**Identifying the optimal *de novo* genome assembly technique that allows identification of genetic variations between different strains of the halophilic archaeon *Haloferax volcanii***

Name: Matthew Gaskins

Names of other project members: Saroj Sharma, Lewis Wood

**Abbreviations:**

BLAST: Basic Local Alignment Search Tool

CENP-T: Centromeric Protein-T

dsDNA: double-stranded DNA

EBI: European Bioinformatics Institute

GC: Guanine-Cytosine

HAN: Hef-associated Nuclease

KMN network: KNL1–MIS12–NDC80 network

MCM complex: MiniChromosome Maintenance protein complex

NCBI: National Center for Biotechnology Information

ORI: Origin of Replication

QC: Quality Control

QUAST: Quality Assessment Tool for Genome Assemblies

RPKM: Reads Per Kilobase of transcript, per Million mapped reads

Sample 1: S1

Sample 5: S5

Sample 6: S6

Sample 7: S7

Sample 8: S8

SMTL: Swiss Model Template Laboratory

UoN: University of Nottingham

**Project Role:**

My first role involved installing CANU into Conda, which proved unsuccessful, therefore I subsequently installed Flye. In combination with Lewis, we produced scripts constructing long-read assemblies for each of our five samples, firstly including both pass and fail-data, then exclusively using pass. I next identified samples as *Haloferax volcanii* using Nucleotide-BLAST, and determined S1 was not significantly different from the *H.volcanii* reference. I employed Nucleotide-BLAST again, comparing S1 with mutant strains and importing alignment hit-tables into Excel to detect differences between genomes.

**Aims and Experimental Question:**

Long-read *de novo* genome assemblers address shortcomings in terms of the difficulty assembling short-reads into large, complex genomes [1]. However, the accuracy and throughput of long-read assemblers remains low relative to short-read alternatives [1]. This report aims to evaluate the benefits and disadvantages of each assembly method, to provide insight into the estimated probability and speed of transition from short- to long-read technologies as the consensus method for genome construction. This report also assesses whether hybrid assemblies, which polish more contiguous long-read assemblies with more-accurate short-reads, could eventually represent the conventional technique for genome assembly in the future.

We aim to achieve our objective of identifying an optimal assembly method by constructing genomes of five sample strains from an unknown microorganism, ultimately characterised as halophilic archaeon *Haloferax volcanii*. S1 presents growth rates comparable to wild-type *Haloferax*, however our four remaining samples - termed S5, S6, S7 and S8 - exhibit reduced development. S7 also displays increased sensitivity to temperature changes. This report aims to assemble genomes for each of these five samples using both a short- and long-read approach, and in some cases a hybrid technique. The quality of assembly produced using each method can then be evaluated to determine our favoured approach for future assemblies.

However, the objective to determine a global, optimised assembly method is limited, as all samples are from one archaeal species. Genome composition is highly disparate between the three domains, and even at the level of species, extreme variability exists regarding genome-size, guanine-cytosine content, and the frequency and length of homopolymeric tracts - these inconsistencies differentially influence the quality of various assembly types [2]. Our preferred method is therefore most suitable for archaea, and specifically halophiles, which are likely to contain comparable homopolymer and guanine-cytosine content to *H.volcanii*.

Sample genomes constructed using our selected technique are then compared against our wild-type S1 control, to identify genomic differences between each of our five samples. Our second aim is therefore to interpret the effect of genomic differences between strains of *Haloferax volcanii* on archaeal dysfunctionality. Functionally characterising these variants provides further insight into thermotolerance and growth regulation within *Haloferax*, a widely-used model organism for studying DNA replication. Whilst our experimental approach may successfully achieve this second objective, the extent to which our optimal assembly method should be extrapolated across taxa requires consideration.

