Genome Informatics assignment 1

303034908

ABSTRACT

Short read paired end data from an unknown species was analysed. First, de novo genome assembly was attempted using the program Velvet on data from an unknown species. The longest contigs from the genome were aligned with BLAST to the NCBI database, giving Buchnera BCc bacterium as the most likely species. Genome to genome alignment using Exonerate was then performed to identify common genes between Buchnera and Escheria Coli. The Gene Ontology was then used to further analyise and compare gene function between the two species.

Introduction

Imagine a newspaper that is placed on a bomb, and is cartoonishly broken into thousands of small snippets of paper. The aim is not just to recreate the newpaper from these snippets, but also to have a semantic understanding of its content, and be able to compare it to other newspapers. This challenge is one of the key struggles of computational biology.

The process of blowing up a newspaper is really describing the Illumina sequencing method. This Next Generation Sequencing (NGS) technology was developed to create huge number of ’short’ sequences in parallel. Data from Illumina can also be paired-end, thus providing further information for analysis.

Velvet, the assembler used in this report, uses single and and paired end reads to create a de Bruijn graph through which a path can be taken; the path becoming the final overall sequence. The k-mer length often has to be decided arbritrarily before hand. Too short there are too many repeats. Too long and what could have been a continuous graph becomes broken up into different paths. In this report the k-mer length was set to 31, the highest typically chosen.

Once a graph is constructed and the path decided, the longest contigs of the graph can be inputted to BLAST to match to other already sequenced organisms.

In the case of this report, a fully sequenced genome of the identified species was already built, and so further analysis of genetic function was done by downloading the species genome. A method for assessing gene function is by sequence comparison. This method makes the assumption that similar genetic sequences lead to similar proteins and therefore similar functions, which sadly is not fully true. However, it is a quick and easy way of understanding an organism’s genome functionally. A key tool for this understanding is the Gene Ontology: a semantic database curated and updated by new experimental evidence which allows for discovery and comparison of gene functions.

Methods and Results

Genome assembly

Genome assembly with Velvet was performed. Analysis of the initial run of Velvetg with no specified parameters was used to estimate the expected kmer coverage and the kmer coverge cutoff. Figure 1 shows a weighted histogram of the node coverage. There are two main peaks, one at around coverage 2 and the other at coverage 22.

Comparison of the results whilst specifying different parameters was required to create an optimal graph (see figure 1. The last row of the table in figure 1 was chosen as the final assembly.

Coverage cutoff	Expected coverage	N50	Number of nodes	Total length of graph	Number of reads used
auto	auto	126	187127	20165915	704095/1129046
8	21	416297	532	435504	149487/1129046
18	21	416297	188	435504	143627/1129046

Table 1. Velvetg parameters and outputs

Alignment for identification

A selection of the longest contigs from the optimal graph were used for local alignment and species identification. Figure 2 shows the top 6 contigs from the optimal graph.

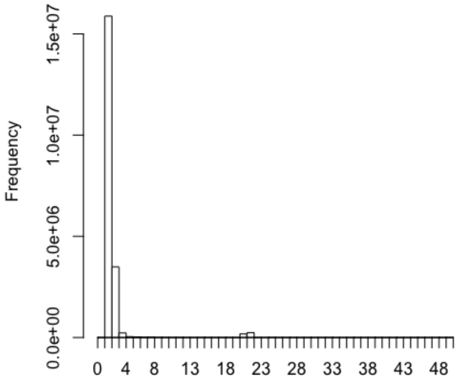

Figure 1. node coverage weighted histogram

Node	Coverage	Contig length
185	21.1	416297
69	24.0	6849
173	12.6	699
51	29.2	543
49	23.0	453

Table 2. Top 5 longest contigs from optimal assembly

As shown in figure 2, the longest contig was around 200 times longer than the next 4 contigs. Figure 2 shows how this contig aligned to the full Buchnera genome with a 100% match.

Alignment was done using the Exonerate program. No special parameters were used in the alignment. The output originally contained 520 alignments. These varied in raw score and similarity. A histogram of the raw scores was plotted (see figure 3 to identify an appropriate threshold for selection of aligned genes. A threshold of raw.score = 125 was chosen as this was the inflection point.

Differences in gene content between species

With the threshold of 125, 117 matches remained. 86 aligned genes were unique to Buchnera and 92 unique to E.Coli. As the E.Coli genome contains 3635 genes and the Buchnera BCc genome contains 365 genes, the percentage of Buchnera genes found in E.Coli was 23.5% whilst the percentage of E.Coli genes found in Buchnera was 2.5%.

The similarity of the aligned genes followed a normal distribution, as shown in figure 6, with a mean of 66.2 and a standard deviation of 5.0. The maximum similarity was 81.8 and the minimum was 57.0.

Gene Ontology analysis was performed on the matched E.Coli genes to infer functional knowledge of the aligned Buchnera genes. Figure 4 shows the most molecular functions inferred from the Gene ontology. The top 3 molecular functions were binding (37%), catalytic activity (17%) and structural molecule activity (13%). Figure 5 instead shows the biological processes inferred by the gene ontology. These were largely dominated by metabolic processes (53%) and cellular processes (47%).

Discussion

Genome assembly

In this report, the histogram in figure 1 showed a very distinct gap between the very frequent low coverage nodes (probably errors) and the comparatively much less frequent higher coverage nodes. The node coverage can be thought of as the number of reads that are used to create a node in the de Brujn graph. Therefore, a higher coverage is a metric for more confidence that that node or contig is correct. By removing low coverage low confidence nodes, the final graph becomes easier to create even if there may be some mistakes or omissions.

