|  |
| --- |
| **BIS7023-B MSc RESEARCH PROJECT 1**  ***INTERIM REPORT***  Name:Mohammed Bilal  UoB number:14006239  MSc Programme:Bioinformatics  Word count:2598 |
| **Title of Research Project** |
| Goals for a new reference-grade *de novo* assembly: A comparative study between existing naked mole-rat genome assemblies and the future of *de novo* genome assembly of the naked-mole rat. |
| **Summary of work done so far** |
| ‘QUAST’ (quality assessment tool) evaluation(s) using the web based platform ‘Galaxy’ (for biomedical research) has provided some insight into the robustness of the two published naked mole-rat/NMR genomes from Kim *et al*, 2011 (alias: HetGla\_1.0) and Keane *et al*, 2014 (alias: HetGla\_female\_1.0). A ‘mauve’ genome alignment of NMR assemblies has also been performed, work in this area is still on-going and I hope to produce a graphical illustration of aligned NMR genomes using the ‘Mauve-Viewer’ in order to identify conserved and homologous regions between the two assemblies. A bioinformatic pipeline has been created outlining the major steps for a future *de novo* genome assembly of raw naked mole-rat sequencing data. The pipeline describes the software tools I intend to use in order to construct a reference-grade genome assembly which will be tested/evaluated against previously published naked mole rat assemblies HetGla\_1.0 and HetGla\_female\_1.0. |
| **Background** |
| Two groups have separately published draft genome assemblies of an individual naked mole-rat (Kim *et al*, 2011 & Keane *et al*, 2014). Both assemblies utilise whole genome shotgun sequencing using high coverage Illumina data with various insert sizes (Lewis *et al*, 2016). One major fault with this approach are that assemblies generated are fragmented and disjointed, containing high levels of unoccupied gaps (15%) preventing genuine analysis of gene regulation and gene expression (Lewis *et al,* 2016). Initial short-read Illumina assemblies yielded major levels of misassemblies and collapsed homologous genes and pseudogenes (Zimin *et al,* 2012). However, with the emergence of long-read sequencing/third-generation technologies (Pacific Biosciences, Oxford Nanopore etc) the quality of the sequencing data will improve, effectively constructing genome assemblies of higher unprecedented qualities and paving the way for new genome, transcriptome and metagenomic studies at higher resolutions (van Dijk *et al,* 2018).  The Current State of Naked Mole-Rat Genomes  Two published genome assemblies of the NMR exist. Their genomic sequences and genome annotations can all freely be downloaded from GenBank or RefSeq databases in multiple file formats (.fna/.gff/.txt etc) from the NCBI website see URL : https://www.ncbi.nlm.nih.gov/assembly/?term=Heterocephalus+glaber.  *‘Assembly\_1’ – refers to the first NMR genome assembly.*   1. From Kim *et al*, 2011 (alias: HetGla\_1.0): An individual male naked mole-rat from a captivity breeding colony in Chicago (Illinois-USA) was sequenced on the Illumina HiSeq 2000 platform at the Beijing Genomics Institute (BGI) to 92-fold coverage with a contig N50 of 19.3Kbp (kilobase pair) and scaffold N50 of 1.6Mbp (megabase pair). The genome was assembled using the novel short-read assembly programme SOAPdenovo (version v. 1.05) with default parameters (Xie *et al,* 2014).   ‘Assembly\_2’ – *refers to the second NMR genome assembly.*   1. From Keane *et al*, 2014 (alias: HetGla\_female\_1.0): High-molecular-weight genomic DNA (gDNA) was obtained from the tissues of an inbred female adult NMR obtained from the University of Rochester, (New York-USA). This was later sequenced on the Hi-Seq Illumina platform at the Broad Institute of Genomics, (Cambridge-USA) to 90-fold coverage with a contig N50 47.8Kbp and scaffold N50 of 20.5Mbp. The genome was assembled using the short-read software programme ALLPATHS-LG (version R39605) with default parameters (Gnerre et al, 2011).  |  |  |  | | --- | --- | --- | |  | ‘Assembly\_1’ - HetGla\_1.0 (2011) | ‘Assembly\_2’ - HetGla\_female\_1.0 (2014) | | Assembly Level | Scaffold | Scaffold | | Assembly Size (Gb) | 2.66 | 2.62 | | Assembly Method | SOAPdenovo v. 1.05 | AllPaths v. R39605 | | Genome Coverage | **x92 (fold)** | x90(fold) | | Number of Contigs | 273,991 | 114, 653 | | Number of Scaffolds | 39,267 | 4,229 | | Contig N50 | 19.3Kb | **47.8Kb** | | Scaffold N50 | 1.6Mb | **20.5Mb** | | Total Sequence Length | 2.64B (Billion/109) | 2.62B | | Total Ungapped Length | **2.43B** | 2.31B | | Spanned Gaps | 234,724 | 110,424 | | Unfilled Gaps | 214 Mb | 303 Mb |   Table 1. Comparison of global statistics for the NMR genome assemblies. “Mb” - Megabase (106). “Kb” - Kilobase (103). Best value indicated in **bold.**  (Data for ‘Assembly\_1’ from NCBI website under accession GCF\_000230445.1)  (Data for ‘Assembly\_2’ from NCBI website under accession GCF\_000247695.1)   |  | | --- | | KEY GENOME ASSEMBLY TERMS  *Contig* – Numerous gel readings connected to one another by overlapping of their sequences. Gel readings in a contig can be totalised to form a contiguous consensus sequence (R Staden, 1980).  *Coverage* - The average number of reads covering a particular position in the genome. Used as a confidence parameter for determining base calls/variant discovery at a particular base position. Higher coverage mean each base is covered by a greater number of aligned sequenced reads.  *Gaps* - Regions within an object where there are no known sequence. Represented as a string of the letter ‘N’.  *N50*-The shortest contig/scaffolds length needed to cover 50% of the genome. To get the N50 -> sort all contigs/scaffolds of a genome by length -> go to the base at 50% of the genome length -> get the contig/scaffold size to which the base belongs to -> your N50 value.  *Read*- The unit of DNA sequence that comes from a sequencing instrument. Single reads can be paired together to form paired-end reads (combination of a forward read and a reverse read).  *Scaffold* - An object consisting of one or more sequence contigs, connected by spanned gaps. Scaffolding is the process of ordering and orientating contigs the preferred way is based on assembling reads into contigs then using paired-end information to join them into scaffolds.  *Spanned* g*aps* - Gaps that have linkage evidence (see NCBI AGP specification v2.1) also called within-scaffold gaps.  *Sequence**length*-Total length of all top-level sequences i.e. the most highly assembled sequences in the assembly.  *Ungapped**length*-Total length of all top-level sequences, Ignoring gaps. 10+ Ns in a sequence is treated like a gap, they occur due to ambiguous base calls. |   Both naked mole-rat assemblies (Assembly\_1 & Assembly\_2) are assembled at the **scaffold** level, where most scaffolds are **unplaced**, this means that the objects (scaffolds) do not have a chromosome, linkage-group, or plasmid assignment. The assemblies are highly fragmented due to the nature of the sequencing technology (Illumina – shotgun sequencing). Assembly\_2 has been described as being a higher quality assembly due to a substantially higher N50 values for scaffold and contigs (Keane *et al,* 2014). However, genomes of the naked mole-rat represent differing levels of completeness, scientist need to exercise caution when comparing NMR genomes to other closely linked species as observation could be due to technical artefacts or limitations of sequencing technology or assembly algorithms rather than true evolutionary changes (Lewis *et al,* 2016). |
| **Hypothesis** |