**Background and Motivation:**

Since the expansion of second-generation DNA sequencing techniques in the early 2000s, comparative genomics has depended on genome assembly to rationalise such biological complexity. In 2006, Illumina developed the Solexa machine, capable of sequencing over 1Gb of data in a single-run by identifying individual nucleotides as complementary dye-terminators pass over sample DNA [3]. Combined with the reduced cost associated with Solexa, this enhanced throughput proved paradigmatic to DNA-sequencing, and prompted development of an assembly pipeline to rationalise these short-reads into completed genomes [3].

Text

Description automatically generated with medium confidenceIn 2017, Wick *et al*. developed Unicycler, which assembles short-reads into completed genomes - Figure 1 displays the stages of this pipeline [4]. The vertices of the de Bruijn assembly graph comprise linked overlaps from short-read *k*-mers produced using SPAdes [5]. Assembled single-copy contigs are then bridged into scaffolds, the quality of which can be improved through long-read supplementation in hybrid assemblies.

The 2015 release of a novel long-read sequencing technique by Oxford Nanopore therefore transformed genome assembly quality. Nanopore-sequencing threads input DNA through a narrow porin, with the change in current quantified to identify the nucleotide traversing the nanopore channel. If mean base quality from each long-read exceeds a predefined threshold, the read is classified as pass-data - otherwise the read is categorised as fail [7].

Timeline

Description automatically generated with medium confidenceThe increased contiguity from this technique provides obvious advantages in terms of assembling larger, complex genomes, prompting development of various long-read assembly pipelines. Figure 2 displays the pipeline for long-read assembler Flye, which concatenates disjoint genomic segments of read-data and constructs them into an assembly graph [8]. Accurate contigs are then generated through bridged and unbridged repeat resolution.

Assembly quality can be evaluated using various software, including QUAST. This program details several metrics describing each assembly, including the number of contigs, the length of the largest contig, and an N50 value reporting the length of the shortest contig for which longer and equal-length contigs constitute over half of the assembly [9]. Moreover, when a reference genome is provided, QUAST evaluates several additional metrics, which quantify similarity between reference and assembly. QUAST metrics therefore disclose assembly quality, in terms of both accuracy and contiguity [9].

Similarity between two assemblies can be identified using Nucleotide-BLAST, which detects homology using a heuristic version of the Smith-Waterman algorithm [10,11]. BLAST attempts local alignment from initial matches, with homologous segment-pairs beneath a user-defined expect-value recorded to display alignment between input sequences [10]. Expect-values state the probability alignment arose due to chance, therefore lower values indicate improved homology. Aligned segment-pairs can be ordered from start-to-end to identify genomic variation between sequences.

**Methods:**

**Sequencing:**

Donated Oxford Nanopore long-read data for each of our five samples were produced using the FLO-MIN106 and LSK109 kit combination. As all data were provided by an external source, the threshold for mean base-quality to be categorised as passed remains unknown. Donated Illumina short-read data were produced in paired-runs using the four-lane NextSeq500/550 flowcell. For each of the five samples, data from each of the four lanes were concatenated into one FASTQ file combining sequence from all lanes.

**Assembly:**

To assemble these five concatenated short-read files, a new Conda(version4.11.0) environment was created on our Azure Cloud Compute server, with Unicycler(version0.4.8) installed [4,12]. Once this new Conda environment had been activated, all five short-read Unicycler assemblies were then constructed individually under default parameters. Unicycler automatically conducts Pilon-polishing(version1.24) to improve the quality of our draft assemblies, which can be observed separately from unpolished short-read assembles [4]. We also employed Unicycler to construct hybrid assemblies for S1 and S5, which were consequently also polished using Pilon. For both samples, short-reads were supplemented solely with pass-only long-reads.

To assemble long-read data for our five samples, another Conda environment was created, this time with Flye(version2.9-b1774) installed [8,12]. Following activation of this new Conda environment, long-read assemblies for each of our five samples were constructed using Flye under default parameters. Flye was firstly run using only passed reads, before separately including both pass and fail-data to produce ten long-read assemblies in total. We then installed Medaka(version1.5.0) into another Conda environment, before exclusively polishing pass-only long-read assemblies [12,13].

