5-2012

Assembly and Automated Annotation of the *Clostridium scatologenes* Genome

Jitesh Tiwari
Western Kentucky University, jitesh.tiwari364@topper.wku.edu

ASSEMBLY AND AUTOMATED ANNOTATION OF THE *CLOSTRIDIUM SCATOLOGENES* GENOME

A Thesis
Presented to
The Faculty of the Department of Biology
Western Kentucky University
Bowling Green, Kentucky

In Partial Fulfillment
Of the Requirement of the Degree
Master of Science

By
Jitesh Tiwari

May 2012

ASSEMBLY AND AUTOMATED ANNOTATION OF
CLOSTRIDIUM SCATOLOGENES GENOME

Date Recommended Dec. 15, 2011

Claire A. Rinehart

Claire Rinehart, Director of Thesis

Sigrid Jacobshagen

Sigrid Jacobshagen

Jonathan Quiton

Jonathan Quiton

Kenchel C Doerner 13-June-2012

Dean, Graduate Studies and Research Date

CONTENTS

INTRODUCTION_____1

MATERIALS AND METHODS_____12

RESULTS_____15

DISCUSSION _____27

REFERENCES_____33

REFERENCE URL'S_____38

LIST OF FIGURES

Figure 1. Flow of data in the SEED (U21) 8

Figure 2. Flow chart depicting the general procedure used to annotate bacterial genome sequences 14

Figure 3. Statistics, for Newbler assembly of the 454 and SOLID data, from Genome Quest. 16

Figure 4. Pie chart showing the number of PEG's involved in different metabolic pathways. 19

Figure 5. Tryptophan biosynthetic pathway. 19

Figure 6. Tryptophan operon synteny. 20

Figure 7. *csd* gene mapped onto contig 2661. 23

Figure 8. The *csd* gene on contig 2661 mapped to other genomes. 23

Figure 9. Tryptophan metabolism pathways. 25

Figure 10. Alignment of contig 7372 with similar regions in other genomes. 25

Figure 11. Alignment of contig 1894 with similar regions in other genomes. 26

Figure 12. Pylogenetic tree showing the relationship between different Clostridium genus (26) 29

Figure 13. Metabolic pathway showing the involvement of butyrate kinase 31

LIST OF TABLES

Table 1. Comparison of the Newbler runs 16

Table 2. Comparison of coverage before and after assembling 18

Table 3. Distribution of PEG's in different metabolic pathway subsystems. 22

ASSEMBLY AND AUTOMATED ANNOTATION OF THE *CLOSTRIDIUM* SCATOLOGENES GENOME

Jitesh Tiwari May 2012 39 Pages

Directed by: Claire A. Rinehart, Sigrid Jacobshagen and Jonathan Quinton

Department of Biology Western Kentucky University

Clostridium scatologenes is an anaerobic bacterium that demonstrates some unusual metabolic traits such as the production of 3-methyl indole. The availability of genome level sequencing has lent itself to the exploration and elucidation of unique metabolic pathways in other organisms such as *Clostridium botulinum*. The *Clostridium scatologenes* genome, with an estimated length 4.2 million bp, was sequenced by the Applied Biosystems Solid method and the Roche 454 pyrosequencing method. The resulting DNA sequences were combined and assembled into 8267 contigs with an average length of 1250 bp with the Newbler Assembler program. Comparision of published subunits of csd gene and assembled contigs identified that one contig contained all three subunits. In addition a gene with similarity to clostridium carboxidivorans butyrate kinase was found lined next to *csd* gene. An alignment of the contig and *csd* gene sequences identified three deletions in the contig within the 4066 bases of the alignment. This implies that there is about 0.07% error rate in the sequencing itself requiring more finishing.

Even without finishing the genome assembly into single contig, contigs were annotated in RAST pipeline predicting 2521 protein encoding genes (PEGs). The PEGs were classified by their metabolic function and compared to classified PEGs found in the closely related clostridium species, *Clostridium carboxidivorans* and *Clostridium. ljungdahlii,* which have similarly sized genomes. According to the RAST analysis,

Clostridium scatologenes had 35% subsystem coverage of all known metabolic processes with its 2521 PEGs. This compares to 41% for *Clostridium carboxidivorans* with 4174 PEGs (29) and 42% for *Clostridium ljungdahlii* with 4184 PEGs (30), indicating that *Clostridium scatologenes* may still have more genes to be identified. Comparison of the percent genes found in the metabolic subsystems was similar except in motility and chemotaxis.

The contigs, on which the *csd* gene and tryptophan metabolizing genes lay, were examined to see if additional genes might support these metabolic pathways. Butyrate kinase was associated with the *csd* genes but no other associations were found for the two tryptophan metabolizing genes. The tryptophan biosynthesis operon genes were all found on one contig (contig 6771) and were syntenic with other bacterial species.

INTRODUCTION

Genome annotation is needed for the proper understanding of genes and the function to which they are associated. We can also say that the main purpose for the annotation is to gain knowledge about the cellular processes in an organism, which helps us understand how these genes work together to direct the growth, development and maintenance of an organism. In the context of pathogenic bacteria, genome sequencing projects are focused on understanding the specific mechanisms underlying bacterial survival under extreme stress environments, surviving the antibacterial drugs, etc. This information about the genome can help in designing vaccines and better drugs for controlling bacterial infections.

The genome annotation of protein encoding genes (PEGs) begins with the identification of coding domains that have the proper codon usage and translation start sites. Potential PEGs are then compared to known genes found in other organisms and the coding domain is adjusted, if necessary. If the matching genes from other organisms have a known function then the function of the PEG can be inferred. Because of high number of bacterial genomes being annotated, proper assembly and high throughput annotation tools are required to aid in identifying the genes and their functions precisely. If these tools are not able to correctly determine the gene sites and their function, the genes may be mis-annotated, which can lead to the propagation of the error as other new genomes are compared to the mis-annotated genes. Therefore, automated annotation evidence needs to undergo human review for each gene before a final annotation is released. By unraveling the gene content of a genome and the corresponding metabolic details, there is the

possibility that new doors may open and lead to the development of novel vaccines or useful antimicrobial compounds or innovative strategies to modify bacteria for applications such as bioremediation (9).

Today there are two commonly used methods for DNA sequencing, Sanger sequencing and a variety of Next Generation high-throughput sequencing methods. In the Sanger approach, DNA is cloned into a plasmid vector and then sequenced from primers specific to the vector by dideoxy chain termination (or Sanger method) (24). One of the drawbacks is that the resulting sequences may include parts of the cloning vector. Another drawback is that the cloned fragment may be toxic to the bacterium in which the recombinant vector is grown and therefore the toxic fragment would not be represented in the final sequence assembly. In the high-throughput approaches, DNA is sequenced without cloning, thus avoiding these two drawbacks. Another plus for the high-throughput sequencing is the large number of reads that can be accomplished per day per instrument and the lower cost (25). Capillary based sequencing instruments can identify up to 307,000 bp from Sanger sequencing reactions per run whereas Next Generation sequencers can generate10 billion bp per run. Because of the sample preparation time needed for cloning and the low number of sequence reads that can be accomplished per day, the Sanger sequencing method is now being replaced by the high-throughput next generation sequencing methods for genomic sequencing. Roche/454's pyrosequencing (2), Illumina's Solexa sequencing (19), Applied Biosystems SOLiD sequencing (26) and Life Technologies' Ion Torrent sequencing (27), are a few of the next generation sequencing technologies available. Next generation sequencing methods have the ability to process

billions of sequence reads in parallel rather than only a few thousand as in Sanger sequencing, thus saving time and money (2).

With millions of sequences being produced by high-throughput sequencing, there is a need to not only track the sequence data but also the quality of the reads. Phred and the 454 pyrosequencer follow the same method of assigning a score to each nucleotide in order to assemble them with a very low error rate.

Phred evolved as one of the best softwares for effective base calling of massive data and was developed during the course of the human genome project. It was the first base calling software with very low error rates. With the recent developments in DNA sequencing technologies, Phred is being replaced by the KB base caller developed by Applied Biosystems. When compared, based on a few microbial genomes, KB base caller gave higher quality reads than Phred (3).

There is a slight difference in the sequencing approach by different next generation high-throughput sequencing methods. The Roche systems (454 pyrosequencer) use native and unmodified DNA bases in its process. In the DNA preparation step, the DNA sample is sheared into small fragments that are then attached to agarose beads, one fragment to one bead. Then the DNA is amplified so that each bead carries 100k copies of the original DNA fragment in a process called Emulsion PCR. Later, the beads are loaded into picotiter plates so that, on average, each well has one bead. The sequencing reagents and one of the four DNA bases are added to start the pyrosequencing process. The intensity of the light signals that are emitted upon addition of bases to the growing sequence indicates the number of specific bases added (24). Illumina sequencing technology is based on arrays of randomly assembled glass beads, which have oligonucleotides and are covalently attached to an

immobile surface. Each bead has about one million copies of a single oligonucleotide attached to its surface. The sequence is read out through polymerase- based extension in a base-by-base fashion using either reversible terminators or sequential nucleotide addition. After the addition of the bases, the incorporated bases are identified by fluorescence (24). There have been improvements observed in the next generation sequencing techniques such as reduced sequencing errors and reduced cost per genome sequencing as well saving a lot of time and manual work. In the SOLiD sequencing system, DNA is sheared into fragments and two adapters are attached to each fragment one at each end. The fragments are then added on to bead with two different primers, one for each adaptor and two complementary strands are synthesized in the PCR Emulsion step. Afterwards, the beads are deposited on a glass slide through covalent bonds of the 3' modified beads. With the help of 8-mer probes, which have a 3' hydroxyl group, a fluorescent dye at the 5' and a cleavage site between the fifth and sixth nucleotide, the positions of the nucleotides are determined by running the cycle of adding the probes 5-7 times with the addition of a complementary universal primer. Then with the help of the fluorescent colors and by knowing the last nucleotide of the 3' end of the primer in the last round helps finds the sequence (U12). Ion semiconductor sequencing determines the release of a hydrogen ion upon addition of each nucleotide to the already existing DNA strand with the help of a hypersensitive ion sensor. Microwells on a semiconductor chip, which contain a single-stranded template DNA molecule and DNA polymerase are flooded with unmodified single nucleotides. If the nucleotides are complementary pair with the strand forming a covalent bond and releasing hydrogen ion, then with the help of this the sequence is determined. The unbound dNTP's are washed out before introducing the next cycle of dNTP's (32).

