International Journal of Pharmacy Research and Technology
2014, Volume 4, Issue 2, 18-22
ISSN 2250 – 0944 (Online)
ISSN 2250 – 1150 (Print)

Research Article

Comparative analysis of de novo assembly of *Bacillus cereus HUB4-4* short reads

Nagaraja .K[1*], Seema J Patel[1], ManjunathDammalli[2]
[1]Dept. of Biotechnology, GMIT, Davangere, Karnataka-577006
[2] Dept. of Biotechnology, SIT, Tumkur, Karnataka-572103
***Corresponding Author:** Email: nagaraj036@gmail.com, Mobile: 8792795979

Received: 28/06/2013, Revised: 17/12/2013 Accepted: 15/01/2014

ABSTRACT

The automated Sanger method is considered as a first generation technology and newer methods are referred to as next generation sequencing. These newer technologies constitute various strategies that rely on a combination of template preparation sequencing and imaging and genome alignment and assembly methods. NGS or high throughput sequencing technologies are intended to lower the cost of DNA sequencing beyond what is possible with standard dye terminator methods. *Bacillus cereus* is a very common dirt-dwelling bacterium which can survive in an oxygen environment with the help of protective endospores. There are several strains like HUB4-4,HUB2-3, CER057etc are cause harm to humans, whereas others are used as helpful probiotics. De novo assembly is mainly used for the purpose of assembling short reads to create full-length sequences. *Bacillus cereus HUB4-4* for this organism there is no contigs and scaffolds are available. So De novo assembly was carried out to find scaffolds and contigs by using the CLC Genomics Workbench and DRA pipeline and N25, N50, N75 contigs were determined. For identified N50 contigs, BLAST is done at NCBI and identified related organism to*Bacillus cereus HUB4-4* in CLC genomics. N50 contigs in DRA pipeline were also determined by using three software tools like soap de novo, velvet, abyss for the *Bacillus cereus HUB4-4* data using short read assembly for paired end data and compare the output of these two tools and identified CLC genomics workbench is more efficient.

Key words: illumina genome analyzer, denovo assembly, *Bacillus cereus HUB4-4,* CLC genomics workbench, DRA pipeline.

INTRODUCTION

Next generation sequencing discover a relationship between un common variation of a gene or diffuse interactions among many genes and a future disease process that makes a difference to the practice of medicine. The automated Sanger method is considered as a first generation technology and newer methods are referred to as next generation sequencing. These newer technologies constitute various strategies that rely on a combination of template preparation sequencing and imaging and genome alignment and assembly methods. The main application of NGS may be the re sequencing of human genomes to enhance our understanding of how genetic differences affect health and disease and other applications of NGS are innovative, modernization, digital marketing, cloud optimize, NGS or high throughput sequencing technologies are intended to lower the cost of DNA sequencing beyond what is possible with standard dye terminator methods. In ultrahigh through put sequencing as many as 500000 by synthesis operations may be run in parallel and also massively parallel signature sequencing typically used for sequencing CDNA for measurements of gene expression levels .NGS enable novel applications such as analysis of ancient DNA samples and also substantially widened the scope of Meta genomic analysis of environmentally derived samples and also helps to characterization of ecological diversity .NGS have certain platforms for helping to open entirely new areas of biological inquiry including the ancient genomes investigation[1].

Bacillus cereus is a very common dirt-dwelling bacterium which can survive in an oxygen environment with the help of protective endospores. There are several strains like HUB4-4 ,HUB2-3,CER057,BAG6X1-1 etc are cause harm to humans, where has others are used as helpful probiotics and also *Bacillus cereus* is an endemic, soil-dwelling, Gram-positive, rod-shaped ,beta hemolytic bacterium and also its responsible [2] for a minority of food borne illnesses (2–5%), causing severe nausea, vomiting and diarrhea. *Bacillus cereus* has been recognized as a food poisoning agent since 1955. There are two forms of food poisoning that occur, one is rapid onset (emetic) and the other is late onset (diarrheal). The rapid onset is characterized by nausea and vomiting while the late onset is characterized by diarrhea and abdominal pain [3]. Denovo assembly of Bacillus cereus HUB4-4 was carried out. In this work which can further helps in SNP detection.

