Research and Testing Laboratory

Data Analysis Methodology

Last Updated: 10/26/2012

The most recent version of this file can be downloaded from
http://www.researchandtesting.com/docs/Data_Analysis_Methodology.pdf

Contents

Term Definitions

Terms used within this guide are defined as follows:

- Tag
  - The term tag refers to the 8-10 bp sequence at the 5' end of the sequence read.
  - The tag is also known as the barcode in some programs.
- Identity Percentage
  - Identity percentage for a read is defined as the length of the HSP Identity divided by the length of the hit HSP.
  - BLAST is run to return 5 hits and 1 HSP per hit.
- HSP – High-Score Pair
  - The highest scoring local alignment between a sequence read and the database sequence such that the score cannot be improved by extension or trimming of the alignment.
- HSP coverage
  - The HSP length divided by the query length, giving the percentage of the query sequence covered by the HSP.

Data Analysis Methodology

Overview of the Data Analysis Process

Once sequencing has completed, the data analysis pipeline will begin processing the data. The data analysis process consists of two major stages, the quality checking and reads denoising stage and the diversity analysis stage. During the read quality checking and denoising stage, denoising and chimera checking is performed on all the reads for each region of data. Then each remaining read is quality scanned to remove poor reads from each sample. The primary output of this stage is a quality checked and denoised FASTA formatted sequence, quality, and mapping file. This stage is performed for all customers whose data we know the encoded tags for. During the diversity analysis stage, each sample is run through our analysis pipeline to determine the taxonomic information for each read and then analyzed to provide the sample's microbial diversity. The output for this stage is a set of files detailing the taxonomic information for each read as well as the number and percentage of each species found within each sample. This stage is performed for all customers whose data is sequenced using primers based within the 16S, 18S, 23S, ITS and SSU regions.

The data analysis pipeline is broken down into the following steps, each of which is discussed more thoroughly in the sections below:

- Raw Data File Generation
    1. SFF File Generation

- Quality Checking and Denoising
    2. Denoising and Chimera Checking
    3. Quality Checking and FASTA Formatted Sequence/Quality File Generation

- Microbial Diversity Analysis
    1. Taxanomic Identification
    2. Data Analysis

SFF File Generation

SFF files are a binary file containing many data about a read in a single file. For each read, the sff contains a flowgram, quality score and sequence with defined lengths from QC measures performed by the machine. The sff represents the raw data and includes many reads that may have been excluded due to length or chimera detection or any other filter requested for custom processing. Since the files are binary, they cannot be opened with standard text editors. Special programs like Mothur [1] or BioPython [2] are able to load their data into human readable formats and output fasta, qual, flowgram or text (sff.txt) versions. Sff files or their derivatives can then be used for further processing of the data. Sff files provided may be of two forms. In the case of an entire region containing a single investigator's samples, the entire region plus mapping file is provided. In cases where multiple investigators had samples on a single region, each sample is demultiplexed from the sff file using the Roche sffinfo tool by providing its barcode, effectively eliminating it from any read extracted. The split sff can then be used for raw data or submitted directly to archives like the NCBI's SRA. In cases where a single sff for all samples is desired but an entire quadrant is not used, an investigator may request a single sff for a nominal charge. Alternatively, it is possible to use the provided split sff files for denoising/chimera removal by modifying the mapping files. Additional instructions are available if you wish to do so.

Denoising and Chimera Checking

The process of denoising is used to correct errors in reads from next-generation sequencing technologies including the Roche 454 technologies. According to the papers "Accuracy and quality of massively parallel DNA pyrosequencing" by Susan Huse, et al. and "Removing noise from pyrosequenced amplicons" by Christopher Quince, et al. the per base error rates from 454 pyrosequencing attain an accuracy rate of 99.5% [3] [4]. However, the large read numbers that the

machine can generate mean that the total number of noisy reads can be substantial. In order to determine true diversity it becomes critical to determine which reads are good and which reads contain noise introduced by the sequencing platform. The Research and Testing Laboratory analysis pipeline attempts to account for this by denoising entire regions of data prior to performing any other steps of the pipeline.

