International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

De novo Genome Assembly of *Yersinia pestis* Strain12

Mandeep Kaur

Student of Department of Biotechnology and Bioinformatics, Guru Nanak Girls College, Model Town, Ludhiana (Punjab), India

Abstract: ***Recent advancement in next generation sequencing technology have made possible to sequence whole genome but assembling a large number of short sequence reads is still big challenge.Yersinia pestis is one of the most prominent human pathogens and the causative agent which cause plague disease. Yersinia contains three species Yersinia pestis, Yersinia pseudotuberculosis, and Yersinia enterocolitica which cause plague disease in humans. In this research work,the comparative study of two assemblers namely Velvet and SPAdes using Yersinia pestis 12 paired end data sets from illumina platform consisting of 358 million base pairs reads. Overall, the best assembly was generated using Velvet which consumed the least amount of memory than any other assembler, total contig are found 3530 and G+C content was 48.40%.The resulting best assembly leads to predication of the gene which reveals 5011 total genes, including a total of 4243 protein coding sequencing in the bacterial genome. We annotated this assembly using three different platforms RAST, BLAST2GO and PROKKA. Additionally functional analysis using KEGG pathway from BLAST2GO provides 108 metabolic pathways. Since Genome Assembly leads to development of new therapeutic targets, it is obligatory to annotate these genomes. Comparison among Yersinia species by de novo assembly makes it possible to annotate the genome for better understanding of pharmacological targets.***

Keywords: Genome assembly, Annotation, Yersinia pestis 12

1. Introduction

1.1 Genome Assembly

Genome assembly refers to the process of reconstructing the original DNA sequence(s) of an organism from the read sequences. Basically, it takes a large number of short DNA sequences and putting them back together to create a representation of the original chromosomes from which the DNA originated [1]. To assemble a whole genome sequence, short fragments are joined to form larger fragments after removing overlaps and these merged sequences are termed contigs, which are usually 5,000 to 10,000 bases long. A number of overlapping contigs can be further merged to form scaffolds, also called supercontigs (30,000–50,000 bases) and these overlapping scaffolds are then connected to create the final highest resolution map of the genome.***De novo* assembly** is the method of combining overlapping sequence reads into contiguous sequences (contigs) without using any guide or reference sequence [2]. For the de novo assembly of short reads, the most commonly used algorithms are based on **De Bruijn graphs**, although other algorithms such as **Overlap Layout Consensus (OLC)** are used [3]. In De Bruijn graphs type of assembly method, each read is broken into a sequence of overlapping k-mers from which the distinct k-mers are added as vertices to the graph, and k-mers that originate from adjacent positions in a read are linked by an edge. The assembly problem can then be created as finding a walk through the graph that visit search edge in the graph once as an Eulerian path problem [4]. Instead, the overlap between reads is implied in the structure of the graph therefore the graph can easily be constructed in two passes over the data firstly, the set of k-mers can be extracted from the reads and added as vertices in the graph, and second, adjacent k-mers can be extracted from the reads and added as edges. This process can be completed nearly as fast as the data can be read from disk with suitable choices of data structures to represent the graph [5]. The Overlap Layout Consensus (OLC) strategy breaks down the assembly into three distinct steps in order to enable a global analysis of the relationships between the reads unlike the inherently localized approach taken by the greedy approaches. The first step is the same as for the greedy approach; the reads are compared to each other for identification of pairwise overlaps between all sequencing reads. This information is then used to construct a contiguous sequence stretches (overlap graph), a graph containing each read as a node, and an edge connects two nodes if an overlap was identified between the corresponding reads [6].

1.1 *Yersinia pestis* strain 12

Yersinia pestis is the causative agent which is extremely virulent in humans, including a systemic invasive infectious disease classically referred to as plague. *Yersinia pestis* is named so because of its discovery by Swiss microbiologist Alexandre Yersin. The classification places it under proteobacteria phylum because it is gram negative facultative anaerobe (can live in absence or presence of oxygen) [7]. *Yersinia pestis* is nonmotile and it's most favorable temperature is 28-30°C. Being purple sulfur bacteria places it in the Gammaproteobacteria class, while it is considered in the Enterobacteriaceae family because it lacks oxidase and is a non-spore forming bacteria.The genome size is about 4.9 Mb, with up to 4 plasmids. Each genome contains 3,161 to 4,419 coding sequences and a GC content of 47 to 48%.Naturally occurring bacterium *Yersinia pestis* primarily found in wild rodents and usually transfer hypodermically to human by the bite of an infected flea, but also transferred by air especially during pandemics of disease [8]. The flea finds a victim and tries to feed by injecting its sucking mechanism, but the bacilli

have blocked the flea's esophagus and pharynx, preventing it from obtaining any blood. Flea continuously punctures its food source; in turn disgorge into the wound and injecting it with the plague bacilli [9].