MATERIALS AND METHODS

Raw sequence was imported from ENA data base. De novo assembly for *Bacillus cereus HUB4-4* was done by using two main tools.

1. CLC genomics work bench
2. DDBJ read annotation pipelines

CLC GENOMICS WORK BENCH

CLC genomics workbench was mainly used for de novo assembly for the *Bacillus cereus HUB4-4* data for the identification of scaffolds and contigs with an efficient procedure.

Procedure:

CLC genomic workbench is opened and then the raw sequence data is opened using the “NGS import “option in the NGS tool box.New folder is created in the navigation area to save all the results and outputs of the *Bacillus cereus HUB4-4* data.In the next step trim sequence option is selected, which is in the NGS tools box, to trim the raw sequence. For trimming process certain parameters are setand trimmed data is obtained. For trimmed data denovo assembly was carried out.After de novo assembly was completed, four new tab pages are showed which are saved. Those four pages are

- De novo assembly log
- Trimmed de novo assembly result
- Trimmed summary report
- Trimmed un mapped reads

Trimmed de novo assembly is the result of de novo assembly. In the trimmed de-novo assembly consensus length row clicked and one of the trimmed contig map is selected from the list which length between 30000 to50000. Then next double clicked on the selected contig map which appeared in a separated window. After wards clicked on extract contig where save in result handling is selected and clicked next. Thensaved it in the folder byclicking finish.

After all the above steps finished, next step is to BLAST the selected contig. BLAST option present in the NGS tools box is selected and then the BLAST at NCBI is chosen. In the BLAST result, it contains the most related organism sequences to *Bacillus cereus HUB4-4* sequences. It is displayed in a graphical manner and one option is given below the BLAST graph named show BLAST table using this option it is known that the name of organism which is most similar to *Bacillus cereus HUB4-4* genome and description of that similar genome are also presented [4].

DDBJ READ ANNOTATION PIPELINE

DDBJ read annotation pipeline is a cloud computing based analytical platform for next generation sequencing data and it is used for running the de novo assembly using different algorithms like velvet, abyss, soap de novo for the unidentified organism *Bacillus cereus HUB4-4* for getting number of contigs and N50 contigs and result of this tool, will be compared with CLC genomics work bench.

Procedure:

For running data, first we have to login DRA pipeline. For that they are given two options, we can run the data by login as guest or else we can create new account and run the data. First uploaded the data through FTP upload or else the data can be uploaded through HTTP upload, private DRA entry, importpublic DRA once uploading is finished. From Menu Column in the left, pressed "de novo assembly" button in step.1 section. Then selected "FTP upload" tab to show uploaded file list. Uploaded files (query files) should be in fastq format. The files to be used are there or notis checked, if those exist pressed "NEXT" button. In the next page selected tool for the de novo assembly (basic analysis) like soap de novo, velvet, abyss and clicked next. Next step is generating query sets from query read files in this selected *Bacillus cereus HUB4-4* query and clicked on set as pair end then clicked next.

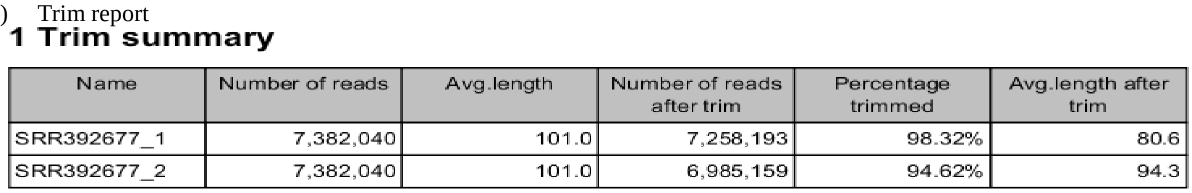
Next the reservation page which contained STATUS and NEXT JOB options. STATUS for knowing the status of the specified tool and next job is for running another job. Detailed view page which contained contig, total contig size, maximum contig size, minimum contig size, N50 contig size (which is also called as result page) is obtained [5].