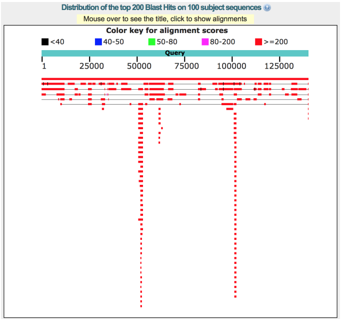
Distribution of the top 200 Blast Hits on 100 subject sequences
Mouse over to see the title, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
25000
50000
75000
100000
125000

Figure 2. BLAST alignment of longest contig to Buchnera Genome

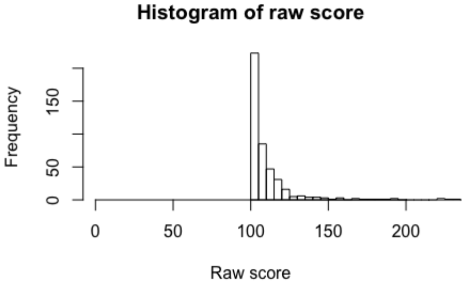

Figure 3. Zoomed in histogram of raw alignment scores

The N50 is the length of the contig that crosses the half-way point when adding all the shortest contigs to each other and comparing to the overall length. A low N50 therefore suggests that the graph is mostly made of very short contigs, which is not very helpful when trying to assemble ideally one continuous sequence. As we can see in figure 1, the N50 in both the second and third row was 416297, around 95% of the total length. This suggests that using those parameters the graph contains one node/contig which is 95% the length of the whole sequence. This is shown to be the case in table 2.

Alignment for identification

Identification using longest contigs is very dependent on the length of the contigs compared to the overall length (which is often unknown when first performing this), as well as any evolutionary similarity with other species. In our results, the longest contig had a 100% match with the whole Buchnera BCc genome. However, normally no one contig contains a whole genome. If the longest contigs happens to be the length of a some genes found in many bacterial species, identification with certainty would not be possible. Moreover, as this data comes originally from experimental scenarios, contamination and methodological artifacts could affect the result. For example if the genome assembly produces two nodes which have equally long contigs and one is from the actual species and the other is from a contaminant species, confidence over the identity will also be lowered.

As shown in figure 3 the raw score for the original unthresholded alignments follows an L shape with an inflection point at score=125. Much of the original many-to-one and one-to-many matching results can be explained by the fact that many alignments were not high scoring enough and therefore likely not actual gene to gene matches. In fact, once the threshold was set, the number of unique Buchnera aligned genes and the number of unique E.Coli aligned genes were very similar: 86 and 92 equivalently.

Differences in gene content between species

The similarity between the alignments, even with the threshold follows a distribution with mean of 66/100. An interesting comparison would be to run exonerate on E.Coli both as query and target. Such a comparison would give a sense of whether a percentage similarity of 66% is low or high for this type of method.

There are different reasons why the gene number of Buchnera is only a tenth of that of E.Coli. The first is the sheer number of the two genomes. The E.Coli genome contains 4,639,221 base pairs whilst the Buchnera genome only 640,681 base pairs. As there is typically a link between size of genome and number of genes, this may be a factor.

Another possible reason is the historical importance of E.Coli. The first E.Coli genome was fully sequenced in 1997 by Blatter et al. 1997 . As a model organism, the quality of annotations and work done on the E.Coli organism would have been

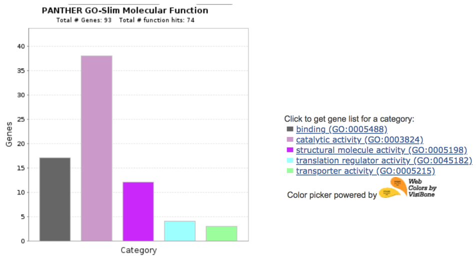

Figure 4. Bar graph of molecular functions of matching Buchenara:E.Coli genes

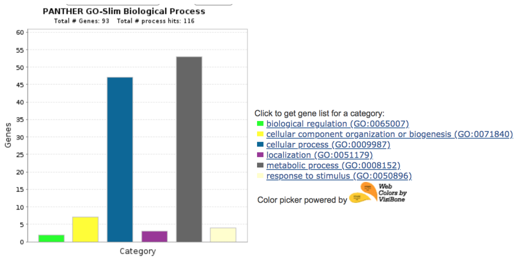

Figure 5. Bar graph of biological processes of matching Buchenara:E.Coli genes

very high compared to the work done on Buchnera genome.

However, the Buchnera genome was sequenced only 3 years later by Shigenobu et al., 2000. In the paper, it was shown that Buchnera is in fact a symbiotic bacteria that lives inside Aphid cells. This could explain the comparatively low number of genes, as, through evolution, many of the unused genes were removed. It is therefore likely that the main reason for the difference in number of genes is due to the different environmental niche that each bacterium evolved in.

The evolutionary niche is further expressed in the results from the Gene Ontology. Much of the inferred function of the matching genes is in translation, ribosomal small and large subunit assembly and alpha-amino acid biosynthesis (see figures 5 and 4). In fact, these genes are part of the housekeeping genes which allow a bacteria to survive. Further analysis would be needed to assess if all of the housekeeping genes are present.

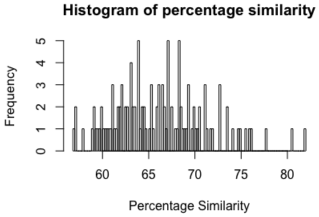

Figure 6. Gene similarity histogram

Bibliography

1. Blattner et al., 1997 "The Complete Genome Sequence of Escherichia coli K-12"
2. Shigenobu et al., 2000 "Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS"