CSC8309: High-Throughput Technologies in Bioinformatics

Assessment

160067554

1 Introduction

Whole genome sequencing is a useful tool for studying infectious bacteria for clinical applications. *Chlamydia trachomatis* is gram-negative bacterium, which is the cause for the most common sexually transmitted infection (STI) (Geneva: WHO 2011), resulting in 100 million cases a year.

Chlamydia trachomatis has been extensively shown to alter host gene expression. In a 2007 study by Schrader (Schrader, *et al*., 2007) *Chlamydia trachomatis* was shown to temporally regulate genes in infected human monocytes.

Recombination is prevalent throughout Chlamydia trachomatis's genome, therefore whole-genome sequencing is crucial for understanding the epidemiology of this bacteria. Monitoring Up and down-regulation of genes during an infection period can be identified through whole-genome sequencing, allowing functional analysis of the genes involved in disease.

Genome assembly refers to the method of merging, by alignment, numerous short DNA sequence fragments into a larger DNA consensus sequence. De novo genome is a computer intensive approach to assembly, assembling a genome without a reference sequence to guide the alignment.

Graph-based genome alignment algorithms handle sequencing data to produce a graph structure which represents pairwise overlaps of contigs. Optimal alignments can be calculated by traversing through the graph structure. Using a de Brujin graphical alignment, each edge in the system characterises a k-mer long sequence, the overlap between the k-1 prefix and k-1 suffix on the connected nodes. The overlap between two reads can be inferred from finding the Eulerian Path: the path connecting each edge only once. This, therefore, provides a consistent explanation for every consecutive k-mer for any sequence read.

2 Methods

2.1 Genome Assembly

FASTQ Chlamydia trachomatis forward and reverse DNA strands were obtained from the Sequence Read Archive (SRA), from experiment SRX1672495 by the University of Vienna, using Illumina HiSeq 2000.

Quality control for these sequences was performed using FastQC, a java program produced by the Babraham Institute in Cambridge. (Andrews S, 2010) in the Linux command line. This was done to ensure there are no biases in the data. The FastQC reports were saved and examined to assess the quality of the Illumina reads.

Calculation of the k value, for a k-mer coverage of 30, required for de Bruijn assembly was performed using Velvet Advisor (Torsten S, 2014) based on equations 1 and 2, where C = genome coverage, N = number of reads (in millions), l = read length, G = genome size (in megabases) and Ck = k-mer coverage. The number of paired-end reads was set to 260, by referring to the FastQC report. Chlamydia trachomatis has an approximate genome size of 1MB. The K-value was estimated at 111.

Equation: 1

$$C = \frac{N \times l}{G}$$

Equation: 2

$$k = (1 + l) - \left(\frac{l}{C} \times C_k\right)$$

Graph-based assembly was carried out in SPAdes (Anton Bankevich 2012) in the linux command line for both forward and reverse reads. k-mer length of 111, as calculated in Velvet, was specified using the -t option. SPAdes was ran with the –careful option to minimise the number of mismatches in the final contigs, and –cov-cutoff was set to auto, allowing SPAdes to automatically compute coverage threshold. The resulting contigs were saved to be used for further analysis.

2.2 Assessment of Assembly

The N25, N50 and N75 statistics were used to assess the quality of the genome assembly, by using the gnx tool at the linux command line. The N statistic for a given assembly is the length of a contig where at least 25%, 50% or 75%, respectively, of the whole assembly is made up of contigs of equal or greater length.

QUAST, a genome assembly assessment tool (Alexey Gurevich 2013), was also used to assess the quality of the assembled genome. QUAST allows comparison of multiple assemblies of the same genome, with a provided reference genome, to compute a range of useful metrics, including misassemblies and mismatches.

2.1 Assembly Annotation and Alignment

Rapid annotation for our assembled genome was performed using Prokka (T Seemann, *et al*., 2014), generating standard output files in gff formats. The contigs produced by SPAdes were required to be reformatted, reducing the contig ID's of the FASTA files to be <20 characters, before running Prokka.

