GENOME ASSEMBLY OF FEMALE NAKED MOLE-RAT WHOLE GENOME SEQUENCING DATA UTILISING DE BRUIJN GRAPH *DE NOVO* GENOME ASSEMBLERS ‘ABYSS’ AND ‘VELVET’

By

Mohammed Bilal

Dissertation submitted in partial

fulfilment of the requirements for the

degree of

Master by Advanced Study in Bioinformatics

School of Chemistry and Biosciences

University of Bradford

Bradford

7th September 2020

DECLARATION OF ACADEMIC INTEGRITY

By signing this page, I confirm that I have read the University policy on Plagiarism, the statement on Academic Integrity and the regulation on the presentation of work for formal assessment on the University website. I state this piece of work is my own and does not contain any unacknowledged work from other sources.

Signed: MOHAMMED BILAL

Date: 7TH SEPTEMBER 2020

TABLE OF CONTENTS

INDEX OF TABLES

INDEX OF FIGURES

ABSTRACT

Living exceptionally long lives and displaying negligible signs of ageing while highly resistant to cancer, diabetes and neurodegenerative diseases, unaffected by certain types of pain despite harsh and hypoxic living environments, the naked mole-rat (*Heterocephalus glaber,* NMR) is indeed a paradox of natural life achieving both an extended health and life span (Saldmann et al, 2019; Lewis et al, 2018). In order for researchers to fully understand the fascinating phenotypic traits presented by the NMR, an adequate genome tool (genome assembly) is required from which genomic data can be drawn out of and interpreted, in order to uncover molecular insights into the NMR. This study aimed to create high quality/optimised genome assemblies for published Illumina HiSeq NMR sequencing data obtained from NCBI SRA database using *de novo* genome assemblers ABySS (Simpson *et al,* 2009) and Velvet (Zerbino, 2008). NMR data files were checked for high quality, optimised for parameter *k* (k-mer) and assembled by esoteric short-read De Bruijn graph *de* novo genome assemblers.

Results: Based on Nx metrics to describe the quality of genome assemblies, results for all split NMR data files using ABySS *de novo* assembler, achieved greater N50 values than the Velvet assembler, advocating ABySS as the ‘better’ genome assembler for next-generation sequencing, short-read, single-ended naked mole-rat fastq data. These finding may aid researchers in choosing a suitable genome assembling software (assembler) or finding optimal *k*-mer lengths for short-read NMR HiSeq data.

ACKNOWLEDGEMENTS

I would like to express my gratefulness for the help and support I received during the completion of the dissertation. I would like to acknowledge my supervisor, Dr Andrew Tedder. This piece of work was made possible by their guidance and provisions. Lastly, I would like to thank my family and friends for their continuous support throughout my university journey.

ABBREVIATIONS

ASIS	ARF Suppression Induced Senescence
BAC	Bacterial Artificial Chromosome
CPU	Central Processing Unit
DBG	De Bruijn Graph
ECI	Early Contact Inhibition
G/GB	Gigabyte
Hi-C	Chromosome Conformation Capture
H_2O_2	Hydrogen Peroxide
HIF1α	Hypoxia Inducible Factor 1 Alpha
HiSeq	High Throughput Sequencing
iPSC	Induced Pluripotent Stem Cell
Kb	Kilobase
Mb	Megabase
ML	Machine Learning
NCBI	National Center for Biotechnology Information
NMR	Naked Mole Rat
NGS	Next Generation Sequencing
NGF	Nerve Growth Factor

OLC	Overlap Layout Consensus
PacBio	Pacific Biosciences
PRD	Proline Rich Domain
QC	Quality Control
RAM	Random Access Memory
SMRT	Single Molecule Real Time
SMS	Single Molecule Sequencing
SRA	Sequence Read Archive
STDOUT	Standard Output
VEGF-A	Vascular Endothelial Growth Factor-A
WGS	Whole Genome Shotgun

INTRODUCTION

The Naked Mole Rat

Heterocephalus glaber or the *blesmols* or more commonly known as the African naked mole-rat (NMR) originate from Sub-Saharan arid and semi-arid regions of Eastern Africa (Ethiopia, Kenya, Somalia). NMRs are subterranean burrowing rodents, living entirely in underground 'labyrinths' which span for many kilometres (Sherman *et al*, 1999). Working co-operatively to build tunnel systems, excavators of these burrows use their sabretooth-like incisors to gnaw (bite) rock-hard soil while other mole-rats transport soil accumulations to the surface and expel them with their hind feet, the whole system functions as a conveyor belt. NMRs are eusocially organised in colonies of up to 295 individuals (mean of 75) with a similar social structure to those seen in ants, termites, and wasps (Brett 2017; Bens *et al,* 2018). These social structures typically consist of a single powerful female "Queen" and several reproductively active males with all other individuals being subordinate. Each subordinate will have a specific job such as food scout, tunnel digger, baby nurturer and burrow protector. Breeding is sequestered to a single reproductive female, creating mole-rat colonies which are "one of the most inbred of all free-living mammals" (Gillespie *et al* 2007). Though non-breeding NMRs are not obligately sterile they can become reproductively active when divided from a colony (Ciszek, 2000). NMRs consume tubers of geophytes (plant storage organs) which can store water, sugars and starch, typically only consuming small portions for quicker regeneration and future feeds (Bennett, 1995). NMRs were first described in 1842 by German naturalist Eduard Rüppell while he was in Ethiopia documenting the African mammals (Edrey *et* *al,* 2011). He named them *Heterocephalus* ("different-headed") *glaber* ("hairless"). However, NMRs are not entirely hairless, they have vibrissae around their snouts and tail that, when stimulated, act as sensory organs. NMRs belong to the family *Bathyergidae,* comprising of five genera with roughly 12 species of African rodent. NMRs are most closely aligned to rodents in the suborder *Hystricomorpha* consisting of chinchillas, guinea pigs and porcupines (Lewis *et al,* 2016). Most rodents in the *Bathyergidae* family have developed adaptions for terrestrial life including shorter limbs, loose skin, tube-shaped bodies, small eyes/ears, and large cutting incisors (Park *et al,* 2010).

