score of 20 was used as the threshold below which read quality was considered too low for analysis. The quality control program Sickle (38) was then used to remove reads with very low Phred scores so as to include only reads with low error rates in the final assembly. The resulting reads were transferred to the UConn High-Power Cluster for assembly and analysis.

Assembly

Assembly of raw reads was done using the genome assembler SPAdes. As a *de Novo* assembler, scaffolds were created without the use of a reference genome using a de Bruijn algorithm. In this method, the reads were separated into segments of a specified length known as *k*-mers. These *k*-mers were then placed in a graph such that *k*-mers with overlapping reads were placed next to one another. These reads were then joined together to create the assembly. Assembly was performed using *k*-mers of length 22, 33, 55 and 77. These *k*-mer values were chosen based on program recommendation and included a combination of short and long length *k*-mers to offset problems with low coverage in the case of long *k*-mers, and potential errors in the case of short *k*-mers.

Identification

Identification of the scaffolds produced in the assembly was first done using NCBI Blast. Three analyses were conducted: [1] A BLAST search of the longest node assembled (Node 1), [2] a BLAST search of the second longest node assembled (Node 2), and [3] a BLAST search of the 16S rRNA region of the genome. The first 10,000 nucleotide bases from the top two nodes were blasted against the nucleotide collection database and the top results along with statistics were recorded. The 16S rRNA region was identified using the known 8F primer (AGAGTTTGATCCTGGCTCAG) for which a perfect match was found in all six assemblies. A BLAST search was performed on the ~30 base pairs following the primer which were aligned against the 16s ribosomal RNA database. The top three BLAST results from each method for each assembly were recorded.

Identification of Keratinase Primers

Keratinase primers were obtained from primary literature on the topics of keratinase and FDBs. Each primer is of a substantial length (at least 19 base pairs long) to reduce the probability of a type 1 errors. The collection was uploaded to Geneious (39), as were each of the 6 assembled scaffolds. The program Geneious was used to identify primers within the assembled genomes and indicate potential keratinase genes. Parameters were set such that matches must match the primer specifically. A spreadsheet consisting of each instance of the primer in the assembly was obtained from Geneious containing all instances in which the primer occurred as well as its location.

