THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

A Multi-Platform de novo Genome Assembly and Comparative Analysis for the Vibrio cholerae -HC-63A1

Vivek Chandramohan
Department of Biotechnology, SIT Engineering College, Tumkur, India
Toufik S. Tatgar
M. Tech, Department of Biotechnology
KLE Dr. M S Sheshgiri College of Engineering and Technology, Belgaum,India

Abstract:
De novo assembly is the process of combining short and repeated reads to large contigs and scaffolds. A contig is a set of overlapping DNA sequence that together represent consensus region of DNA. Scaffolds consist of overlapping contigs separated by gaps of unknown length. Vibrio cholerae was isolated using Next Generation Sequencing Techniques. Insilico genome assembly had been done. NGS is a disruptive technology that has found widespread acceptance in the life sciences research community. The high throughput and low cost of sequencing has encouraged researchers to undertake ambitious genomic projects, especially in de novo genome sequencing. The sequence data of the new strain HC-63A1 was derived from public database, analyzed and assembled using CLC, DNAnexus, DDBJpipeline. This has led to the generation of several genome sequences based exclusively on short sequence Illumina sequence reads, recently culminating in the assembly of 3.13GB genome of Vibrio cholerae HC-63A1 from Illumina sequence reads with an average length of just 99 nucleotides . Comparative analyses involving results of CLC, Velvet, Abyss and Soap denovo tools from di_erent platforms. Among these four tools CLC genome workbench is better tool for denovo assembly.

Keywords*: Contig, scaffold, Insilico*

1. Introduction

1.1. Introduction Vibrio Cholerae
Vibrio cholerae is a comma-shaped Gram-negative bacterium. Some strains of V.cholerae cause the disease cholera. V. Cholerae is a facultative anaerobe and has a fagellum at one cell pole. V. cholerae was first isolated as the cause of cholera by Italian anatomist Filippo Pacini in 1854, but his discovery was not widely known until Robert Koch, working independently for 30 years. Later publicized the knowledge and the means of _ghting the disease.

1.2. Introduction Next-Generation Sequencing
Next-generation sequencing (NGS), also known as high-throughput sequencing, is the catch-all term used to describe a number of di_erent modern sequencing technologies including (NGS PLATFORMS):

- Roche 454 sequencing
- Illumina (Solexa) sequencing
- SOLiD sequencing
- Iontorrent

These recent technologies allow us to sequence DNA and RNA much more quickly and cheaply than the previously used Sanger sequencing, and as such have revolutionized the study of genomics and molecular biology.

2. Denovo Assembly Using Illumina Reads

In this study, we systematically studied and compared the performance of four commonly used de novo short read assembly tools including Velvet ,ABySS, SOAP denovo, CLC Genomics Workbench (commercially available) and Trinity following different approaches. Most of these assemblers are based on de Bruijn graphs wherein sequence reads are broken into smaller sequences of DNA, referred to as k-means, where k-denotes the length of these smaller sequences.
De novo transcriptome assembly was performed following two different approaches. In the first approach
(best k-mean strategy), redundant (approx. 86 million) and non-redundant (approx. 24 million) high quality short reads were assembled using short read assemblers mentioned above. The k-value ranged from 19 to 61 for Velvet, ABySS and SOAPdenovo, whereas CLC and Trinity softwares were run

at default settings. The best k-mer assembly was identified on the basis of various assembly parameters. In the second approach (additive k-mer followed by TGICL), a two-step strategy was employed. Initially, contigs generated for all k-mer values by respective assembler were merged and redundant sequences were removed by cd-hit tool . The non-redundant contigs thus generated were assembled again using TGICL suite , as TGICL suite is effective in assembly of long reads rather than short sequences generated by NGS platforms. To check the integrity of assembled transcriptome generated using various assembly tools and following different strategies, high quality short reads were mapped back onto respective assembled Unigenes using CLC workbench software and compared for total mapped back reads and unique mapped back reads. Total mapped back reads depict the number of reads involved in the formation of Unigenes, while uniquely mapped back reads signifies the number of short reads, each showing single target in the assembled transcriptome. Ideally, assembly finalized for further annotation should have high percentage of total as well as uniquely mapped back reads. Various other parameters like total number of contigs (> 100 bp), N50 length and average contig length were also taken into consideration as a function of k-mer length to select the best possible transcriptome assembly. N50 length is a weighted median statistics such that 50% of the entire assembly is contained in contigs equal to or larger than this value. Since higher N50 length indicates better performance of the assembly tool, more weightage was given to N50 value rather than average read length wherever a marginal difference was present in average read length of compared assemblies.

3. Proposed Method

Vibrio cholerae was isolated using Next Generation Sequencing Techniques. NGS is a disruptive technology that has found widespread acceptance in the life sciences research community. The high throughput and low cost of sequencing has encouraged researchers to undertake ambitious genomic projects, especially in de novo genome sequencing.