Fascinating phenotypic traits presented by the naked mole-rat have gained the attention of academics and researchers involved in studying the underlying genetics of long-life, cancer resistance and genome maintenance (Petruseva *et al,* 2017). Living exceptionally long lives, displaying negligible signs of ageing while highly resistant to cancer, diabetes and neurodegenerative diseases, unaffected by certain types of pain despite harsh, stressful, and hypoxic living environments, the NMR is indeed a paradox of natural life achieving both an extended health and life span (Saldmann *et al,* 2019; Lewis *et al,* 2018). By sequencing and analysing the NMR genome and transcriptome, scientists have uncovered insights into the extraordinary traits of the NMR consistent with cancer resistance, high tolerance to hypoxic ecosystems and exceptional longevity (Kim *et al,* 2011; Keane *et al,* 2014, Fang *et al,* 2014). NMRs are the longest-lived rodents with a maximum lifespan of more than 31 years in laboratory conditions, five times greater than projections based on its body size (Edrey *et al,* 2011). It is well understood that NMRs live up to 80% of their lives with no age-related increase in mortality or morbidity (Edrey *et al,* 2011). NMRs do not conform to the biological law for ageing and transgress Gompertzian laws of mortality which describes ageing, mathematically, the law which applies to all mammals (after maturity) that the risk of dying increase as age increases (Gompertz, 1825). One reason for their extended lifespan may be due to a greater capability of NMR mitochondria to detoxify hydrogen peroxide (H_20_2) however, further efforts are required to confirm if greater H_20_2 detoxification is a universal trait found amongst long-lived rodents relative to shorter-lived ones (Munro *et al*, 2019). The NMR presents the perfect organism to study the effects of oxidative stress and its resistance, in relation to understanding healthy ageing (Saldmann *et al*, 2019). Cancer resistance is another extraordinary feature that NMRs demonstrate, In the last two decades, researchers monitoring NMRs have found only six cases of neoplasia from amongst thousands of NMRs found in zoo and laboratory colonies (Delaney *et al*, 2013; Delaney *et al*, 2016). As NMR cells do not display replicative senescence, they depend on early-acting anti-cancer mechanisms, one well-studied mechanism, discovered in 2009, is the phenomenon of 'early contact inhibition' (ECI) where cell division is arrested in high cell densities (Seluanov *et al*, 2009). Typically, adherent cells display contact inhibition whereby non-cancerous cells halt their growth when in contact with each other (Pavel *et al*, 2018). Cancerous cells lose this feature, so continue to proliferate uncontrollably. ECI in NMRs was first thought to be connected with high levels of p16INK4a (Seluanov *et al*, 2009) however, more recent studies using RNA sequencing data have suggested a two-tier ECI protection model with the expression of protein pALTINK4a/b inducing cell cycle arrest both under ECI and genotoxic stress, advocating an increased resistance to uncontrollable cell division in NMRs (Petruseva *et al*, 2017). Also, NMR cells were shown to be impervious to induced pluripotent stem cell (iPSC) reprogramming via Yamanaka factors (*Oct4*, *Sox2*, *Klf4*, *c-Myc*). NMR-iPSCs transplanted into mice were investigated for teratoma-forming potential, however, teratomas did not form due to NMR-specific tumour-suppression mechanism 'ARF suppression-induced senescence' (ASIS) which acts as a defence against iPSC tumorigenicity (Miyawaki *et al,* 2016). However, further research into the mechanisms underlying ASIS is needed to produce safe iPSCs for cell-based cancer treatment therapies In humans (Miyawaki *et al,* 2016). NMRs are an extremely hypoxia-tolerant species and have adapted to live in overcrowded colonies deep underground where conditions are harsh, O_2 levels can plummet to around 2% while CO_2 levels are high at around 7-10% (Xiao *et al,* 2017; McNab, 1966). Under experimental conditions, NMRs have been reported to survive for up to 18 minutes in complete anoxia (absence of oxygen) via atmospheric chamber without sustaining injury (Park *et al,* 2017). This feat likely is due to NMR adaptions to utilise fructose in an oxygen-independent manner, delivering fuel to vital organs such as the brain and the heart, via GLUT5 transporter (Browe *et al,* 2020). Interestingly, fructose metabolism coupled with hypoxic stress is linked to cancer malignancy and heart failure (Port *et al,* 2012). Therefore, It is vital to understand how NMRs use this system with zero psychosomatic drawbacks, molecular insights could benefit researchers in ischaemia related human diseases such as stroke and coronary heart disease/myocardial ischemia (Park *et al,* 2017). NMRs also use behavioural tactics consistent with reducing energy demand via metabolic rate depression to be able to tolerate prolonged hypoxia and in some cases ceasing all activity and entering into a coma-like state (Ilacqua *et al,* 2017). Xiao *et al* (2017) investigated the role of homeostasis master-regulator 'HIF-1-α' ("hypoxia inducible factor-1-alpha") and a protein-coding gene/growth factor 'VEGF-A' ("vascular endothelial growth factor-A") in NMR-HSC's (NMR-hepatic stellate cells) exposed to HIF-1-α inhibitors before and after hypoxic contact. HIF-1-α triggers biochemical reactions caused by hypoxic stress which help decrease the negative impact it causes (Xio *et al*, 2017). According to the study, hypoxia intensified upregulation of HIF-1-α and VEGF in NMR-HSC’s and a larger rise in apoptosis was observed when the homeostasis master-regulator (HIF-1-α) was inhibited in NMR-HSC’s exposed to hypoxic conditions (Xiao *et al,* 2017). Further studies with HIF-1-α is required to help explore the underlying mechanisms which help NMR-HSCs survive hypoxic environments. Additional noteworthy phenotypic traits NMRs display are their heightened pain thresholds for noxious stimuli (i.e. acid) and diminished evasion strategies to acidic fumes, carbon dioxide and ammonia (Schuhmacher *et al*, 2018). NMRs fail to produce a thermal-hyperalgesia response due to nerve growth factor (NGF) induced TRPV1 sensitisation in sensory neurones (Gebhart & Schmidt, 2013; Omerbašić *et al,* 2016). Although the loss of thermal hyperalgesia may not be a necessary trait for NMRs, who have adapted to ecosystems close to the equator where temperatures have remained stable (Omerbašić *et al,* 2016). Lastly, NMRs are unable to thermoregulate and are effectively poikilotherms (“cold-blooded”) in low temperatures they rely on behavioural thermoregulation, thigmothermy (heat transfer via warm objects), insulation and winter dormancy (Buffenstein *et al,* 2001). Normally homeotherms (warm-blooded vertebrates) release thyroid hormone in response to low temperatures to self-regulate and produce heat, however thyroid activity and basal metabolism is reduced during exposure to colder conditions, for poikilotherms (Buffenstein *et al*, 2001).