Following the generation of each sequence read and the scoring of its quality, the reads need to be assembled into contiguous sequence (contigs). Due to the limited knowledge on how these software tools are designed, choosing a befitting assembler becomes a difficult task. Newbler is the assembler developed by Roche to specially assemble the sequencing data generated by the 454 pyrosequencer. Newbler can process both single and paired end reads. It accepts only two input formats, Roche's .sff (standard flowgram format) and fasta files (U10) with or without quality files. Paired end reads can be referred to as two short sequences each at one end of the DNA molecule of interest or a little away from each other on the same DNA molecule. Many of the assembly software's identifies the paired end libraries and picks the nodes which can be a part of the same DNA molecule or a contig and based on the paired distance values joins them using the set default algorithms. The major challenges that the genome assembler's faces today are the high number of repetitive reads being produced and to join them into one long sequence. Information about the related genomes and the distribution of the genome size can help assembling software's to easily detect the repeat regions and exclude them from assembly. Other challenges faced are absence of reads or low quality reads in overlapping regions that prevent the assembly and the downstream analysis of short read datasets. Relatively high error rates further complicate the analysis of next generation sequencing data are making it the biggest challenge to assemble the whole genome into one contig (28).

There are many annotated gene databases for eukaryotic organisms but not as many for bacterial genomes. In 2006, more than 300 bacterial genome sequences were available publicly (1) and their number was increasing rapidly. With this rapid increase there was a need to standardize the representation of gene names and functions, since many of these

genes and functions were shared between organisms. The Gene Ontology (GO) consortium was formed in 1999 to provide a controlled vocabulary that could be applied across all organisms (4) (U8). At the initial stages of the consortium, all the protein coding genes from the three model systems *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Drosophila melanogaster* were included and the protein coding genes were given a unique GO ID to identify them (4). These GO IDs were hierarchically linked to provide information about the biological process, molecular function, and the functions of the genes. The GO IDs have also been linked to other existing gene and protein databases such as SwissPROT (12) (U6), Gen-Bank (13) (U14), EMBL (14) (U15), DDBJ (15) (U16), PIR (16), Pfam (17) (U7) and many more. To receive a GO ID, a gene was required to have published confirmation of its function. Therefore, not all genes have a GO ID assignment but the number is growing steadily.

Gene annotation begins by identifying the location of PEGs on the genome sequence. Several programs have been used to locate potential genes including GENSCAN and GeneMark. Many of these programs examine the statistical properties of the codon usage to identify potential genes. These are confirmed as potential PEGs by collecting additional evidence, such as the strength and adjacency of a ribosomal binding site in prokaryotes. When matches are found between the predicted PEG protein and proteins from other organisms, this strengthens the evidence that the predicted PEG is indeed a real PEG. There are several databases that can be used to check for protein matches. Each of these carries information about protein structure and function that can be inferred upon a matching PEG in need of annotation. Because the different databases have not all been manually curated for consistency, it is important to crosscheck annotations on multiple

databases. Many databases are linked to genome curation tools like Apollo (U18), Manatee (U19) and Artemis (U20) etc, allow the manual curation of the data which is being updated on a regular basis

With the rapid increase in the number of genomes being sequenced, there is a need to provide not only a consistent vocabulary (GO IDs) to also provide a mechanism to support the comparison of genes and annotations across many genomes. The SEED project has developed such a mechanism of curation by using subsystems of genes that have a related functional roles, such as a metabolic pathway, to annotate the overall function of a group of genes into their respective families (23) (U13). Figure 1 depicts the flow of annotation in the SEED framework. In order to provide SEED-quality annotations for bacterial and archaeal genomes, the RAST (Rapid Annotation using Subsystems Technology) server was developed (14)(U5). FIGfams are a set of protein sequences which are considered to follow out the same functional role with over 70% sequence similarity. RAST serves not only to provide annotation of new sequences to known subsystems, it also serves as a collection point for new subsystems that, when subsequently curated, will become new FIGfams (Fellowship for Interpretation of Genomes) that are used to classify genes entered into RAST data cycle.

An additional annotation refinement tool used to curate already existing bacterial genomes found in a number of publicly available databases is available from the Joint Genome Initiative. The Integrated Microbial Genome (IMG) Expert Review (ER) website has a rich framework of tools and publicly available genomes that can be reviewed and annotated (5) (U4). IMG ER can also be used on new bacterial annotation projects after being processed through other annotation pipelines and before they are published in

GENBANK (U15). The IMG ER addresses all the annotation problems detected by the IMG's analysis tools, like genes without assigned functions.

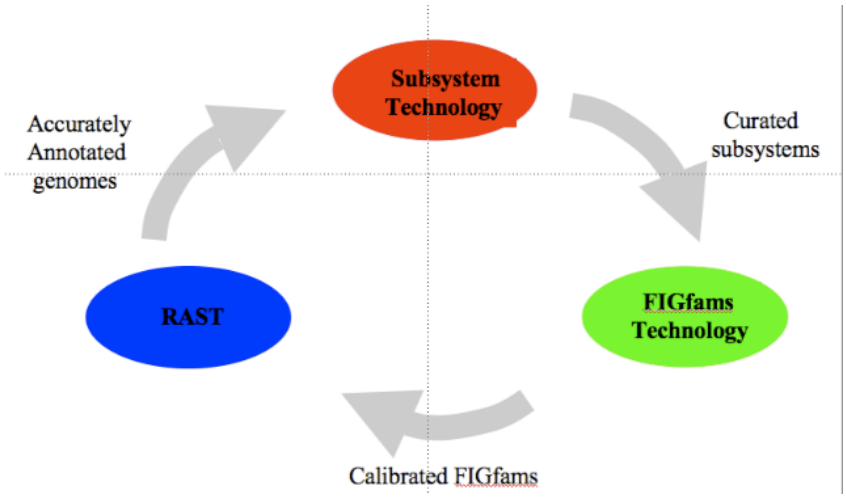

Figure 1. Flow of data in the SEED (U13). SEED framework showing the workflow on how the new genomes are being annotated and grouped under FIGfams.

With the high number of bacterial genomes being annotated, there is a need for high-throughput annotation pipelines that can automate the annotation process and compare evidences from many different database sources. These pipelines should also provide the ability to manually curate the genes and resolve conflicts between databases. IMG ER and RAST are two such high-throughput annotation pipelines available to support microbial genomes. They are online open-source high-throughput annotation pipelines, which provide service to all the users free of cost after a simple registration process. They also have a password protection system, which maintains the data privacy but allows the formation of workgroups that can share private data between several individual investigators (10). RAST has over 120 external users registered that have submitted over

350 genomes to a database of more than 1200 genomes (11). Where as “The IMG system contains a rich collection of genomes from all three domains of life: as of April 2009, IMG included 1,284 bacterial, 59 archaeal, 49 eukaryotic genomes, as well as 2,524 viruses and 924 plasmids” (U11).

Both the open source services RAST and IMG use the SEED framework for the comparative genomics approach (14) (5). The first step for the annotation process would be to import the files into the RAST pipeline, it can take a number of different input formats e.g.: fasta, .fna, 454 reads, etc. Once the files are uploaded, they are first normalized by removing duplicate copies of sequences and then a unique internal id is generated for each of the sequences. Later these unique sequences are screened for PEG’s (Protein Encoding Genes) via BlastX (18) (U5) to the linked genome databases. After processing, the results are available to view and download. The web based interface provides access to browse and further analyze the data. The output files are available as tab delimited files, which can be saved on a personal computer for further review and curation of the genes.

Another annotation service with a good user interface is xBASE (7) (U1). It currently has almost 1400 complete bacterial genomes in its database. The xBASE schema allows the addition of bacterial genomes from NCBI, GenBank and other public databases (8). The user interface allows the entry of new bacterial sequence files in fasta or text format. The bacterial genome with which the new sequence files will be compared can also be selected. Therefore, if a closely related family member is known, it can be chosen rather than including all the available genomes in the comparison.