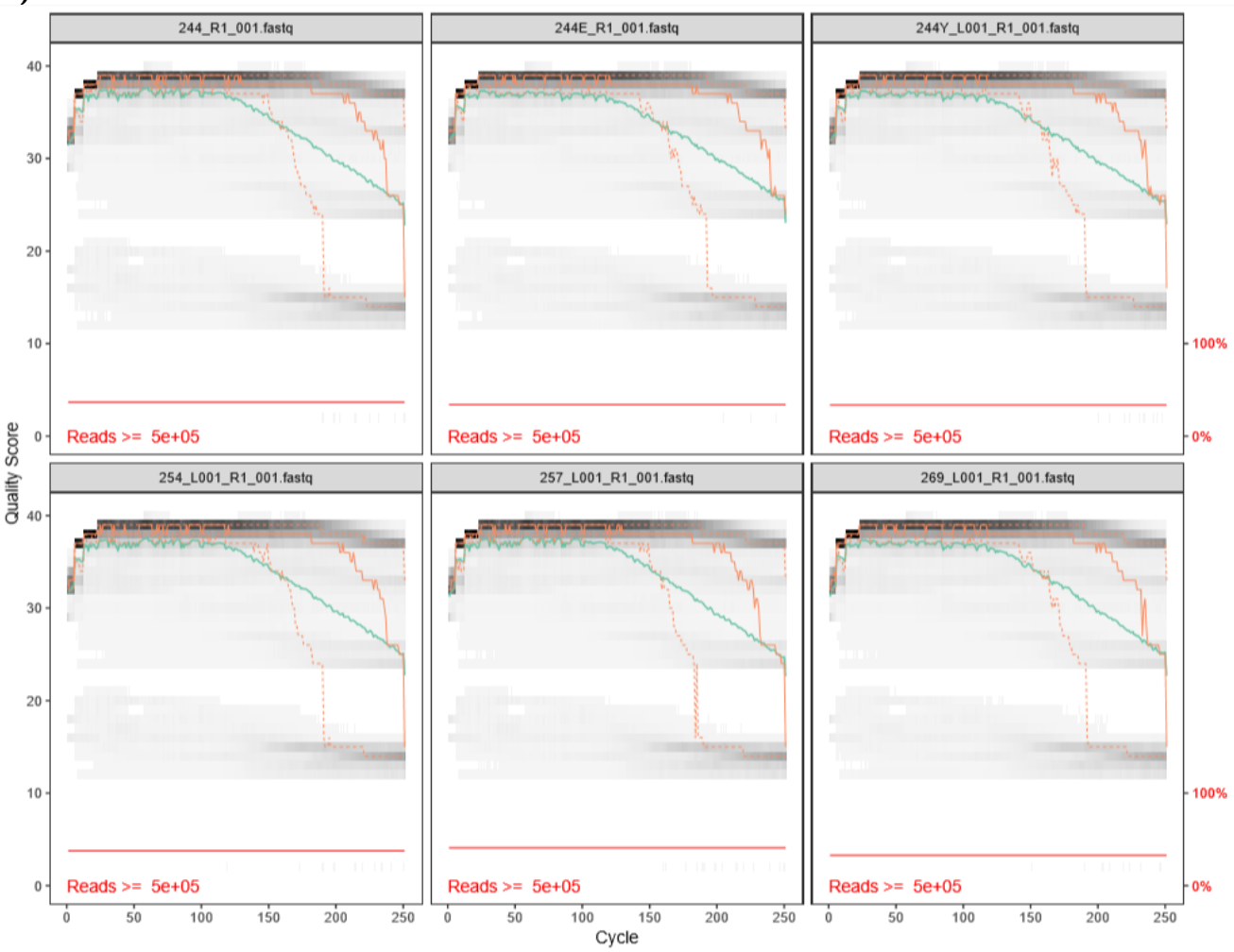
Table 1: Keratinase Primers

Primer Name	Sequence	Primer Length
XMAL (40)	CGACCCGGGATGATGAGGAAAAAGAGTTTTTGGC	34
XHOL (40)	CGACTCGAGTTGAGCGGCAGCTTCGACATTGAT	33
Pker (40)	GCACCCGGGGCCAAGCTGAAGCGGTCTATT	30
KPN (40)	GTCAGGTACCCACCCGTTTTACCTTCGCAGTTACT	35
Ker1 (41)	GCGCAAACCGTTCCTTACGGCATTCCTCTCATTAA AGCGGACAAAGTGCAGGCTCAAGGCTTTAAG	66
F1 Forward (41)	TTAGAAGCCGCTTGAACGTTA	21
F1 Reverse (41)	ATGTGCGTGAAAAAGAAAAATGTG	24
F2 Forward (41)	AAGTATTAGATCGTTACGGCGATGGAC	27
F2 Reverse (41)	CCAAGAACACCAATCGTGTTATCAAGG	27
BPI1 (41)	GAACACGACCCTAGCATTGC	20
BPI2 (41)	AGTACAGTACCAAGCAGTGG	20
BPE1 (41)	TCAGGTCTACTCTTATTTGC	20
BPE2 (41)	ATGATTCTCTCCATCATCG	19

Results and Discussion

Quality Control

A)



B)