Genome Assembly

An assembly is a structure that helps map sequencing data (reads) originating from biological material, to a presumptive re-construction of a target organism, with the assistance of automated sequencing machines and genome assembly algorithms, reads are grouped into long contiguous sequences (contigs) and contigs grouped into ordered and orientated scaffolds (supercontigs) creating the assembly (Miller et al, 2010). Genome assembly algorithms are essentially well-defined computer software programmes, specifically designed to reconstruct DNA sequences from short, fragmented reads/read-pairs. These algorithms incorporate the mathematical model '*k'* (*K*-mer) originating from mathematicians Leonhard Euler (Euler, 1736) and then adapted by Nicolaas de Bruijn (Bruijn, 1946). A *k*-mer refers to all possible substrings of length *k* that are contained in a string exactly once e.g. If *k* = 3, for the following DNA string 'A T G A T T A C T A T T A G', all possible *k*-mers are: ATG, TGA, GAT, ATT, TTA, TAC, ACT, CTA, TAT, ATT, TTA, TAG. The process of arranging sequencing data into an assembly for the first time, without any reference to existing assembled data, is called '*de novo'* assembly (Miller *et al*, 2010). The alternative approach is 'reference-guided' genome assembly, where reads are mapped against a previously assembled (often closely related) reference-genome in order to construct an alternative consensus sequence (Vezzi *et al*, 2011). No genome assembly or sequence assembling software (also known as 'assembler') is perfect, nevertheless in order to gauge how good an assembly actually is we use a quantifiable system of measurement called the N50 (Nx) metric. The N(x) metric for a genome assembly is usually calculated in increments such as 'N20', 'N50', 'N80' etc. N50, although not perfect, is the most widely used metric to describe how well an assembler has succeeded in forming together contigs and scaffolds in a *de novo* genome assembly (Mäkinen et al, 2012). N50 is defined as a 'weighted median statistic' such that 50% of the entire assembly is contained in contigs that are equal to or larger than this value (Castro & Ng, 2017). The definition above is interchangeable for other N(x) values i.e. N20 = 20% of the entire assembly, N80 = 80% of the entire assembly. The N50 statistic was first used to quantify contiguity for the initial draft Human genome, as drawbacks were seen in earlier quantification methods which either deflated or inflated statistics based on small or large contiguous sequences (Lander *et al,* 2001). This study aims to follow the Human Genome Sequencing Consortium (Lander *et al,* 2001) and use N50 (Nx metric) for genome assembly assessment/quantification.

It is crucial to produce a high-quality comprehensive NMR reference genome, to further genomic research in the underlying genetics of cancer resistance and healthy ageing. A recent example of a high-quality reference genome produced by Third-Generation SMRT (Single-Molecule Real-Time Sequencing) Pacific Biosciences (PacBio) long-reads and chromosome conformation capture (Hi-C) technology (for scaffolding), revealed new gene families and gene family expansions during *Brassica oleracea* (Cabbage) evolution, thus providing a valuable high quality reference genome which will facilitate research towards an improved crop (Lv *et al,* 2020). The NMR is a model organism for human aging research (Buffenstein, 2005). But there are gaps present in methodology and resources to study. NMRs provide an opportunity to answer central questions in biology (and further) which remain unanswered to this day, such as mechanisms of ageing, adaptions to extreme environment and resistance to cancer (Russell *et al,* 2017).

‘Shotgun sequencing’ is the most commonly used strategy in establishing an organism DNA sequence, DNA is sheared into millions of random fragments then inserted into cloning vectors (i.e. circular-plasmids) where the ends of the inserted DNA are sequenced, creating reads (Pop, 2004). Sequencers can interpret fragments by way of single-end or paired-end reading. The difference being in single-end reading, fragments are sequenced in one direction creating a sequence of base pairs, paired-end reading allows fragments to be sequenced in both directions according to the read length. Sequences aligned as read *pairs* allow more accurate read alignments and indel predictions (Grimm *et al, 2013*). At the same time, linkage information between two contigs can be provided from reads derived from the same fragment (Pop, 2004). Fragmented reads are restored to sequences using esoteric software programs called 'assemblers', which rely on complex algorithms derived from fields such as computer science, mathematics and biology (Pop, 2004). The drawback of the shotgun strategy is dealing with high copy repeats in multicellular organism genomes, which increase the risk of mis-assembly (Zhong *et al,* 2003). Published NMR genome assemblies 'HetGla_1.0' from Kim *et al,* (2011) & 'HetGla_female_1.0' from Keane *et al,* (2014) have used a shotgun whole-genome sequencing strategy with high coverage Illumina data (Lewis *et al,* 2016). Please refer to Table 1 for a summary of contiguity statistics for both NMR genome assembles.

Next-generation sequencing (NGS) technologies (Roche/454, Solexa/Illumina, Ion Torrent etc) yield millions of short reads, therefore hundreds of powerful central processing units (CPUs) and veracious assembling algorithms are needed to assemble fragmented reads into longer contigs, however without a reference genome, this can be a challenging task for any *de novo* assembler (Shendure J, 2008). The current *de novo* genome assembly algorithms include Greedy, Overlap-Layout Consensus (OLC), De Bruijn graph (DBG), String graph and Hybrid algorithms (Miller et al, 2010). All *de novo* assemblers utilise one of these algorithms, Greedy algorithms start by joining reads that are most similar to each other in a pairwise, iterative fashion until all overlaps are combined. The process ends when no more overlaps are left or all the available overlaps conflict with already constructed segments (Pop et al, 2002). The problem with this approach is that it is fundamentally local in nature and ignores long-range relations between pairs of reads, useful for solving the problem of repeats (Pop et al, 2002). This algorithm can become lodged at ‘Local maxima’ if contigs consist of reads which would help other contigs grow in size (Miller et al, 2010). OLC is another major class of assembly algorithms working in three stages, first, all the overlaps in all reads are located (‘O’) then the overlaps are laid out (‘L’) on an overlap graph and lastly, a consensus (‘C’) string is extrapolated via multiple sequence alignment (Li *et al,* 2011). OLC works better with low coverage longer reads to overcome the problem of repeats (Li *et al,* 2011). Interesting to note that OLC algorithms are being used to assemble third-generation Pacific Biosciences/PacBio and Oxford Nanopore sequences, ‘Canu’ being a recent example of an OLC based assembler (Koren *et al,* 2017). The last major class of assembly algorithm is the De Bruijn graph (DBG). DBG functions by chopping down each sequence into smaller reads called *k*-mers to first solve the problem of scalability and address differences in initial read lengths (Khan *et al,* 2018). DBGs are structured in a way where each read is represented as a node (also known as ‘vertices’) and overlaps between reads are represented by directed-edges, connecting two reads together. DBGs employ sub-strategies for genome assembly, they are known as Eulerian and Hamiltonian cycles (Compeau *et al,* 2011). In a Hamiltonian cycle, vertices represent nodes and ‘edges’ are the pairwise alignments. Following the edges in numerical order, enables one to create a genome by combining pairwise alignments between consecutive reads (Compeau et al, 2011). Important to note, each vertex is only visited once, where *k*-mers are produced from vertices. In a Eulerian cycle, vertices are (*k*-1)-mers and edges are *k*-mers. A graph is constructed by taking all edges (*k*-mers) and connecting them to ‘prefix-suffix’ nodes e.g. ATC has a prefix AT and a suffix TC therefore ‘AT’ and ‘TC’ would be separate vertices. To recover the genome sequence, all edges are transversed, shifting each consecutive *k*-mer by one position. Each edge is visited once in a Eulerian cycle (Compeau *et al,* 2011). Assemblers 'ALLPATHS' (Butler *et al,* 2008) 'ABySS' (Simpson *et al,* 2009), 'SOAPdenovo' (Li *et al*¸ 2010), 'Velvet' (Zerbino, 2008) and 'VelvetOptimiser' (Zerbino, 2010) are all based on DBG algorithms for *de novo* short read assembly. Both genome assemblies for the NMR have been created with DBG based assemblers SOAPdenovo (assembly: HetGla_1.0) and ALLPATHS (assembly:HetGla_female_1.0).

