POLITECNICO
MILANO 1863

PROJECT FOR THE GENOMICS COURSE OF THE DEGREE IN BIOINFORMATICS FOR COMPUTATIONAL GENOMICS

REFERENCE BASED ASSEMBLY OF BRASSICA RUPESTRIS

Professor: Aureliano Bombarely

Author: Francesco Gazzo

ABSTRACT

The rapid development of high throughput sequencing technologies has given the possibility to have easy access to large amount of genome sequencing data for many organisms.

It is here presented an easy and quick pipeline to make a reference based assembly of chloroplasts. The analysis has been done on *Brassica rupestris*, using two different references, *Brassica napus* and *Cakile arabica*, taking the data from NCBI databases.

The whole project has been done with different Bash programs, in particular with AbySS, a program that creates consensus sequences in order to construct scaffolds, starting with short reads.

Having used a k-mer of 31 and 63, and a high size of coverage (from 100X to 500X), the general output shows an average length of the scaffolds of 127Kbp. At the end, the best scaffold (127520 bp) has been plotted and compared with the *B. napus* reference (152860 bp), showing that the pipeline was not able to assemble the whole genome, missing one of the two inverted repeats, typical of the chloroplast DNA.

In conclusion, it is possible to assert that the developed pipeline is good enough for an initial understanding of the chloroplast genome structure and of the genes in it. A further step would be to implement a better optimization or to use long reads instead of short ones.

INTRODUCTION

In green plants, organelles contain chloroplast and mitochondrion genomes. Generally, chloroplast genome maintains a conserved circular and quadripartite structure, with a pair of invert repeat regions that separate large single copy (LSC) and small single copy (SSC) regions [1]. Since the rapid development of high throughput sequencing technologies, sequencing cost become cheaper and cheaper. Due to high number of copies of plastid genome in a single cell, it is easy to get enough reads to assemble a whole chloroplast genome from a low coverage of the whole genome sequencing data [2].

In this study, it has been used Bash programs such as Bowtie2 [3], samtools [4] and ABySS [5] to compute the reference based genome assembly of a chloroplast, Blast [6] and GeSeq [7] to compute comparisons and annotation.

MATERIAL AND METHODS

Illumina reads of *Brassica rupestris (SRR6453617)* were downloaded from the Sequence Reads Archive (SRA) thanks to fastq-dump (Version: 2.8.2) directly on the server. Briefly, the DNA of the SRA was isolated with Urea method and then libraries were made with Illumina TruSeq LT Genomic kit [8]; 2*150bp.

Chloroplast reference genomes were downloaded from the NCBI nucleotide database. They were *Brassica napus* NC_016734.1 and *Cakile arabica* NC_030775.1.

The whole coding project has been done on Annona server using Bash programs and it has been used Blast and GeSeq for, respectively, the sequences comparison and the assembly annotation.

The project is divided into the following steps:

1. Removing of the adapters from the Illumina reads using fastq-mfc (Version: 1.05) [9]
2. Indexing of the repository using bowtie2-build (Version: 2.3.4.1)
3. Mapping the Illumina reads, after having taken the same amount of reads, using bowtie2 and then samtools view (Version: 1.7) to transform the file from Sam to Bam format, in order to have a binary version, and to discard the reads unmapped
4. Sorting and transforming into fastq file with samtools sort and fastq
5. Calculating the coverage per nucleotide positions with bedtools genomecov (Version: v2.26.0) [9] and then plotting it with R
6. Sampling different sizes with seqtk sample [10] and assembling them with abyss-pe using different k-mers (Version: GNU Make 4.1)
7. Using Blast and GeSeq for visual comparison and assembly annotation

RESULTS

From the ~9600000 processed reads per file (R1 and R2), after having removed the adapters, it has been done the mapping on the two references, *B. napus (*same genus) and *C. arabica* (same family, different genus), resulting

in the 17% and 12% of the overall alignment rate, respectively, showing a quite good coverage with no gaps (Table 1 and Fig. 1), allowing a general good posterior assembly.

At this point it has been performed the assembly using 100X, 200X, 300X, 400X and 500X with different k-mer. In table 2 there are shown only the scaffolds made with k-mer 31 and 63, because of they are in general the better results. The average length of the scaffolds, are 130Kbp, and it has been selected the one generate from the 300X sample with a k-mer of 31 for the further analysis, because it has been assembled in a single sequence of 127,520 bp.