**Species Identification:**

We then identified our sample species by inputting two random assemblies from each assembly method into Nucleotide-BLAST [10]. Our search was optimised for identifying highly similar sequence with megablast, and used default parameters barring an adjusted expect-value threshold of 1. We identified the microorganism with highest percentage identity to our samples - *Haloferax volcanii* (strain DS2).

**Quality Control:**

The quality of each of our thirty-four total assemblies was assessed against the concatenated reference NCBI *Haloferax volcanii* genome (EBI Accession-ID: GCA\_000025685) using QUAST(version5.1) [9,14]. Our preferred assembly method was selected favouring several characteristics, including but not limited to rare misassemblies, high N50/N75 values, and similarity to the reference genome, in terms of length, guanine-cytosine content and the percentage of bases aligned with the reference.

**Evaluating the suitability of S1 as a control for *H.volcanii*:**

To ensure S1 was suitable as a control for the *H.volcanii* genome, we used Nucleotide-BLAST to compare this sample to the reference under identical parameters as above [10]. Contig headers were removed from each sample to concatenate assemblies into uninterrupted nucleotide sequence. Provided no variation existed in regions of the genome associated with growth or thermotolerance, and a 99% identity threshold was exceeded for the remainder of the two genomes, we would accept S1 as a control sample, which proved accurate.

**Identifying genomic differences between strains:**

Genomic differences between S1 and the reference, and subsequently S1 and our four mutant strains, were identified by importing Nucleotide-BLAST-generated alignment hit-tables into Excel(version2201), ordering alignments from start-to-end of the relevant reference, and identifying deletions surpassing 50 base-pairs in length. Nucleotide sequence for genomic differences between S1 and mutant strains was extracted from S1 using SAMtools(version1.9) [15].

Extracted sequence was then input to Nucleotide-BLAST; homologous sequences with >95% identity to *H.volcanii* were characterised as functionally equivalent to DS2, with the identity of proteins encoded by these sequences recorded. If the protein was uncharacterised, a homology model was constructed with SwissModel [16]. Alternatively, if deletions are located within intergenic regions, characterised genes in the 1,000 nucleotides either side of the identified sequence are recorded.

**Code:**

All scripts and output assemblies can be found inside our GitHub repository [<https://github.com/stymwg/Project1>). Provided users can access the 10.102.161.8 Cloud Compute, all assemblies can also be located in the directory shared/Project1\_Resources/GROUP2\_Workspace/, alongside our QC reports.

A master script for each section of the project is located within our GitHub repository under filename master\_script. Read alongside the masterscript\_readme file to run code.

**Results and Discussion:**

To achieve our aims of identifying the optimum method of assembling genomes using provided sequence data, we selected Unicycler to assemble short-reads, as this program employs SPAdes [4,5]. The accuracy of short-read data promotes its use in polishing more-contiguous assemblies constructed from less-accurate long-reads; low mismatch-substitution error-rates are therefore particularly favourable for our short-read assembly algorithm. SPAdes exhibits the lowest mismatch error-rate within correctly-assembled contigs compared to other popular short-read assembly software, including Velvet and SOAPdenovo [17]. SPAdes also constructs highly contiguous genomes, increasing contig lengths without misassembly relative to alternative algorithms [17]. Moreover, integrated Pilon-polishing enhances the contiguity of Unicycler draft assemblies, further supporting use of this short-read assembler.

We initially selected CANU as our long-read assembler due to its well-characterised contiguity and accuracy [18]. However, we were unable to install CANU into a new Conda environment, therefore we selected Flye to assemble long-read data. This program produces fewer misassemblies than alternative long-read assemblers including CANU, although the contiguity of its assemblies is comparatively poor [19]. We first assembled passed long-read data into completed genomes, but then included additional fail-data to produce a second set of assemblies, as this increased the amount of input sequence available for assembly. The mean base quality of fail-only data is too low to construct reliable genomes, therefore fail-only assemblies are redundant to our aims, and we did not complete any such assembly.