Current NMR Genomes

Two draft genome assemblies, using Illumina HiSeq 2000-based short sequence reads, currently exits for the NMR. Both genomes are publicly available from the National Center for Biotechnology Information Assembly (NCBI Assembly) database. The first NMR genome assembly 'HetGla_1.0' from Kim *et al* (2011) was sequenced on the Illumina HiSeq 2000 sequencing platform. 18 paired-end sequencing libraries with insert sizes of 170bp (base pair), 350bp, 500bp, 800bp, 2kbp (kilobase pair), 5kbp, 10kbp and 20kbp were built and sequenced creating 475G (gigabytes) of sequence, which were filtered down to 247G high quality reads with a coverage of 91.55x. The genome size of the NMR was estimated to be 2.74G. The genome was assembled using *de novo* assembler SOAPdenovo (v1.05) which uses a DBG algorithm. Approximately 98.6% of the genome was covered by at least 20 reads. The N50 value (the shortest contig/scaffolds length needed to cover 50% of the genome) for contigs and scaffolds were **19.3kb** (kilobase pair) **and 1.6Mb** (megabase pair). Total contigs and scaffolds for the assembly were 273,991 and 39,226 with a prediction of 22,561 total genes (Table 1). RNA-sequencing data, transcriptome sequencing and differential expression analysis revealed putative mechanisms for ageing and longevity in NMRs (Kim *et al,* 2011). The second NMR genome assembly 'HetGla_female_1.0' from

Keane *et al* (2014) was sequenced on the Illumina HiSeq platform. The specimen was an individual partially inbred female adult NMR originating from a captive breeding colony at the University of Rochester (USA). The assembly was built from 180bp paired-end fragment libraries, 3kb jumping libraries, 6-14kb sheared jumping libraries and 40kb FOSILLs (Williams *et al*, 2012) with a coverage of 90x. DBG based assembler ALLPATHS-LG (R38830) was used to assemble reads, using default parameters. The N50 value for contigs and scaffolds were **47.5kb** and **20.5Mb**, considerably higher than the first genome NMR assembly (Keane *et al*, 2014). Total contigs and scaffolds for the assembly were 114,653 and 4,229 with a prediction of 26,992 total genes (Table 1). The second genome assembly by Keane *et al*, (2014) revealed evidence for adaptive evolution of NMR p53 proline-rich domains (PRD) playing a role in the evolution of NMR traits and further supporting the novel anti-cancer mechanism, early contact inhibition (ECI) further suggesting high-molecular mass hyaluronan, HMMR contributes to ECI signalling (Keane *et al* 2014; Seluanov *et al*, 2009). Out of the two published NMR assemblies, HetGla_female_1.0 has higher N50 values for contigs (47.8kb) and scaffolds (20.5Mb) than HetGla_1.0 N50 values for contigs (19.3kb) and scaffolds (1.6Mb) suggesting HetGla_female_1.0 to be a better quality assembly (Table 1). 'L50' is defined as the minimum length of all contigs/scaffolds that together account for 50% of the genome (Schneeberger *et al*, 2011). HetGla_female_1.0 has a scaffold L50 of 42 while HetGla_1.0 has a scaffold L50 of 502 (Table1) advocating the female NMR genome assembly as more 'efficient'. An important point to note, both NMR genomes to date are assembled to the *scaffold* level (Table 1) where most scaffolds are 'unplaced' meaning the objects (scaffolds) do not have a chromosome, plasmid, or linkage-group assignment. To put this into comparison to other Rodentia genome assemblies, reference genome assembly for the *Mus musculus/GRCm39* (house mouse) is of *chromosome* level assembly, comprising of assembled molecules for 19 chromosomes including the *X* and *Y* chromosome, with a contig and scaffold N50 of **59.5Mb** and **106.1Mb** (www.ncbi.nlm.nih.gov/assembly/GCA_000001635.9). However, no mammalian genome is totally assembled and gap free, but it is possible using a variety of sequencing method (i.e. single molecule sequencing, chromatin conformation capture technologies) to produce high quality mammalian genomes with no pre-existing reference, as seen for the *Babulus bubalis* (water buffalo) genome assembly with contig and scaffold N50 of **22.4Mb** and **117.2Mb** (Low *et al*, 2019) and the *Capra hircus* (goat) genome assembly with contig and scaffold N50 of **26.2Mb** and **87.3Mb** (Bickhart *et al*, 2017).

Table 1.

Comparison of genome assembly contiguity statistics from the Illumina NMR sequencing projects: 'HetGla_1.0'; Kim *et al* (2011) & 'HetGla_female_1.0'; Keane *et al* (2014).

Features	**HetGla_1.0**	**HetGla_female_1.0**
Depositor:	Beijing Genomics Institute, CHN	Broad Institute, USA
NCBI Accession:	GCF_000230445.1	GCF_000247695.1
Assembly Level:	Unplaced-Scaffold	Unplaced-Scaffold
Assembler:	SOAPdenovo v1.05	ALLPATHS-LG v.r38830
Assembly Size:	2.74G	2.6G
Coverage:	X90	X90
Contig N50:	19.3kb	47.8kb
Scaffold N50:	1.6Mb	20.5Mb
Contig L50:	33,794	13,150
Scaffold L50:	502	42
Total Contig:	273,991	114,653
Total Scaffold:	39,266	4,229
Total Gene:	22,561	26,992

G = Gigabyte , Mb = Megabase (10^6), Kb = Kilobase (10^3).