2. Review of Literature

Yersinia pestis has been responsible for three pandemics throughout history. The first pandemic, the Justinian plague (541-767 AD), was originated in east or central Africa and spread from Egypt to the Mediterranean basin which may result in the deaths of 40 to 100 million people. Plague originated in central Asia, spread to the Caspian Sea and then throughout Europe, during second pandemic, the Black Death from 1346 to the 1800s and resulted in the deaths of one third of the population of Europe [10].During early1330s, the Black Death began with an outbreak of deadly bubonic plague in China, which is one of the busiest of the world's trading nations. Several Italian merchant ships returning from a trip to the Black Sea in 1347, docked in Sicily with numerous passengers on board already dying of plague. Within days, the disease had spread to surrounding countryside, where it continued to spread slowly from village to village [11].In 1546, the first complete theory of infection was published by Girolamo Fracastoro, according to her speculation the plague was caused by an infective agent of minute size that he called seminaria contagionis to cause spoiling and were transmitted by minute particles [12]. However, the third pandemic began in the mid-1800s in the Yunnan region of China and spread worldwide through marine shipping from Hong Kong and resulted in the deaths of 12 million people, mainly in India. During the next 2 years, 1899 and 1900, it was widely spread to Africa, Australia, Europe, Hawaii, India, Japan, the Middle East, the Philippines, North America, and South America. Plague arrived officially in the United States in March1900, an initial event was discovered by the lifeless body of a Chinese laborer in a hotel basement in San Francisco, California [13]. Another epidemic occurred after the 1906 earthquake in San Francisco, which destroyed buildings that left rats as well as humans homeless. In 1907, cases of plague were reported again, but this time, experience in dealing with the rats led to a halt of the epidemic, at least in that area. *Yersinia pestis* has been subdivided into four biovars: Orientalis, Medievalis, Antiqua, and Microtus on the basis of minor phenotypic differences [14]. In Zambia, there have been three outbreaks of human plague in North-Western region occurred in December 1993, while in Southern zone, the outbreak occurred in December 1996 and a number of cases were recorded. Another outbreak in Eastern zone occurred in January 2001, where 436 human cases of bubonic plague with 11 deaths in 2007 were reported and, 32 cases were involved. In India, plague has been known to occur a period of total silence, which experienced a large outbreak in 1994 after 30 years [15].On 23 August 2017, a 31-year-old male from Toamasina developed malaria-like symptoms while visiting the Ankazobe district in the central highlands of Madagascar. His condition worsened and he died on 27 August 2017. His body was prepared for a funeral at the nearest hospital in the Moramanga district hospital located between Antananarivo and Toamasina. He was buried in a village close to Toamasina without safety procedures. Subsequently, 31 people who had been in contact with this case fell ill and four of them died. From the 1 August to 22 November 2017, a total of 2348 confirmed, including 202 deaths, were reported by the Ministry of Health of Madagascar to World Health Organization. There were 1791 cases of pneumonic plague, of which 22% were confirmed. In addition to pneumonic cases, there were reports of 341 cases of bubonic plague, one case of septicemic plague. In the USA, considerable resources have been devoted to preparing for a plague bioweapon attack, including emergency response exercises and stockpiling of antimicrobials for treatment, post-exposure, and pre-exposure prophylaxis [16].

3. Pathogenesis

Yersinia pestis is most commonly transmitted to humans through the bites of infected fleas, resulting in either primary bubonic plague or septicemic plague. The flea draws viable *Yersinia pestis* organisms into its intestinal tract and organisms multiply in the flea and block the flea's proventriculus (an organ between the stomach and esophagus of the flea). In the midgut of it flea vector (Xenopsylla cheopis), *Yersinia pestis* survives cytotoxic digestion of blood plasma through the action of Yersinia murine toxin (Ymt). Plasmid-encoded phospholipase D (PlD) also referred to as Ymt (Yersinia murine toxin)[17].The hemin storage system locus (hms) also contributes to the pathogenicity of *Y. pestis* in fleas. The formation of biofilms in the proventriculus contributes to the transmission of the plague in fleas because of the biofilms growing in the proventriculus can extend into and block the esophagus of the flea. Consequently, when the flea attempts to feed, the blood enters the esophagus, mixes with the biofilms and returns to the host animal when the flea stops feeding this blockage of the flea's esophagus is a key step in Yersinia pestis transmission [18]. Once *Y. pestis* has entered the human host, within 2-6 days the bacterium spreads throughout the lymphatic system and enters the bloodstream. *Y. pestis* spread throughout the lymphatic system triggers a large-scale immune response with the appearance of buboes on the armpits; neck and also increase in numbers of bacteria in the bloodstream promote the odds of human-human transmission [19]. Mostly mammals can be infected by a flea, *Y. pestis* multiplies in the flea gut and expresses a coagulase that clots ingested blood, block the proventriculus, that is, unable to move food (blood) from its esophagus to its midgut. As flea repeatedly attempts to feed, and because it is unable to digest the blood, it bring up the newly infected blood back into the bloodstream of the mammal on which it is feeding, therefore transferring the microorganism from the flea to the mammal [20]. Approximately 25,000 to 100,000 *Y pestis* organisms are injected into the skin of the mammal host during this process. Historically, the antibiotics most commonly used have included streptomycin, gentamicin, tetracycline or doxycycline, and chloramphenicol. A live attenuated vaccine, EV76, also was in use in humans in some areas of the world, but it also is not commercially available at present [21].

4. Methodology

1) From NCBI WGS projects (https://www.ncbi.nlm.nih.gov/Traces/wgs/) is used for selection organism *Yersinia pestis* 12having 1,092 contigs is taken in account for its assembly and annotation.
2) EBI-ENA(https://www.ebi.ac.uk/ena)is used for retrieval of raw reads of illumina experiment(358 million base pairs)for genome assembly of *Yersinia pestis* 12 in FastQ format.
3) FastQC(https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)is used for assessing quality of raw reads providing good quality score.
4) Velvet (https://usegalaxy.org.au/)de novo assembler can be used to quickly build long continuous sequences, or contigs(total contigs 3,530 and greater than 1000bp are 870) of short-read datasets as produced by next-generation sequencing technologies.
5) SPAdes Genome Assembler (https://usegalaxy.org.au/) is another tool for de novo sequencing providing total contigs 1,150 and greater than 1000bp are 352.
6) QUAST (https://usegalaxy.org.au/)is a quality assessment tool for evaluating and comparing genome assemblies.
7) GeneMarkS(https://exon.gatech.edu/genmark/genemarks.cgi)is used for 5011gene prediction in sequenced prokaryotic genome.
8) Blast2GO (https://www.blast2go.com/)is used for functional annotation of 784 contigs from 870 assembled contigs and providing role in 108 metabolic pathways.
9) Rapid Annotation using Subsystem Technology (RAST)(http://rast.theseed.org/FIG/rast.cgi)Server provides high-quality genome annotations for prokaryotes across the whole phylogenetic tree.
10) Prokaryotic Genome Annotation (Prokka)(https://usegalaxy.org.au/)is a software tool to rapidly annotate genes and identify coding sequences in prokaryotic genomes.

5. Results and Discussion

Fast QC assessing quality of raw read

Table 1: FastQC summary of statistics

Parameters	*forward reads*	*reverse reads*
Basic Statistics	PASS	PASS
Per base sequence quality	PASS	PASS
Per tile sequence quality	FAIL	FAIL
Per sequence quality scores	PASS	PASS
Per base sequence content	PASS	PASS
Per sequence GC content	PASS	PASS
Per base N content	PASS	PASS
Sequence Length Distribution	PASS	PASS
Sequence Duplication Levels	PASS	PASS
Overrepresented sequences	PASS	PASS
Adapter Content	PASS	PASS

The basic statistics showed that both files have the same number of reads (3,581,097) which they should have because they contain paired-end reads (one file has forwards reads, the second reverse reads). No reads were flagged for bad quality and they had sequence length of 35 and 44 bp in forward and reverse strand files. Similarly, GC content was around 47 and 49 percent.Both FASTQ files had overall good per base quality of reads with declining quality towards the end (especially in the case of reverse reads).