Table

Description automatically generatedAs displayed in Table-1, including fail-data alongside passed-reads can reduce disparity between the length of our assembly and the reference genome, albeit inconsistently, without increasing misassembly frequencies. However, incorporating fail-data also increases the frequency of both mismatches and indels, therefore we decided to exclusively polish pass-only long-read assemblies with Medaka, due to their improved accuracy. We selected Medaka as this program reduces local mismatch frequency compared to alternative polishing software such as Racon [20]. Indeed, Medaka reduces mismatch frequency more significantly than Pilon.

Due to the proficiency of SPAdes, we also selected Unicycler to build both Illumina and pass-only Nanopore data into hybrid assemblies [4,5]. As previously described, including fail-data reduces output accuracy, therefore these data were excluded from hybrid assemblies. Whilst Trycycler provides more accuracy than Unicycler, this software requires more time to run, and necessitates multiple command-line entries during assembly [21]. Indeed, even using faster Unicycler software, time-constraints only allowed hybrid assembly of S1 and S5. We also employed QUAST to evaluate assembly quality due to its speed; this program parallelises its most time-consuming stages to rapidly generate metrics that evaluate assembly quality [9]. A reference *H.volcanii* genome was included in our QUAST command to increase the number of metrics available to compare assemblies [14].

Table

Description automatically generatedThe QUAST report displayed in Table-2 justifies application of polishing, but only for specific methods of assembly. Interestingly, polishing our hybrid assembly decreases accuracy. From the improved post-polishing alignment with the reference, it is possible to infer the corresponding increase in coverage reveals the true level of read-duplication within polymorphic regions, which may introduce mismatches and indels that reduce accuracy. This proves less problematic for short-read assemblies; false duplications are less likely to persist within shorter and more numerous contigs. Polished long-read assemblies also display increased relative mismatch frequency, however this reduction in accuracy is outweighed by a more significant decrease in the frequency of indels. However, improved sample-reference alignment prompted our decision to employ polished long-read assemblies to compare assembly methods and identify our model approach.

Table

Description automatically generatedTable-3 presents a comparison between polished short-read, polished long-read and unpolished hybrid assemblies - the most accurate version of each assembly type - which we used to identify our model approach. To improve comprehensibility, Table-3 only displays quality metrics for S1, however similar patterns regarding contiguity and accuracy were observable across all five samples.

As anticipated, polished short-read assemblies contain significantly fewer mismatches and indels than long-read and hybrid equivalents. This indicates short-read assemblies are our most accurate technique, although local misassemblies are less frequent in long-read genomes. However, accuracy is not our only favourable criterium - assembly contiguity is essential for identifying significant structural and functional elements of the genome. As short-read assemblies contain significantly more contigs than hybrid or long-read genomes, their N50 and N75 values are correspondingly lower than these alternative techniques. This inferior contiguity therefore opposes selecting short-reads for our preferred method of assembly.

Shape

Description automatically generatedThe advantage of improved assembly contiguity is perhaps best-exhibited by visualising genomes using Bandage [22]. As displayed in the reference assembly of Figure 3a, the *Haloferax volcanii* genome constitutes five components: one main 2.848Mb chromosome, three secondary pHV4, phV3 and phV1 chromosomes (listed from longest-to-shortest), and one pHV2-plasmid [23]. However, as shown in Figures 3b-d, the maximum number of components present in experimental S1 assemblies is three. Laboratory-cultured *H.volcanii* has been characterised to integrate pHV4 onto the main chromosome, and pHV2 is absent, therefore using this observation we infer *H.volcanii* samples have been cultured in the laboratory [24]. The existence of only two constituent genomic components within our hybrid assembly therefore represents a significant error - pHV3 has been inappropriately amalgamated with the main chromosome and pHV4. Bandage also highlights how the incontiguous nature of short-read assembly is insufficient to resolve the shape of archaeal replicons. Only long-read assembly accurately determines the number and orientation of sample replicons, which contain one main and two secondary circular chromosomes - a significant advantage of this methodology.