Major limitations exist with the current NMR genome assemblies which limit their utility for further genomic research. Both NMR assemblies make use of shotgun whole-genome sequencing, this strategy results in fragmented assemblies with high percentages of unoccupied gaps, impeding analysis of NMR gene expression and function (Lewis *et al*, 2016). Limitations with NMR genomes exist due to the sequencing technology, next generation sequencing (NGS) technologies made by Illumina typically generate millions of short sequences per run (between 25-100bp), complex genomes with higher repeats and duplications suffer from the short read length (Trapnell & Salzberg, 2009; Alkan *et al*, 2011). A high-quality reference genome at this time does not exist for the NMR, current assemblies are of 'draft' quality, further complicated by sequence alignment of short-reads ('highest' record of N50 for contigs and scaffolds = 47.8kb and 20.5Mb from 'HetGla_female_1.0'). If for example, a high-quality reference genome existed, this would support the 'read-mapping' problem associated with *de novo* assembly, allowing one to accurately infer the positions of reads within the reference sequence (Trapnell & Salzberg, 2009). In order to create a reference genome to better confer downstream genetic analysis of the NMR, data produced from multiple sequencing technologies/methods should be incorporated to support the assembly process. For example, genome assemblies of related species rat, mouse and guinea pig have incorporated Illumina sequencing, Sanger sequencing (Sanger & Coulson, 1977) and bacterial artificial chromosome (BAC) cloning, producing less fragmented assemblies which are more complete and yield greater N50 values than current NMR genomes which only make use of one sequencing technology (Lewis *et al*, 2016). A high-quality reference genome has helped implement marker-assisted genomic selection in plant breeding programmes, which can help generate higher genetic gains in shorter breeding cycles (Benevenuto *et al*, 2019) highlighting the importance of reference genomes in eukaryotes. Another limitation with both the NMR genome are assemblies have been assembled using only DBG based *de novo* assemblers SOAPdenovo and ALLPATHS-LG, although DBG algorithms seem to be a better choice from genome assembly of larger genomes using NGS short sequencing reads (Li *et al*, 2011). However, NMR genomes generated by NGS technologies and assembled with DBG based assemblers manifest gapped assemblies with poor genome annotations, which are prone to miscalls causing frameshifts and truncated genes (Zhang *et al*, 2012). Other limitation with the current NMR genomes are concerning the effects of genetic diversity within "inbred" vs. "outbred" colonies. Both assemblies of the NMR were created from either inbred or partially inbred colonies. Institutional NMR colonies within Universities are small and can be subjected to genetic drift and the founder effect (Brekke *et al*, 2018). Laboratory-maintained NMRs may display a reduced allelic diversity compared to wild NMRs who display higher variation and natural diversity, therefore affecting the reproducibility of experiments and the validity of current NMR assembles (Brekke *et al*, 2018). Inbreeding is a widespread problem for mammalian populations which can often result in a substantial decline in fitness, known as 'inbreeding depression', due to the expression of deleterious traits, attributed to loss of heterozygosity (Pusey & Wolf, 1996). To deal with this, dispersal from the family group, and avoidance of kin as mates are used as avoidance mechanisms to lessen the effects of inbreeding eventually leading to purging of harmful/deleterious alleles (Clarke & Faulkes, 1999). This reproductive biology can have major complications when generating a *de novo* genome assembly. Because it is difficult to make highly inbred individuals, for the purpose of genome assembly, constructed/assembled alleles (i.e. haplotypes) can be highly diverse and contain large amounts of variation (Hahn *et al*, 2014). Natural allelic variation makes it difficult for standard assembly methods to distinguish between allelic sequences (sequences that code for a gene) and truly paralogous loci (gene duplications occurring at different chromosomal locations in the same organism, derived from a common ancestral gene) at low divergence levels, which can result in large-scale systematic bias in genome assembly (Hahn *et al*, 2014).

Aims & Objectives

In order for researches to fully understand the fascinating phenotypic traits presented by the NMR, an adequate genome tool (genome assembly) is required from which genomic data can be drawn out of and interpreted, in order to uncover molecular insights into the NMR. Currently NMR genomes consist of Illumina short-read sequencing data, assembled by DBG (De Bruijn graph) algorithm based *de novo* assemblers SOAPdenovo (Xie *et al, 2014*) and ALLPATHS-LG (Butler *et al*, 2008). This study aims to build an improved *de novo* genome assembly of the NMR, using published Illumina HiSeq sequencing data obtained from the National Centre for Biotechnology Information (NCBI) database and *de novo* sequence assemblers 'ABySS' (Simpson *et al*, 2009) version 2.0.2 and 'Velvet' (Zerbino, 2008) version 1.2.09. The broad aims can be separated into three main objectives:

1. To perform quality cleaning/control (QC) steps with Illumina short-read naked mole-rat (NMR) paired-end runs 'SRR363832', 'SRR363833' and single-end runs 'SRR530529', 'SRR530530', 'SRR530531', 'SRR530532' from Accession/BioProject PRJNA72441.

2. To computationally perform genome assembly with all (mentioned) NMR Illumina short-reads using De Bruijn graph based sequence assemblers ‘ABySS’ (Simpson *et al,* 2009) and ‘Velvet’ (Zerbino *et al,* 2008) and comparatively assess contiguity statistics from each assembly (such as N50 for contigs and scaffolds).

3. To evaluate the methods used to generate NMR assemblies in this study and to recommend further technical methods, in relation to sequencing, which would benefit *de novo* genome assembly of the NMR in the future.