Work Flow

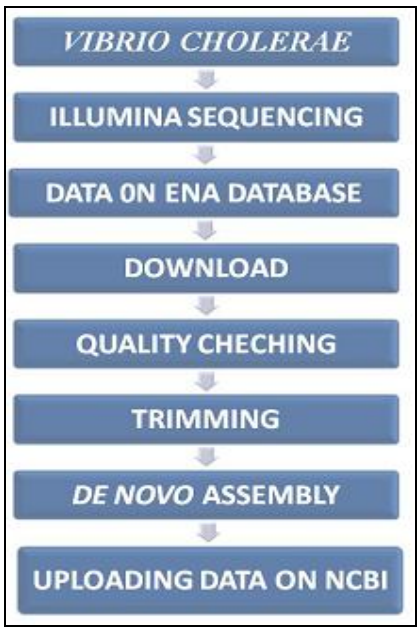

Figure 1: Work Flow Chart of data flow

The sequence data of the new strain HC-63A1 was derived from public database such as ENA, NCBI. Sequenced data is downloaded using these databases(We used DNAnexus to download the data).This data is in zip format so it extracted to get original sequenced file. Quality checking of data is done using DNAnexus which computes extensive sample statistics and quality control metrics to help judge whether given sequencing run was good, and if not, why. Determining the source of the problem allows center and users to work together to correct the issue for future runs. After quality checking will go for triming the data; CLC Genomics Workbench offers a number of ways to trim sequence reads prior to assembly and
mapping, including adapter trimming, quality trimming and length trimming. For each original read, the regions of the sequence to be removed for each type of trimming operation are determined independently according to choices made in the trim dialogs. Next step is de novo assembly, The de novo assembly algorithm of CLC Genomics Workbench offers comprehensive support for a variety of data formats, including both short and long reads, and mixing of paired reads.
The de novo assembly process has two stages:

- First, simple contig sequences are created by using all the information that are in the read sequences. This is the actual de novo part of the process. These simple contig sequences do not contain any information about which reads the contigs are built from.
- Second, all the reads are mapped using the simple contig sequence as reference. This is done in order to show e.g. coverage levels along the contigs and enabling more downstream analysis like SNP detection and creating mapping reports. Note that although a read aligns to a certain position on the contig, it does not mean that the information from

this read was used for building the contig, because the mapping of the reads is a completely separate part of the algorithm.

To only have the simple contig sequences as output, this can be chosen when starting the de novo assembly and we also assembled using tools like DNAnexus, DDBJpipeline. Comparative analyses involving results of CLC, Velvet, Abyss and SOAPdenovo tools from different platforms.

4. Assembly Algorithms

The sequence assembly problem is essentially one of constructing a DNA sequence superstring that explains the observed set of sequence reads. This superstring might represent the DNA that was subjected to sequencing. If the data were completely error-free, then we would expect every sequence read to be contained within the superstring. So, we might be tempted to formulate sequence assembly as finding a superstring that contains all sequence-read strings as substrings. In real biological sequence data, the problem is more complicated. Sequencing error rates may be as high as 1_4% per nucleotide implying that many of the sequence reads contain mismatches with respect to the solution superstring. Furthermore, there will inevitably be multiple solutions; that is, we could propose many possible superstrings that satisfy the criterion of containing all the observed sequence reads. So which superstring is the best one? In a spirit of parsimony, we might choose the shortest superstring, but this would likely not be the biologically correct one. The reason is that real genomic DNA sequences tend to contain large numbers of perfectly and/or imperfectly

repeated sequences, which would be erroneously collapsed in the shortest superstring. Some of the major differences between the popular assembly programs are to be found in their strategies for dealing with repeats. A further complication is that sequence reads can originate from the reverse complement as well as from the forward orientation of the template DNA sequence.

Myers has proposed a third alternative to the overlap graph and the de Bruijn/spectrum graph. He introduced the concept of a string graph, which is a simplified graph derived from one in which the vertices are sequence reads and the edges are overlaps, and described algorithms for efficiently generating a string graph from a set of sequence reads. The method consists of four steps:1 An overlap graph is generated from an all-against all alignment of the sequence reads.2 The overlap converted to a string graph by merging and reducing redundant overlaps and edges.3 False vertices and edges are identified and eliminated using a network flow approach.4 An Eulerian path or circuit is found, which defines the assembly. This is much easier to find than the Hamiltonian path of an overlap graph.