Genome assembly (Velvet)
It works mainly with fasta and fastq formats. For paired-end reads, the assumption is that each read is next to its mate read. In other words, if the reads are indexed from 0, then reads 0 and 1 are paired, 2 and 3, 4 and 5, etc. Concerning read orientation, Velvet expects paired-end reads to come from opposite strands facing each other, as in the traditional Sanger format. If we have paired-end reads produced from circularization, it will be necessary to replace the first read in each pair by its reverse complement before running velveth. Velvet produces a number of files in its output directory, including

The *Contigs* file: Contig sequences in FASTA format. It will show each contig with the k-mer length and k-mer coverage (for k-mer size of 31).

The *Contigs stats* file: a tab-separated table with statistics on the contigs.

The velvet *Log* file and Assembly *Last Graph* file.

Genome assembly (SPAdes)
It usesk-mers for building the initial de Bruijn graph and it performs graph-theoretical operations which are based on graph structure, coverage and sequence lengths. Moreover, it adjusts errors iteratively. The assembly in SPAdes involvesassembly graph construction, k-bimer (pairs of k-mers) adjustment which means the exact distances between k-mers in the genome (edges in the assembly graph) are estimated, scaffolds and contig construction. In its output directory SPAdes produces a number of files, including:

The contigs.fasta file contains the resulting contigs in fasta format with its k-mer length and k-mer coverage (for k-mer size of 31).

The scaffolds.fasta file contains the resulting contigs in fasta format with its k-mer length and k-mer coverage.

The contigs.stats file contains results in tabular form with its k-mer length and k-mer coverage.

The scaffolds.stats file contains results in tabular form with its k-mer length and k-mer coverage.

The log file containswith the Advanced Parameters section for more information on how to create and use it.

QUAST (Quality assessment)
Although the assembly metrics such as N50 and contig numbers are widely used for the assembly evaluation, they

may not always correlate well with the actual quality of the assembly and several other bioinformatics approaches and metrics have been developed to assess assembly quality.In output directory of QUAST produces a number of files.

Table 2: Description of QUAST output files

QUAST output	*Description of file contents*
report.txt	Assessment summary in plain text format.
report.tsv	Tab-separated version of the summary, suitable for spreadsheets.
report.tex	LaTeX version of the summary.
icarus.html	Icarus main menu with links to interactive viewers.
report.html	HTML version of the report with interactive plots inside.

Comparison of Genome Assembly

To assemble paired end reads file of Yersinia pestis 12, three assemblers are used. The assemblers are Velvet, SPAdes and CLC genomics workbench and their assembly comparison is shown in table as follows:

Table 3: Comparison of Genome Assembly

Parameters	*Velvet*	*SPAdes*
# contigs	3,530	1,150
>1000(bp)	870	352
N50 (bp)	6,537	20,241
Largest contig (bp)	34,433	88,856
Total (bp)	4,223,721	4,455,822

The Velvet, SPAdes and CLC Genomics Workbench are de novo assembler (de novo assembly algorithm works by using de Bruijn graphs) was designed to build contigs and eventually scaffolds from short read sequencing data. As the total base pair coverage is approximately similar in all three assemblers. But the overall similar results are produced by the SPAdes and CLC Genomics Workbench. Velvet found the 870 greater numbers of contigs in 1000 bp length. Velvet is a reliable and easy to use DBG assembler. Velvet makes extensive use of graph simplification to reduce simple non-intersecting paths to single nodes. Simplification compresses the graph without loss of information. It does not use an error-correction pre-processor, though it does have an error-avoidance read filter. It applies a series of heuristics that reduce graph complexity. The heuristics exploit local graph topology, read coverage, sequence identity, and paired-end constraints. Velvet can already convert high-coverage very short reads into reasonably sized contigs with no additional information. The Velvet framework will provide a rich set of different algorithmic options tailored to different tasks and thus provide a platform for cheap de novo sequence assemblies. Therefore, resulted contigs from velvet algorithm are annotated using different annotation software.

GeneMark S for gene prediction

The iterative method, GeneMark S program is used for gene prediction in prokaryotic genomic DNA sequence (i.e. for Yersinia pestis 12) with a specific focus on identifying gene starts.The output of the GeneMark program consists of a list of predicted as genes, protein sequence and nucleotide sequence file of genes. In GeneMark S (GeneMark.hmm) reports all **predicted genes** in a format that includes the strand of gene resides on, its boundaries, length in nucleotides and gene class. Class indicates which of the two Markov chain models (Typical or Atypical gene model) used in GeneMark. Genes of the Typical class exhibit codon usage patterns specific to the majority of genes in the given species, while Atypical class genes may not follow such patterns and frequently contain significant numbers of laterally transferred genes.The nucleotide sequences of predicted genes and translated protein sequences are available as an output to facilitate further analysis, such as BLAST searching.As a result for 870 contigs, GeneMark S predicts 5011 genes that includes the strand of gene resides on, its boundaries, length in nucleotides and gene class.