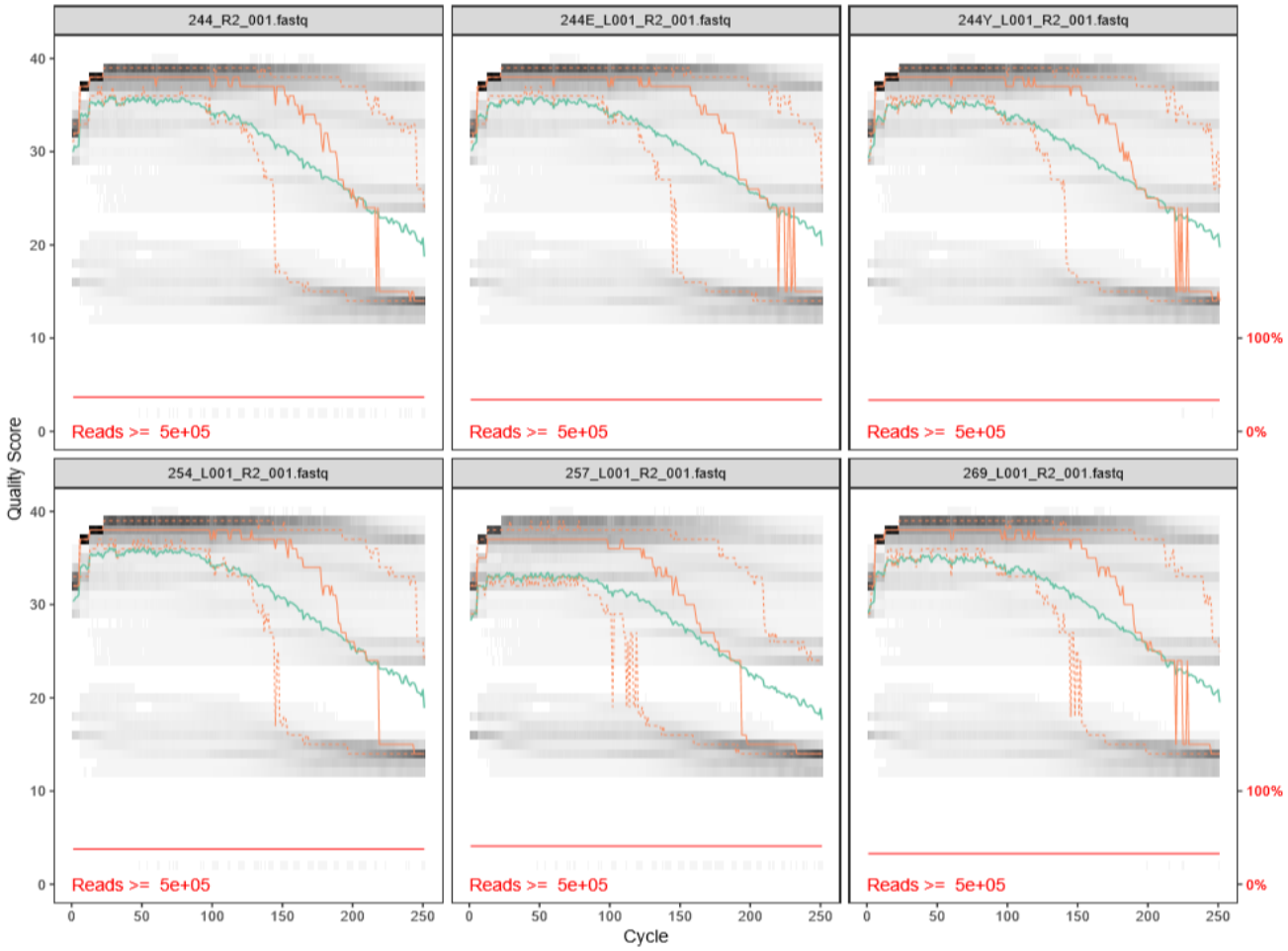


Figure 1: Quality plots representing the average Phred score of the reads over their entire length were developed for [A] the forward reads for each of the six bacterial samples and [B] the reverse reads for each of the six bacterial samples. The grey-scale heat map represents the frequency of each Phred score at a given position and the continuous green line indicates mean score.

Quality analysis was performed on raw reads using the R package Dada2. Quality plots were developed representing the Phred Score of the reads over their entire length. A Phred Score is a logarithmic value representing the prevalence of errors in a sequence. A Phred score of 10 will have an accuracy of 90%, a score of 20 will have an accuracy of 99% and so forth. Phred scores above a threshold of 20 typically indicate a sufficient accuracy for use in assembly and analysis and examination of the six plots shows that in general, the reads tend to be over 20 with the exception of the very end when it dips slightly below in some cases. Additionally, the reverse reads contain more errors than the forward reads over the entire sequence and especially towards the end, however, this is expected with illumina reads. Since some reads had Phred scores lower than 20, the reads were trimmed and the segments with extremely low quality scores were removed. Trimming was performed with sickle on the UConn HPC and the altered reads were ultimately used for analysis.

BLAST Identifications

Table 2: BLAST Identification Results - Sample 243

Section	Bacteria/Strain	Accession #	Query Cover	Percent Identity
Node 1	Bacillus pumilus strain SH-B9	CP011007.1	60%	89.87%
	Bacillus altitudinis strain DF48	CP025643.1	37%	94.69%
	Bacillus pumilus strain SF-4	CP047089.1	67%	86.20%
Node 2	Bacillus pumilus Strain 145	CP027116.1	99%	98.04%
	Bacillus pumilus strain ZB201701	CP029464.1	99%	93.00%
	Bacillus pumilus Strain PDSLzg-1	CP016784.1	99%	92.93%
16s rRNA	Anaerococcus lactolyticus Strain JCM 8140	NR_113565.1	100%	91.20%
	Anaerococcus degeneri Strain gpac104	NR_146834.1	100%	90.74%
	Anaerococcus mediterraneensis	NR147392.1	100%	89.35%

Table 3: BLAST Identification Results - Sample 244

Section	Bacteria/Strain	Accession #	Query Cover	Percent Identity
Node 1	Bacillus pumilus Strain 145	CP027116.1	100%	99.74%
	Bacillus pumilus Strain PDSLzg-1	CP016784.1	100%	95.59%
	Bacillus pumilus Strain ZB201701	CP029464.1	100%	95.54%
Node 2	Bacillus pumilus Strain 145	CP027116.1	93%	99.26%
	Bacillus pumilus strain SH-B9	CP011007.1	92%	92.11%
	Bacillus pumilus Strain SAFR-032	CP000813.4	92%	91.33%
16s rRNA	Bacillus zhangzhouensis Strain MCCC 1A08372	NR_148786.1	100%	99.59%
	Bacillus australimaris Strain MCCC 1A05787	NR_148787.1	100%	99.59%
	Bacillus pumilus Strain NBRC 12092	NR_113945.1	100%	99.59%

Table 4: BLAST Identification Results - Sample 247

Section	Bacteria/Strain	Accession #	Query Cover	Percent Identity
Node 1	Bacillus safensis Strain PgKB20	CP043404.1	45%	92.03%
	Bacillus pumilus Strain 145	CP027116.1	51%	94.53%
	Bacillus safensis IDN1	AP021906.1	44%	89.83%
Node 2	Bacillus pumilus Strain 145	CP027116.1	100%	99.53%
	Bacillus pumilus strain SH-B9	CP011007.1	96%	92.03%
	Bacillus pumilus Strain PDSLzg-1	CP016784.1	96%	93.06%
16s rRNA	Xylanibacillus composti Strain K13	NR_159902.1	100%	100.00%
	Bacillus marinisedimentorum Strain NC2-31	NR_159293.1	100%	100.00%
	Lactobacillus allii Strain WiKim39	NR_159082.1	100%	100.00%

