RESEARCH

Introduction to Focus Areas in Bioinformatics Project Week 6

Sina Glöckner[*], Christina Kirschbaum, Swetha Rose Maliyakal Sebastian and Gokul Thothathri

[*]Correspondence:
sina.gloeckner@fu-berlin.de
Institute for Informatics, Freie Universität Berlin, Takustr. 9, Berlin, DE
Full list of author information is available at the end of the article

Abstract

Goal of the project: To assemble the plastid genome of a water lily under two assembly routes, lists the number of generated contigs and visually compare the best contigs.

Main results of the project: The depth normalization improved the coverage and the assembly process for the genome.

Personal key learnings:
Sina: Implementing bash scripts
Christina: Overview NOVOPlasty assembly
Swetha: Visualization using MView, types of formats
Gokul: Basic bash scripts, multiple sequence alignment using MAFFT

Estimation of the time: Everyone invested about 6 hours for this week's project.

Project evaluation: 2

Number of words: 838

1 Scientific Background

Through next-generation sequencing, major advances have been made in genomic research, with the possibility to sequence a whole human genome within one day. Next-generation sequencing methods are sequencing millions of fragments in parallel and joining them by mapping these reads to a reference genome. During this process, every base is read several times to deliver accurate data through the depth. [1]

One particular field of interest in genomic research is the assembly of plastid genomes, which are used in phylogenetic studies or to identify food. An assembly algorithm that was developed specifically for organelle genomes is NOVOPlasty. [2]

NOVOPlasty uses the *de novo* assembly. For the assembly, k-mers are built and overlapping k-mers are searched. This guarantees an accurate assembly without a reference genome. However, the computational costs are higher than for a reference assembly. [3]

2 Goal

This project aimed to assemble the plastid genome of the water lily *Nymphaea odorata* subsp. *odorata* with and without applying sequence depth normalization. Afterwards, the number of contigs generated was to be listed and the best contigs of each method were visually compared.

3 Data

To assemble the plastid genome, we started with the raw reads of *Nymphaea odorata* subsp. *odorata* of the run SRR12134661, which was downloaded from NCBI SRA. NCBI SRA is a repository of raw sequence data that aims to balance the cost of long-term archival by storing sufficient information to support re-use of the submitted data [4].

The read set was reduced with samtools [5] for properly paired and fully mapped reads. That way, 4,536,052 reads remained. To further simplify the data, the first 304,984 reads were extracted.

As reference genomes, *Nymphaea odorata* with the accession number NC_057567 and *Nymphaea ampla* with the accession number NC_035680 were chosen. They were downloaded with Entrez Direct [6]. The sequence records were taken from NCBI RefSeq, a database where curated, stable, and non-redundant reference sequences are collected [7].

4 Methods

After this reduction, the reads were mapped against the reference genomes with Bowtie2 [8]. The reference genomes were formatted, then the first one was extracted and a small index was built. Afterward, the result file was sorted with samtools [5] and the coverage was computed using bamcov [9].

In the next step, a kmer-based sequencing depth normalization was performed with a read depth of 200, a read length of 151, and a best kmer size of 31. The normalization was applied with BBTools [10] and the new sequencing depth was plotted. The coverage was computed and compared to the coverage before the adjustment in a histogram using R [11].

Finally, the genome was assembled with NOVOPlasty [2] for the original files and the files with the adjusted depths. The first read of the depth normalized plastome reads was used as a seed sequence. The headers of the results were renamed to better differentiate between both runs. Then, the first contigs of both runs were aligned with MAFFT [12] and the alignment was visualized with MView [13].