BLAST2GO (Functional Annotation)

Blast2GO is a comprehensive bioinformatics tool used for the functional annotation of velvet contigs. It uses the Basic Local Alignment Search Tool (BLAST) to find sequences similar to query set. As the BLAST search progresses, sequences with successful BLAST results change their color on the Main Sequence Table from white to orange and the BLAST result related columns will be filled. In case no results could be retrieved for a given sequence, this row will turn dark-red. Similarly, Interpro (purple), mapping (green) and annotation results in blue color of Main Sequence Table.The **Gene Ontology** consortium has developed a vocabulary of defined terms that describe gene products in the context of three domains: biological process, molecular function and cellular component in a species-independent manner.A pie chart was generated for each category; biological process (Figure 1), molecular function (Figure 2) and cellular component (Figure 3).In summary, these representations consist of 2298, 2441 and 918 hits respectively.

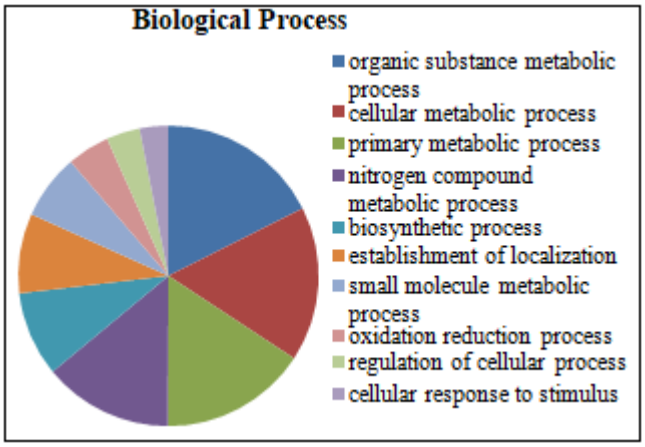

Figure 1: Biological process categories for the *Yersinia pestis 12*

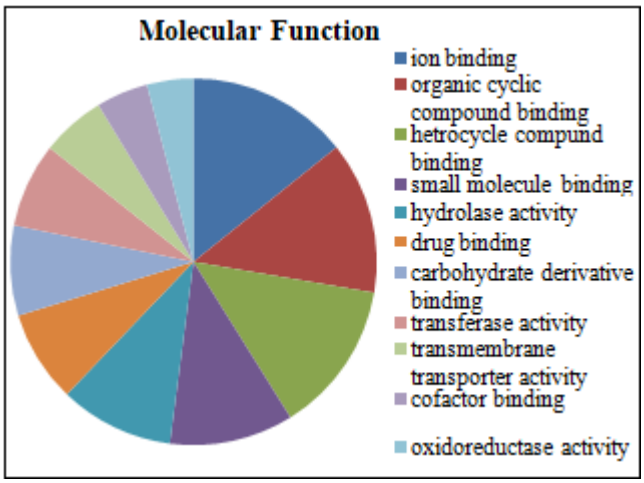

Figure 2: Molecular function categories for the *Yersinia pestis 12*

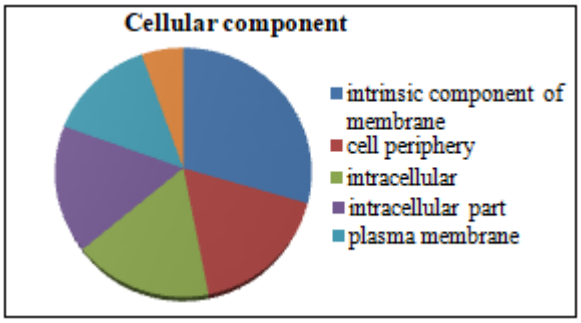

Figure 3: Cellular component categories for the *Yersinia pestis 12*

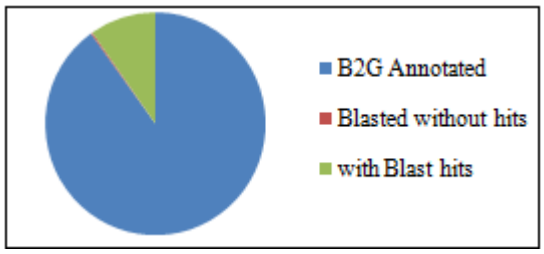

Figure 4: Data Distribution for annotation

*Functional Annotation using BLAST2GO:*Over 870 contigs annotations have been performed, 784(90%) contigs (Figure 4) were found annotated. From which 2% were found blasted without hits and 10% are blasted with hits.