RESULTS AND DISCUSSION

CLC genomics

CLC genomics workbench is opened, sequences are imported through illumine high throughput sequence option. Once importing is finished. Next step is to trim the sequences. After trimming process, trim result is as follows:

a) Trim report



1 Trim summary

Name	Number of reads	Avg.length	Number of reads after trim	Percentage trimmed	Avg.length after trim
SRR392677_1	7,382,040	101.0	7,258,193	98.32%	80.6
SRR392677_2	7,382,040	101.0	6,985,159	94.62%	94.3

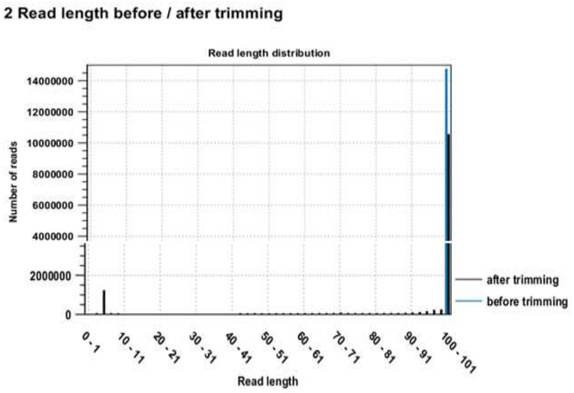


Figure 2 Trim report

1) Trim summary

Name: The name of the sequence list used as input. In this project we use sequences SRR3926677_1 (forward) and SRR392677_2 (backward) of "*Bacillus cereus HUB4-4*".

Number of reads: it retains number of reads in the input file is 7,382,040 for both forward and backward sequence.

Average length: average length of the reads in the input file is 101.0.

Numbers of reads after trim: number of reads retained after trimming process are 7,258,193 for forward and 6,985,159 for backward.

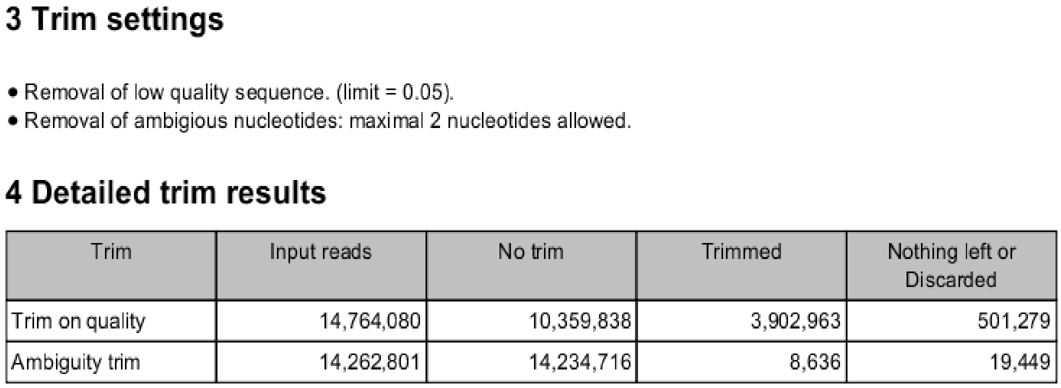
Percentage trimmed: percentage of the input reads trimmed are retained 98.32% for forward and 96.64% for reverse.

Avg. length after trim: the average length of the retained sequences after trim are 80.6 for forward and 94.3 for reverse.