At the end, the best assembly has been compared with the *B. napus* reference using Blast and GeSeq (Fig 4 and 5). According to the reference selected, the assembly annotation shows the majority of the genes present in the reference (12 genes less on the 128 in total).

DISCUSSION

Fastq-mcf and bowtie2 worked properly for this project. The reduction of the starting reads are consistently and the coverage reached is good enough for an assembly.

The problem with the *B. napus* reference started with ABySS and its inability to reconstruct a perfect assembly of the chloroplast genome, probably due to the composition of it. In fact, it is divided into four parts: Long Single Copy (LSC) section, Short Single Copy (SSC) and two Inverted Repeats: IRa and IRb. According to the Blast and GeSeq plots (Fig 4 and 5), it seems that abyss was not able to reconstruct one of the two IR, resulting in ~25000 bp less than the reference and 12 genes less. Indeed, it is possible to see on the Dot Plot, that there are two sequences, separated by a gap, that represent the LSC and SSC, and it is present a diagonal sequence that represents one of the two IR, but it is missing the other one.

On the contrary, using the *C. arabica* reference showed a definitely lower coverage per position (Fig.2), resulting in a poor scaffold, where the best was composed of 6 different sequences.

In general it is possible to say that, because of the short length of the starting reads (150 bp on average), the best k-mers are from 31 to 63. Bigger ones create graphs with many separate disconnected sub graphs (i.e. there are many small groups of contigs that have no connections to the rest of the graph), and smaller ones result in very dense and tangles graph (i.e. all contigs are tied together in a single graph structure).

CONCLUSIONS

The described procedure showed an average good results, not being able to reconstruct a perfect assembly genome, but being able to give a first impression on how it should be composed and also which genes are in. A general better optimization of the used programs could give better results, indeed they could be reached using long reads instead of short reads for what concern the assembly genome of a chloroplast. It is also possible to say that the use of a reference of the same genus, helps in finding a better quality scaffold.

TABLES AND FIGURES

Table 1 Summary of the *B. rupestris* processing and mapping reads

Type of reads	Reference	Number of reads	Total bases (bp)
Raw R1		9.667.463	1.448.993.600
Raw R2		9.667.463	1.450.166.982
Processed R1		9.667.383 (-80)	1.448.992.417 (-1183)
Processed R2		9.667.132 (-331)	1.450.162.376 (-4606)
Mapped R1+R2	*B. napus*	2615525 (17.02%)	394492373
Mapped R1+R2	*C. arabica*	2308135 (12.02%)	348124738

Table 2 Summary of scaffolds of *B. rupestris*

Size of the sample	Reference	Number of sequences with k-mer 31	Total bases (bp) with k-mer 31	Number of sequences with k-mer 63	Total bases (bp) with k-mer 63
100K	*B. napus*	17	127844	21	128319
200K		10	127822	16	128113
300K		**1 ***	**127520**	8	133550
400K		4	130545	3	133320
500K		2	127471	3	134277
100K	*C. arabica*	20	127044	24	127367
200K		14	126945	23	159203
300K		13	127010	16	127426
400K		8	127078	11	159126
500K		5	126846	6	158927