Another beneficial characteristic of long-read assemblies is their relative similarity to the reference compared to alternative techniques, in terms of size, GC content and alignment. This demonstrates long-read assemblies accurately approximate well-characterised genomes from the same species. The nature of Nanopore sequencing itself represents another advantage; sequencing our five samples in a single Nanopore run is approximately half as expensive as Illumina, whilst Nanopore also allows real-time data analysis that enables termination of erroneous runs [25]. We therefore select polished long-read assemblies as our preferred method for this investigation, due to their cost-effectiveness, strong alignment to the reference, and superior contiguity.

Table

Description automatically generatedUsing Nucleotide-BLAST, we identified our sample species as *Haloferax volcanii*; this organism contained over 99.9% identity with assemblies from all five strains [10,14]. Nucleotide-BLAST was run under default parameters, except the expect-threshold was increased to 1 to allow more alignments to be identified. Nucleotide-BLAST searches were optimized for megablast, which aligns highly similar sequences such as our assemblies more quickly than alternative programs. We then attempted to ascertain whether S1 was a suitable control for *H.volcanii* - due to its wild-type phenotype, we hypothesised S1 was genetically similar to the reference, and therefore compared this polished long-read assembly to the NCBI genome [14]. Table-4 displays major differences between the two genomes, none of which exist in regions likely to alter archaeon growth or thermotolerance. Indeed, the only two notable S1-deletions are located in genes associated with a well-characterised knockout system used to introduce deletions into *H.volcanii* [26]. A plasmid carrying the *pyrE2* marker-gene, *trpA* and flanking sequence of the desired deletions is integrated into the chromosome, enabling *H.volcanii* to synthesise both uracil and tryptophan; *pyrE2* and *trpA* encode key uracil and tryptophan biosynthesis genes respectively. Subsequent loss of the plasmid through intrachromosomal cross-over introduces desired deletions, as genes are lost alongside artificially-introduced flanking sequences [26]. This indicates deletions have been introduced artificially, and also suggests data were provided by UoN researcher Thorsten Allers, who developed the *pyrE2*/*trpA* knock-out system [26].

Table

Description automatically generatedWe therefore concluded S1 represented a suitable control for *H.volcanii*; our four mutant genomes were subsequently compared to this sample, again using Nucleotide-BLAST [10]. The alignment hit-table between S1 and mutant strains was ordered from start-to-end of S1 to identify deletions in sample assemblies (>50base-pairs). Deleted sequence from mutant strains was extracted from S1 and compared against the Nucleotide-BLAST database to identify homologous sequence in the *H.volcanii* reference. Table-5 outlines all deletions observed in mutant strains relative to S1, and their wild-type function.

Four deletions are consistent across all mutant strains - Δ*orc5*, Δ*orc2,* Δ*A0638,* and ΔORI1. Three of these deletions are associated with initiating DNA replication, partially explaining reduced growth-rates in mutant samples. The main *H.volcanii* chromosome features three origins of replication, each requiring a corresponding Orc-initiator protein - ORI1 is associated with Orc1, ORI2 with Orc5 and ORI3 with Orc2 [23]. All four mutant strains have truncated or absent versions of either one replication origin, or an associated initiator protein, assuming deletions proximal to *orc2* and *orc5* impact their biological activity. The existence of a *pyrE2*/*trpA* gene-knockout system indicates deletions are artificially engineered - considered alongside a consistent deletion-size of approximately 1,000 base-pairs, this assumption appears valid, as multiple circumstantial deletions are unlikely to have been introduced accidentally.