The Research and Testing pipeline broken down into four major stages, with each stage defined in-depth in the following sections. Each of these stages is performed on each of the regions generated by the sequencer separately. The four primary stages are:

- Quality Trimming
- Clustering
- Chimera Checking
- Denoising

Quality Trimming

The Quality Trimming stage is used to attempt to clean up potential poor quality ends of each read. Quality trimming uses the Quality Scores provided by the sequencer to determine where the sequencer data has become too noisy. The Research and Testing denoiser uses the following algorithm in order to perform the quality trimming stage:

- All the reads from a sequencing region are read out of the FASTA file directly from the sequencer.
- The Quality Scores for each read are then read in.
- The quality score for each base is then added to a running sum, which is then divided by the current number of bases read in to determine the running average up until that base.
- The running average is then compared so a threshold (T) to determine if the average that position has fallen below the allowed average threshold.
- Once the entire read has been averaged, the read will be trimmed back to the last base that caused the read to fall below T, unless the overall average is greater than T.
- For example, if a read ended with the following 5 averages - 29 30 31 28 27, then the read would be trimmed back to the 28 mark, making the base that pulled the average up to 31 the final base on the new read.
- Currently our pipeline sets T = 25.

Clustering

The Clustering stage attempts to classify reads into clusters and then remove any reads that do not manage to join one of these clusters. This stage uses the trimmed sequences and quality scores provided from the Quality Trimming stage and will output the clustered sequences along with the information required to determine how each read joined the cluster. The Research and Testing denoiser uses the following algorithm in order to perform the clustering stage:

- Reads within the data set are sorted from the longest to the shortest read.
- Using USEARCH [5], reads are then dereplicated, meaning they are clustered together into groups such that each sequence is an exact match to a portion of the seed sequence for the cluster. Each cluster is marked with the total number of member sequences.
- The seed sequences are then sorted again by length, from longest seed to shortest seed. Keep in mind that no minimum size restrictions exist on the clusters, thus single member clusters will exist.
- Clustering at a 4% divergence using the USEARCH [5] application is performed on the seed sequences in order to determine similar clusters. The result of this stage is the consensus sequence from each new cluster, each tagged to show their total number of member sequences (dereplicated + clustered).
- The consensus sequences are re-sorted again based upon their length, from the longest to shortest read. A minimum size restriction then removes any cluster that does not contain at least two member sequences is removed from consideration. These removes sequences are referred to as singletons and tend to indicate a particularly noisy read.

Once this process has completed, we have removed all reads that failed to have a similar or exact match elsewhere on the region. Moreover, through the use of consensus sequences, we have generated a set of reads that can be used to help correct base pair errors that occurred during sequencing.

Chimera Checking

As discussed in the paper "Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons" by Brian Haas, et al. the formation of chimeric sequences occurs when an aborted sequence extension is misidentified as a primer and is extended upon incorrectly in subsequent PCR cycles. This can be seen in Figure 1, shown below.

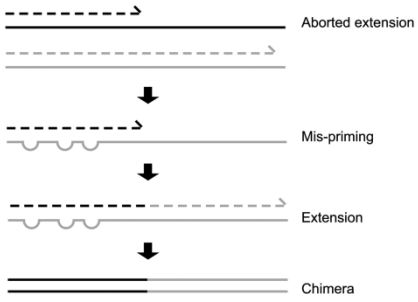

Figure 1.

Formation of chimeric sequences during PCR. An aborted extension product from an earlier cycle of PCR can function as a primer in a subsequent PCR cycle. If this aborted extension product anneals to and primes DNA synthesis from an improper template, a chimeric molecule is formed. Figure and description taken directly from "Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons" by Brian Haas, et al.

Because amplification produces chimeric sequences that stem from the combination of two or more original sequences [6], we will perform chimera detection using the *de novo* method built into UCHIIME.