1) **Read length before / after trimming:** This graph shows the number of reads of various lengths. The numbers before and after are overlaid so that it can be easily seen how the trimming has affected the read lengths.
2) **Trim settings:** A summary of the settings used for trimming process. Here low quality 14,764,080,

14,262,801sequence (limit = 0.05) and ambiguous nucleotides (maximal 2 nucleotides allowed) are removed.

3) Detailed trim results: A table with one row for Quality trimming based on quality scores and Ambiguity trimming to trim off.



3 Trim settings

- Removal of low quality sequence. (limit = 0.05).
- Removal of ambigious nucleotides: maximal 2 nucleotides allowed.

4 Detailed trim results

Trim	Input reads	No trim	Trimmed	Nothing left or Discarded
Trim on quality	14,764,080	10,359,838	3,902,963	501,279
Ambiguity trim	14,262,801	14,234,716	8,636	19,449

Input reads: The numbers of reads used as input are 14,764,080 for quality and 14,262,801 ambiguity trim. Since the trimming is done sequentially, the number of retained reads from the first type of trim is also the number of input reads for the next type of trimming.

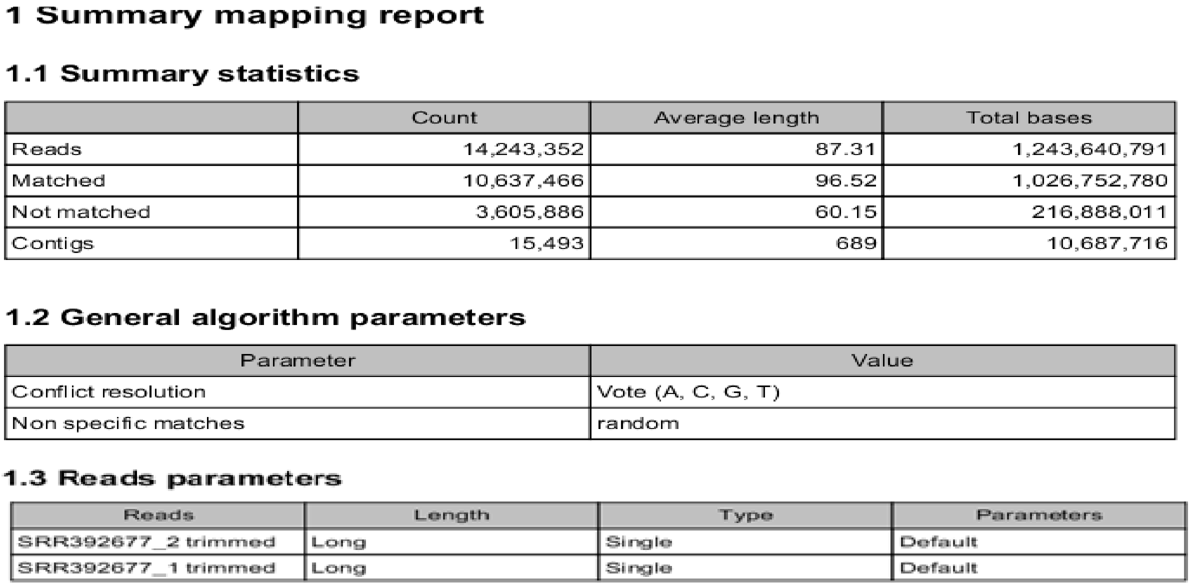
No trim: The number of reads that have been retained, unaffected by the trimming process are 10,359,838 for quality and 14,234,716 ambiguity trim.

Trimmed: The numbers of reads that have been partly trimmed are 3,902,963 for quality and 8,636 for ambiguity trimming. This number plus the number from No trim is the total number of retained reads.

Nothing left or discarded: The number of reads that have been discarded either because the full read was trimmed off or because they did not pass the length trim (e.g. too short) or adapter trim (if discard when not found was chosen for the adapter trimming).Thus we got 5,012,79 for quality trim and 19,449 for ambiguity trim.The trimmed sequence is obtained by removing a specified number of bases at either 3' or 5' end of the reads and reads shorter or longer than a specified threshold. Trimming helps in minimizing the errors in data.