*Metabolic Pathways using KEGG:*To better understand functions and interactions, all annotated contigs were mapped against the KEGG database for a pathway-based analysis. As a result, a total of 607 enzymes were assigned to a KEGG pathway. This relatively low number of enzymes assigned to a pathway is likely the result of imperfect annotation caused by Blast2GO, although enzymes were present in 108 different KEGG pathways..KEGG pathways associated with metabolism had the highest representation, with a large number of the sequences associated with 'carbohydrate metabolism', 'energy metabolism', 'lipid metabolism' and 'amino acid metabolism' pathways.

Rapid Annotation using Subsystem Technology (RAST)

The **RAST** server annotation and summary of run in terms of quality check in given for RASTtk and Classic RAST.

Table 4: Summary table for RAST quality check run

Parameters	*RASTtk*	*Classic RAST*
Number of features	4258	3753
Number of warnings	0	2
Number of fatal problems	0	0
Possibly missing genes	-	124

RAST Annotation showing some features of organism (*Yersinia pestis 12*).

Table 5: Organism Overview for *Yersinia pestis 12* (755859.8)

Features	*RASTtk*	*Classic RAST*
L50	186	186
Number of contigs with protein encoding genes (PEGs)	870	870
Number of subsystems	377	511
Number of coding sequences	4243	3736
Number of RNAs	15	17

Once assignments of function have been made, an initial metabolic reconstruction is formed. Subsystem coverage in Figure 5 shows that 62% of the coding sequences are not in subsystems and 38% of sequences are in subsystems. The pie chart shows the distribution of subsystem categories in the genome. The most prominent categories are Carbohydrates (287 sequences), Amino Acids and Derivatives (364 sequences), and Cofactors, Vitamins, Prosthetic Groups, Pigments (170 sequences). As a bacterium that must survive in Arctic, unsurprisingly significant portion of coding sequences are in the category Stress Response (71 sequences) and 1 sequences were also identified in the Dormancy and Sporulation category.

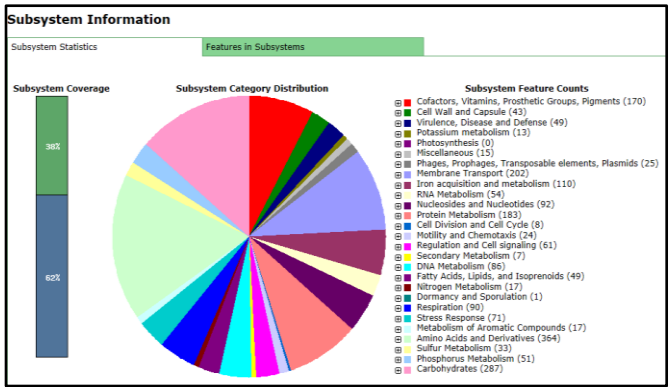

Figure 5: Genes connected to subsystems and their distribution in different category (RASTtk)

Subsystem coverage in Figure 6 shows that 41% of the coding sequences are not in subsystems and 59% of sequences are in subsystems. The most prominent categories are Carbohydrates (450 sequences), Amino Acids and Derivatives (403 sequences), and Cofactors, Vitamins, Prosthetic Groups, Pigments (232 sequences). As a bacterium that must survive in

Arctic, unsurprisingly significant portion of coding sequences are in the category Stress Response (129 sequences) and 1 sequences were also identified in the Dormancy and Sporulation category.