5 Results

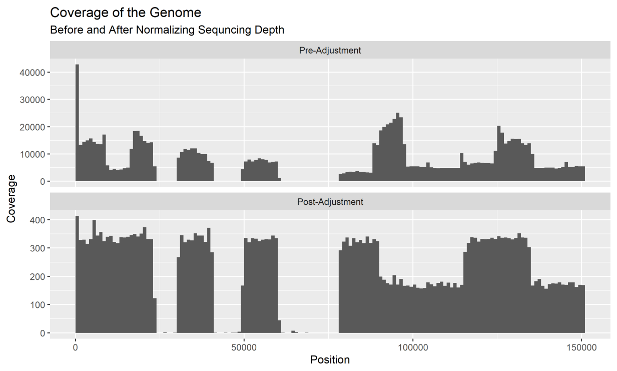


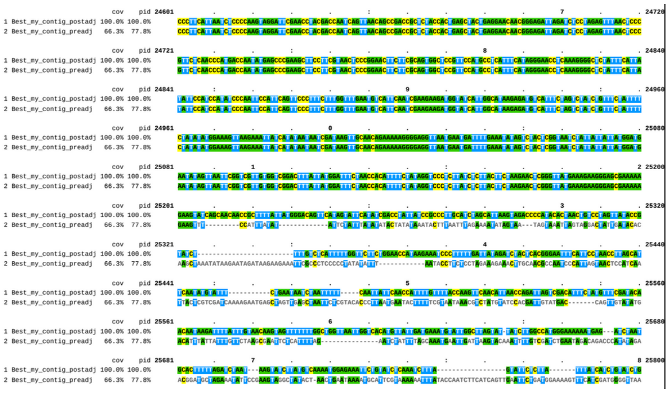
Figure 1: **Coverage plot of the genome.** The coverage of the genome before and after the depth normalization.

The coverage changed drastically after adjusting the depth. Figure 1 shows, that the span, as well as the overall values, are reduced after depth normalization.

This is also reflected in the results of the mapping process. Before adjusting the depth, there were 2,605,548 paired reads, out of which 87.67% were aligned. In comparison, 61,006 were paired with an alignment rate of 99.9%.

Additionally, the assembly results changed. Pre-Adjustment, five contigs were found with lengths ranging from 306 base pairs (bp) to 56,405 bp. Contrary, after adjusting the depth, one contig was created with a length of 81,877 bp.

The longest contigs were aligned. This can be seen in Figure 2.



```
                            cov    pid 24601          .         .         .         .         :         .         .         .         .         7         .         . 24720
1 Best_my_contig_postadj 100.0% 100.0%       CCCTTCATTAATCTCCCCAAGTAGGATTCGAACCTACGACCAATCAGTTAACAGCCGACCGCTCTACCACTGAGCTACTGAGGAACAACGGGAGATTAGATCTCCTAGAGTTTAACTCCC
2 Best_my_contig_preadj   66.3%  77.8%       CCCTTCATTAATCTCCCCAAGTAGGATTCGAACCTACGACCAATCAGTTAACAGCCGACCGCTCTACCACTGAGCTACTGAGGAACAACGGGAGATTAGATCTCCTAGAGTTTAACTCCC

                            cov    pid 24721          .         .         :         .         .         .         .         8         .         .         .         . 24840
1 Best_my_contig_postadj 100.0% 100.0%       GTTCTCAACCCATGACCAATATGAGCCCGAAGCTTCCTTCGTAACTCCCGGAACTTCTTCGCAGTGGCTCCGTTCCATGCCTCATTTCATAGGGAACCTCAAAGGGGCTCTATTTCATTA
2 Best_my_contig_preadj   66.3%  77.8%       GTTCTCAACCCATGACCAATATGAGCCCGAAGCTTCCTTCGTAACTCCCGGAACTTCTTCGCAGTGGCTCCGTTCCATGCCTCATTTCATAGGGAACCTCAAAGGGGCTCTATTTCATTA

                            cov    pid 24841          :         .         .         .         .         9         .         .         .         .         :         . 24960
1 Best_my_contig_postadj 100.0% 100.0%       TATTCCATCCATATCCCAATTCCATTCAGTTCCCTTTCTTTGGTTTGAATGTCATTCAATCGAAGAAGATGGTATCATTGGCATAAGAGATGTCATTTCTAGTCTATCTGTTTCTATTTT
2 Best_my_contig_preadj   66.3%  77.8%       TATTCCATCCATATCCCAATTCCATTCAGTTCCCTTTCTTTGGTTTGAATGTCATTCAATCGAAGAAGATGGTATCATTGGCATAAGAGATGTCATTTCTAGTCTATCTGTTTCTATTTT

                            cov    pid 24961          .         .         .         0         .         .         .         .         :         .         .         . 25080
1 Best_my_contig_postadj 100.0% 100.0%       CTATATATGGAAAGTTAAGAAATTATCATATAATAATCGATAAGTTGCAACAGAAAAAGGGGAGGTTAATGAATGATTTTGAAATATAGTCTACTCGGTAATCTATTATTATTATGGATG
2 Best_my_contig_preadj   66.3%  77.8%       CTATATATGGAAAGTTAAGAAATTATCATATAATAATCGATAAGTTGCAACAGAAAAAGGGGAGGTTAATGAATGATTTTGAAATATAGTCTACTCGGTAATCTATTATTATTATGGATG

                            cov    pid 25081          .         1         .         .         .         .         :         .         .         .         .         2 25200
1 Best_my_contig_postadj 100.0% 100.0%       AATATAGTTAATTCGGTCGTTGTGGTCGGACTTTATTATGGATTTCTAACCACATTTTCTATAGGTCCCTCTTATCTCTTACTTCTAAGAACTCGGGTTATGAAAGAAGGGAGCGAAAAA
2 Best_my_contig_preadj   66.3%  77.8%       AATATAGTTAATTCGGTCGTTGTGGTCGGACTTTATTATGGATTTCTAACCACATTTTCTATAGGTCCCTCTTATCTCTTACTTCTAAGAACTCGGGTTATGAAAGAAGGGAGCGAAAAA

                            cov    pid 25201          .         .         .         .         :         .         .         .         .         3         .         . 25320
1 Best_my_contig_postadj 100.0% 100.0%       GAAGTATCAGCAACAACCGCTTTTATTATGGGACAGTTCATAGTATTCATATCGACCTATTATCCGCCCTTGCATCTAGCATTAAGTAGACCCCATACACTAACTGTCCTAGTTATACCG
2 Best_my_contig_preadj   66.3%  77.8%       GAAGTTT---------CCATTTATAT-------------ATTCTATTTATATATACTATATAAATACTTTAATTTAGAAAATATAGTAA---TAGTAAATTAGTAGGACTATTGATACAC

                            cov    pid 25321          .         .         :         .         .         .         .         4         .         .         .         . 25440
1 Best_my_contig_postadj 100.0% 100.0%       TATCT-------------------------TTTGTCTCATTTTTGGTTCTTCTGGAACCATAAGAAATCCCTTTTTGATTATAGATCTACTCACGGGAATTTCATTCCTAACCTTAGCAT
2 Best_my_contig_preadj   66.3%  77.8%       AAGCTAAATATAAGAATAGATAAGAAGAAATTCGCCCTCCCCCTATATATTT------------AATACCTTCTCCTAGAAAGAAACTTGCAACGCCAATCCCATTAGTAACTCCATCAA

                            cov    pid 25441          :         .         .         .         .         5         .         .         .         .         :         . 25560
1 Best_my_contig_postadj 100.0% 100.0%       TCAATATGTATTT-----------CTGAATAATCTAATTTTT-----CAATTATTCAACCATTTTGTTTTACCAAGTTCAACATTAACCAGATTAGTCGACATTTCTATGTTTCGATACA
2 Best_my_contig_preadj   66.3%  77.8%       TTACTCGTCGATCAAAAGAATGAGCTAGTTGAGCTAATTCTCGTACACCCTTAATGAATACTTTTTCGTAATAAACGTCTATGTATCCACGATTGTATGAC-------CAGTTGTATATG

                            cov    pid 25561          .         .         .         6         .         .         .         .         :         .         .         . 25680
1 Best_my_contig_postadj 100.0% 100.0%       ACAATAAGATTTTATTTGTAACAAGTAGTTTTTTTGGCTGGTTAATTGGTCACATGTTATTGATGAAATGTATTGGCTTAGTAT-TATCTTGGCCATGGGAAAAAATGAG---ATCTAAT
2 Best_my_contig_preadj   66.3%  77.8%       ACATTTATTATTTGTTCTAAGCGAATTCTCATTTTAG---------------AATCTATTTTAGCAAATGAATTGATTAAGTACAAATTTTGTCGATCTGAATAGACAGACCCATATAGA

                            cov    pid 25681          .         7         .         .         .         .         :         .         .         .         .         8 25800
1 Best_my_contig_postadj 100.0% 100.0%       GCACTTTTTAGATCTAAT---AAGTATCTTATGTCAAAATGGAGAAATTCTGTATCTCAAATCTTTA------------------GTATTCTCTTA-------TTTATCATCTGTATCTG
2 Best_my_contig_preadj   66.3%  77.8%       ACGGATGCTAGAAATATTCCGAAGTAGGCTATACT-AACTGAATAAAATGCATTCGTAAAAAATTTATACCAATCTTCATCAGTTGAATTCTGATGGAAAAGTTTCATCGATGGGGTTAA
```

Figure 2: **Screenshot of the alignment.** The screenshot shows a part of the resulting alignment generated with MView.