De novo assembly

De novo assembly is performed for trimmed sequence which results in summary mapping report. That contains the following information when both scaffolding and read mapping is performed:



1 Summary mapping report

1.1 Summary statistics

	Count	Average length	Total bases
Reads	14,243,352	87.31	1,243,640,791
Matched	10,637,466	96.52	1,026,752,780
Not matched	3,605,886	60.15	216,888,011
Contigs	15,493	689	10,687,716

1.2 General algorithm parameters

Parameter	Value
Conflict resolution	Vote (A, C, G, T)
Non specific matches	random

1.3 Reads parameters

Reads	Length	Type	Parameters
SRR392677_2 trimmed	Long	Single	Default
SRR392677_1 trimmed	Long	Single	Default

Figure 4 Summary mapping report

1) Summary statistics: A summary of the mapping statistics

➢ Reads: The number of reads after trimming is 14,243,352, average length is87.31 and total bases are 1,243,640,791.

➢ Matched: The number of reads that are mapped is 10,637,466, their average length is 96.52 and total bases are 1,026,752,780.

➢ Not matched: The number of reads that do not match is 3,605,886, their average length60.15 and total bases are 216,888,011.

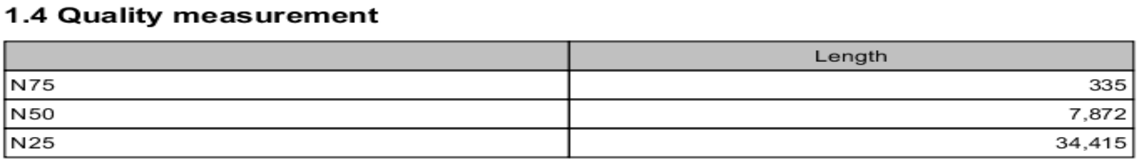
➢ Contigs: The number of overlapping reads is 15,493, their average length 689 and total bases are 10,687,716.

2) General algorithm parameters:

➢ Conflict resolution: If there is a conflict between reads than Vote (A, C, G, T) value is used that solve the conflict by counting instances of each nucleotide and then letting the majority decide the nucleotide in the consensus. In case of equality, ACGT are given priority over one another in the stated order.

- Non-specific matches: It refers to a situation where a read aligns at more than one position. This uses Random value that place the read in one of the positions randomly.

3) Read parameters: This includes assembly and mapping of both short reads, long reads, single reads and paired reads in one go. This makes it easier to combine the information from different sources. Thus we obtain long (length), single (type) and default (parameters) for SRR392677_1 (forward) and SRR392677_2 (reverse) sequences of "*Bacillus cereus HUB4-4*".



1.4 Quality measurement

	Length
N75	335
N50	7,872
N25	34,415

Figure 5 Quality measurement

4) Quality measurement: This includes

- N25 contigs: The N25 contig set is calculated by summarizing the lengths of the biggest contigs until 25 % of the total contig length is reached. The minimum contig length in this set is the number that is usually used to report the N25 value of a de novo assembly. Thus we obtained 34,415 N25 contig.
- N50 contigs: This measure is similar to N25 just with 50 % instead of 25 %. We obtained 7,872 N50 contig. It is a more informative way of measuring the lengths of contigs.
- N75 contigs: Similar to the ones above, just with 75 %. We obtained 335.