The benefits of doing bacterial genome annotation in *Clostridium scatologenes*

Members of the genus *Clostridium* are gram-positive, spore forming bacteria that are anaerobic. Scatologenes means either an organism that produces a dung-like odor or an organism that produces skatole. *Clostridium scatologenes* is isolated from soil, contaminated food and feces of infants undersized at birth and are proposed to have mol% G+C of the DNA is 27. The 3-methyl indole molecule is malodors chemical, which is produced by the anaerobic degradation of tryptophan in few bacterial species including *Clostridium scatalogenes* and is found in stored swine manure and is intended to be responsible for foul tasting of pork. It is proposed to be produced by the metabolism of IAA (Indole Acetic Acid) and tryptophan. Sequencing and annotation of the *Clostridium scatalogenes* genome particularly its gene products involved in the tryptophan degradation pathway may prove useful in identifying out the factors responsible for the production of 3-methyl indole (6) as the pathway by which 3-methylindole is produced from tryptophan is not yet explained.

The protein product of the *csd* gene, 4-hydroxyphenylacetate decarboxylase, belongs to the family of lyases which specifically cleaves carbon-carbon bonds. The enzyme 4-hydroxyphenylacetate decarboxylase from *Clostridium scatologenes* catalyzes the decarboxylation of p-hydroxyphenylacetate to yield the cytotoxic compound p-cresol. The metabolic toxicity of p-cresol allows the suppression of other microbes and thus provides a growth advantage in highly competitive environments (20). This enzyme is synthesized as large precursor polypeptides and requires post-translational activation by dedicated iron-sulfur proteins for catalytic activity. The *csd* gene is the only sequenced gene that is publicly available for *Clostridium scatologenes*.

The genomic context of the genes used in tryptophan degradation and the *csd* genes are of interest in order to discover the metabolic pathways in which they are expressed. In this study the assembly of the genome along with the annotation and classification of its genes into metabolic pathways was used to examine the context of these genes for clues about the metabolic role that they play in the production of malordorants and cytotoxic compounds.

MATERIALS AND METHODS

Geneious

Geneious is a multipurpose informatics analysis program that manages sequences, translates, compares, aligns, and assembles DNA and protein sequences (U9). The sequence assembler in Geneious can handle reads of any length and from any sequencing software, the assembler also reports read errors consisting of incorrect bases. The de novo assembly algorithm is similar to the one used in multiple sequence alignment.

GenomeQuest

GenomeQuest is a commercial online web based tool kit that provides access to use Newbler and other Metagenomic annotation pipeline tools (U22). Newbler was used for assembling the sequences from SOLiD and 454 pyrosequencer together. The minimum overlap parameter was set to 25 and the percentage identity was set to 80%. Once the assembly was accomplished, a Metagenomics annotation search was performed on Genome Quest (U14) to determine the bacterial genomes most closely matching *Clostridium scatologenes*.

BLAST

BLAST (Basic Local Alignment Search Tool) is a tool used to compare DNA and protein query sequences to the sequences found in databases or local search files (U2). It uses a rapid search algorithm and gives a list of pairwise alignments which are ranked by a score.

IMG ER

IMG ER is an online annotation pipeline that allows for the functional annotation and manual curation of genomes (5) (U6). Their expert reviews tools help refine annotation before submission of a genome to Genbank.

RAST

RAST (Rapid Annotation Using Subsystem Technology) is a fully-automated online service for annotating bacterial genomes (14) (U3). It uses the SEED framework which compares families of genes called FIGfams to input sequences and assigns annotations. The curation of annotations is done by experts, who annotate subsystems of genes across many genomes and then recompile the curated annotations into FIGfams (Figure 1) (28) (U21).

The RAST annotation pipeline automates the workflow illustrated in Figure 2, starting with gene prediction and compiling annotation evidence to be presented to the human annotator at the data visualization end of the pipeline. The human annotator then has the option to accept the annotation evidence compiled in the pipeline or editing it. RAST and IMG ER both have similar pipeline algorithms and are both based on the SEED framework. A rich set of comparison tools in RAST allows multiple genes, subsystems and genomes to be simultaneously compared and visualized.

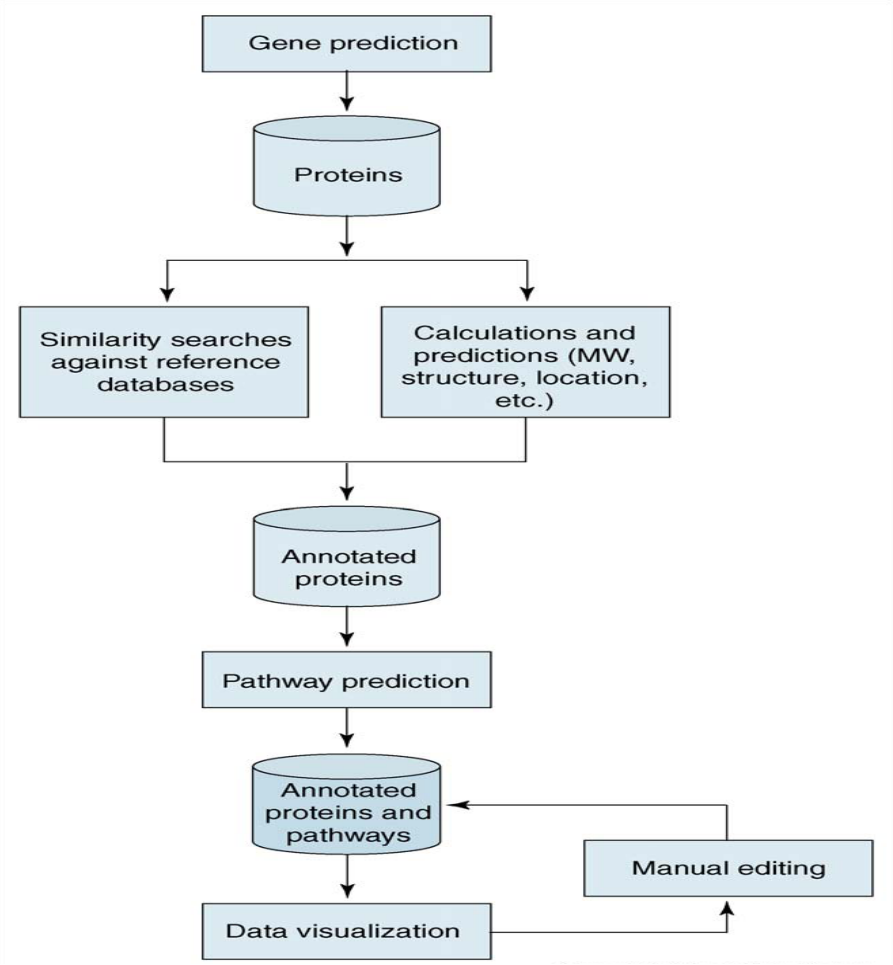

Figure 2: Flow chart depicting the general procedure used to annotate bacterial genome sequences in general (1).

RESULTS

Clostridium scatologenes genomic DNA has been sent to the Laboratory of Genomics and Bioinformatics at University of Oklahoma Health Sciences Center and sequenced using the Applied Biosystems SOLiD technology. The assembled genome was returned as 10,096 contigs. A second round of sequencing was done using the Roche/454 sequencing technology at the University of Illinois, which yielded 227,000 reads. The files received contained both the sequence files and the quality files. Based on the total number of bases sequenced recovered from the 454 sequencing and from estimating the genome length of *Clostridium scatologenes* to be about 4 million base pairs, the depth of coverage at each position in the genome should average about 16.7 folds.

Assembly

Newbler was the read assembler software used for genome assembly. Newbler was designed to be used with 454 sequence data. After running the Newbler assembler on the Genome Quest website for the 454 pyrosequenced reads, it gave 10,304 contigs with an average length of 1059 bp (Table 1). The Newbler assembly of the 454 reads and the contigs from the SOLID sequencing gave 8,267 contigs that averaged 1250 bp in length as shown in figure 3 and Table 1. These 8,267 contigs resulting from the combined assembly runs were used for the annotation process.

NEWBLER Run's	Average length of the contig's.	Number of contig's.	Length of the longest contig.
Newbler 454	1059	10304	7124
Newbler 454 and SOLID	1250	8267	10061

Table 1: Comparison of the read assembler software Newbler. Comparing the average length and the number of contigs being produced by Newbler on 454 reads and combined 454 and SOLID reads.