| Out of the two published genome assemblies for the naked mole-rat, female genome ‘Assembly\_2’ from Keane *et al (*2014) is the more superior genome assembly (for data mining/future NMR gene study) than the male genome ‘Assembly\_1’ from Kim *et al (*2011) because of higher N50 values for scaffold and contig. |
| **Aims and Objectives** |
| 1. To encapsulate the state of current sequencing technologies, algorithms, and the naked mole-rat genome assemblies. 2. To create a ‘mauve’ genome alignment in order to visualise the two naked mole-rat genomes. 3. To critically assess the quality of the naked mole-rat genome assemblies using ‘QUAST’ - a quality assessment tool for assessing genome. 4. To build a full bioinformatic pipeline for a new *de novo* genome assembly using existing Illumina NMR reads and the software assemblers ‘ABySS’, ‘Velvet’ and ‘SPAdes’ to generate an improved genome assembly. |
| **Methods** |
| This section describes the data analysis methods used in order to generate meaningful results from the naked mole-rat genome assemblies which can support the hypothesis and the aims and objectives set above.  Method 1: Constructing a Mauve Genome Alignment  Mauve is a free open-source software that delivers a platform for comparative genomics and population genomic studies (Darling *et al,* 2010). A mauve multiple genome alignment (version 2.4.0) between the two published genomes for the naked mole-rat (HetGla\_1.0 & HetGla\_female\_1.0) was created in a Linux terminal from which I had remote access to a powerful Bradford University SSH server.  The command below in the green box was typed into the Linux terminal to generate a mauve alignment:  progressiveMauve --output=H\_glaber.xmfa /home/bilalm/H\_glaber\_genomes/GCA\_000230445.1\_HetGla\_1.0\_genomic.fna /home/bilalm/H\_glaber\_genomes/ GCA\_000247695.1\_HetGla\_female\_1.0\_genomic.fna  The alignment consisting of both NMR assemblies (HetGla\_1.0 & HetGla\_female\_1.0) was created with the progressiveMauve algorithm which produced three output files. The outputs were: (1) .XMFA file – containing the mauve genome alignments (2) .backbone file – records of conserved regions (3) .bbcols file – records of (predicted) alignment positions.    **Figure 2** The backbone file (.xmfa.backbone)  **Figure 3** The alignment positions file (.xmfa.bbcols)  **Figure 1** The alignment file (.xmfa)  It is possible for the alignment file (.XMFA) to be graphically visualised using a freely available genome alignment visualisation tool called Mauve-Viewer (version 20150226). However, this is not included within this report. In the Mauve-Viewer, genomic regions are represented as unique coloured blocks. Coloured blocks in the first genome are connected by lines to similar coloured blocks in the second genome. The lines specify which regions in each genome are homologous.  Method 2: Quality Assessment Tool for Genome Assemblies - QUAST  QUAST is a quality assessment tool for evaluating and comparing genome assemblies (Gurevich *et al,* 2013). I used the QUAST tool in Galaxy (Galaxy Version 5.0.2+galaxy1) to generate a number of summary statistics in numerical and graphical format, in order to gauge an assembly quality for the naked mole-rat genomes. Firstly, both genome (HetGla\_1.0 & HetGla\_female\_1.0) were downloaded from the NCBI website under accession numbers ‘GCF\_000230445.1’ & ‘GCF\_000247695.1’ in FASTA/.fna format. Before I could feed these files into QUAST, a decompression step was required as the newly downloaded .fna files were in a compressed .gz format. I used the command ‘$gunzip’ in my command-line to expand the compressed file and release the raw FASTA files for both NMR assemblies. Two FASTA files were easily imported into Galaxy using a drag and drop box and now I was able to feed the files directly