However, in 2013 Allers *et al.* demonstrated in absence of a functionally active origin of replication on the main chromosome of *H.volcanii*, homologous recombination is sufficient to reconstitute and indeed expedite DNA replication, to such an extent that archaeal growth accelerates rather than declines [27]. Disruption to these three replication origins is therefore insufficient to explain the reduced growth-rate of our four samples. Correspondingly, S6, S7 and S8 also contain deletions within genes encoding RecJ homologues, nucleases that repair dsDNA breaks during homologous recombination [28]. For S6 and S8, the gene encoding RecJ1-exonuclease is deleted. RecJ1 associates with GINS51, therefore MCM-GINS interactions are altered in its absence. As MCM-helicase activity declines, homologous recombination stalls, reducing archaeal growth [28]. For S7, the gene encoding 3’-exonuclease HAN has instead been deleted. This protein associates with stalled-fork repair protein Hef, which resolves homologous recombination intermediates; deletion of this protein restricts archaeal growth in absence of replicative origins within the main chromosome [28]. HAN deficiency also explains increased S7 thermosensitivity; replication forks that stall due to cellular-stress responses caused by temperature change are not repaired and eventually collapse, which can reduce genome integrity to a deleterious extent [29].

Nevertheless, the decreased growth-rate of S5 remains unresolved. An explanation is not straightforward, however one answer lies with the Δ*A0638* mutation, present in all four samples. Whilst *A0638*-encoded protein is uncharacterised, homology modelling from its primary sequence identifies 26.67% similarity to two distinctive C-terminal alpha-helices of the CENP-T protein (SMTL ID:3vh5.1.C). In eukaryotes, CENP-T anchors the KMN network to kinetochores in replicating cells [30]. Subsequent kinetochore-microtubule attachment allows chromosomal segregation, and the cell-cycle to proceed beyond S-phase [30] Inactivation of a CENP-T homologue in *Haloferax* may equivalently stall archaeal cell-cycles, conferring reduced growth as presented in our four samples.

The objective to identify genomic differences that provoke varying phenotypes within our five samples has therefore been achieved. We also successfully determined polished long-read assemblies represent the optimal method of constructing genomes from halophilic archaea. However, the breadth of our conclusions could be widened across taxa by analysing additional samples from alternative archaea, and even from different domains. Moreover, assembling each sample genome was unnecessary to identifying variation between strains. Genomic differences could be more efficiently detected through immediately mapping our reads to an NCBI reference, with misalignments characterised to determine their functional consequences.

Lastly, differential gene-expression analysis through RNA-Seq represents a more precise approach to rationalising phenotypic variation, which can arise from gene expression changes instead of nucleotide content. Therefore, whilst our approach sufficiently achieves our two aims of the project, multiple improvements could be introduced to increase both the importance and breadth of its conclusions.

**References:**

1. Amarasinghe, S., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome biology,* **21(1):**1-16.

2. Goldstein, S., Beka, L., Graf, J. & Klassen, J. L. (2019). Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. *BMC genomics,* **20(1):** 1-17.

3. Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., & Law, M. (2012). Comparison of next-generation sequencing systems. *Journal of biomedicine & biotechnology,* **2012:** 251364.

4. Wick, R., Judd, L., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS computational biology,* **13(6):**1005595.

5. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S. & Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology,* **19(5):** 455–477.

6. Simpson, J. (Aug 2016). *De novo* assembly. Ontario Institute for Cancer Research.

7. Jain, M., Olsen, H.E., Paten, B. *et al.* (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology,* **17:**239.

8. Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature biotechnology,* **37(5):** 540-546.

9. Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics,* **29(8):** 1072–1075.

10. McGinnis, S., & Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic acids research,* **32:** W20–W25.

11. Okada, D., Ino, F., & Hagihara, K. (2015). Accelerating the Smith-Waterman algorithm with interpair pruning and band optimization for the all-pairs comparison of base sequences. *BMC bioinformatics,* **16:** 321.