6 Discussion

This week's project focused on data processing and data transformation. The results showed a remarkable difference between the data before and after adjustment. The coverage, as well as the assembly results, were significantly improved after depth normalization.

Additionally, the handling of a big amount of data was included, which is also an everyday occurrence for a data scientist.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

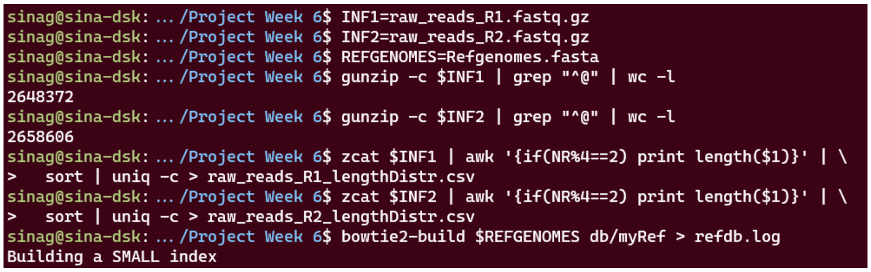
Sina Glöckner implemented the plastid genome assembly with and without sequence depth normalization. Gokul Thothathri aligned the results with MAFFT. Swetha Rose Maliyakal Sebastian visualized the alignment with MView. Christina Kirschbaum wrote the report.

References

1. Behjati S, Tarpey P. What is next generation sequencing? Archives of disease in childhood Education and practice edition. 2013 08;98.
2. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: De novo assembly of organelle genomes from whole genome data. Nucleic Acids Research. 2016 10;doi: 10.1093/nar/gkw955.
3. Moreton J, Izquierdo Barraza A, Emes R. Assembly, Assessment, and Availability of De novo Generated Eukaryotic Transcriptomes. Frontiers in Genetics. 2016 01;6.
4. Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. Nucleic Acids Res. 2011 Jan;39(Database issue):D19–D21. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3013647/.
5. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 06;25(16):2078–2079.

6. Baxevanis A. Searching the NCBI databases using Entrez. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]. 2006 04;Chapter 1:Unit 1.3.
7. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016 Jan;44(D1):D733–745.
8. Langmead B, Salzberg S. Langmead B, Salzberg SL.. Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357-359. Nature methods. 2012 03;9:357–9.
9. bamcov/bamcov.c at master · fbreitwieser/bamcov;. Available from: https://github.com/fbreitwieser/bamcov.
10. BBTools: Multithreaded bioinformatics tools for analysis of DNA and RNA sequence data;. Available from: https://jgi.doe.gov/data-and-tools/bbtools/.
11. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the tidyverse. Journal of Open Source Software. 2019;4(43):1686.
12. Katoh K, Misawa K, Miyata T. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research. 2002 08;30:3059–66.
13. Brown N, Leroy C, Sander C. MView: A web-compatible database search or multiple alignment viewer. Bioinformatics (Oxford, England). 1998 02;14:380–1.

Figures



```
sinag@sina-dsk: .../Project Week 6$ INF1=raw_reads_R1.fastq.gz
sinag@sina-dsk: .../Project Week 6$ INF2=raw_reads_R2.fastq.gz
sinag@sina-dsk: .../Project Week 6$ REFGENOMES=Refgenomes.fasta
sinag@sina-dsk: .../Project Week 6$ gunzip -c $INF1 | grep "^@" | wc -l
2648372
sinag@sina-dsk: .../Project Week 6$ gunzip -c $INF2 | grep "^@" | wc -l
2658606
sinag@sina-dsk: .../Project Week 6$ zcat $INF1 | awk '{if(NR%4==2) print length($1)}' | \
>   sort | uniq -c > raw_reads_R1_lengthDistr.csv
sinag@sina-dsk: .../Project Week 6$ zcat $INF2 | awk '{if(NR%4==2) print length($1)}' | \
>   sort | uniq -c > raw_reads_R2_lengthDistr.csv
sinag@sina-dsk: .../Project Week 6$ bowtie2-build $REFGENOMES db/myRef > refdb.log
Building a SMALL index
```

Figure 3: **Screenshot Terminal.** The screenshot shows the very first steps of the computations.