into QUAST. I conducted **two** QUAST evaluation of NMR genomes. The first QUAST analysis was made up of just the two existing NMR genomes in FASTA format, as I wanted to verify global statistics e.g. N50 for contig and scaffold, total length, GC% etc. The second more in-depth QUAST analysis required a reference genome and an annotation file in .gff (general feature) format. In this case I used the female NMR assembly (HetGla\_female\_1.0) for the reference, as specified on the NCBI website (see: https://www.ncbi.nlm.nih.gov/genome/?term=Naked+mole-rat) and for the genome annotation file I also used the NCBI directed .gff annotation file for HetGla\_female\_1.0 genome. The outputs of the second QUAST analysis were many more genome assembly statistics (NGx, NAx, NGAx graphs – only available if reference inputted), misassemblies & unaligned reports. However, one could speculate that the results for the second QUAST analysis were more charitable/biased towards the HetGla\_female\_1.0 assembly. For full QUAST results visit: https://github.com/mbilal1995/Research-Project\_1 |
| **Results obtained so far** |
| Results obtained from Methods 2:  Figure 4. This is a tabular report presenting statistics for the first QUAST evaluation of the two genomes assemblies for the naked mole-rat (HetGla\_female\_1.0 & HetGla\_1.0) using the Galaxy GUI.    Figure 5. This is a QUAST HTML report comparing the two NMR genome assemblies’ side-by-side. This has been generated using the Galaxy GUI. The heatmap illustrates the best values in blue and worst values in red.    Figure 6. This is a ‘cumulative length’ plot showing the number of bases in the first *x* contigs, as *x* varies from zero to the number of contigs. The y-axis is the size of the assembly & the x-axis is the number of contigs.    Figure 7. This is a ‘N*x’* plot, a variation of N50 based on aligned blocks. This plot shows the changes of N*x* value in dependence of x value. The y-axis is the N*x* value and x-axis is the x value.    Figure 8. This is a ‘GC content’ plot. This shows the distribution of GC content in the contigs. X-axis shows the percent of GC from (0-100%) and the y-axis shows the number of non-overlapping 100bp windows whose GC content is *x.* |
| **Discussion of current state of the work and future directions** |
| *De novo* assembly refers to assembling a novel genome where no reference sequence is available for alignment. In *de novo* genome assembly, the end result may have DNA lines containing runs of ‘N’s these represent regions within an object (scaffold) where there are no known sequence or gap regions. N50 is a metric widely used to describe how well a software assembly program has succeeded with forming together contigs and scaffolds in a genome assemblies. N50 is defined as a weighted median statistic such that 50% of the assembly is contained in contigs that are equal to or larger than this value (Castro & Ng, 2017). Performance and completeness assessment of a *de novo* genome assembly is presently measured by its N50 (Mäkinen *et al,* 2012). All in all, it is understood that the greater the N50, the more accurate the assembly (Mäkinen *et al,* 2012). With this information in mind, the tabular report from a QUAST evaluation (fig.4) of the two published naked mole-rat assemblies confirmed a higher N50 value for HetGla\_female\_1.0 (20,532,749) vs HetGla\_1.0 (1,604,037) signifying the female NMR assembly, from Keane *et al* (2014) as the more superior assembly. Armed with this information, we can now accept the hypothesis. In addition to the N50, it is important to report the number of gaps (N’s). QUAST defines ‘N’ as the total number of uncalled bases in the assembly. HTML QUAST reports (fig.5) a higher number of N’s for the HetGla\_female\_1.0 (303,433,536) compared