* Best result

Fig 1 Coverage for the Illumina read mapping on *B. napus*

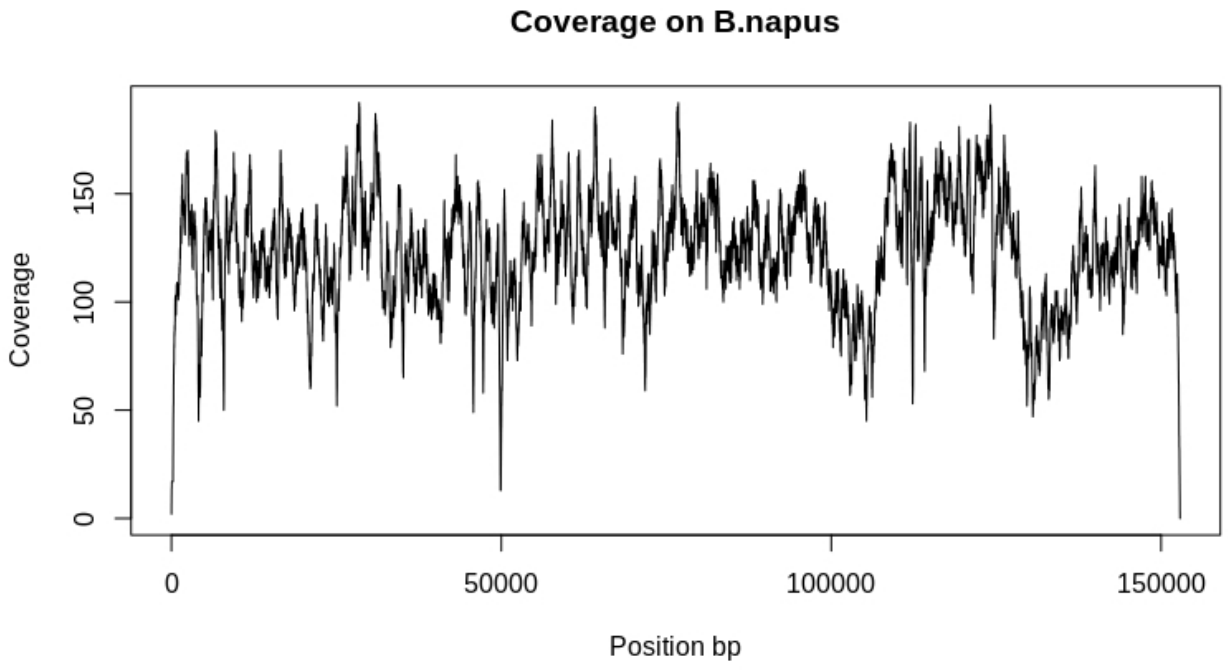


Fig 2 Coverage for the Illumina read mapping on *C. arabica*