Table 5: BLAST Identification Results - Sample 254

Section	Bacteria/Strain	Accession #	Query Cover	Percent Identity
Node 1	Bacillus pumilus Strain MTCC B6033	CP007436.1	100%	99.38%
	Bacillus altitudinis Strain GR-8	CP009108.1	100%	98.68%
	Bacillus altitudinis Strain W3	CP011150.1	100%	98.67%
Node 2	Bacillus pumilus Strain TUAT1	AP014928.1	100%	95.85%
	Bacillus cellulasensis Strain GLB197	CP018574.1	100%	95.76%
	Bacillus altitudinis Strain W3	CP011150.1	100%	95.71%
16s rRNA	Bacillus velezensis Strain FZB42	NR_075005.2	100%	100.00%
	Bacillus subtilus Strain 168	NR_102783.2	100%	100.00%
	Bacillus subtilus Strain IAM 12118	NR_112116.2	100%	100.00%

Table 6: BLAST Identification Results - Sample 257

Section	Bacteria/Strain	Accession #	Query Cover	Percent Identity
Node 1	Bacillus pumilus Strain 145	CP027116.1	99%	98.55%
	Bacillus pumilus Strain ZB201701	CP029464.1	99%	94.34%
	Bacillus pumilus Strain PDSLzg-1	CP016784.1	99%	94.34%
Node 2	Bacillus pumilus Strain 145	CP027116.1	99%	98.18%
	Bacillus pumilus Strain ZB201701	CP029464.1	99%	92.84%
	Bacillus pumilus Strain NCTC10337	LT906438.1	99%	91.44%
16s rRNA	Xylanibacillus composti Strain K13	NR_159902.1	100%	100.00%
	Bacillus marinisedimentorum Strain NC2-31	NR_159293.1	100%	100.00%
	Lactobacillus allii Strain WiKim39	NR_159082.1	100%	100.00%

Table 7: BLAST Identification Results - Sample 269

Section	Bacteria/Strain	Accession #	Query Cover	Percent Identity
Node 1	Bacillus pumilus Strain 145	CP027116.1	100%	98.80%
	Bacillus pumilus strain SH-B9	CP011007.1	100%	93.63%
	Bacillus pumilus Strain ZB201701	CP029464.1	96%	94.12%
Node 2	Bacillus pumilus Strain 145	CP027116.1	100%	98.80%
	Bacillus pumilus strain SH-B9	CP011007.1	100%	93.63%
	Bacillus pumilus Strain ZB201701	CP029464.1	96%	94.12%
16s rRNA	Bacillus zhangzhouensis Strain MCCC 1A08372	NR_148786.1	100%	99.83%
	Bacillus safensis Strain 100820	NR_113945	100%	99.83%
	Bacillus pumilus String NBRC 12092`	NR_115334.1	100%	99.83%

Results of the BLAST Identification search were indicated in tables 2-7. In five of the six bacteria BLASTed, *Bacillus pumilus* was the number one hit for both node 1 and node 2. In the sixth sample, sample 247, *Bacillus safensis* was the top hit for node 1, however *Bacillus pumilus* was the top hit for node 2. In most of these cases, *Bacillus pumilus* matched the genome with a high query cover, meaning the sequences matched along most of their length, and high percent identity, meaning that the segments which matched were very similar. The prevalence of *Bacillus pumilus* as the identification for these bacteria is very surprising since many prior studies have highlighted *Bacillus licheniformis* as the primary FDB found in birds. Recently however, studies have been released which have highlighted the role of *Bacillus pumilus* as an efficient feather degrading bacterium, especially for use with poultry waste (42), and have isolated the bacteria from feathers (43) which supports it as the identity of the six bacteria.

In addition to *Bacillus pumilus*, several other forms of *Bacillus* were identified as hits when BLASTing 10,000 bases of the two longest nodes. *Bacillus altitudinis* was suggested as one of the top hits for both node 1 and node 2 of Sample 254 (hit 2 and 3 for node 1 and hit 1 for node 2), and node 1 of Sample 243 (hit 3). For both node 1 and node 2 of Sample 254, the sequence has a very high identity (<95%) and complete query coverage (100%) indicating that it was a possible identity of the bacteria in addition to *Bacillus pumilus*. Conversely, the match with Sample 243 was found to have a high sequence identity of 94.69%, but only a query cover of 37% compared to the query cover of *Bacillus pumilus* were both above 60% indicating that it was not a reliable match to the genome. *Bacillus altitudinis* is a species of bacteria that was first isolated from air samples high in the air over Singapore (44). It was also isolated in lakes, soil and silt (44) which when combined with having been found in high altitudes may suggest methods by which it could be introduced into the feather microbiome of these sparrows. Another

bacterium which was commonly identified by blasting the longest two nodes was *Bacillus safensis*. This species was listed as a top hit of node 2 of Sample 254 (hit 2), and node 1 of sample 247 (hit 1 and hit 3). Both hits for Sample 247 node 1 have relatively low query covers (45% and 44%) indicating that they might not be reliable identifications. Conversely, the hit on node 2 of sample 254 had full query cover and over 95% identity meaning that it is more likely to be the true identity, however still unlikely due to the prevalence of *Bacilus pumilus*.