The Research and Testing analysis pipeline performs chimera detection and removal by executing UCHIIME in *de novo* mode on the clustered data that was output by our denoising methods. By using this method we can determine chimeras across entire region of data even after accounting for noise and removing low quality sequences.

Once UCHIIME has completed execution we are left with a file informing us which sequences are possibly chimeric and which sequences are definitely not chimeric. All possibly chimeric sequences are removed at this point.

Denoising

The final stage of the pipeline is Denoising, which takes the data generated by all previous steps and attempts to correct base pair errors and remove bad sequences, such as chimeric sequences and noisy reads. The Research and Testing denoiser uses the following algorithm in order to perform the clustering stage:

- The system will read in the file containing a list of all non-chimeric sequences and then will back track through to collect the list of all reads that are to be kept. This process takes all of the names of each trimmed read and maps it to the dereplicated seed that it was a part of, maps each dereplicated seed name to the name of the clustered consensus sequence, and then maps each consensus sequence to either being chimeric or non-chimeric.
- Working back through the list allows us to determine which trimmed sequences to keep and which ones violated one of the rules during clustering or chimera checking.
- The remaining sequences are then grouped back up according to their consensus cluster and have their alignment pattern and quality scores read in to the system.

- Each sequence in a cluster then compares itself to the consensus sequence using the alignment pattern and their quality scores to determine what should occur. The algorithm uses the following logic:
    - If the base is marked for deletion and the quality score is below 30, the base is deleted.
    - If the base is marked for deletion and the quality score is greater than or equal to 30, then the base is retained.
    - If the base is marked for alteration and the quality score is less than 30, the base is changed to whatever the consensus sequence is, unless the consensus states to put an N.
    - If the base is marked for alteration and the quality score is greater than or equal to 30, then the base is retained.
    - If the pattern states that a base should inserted into the sequence, then the base is inserted into the sequence at that position.
- For each base pair that is changed or added, a new quality score is generated by taking the lower median value of all quality scores from the sequences that showed that base. i.e. if sequences 1 and 3 show an A should be in a position (with scores 30 and 32) and sequence 2 as a G in that position (with a score of 15), then the G is replaced with an A and the quality score for this A is changed to 30 (the median value for the set. In this case median values will always take the lower of the two options to keep from artificially inflating bases).
- For each base that is deleted, the quality score for that base is also removed.
- At this point each read and quality score is now written to the denoised FNA and Qual files for use in our taxonomic analysis pipeline.

Quality Checking and FASTA Formatted Sequence/Quality File Generation

The denoised and chimera checked reads generated during sequencing are condensed into a single FASTA formatted file such that each read contains a one line descriptor and one to many lines of sequence/quality scores. The Research and Testing Laboratory analysis pipeline takes the FASTA formatted sequence and quality files and removes all sequences that meet the following quality control requirements:

1. Failed sequence reads,
2. Sequences that have low quality tags and
3. Sequences that fail to be at least half the expected amplicon length or 250 bp in length, whichever is shortest.

Sequences that pass the quality control screening are condensed into a single FASTA formatted sequence and quality file such that each read has a one line descriptor followed by a single line of sequence/quality data. The descriptor line in both files has been altered to contain the samples name followed by the original descriptor line, separated with a unique delimiter (::).

This stage of the pipeline creates the FASTA reads archive which contains the following files:

1. The sequence reads from all samples concatenated into a single sequence file. The original tags have been removed from each sequence and an "artificial tag" has been added in its place. The title of the file will be <name>.fas.
2. The quality scores from all samples concatenated into a single quality file. The scores are labeled with the corresponding sample name and will have a matching line in the .fas file. Since the original tags were removed from the sequence and an "artificial tag" was put into its place, the quality scores have been similarly altered such that the original scores for the tag have been removed and an "artificial quality tag" has been added in its place. The artificial quality tag consists of Q30s for the length of the tag. This file will be labeled <name>.qual.
3. A mapping file consisting of sample names included in the analysis. This file contains the information for each sample such that each line has the sample name, tag and primer used for the sample. This file will be labeled as: <name>.txt

