Genome Informatics: Assignment 1

University of Cambridge

Henrik Åhl

November 1, 2016

Abstract

The current pace of development for *Next Generation Sequencing* (NGS) has allowed for a rapid decrease in cost for genome sequencing from short-read data. Naturally, this provides a high increase in the amount of data available for analysis, which subsequently can aid in understanding the complex patterns and dependencies of the genome on a nucleotide level. More specifically, the amount of data provides a firm basis for novel gene detection and comparison between species.

In this report, Illumina HiSeq sequencing has been performed on an unknown bacterial sample, which we assemble using the Velvet *de novo* assembler, developed by Zerbino and Berney [1]. The assembled product is identified and compared by alignment to bacterial model organism *Escherichia coli* (E. coli).

We find that the species, identified as *Buchnera Aphidicola* in *Cinara cedri* (BCc), contains a significantly lower amount of overall genes, and likewise a smaller genome than E. coli. However, a large portion of the BCc sample is present in the E. coli genome, implying that the BCc genome consists to a larger extent of genes strictly necessary for survival.

Preface

This is an assignment report in connection to the *Genome Informatics* module in the Computational Biology course at the University of Cambridge, Michaelmas term 2016. All related code is as of November 1, 2016 available on `https://github.com/supersubscript/compbio/tree/master/src/sp_assignments/assignment_1/`, or available per request by contacting hpa22@cam.ac.uk. Likewise, the corresponding assignment can be found on `https://github.com/supersubscript/compbio/tree/master/general/gi_assignment_1.pdf`.

1 Introduction

Buchnera Aphidicola is a bacterial obligate endosymbiont found primarily in connection to aphids. It contains one of the smallest known genomes of living organisms, hosting around 422000 base pairs and 400 genes. However, the genome is also one of the most genetically stable [2].

The Buchnera Aphidicola strain found in aphid *Cinara cedri* (henceforth abbreviated BCc) was at the point of sequencing found to be roughly 200 kilobases smaller than other sequenced B. Aphidicola strains, furthering the notion that the species and its various strains are undergoing a genome reduction process over time [3]. In particular, B. Aphidicola is known to lack the genes necessary for production of lipopolysaccharides for its outer membrane, which is done by gene rfaS in related species *Escherichia coli* K-12 [4]. Notably, BCc also lacks most metabolic functions, is largely unable to supply for its own nutritional needs, and relies heavily on its symbiotic status in order to survive. The species thus forms an interesting model organism for studying the process of genome reduction and the possibility of a lower threshold of genes required for prolonged existence. It has, however, been hypothesised that the bacterium with its continuing genome evolution is destined for future replacement by another endosymbiont in its aphid host [3].

By use of the Velvet assembler for *de novo* short read sequencing, we assemble our bacterial sample under a range of various configurations of k-mer lengths, ultimately choosing the parameter configuration which renders the best N50 score for future investigations. We utilise this choice to go into a more fine-grained analysis by finding the optimal cut-off lengths and our estimated coverage, consequently attaining our optimised assembly. The genome is then identified and compared to E. coli with respect to genomic qualities such as gene number and gene similarities. Finding that the genomes are considerably unlike each other, we pinpoint differences and their possible causes. Finally, we perform an enrichment analysis on common genes, thereby determining their biological functions.

2 Methods and Results

We perform the initial assembly by running Velvet on a set of different k-mer lengths. Our data consist of roughly 1129000 reads of paired-end data of length 100, with an average read-length of ca. 500 base pairs. Investigating k-mer lengths between 17–31 in size, with an expected coverage of length 22, we find that the N50 score reaches a maximum at a k-mer length of 25, as can be seen in fig. 1. As we want to exclude the nodes which likely are due to mutations, we want to trim our assembly for low-coverage data points. We also want to set the expected coverage to the number of times we expect each subsequence to be covered, as it allows us to exclude repeats.

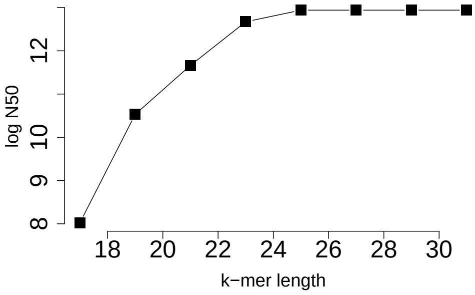


Figure 1: N50 distribution for different k-mer lengths.

Given this, we perform another assembly with k-mer length 25, without cut-off. The outcome of this can be seen in figs. 2 and 3. Noting that there is a bias towards sub-20 coverage in the non-normalised case, and a grouping of length-adjusted coverage nodes in the sub-6 region, we choose to disregard these nodes and repeat the process. Ultimately, we find that an expected coverage of 23, along with a coverage cut-off of 19 gives the most sensible assembly, with the highest N50, and a summed contig length of about 439000 base pairs.