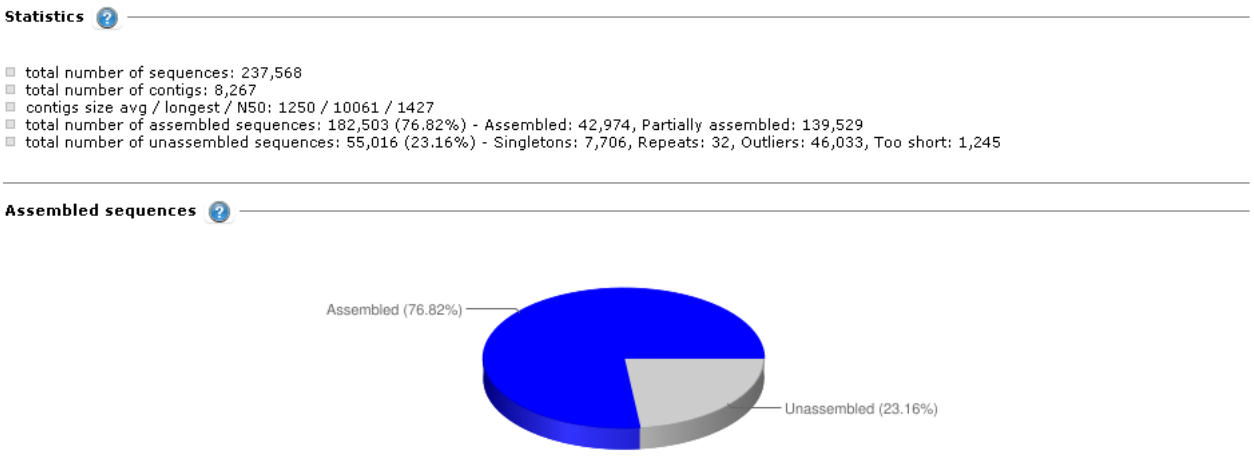

Figure 3: Statistics, for Newbler assembly of the 454 and SOLID sequencing data as obtained from Genome Quest. Partially assemble: the sequences which were not included wholly in the assembly, singletons- those sequences that could not be assembled to another sequence from the input; outliers- the read was identified as problematic and excluded from the final assembly; repeats- reads that were either inferred to be repetitive when compared to other reads from the input, or reads that partially overlapped a contig; too short- the trimmed read was too short to be used in the assembly.

Metagenomics annotation

The assembled files from the Newbler were used to run a metagenomics search to find the most closely related genomes to *Clostridium scatologenes.* Assembled files were uploaded to the metagenomics pipeline available on Genome Quest. The output revealed that the most closely related organism to *Clostridium scatologenes* in the genome database is *Clostridium carboxidivorans*. This is in accordance with the metagenomics annotation from RAST which also revealed that *Clostridium scatologenes* is closely related to *Clostridium carboxidivorans*.

Annotation

Following assembly, the 8,267 contigs needed to be annotated. Initially, a small subset of 3500 assembled contigs was submitted to three annotation pipelines: RAST, IMG ER, and xBASE. The most useful results were generated from RAST and IMG ER, which identified transcription start and stop information and gene names. The xBASE pipeline would only give the protein IDs of the genes and no information about where the genes started and stopped on the contigs. IMG ER took almost twenty days to complete the annotation pipeline with the subset of the contigs. RAST was able to do the same job within 24 hrs with nearly the same precision and accuracy. When comparing the output from all the three annotation pipelines, all three gave a similar number of the genes ranging from 624-721. Hence, RAST was used for further runs on the on all of the 8,267 contigs. This resulted in the identification of 2521 PEGs

Annotation Pipeline	Average run time	No. of genes predicted
RAST	~24hrs	624
xBASE	~24 hrs	721
IMG ER	15 days	711

Table 2: Comparison of different annotation pipelines. Comparing the average run time and the number of genes predicted by different annotation pipelines. Numbers of genes predicted are excluding duplicates.

The 2521 PEGs for *Clostridium scatologenes* can be sorted into several functional subsystem pathways as shown in Figure 4. The hierarchal arrangement shows the number of genes that are found in each subsystem. For example, tryptophan synthesis has 11 genes listed. A map of the tryptophan biosynthetic pathway (Figure 5) shows that *Clostridium scatologenes* contains all the necessary genes to code for the enzymes required for the products of glycolysis to be converted to tryptophan product (green highlighted EC numbers). When the tryptophan operon of other closely related was compared to *Clostridium scatologenes* there was a distinct synteny, conservation of gene order, found for the genes located on contig 6771 as shown in Figure 6.

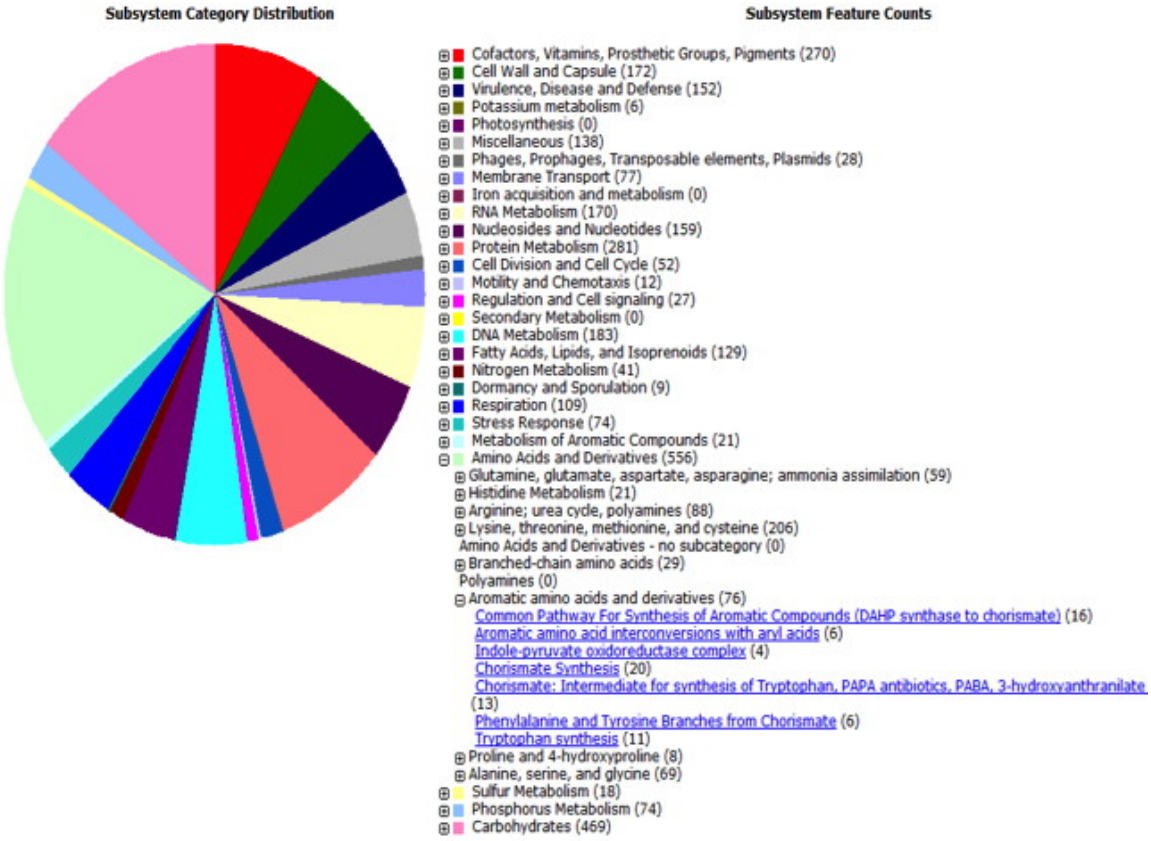

Figure 4: Pie chart showing the number of PEG's (Protein Encoding Genes) involved in different subsystem pathways. Subsystem category distribution with the number of genes involved in different metabolic pathways is depicted, with 11 of them involved in Tryptophan synthesis.