Taxonomic Identification

In order to determine the identity of each remaining sequence, the sequences will first be sorted such that the FASTA formatted file will contain reads from longest to shortest. These sequences are then clustered into OTU clusters with 100% identity (0% divergence) using USEARCH [5]. For each cluster the seed sequence will be put into a FASTA formatted sequence file. This file is then queried against a database of high quality sequences derived from NCBI using a distributed .NET algorithm that utilizes BLASTN+ (KrakenBLAST www.krakenblast.com). Using a .NET and C# analysis pipeline the resulting BLASTN+ outputs were compiled and data reduction analysis performed as described previously [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17] [18] [19] [20] [21] [22] [23] [24] [25] [26] [27] [28] [29].

Based upon the above BLASTn+ derived sequence identity percentage the sequences were classified at the appropriate taxonomic levels based upon the following criteria. Sequences with identity scores, to well characterized database sequences, greater than 97% identity (<3% divergence) were resolved at the species level, between 95% and 97% at the genus level, between 90% and 95% at the family and between 85% and 90% at the order level , 80 and 85% at the class and 77% to 80% at phyla. Any match below this percent identity is discarded. In addition, the HSP must be at least 75% of the query sequence or it will be discarded, regardless of identity.

After resolving based upon these parameters, the percentage of each organism will be individually analyzed for each sample providing relative abundance information within and among the individual samples based upon relative numbers of reads within each. Evaluations presented at each taxonomic level, including percentage compilations represent all sequences resolved to their primary identification or their closest relative [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17] [18] [19] [20] [21] [22] [23] [24] [25] [26] [27] [28] [29].

Analysis description

These folders contain the actual result files from analysis. The folder contains the results for blasting the seed reads against the appropriate Research and Testing Laboratory database. It also contains a seqs_otu_table.txt file showing the dereplicated sequences, clustering at 0% divergence, with the confidence interval used to give the level of classification for it. Also present is the seqs_otus.txt file which gives the clusters formed by USEARCH. The seed read used for a cluster is the first listed and is the longest sequence of the cluster.

The .csv files are in a comma separated format and can be opened with Excel or another text editor program. Each file can be dragged and dropped into Excel or you may choose to right click on the file name, select “Open With” and choose Excel as the program. Each file contains information about all the samples. Sample names span the first row with the bacterial/fungal designations at each respective taxonomic level are listed in the first column. Counts (or Percentages- if looking at the …Percent.csv file) of each of the respective taxonomic levels found within the sample are listed below the sample name.

Generally, the most relevant files are the Percent composition files, although the PercentTraceback can be used when confidence intervals are desired. The files include composition information forced to the top blast hit at a specific taxonomic level. For example, in the species files, the nearest well described species for each sequence is listed; similarly, the genus files contain the genus of the species as catalogued in the database, and so on for each taxonomic level. Research and Testing Laboratory uses 7 taxonomic levels for each organism: Kingdom, Phylum, Class, Order, Family, Genus, Species.

The information is organized by taxonomic level (each file specifies for which taxonomic level the information included is for). Files names include:

<name>Kingdom
<name>BelowMinimum
<name>Excluded
<name><taxa level>Counts
<name><taxa level>Percent

<name>SpeciesOptions
<name>SpeciesFullTaxa
<name>SpeciesTraceback
<name>SpeciesPercentTraceback

File descriptions:

<name>Kingdom.csv with the following columns followed by samples:

1. Query sequence name (with sample label) and its cluster information
2. The hit name with "-I" followed by the identity percentage and "Q" followed by the query length,
3. Identity column, indicating the identity percentage of the query to the hit sequence along the HSP (High Scoring Pair) region.
4. The count for the hit, based on the number of cluster members. If clustering is not performed, this is always 1.