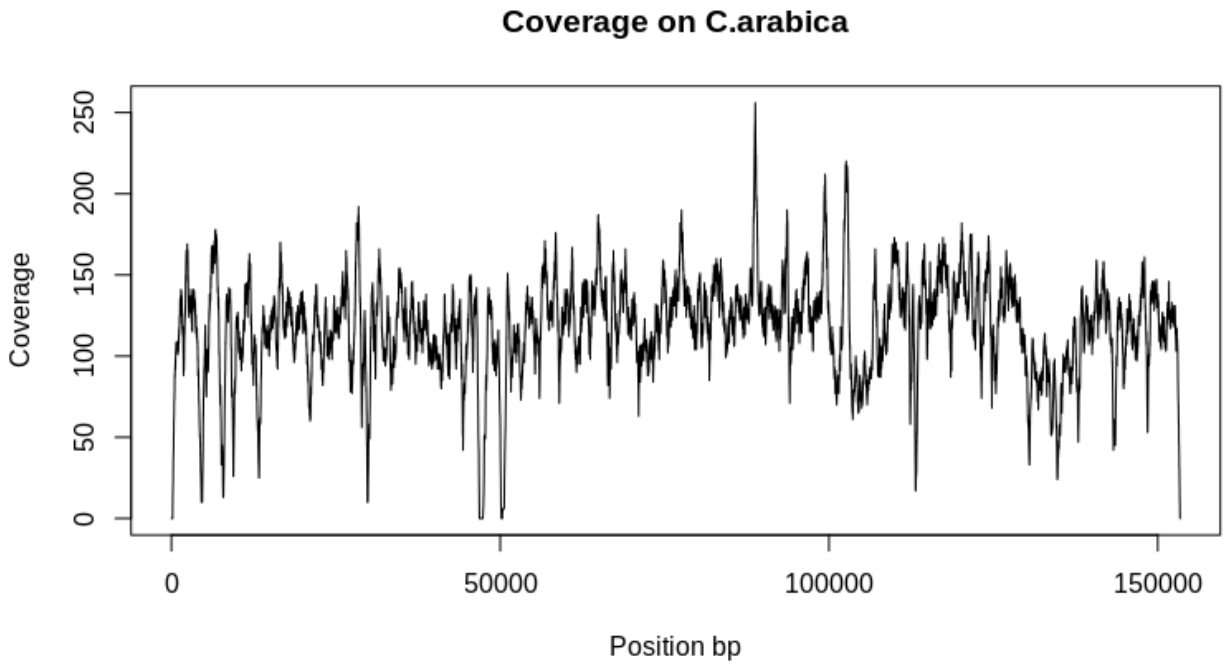


Fig.3 Dot Plot: *B. rupestris* assembly on x-axis and *B. napus* on y-axis

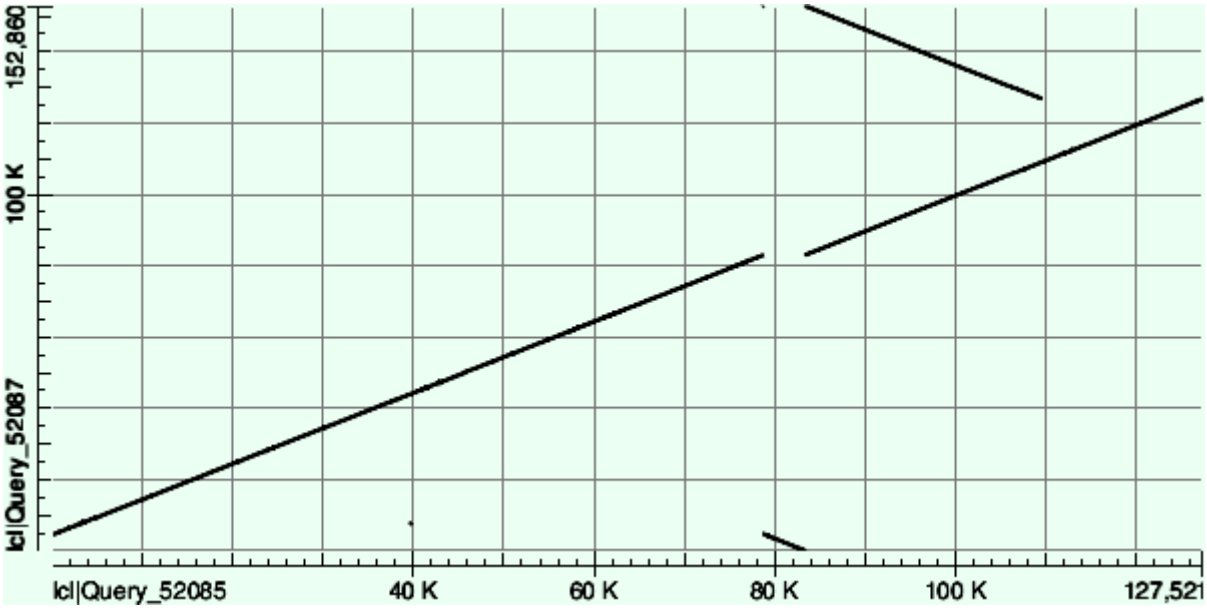
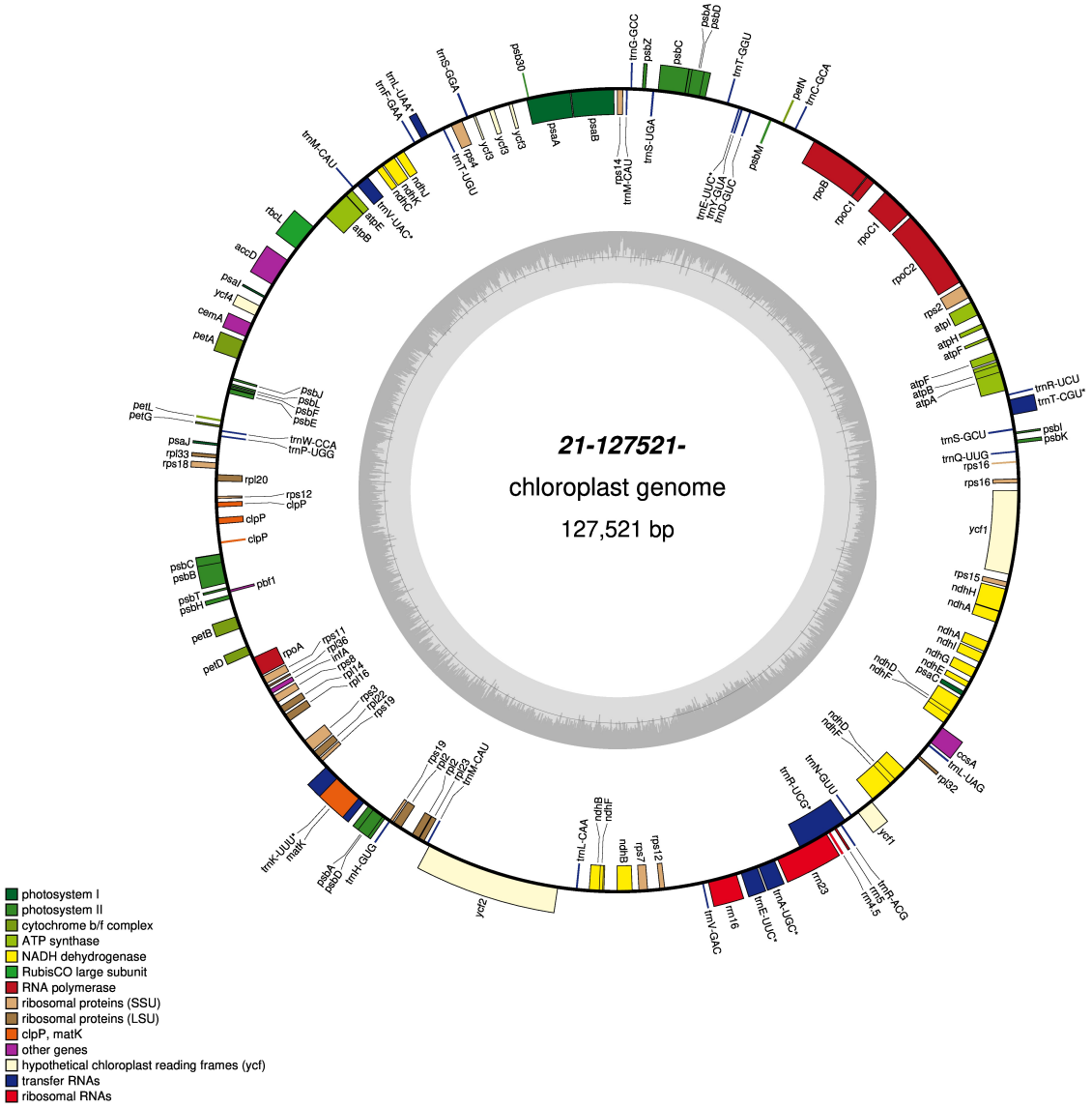