Genome alignment was carried out with Bowtie (Langmead and Salzberg 2012) in the linux command line. Bowtie uses Burrows-Wheeler Transform implementation methods for seeding the alignment before extending the seeds using the gapped Smith-Waterman algorithm. The assembled Chlamydia trachomatis genome was aligned to RNA-Seq reads retrieved from Gene Expression Omnibus experiment GSE44253 (Humphrys MS, *et al*., 2013).

Before aligning with Bowtie, the target genome was first indexed using the Burrows-Wheeler Transform in the linux command line. Bowtie was then ran on each of the four pairs of FASTQ RNA-Seq files provided by experiment GSE44253. The resulting SAM (Sequence Alignment Map)

files were compressed to BAM files, using Samtools in the linux command line, to reduce file size.

The number of aligned reads which can be assigned to features in the annotated genome produced by Prokka were quantified using HTSeq (Simon Anders 2014), a Python package for processing high-throughput sequencing data. The Prokka annotated genome was first reformatted using an executable script, before using the "htseq-count" method in the linux command line, on each of the BAM files, for counting the overlapping reads with annotation features.

2.2 Data-Analysis of Expression

Since the GEO experiment GSE44253 contains incomplete sample data, data analysis was carried out in R Studio (RStudio Team 2015) with simulated data based on available literature. Analysis was carried out using Bioconductor packages. DESeq2 (M.I. Love 2014) is a specialised RNA-Seq count data analysis package, using a negative binomial distribution model for testing differential expression. Due to the nature of negative binomial distributions, variance can be adjusted independently from the mean, allowing for suitable modelling of RNA-Seq count data over a large dynamic range. Normalisation and stabilisation steps for handling the raw count output was carried out in DESeq2.

DESeq2 was also used to measure the dispersion of the count data, a measurement of the variance of the count data, modelling the "biological noise" within the system. The variance was calculated as the sum of the measure of biological variance and the measure of uncertainty of read counting as a quantitative metric (shot noise). When genes are highly expressed the biological noise dominates, whereas the shot noise dominates for lowly expressed genes.

Determination of which genes were differentially expressed was also performed by DESeq2, using Wald's Negative Bionomical test, on the RNA-Seq data. Benjamini-Hochberg correction methods implemented in DESeq2 were used to control the false discover-rate.
The differential expression = the contrast is 24 hours post-infection against 1-hour post-infection, with T24 as the numerator for the fold change, and T1 as the denominator for the fold change. This means that a fold change greater than 1 (i.e. a positive log fold change) means expression is greater (upregulated) in T24 compared to T1, and a fold change between 0 and 1 (i.e. a negative log fold change) means expression is lower (downregulated) in T24 compared to T1.

A MA-Plot was created in R by DESeq2, resulting in a scatter plot of the log2 fold changes against the mean on the normalised counts. Datapoints with a P-value lower than the passed alpha argument (0.01) are coloured red.

Principal Components Analysis (PCA) was also carried out by DESeq2, allowing for simple outlier detection and to determine whether relationships between samples are as expected.
To visualise the Euclidean distance between the samples, a sample distance Heatmap was produced using gplots in R.

PLSDB – a plasmid database (Valentina Galata, *et al*., 2018) was used to initially suggest the plasmid identity in the assembled genome from the SPAdes-produced contigs FASTA file.

Individual BLAST (Altschul, S.F, *et al*., 1990) searches were performed on each of the Prokka annotated nodes to locate which node the plasmid is situated, and to verify the accuracy of the plasmid identity as suggested by PLSDB.

The protein sequences in from Prokka annotated NODE_3 were directly compared with the suggested plasmid (CtrE-103) to further assess similarity, and verify the genes present on the plasmid.

Assessment of differential expression between 1-hour post-infection samples and 24 hours post-infection samples in the simulated data for the plasmid genes was performed through comparison of the Log2FoldChange of expression for the previously identified genes on NODE_3. The results were recorded to be later analysed.