<name>BelowMinimum.csv

Contains all hits from the blast results that fell below 77% identity OR had HSP coverage below 70%. These are given a classification at the closest species, but do not appear in any other file. These hits did not have sufficient similarity to any reference sample to have confidence in assigning to an organism. If a sample had all its reads fall into this file, all values in the Counts file will show as 0 and all the values in the Percent files will show as NaN for the sample as a result of division by zero.

<name>Excluded.csv

Contains hits against organisms that have been requested to be excluded from results by the sender. By default, this is always empty. Some items may be dropped from the blastout and will not appear in the excluded file. This currently includes plastid and mitochondrial sequences that may be present in the database.

Additionally, some reads generated may be present in none of the above files if they were not sufficiently similar to any of the reference sequences in the database. This may occur due to a spurious amplification or poor quality read. These reads are considered to be further noise in the data.

Each of the taxa levels contains 2 files

<name><taxa level>Counts.csv

Counts are merged on the organism term, with separation given to those with top hits among different species, condensing all identity scores and query lengths. A set of hits by a read against multiple organisms but with identical similarity have each organism listed, separated by "::". A read may have up to 5 hits, but only the best are used for determination. If a non-specific species is found as the top hit, a similar quality hit on a full species is set as the top hit for the Percent computations. Each unique set of terms has a single line with summed counts for all of its samples.

<name><taxa level>Percent.csv

Converts the previous file of raw read counts into percentages of composition per sample per organism, for the top hit only. Multiple organism hits are converted to top hit only for merging with other reads with similar top hits, except those with generic species as the top hit, which are changed to the next hit with a specific name. Some entries here may appear as NaN, indicating that the count was zero as a result of discarding all hits for the sample due to percent identity or filtering.

Additional species files

<name>SpeciesOptions.csv

Shows the percent file without discarding any top hits and calculating percents based on this. This shows what percent of the reads had only a single similar species based on sequence or multiple similarities.

<name>SpeciesFullTaxa.csv

Shows the Percent file with the full taxa of every organism listed. This way, multiple files do not need to be examined to determine the catalogued taxonomy used in the groupings

Due to incomplete taxonomic data for some organisms, the top hits may have unusual naming conventions. Some entries have no specific name, so they are abbreviated <Genus> sp. If a sample from the database had missing taxonomic data at the class level e.g. *Cyanobacteria*, then the class of the organism would be *Cyanobacteria (class)*. Others may be named with an "unclassified" e.g. *Clostridiaceae unclassified.* This occurs in many places, but the naming convention remains the same for any taxonomic level. Those without at least a defined class are not added to the database for bacteria and fungi. Also seen are species like *Pseudomonas sp Rhizobiaceae* or a genus like *Pseudomonas (Rhizobaceae)*. This is a generic Pseudomas species belonging to the Rhizobiaceae family. Since there are multiple entries for Pseudomonas sp in the NCBI Taxonomy, the suffix has been added to distinguish at each level where there may be confusion over which species it is, if only the usual nomenclature had been used.

Also included in the **Percent** folder is a "traceback" file, *SpeciesPercentTraceback*. This file reflects taxonomic information for organisms found in the sample based on the goodness of the alignment against reference sequences. This file is similar to the previously described percent file, however, the nearest neighbor isn't forced, but instead based on the percent identity of the query sequence to the reference sequence; the "most certain" taxonomic level is listed. A summary table of the level which is used for identification based on identity score is provided below.

Identity to reference sequence	Traceback Designation
I > 97%	
97% ≥ I > 95%	(unk species)
95% ≥ I > 90%	(unk genus)
90% ≥ I > 85%	(unk family)
85% ≥ I > 80%	(unk order)
80% ≥ I > 77%	(unk class)

For example, if one of the sequences was identified as a *Staphylococcus aureus* with an identity of 98%, it would still be labeled as *Staphylococcus aureus* in the Traceback file. However, if the identity was 88%, it would be grouped with other sequences with identities between 85-89% belonging to the Bacillales order and labeled as *Bacillales (unk family)*. The traceback changes are made to all entries, and then additional merging is done to combine now repeated terms. This may also split hits against an organism into several levels of confidence giving a *Staphylococcus aureus, Staphylococcus (unk species), Staphylococcaceae (unk genus)* from the *S. aureus* hits. Since taxonomic database entries are used, there may be entries similar to *Bacillales(family)(unk genus)* which is a hit against an organism with incomplete taxa (only up to order) and an identity above 89% but below 94%. Any entry with a full species name and no parenthetical notation means the hit was above 96% identity.