The BLAST results of the 16s rRNA region were far more varied than those of the nodes. This was likely in part due to the number of base pairs after the primer being much lower (~30) than those from the nodes. Ultimately, there was very little consistency among the top three 16s rRNA BLAST hits for any of the six isolates, and only in Sample 244 and Sample 269 was *Bacillus pumilus* among the top three hits. Sample 243 was particularly different since not only the top three hits, but also the top 100 hits all had the genus *Anaerococcus* with very strong association values. *Anaerococcus* is a genera of coccus shaped bacteria which are present in the human microbiome. As both nodes from this sample were identified as *Bacillus pumilus*, this identification is most likely incorrect. More likely, this identification is a result of one of several alternate explanations. One possible explanation is contamination of the bacterial sample. Since *Anaerococcus* is found in the human microbiome, the sample may have been contaminated during collection or testing. Another explanation is horizontal gene transfer from an *Anaerococcus* bacteria into *Bacillus pumilus*. Additionally, an error in sequencing may have occurred which altered the 16s region to more closely match that of *Anaerococcus*, however this is unlikely due to the strength of the alignment. Lastly, it could be due to an error in the database, however this is also unlikely due to the number of *Anaerococcus* matches.

Strain Seeker Identification

Table 8: StrainSeeker Identification Results - Sample 243

Read Direction	Bacteria/Strain
Forward	*Bacillus pumilus SAFR-032*
	Bacillus pumilus W3
	Bacillus pumilus GR-8
Reverse	*Bacillus pumilus SAFR-032*
	Bacillus pumilus W3
	Bacillus pumilus GR-8

Table 9: StrainSeeker Identification Results - Sample 244

Read Direction	Bacteria/Strain
Forward	*Bacillus pumilus SAFR-032*
	Bacillus pumilus W3
	Bacillus pumilus GR-8
Reverse	*Bacillus pumilus SAFR-032*
	Bacillus pumilus W3
	Bacillus pumilus GR-8

Table 10: StrainSeeker Identification Results - Sample 247

Read Direction	Bacteria/Strain
Forward	*Bacillus pumilus SAFR-032*
	Bacillus pumilus W3
	Bacillus pumilus GR-8
Reverse	*Bacillus pumilus SAFR-032*
	N/A
	N/A

Table 11: StrainSeeker Identification Results - Sample 254

Read Direction	Bacteria/Strain
Forward	*Bacillus pumilus GR-8*
	Bacillus pumilus W3
	Bacillus pumilus MTCC B6033
Reverse	*Bacillus pumilus GR-8*
	Bacillus pumilus W3
	Bacillus pumilus MTCC B6033

Table 12: StrainSeeker Identification Results - Sample 257

Read Direction	Bacteria/Strain
Forward	*Bacillus pumilus SAFR-032*
	Bacillus pumilus W3
	Bacillus pumilus GR-8
Reverse	*Bacillus pumilus SAFR-032*
	Bacillus pumilus W3
	Bacillus pumilus GR-8

Table 13: StrainSeeker Identification Results - Sample 269

Read Direction	Bacteria/Strain
Forward	*Bacillus pumilus SAFR-032*
	Bacillus pumilus NJ-M2
	Bacillus pumilus W3
Reverse	*Bacillus pumilus SAFR-032*
	Bacillus pumilus NJ-M2
	Bacillus pumilus W3

An additional attempt at identification was performed using StrainSeeker. The StrainSkeeker results were much more uniform across all six bacterial samples than the BLAST results. Both the forward and reverse sequence reads were matched against the database and each of the twelve searches yielded exclusively different strains of *Bacillus pumilus*. This strongly suggests that the identity of each of the six bacterial isolates is *Bacillus pumilus.* Of the different strains implicated in the search, the SAFR-032 strain was the most common top hit and was the closest match in five of the six isolates (Samples 243, 244, 247, 257 and 269). Interestingly, in the sixth isolate (Sample 254), it was not listed at all as a strong match with the GR-8 strain being the strongest match to the raw reads.

Also interesting was the fact that the SAFR-032 strain was only suggested as a match once during the BLAST search and even that one time was as the third hit of the second node of sample 244. The closeness in matches between the different strains to the assembled reads suggests that the differences between strains are very minimal and it is therefore very difficult to determine which strain the isolates are, or if they are a new strain altogether.

Keratinase Gene Search

Table 14: Geneious Search Results

Sample	Name	Direction	Length	Node
243	BPE2	forward	19	3
254	BPE2	reverse	19	7
254	F1 Forward	forward	21	7
254	BPI2	reverse	20	7
254	F2 Reverse	forward	27	7
254	BPI1	reverse	20	7
254	F1 Reverse	reverse	24	7
257	BPI2	reverse	20	4
257	F1 Reverse	reverse	24	4
269	BPI2	reverse	20	5
269	F1 Reverse	reverse	24	5

As a follow-up to the study above, Geneious was used to search for a group of known keratinase primers (Table 1) within the assembled genomes. Only primers with a length of at least 19 base pairs were chosen in order to reduce the possibility that a primer might match the genomes by chance. Perfect matches to the primers were observed in four of the six assembled genomes (Samples 243, 254, 257, and 269). The presence of these perfectly matching primers suggests that the four samples contain genes coding for keratinase.

For samples 244 and 247, the search was re-run twice with slightly less stringent conditions. For the first instance, one mismatch was allowed between the primer and the actual sequence, however, no matches were found. A second search was then conducted allowing for two mismatches, however no results were obtained from this search either. Although no matches

were found, Bacterial samples 244 and 247 may still contain keratinase genes. Only a small sample of primers were used in the search and the bacteria may contain undocumented or even novel keratinase genes.