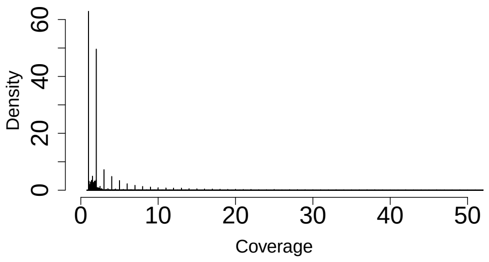


Figure 2: Distribution of coverage for non-normalised contigs.

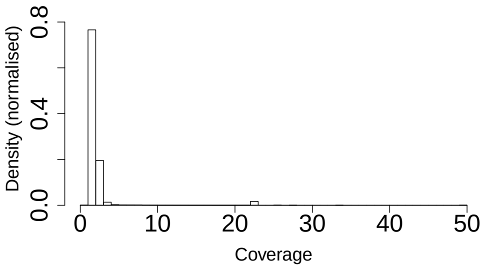


Figure 3: Coverages for contigs normalised with respect to sequence length. Note how a a group of low-coverage nodes are present also after normalisation, and that some are found in the >20 region.

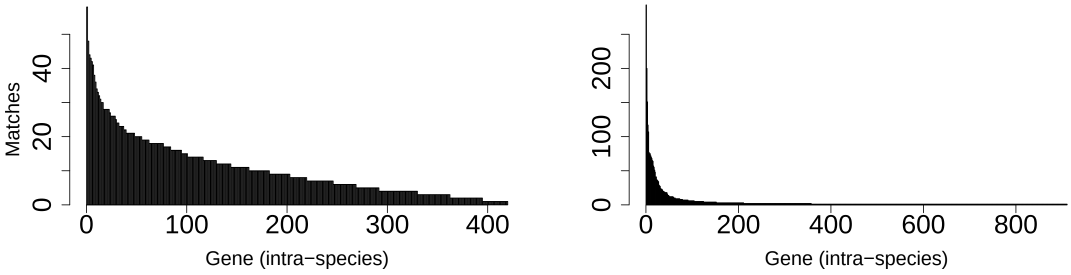
Using the *NCBI-BLAST* interface to identify the species given our assembly, we indeed find a strong (99 %) identity correspondence with B. Aphidicola in *Cinara cedri*, as well as a 99 % query cover. For the closest competitor, Buchnera strain *Cinara tujafilina*, the match does not exceed 75 %, and only reaches a query cover of 31 % in the local alignment.

Retrieving the annotated BCc genes from bacterial genome-portal EnsemblBacteria, we use sequence comparison software Exonerate to align BCc with the closely related species Escherichia coli K-12. Performing a local alignment with affine gaps [5], we are able to identify similar genes in the two species. Table 1 shows the values extracted from the Exonerate alignment after curation. Note in particular that a large fraction of the BCc gene pool is matched by one or several genes in E. coli K-12. Additionally, BCc contains significantly fewer genes overall than the E. coli strain. It also finds matches for almost all of its genes, though it should be noted that the local alignment does not account for the overall similarities between genes, but rather included subsequences. In other words, two genes who for example might have similar binding site sequences has a chance of becoming matched in this process. Moreover, the BCc genes find matches for more genes than the number that are shared.

Table 1: Extracted values from Exonerate run for local alignment with affine gap penalties.

Parameter	*Value*
Total number of genes in E. coli K-12	3635
Total number of genes in BCc	365
Number of genes in both species	317
Matched genes in BCc	350
Matched genes in E. coli K-12	759
Number of matched genes in both species	97
Fraction of BCc genes shared	0.87
Fraction of E. coli K-12 genes shared	0.09

Also due to the local alignment, it is inevitable that many genes will have several matches, particularly if they contain genomic features which are typical for genes of its given kind. Figures 4a and 4b shows this phenomenon, where it is evident that some genes are frequently paired with other genes in the corresponding species. For example, the most frequently found E. coli gene, rfaS, is matched 293 times, likely due to its relatively long nucleotide sequence.



(a) Buchnera Aphidicola in *Cinara cedri*

(b) Escherichia coli K-12

Figure 4: Number of matches per gene for BCc and E. coli K-12 respectively. Note that genes are sorted by number of matches, so that the horizontal position of specific genes differs between the graphs.

The distribution of fraction identity between the alignments is featured in fig. 5. Most alignments have a match fraction around 65 %, meaning that the differences are mostly fairly big. This stands in relation to the fact that 350 out of the 365 genes in BCc find a corresponding match in the E. coli K-12 gene pool. A set of alignments also match with higher accuracy, rendering identity matches of up towards 100 %. However, these matches mostly consist of shorter subsequences and not necessarily the whole genes. In conclusion, most genes that match between species contain significant differences.