3 Results

3.1 Genome Assembly

In the FastQC report for both forward and reverse strands in the assembled genome, per base sequence content is highlighted red. The initial ~20 bases for each contig being are shown to be bias, but after the initial bases the per base sequence content are consistent.

The N25, N50 and N75 statistics calculated in SPAdes, for the contig FASTA files, were all shown to have the same number: 1043176.

3.2 Identification of plasmid in assembled genome

By submitting the SPAdes produced contig FASTA files to PLSDB, the identity of the plasmid was suggested to be Plasmid NZ_CP015295.1.

The BLAST result from NODE_3 FASTA sequence showed Chlamydia trachomatis strain E-103 plasmid CtrE-103 with 100% Per. Identity and highest Max score, with a sequence length of 7502; the length of NODE_3 is 7613.

The genes present on plasmid CtrE-103 were viewed in the GenBank record of the sequence. The genes annotated in GenBank for plasmid CtrE-103 were compared with the Prokka annotated genes on NODE_3 to assess similarities as shown in Table 1.

(A)

PROKKA_ID	Proteins present in NODE_3	Gene
PROKKA_00949	Phage integrase family protein	N/A
PROKKA_00950	Tyrosine recombinase XerC	XerC
PROKKA_00951	Replicative DNA helicase	dnaB
PROKKA_00953	Virulence plasmid protein pGP3-D	N/A
PROKKA_00954	Hypothetical Protein	N/A
PROKKA_00947	Hypothetical Protein	N/A
PROKKA_00948	Hypothetical Protein	N/A
PROKKA_00952	Hypothetical Protein	N/A

(B)

Proteins present in CtrE-103	Gene	PROKKA_ID of matching proteins in NODE_3
Tyrosine-type recombinase/integrase	XerC	N/A
site-specific integrase	DNA_BRE_C	PROKKA_00950
Replicative DNA helicase	DnaB	PROKKA_00951
DUF5597 domain-containing protein	N/A	PROKKA_00952
hypothetical protein	N/A	PROKKA_00953
hypothetical protein	N/A	PROKKA_00954
ParA family protein	ParA	N/A
Virulence plasmid protein pGP6-D	N/A	PROKKA_00948

Table 1. (A) PROKKA_IDs of the genes present in NODE_3, with associated protein. **(B)** Genes present in CtrE-103 plasmid, with associated PROKKA_ID with matching genes in NODE_3.

Comparing translations of the suggested genes in NODE_3 of the prokka.results GenBank file, and the GenBank record for CtrE-103 showed tyrosine-type recombinase/integrase from CtrE-103 as having no matching protein sequence with node 3, however XerC tyrosine recombinases are seen in both DNA strands.

Site-specific integrase from ctrlE-103 has the same protein sequence as prokka annotated Tyrosine recombinase XerC (PROKKA_00950). Replicative DNA helicases in both DNA sequences have the same protein sequence (PROKKA_00951). DUF5597 domain containing protein from CtrE-103 has a matching protein sequence as hypothetical protein PROKKA_00952. Virulence plasmid protein pGP3-D (PROKKA_00953) in Prokka results has same protein sequence as hypothetical protein in CtrE-103. Hypothetical protein PROKKA_00954 sequence matches hypothetical CtrE-103 protein sequence. The ParA family protein in CtrE-103 does not match any protein sequence seen in the Prokka annotated sequence. Hypothetical protein PROKKA_00948 sequence matches the CtrE-103 Virulence plasmid protein pGP6-D sequence. PROKKA_00947 hypothetical protein is not seen in CtrE-103 plasmid.

Searching for "plasmid" in the Prokka annotated table highlighted a plasmid gene, pepF1, present in NODE_1. Subsequent BLAST searches of the NODE_1 DNA FASTA sequence were performed to assess where the sequence was derived.

3.3 Expression data analysis

The DESeq2 produced MA-Plot for the simulated data in R is shown in Figure 1. Numerous red-data points are shown indicating significant changes in expression when compared to their expected expression. Black datapoints shown near the threshold line represent genes with expected expression.