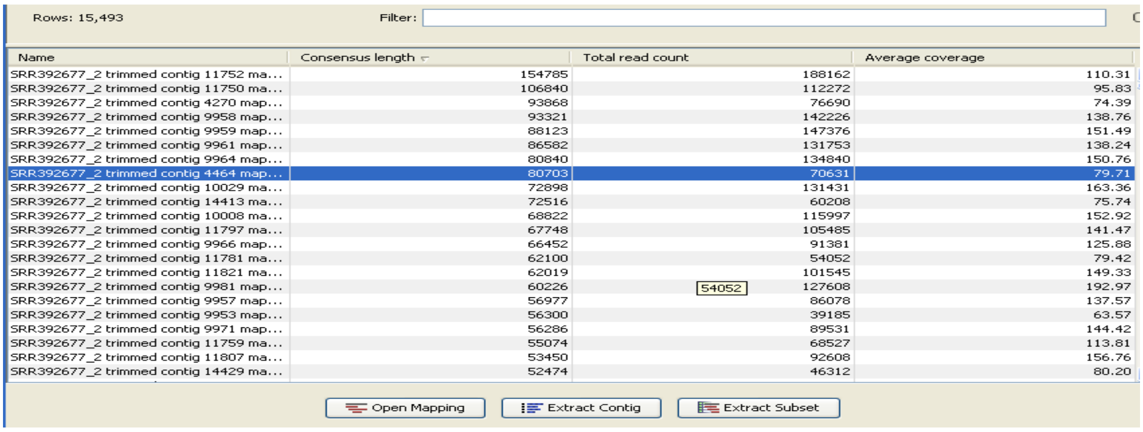
5) Distribution of read length: For each sequence length, the number of reads and the distribution in percent are seen. This is mainly useful if there is no too much variance in the lengths for e.g. Sanger sequencing data.

6) Distribution of matched reads lengths: Equivalent to the above, except that this includes only the reads that have been matched to a contig.

7) Distribution of non-matched reads lengths: Show the distribution of lengths of the rest of the sequences.

De novo assembly helps in creating simple contig sequences by using all the information that is in the read sequences.

This does not contain any information about which reads the contigs are built from. These contigs are extracted and linked together to form scaffolds. After identifying scaffolds and contigs, it has been used in resequencing. In this project importance is given to N50 contig having 201 lengths because it is the most well-known measure of De novo assembly quality. N50 is a statistical measure of average length of a set of sequences. The N50 length is defined as the length *N* for which 50% of all bases in the sequences are in a sequence of length $L < N$.