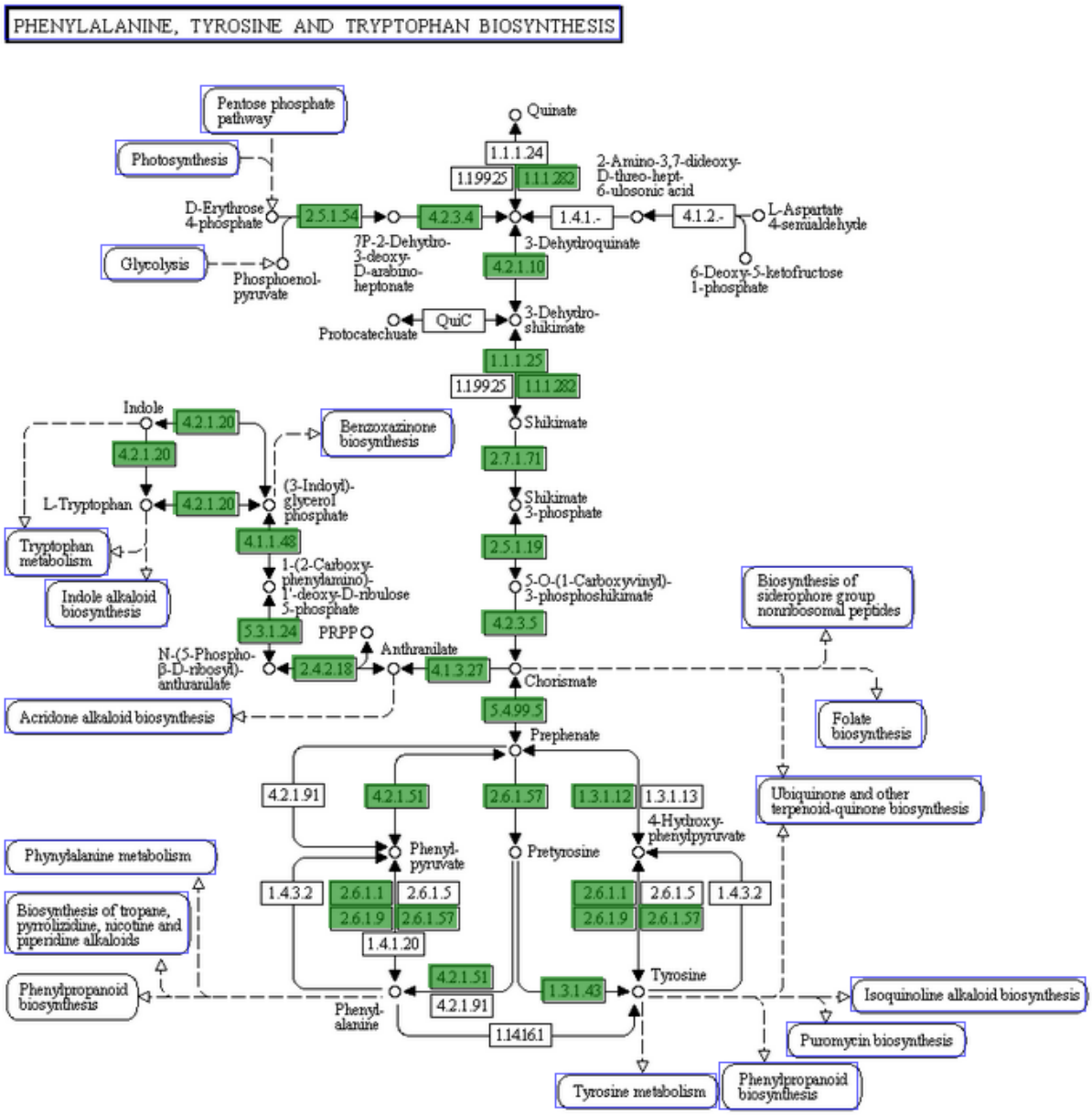

Figure 5: Tryptophan biosynthetic pathway. The numbers in boxes are the EC numbers of the enzymes needed to convert substrates to products in the direction of the arrows. The green boxes show that the genes for these enzymes are found in *Clostridium scatologenes*.

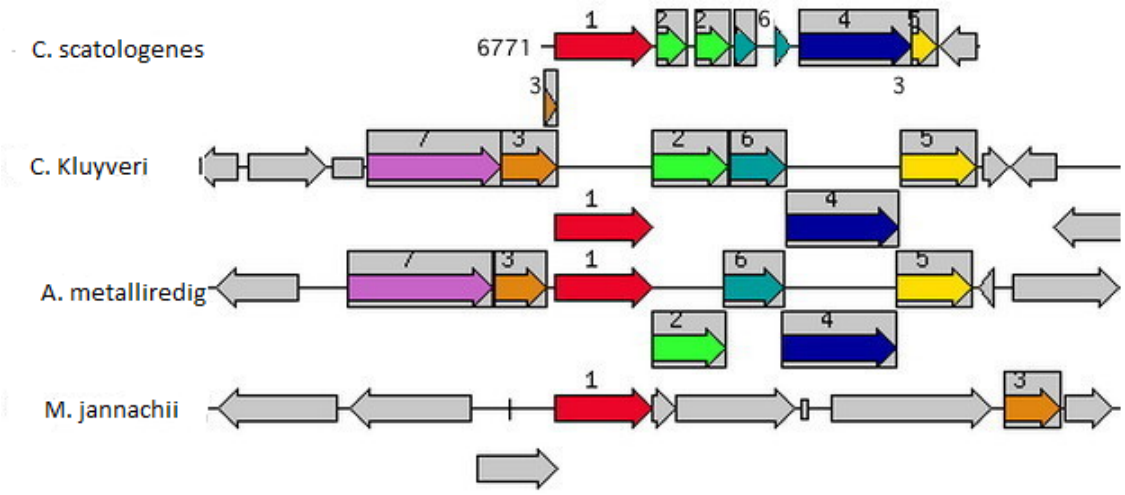

Figure 6: Tryptophan operon synteny. EC 2.4.2.18 (Red 1), EC 4.1.1.48 (Green 2), EC 4.1.3.27 (Brown 3 and Purple 7), EC 5.3.1.24 (Turquoise 6) EC 4.2.1.20 (Blue 4 and Yellow 5). The genes for protein 2 and protein 6 are shown a split due to indels (insertions or deletions) when aligned with *Clostridium kluyveri* in the contig sequence that shifted the reading frames.

The protein encoding genes identified by RAST for *Clostridium scatologenes*, *Clostridium carboxidivorans* and *Clostridium ljungdahlii*, were classified by their metabolic functions and the percentage of the genes involved in each of those metabolic pathways were compared to check for any significant deficiencies (Table 3). "Motility and Chemotaxis" subsystem seems to be underrepresented in *Clostridium scatologenes*. *Clostridium ljungdahlii* seems to be overrepresented in the "Cofactors, Vitamins, Prosthetic Groups, Pigments" and "Dormancy and Sporulation" subsystems. Otherwise the 2457 PEGs seem to be evenly represented across the subsystems. When looking at individual genes, 1685 of the 2457 PEG's found in *Clostridium scatologenes* were also found in *Clostridium carboxidivorans*. RAST estimated that there may be up to an additional 58 PEGs that have not been called based on the number of base pairs in gaps longer than 2 kbp and an estimate that each PEG occupies about 1 kbp.

Subsystem Categories	**C.scatologenes**	**C.carboxidivorans**	**C.ljungdahlii**
Cofactors, Vitamins, Prosthetic Groups, Pigments	4.08%	3.60%	**5.96%**
Cell Wall and Capsule	2.60%	2.28%	4.13%
Virulence, Disease and Defense	2.295%	2.40%	2.39%
Potassium metabolism	0.090%	0.095%	0.16%
Photosynthesis	0%	0%	0%
Miscellaneous	2.08%	2.85%	3.57%
Phages, Prophages, Transposable elements, Plasmids	0.42%	0.02%	0.46%
Membrane Transport	1.16%	1.65%	1.25%
Iron acquisition and metabolism	0%	0.04%	0%
RNA Metabolism	2.57%	2.90%	3.73%
Nucleosides and Nucleotides	2.4%	3.285%	3.62%
Protein Metabolism	4.24%	5.08%	5.80%
Cell Division and Cell Cycle	0.78%	0.67%	0.69%
Motility and Chemotaxis	**0.18%**	**2.59%**	**2.36%**
Regulation and Cell signaling	0.41%	0.53%	0.55%
Secondary Metabolism	0%	0%	0%
DNA Metabolism	2.76%	2.66%	2.62%
Fatty Acids, Lipids, and Isoprenoids	1.95%	1.73%	1.67%
Nitrogen Metabolism	0.62%	0.815%	0.69%
Dormancy and Sporulation	0.135%	0.89%	**1.81%**

Respiration	1.645%	2.01%	1.76%
Stress Response	1.18%	1.80%	1.11%
Metabolism of Aromatic Compounds	0.32%	0.215%	0.16%
Amino Acids and Derivatives	8.39%	9.16%	9.8%
Sulfur Metabolism	0.27%	0.96%	0.32%
Phosphorus Metabolism	1.18%	1.03%	1.02%
Carbohydrates	7.08%	7.36%	5.78%

Table 3: Distribution of PEG's in different metabolic pathway subsystems. Those subsystems with gene contributions significantly different from *Clostridium scatologenes* are shown in bold

csd Gene

The published *csd* gene from *Clostridium scatologenes* has three different subunits: 4- hydroxyphenylacetate decarboxylase glycyl radical subunit, 4- hydroxyphelyacetate decarboxylase small subunit and 4- hydroxyphenylacetate decarboxylase activating enzyme (20). A BLAST comparison of the *csd* gene to the 8,267 assembled contigs identified one, contig 2166, that contained all the three subunits lined up next to each other as shown in Figure 7. A DNA sequence alignment of contig 2166 with the nucleotide sequence for the *csd* genes from *Clostridium scatologenes* retrieved from the NCBI GenBank showed that there were two deletions and one insertion (indels) of 1basepair each in contig 2166 that were not found in the published *csd* genes. This gives a sequencing error rate of 0.07 % (= 3/4066). A BLAST search of the region of contig 2166 that did not align with the *csd* genes identified an additional gene on this contig matching butyrate kinase from *Clostridium carboxidivorans*. Contig 2166 was aligned with genes similar to the *csd* genes in other genomes (Figure 8) to see if there was any synteny, or conservation in gene order, with the

surrounding genes. The *csd* genes mapped to pyruvate formate lyase, pyruvate formate lyase activating enzyme, 4-hydroxyphenylacetate decarboxylase regulatory subunit in *Clostridium botulinum* and the other similar genomes. Note that the names of the first two subunits are different in *Clostridium scatologenes* than in the similar genomes due to different naming conventions between the databases. Figure 5 shows no synteny beyond the *csd* genes.

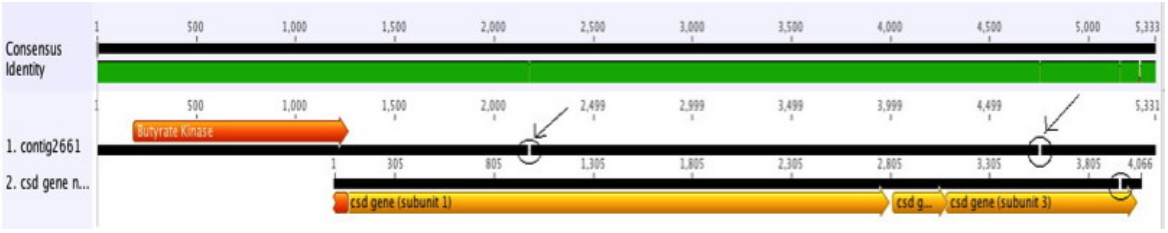

Figure 7: *csd* gene mapped onto contig 2661. All the three subunits of the *csd* gene: 4-hydroxyphenylacetate decarboxylase glycyl radical subunit, 4- hydroxyphelyacetate decarboxylase small subunit and the 4- hydroxyphenylacetate decarboxylase activating enzyme, mapped on to contig 2661 along with the butyrate kinase enzyme which had the top BLAST hit from *Clostridium carboxidivorans*. The three observed deletions are marked with circles.