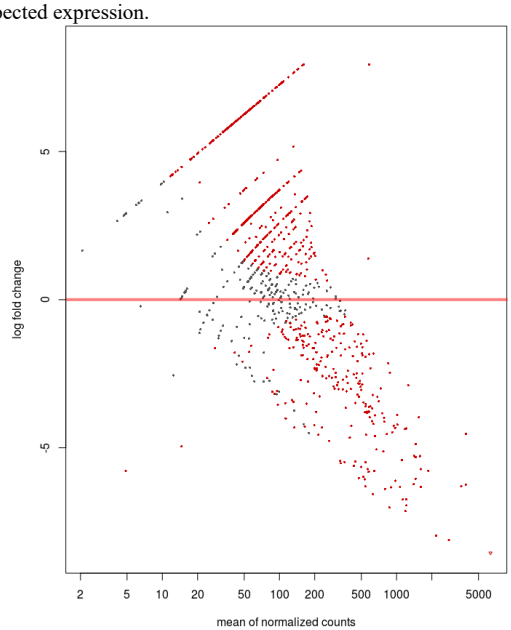

Figure 1. MA-Plot representing Log fold change vs mean of normalised counts for the simulated data in R. The threshold line is shown at log fold change of 0, representing the expected value of gene expression. Black datapoints are genes with expected expression, whereas datapoints are coloured red if their expression is not as expected in healthy cells.

The results from Principle Component Analysis are noted in Figure 2. T1 samples were shown to have little variance, and tight clusters. T24 samples were demonstrated to have a spread of ~25 5% variance, where two of the three nodes having similar variance. The third node has a variance of ~15 5% variance.

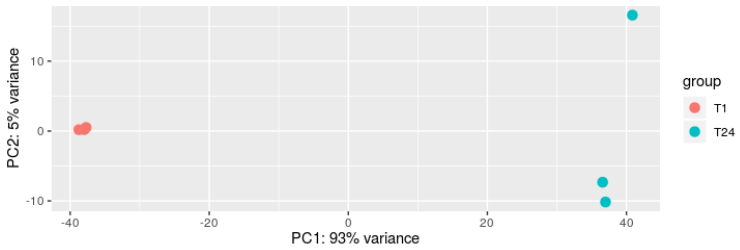

Figure 2. Principle Component Analysis (PCA) plot of the variance of expression between 1-hour post infection samples and 24 hours post-infection sample repeats. T1 samples are shown to be clustered, and have little variance, whereas T24 samples are spread between -10 and ~15 5% variance.

Dispersion estimation plot is shown in Figure 3. Most genes are shown to lie within the fitted estimate, however genes with abnormal expression strength are encircled blue.