to the HetGla\_1.0 assembly (213, 913, 259). QUAST defines ‘Ns per 100Kbp’ (kilobase pair) as the average number of uncalled bases per 100,000 assembly bases. This metric is also higher in the HetGla\_female\_1.0 (11,589) than its counterpart (8111.31). In this regard the HetGla\_1.0 is a better genome assembly as it has less ambiguous bases (N’s). Fig6 is a ‘cumulative length’ graph describing the number of total bases in the 1000th contigs as being higher in HetGla\_female\_1.0 (2,612,076,708) than in HetGla\_1.0 (1,911,909,966). There are less contigs in the HetGla\_female\_1.0 assembly (4229) compared to HetGla\_1.0 (39,267). Having less contigs does not make one assembly better than the other, other factors play more important roles like scaffold number, sequence coverage, read length, sequencing platform, and assembler algorithm. L50 is defined as the minimum length of all scaffold that together account for 50% of the genome (Schneeberger *et al*, 2011). HetGla\_female\_1.0 has an L50 of 42 and HetGla\_1.0, an L50 of 500 (fig.5) essentially making the female NMR genome assembly more efficient. Contiguous sequences for the female NMR assembly appear to be larger in length (Fig.7) and have higher base counts within the contigs(Fig.6) further supporting the superiority of HetGla\_female\_1.0 over HetGla\_1.0.  Future Directions  In the future I intend to build a full *de novo* genome assembly of naked mole-rat reads based off Illumina sequencing data from the NCBI website. The research question will be: “Working with the existing Illumina sequencing data, is It possible to make an improved *de novo* assembly from either the reads of one or both of the NMR genomes?” The bioinformatic pipeline consisting of the major steps I will use to solve the research question are set out in the black box below:   1. Download HetGla\_female\_1.0 & HetGla\_1.0 naked mole-rat Illumina reads from the NCBI website: (https://www.ncbi.nlm.nih.gov/assembly/?term=naked+mole-rat). 2. Examine the quality of the raw reads using the software programme *FastQC* with in a Linux terminal. 3. Trim poor quality Illumina reads/adapters by using the programme *Trimmomatic* with in the Linux terminal.   (Note: Use the output files from the *Trimmomatic* tool for the next steps).   1. Download and compile the ‘ABySS’, ‘Velvet’ and ‘SPAdes’ genome assembler(s) in the Linux terminal. 2. Assemble raw NMR Illumina reads into a draft genome assembly using each of the assemblers. Use the tool *SSPACE* to control the scaffolding process. *SAM tools* can be used for manipulating alignments in the SAM format. 3. Examine the contiguity statistics of the ‘new’ genome assemblies use the tool *SAMStat* and compare them to the HetGla\_female\_1.0 (Keane *et al,* 2014) to see if real improvements have been made. 4. Perform QUAST evaluation(s) of the new genomes to gauge the best and worst assemblies.   Fig9. *De novo* genome assembly for the naked mole-rat.  Having completed a literature review around genome assembly specifically sequencing technologies, sequencing assemblers, and algorithms I feel as though limitations with both NMR genome assemblies stem from the nature of the sequencing technology i.e. whole-genome shotgun sequencing producing short-reads. NMR assemblies have been shown to be less contiguous (for N50) and more gapped than the Mouse (GRCm38.p4) and Human (GRCh38.ph) genomes (Lewis *et al*, 2016). One reason for this may be the inclusion or the mixture of different types of sequencing technologies to help aid assembly i.e. Nanopore sequencing/ultra-long reads etc. For Literature review: https://github.com/mbilal1995/Research-Project\_1 . |

|  |
| --- |
| **References** *no word limit* |