Fig 4 GeSeq assembly of *B. rupestris* assembly



AVAILABILITY OF DATA AND MATERIALS

- For *Brassica rupestris:* https://www.ncbi.nlm.nih.gov/sra/SRX3544562[accn]
- For *Brassica napus*: https://www.ncbi.nlm.nih.gov/nuccore/NC_016734.1
- For *Cakile arabica*: https://www.ncbi.nlm.nih.gov/nuccore/NC_030775.1

REFERENCES

1. Bock, R. and V. Knoop. 2012. Genomics of chloroplasts and mitochondria. Berlin: Springer https://scholar.google.com/scholar?q=Bock%2C+R.+and+V.+Knoop.+2012.+Genomics+of+chloroplasts+and+mitochondria.+Berlin%3A+Springer.
2. Twyford, A. D. and R. W. Ness. 2016. Strategies for complete plastid genome sequencing. https://onlinelibrary.wiley.com/doi/full/10.1111/1755-0998.12626
3. Bowtie2: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
4. Samtools: http://samtools.sourceforge.net/
5. Abyss: Shaun D Jackman, Benjamin P Vandervalk, Hamid Mohamadi, Justin Chu, Sarah Yeo, S Austin Hammond, Golnaz Jahesh, Hamza Khan, Lauren Coombe, René L Warren, and Inanc Birol (2017). ABySS 2.0: Resource-efficient assembly of large genomes using a Bloom filter. *Genome research*, 27(5), 768-777. doi:10.1101/gr.214346.116
6. Blast: https://blast.ncbi.nlm.nih.gov/Blast.cgi
7. GeSeq: Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R and Greiner S (2017) GeSeq – versatile and accurate annotation of organelle genomes. Nucleic Acids Research 45: W6-W11
8. Illumina TruSeq LT Genomic kit https://www.illumina.com/documents/products/datasheets/datasheet_truseq_dna_sample_prep_kits.pdf
9. Fastq-mcf: https://github.com/ExpressionAnalysis/ea-utils/blob/master/clipper/fastq-mcf.cpp
10. Bedtools: https://bedtools.readthedocs.io/en/latest/
11. Seqtk: https://github.com/lh3/seqtk