12. Anaconda Software Distribution, Conda, version 4.1.1.0. (Austin, TX: Anaconda, 2016), accessed Jan 28, 2022, <https://www.anaconda.com>

13. medaka: sequence correction provided by ONT Research. <https://github.com/nanoporetech/medaka>. Accessed 22 Feb 2022.

14. Hartman, A. L., Norais, C., Badger, J. H., *et al.* (2010). The complete genome sequence of *Haloferax volcanii* DS2, a model archaeon. *PloS one,* **5(3):** 9605.

15. Li, H., Handsaker, B., & 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* **25(16):** 2078–2079.

16. Kiefer, F., Arnold, K., Künzli, M., Bordoli, L., & Schwede, T. (2009). The SWISS-MODEL Repository and associated resources. *Nucleic acids research,* **37**:387–392.

17. Chen, Y., Zhang, Y., Wang, A. Y., Gao, M., & Chong, Z. (2021). Accurate long-read *de novo* assembly evaluation with Inspector. *Genome biology,* **22(1):** 1-21.

18. Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome research,* **27(5):** 722–736.

19. Murigneux, V., Rai, S. K., Furtado, A. *et al.* (2020). Comparison of long-read methods for sequencing and assembly of a plant genome. *GigaScience,* **9(12):** 146.

20. Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome research,* **27(5):** 737–746.

21. Wick, R. R., Judd, L. M., Cerdeira, L. T., Hawkey, J., Méric, G., Vezina, B., Wyres, K. L., & Holt, K. E. (2021). Trycycler: consensus long-read assemblies for bacterial genomes. *Genome biology,* **22(1):** 266.

22. Wick R.R., Schultz M.B., Zobel J. & Holt K.E. (2015). Bandage: interactive visualisation of *de novo* genome assemblies. *Bioinformatics,* **31(20):** 3350-3352.

23. Norais, C., Hawkins, M., Hartman, A. L., Eisen, J. A., Myllykallio, H., & Allers, T. (2007). Genetic and physical mapping of DNA replication origins in *Haloferax volcanii.* *PLOS genetics,* **3(5):** 77.

24. Wendoloski, D., Ferrer, C., & Dyall-Smith, M. L. (2001). A new simvastatin (mevinolin)-resistance marker from *Haloarcula hispanica* and a new *Haloferax volcanii* strain cured of plasmid pHV2. *Microbiology,* **147(4):** 959–964.

25. Wang, Y., Zhao, Y., Bollas, A. *et al.* (2021). Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology,* **39:**1348–1365.

26. Allers, T. *et al.* (2004). Development of additional selectable markers for the halophilic archaeon *Haloferax volcanii* based on the *leuB* and *trpA* genes. *Applied and environmental microbiology,* **70(2):** 943-53.

27. Hawkins, M. *et al.* (2013). Accelerated growth in the absence of DNA replication origins. *Nature,* **503(7477):**544-547.

28. Pérez-Arnaiz, P., Dattani, A., Smith, V., & Allers, T. (2020). *Haloferax volcanii* -a model archaeon for studying DNA replication and repair. *Open biology,* **10(12):**200293.

29. Nagata, M., Ishino, S., Yamagami, T., Simons, J. R., Kanai, T., Atomi, H., & Ishino, Y. (2017). Possible function of the second RecJ-like protein in stalled replication-fork repair by interacting with Hef. *Scientific reports,* **7(1):** 16949.

30. Foley, E. & Kapoor, T. (2013). Microtubule attachment and spindle assembly checkpoint signalling at the kinetochore. *Nature reviews, molecular cell biology*, **14(1):** 25-37.

**Word Count:**

Project Role: 82

Aims & Experimental Question: 385

Background & Motivation: 425

Methods: 575

Results and Discussion: 1866

**Total**: 3333