Of the eleven matches found, eight were in the reverse direction whereas only three were in the forward direction. In the case of the F1 and F2 primers, both forward and reverse primers were searched and in three of the four instances in which they were found the direction of the sequence matched the direction of the primer. Additionally, the primers found in each sample were all grouped together on the same node which may indicate that a keratinase gene was identified in that area. More research is needed to confirm whether or not this is true and if any novel genes were identified.

Conclusion

The results of the BLAST identification found that the most common match to each of the six sample genomes was *Bacillus pumilus,* a well-documented FDB with potential for use in repurposing feather waste. This was further supported by the results of the StrainSeeker identification which matched the raw genomic reads exclusively to strains of *Bacillus pumilus*. Therefore, due to the evidence provided by the two tools we can safely conclude that *Bacillus pumilus* is the identity of each of the six isolates, however, there is not enough evidence to determine specific strains for any of the six isolates.

A follow-up study identified previously documented keratinase primers in four of the six genomes (Sample 243, 254, 257, and 269). The primers in these samples were all found very close to one another on the same node supporting the presence of a keratinase gene in all four bacteria. Although no keratinase primers were identified in Samples 244 and 247, there is still a possibility that the bacteria contain a different version of this gene since only a small sample of primers was actually searched. Additionally, since keratinases refer to a range of proteins which perform the task of degrading keratin, these bacteria may contain a gene for a slightly different keratinase protein.

Overall, FDBs have the potential to reduce and repurpose tons of feather waste each year. However, more study is still needed to determine the process by which keratinase degrades feathers and which FDB or combination of FDBs is the most effective. Strategies must also be developed to implement FDBs on a national scale.

Acknowledgements

I would like to thank all the members of the Hird Laboratory at the University of Connecticut for their advice and guidance over the past two years. I would like to give a special thank you to Elizabeth Herder for teaching me the skills I would need to complete this project and working with me as a mentor. I would also like to give a very special thank you to Dr. Sarah Hird for her support and instruction throughout my time working in her lab, and for giving me an amazing first research experience. Finally, I would like to thank my family for their love and support and for pushing me to be the best version of myself. Without all of your help, I would not have been able to complete this thesis.

References

1. Ley, Ruth E., Daniel A. Peterson, and Jeffrey I. Gordon. “Ecological and Evolutionary Forces Shaping Microbial Diversity in the Human Intestine.” *Cell* 124, no. 4 (February 24, 2006): 837–48. https://doi.org/10.1016/j.cell.2006.02.017.
2. Qin, Junjie, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, et al. “A Human Gut Microbial Gene Catalog Established by Metagenomic Sequencing.” *Nature* 464, no. 7285 (March 4, 2010): 59–65. https://doi.org/10.1038/nature08821.
3. Shreiner, Andrew B., John Y. Kao, and Vincent B. Young. “The Gut Microbiome in Health and in Disease.” *Current Opinion in Gastroenterology* 31, no. 1 (January 2015): 69–75. https://doi.org/10.1097/MOG.0000000000000139.
4. Kamada, Nobuhiko, and Gabriel Núñez. “Regulation of the Immune System by the Resident Intestinal Bacteria.” *Gastroenterology* 146, no. 6 (May 2014): 1477–88. https://doi.org/10.1053/j.gastro.2014.01.060.
5. Wang, Zeneng, Elizabeth Klipfell, Brian J. Bennett, Robert Koeth, Bruce S. Levison, Brandon Dugar, Ariel E. Feldstein, et al. “Gut Flora Metabolism of Phosphatidylcholine Promotes Cardiovascular Disease.” *Nature* 472, no. 7341 (April 7, 2011): 57–63. https://doi.org/10.1038/nature09922.
6. Kostic, Aleksandar D., Ramnik J. Xavier, and Dirk Gevers. “The Microbiome in Inflammatory Bowel Disease: Current Status and the Future Ahead.” *Gastroenterology* 146, no. 6 (May 2014): 1489–99. https://doi.org/10.1053/j.gastro.2014.02.009.
7. Ley, Ruth E., Fredrik Bäckhed, Peter Turnbaugh, Catherine A. Lozupone, Robin D. Knight, and Jeffrey I. Gordon. “Obesity Alters Gut Microbial Ecology.” *Proceedings of the National Academy of Sciences* 102, no. 31 (August 2, 2005): 11070–75. https://doi.org/10.1073/pnas.0504978102.
8. Qin, Junjie, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, et al. “A Metagenome-Wide Association Study of Gut Microbiota in Type 2 Diabetes.” *Nature* 490, no. 7418 (October 4, 2012): 55–60. https://doi.org/10.1038/nature11450.