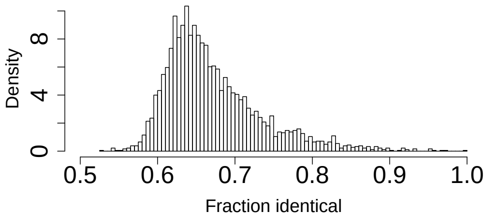


Figure 5: Distribution of fraction identity between nucleotide bases between the compared subsequences in the Exonerate local alignment.

Finally, by performing an enrichment analysis on the genes which occur in both species, it is possible to determine which features the species in particular have in common. Using the Gene Ontology (GO) platform, we find that genes overlapping are related in particular to primary metabolic processes, gene expression, translation, various forms of binding, and processes related to maintenance of intra-cellular structures such as organelles. In short, crucial genetic components for living organisms.

3 Discussion

Accuracy in results are limited by the assembly, but marginally so

Initially, it must by necessity be stated that the results obtained to a slight degree depend on the parameters used for the assembly. The fact that the identification of the species is done by sequence comparison is inherently dependent on our sequenced genome. Still, the fact that such an adequate match is attained in out investigation is reassuring of our method. In more obscure cases it is however a significant limiting factor which ought to be taken into account. In principle, a more systematic approach to reach optimal parameters is clearly favourable, which can be done by performing a more thorough analysis over a larger set of k-mer lengths.

In our particular case, this can of course be seen as a minor problem due to the adequate match of species. Nevertheless, also in the process of identifying the species, we lose some resolution as this is done based on the sequence of the largest assembled contig. This approach is undoubtedly non-stringent, as it completely over-glosses the possibility of novel findings and in practice always relies on partial matching. We can, for example, happen to misidentify more complex sequences as singular, well-known genes or structures.

In addition to this, using the BLAST interface was here done in a fairly offhanded way. As the typical BLAST alignment relies on heuristic methodics and local alignments, also this forms a procedural speed-bump.

Possibly faulty annotations allow for an increase in multiply matched genes

We are, however, able see even in our rough comparison between the species that indeed many of the genes found in BCc also occur in E. coli K-12. Our local alignment hampers the analysis somewhat by rendering multiple matches for many of the genes in both species. In particular in Eċoli K-12, which has a significantly longer genome and larger gene pool, the consequences of this can be clearly seen. The fact that the rfaS gene is matched so frequently is likely due to its known role in the production of lipopolysaccharides – a capability we recall that the BCc strain lacks. This also raises the issue of annotation, as such a significant amount of genes in BCc match this gene signifies that they may in fact rather be pseudogenes, having arisen as a consequence of the degradation of the genome over time. In contrast, a few genes in the BCc genome are not found at all in E. coli K-12, possibly being due to novel genes having developed in a similar fashion.

In E. coli, many genes lack matches at all, which plays according to expectations as the BCc genome is considerably smaller. These have likely fallen out under the evolutionary trajectory due to genomic deletions.

Another possible explanation of genes with multiple matches is that repetitive regions may have been included over several differently annotated genes. Also, as previously mentioned, genes containing common functional subsequences, such as those relating to binding and of certain proteins and the likes, may cause an increase in multiply matching genes. In principle we ought to be able to circumvent these problems by applying a score threshold for the alignment, although with such an approach we obfuscate a significant portion of genomic features which may be of interest, such as our partwise alignment to rfaS.

Common genes with Escherichia coli K-12 tell of essential processes for survival

As many of the genes which exist in both the species relate to functions we can largely consider to be essential, it is not far-fetched to infer that the BCc genome has been scraped down to contain mainly genes strictly necessary for its continued existence. However, even though the strain is known to have lost most metabolic functions, many genes still relate to metabolic processes. Many genes also naturally relate to processes in connection to cellular components such as ribosomal structures and the likes, for which it is hard to consider an organism.

4 Conclusion

We have seen that using sophisticated analytical tools for sequence manipulation and comparison, we are able to assemble a genome with high enough accuracy that we can identify the corresponding species. In comparing our result with reference organism E. coli K-12, we are able to see previously identified differences and reasons for the genomic difference. We can also infer possible reasons for the structure and general function of the genome of our species.

Acknowledgements

A great many thanks to Julian Melgar for no particular reason. Thank you, Julian!

References

[1] Daniel R Zerbino and Ewan Birney. "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." In: *Genome Res.* 18.5 (2008), pp. 821–9. DOI: 10.1101/gr.074492.107.