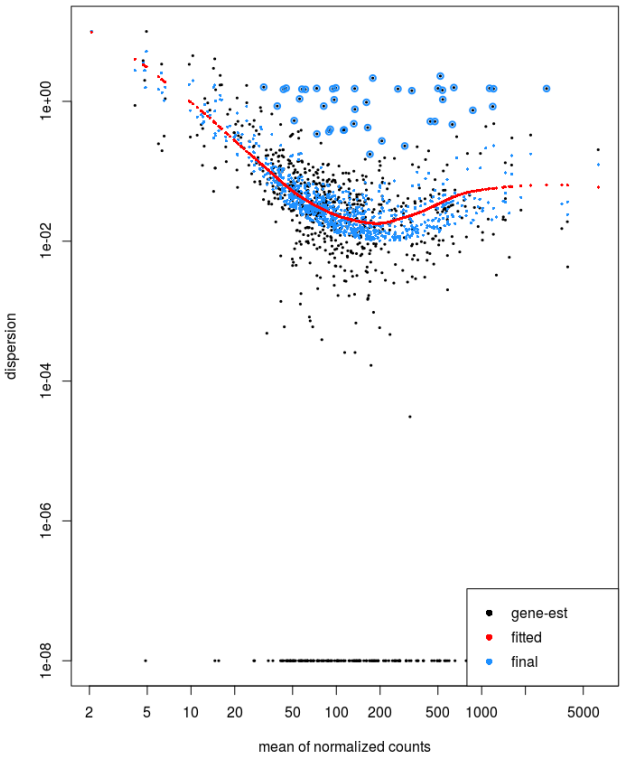

Figure 3. Plot of dispersion estimates over the average strength of expression for the simulated dataset. Black dots represent gene-wise maximum likelihood estimates. The red line denotes the overall trend of the dispersion-mean dependence. The fitted trend is used as a prior mean for a second estimation round, resulting in final estimates, noted by blue dots. Black points encircled in blue represent dispersion outliers.

Sample distance heatmap between 1-hour post-infection samples (T1A, T1B, T1C) and 24 hours post-infection samples (T24A, T24B, T24C) shown in Figure 4. T1 samples are all shown to have significant differential expression to T24 samples, as represented by yellow colouration.

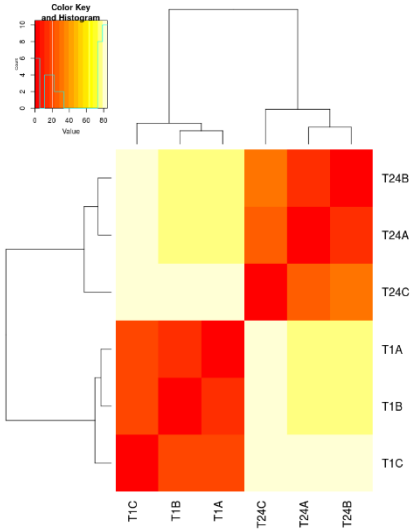

Figure 4. Sample distance heatmap between 1-hour post-infection samples (T1A, T1B, T1C) and 24 hours post-infection samples (T24A, T24B, T24C). The colour intensity corresponds to differential expression, where the redder the colour, the more similar the expression. Samples taken 1-hour post infections are shown to have significant differential expression to samples taken 24-hours post-

3.4 Comparison of 1-hour post-infection and 24 hours post-infection for simulated data of plasmid genes

Log2 fold changes of differential expression between 1-hour post-infection samples and 24 hours post-infection samples in the simulated data for the plasmid genes are noted in Table 2. PROKKA_00953 is shown to have the highest Log2FoldChange of 7.936576.

PROKKA_ID	Proteins present in NODE_3	Log2FoldChange
PROKKA_00949	Phage integrase family protein	3.410331
PROKKA_00950	Tyrosine recombinase XerC	1.391809
PROKKA_00951	Replicative DNA helicase	-0.674441
PROKKA_00953	Virulence plasmid protein pGP3-D	7.936576
PROKKA_00954	Hypothetical Protein	2.189226
PROKKA_00947	Hypothetical Protein	2.951134
PROKKA_00948	Hypothetical Protein	3.953436
PROKKA_00952	Hypothetical Protein	-0.2236657

Table 2. Table showing PROKKA_ID's of genes present in NODE_3, and their respective Log2FoldChange between 1-hour post-infection samples and 24 hours post-infection samples for the simulated data for the plasmid genes.

4 Discussion

3.5 Genome assembly and alignment

The *C. trachomatis* genome assembly as shown to be a success as described in the FastQC report, where only issues were present in the per base sequence content. This may be caused by steps in sequence preparation, where enzymes used to fragment the reads create bias, by cutting in slightly different places. The bias is consistent, only present within the first 15 bases, and so further supporting the bias may be due to enzymatic factors, rather than intrinsic issues such as contamination.

Assessment of genome assembly by N statistics showed N25, N50 and N75 to have the same value: 1043176. This indicates that the contig that takes us over the 25% length mark is the same as the contig that takes us over the 75% length mark, which may be attributed to the atypical sequencing of the entire *Chlamydia trachomatis* genome in one read. The subsequent QUAST report showed the contigs to have extremely similar cumulative lengths to the reference sequence.