Rows: 15,493

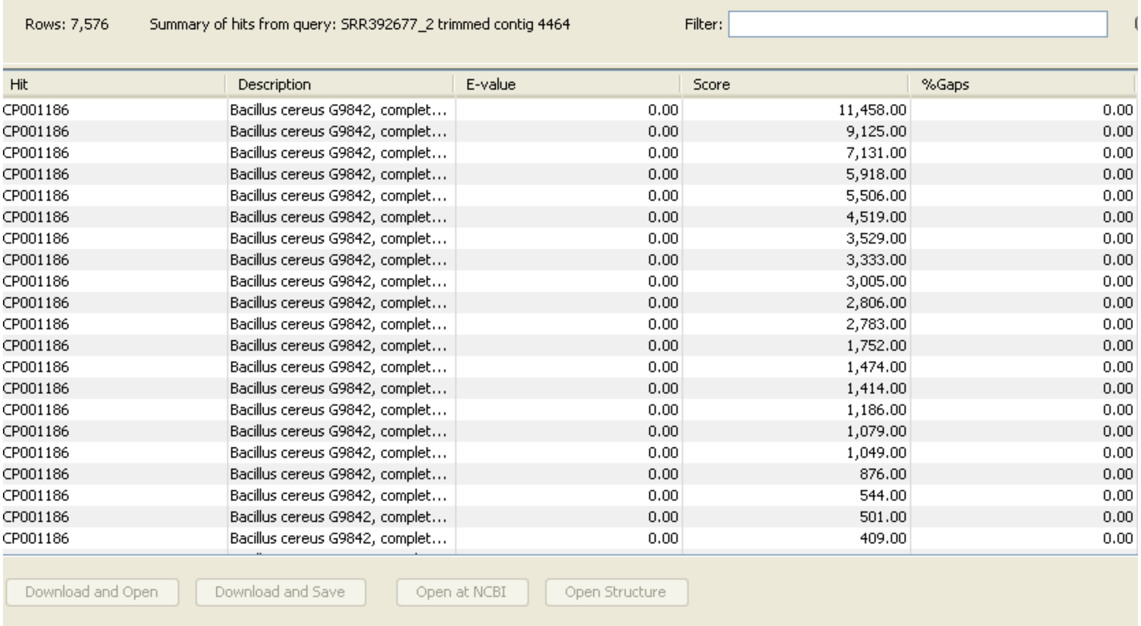
Name	Consensus length	Total read count	Average coverage
SRR392677_2 trimmed contig 11752 ma...	154785	188162	110.31
SRR392677_2 trimmed contig 11750 ma...	106840	112272	95.83
SRR392677_2 trimmed contig 4270 map...	93868	76690	74.39
SRR392677_2 trimmed contig 9958 map...	93321	142226	138.76
SRR392677_2 trimmed contig 9959 map...	88123	147376	151.49
SRR392677_2 trimmed contig 9961 map...	86582	131753	138.24
SRR392677_2 trimmed contig 9964 map...	80840	134840	150.76
SRR392677_2 trimmed contig 4464 map...	80703	70631	79.71
SRR392677_2 trimmed contig 10029 ma...	72898	131431	163.36
SRR392677_2 trimmed contig 14413 ma...	72516	60208	75.74
SRR392677_2 trimmed contig 10008 ma...	68822	115997	152.92
SRR392677_2 trimmed contig 11797 ma...	67748	105485	141.47
SRR392677_2 trimmed contig 9966 map...	66452	91381	125.88
SRR392677_2 trimmed contig 11781 ma...	62100	54052	79.42
SRR392677_2 trimmed contig 11821 ma...	62019	101545	149.33
SRR392677_2 trimmed contig 9981 map...	60226	127608	192.97
SRR392677_2 trimmed contig 9957 map...	56977	86078	137.57
SRR392677_2 trimmed contig 9953 map...	56300	39185	63.57
SRR392677_2 trimmed contig 9971 map...	56286	89531	144.42
SRR392677_2 trimmed contig 11759 ma...	55074	68527	113.81
SRR392677_2 trimmed contig 11807 ma...	53450	92608	156.76
SRR392677_2 trimmed contig 14429 ma...	52474	46312	80.20

Figure 6 N50 contigs

BLAST at NCBI

For N50 contig BLAST at NCBI is performed. Result obtained is as follows:

After extracting the N50 contig, BLAST at NCBI is performed which results in the related sequences which is most similar to our genome. In this project after BLAST we obtained "*Bacillus cereus G9842*" which is most similar to our genome "*Bacillus cereus HUB4-4*". Here E-value plays a significant role. The E-value or Expect value (E) is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size. It decreases exponentially as the Score (S) of the match increases. Essentially, the E value describes the random background noise. The lower the E-value, or the closer it is to zero, the more "significant" the match is. The calculation of the E value takes into account the length of the query sequence [9]. For "*Bacillus cereus G9842*" lowest E-value obtained is 0.00. This shows "*Bacillus cereus G9842*" more significant match with"*Bacillus cereus HUB4-4*".Similarly multiple contigs (6contigs) are selected after de novo assembly and those contigs are extracted which linked together to form scaffolds. Then BLAST at NCBI is performed which results in "*Bacillus* *cereus G9842”* which is most similar to our genome *“Bacillus cereus HUB4-4”*.