[2] Francisco J. Silva, Amparo Latorre, and Andrés Moya. "Why are the genomes of endosymbiotic bacteria so stable?" In: *Trends in Genetics* 19.4 (2003), pp. 176 –180. ISSN: 0168-9525. DOI: http://dx.doi.org/10.1016/S0168-9525(03)00041-6. URL: http://www.sciencedirect.com/science/article/pii/S0168952503000416.

[3] V. Pérez-Brocal et al. "A Small Microbial Genome: The End of a Long Symbiotic Relationship?" In: *Science* 314 (Oct. 2006), pp. 312–313. DOI: 10.1126/science.1130441.

[4] Shuji Shigenobu et al. "Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS". In: *Nature* 407.6800 (2000), pp. 81–86.

[5] Guy St C Slater and Ewan Birney. "Automated generation of heuristics for biological sequence comparison". In: *BMC bioinformatics* 6 (2005), p. 31. ISSN: 1471-2105. DOI: 10.1186/1471-2105-6-31. URL: http://europepmc.org/articles/PMC553969.

A Genes annotated in both species

```
##   [1] thrA thrB thrC dnaK dnaJ rpsT ileS dapB carA carB ksgA leuD leuC leuB
##  [15] leuA ilvI ilvH ftsA ftsZ secA aceE aceF pcnB yadR dapD map  rpsB tsf
##  [29] pyrH frr  yaeT fabZ dnaE proS dnaQ nusB cyoE cyoD cyoC cyoB cyoA bolA
##  [43] clpP clpX lon  ppiD mdlA mdlB dnaX ybaB htpG adk  cysS folD ahpC cspE
##  [57] lipA lipB ybeD holA leuS ybeY miaB glnS fldA ybgI sucA sucB gpmA infA
##  [71] trxB serS serC aroA rpsA asnS ompA yccK yceA rluC rpmF fabD fabG acpP
##  [85] holB ycfH mfd  trmU minE minD minC ychF pth  prsA prfA ychA yciA yciC
##  [99] rnb  fabI tyrS rnt  ydhD sufA aroH pheT pheS rplT rpmI infC thrS sppA
## [113] gapA yeaZ yoaE htpX zwf  pykA yebC aspS argS sbcB hisG hisD hisC hisB
## [127] hisH hisA hisF hisI metG nfo  rplY gyrA nuoN nuoM nuoL nuoK nuoJ nuoI
## [141] nuoH nuoG nuoF nuoE nuoB nuoA ackA pta  truA fabB aroC yfcN gltX talA
## [155] tktB dapE dapA yfgM hisS yfgB fdx  hscA hscB iscU iscS suhB glyA tadA
## [169] acpS era  rnc  lepB lepA ung  rluD pheA rplS trmD rimM rpsP ffh  grpE
## [183] smpB csrA alaS eno  recD recB recC lysA lysS ygfZ zapA rpiA pgk  yqgF
## [197] yggW mutY yggX cca  rpsU rpoD yraL deaD pnp  rpsO truB rbfA infB nusA
## [211] argG secG greA rpmA rplU yrbA rpsI rplM tldD fis  aroE def  fmt  rplQ
## [225] rpoA rpsD rpsK rpsM rpmJ secY rplO rpmD rpsE rplR rplF rpsH rpsN rplE
## [239] rplX rplN rpsQ rpmC rplP rpsC rplV rpsS rplB rplW rplD rplC rpsJ fusA
## [253] rpsG rpsL yheL yheM yheN trpS rpe  aroB aroK bioH asd  rpoH ftsY yhhF
## [267] yhhP glyS glyQ rpmG rpmB dut  gyrB dnaN dnaA rpmH rnpA yidC gidA ilvD
## [281] ilvC rep  trxA rho  cyaA dapF metE polA pfkA tpiA fpr  hslV rpmE metF
## [295] argB argH secE nusG rplK rplA rplJ rplL rpoB rpoC pgi  dnaB ssb  efp
## [309] orn  miaA rpsF rpsR rplI ppa  valS dnaC rsmC
```

B Genes matched in both species

```
##  [1] thrA thrC talA dnaK dnaJ ileS dapB carA prsA sucA asnS ilvI ilvH ftsA
## [15] secA aceE aspS lepA rpoB rluC dnaA dnaE gyrA pheT metG minD tktB nuoH
## [29] nuoM hisS nuoB hisB ahpC rpsA gidA dapE nuoF argH aroA map  pfkA infC
## [43] yciA nuoA thrS aroH serC fabB metF lipB recC alaS leuD leuC leuB rho
## [57] mdlA polA fusA infB trpS lon  nusA rpiA pnp  rep  rpsK yggW ilvC deaD
## [71] lysS cyoB rpsD dnaB rpsH folD cspE lysA ybeY rplB ppiD iscS rplN mutY
## [85] asd  def  rpmA rpsC iscU rpsG rplD rpsI rplQ rpsN rplX rplV rpsS
```