| Board on Agriculture and Natural, R., Waddell, K., Pool, R., National Research, C., Division on Earth and Life, S. and Board on Life, S. (2002) *Exploring Horizons for Domestic Animal Genomics : Workshop Summary.* Washington, D.C., UNITED STATES: National Academies Press.  Bruijn, d. N. G. (1946) A combinatorial problem. *Proceedings of the Section of Sciences of the Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam* 49 (7), 758 - 764.  Castro, C. J. and Ng, T. F. F. (2017) U(50): A New Metric for Measuring Assembly Output Based on Non-Overlapping, Target-Specific Contigs. *J Comput Biol* 24 (11), 1071-1080.  Choudhuri, S. (2014) Chapter 7 - Additional Bioinformatic Analyses Involving Nucleic-Acid Sequences\*\*The opinions expressed in this chapter are the author’s own and they do not necessarily reflect the opinions of the FDA, the DHHS, or the Federal Government. In Choudhuri, S. (editor) *Bioinformatics for Beginners.*  Oxford: Academic Press. 157-181.  Compeau, P. E., Pevzner, P. A. and Tesler, G. (2011) How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol* 29 (11), 987-91.  Darling, A. E., Mau, B. and Perna, N. T. (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PloS one* 5 (6), e11147.  Euler, L. (1736) Solutio problematis ad geometriam situs pertinensis. *Comm. Acad. Sci. Imper. Petropol.* 8, 128-140.  Fan, H., Wu, Q., Wei, F., Yang, F., Ng, B. L. and Hu, Y. (2019) Chromosome-level genome assembly for giant panda provides novel insights into Carnivora chromosome evolution. *Genome Biology* 20 (1), 267.  Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., Sharpe, T., Hall, G., Shea, T. P., Sykes, S., Berlin, A. M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E. S. and Jaffe, D. B. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A* 108 (4), 1513-8.  Goodwin, S., McPherson, J. D. and McCombie, W. R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* 17 (6), 333-351.  Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29 (8), 1072-5.  Hamilton, J. P. and Robin Buell, C. (2012) Advances in plant genome sequencing. *The Plant Journal* 70 (1), 177-190.  Human Genome Sequencing Consortium, I., International Human Genome Sequencing, C. and Joint Genome, I. (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431 (7011), 931-945.  Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., Snutch, T. P., Tee, L., Paten, B., Phillippy, A. M., Simpson, J. T., Loman, N. J. and Loose, M. (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* 36 (4), 338-345.  Khan, A. R., Pervez, M. T., Babar, M. E., Naveed, N. and Shoaib, M. (2018) A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective. *Evolutionary bioinformatics online* 14, 1176934318758650-1176934318758650.  Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H. and Phillippy, A. M. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* 27 (5), 722-736.  Lewis, K. N., Soifer, I., Melamud, E., Roy, M., McIsaac, R. S., Hibbs, M. and Buffenstein, R. (2016) Unraveling the message: insights into comparative genomics of the naked mole-rat. *Mammalian Genome* 27 (7), 259-278.  Li, R. Fan, W. Tian, G. Zhu, H. He, L. Cai, J. Huang, Q. Cai, Q. Li, B. Bai, Y. Zhang, Z. Zhang, Y. Wang, W. Li, J. Wei, F. Li, H. Jian, M. Li, J. Zhang, Z. Nielsen, R. Li, D. Gu, W. Yang, Z. Xuan, Z. Ryder, O. A. Leung, F. C.-C. Zhou, Y. Cao, J. Sun, X. Fu, Y. Fang, X. Guo, X. Wang, B. Hou, R. Shen, F. Mu, B. Ni, P. Lin, R. Qian, W. Wang, G. Yu, C. Nie, W. Wang, J. Wu, Z. Liang, H. Min, J. Wu, Q. Cheng, S. Ruan, J. Wang, M. Shi, Z. Wen, M. Liu, B. Ren, X. Zheng, H. Dong, D. Cook, K. Shan, G. Zhang, H. Kosiol, C. Xie, X. Lu, Z. Zheng, H. Li, Y. Steiner, C. C. Lam, T. T.-Y. Lin, S. Zhang, Q. Li, G. Tian, J. Gong, T. Liu, H. Zhang, D. Fang, L. Ye, C. Zhang, J. Hu, W. Xu, A. Ren, Y. Zhang, G. Bruford, M. W. Li, Q. Ma, L. Guo, Y. An, N. Hu, Y. Zheng, Y. Shi, Y. Li, Z. Liu, Q. Chen, Y. Zhao, J. Qu, N. Zhao, S. Tian, F. Wang, X. Wang, H. Xu, L. Liu, X. Vinar, T. Wang, Y. Lam, T.