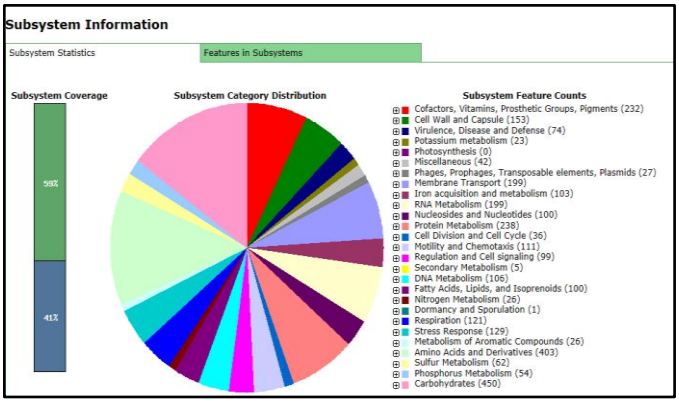

Figure 6: Genes connected to subsystems and their distribution in different category (Classic RAST)

The whole Genome Browser allows the users to zoom from a graphic whole genome presentation into any desired area of the genome down to a gene. All annotated feature can be viewed and downloaded from the View Features Page. For each peg the location on the contig, the functional role assignment, its EC number and GO category, the connection to a subsystem.

PROKKA (Prokaryotic annotation)

Prokka expects preassembled genomic DNA sequences in FASTA format. Finished sequences without gaps are the ideal input, but it is expected that the typical input will be a set of scaffold sequences produced by *de novo* assembly software. This sequence file is the only mandatory parameter to the software. Prokka relies on external feature prediction tools to identify the coordinates of genomic features within contigs. Prokka produces number of files in the specified output directory, all with a common prefix. These are described in Table 6.

Table 6: Description of Prokka output files

Suffix	*Description of file contents*
.fna	FASTA file of original input contigs (nucleotide)
.faa	FASTA file of translated coding genes (protein)
.ffn	FASTA file of all genomic features (nucleotide)
.fsa	Contig sequences for submission (nucleotide)
.tbl	Feature table for submission.
.gbk	GenBank file containing sequences and annotations
.gff	GFF v3 file containing sequences and annotations
.log	Log file of Prokka processing output
.txt	Annotation summary statistics

Comparison of Genome Annotation

To assess accuracy, the annotations are compared for Blast2GO, RAST and Prokka for the highly curated *Yersinia pestis* 12 genome.

Table 7: Comparison of annotation of *Yersinia pestis* 12

Feature	*Blast2GO*	*RASTtk*	*Classic RAST*	*Prokka*
Total CDS	-	4244	3737	3512
Hypothetical protein	14	607	482	991
With EC number	420	1040	795	1156
Total tRNA	-	15	17	14
Total tmRNA	-	-	-	1
Total rRNA	-	-	-	3
KEGG Pathways	108	168	170	-

Prokka comprehensive toolbox used for annotation of prokaryotic genomes (i.e. *Yersinia pestis* 12) takes contigs as input and finds and annotates: 3512 CDSs, 3 rRNAs, 14 tRNA. Rast fully automated annotation service for archaeal and bacterial genomes. It identifies protein-encoding, rRNA, tRNA genes and assigns functions to genes. It seeks to rapidly produce high-quality assessment of gene functions and an initial metabolic reconstruction.Annotation comparison found that the Blast2GO is an all in one bioinformatics solution for functional annotation of novel sequences and the analysis of annotation data. Its main function is to assign information about the biological function of gene or protein sequences by making use of diverse public resources like comparison algorithms and databases. The software identifies already characterized similar sequences and transfers its functional labels to the uncharacterized sequence in Yersinia pestis and 784 contigs are found annotated out of 870 assembled contigs.

6. Conclusion

De novo assemblies of *Yersinia pestis* strain 12 by two different de novo DNA assemblers: Velvet and SPAdes based on illumina paired end reads. Applications were compared on two sets of bacterial DNA reads obtained from the National Center for Biotechnology Information (WGS). The benchmark dataset contains DNA reads from SRR69094 for *Yersinia pestis 12*. The results were compared in terms of the number of contigs longer than 1000 bp, the length of N50 contig, the length of the longest contig. Perhaps, the best resulted 870 contigs found longer than 1000 bp from Velvet assembler involves the quality assessment which can be been done by using Quast tool.Open-access quality assessment tool QUAST will help to choose the best assembler for research, and it will help to improve their software used for genome assembly. The resulting 870 contigs leads to predication of the gene by GeneMark S which includes 5011 genes in the bacterial genome. Functional Annotation of the assembled genome is performed by comparing three different tools. Firstly, the Rast server provides initial annotations and also identifies protein-encoding, rRNA and tRNA genes, assigns functions to the genes and predicts which subsystems are represented in the genome.Another tool Prokka produced an overall better annotation than RAST and to identify the coordinates of genomic features within contigs. Blast2GO combines high-throughput analysis, statistical evaluation and biology framed visualization with a high degree of user interaction. Annotation of 784 contigs which had good depth coverage and large length provided biological process (cellular metabolic

process, primary metabolic process, biosynthetic process) molecular function (ion binding hydrolase activity, drug binding, transferase activity, trans membrane transferase activity) and cellular component (cell periphery, intracellular, plasma membrane, intracellular part). The data will provide a valuable resource for future studies of the bacteria *Yersinia pestis* strain *12* genomefor providing a target drug to diagnose the plague disease. In this project, the part of genome analysis that is customarily performed before a genome sequence is deposited in GenBank, is done on *Yersinia pestis* 12.