MATERIALS AND METHODS

NMR Sequencing Data

Six Illumina HiSeq 2000 female naked mole-rat (NMR), whole genome shotgun (WGS) sequenced reads, submitted by the Broad Institute (USA) were selected, under BioProject/accession number 'PRJNA72441' and downloaded from the NCBI SRA (National Centre for Biotechnology Information Sequence Read Archive) database (ncbi.nlm.gov/sra). BioProject PRJNA72441 contained a total of thirty sequencing runs consisting of twenty-two paired-end reads and eight single-end reads. The following six BioSamples/runs, including their gigabyte (G) size, were selected for this study: 1. SRR363832 (9G), 2. SRR363833 (22G), 3. SRR530529 (58G), 4. SRR530530 (69G), 5. SRR530531 (64G), 6. SRR530532 (66G). 1 and 2 were paired-end, and 3, 4, 5 and 6 were single-ended. Every BioSample/run was available as a sperate *fastq* file. All six BioSamples/runs were generic samples from *Heterocephalus glaber.*

Sequence Data Quality Control

Fastq files SRR363832, SRR363833, SRR530529, SRR530530, SRR530531, SRR530532 were quality checked using standalone Java programme 'FastQC' (Brown *et al*, 2017) version 0.11.8. The FastQC output file (*html*) for 'SRR363832' and 'SRR363833' contained identical 'hiccups' at positions 94 to 110bp (base pair). To automate base correction and error removal, we decided to use 'AfterQC' (Chen *et al*, 2017) version 0.9.6. AfterQC is a C-Python programme for automated filtering, trimming, error correction and quality control of *fastq* data. Six *fastq* files were fed into/checked by AfterQC generating three folders, (1) ‘good’ folder - containing the cleaned data, (2) ‘bad’ folder - containing the poor quality reads, (3) ‘QC’ folder - containing a report (*html*) detailing a summary of filtered statistics for each file analysed (Table 2).

Table 2.

File sizes and read counts for NMR *fastq* files derived from AfterQC analysis.

Files:	Raw File Size	“Good” folder (AfterQC)	“Bad” folder (AfterQC)	Good Reads	Filtered Out Reads
SRR363832	9G	5.7G	2.3G	71.62%	28.37%
SRR363833	22G	13G	6.5G	66.27%	33.70%
SRR530529	58G	52G	5.9G	89.95%	10.04%
SRR530530	69G	61G	7.8G	88.86%	11.13%
SRR530531	64G	57G	6.4G	90.17%	9.82%
SRR530532	66G	60G	6.8G	89.97%	10.02%

G - Gigabytes

K-mer Size Selection

AfterQC-cleaned *fastq* files(s) SRR530529.good.fq, SRR530530.good.fq, SRR530531.good.fq and SRR530532.good.fq were examined using standalone programme ‘KmerGenie’ (Chikhi & Medvedev, 2014) v1.7051. KmerGenie is an automated size selection software that putatively estimates the ‘best’ *k*-mer size for De Bruijn graph (DBG) algorithm based *de novo* genome assemblers (Medvedev, 2014). KmerGenie predicted optimal *k*-mer sizes for all NMR *fastq* files and even predicted assembly sizes from data files (Table 3).

Table 3.

KmerGenie estimation of ‘best’ *k*-mer length and assembly size for AfterQC-cleaned *fastq* files.

AfterQC-cleaned *fastq* file:	KmerGenie ‘Best’ *K*-mer length	KmerGenie Predicted Assembly Size
SRR530529	27	1883246855 bp
SRR530530	33	1992860663 bp
SRR530531	31	1962429397 bp
SRR530532	31	1987022478 bp

bp – base pair.

Creating Manageable Files

Illumina HiSeq 2000 female NMR *fastq* files SRR530529, SRR530530, SRR530531 and SRR530532 had an original byte size of between 58G and 69G (Gigabytes) (Table 4). Due to the limits of the Linux server (32 CPUs, 1.80GHz, 125G RAM) it was decided to split all AfterQC-cleaned *fastq* files into smaller 10G files, to make genome assembly more ‘manageable’. However, there was a problem with this method, the command used to create the smaller files, split in the middle of a line (sequence) if the byte limit was reached i.e. 10G. Producing files that did not follow normal *fastq* formatting of ‘@’ in every fourth line thus rejected by the sequence assembling programme (assembler) as a legitimate *fastq* file, due to the splitting method. To get around this problem, we re-split each AfterQC-cleaned *fastq* file, based on line number by passing the `split -l` flag. But before using this method, it was necessary to find the total line count of each AfterQC-cleaned file, using the Linux/UNIX command `wc -l`, in order to come up with a suitable line number to split each file by. AfterQC-cleaned *fastq* files, up until now, had a total line count of between 774,551,052 and 909,147,492 (Table 4). AfterQC-cleaned *fastq* files SRR530529.good.fq, SRR530531.good.fq, SRR530532.good.fq were split with a line number of '-l 140,000,000' while SRR530530.good.fq (the largest file, based on byte size) was split with a line number of '-l 180,000,000' (Table 4) producing between 6 to 7 'miniature' files (xaa, xab, xac, xad, xae, xaf, +/- xag) with file byte sizes ranging between 629M (megabytes) - 13G (Table 5). It is important to note that the number after '-l' flag had to be divisible by 4, this is because a single *fastq* record consists of four lines (Identifier, Sequence, + sign and Quality scores).

Table 4.

Splitting AfterQC-cleaned *fastq* files, based on line numbers of 140 million or 180 million.

AfterQC-cleaned *fastq* file:	Original File Size	AfterQC File Size	Total Lines in AfterQC File	Split By (lines)	No. of Files Generated
SRR530529	58G	52G	774,551,052	140,000,000	6
SRR530530	69G	61G	909,147,492	180,000,000	6
SRR530531	64G	57G	855,002,584	140,000,000	7
SRR530532	66G	60G	886,947,940	140,000,000	7

G – Gigabytes.

Table 5.

File sizes of all splits from AfterQC-cleaned *fastq* files.

Split File:	AfterQC File: SRR530529.good.fq	AfterQC File: SRR530530.good.fq	AfterQC File: SRR530531.good.fq	AfterQC File: SRR530532.good.fq
1. xaa	9.3G	12G	9.3G	9.3G
2. xab	9.3G	12G	9.3G	9.3G
3. xac	9.4G	13G	9.4G	9.4G
4. xad	9.4G	13G	9.4G	9.4G
5. xae	9.4G	13G	9.4G	9.4G
6. xaf	5.0G	629M (megabyte)	9.4G	9.4G
7. xag	N/A	N/A	1.1G	3.2G

G – Gigabytes, M – megabytes.

Genome Assembly using ABySS/Velvet

The first genome assembling programme (assembler) used in this study was ‘ABySS’ (Simpson *et al,* 2009) version 2.0.2 (Table 6). ABySS is a *de novo* sequence assembler intended for short paired-end/single-end reads and large genomes (Simpson *et al,* 2009). All smaller, more manageable, single-ended split data files, derived from splitting AfterQC-cleaned NMR *fastq* file, SRR530529 (xaa, xab, xac, xad, xae, xaf), SRR530531 (xaa, xab, xac, xad, xae, xaf, xag) and SRR530532 (xaa, xab, xac, xad, xae, xaf, xag) were assembled using the ABySS single-ended assembly command ‘se=/read.fq’ (option for singe-end reads) and ‘unitigs’ command, since all reads were single-ended hence the name ‘unitigs’ (unitigs = “sequences assembled without using paired-end information”). KmerGenie’s prediction models/estimation for ‘best’ *k*-mers size (Table 3) were followed in all cases for each NMR *fastq* file assembled (i.e. …SRR530529 = *k*27, SRR530531= *k31*, SRR530532= *k*31). Contiguity statistics for each ABySS assembled file were calculated using an ABySS built-in function (`abyss-fac <unitigs.fa>`).