-W. Yiu, S.-M. Liu, S. Zhang, H. Li, D. Huang, Y. Wang, X. Yang, G. Jiang, Z. Wang, J. Qin, N. Li, L. Li, J. Bolund, L. Kristiansen, K. Wong, G. K.-S. Olson, M. Zhang, X. Li, S. Yang, H. Wang, J. and Wang, J. (2010) The sequence and de novo assembly of the giant panda genome. *Nature* 463 (7279), 311-317.  Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., Yang, B. and Fan, W. (2011) Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Briefings in Functional Genomics* 11 (1), 25-37.  Metzker, M. L. (2010) Sequencing technologies — the next generation. *Nature Reviews Genetics* 11 (1), 31-46.  Miller, J. R., Koren, S. and Sutton, G. (2010) Assembly algorithms for next-generation sequencing data. *Genomics* 95 (6), 315-27.  Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Knut, H. J. R., Remington, K. A., Anson, E. L., Bolanos, R. A., Chou, H.-H., Jordan, C. M., Halpern, A. L., Lonardi, S., Beasley, E. M., Brandon, R. C., Chen, L., Dunn, P. J., Lai, Z., Liang, Y., Nusskern, D. R., Zhan, M., Zhang, Q., Zheng, X., Rubin, G. M., Adams, M. D. and Venter, J. C. (2000) A Whole-Genome Assembly of Drosophila. *Science* 287 (5461), 2196-2204.  Orr, M. E., Garbarino, V. R., Salinas, A. and Buffenstein, R. (2016) Extended Postnatal Brain Development in the Longest-Lived Rodent: Prolonged Maintenance of Neotenous Traits in the Naked Mole-Rat Brain. *Frontiers in neuroscience* 10, 504.  Pop, M., Salzberg, S. L. and Shumway, M. (2002) Genome sequence assembly: algorithms and issues. *Computer* 35 (7), 47-54.  Sanger, F. and Coulson, A. R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* 94 (3), 441-8.  Sanger, F., Nicklen, S. and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* 74 (12), 5463-5467.  Schneeberger, K., Ossowski, S., Ott, F., Klein, J. D., Wang, X., Lanz, C., Smith, L. M., Cao, J., Fitz, J., Warthmann, N., Henz, S. R., Huson, D. H. and Weigel, D. (2011) Reference-guided assembly of four diverse <em>Arabidopsis thaliana</em> genomes. *Proceedings of the National Academy of Sciences* 108 (25), 10249.  Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J. and Timp, W. (2017) Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods* 14 (4), 407-410.  Staden, R. (1980) A new computer method for the storage and manipulation of DNA gel reading data. *Nucleic Acids Res* 8 (16), 3673-94.  van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. and Thermes, C. (2018) The Third Revolution in Sequencing Technology. *Trends Genet* 34 (9), 666-681.  Vassy, J. L., Christensen, K. D., Schonman, E. F., Blout, C. L., Robinson, J. O., Krier, J. B., Diamond, P. M., Lebo, M., Machini, K., Azzariti, D. R., Dukhovny, D., Bates, D. W., MacRae, C. A., Murray, M. F., Rehm, H. L., McGuire, A. L., Green, R. C., MedSeq, P. and for the MedSeq, P. (2017) The Impact of Whole-Genome Sequencing on the Primary Care and Outcomes of Healthy Adult Patients: A Pilot Randomized Trial. *Annals of internal medicine* 167 (3), 159-169.  Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., Zhou, X., Lam, T.-W., Li, Y., Xu, X., Wong, G. K.-S. and Wang, J. (2014) SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics (Oxford, England)* 30 (12), 1660-1666.  Yoon, S., Kim, D., Kang, K. and Park, W. J. (2018) TraRECo: a greedy approach based de novo transcriptome assembler with read error correction using consensus matrix. *BMC Genomics* 19 (1), 653.  Zerbino, D. R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18 (5), 821-9.  Zimin, A. V., Kelley, D. R., Roberts, M., Marçais, G., Salzberg, S. L. and Yorke, J. A. (2012) Mis-assembled "segmental duplications" in two versions of the Bos taurus genome. *PLoS One* 7 (8), e42680. |