5. Methods for Scaffolding

Once assemblers have generated contigs using one of the above methods, it is necessary to group contigs together in the correct orientation and order. Typically this is done by exploiting the additional information offered by paired-end reads. A read pair that spans two contigs is taken as evidence for the juxtaposition of those two contigs within the genome. Again, this problem can be formulated using a graph-based approach; this time, the contigs are modeled as nodes (vertices) and matching read pairs are modeled as edges connecting the pair of contigs. Again the algorithm consists of finding an optimal path through the graph. In practice, there is often inconsistency in the pattern of read-pair links arising from formation of chimaeric DNA molecules and other technical issues. SOAPdenovo assembler uses a minimum of three read pairs as the criterion for defining order and distance between contigs ,whilst for ABySS the default criterion is five read pairs Scaffolding algorithms may attempt to minimize incongruence between the proposed assembly and the observed read-pair data using consensus from large numbers of read-pairs. The estimated clone length should be approximately equal to the distance between the reads; this criterion is used to identify erroneous pairing data and to choose between alternative paths. Any scaffolding algorithm must solve three main problems such as(i) find all connected components in the defined graph, (ii) find a consistent orientation for all nodes in the graph and (iii) given length estimates of the edges, embed the graph on a line/circle to minimize the number of constraints that are invalidated. The last two problems are defined as NP-complete, but there are good heuristics available to overcome this. Also, in the final step, it is not necessary to minimize completely as the presence of multiple edges from a single node can indicate the presence of mis-assemblies that may require manual intervention and/or additional data to correct. Short reads make the added information offered by paired-end data essential for denovo assembly of all but the simplest genomes. The fundamental limitation of any read length is that any contigs resulting from an assembly will necessarily be limited to regions of the genome that do not have repetitive elements

longer than the read length. Shorter read lengths make this problem particularly difficult. Additionally non-unique elements arising from gene-duplication or multi-copy domains can lead to mis-assemblies of contigs. To resolve these problems paired-end data can be used. Essentially each section of DNA is sequenced twice-once from each end. By carefully controlling this length of DNA (typically either 200_800 nucleotides, or with a circularization protocol 1_10 kb, the so-called insert size) it is possible to obtain an estimate for how far apart each read should appear in the final assembly. In this way, assuming that at least one read can be mapped to a unique position it is possible to assign at least an approximate

location for its partner. A combination of coverage deviation from the median level of coverage and paired-end information has been found to be sufficient in most cases to obtain a good assembly for even relatively long repetitive regions. Paired reads are not the only source of long-range information useful

for scaffolding. Another approach is to use long reads or contigs from another sequencing technology such as 454 of capillary sequencing. The Velvet assembler, for example, can take as input mixtures of long and short sequences. Alternatively, contigs assembled from short reads can be combined with longer sequences using a long-read assembler such as Minimus .

6. Choosing an Assembler

For the biologist faced with assembling real data, which of the programs is the 'best'? First, we should point out that any answer to that question might soon become out of date as this is an active field and existing software is continually being improved whilst new programs are being developed. However, the key issues likely to remain are usability and assembly quality. Usability comprises a number of factors including hardware and software requirements, ease of installation and execution, and speed. The quality of an assembly comprises both the contiguity (lengths of the contigs or scaffolds) and the accuracy of the assembly. Cultural issues may also be important. A survey would not be a straightforward undertaking, since the quality of the resulting assemblies is likely a function of both the choice of program and the particulars of the dataset; that is some programs may perform better on some datasets than on others. Different programs vary in the details of how they resolve errors and inconsistencies in the data. The nature of these errors and consistencies likely vary between haploid versus diploid genomes. An additional complication is that each assembly program takes several options and parameters. For example, in algorithms based on the de Bruijn graph, results are highly dependent on the choice of k-mer (i.e. k-tuple) size. This means that, even after having chosen which software to use,

it is equally important to choose the optimal parameter values. So there is no definitive answer as to which is the 'best' short-read assembler and there is an urgent need for a comprehensive comparison (or competition) between the candidates on a suitably broad selection of datasets. Such a study should utilize a range of different datasets varying in factors such as size, error-rate, heterozygosity, repeat structure, sequence complexity, sampling bias, read lengths, insert lengths (for paired reads or mate pairs), etc. In the meantime, we would make some suggestions based on head-to-head comparisons in the literature as well as our own experience. .Large memory requirements mean that assembly of non-hierarchical reads from large (e.g. mammalian) genomes is only practically feasible using a parallelization strategy such. However, a better solution might be to generate hierarchical sequence data from such genomes, though these methods are more laborious and may require larger quantities of expensive reagents.

7. Results And Discussion

From DNAnexus platform we obtained total number of sequences 13,243,534 sequence length is 99 and %GC is 47%. Once quality checking is done, the next step is to perform trimming where total number of reads is 13,243,534 , average length is 99, Number of reads after trim is 13,175,586,percentage trim is 99.49% and Average length after trim is 93.3 . Next, we performed repeated denovo assembly with different K-mer values using different tools from 19 to 61 in order to obtain N50 contig value which is called pre-denovo assembly. Later, All the results are compared to find less no. of countings for good de novo assembly. Finally, the best result from good tool data is submitted to NCBI where NCBI staff will verify the data and provide results.