The second assembler used in this study was ‘Velvet’ (Zerbino, 2008) v1.2.09 (Table 6). Velvet is a *de novo* assembler, intended for short-read sequencing technologies (Illumina, Solexa, 454 etc.) based on De Bruijn graph (DBG) algorithms (Zerbino, 2008). Velvet assembles *fastq* files in two steps, first ‘Velveth’ was used, a simple hashing program (sometimes referred to as *k*-mer length) which stores and compares *k*-mers for the construction of De Bruijn graph. For step two we used ‘Velvetg’, for DBG construction, error removal and repeat resolution in order to assemble sequencing reads, also known as the core of Velvet (Zerbino, 2008). Velveth produced ‘Log’, ‘Roadmaps’ and ‘Sequences’ files which were all used by Velvetg to construct a single contigs.fa (*fasta*) file containing the sequences of contigs. Both Velveth and Velvetg, were utilised to assembly single-ended split data files derived from splitting AfterQC-cleaned NMR *fastq* file, SRR530529 (xaa, xab, xac, xad, xae, xaf), SRR530531 (xaa, xab, xac, xad, xae, xaf, xag) and SRR530532 (xaa, xab, xac, xad, xae, xaf, xag). KmerGenie's estimations for 'best' *k*-mers size (Table 3) were followed in all cases for each NMR *fastq* file assembled. Contiguity statistics for each Velvet assembled file (*fasta)* were stored in the Velvetg produced 'Log' files along with each assemblies' original parameter. It was further sought to use the in-built abyss-fac function which generated detailed contiguity statistics, based of sequences with a threshold size of 500bp (base pair) or larger.

Table 6.

Assembling programme and *K*-mer lengths for NMR *fastq* file genome assembly.

Fastq File:	Genome Assembler & *K*-mer Size	Genome Assembler & *K*-mer size
SRR530529	ABySS (se) *k*=27	Velveth Velvetg *k*=27
SRR530531	ABySS (se) *k*=31	Velveth Velvetg *k*=31
SRR530532	ABySS (se) *k*=31	Velveth Velvetg *k*=31

se = single-ended.

RESULTS

FastQC

Six *fastq* files (SRR363832, SRR363833, SRR530529, SRR530530, SRR530531, SRR530532) were quality checked using Java programme ‘FastQC’ version .0.11.8.

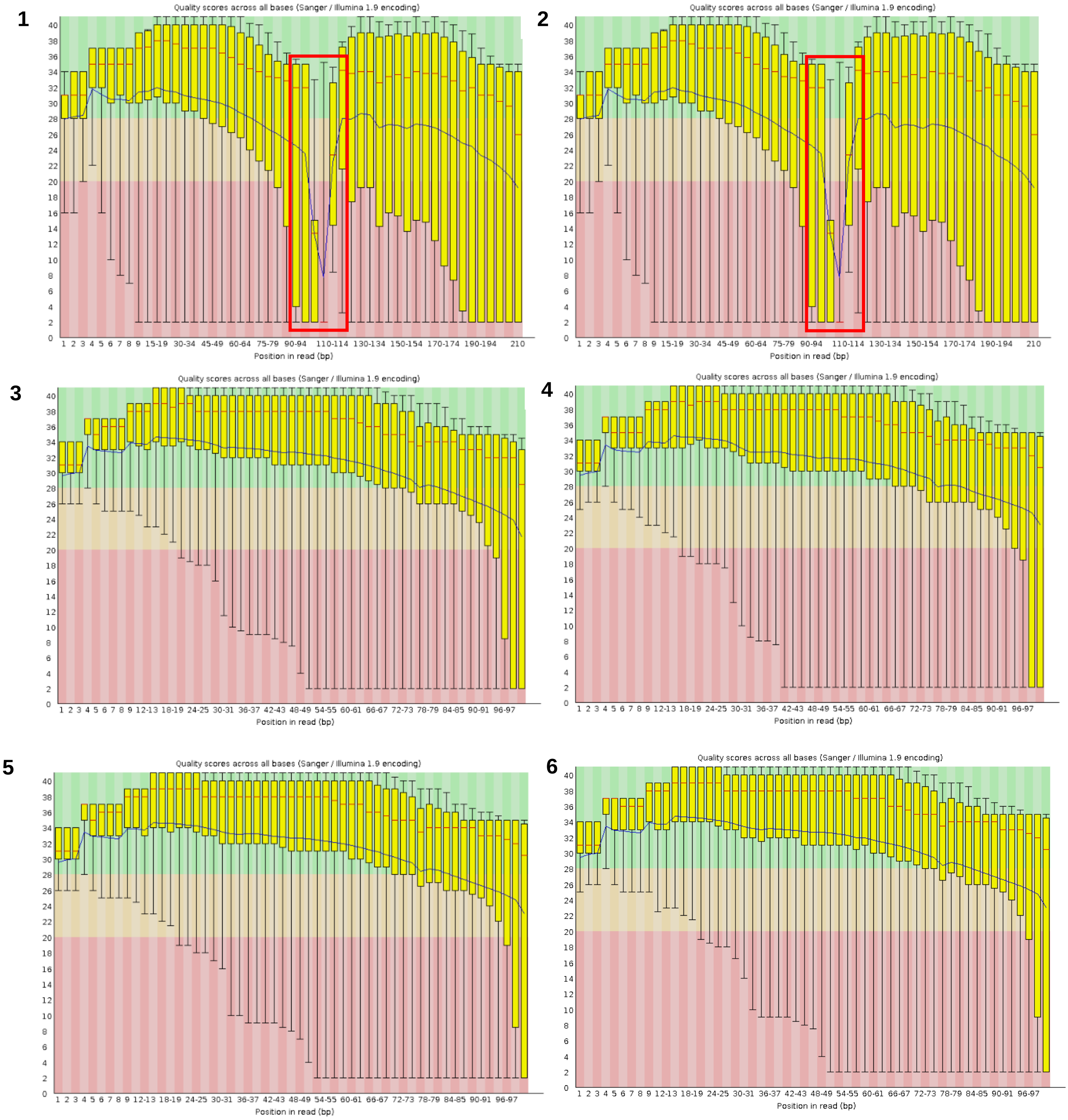


Figure 1. Per base sequence quality (scores) for *fastq* files SRR363832 (1), SRR363833 (2), SRR530529 (3), SRR530530 (4), SRR530531 (5) and SRR530532.fastq (6). X-axis represent the base positions (bp) in the read. Y-axis represent the Phred quality scores (0-40). Median values are represented by the central red line. Yellow boxes represents the inter-quartile range (25%-75%). Upper and lower whiskers represent 10% and 90% values. Mean quality is represented by the blue line. Red rectangles in SRR363832(1) & SRR363833 (2) represent ‘hiccup’ portions.