Due to the changes made between the *SpeciesPercent* file and the *SpeciesTracebackPercent* file, the two are not directly comparable. The Percent file shows only the top hit result while the Traceback further divides this into confidence measures based on percent identity.

Please feel free to contact bioinformatics@researchandtesting.com for any other clarification or concerns.

References

[1] P. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B.

Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. V. Horn and C. F. Weber, "Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities," *Appl Environ Microbiol,* vol. 75, no. 23, pp. 7537-41, 2009.

[2] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski and M. J. L. d. Hoon, "Biopython: freely available Python tools for computational molecular biology and bioinformatics," *Bioinformatics,* vol. 25, no. 11, 2009.

[3] S. M. Huse, J. A. Huber, H. G. Morrison, M. L. Sogin and D. M. Welch, "Accuracy and quality of massively parallel DNA pyrosequencing.," *Genome Biology,* vol. 8, no. 7, 2007.

[4] C. Quince, A. Lanzen, R. J. Davenport and P. J. Turnbaugh, "Removing Noise From Pyrosequenced Amplicons," *BMC Bioinformatics ,* vol. 12, no. 38, 2011.

[5] R. C. Edgar, "Search and clustering orders of magnitude faster than BLAST," *Bioinformatics,* pp. 1-3, 12 August 2010.

[6] R. C. Edgar, B. J. Haas, J. C. Clemente, C. Quince and R. Knight, "UCHIIME improves sensitivity and speed of chimera detection," *Oxford Journal of Bioinformatics,* vol. 27, no. 16, pp. 2194-2200, 2011.

[7] R. Andreotti, A. Pérez de León, S. Dowd, F. Guerrero, K. Bendele and G. Scoles, "Assessment of bacterial diversity in the cattle tick Rhipicephalus (Boophilus) microplus through tag-encoded pyrosequencing," *BMC Microbiology,* vol. 11, no. 1, p. 6, 2011.

[8] M. T. Bailey, S. E. Dowd, N. M. .. Parry, J. D. Galley, D. B. Schauer and M. Lyte, "Stressor Exposure Disrupts Commensal Microbial Populations in the Intestines and Leads to Increased Colonization by Citrobacter rodentium," *Infection and Immunity,* vol. 78, no. 4, pp. 1509-1519, April 2010.

[9] M. T. Bailey, J. C. Waltonc, S. E. Dowd, Z. M. Weil and R. J. Nelson, "Photoperiod modulates gut bacteria composition in male Siberian hamsters (Phodopus sungorus)," *Brain, Behavior, and Immunity,* vol. 24, no. 4, pp. 577-584, May 2010.

[10] T. R. Callaway, S. E. Dowd, T. S. Edrington, R. C. Anderson, N. Krueger, N. Bauer, P. J. Kononoff and D. J. Nisbet, "Evaluation of bacterial diversity in the rumen and feces of cattle fed different levels of dried distillers grains plus solubles using bacterial tag-encoded FLX amplicon pyrosequencing," *Journal of Animal Science,* vol. 88, no. 12, pp. 3977-3983, 2010.

[11] R. Wolcott, V. Gontcharova, Y. Sun, A. Zischakau and S. Dowd, "Bacterial diversity in surgical site infections: not just aerobic cocci any more.," *Journal of Wound Care,* vol. 18, no. 8, pp. 317-323, 2009.