8. Comparative Analysis

Assembly metrics for vibri cholerea genome assembled from illumina paired end data with CLC,

VELVET, SOAPDENOVO & abyss. We performed several assemblies with CLC, velvet, soap denovo and abyss, using di_erent values of Kmer, and we determined that CLC works best with K=23 for Vibrio cholerea data set. Velvet works best with K=49 for Vibrio cholerea data set. Abyss works best with K=59 for Vibrio cholerea data set. Soap denovo gives same result for all values of K for Vibrio cholerea data set.

9. Maximum Contig Length

KMER	CLC	VELVET	ABySS	SOAPdenvo
19	2942	237,790	9,871	598,463
21	2811	36,862	10,600	598,463
23	2664	11,558	6,202	598,463
25	2754	9,587	5,675	598,463
27	2756	9,262	8,387	598,463
29	2768	8,731	4,692	598,463
31	2743	7,998	4,609	598,463
33	2764	6,845	7,443	598,463
35	2753	5,935	3,828	598,463
37	2754	5,173	3,537	598,463
39	2740	4,773	5,838	598,463
41	2764	4,289	3,275	598,463
43	2760	4,017	2,975	598,463
45	2762	3,658	4,839	598,463
47	2764	3,154	2,658	598,463
51	2754	2,927	2,400	598,463
53	2754	3,150	4,022	598,463
55	2788	3,935	2,211	598,463
57	2788	4,894	2,047	598,463

Table 2: Shows no. of Contigs on different tools

10. Conclusions and Future Studies

Next generation sequencing technologies are becoming increasingly affordable, accessible and robust for non-model organisms and also for enabling faster and cheaper data generation. However, the task of extracting meaningful information from such a huge amount of data repositories has become more challenging. Genome assembly carried out on different platforms like CLC, DNAexus, DDBJpipeline. Comparative analyses involving results of CLC, Velvet, Abyss and Soap denovo tools from different platforms. With the current project work it was seen that CLC platform is good for denovo assembly. Vibrio cholerea HC-63A1 genome sequencing contain 47% of GC. In future studies we can go for reference genome assembly to know the chromosomes, genes and proteins.

11. References

1. Ghangal R, Chaudhary S, Jain M, Purty RS, Chand Sharma P (2013) Optimization of De Novo Short Read Assembly of Seabuckthorn (Hippophae rhamnoides L.) Transcriptome. PLoS ONE 8(8): e72516. doi:10.1371/journal.pone.0072516
2. Elena T, Capraru G, Rosu CM, Zam_rache MM, Olteanu Z et al. (2011) Morphometric pattern of somatic chromosomes in three Romanian seabuckthorn genotypes. Caryologia 64: 189-196.
3. Lu R (1997) Eco-geographical distribution of seabuckthorn and prospects of international cooperation. In: S LuM LiJ HuS Liu.orldwide Research & Development of Seabuckthorn. Beijing, China:Science Publishing House & Technology Press. pp. 11-22.
4. Lian YS, Chen XL (2000) The regular patterns of distribution on the natural components in plants of the genus Hippophae L. J Northwest Normal University (Natural Science Edition) 36: 113-128
5. Stobdan T, Angchuk D, Singh SB (2008) Seabuckthorn: An emerging storehouse for researchers in India. Curr Sci 94: 1236-1237.
6. Sezik E, Yesilada E, Shadidoyatov H, Kulivey Z, Nigmatullaev AM et al. (2004) Folk medicine in UzbekistanI. Toshkent, Djizzax, and Samarqand provinces J Ethnopharmacol 92: 197-207.
7. Shinwari ZK, Gilani SS (2003) Sustainable harvest of medicinal plants at Bulashbar Nullah, Astore (Northern Pakistan). J Ethnopharmacol 84: 289-298. doi:10.1016/S0378-8741(02)00333-1. PubMed: 12648828.
8. Dhyani D, Maikhuri RK, Rao KS, Kumar L, Purohit VK et al. (2007) Basic nutritional attributes of Hippophae rhamnoides (Seabuckthorn) populations from Uttarakhand Himalaya, India. Curr Sci 92: 1148-1152.
9. Singh KN, Lal B (2008) Ethnomedicines used against four common ailments by the tribal communities of Lahaul-Spiti in western Himalaya. J Ethnopharmacol 115: 147-159. doi:10.1016/j.jep.2007.09.017. PubMed: 17980527.
10. Ledwood JS, Shimwell DW (1971) Growth rates of Hippophae rhamnoides L. Ann Bot 35: 1053-1058