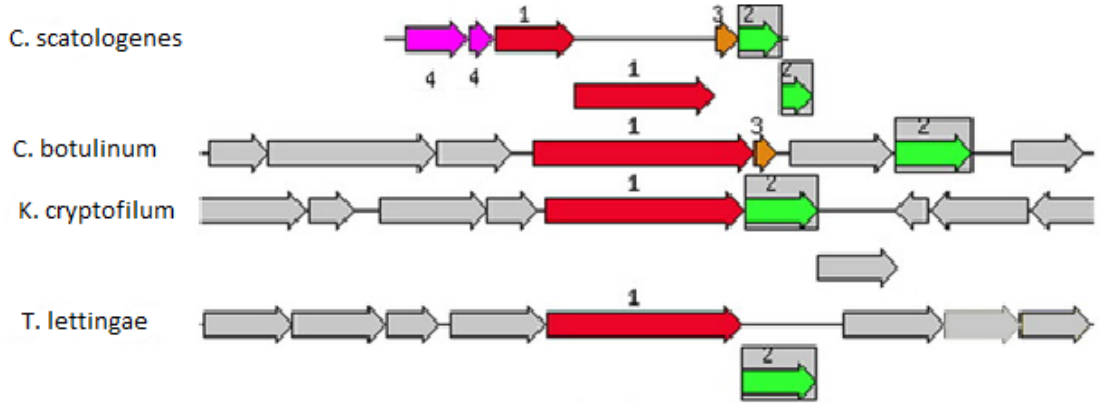

Figure 8: The *csd* gene on contig 2661 mapped to other genomes. All the three subunits of the *csd* gene (1 = 4- hydroxyphenylacetate decarboxylase glycyl radical subunit, 2 = 4-hydroxyphelyacetate decarboxylase small subunit and 3 = 4- hydroxyphenylacetate decarboxylase activating) and the gene for butyrate kinase (4) were mapped onto other closely related genomes to check for synteny. Genes 1, 2 and 4 in Clostridium scatologenes are shown as split due to a shift in reading frame resulting from indels (insertion or deletions) in the contig sequence.

Tryptophan Metabolism Genes

One of the purposes for sequencing the *Clostridium scatologenes* genome was to try to elucidate the enzymes that may be involved in the metabolism of tryptophan into the malordorant 3-methyl indole. Figure 9 shows the metabolic pathways that use tryptophan. Two of the enzymes, EC 1.13.12.3 and EC 3.5.5.1 have genes on *Clostridium scatologenes* and appear to be part of a pathway leading to indoleacetate. Since the genes that actually produce 3-methyl indole may not have been identified with this function, the contigs that these two genes are located on were examined to see if there were adjacent hypothetical genes that may also be related to this function.

Figure 10 shows the alignment of contig 7372, which contains the gene for EC 1.13.12.3, with related genomes. Contig 7372 was not large enough to show other genes or potential genes. However, if this region of the genome is syntenic then it is interesting to note that the adjacent two genes in *Clostridium beijerincki* are both hypothetical proteins. A BLAST search of each of these hypothetical proteins against the 8267 contigs from *Clostridium scatologenes* identified conserved sequences on contigs 4092 and 4932. The genes on both of these contigs are annotated as “secreted protein containing uncharacterized conserved protein of ErfK family”.

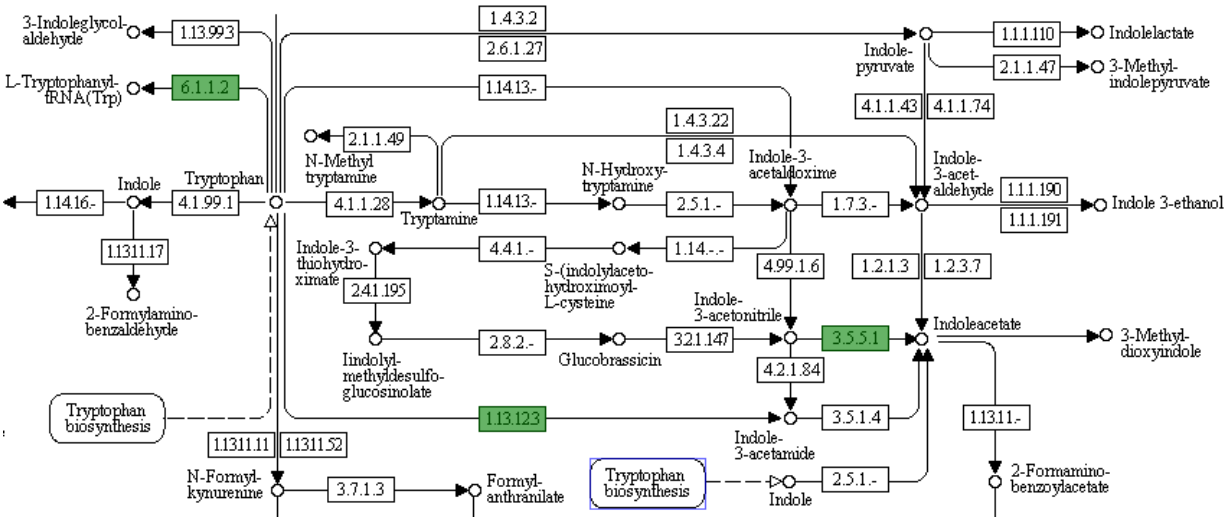

Figure 9. Tryptophan metabolism pathways. Three *Clostridium scatologenes* genes are found in pathways that metabolize tryptophan (EC numbers shown in green boxes). EC 6.1.1.2 is used to charge tryptophan onto tRNAs. EC 1.13.12.3 and EC 3.5.5.1 seem to be part of a metabolic pathway leading to indoleacetate.

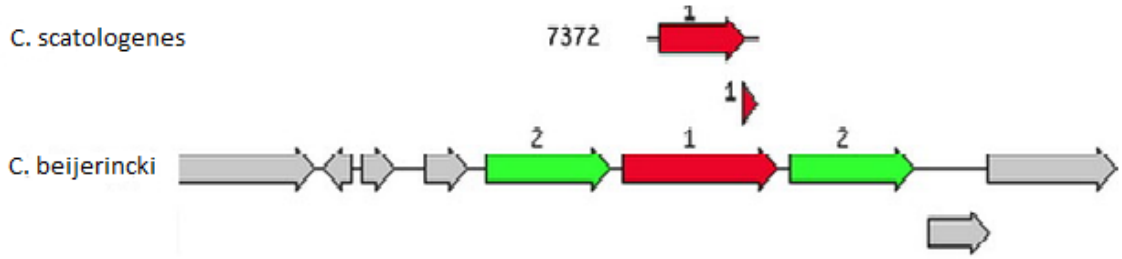

Figure 10. Alignment of contig 7372 with similar regions in other genomes. Only one gene was found on contig 7372 corresponding to EC 1.13.12.3. Genes labeled 2 in green are both hypothetical proteins with undefined functions. Gene 1 for *Clostridium scatologenes* is split due to an indel (insertion or deletion) in the contig sequence.

Figure 11 shows the alignment of contig 1894, which contains the gene for EC 3.5.5.1, with related genomes. Five genes are contained on this contig. Two of them are hypothetical proteins but could not be part of an operon with EC 3.5.5.1, because they are transcribed off from the opposite strand.

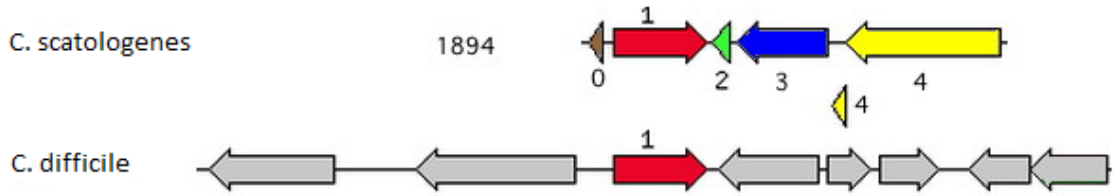

Figure 11. Alignment of contig 1894 with similar regions in other genomes. Five genes were identified on contig 1894 corresponding to: predicted transcriptional regulator of pyridoxine metabolism (Brown 0), EC 3.5.5.1 (Red 1), hypothetical protein (Green 2), hypothetical protein (Blue 3), EC 1.12.7.2 periplasmic hydrogenase (Yellow 4). Gene 4 for *Clostridium scatologenes* is split due to an indel (insertion or deletion) in the contig sequence.

DISCUSSION

The main goal of the genome sequencing projects is to gain knowledge about the cellular processes that occur in an organism. Understanding how an organism develops, survives and adapts to different environmental conditions can be aided by trying to comprehend the pathways and regulation networks. An important step in gaining this valuable information is annotation, which primarily involves identifying the genes, predicting their function and most importantly their functional relationships. If genes are mis-annotated and persist in databases, they have the potential of being propagated during the annotation of new genes that match them. Hence, there have been many different annotation methods developed to ensure that new genes are not falsely annotated, most of which rely on hand curation of key datasets or the use of many confirming datasets.