[12] R. D. Wolcott, V. Gontcharova, Y. Sun and S. E. Dowd, "Evaluation of the bacterial diversity among and within individual venous leg ulcers using bacterial tag-encoded FLX and Titanium amplicon pyrosequencing and metagenomic approaches," *BMC Microbiology,* vol. 9, no. 226, 2009.

[13] T. R. Callaway, S. E. Dowd, R. D. Wolcott, Y. Sun, J. L. McReynolds, T. S. Edrington, J. A. Byrd, R. C. Anderson, N. Krueger and D. J. Nisbet, "Evaluation of the bacterial diversity in cecal contents of laying hens fed various molting diets by using bacterial tag-encoded FLX amplicon pyrosequencing," *Poultry Science,* vol. 88, no. 2, pp. 298-302, 2009.

[14] S. E. Dowd, T. R. Callaway, R. D. Wolcott, Y. Sun, T. McKeehan, R. G. Hagevoort and T. S. Edrington, "Evaluation of the bacterial diversity in the feces of cattle using 16S rDNA bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP)," *BMC Microbiology,* vol. 8, no. 125, 2008.

[15] S. E. Dowd, Y. Sun, P. R. Secor, D. D. Rhoads, B. M. Wolcott, G. A. James and R. D. Wolcott, "Survey of bacterial diversity in chronic wounds using Pyrosequencing, DGGE, and full ribosome shotgun sequencing," *BMC Microbiology,* vol. 8, no. 43, 2008.

[16] S. E. Dowd, Y. Sun, R. D. Wolcott, A. Domingo and J. A. Carroll, "Bacterial Tag–Encoded FLX Amplicon Pyrosequencing (bTEFAP) for Microbiome Studies: Bacterial Diversity in the Ileum of Newly Weaned Salmonella-Infected Pigs," *Foodborne Pathogens and Disease,* vol. 5, no. 4, pp. 459-472, 2008.

[17] F. D. Guerrero, S. E. Dowd, Y. Sun, L. Saldivar, G. B. Wiley, S. L. Macmil, F. Najar, B. A. Roe and L. D. Foil, "Microarray Analysis of Female- and Larval-Specific Gene Expression in the Horn Fly (Diptera: Muscidae)," *Journal of Medical Entomology,* vol. 46, no. 2, pp. 257-270, 2009.

[18] F. D. Guerrero, S. E. Dowd, A. Djikeng, G. Wiley, S. Macmil, L. Saldivar, F. Najar and B. A. Roe, "A Database of Expressed Genes From Cochliomyia hominivorax (Diptera: Calliphoridae)," *Journal of Medical Entomology,* vol. 46, no. 5, pp. 1109-1116, 2008.

[19] S. Handl, S. E. Dowd, J. F. Garcia-Mazcorro, J. M. Steiner and J. S. Suchodolski, "Massive parallel 16S rRNA gene pyrosequencing reveals highly diverse fecal bacterial and fungal communities in healthy dogs and cats," *FEMS Microbiology Ecology,* vol. 76, no. 2, pp. 301-310, 2011.

[20] H. D. Ishak, R. Plowes, R. Sen, K. Kellner, E. Meyer, D. A. Estrada, S. E. Dowd and U. G. Mueller, "Bacterial Diversity in Solenopsis invicta and Solenopsis geminata Ant Colonies Characterized by 16S amplicon 454 Pyrosequencing," *Microbial Ecology,* vol. 61, no. 4, pp. 821-831, 2011.

[21] J. Leake, S. Dowd, R. Wolcott and A. Zischkau, "Identification of yeast in chronic wounds using new pathogen-detection technologies," *Journal of Wound Care,* vol. 18, no. 3, pp. 103-108, 2009.

[22] W. Li, S. E. Dowd, B. Scurlock, V. Acosta-Martinez and M. Lyte, "Memory and learning behavior in mice is temporally associated with diet-induced alterations in gut bacteria," *Physiology & Behavior,* vol. 96, no. 4-5, pp. 557-567, 2009.