The quality of alignment of the *C. trachomatis* genome to the RNA-Seq reads from GEO experiment GSE44253 depends on the quality of the reads from the experiment. Interestingly, in the described experiment for GSE44253, over 100 million transcriptomes of Chlamydia and Chlamydia-infected cells were retrieved, not the 4 uploaded in GEO. Due to this disparity, the data generated from the alignment of these sequences with our assembled genome may not be fully representative of differential expression in host-infected. Further work would include obtaining the full dataset produced in experiment GSE44253, to be used to more accurately model changes in gene expression.

3.6 Plasmid Identification

The most likely location on the plasmid, as suggested by Table 1, is NODE_3. NODE_3 shares 75% of the genes shown in plasmid ctrlE-103, including replicative helicases and phage integrases – genes commonly present in plasmids.

The plasmid gene present in NODE1 may be remnants from evolutionary history. NODE_1 was too large to be passed through BLAST, so further work would include identifying the origin of NODE_1.

The disparities shown in protein names between the annotation via Prokka and the GenBank record for *Chlamydia trachomatis* strain E-103 plasmid CtrE-103 (Table 1), may be due to the rapid and shallow nature of Prokka annotation. As seen by the results some proteins are annotated as hypothetical, even if the protein has been characterised. Improving annotation would include using updated proteins names, when applicable, such as assigning Virulence plasmid protein pGP6-D to hypothetical protein PROKKA_00948.

3.7 Changes in gene-expression between 1-hour post-infection and 24-hours post infection for simulated data of plasmid genes

Genes are clearly shown to be differential expressed between 1-hour post-infection and 24-hours post-infection, as shown in the heatmap in Figure 4.

The changes in expression of the genes present in NODE_3 of the Prokka annotated sequence between 1-hour post-infection and 24-hours post infection are shown in Table 2. All NODE_3 proteins were shown to change expression, suggesting they are involved during infection. Virulence plasmid protein pGP3-D expression was shown to increase the most, with a Log2FoldChange of 7.936576. This suggest the protein is more important for the pathogenesis of *Chlamydia trachomatis* than the other proteins. Further work could include immunization studies with pGP3 gene and evaluating the infection progress of immunized cells with *Chlamydia trachomatis*.

References

Geneva: World Health Organization; 2011. Global prevalence and incidence of selected curable sexually transmitted diseases: Overview and estimates.

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Schrader, Sina et al. “Expression of inflammatory host genes in Chlamydia trachomatis-infected human monocytes.” Arthritis research & therapy vol. 9,3 (2007): R54. doi:10.1186/ar2209

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Torsten Seemann, Victorian Bioinformatics Consortium, Monash University, Available online at: http://dna.med.monash.edu.au/~torsten/velvet_advisor/

Bankevich, Anton et al. “SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.” Journal of computational biology : a journal of computational molecular cell biology vol. 19,5 (2012): 455-77. doi:10.1089/cmb.2012.0021

Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi and Glenn Tesler, QUAST: quality assessment tool for genome assemblies, Bioinformatics (2013) 29 (8): 1072-1075

Torsten Seemann, Prokka: rapid prokaryotic genome annotation, Bioinformatics, Volume 30, Issue 14, 15 July 2014, Pages 2068–2069, https://doi.org/10.1093/bioinformatics/btu153

Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012, 9:357-359.

Humphrys MS, Creasy T, Sun Y, Shetty AC et al. Simultaneous transcriptional profiling of bacteria and their host cells. PLoS One 2013;8(12):e80597. PMID: 24324615

S Anders, T P Pyl, W Huber: HTSeq — A Python framework to work with high-throughput sequencing data. bioRxiv 2014

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Love MI, Huber W, Anders S (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” Genome Biology, 15, 550. doi: 10.1186/s13059-014-0550-8.

Valentina Galata, Tobias Fehlmann, Christina Backes, Andreas Keller; PLSDB: a resource of complete bacterial plasmids, Nucleic Acids Res., 2018 Oct 31, doi: 10.1093/nar/gky1050

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.