9. Bull, Matthew J., and Nigel T. Plummer. “Part 1: The Human Gut Microbiome in Health and Disease.” *Integrative Medicine: A Clinician’s Journal* 13, no. 6 (December 2014): 17–22.
10. Forum, Institute of Medicine (US) Food. *Study of the Human Microbiome*. National Academies Press (US), 2013. https://www.ncbi.nlm.nih.gov/books/NBK154091/.
11. Turnbaugh, Peter J., Ruth E. Ley, Micah Hamady, Claire Fraser-Liggett, Rob Knight, and Jeffrey I. Gordon. “The Human Microbiome Project: Exploring the Microbial Part of Ourselves in a Changing World.” *Nature* 449, no. 7164 (October 18, 2007): 804–10. https://doi.org/10.1038/nature06244.
12. Reuter, Jason A., Damek Spacek, and Michael P. Snyder. “High-Throughput Sequencing Technologies.” *Molecular Cell* 58, no. 4 (May 21, 2015): 586–97. https://doi.org/10.1016/j.molcel.2015.05.004.
13. Simpson, Jared T., Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven J.M. Jones, and İnanç Birol. “ABySS: A Parallel Assembler for Short Read Sequence Data.” *Genome Research* 19, no. 6 (June 2009): 1117–23. https://doi.org/10.1101/gr.089532.108.
14. Luo, Ruibang, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, et al. “SOAPdenovo2: An Empirically Improved Memory-Efficient Short-Read de Novo Assembler.” *GigaScience* 1 (December 27, 2012): 18. https://doi.org/10.1186/2047-217X-1-18.
15. Boisvert, Sébastien, François Laviolette, and Jacques Corbeil. “Ray: Simultaneous Assembly of Reads from a Mix of High-Throughput Sequencing Technologies.” *Journal of Computational Biology* 17, no. 11 (November 2010): 1519–33. https://doi.org/10.1089/cmb.2009.0238.
16. Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, et al. “SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing.” *Journal of Computational Biology* 19, no. 5 (May 2012): 455–77. https://doi.org/10.1089/cmb.2012.0021.
17. Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215, no. 3 (October 5, 1990): 403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

18. Chakravorty, Soumitesh, Danica Helb, Michele Burday, Nancy Connell, and David Alland. “A Detailed Analysis of 16S Ribosomal RNA Gene Segments for the Diagnosis of Pathogenic Bacteria.” *Journal of Microbiological Methods* 69, no. 2 (May 2007): 330–39. https://doi.org/10.1016/j.mimet.2007.02.005.
19. Galkiewicz, Julia P., and Christina A. Kellogg. “Cross-Kingdom Amplification Using Bacteria-Specific Primers: Complications for Studies of Coral Microbial Ecology.” *Applied and Environmental Microbiology* 74, no. 24 (December 2008): 7828–31. https://doi.org/10.1128/AEM.01303-08.
20. Kent, W. James. “BLAT--the BLAST-like Alignment Tool.” *Genome Research* 12, no. 4 (April 2002): 656–64. https://doi.org/10.1101/gr.229202.
21. Roosaare, Märt, Mihkel Vaher, Lauris Kaplinski, Märt Möls, Reidar Andreson, Maarja Lepamets, Triinu Kõressaar, Paul Naaber, Siiri Kõljalg, and Maido Remm. “StrainSeeker: Fast Identification of Bacterial Strains from Raw Sequencing Reads Using User-Provided Guide Trees.” *PeerJ* 5 (2017): e3353. https://doi.org/10.7717/peerj.3353.
22. Grice, Elizabeth A., and Julia A. Segre. “The Skin Microbiome.” *Nature Reviews. Microbiology* 9, no. 4 (April 2011): 244–53. https://doi.org/10.1038/nrmicro2537.
23. Eaton, Muir D, and Scott M Lanyon. “The Ubiquity of Avian Ultraviolet Plumage Reflectance.” *Proceedings of the Royal Society B: Biological Sciences* 270, no. 1525 (August 22, 2003): 1721–26. https://doi.org/10.1098/rspb.2003.2431.
24. “Wayback Machine,” December 3, 2007. https://web.archive.org/web/20071203190459/http://www.orn.mpg.de/documents/peters/Delhey_AmNat2007_copy.pdf.
25. Gunderson, Alex R. “Feather-Degrading Bacteria: A New Frontier in Avian and Host–Parasite Research?” *The Auk* 125, no. 4 (October 2008): 972–79. https://doi.org/10.1525/auk.2008.91008.
26. Bragulla, Hermann H, and Dominique G Homberger. “Structure and Functions of Keratin Proteins in Simple, Stratified, Keratinized and Cornified Epithelia.” *Journal of Anatomy* 214, no. 4 (April 2009): 516–59. https://doi.org/10.1111/j.1469-7580.2009.01066.x.
27. Brandelli, Adriano, Daniel J. Daroit, and Alessandro Riffel. “Biochemical Features of Microbial Keratinases and Their Production and Applications.” *Applied Microbiology*

and Biotechnology 85, no. 6 (February 2010): 1735–50. https://doi.org/10.1007/s00253-009-2398-5.

28. Yamamura, Shohei, Yasutaka Morita, Quamrul Hasan, Kenji Yokoyama, and Eiichi Tamiya. "Keratin Degradation: A Cooperative Action of Two Enzymes from Stenotrophomonas Sp." *Biochemical and Biophysical Research Communications* 294, no. 5 (June 28, 2002): 1138–43. https://doi.org/10.1016/S0006-291X(02)00580-6.
29. Gupta, Rani, and Priya Ramnani. "Microbial Keratinases and Their Prospective Applications: An Overview." *Applied Microbiology and Biotechnology* 70, no. 1 (March 2006): 21–33. https://doi.org/10.1007/s00253-005-0239-8.
30. Li, Qingxin. "Progress in Microbial Degradation of Feather Waste." *Frontiers in Microbiology* 10 (December 5, 2019). https://doi.org/10.3389/fmicb.2019.02717.
31. Verma, Amit, Hukum Singh, Shahbaz Anwar, Anirudha Chattopadhyay, Kapil K. Tiwari, Surinder Kaur, and Gurpreet Singh Dhilon. "Microbial Keratinases: Industrial Enzymes with Waste Management Potential." *Critical Reviews in Biotechnology* 37, no. 4 (June 2017): 476–91. https://doi.org/10.1080/07388551.2016.1185388.
32. Tamreihao, K., Saikat Mukherjee, Rakhi Khunjamayum, Laishram Jaya Devi, Roshan Singh Asem, and Debananda S. Ningthoujam. "Feather Degradation by Keratinolytic Bacteria and Biofertilizing Potential for Sustainable Agricultural Production." *Journal of Basic Microbiology* 59, no. 1 (2019): 4–13. https://doi.org/10.1002/jobm.201800434.
33. Kshetri, Pintubala, Subhra S. Roy, Susheel K. Sharma, Thangjam S. Singh, Meraj A. Ansari, Blessa Sailo, Surender Singh, and Narendra Prakash. "Feather Degrading, Phytostimulating, and Biocontrol Potential of Native Actinobacteria from North Eastern Indian Himalayan Region." *Journal of Basic Microbiology* 58, no. 9 (September 2018): 730–38. https://doi.org/10.1002/jobm.201800169.
34. Cai, Cheng-gang, Bing-gan Lou, and Xiao-dong Zheng. "Keratinase Production and Keratin Degradation by a Mutant Strain of Bacillus Subtilis." *Journal of Zhejiang University. Science. B* 9, no. 1 (January 2008): 60–67. https://doi.org/10.1631/jzus.B061620.
35. Yamamura, Shohei, Yasutaka Morita, Quamrul Hasan, Sathuluri Ramachandra Rao, Yuji Murakami, Kenji Yokoyama, and Eiichi Tamiya. "Characterization of a New Keratin-

Degrading Bacterium Isolated from Deer Fur." *Journal of Bioscience and Bioengineering* 93, no. 6 (2002): 595–600.

36. Bhange, Khushboo, Venkatesh Chaturvedi, and Renu Bhatt. "Feather Degradation Potential of Stenotrophomonasmaltophilia KB13 and Feather Protein Hydrolysate (FPH) Mediated Reduction of Hexavalent Chromium." *3 Biotech* 6, no. 1 (June 2016). https://doi.org/10.1007/s13205-016-0370-5.
37. Callahan, Benjamin J., Paul J. McMurdie, Michael J. Rosen, Andrew W. Han, Amy Jo A. Johnson, and Susan P. Holmes. "DADA2: High-Resolution Sample Inference from Illumina Amplicon Data." *Nature Methods* 13, no. 7 (July 2016): 581–83. https://doi.org/10.1038/nmeth.3869.
38. Najoshi. *Najoshi/Sickle*. C, 2020. https://github.com/najoshi/sickle.
39. Kearse, Matthew, Richard Moir, Amy Wilson, Steven Stones-Havas, Matthew Cheung, Shane Sturrock, Simon Buxton, et al. "Geneious Basic: An Integrated and Extendable Desktop Software Platform for the Organization and Analysis of Sequence Data." *Bioinformatics* 28, no. 12 (June 15, 2012): 1647–49. https://doi.org/10.1093/bioinformatics/bts199.
40. Radha, S., and P. Gunasekaran. "Sustained Expression of Keratinase Gene under PxylA and PamyL Promoters in the Recombinant Bacillus Megaterium MS941." *Bioresource Technology* 99, no. 13 (n.d.): 5528–37.
41. Fellahi, Soltana, Abdelwaheb Chibani, Elisabeth Feuk-Lagerstedt, and Mohammad J. Taherzadeh. "Identification of Two New Keratinolytic Proteases from a Bacillus Pumilus Strain Using Protein Analysis and Gene Sequencing." *AMB Express* 6 (June 30, 2016). https://doi.org/10.1186/s13568-016-0213-0.
42. Vettath, Vineeth Kodengil, Ana Carolina M. Junqueira, Akira Uchida, Rikky W. Purbojati, James N. I. Houghton, Caroline Chénard, Daniela I. Drautz-Moses, et al. "Complete Genome Sequence of Bacillus Altitudinis Type Strain SGAir0031 Isolated from Tropical Air Collected in Singapore." *Genome Announcements* 5, no. 45 (November 9, 2017). https://doi.org/10.1128/genomeA.01260-17.
43. Ramakrishna Reddy, M., K. Sathi Reddy, Y. Ranjita Chouhan, Hameeda Bee, and Gopal Reddy. "Effective Feather Degradation and Keratinase Production by Bacillus Pumilus

GRK for Its Application as Bio-Detergent Additive." *Bioresource Technology* 243 (November 2017): 254–63. https://doi.org/10.1016/j.biortech.2017.06.067.

44. Nagal, Swetlana, and P. C. Jain. "Feather Degradation by Strains of Bacillus Isolated from Decomposing Feathers." *Brazilian Journal of Microbiology* 41, no. 1 (2010): 196–200. https://doi.org/10.1590/S1517-838220100001000028.