[23] P. U. Olafson, K. H. Lohmeyer and S. E. Dowd, "Analysis of expressed sequence tags from a significant livestock pest, the stable fly (Stomoxys calcitrans), identifies transcripts with a putative role in chemosensation and sex determination," *Insect Biochemistry and Physiology,* vol. 74, no. 3, pp. 179-204, 2010.

[24] D. W. Pitta, W. E. Pinchak, S. E. Dowd, J. Osterstock, V. Gontcharova, E. Youn, K. Dorton, I. Yoon, B. R. Min, J. D. Fulford, T. A. Wickersham and D. P. Malinowski, "Rumen Bacterial Diversity Dynamics Associated with Changing from Bermudagrass Hay to Grazed Winter Wheat Diets," *Microbial Ecology,* vol. 59, no. 3, pp. 511-522, 2010.

[25] R. Sen, H. D. Ishak, D. Estrada, S. E. Dowd, E. Hong and U. G. Mueller, "Generalized antifungal activity and 454-screening of Pseudonocardia and Amycolatopsis bacteria in nests of fungus-growing ants," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 106, no. 42, pp. 17805-17810, 2009.

[26] D. Smith, D. Snow, E. Rees, A. Zischkau, J. D. Hanson, R. Wolcott, Y. Sun, J. White, S. Kumar and S. Dowd, "Evaluation of the bacterial diversity of Pressure ulcers using bTEFAP pyrosequencing," *BMC Medical Genomics,* vol. 3, no. 1, p. 41, 2010.

[27] M.-F. Stephenson, L. Mfuna, S. E. Dowd, R. D. Wolcott, J. Barbeau, M. Poisson, G. James and M. Desrosiers, "Molecular characterization of the polymicrobial flora in chronic rhinosinusitis," *Journal of Otolaryngology-Head And Neck Surgery,* vol. 39, no. 2, pp. 182-187, 2010.

[28] J. S. Suchodolski, S. E. Dowd, E. Westermarck, J. M. Steiner, R. D. Wolcott, T. Spillmann and J. A. Harmoinen, "The effect of the macrolide antibiotic tylosin on microbial diversity in the canine small intestine as demonstrated by massive parallel 16S rRNA gene sequencing," *BMC Microbiology,* vol.

9, no. 210, 2009.

[29] W. Williams, L. Tedeschi, P. Kononoff, T. Callaway, S. Dowd, K. Karges and M. Gibson, "Evaluation of in vitro gas production and rumen bacterial populations fermenting corn milling (co)products," *Journal of Dairy Science,* vol. 93, no. 10, pp. 4735-4743, 2010.

[30] S. E. Dowd, R. D. Wolcott, Y. Sun, T. McKeehan, E. Smith and D. Rhoads, "Polymicrobial Nature of Chronic Diabetic Foot Ulcer Biofilm Infections Determined Using Bacterial Tag Encoded FLX Amplicon Pyrosequencing (bTEFAP)," *PLoS One,* vol. 3, no. 10, 2008.

[31] S. M. Finegold, S. E. Dowd, V. Gontcharova, C. Liu, K. E. Henley, R. D. Wolcott, E. Youn, P. H. Summanen, D. Granpeesheh, D. Dixon, M. Liu, D. R. Molitoris and J. A. G. III, "Pyrosequencing study of fecal microflora of autistic and control children," *Anaerobe,* vol. 16, no. 4, pp. 444-453, 2010.

[32] V. Gontcharova, E. Youn, Y. Sun, R. D. Wolcott and a. S. E. Dowd, "A Comparison of Bacterial Composition in Diabetic Ulcers and Contralateral Intact Skin," *Open Microbiology Journal,* vol. 4, pp. 8-19, 2010.

[33] V. Gontcharova, E. Youn, R. D. Wolcott, E. B. Hollister, T. J. Gentry and S. E. Dowd, "Black Box Chimera Check (B2C2): a Windows-Based Software for Batch Depletion of Chimeras from Bacterial 16S rRNA Gene Datasets," *Open Microbiology Journal,* vol. 4, pp. 47-52, 2010.