Genome annotation includes the identification of the location of protein coding and non-protein coding genes, the functional identity of genes and the manual curation of these identities. The objective behind annotation pipelines is to automate much of the identification process as accurately and rapidly as possible and present evidence for the confirmation of the annotations. Every dataset has some incorrect information and therefore requires either regular maintenance or the dynamics of adding new information to the dataset while improving the overall quality. The RAST annotation pipeline, coupled with the curation cycle found in the SEED technology that drives it, provides a cycle of continued improvement as new data are compared, curated and incorporated into the workflow (Figure 1). Studies on the RAST server have estimated the false-negative prediction rate at between 1.3% and 2.1% (14).

Many bacterial genomes have been sequenced and annotated. Most free-living bacteria have at least 1000 genes. RAST identified 2457 protein encoding genes in *Clostridium scatologenes*. Other closely related clostridium species such as *Clostridium carboxidivorans* and *Clostridium ljungdahlii* have 4174 and 4184 PEGs respectively and therefore *Clostridium scatologenes* probably has some additional genes. Based on the number of bases in gaps between identified genomic features (PEGs or RNAs) and the average length of a PEG, RAST has estimated that the current set of contigs has the potential to code for an additional 58 PEGs. Therefore, even though there is an even coverage of subsystems within *Clostridium scatologenes* (Table 3), there is a strong suggestion that the number of PEGs will expand as sequencing is finished.

The deletions observed while aligning the contig 2661 and the whole nucleotide sequence of *csd* gene shows that there are some indels (insertions or deletions) in the contigs (Figure 7). The estimated error rate for indels is 0.07% for the current project. Indels are most obvious in coding regions because they shift the reading frame and BLAST searches as well as alignments will identify both segments of the split gene. This is clearly illustrated in the alignments of the *csd* genes, the butyrate kinase gene, some of the tryptophan biosynthesis genes and the degradation genes as shown in figures 8, 9, 11 and 12.

The Metagenomics search done with GenomeQuest and the subsystem comparisons done by RAST identified *Clostridium carboxidivorans* as one of the closest sequenced genomes to *Clostridium scatologenes*. Furthermore, *Clostridium scatologenes* and *Clostridium carboxidivorans* share 1685 PEGs in common. This confirms at the genome

level the work of Liou *et al*. (21), who analysed the16S rRNA gene from several species to determine their phylogenetic relationships and found that *Clostridium carboxidivorans* P7 is closely related to two *Clostridium scatologenes* strains, ATCC 25775 (99.75% identity) and SL1 (99.85% identity) (26). DNA-DNA reassociation and 16S rRNA were used to investigate the taxonomic relationship between the bacterial strains (22). The DNA-DNA reassociation value of *Clostridium carboxidivorans* with respect to *Clostridium scatologenes* strain ATCC 25775 was 50.2% and with respect to SL1 strain ATCC 25775 was 53%, with the threshold value of 70%. Based on the 16S rRNA and DNA-DNA reassociation values the strain SL1 was renamed *Clostridium drakei*. The above results demonstrate that the strains P7, ATCC 25775 and SL1represent distinct but closely related species within the genus *clostridium* (Figure 12).

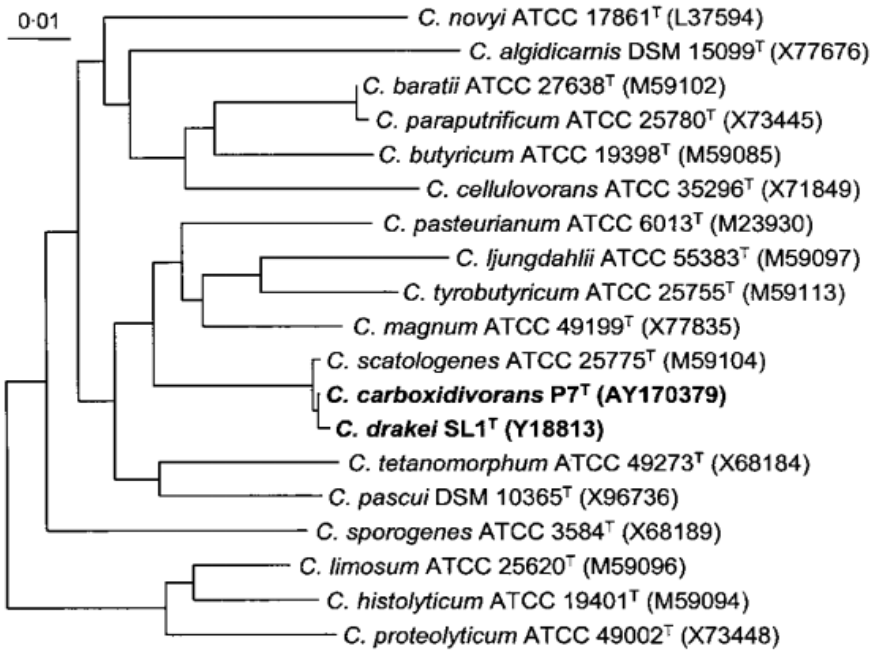

Figure 12: Pylogenetic tree showing the relationship between different species of the genus *Clostridium* (21). Tree showing *Clostridium scatologenes, Clostridium carboxidivorans,* and *clostridium drakei* as closely related family members.

The other relevant questions of this project were which genes are involved in the production of p-cresol and the malodorant 3-methyl indole in *Clostridium scatologenes*? One way to answer these questions is to look at the context that genes are located in the metabolic pathways. Often, the genes for the same metabolic pathways are present on an operon that is transcribed as a single unit. Identifying all the genes on an operon greatly aids in the definition of the pathway of metabolic transformation that turn substrates into products. Another way is to look for synteny between genomes from different organisms. The conservation of gene order or placement (synteny) is often associated with conservation of functions that may operate together. Another approach would be to map sequences onto a closely related genome and look at the coverage and ordering of the genes which can help in mapping any similar genes which might not have got mapped on the test genome.

In this study, the genome sequence for *Clostridium scatologenes* was represented in 8,267 contigs. When looking for contextual genes, the size of the contigs limited the extent to which associations could be made. For the *csd* genes, butyrate kinase was found to be associated and expressed in the same direction and therefore could be transcribed as a member of a polycistronic operon. The metabolic pathway containing this enzyme is shown in Figure 13. It does not have any obvious link to the synthesis of p-cresol.

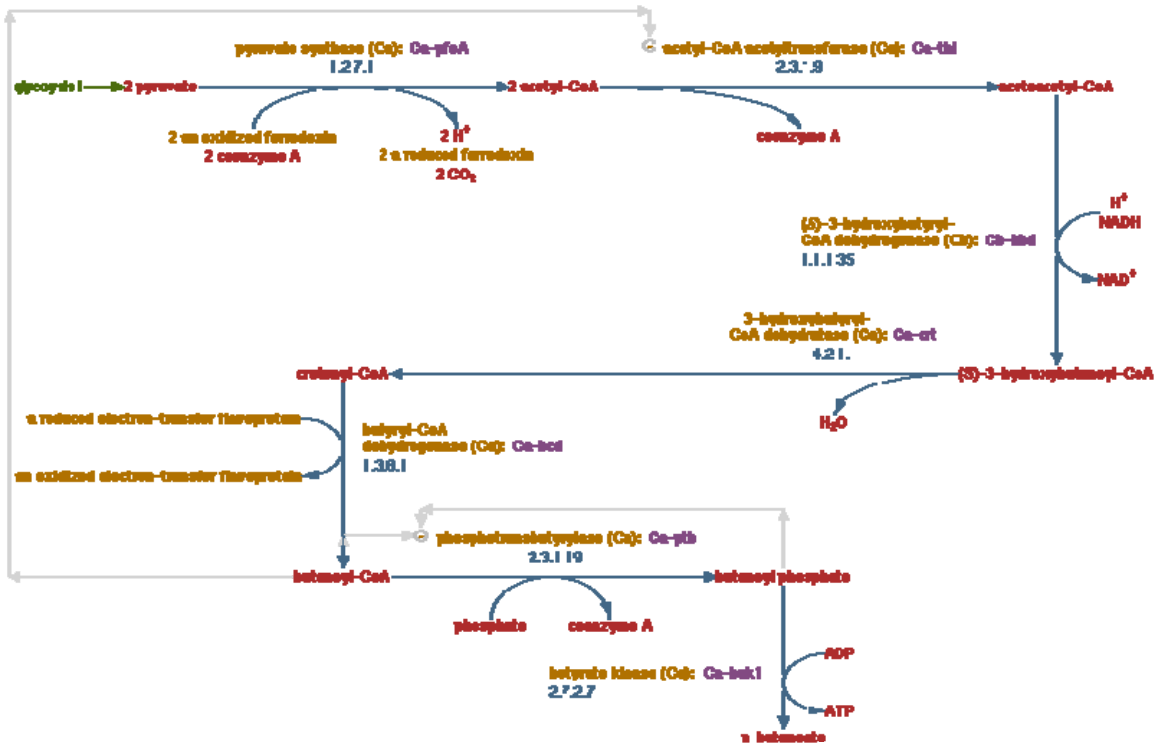

Figure 13: Metabolic pathway showing the involvement of butyrate kinase.