References

[1] J. Miller, S. Koren, G. Sutton. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, 95:315-327.
[2] Chaisson MJ, Pevzner PA. (2008). Short read fragment assembly of bacterial genomes. *Genome Research*, 18: 324–330.
[3] Shendure J, Ji H.(2008). Next-generation DNA sequencing. *Nature Biotechnology*, 26: 1135–1145.
[4] Pevzner PA, Tang H. (2001). Fragment assembly with double-barreled data. *Bioinformatics*, 17.S225:33.
[5] Myers EW, Sutton GG, Delcher AL, *et al.* (2000). A whole-genome assembly of Drosophila. *Science*, 287:2196–204.
[6] Mihai Pop. (2009). Genome assembly reborn: recent computational challenges. *Briefings in bioinformatics*, 10(4):354-366.
[7] Vladimir V. Kutyrev, Galina A. Eroshenko, Vladimir L. Motin, *et.al.* (2018). Phylogeny and Classification of Yersinia pestis Through the Lens of Strains from the Plague Foci of Commonwealth of Independent States. 9:1-11
[8] Achtman, M. *et al.* (1999). Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis. *Proc. Natl Acad. Sci. USA.*, 96:14043-14048.
[9] Gross L. (1995). How the plague bacillus and its transmission through fleas were discovered: reminiscences from my years at the Pasteur Institute in Paris. *ProcNatlAcadSci USA*., 92:7609-7611.
[10] Morelli, G., Song, Y., Mazzoni, C. J., Eppinger, M., Roumagnac, P., Wagner, D. M., Feldkamp, M., Kusecek, B., Vogler, A. J., *et.al.* (2010). Yersinia pestis genome sequencing identifies patterns of global phylogenetic diversity. *Nat Genet 42.*, 1140–1143.
[11] Mee C. (1990). How a mysterious disease laid Europe's masses. *Smithsonian*, 20:66-69.
[12] Riedel S. (2005). Plague: from natural disease to bioterrorism. *ProcBaylUniv Med Cent., 18:116-124.*
[13] Caten JL, Kartman L. (1968). Human plague in the united states,1900-1966. *JAMA.*,205:333-336.
[14] S. S. Nyirenda, B. M. Hang'ombe, E. Mulenga and B. S. Kilonzo. (2017). Serological and PCR investigation of Yersinia pestis in potential reservoir hosts from a plague outbreak focus in Zambia. *BMC Res Notes.*, 10:345-6.
[15] Inglesby TV, Grossman R, O'Toole T. (2001). A plague on your city: observations from TOPOFF. *Clinical Infectious Diseases*, 32: 436–445
[16] Hinnebusch BJ, Rudolph AE, Cherepanov P, *et al.* (2002). Role of Yersinia murine toxin in survival of Yersinia pestis in the midgut of the flea vector. *Science*, 296:733-735
[17] Rebecca J. Eisen, David T. Dennis, and Kenneth L. Gage. (2015). The Role of Early-Phase Transmission in the Spread of Yersinia pestis. *J. Med. Entomol.*, 52(6):1183–1192.
[18] Bacot, A. W., and C. J. Martin. (1914). Observations on the mechanism of the transmission of plague by fleas. *J. Hyg*., 13: 423–439.
[19] Hinnebusch BJ. (1997). Bubonic plague: a molecular genetic case history of the emergence of an infectious disease. *J Mol Med*., 75:645-652.
[20] Cornelis GR. (2000) Molecular and cell biology aspects of plague. *ProcNatlAcadSci USA*., 97:8778-8783.
[21] Williamson ED. (2001). Plague vaccine and development. *J ApplMicrobiol*., 91:606-608.

Author Profile

Mandeep Kaur received the B.Sc. and M.Sc. in Biotechnology degrees from Guru Nanak Girls College, Model Town Ludhiana in 2017 and 2019respectively.