Rows: 7,576 Summary of hits from query: SRR392677_2 trimmed contig 4464 Filter:

Hit	Description	E-value	Score	%Gaps
CP001186	Bacillus cereus G9842, complet...	0.00	11,458.00	0.00
CP001186	Bacillus cereus G9842, complet...	0.00	9,125.00	0.00
CP001186	Bacillus cereus G9842, complet...	0.00	7,131.00	0.00
CP001186	Bacillus cereus G9842, complet...	0.00	5,918.00	0.00
CP001186	Bacillus cereus G9842, complet...	0.00	5,506.00	0.00
CP001186	Bacillus cereus G9842, complet...	0.00	4,519.00	0.00
CP001186	Bacillus cereus G9842, complet...	0.00	3,529.00	0.00
CP001186	Bacillus cereus G9842, complet...	0.00	3,333.00	0.00
CP001186	Bacillus cereus G9842, complet...	0.00	3,005.00	0.00
CP001186	Bacillus cereus G9842, complet...	0.00	2,806.00	0.00
CP001186	Bacillus cereus G9842, complet...	0.00	2,783.00	0.00
CP001186	Bacillus cereus G9842, complet...	0.00	1,752.00	0.00
CP001186	Bacillus cereus G9842, complet...	0.00	1,474.00	0.00
CP001186	Bacillus cereus G9842, complet...	0.00	1,414.00	0.00
CP001186	Bacillus cereus G9842, complet...	0.00	1,186.00	0.00
CP001186	Bacillus cereus G9842, complet...	0.00	1,079.00	0.00
CP001186	Bacillus cereus G9842, complet...	0.00	1,049.00	0.00
CP001186	Bacillus cereus G9842, complet...	0.00	876.00	0.00
CP001186	Bacillus cereus G9842, complet...	0.00	544.00	0.00
CP001186	Bacillus cereus G9842, complet...	0.00	501.00	0.00
CP001186	Bacillus cereus G9842, complet...	0.00	409.00	0.00

Figure 7 Blast result

Table 1 Comparison of two tools result

	CLC genomics workbench	DRA pipeline		
		Soap de novo	Abyss	Velvet
Contig	15,493	1,11,463	5,311	4,01,078
Total contig size	10,687,716	11,436,626	6,602,095	28,54,840
Max contig size	154,785	4,543	1,06,348	1,203
Min contig size	72	24	31	45
N50 contig size	201	164	19,727	73
N75contig size	4780	------	--------	---------
N25 contig size	46	------	------	-------
Mean contig length	689,80	--------	------	------
Standard deviation	3,867,80	------	-------	---------
%GC	39.76	------	------	-------

Above table indicates the comparison of two tools output.In those two tools CLC genomics is more desirable tool compared to DRA pipeline. Because ABySS tool take more time to run data, takes more time to upload data to DRA pipeline, errors are more while uploading data, less user friendly compared to CLC genomics. CLC genomics is platform independent. It can run on windows, Max OS X, Linux and in CLC genomics work bench N75, N25 contigs are also obtained. These N75,N25 contigs will be useful when N50 contigs are too large. But in DRA pipeline N75, N25 contigs are not obtained.The difference in results between ABySS and the CLC genomics workbench isminimal, but when it comes to hardware requirements, speed and memory consumption, the CLC genomics workbench performs better than ABySS. In CLC genomics workbench, related organism has been identified but in DRA pipeline it has not been identified. Mean contig length, standard deviation and the percentage of GC contents are identified in CLC genomics workbench, but not in DRA pipeline results. In CLC genomics workbench, scaffolds are visualized within the tool, but in DRA pipeline visualization software is required for scaffolds visualization. Because of above features we can come to conclusion that CLC genomics workbench is more efficient.

References

1. Micheal L and metzker. “*Sequencing technologies- the next Generation*“. 11, pp 31-35 (2010).
2. Ryan KJ; Ray CG (editors). “*Sherris Medical Microbiology*”,McGraw Hill,4th Edition, 2004.
3. Kotiranta A,Lounatmaa K, Haapasalo M. "*Epidemiology and pathogenesis ofBacillus cereusinfections*". 2(2), pp 189–98 (2000).
4. www.clcbio.com
5. www.p.ddbj.nig.ac.jp