Two genes were found associated with enzymes for tryptophan degradation leading to the synthesis of indoleacetate, EC 1.13.12.3 and EC 3.5.5.1. Since the enzymes that actually produce 3-methyl indole may not have been identified with this function, observation of hypothetical protein, whose genes lie adjacent to the genes for these two enzymes may require further investigation.

The gene for the enzyme EC 1.13.12.3 lays on a short contig with not even the whole gene represented (Figure 10). However, if both the adjacent genes in *Clostridium beijerincki* are syntenic to this genomic region then it would be interesting to note that the genes are hypothetical proteins. A BLAST search of each of these hypothetical proteins against the 8267 contigs from *Clostridium scatologenes* identified conserved sequences on

contigs 4092 and 4932. The genes on both of these contigs are annotated as “secreted protein containing uncharacterized conserved protein of ErfK family”.

The gene for the enzyme EC 3.5.5.1 is located on a contig with four other genes, all of which are transcribed from the opposite strand and therefore could not be part of an operon with this gene (Figure 11). Two of the genes encode hypothetical proteins. One of the genes codes for the enzyme EC 1.12.7.2, a periplasmic hydrogenase. The last gene codes for a predicted transcriptional regulator of pyridoxine metabolism.

In conclusion, the context of the genes involved in the synthesis of p-cresol and 3-methyl indole have been examined in the *Clostridium scatologenes* genome and other genes were found to be in the adjacent context or inferred to be in the adjacent context. None of them seem to be directly related to the synthesis or regulation of these two compounds. However, further sequence finishing and contig joining will yield longer sequences covering these regions. Then perhaps syntenic conservation of genes within these regions may shed further light on the pathways responsible for p-cresol and 3-methyl indole synthesis and regulation.

REFERENCES

1. Stothard, P., & Wishart, D. S. (2006). Automated bacterial genome analysis and annotation. Current Opinion in Microbiology. 9:505–510.
2. Balzer, S., Malde, K., Lanzén, A., Sharma, A., & Jonassen, I. (2010). Characteristics of 454 pyrosequencing data—enabling realistic simulation with flowsim. Oxford Journal. 26:i420-i425.
3. Ewing, B., Hillier, L., Wendl, Michael C., & Green, P. (1998). Base-calling of automated sequencer traces using phred. i. accuracy assessment. Genome Res. 8:175-185.
4. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., Sherlock, G. (2000) Gene ontology: tool for the unification of biology. Nat. Genet. 25:25-9.
5. Markowitz, V. M., Mavromatis, K., Ivanova, N. N., Chen, I. A., Chu, K., & Kyrpides, N. C. (2009). Img er: a system for microbial genome annotation expert review and curation. Oxford Bioinformatics. 25:2271–2278.
6. Doerner, K. C., Cook, K. L., & Mason, B. P. (2008). 3-methylindole production is regulated in clostridium scatologenes atcc 2577. Letters in Applied Microbiology. 48: 125-132.
7. Chaudhuri, R., & Palle, M. J. (2006). xbase, a collection of online databases for bacterial comparative genomics. Nucleic acids Res 34:D335–D33.

8. Chaudhuri, R. R., Loman, N. J., Snyder, L. A. S., Bailey, C. M., Stekel, D. J., & Pallen, M. J. (2008) xBASE2: a comprehensive resource for comparative bacterial genomics. Nucleic Acids Res. 36:D543-D546.

9. Stothard, P., & Wishart, D. S. (2006). Automated bacterial genome analysis and annotation. Current Opinion in Microbiology. 9:505–510.

10. Meye, F., Paarman, D., M. D., Olso, R., Glas, E., Rodrigue, A., Steven, R., & Wilk, A. (2008). The metagenomics rast server – a public resource for the automatic phylogenetic and functional analysis of metagenome. BMC Bioinformatics. 9:386.

11. Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G. D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., Zagnitko, O. (2008). The RAST server: Rapid annotations using subsystems technology. BMC Genomics 9:75.

12. Bairoch, A., Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 28:45–48.

13. Benson, D. A., Karsch-Mizrachi, L., Lipman, David J., Ostell, J., Rapp, Barbara J., & Wheeler, David L. (2000) GenBank. Nucleic Acids Res 28:15–18.

14. Baker, W., Broek, Alexandra van den., Camon, E., Hingamp, P., Sterk, P., Stoesser, G., & Tuli, Mary A. (2000) The EMBL Nucleotide Sequence Database. Nucleic Acids Res 28:19–23.

15. Tateno, Y., Miyazaki, S., Ota, M., Sugawara, H., & Gojobori, H. (2000) DNA Data Bank of Japan (DDBJ) in collaboration with mass sequencing teams. Nucleic Acids Res 28:24–26.

16. Barker, Winoma C., Garavelli, John S., Huang, H., McGarvey, Peter B., Orcutt, Bruce C., Srinivasarao, Geetha Y., Xiao, C., Yeh, Lai-su L., Ledley, Robert S., Janda, Joseph F., Pfeiffer, F., Mewes, Hans-werner., Tsugita, A., & Wu, C. (2000) The Protein Information Resource (PIR). Nucleic Acids Res 28:41–44.

17. Bateman, A., Birney, E., Durbin, R., Eddy, Sean R., Howe, Kevin L., & Sonnhammer, Erik L L. (2000) The Pfam protein families database. Nucleic Acids Res 28:263–266.

18. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipmann, D. J. (1990) Basic local alignment search tool. J Mol Biol 215:403-420.

19. Fan, J., Chee, M. S., & Gunderson, K. L. (2006). Highly parallel genomic assays. Nat Rev Genet 7(8):632-644.

20. Yu, L., Blaser, M., Andrei, P. I., Pierik, A. J., & Selmer, T. (2006). 4-hydroxyphenylacetate decarboxylases: properties of a novel subclass of glycyl radical enzyme systems. Biochemistry. 45:9584–9592.

21. Liou, J. S., Balkwill, D. L., Drake, G. R., & Tanner, R. S. (2005). Clostridium carboxidivorans sp. nov., a solvent-producing clostridium isolated from an agricultural settling lagoon, and reclassification of the acetogen clostridium scatologenes strain sl1 as clostridium drakei sp. IJSEM. 55:2085–2091.

22. Stachkebrand, E., & Goebe, B. M. (1994). Taxonomic note: A place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriolog. IJSEM. 44:846-849.

23. Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E. D., Gerdes. S., Glass, E. M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A. C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G. D., Rodionov, D. A., Rückert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O., Vonstein, V. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 33:5691-702.

24. Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA. 74:5463-5467.

25. Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. Nat Biotechnol. 26:1135 - 1145.

26. Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J. A., Costa, G., McKernan, K., Sidow, A., Fire, A., & Johnson, Steven M. (2008). A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. Genome Res. 18:1051–1063.

27. Pareek, C. S., Smoczynski, R., & Tretyn, A. (2011). Sequencing technologies and genome sequencing. Journal of Applied Genetics. 52:413–435.

28. Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J., & Shen, B. (2011). A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. Plos One 6:e17915.

29. Paul, D., Austin, F.W., Arick, T., Bridges, S.M., Burgess, S.C., Dandass, Y. S., Lawrence, M. L. (2010) Genome sequence of the solvent-producing bacterium Clostridium carboxidivorans strain P7T. Journal of Bacteriology. 192:5554-5.

30. Kopke, M., Held, C., Hujer, S., Liesegang, H., Wiezer, A., Wollher, A., Ehrenreich, A., Liebl, W., Gottschalk, G., Durre, P., (2010) Clostridium ljungdahlii represents a microbial production platform based on syngas. Proc Natl Acad Sci USA 107:13087-92.

REFERENCE URL'S

U#	Tool Name	URL
U1	xBASE	http://www.xbase.ac.uk/annotation/
U2	BLASTN	ftp://ftp.ncbi.nlm.nih.gov/blast
U3	RAST	http://rast.nmpdr.org/rast.cgi
U4	IMG ER	https://img.jgi.doe.gov/cgi-bin/er/main.cgi
U5	Blast X	ftp://ftp.ncbi.nlm.nih.gov/blast
U6	SWISS-PROT	http://www.expasy.org/sprot/
U7	Pfam	http://pfam.sanger.ac.uk/
U8	Gene Ontology	http://www.geneontology.org/
U9	GENEIOUS	http://www.geneious.com/assets/documentation/geneious/Geneious Manual.pdf
U10	FASTA	http://www.ebi.ac.uk/fasta33/
U11	IMG	http://www.jgi.doe.gov/News/news_09_05_18.html
U12	SOLiD	http://en.wikipedia.org/wiki/2_Base_Encoding
U13	SEED	http://www.theseed.org/wiki/Home_of_the_SEED
U14	Genome Quest	http://www.genomequest.com/
U15	GenBank	http://www.ncbi.nlm.nih.gov/genbank/
U16	EMBL	http://www.ebi.ac.uk/embl/

U17	DDBJ	http://www.ddbj.nig.ac.jp/
U18	Apollo	http://apollo.berkeleybop.org/current/index.html
U19	Manatee	http://manatee.sourceforge.net/
U20	Artemis	http://www.sanger.ac.uk/resources/software/artemis/