The average PHRED base quality of *fastq* file SRR363832 & SRR363833 was low (<30), with a 'hiccup' at position 94 to 110 base pair (Figure 1). These hiccups result in lower quality scores across the entire read. Hiccups could potentially be caused by signal decay, phasing, overclustering of flow cells or instrumental breakdown. The average PHRED base quality scores for *fastq* file SRR530529, SRR530530, SRR530531 and SRR530532 was high (>30), with no apparent bias and an expected decrease in quality towards the end of the read due to signal decay or phasing during the sequencing run (Figure 1).

AfterQC

Six *fastq* files, SRR363832 SRR363833 SRR530529 SRR530530 SRR530531 SRR530532 were evaluated by AfterQC, version 0.9.6 (Table 7).

Table 7.
AfterQC summary statistics for six *fastq* files.

FASTQ Files:	"Good" *fastq* file size	"Bad" *fastq* file size	Good Reads	Bad Reads
SRR363832	5.7G	2.3G	71.62%	28.37%
SRR363833	13G	6.5G	66.27%	33.73%
SRR530529	52G	5.9G	89.95%	10.05%
SRR530530	61G	7.8G	88.87%	11.13%
SRR530531	57G	6.4G	90.17%	9.83%
SRR530532	60G	6.8G	89.98%	10.02%

G – Gigabyte.

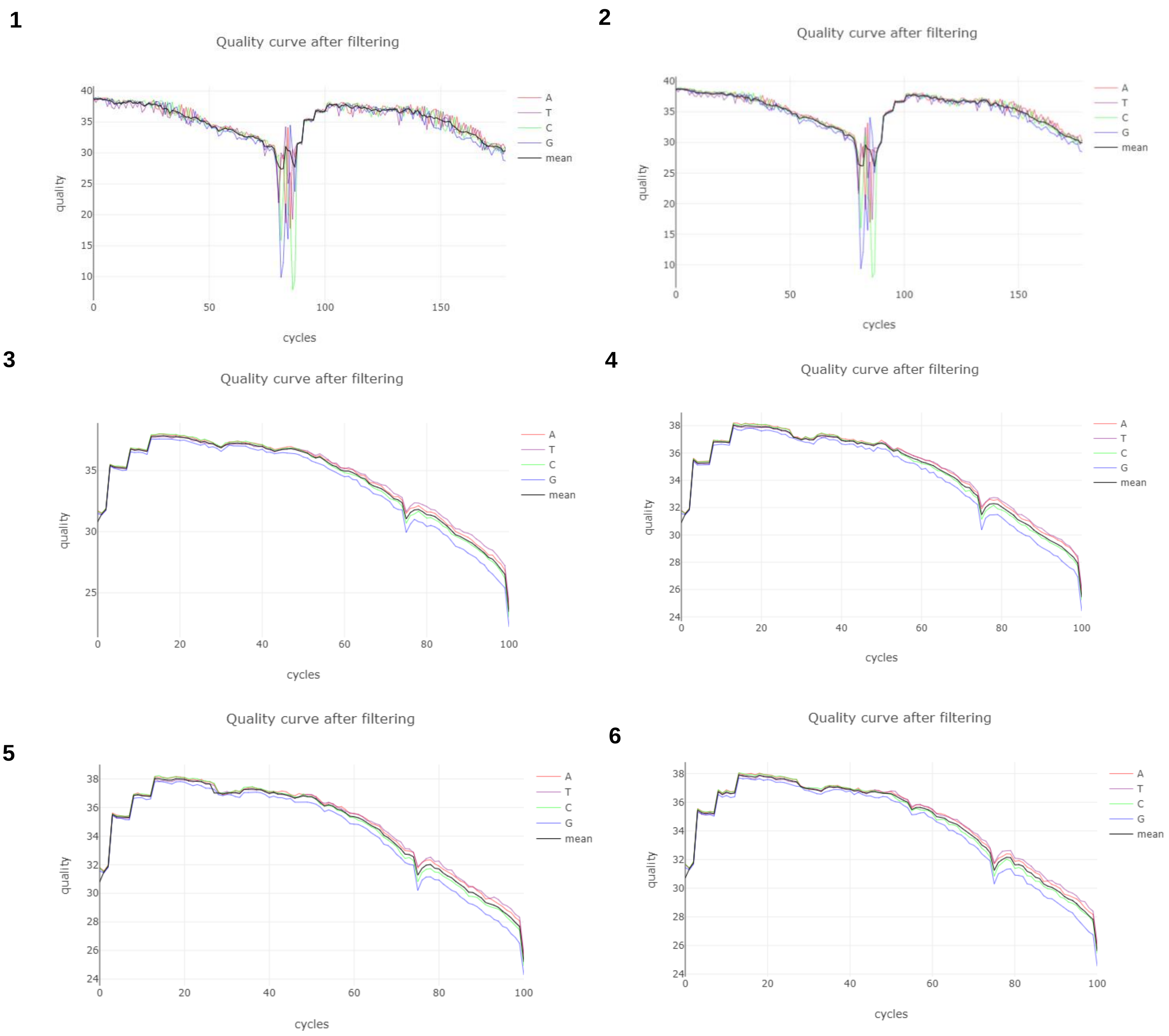


Figure 2. AfterQC quality control (QC) profiles for *fastq* files SRR363832 (1), SRR363833 (2), SRR530529 (3), SRR530530 (4), SRR530531 (5) and SRR530532 (6). X-axis represents the number of sequencing cycles; Y-axis represents the quality scores (0-38). The black line represents the mean number of base.

Approximately 9.8% - 33.7% of sequences in *fastq* files had been filtered away, leaving 90.2% - 66.3% sequences in the post *fastq* “Good” files (Table 7). Approximately 2.3G (Gigabytes) to 7.8G were “Bad” files (i.e. low quality) (Table 7). Mean base quality for *fastq* files SRR530529, SRR530530, SRR530531 and SRR530532 was high (≥30) with no apparent bias and an expected decrease in quality towards the end of the cycle due to signal decay or phasing during the sequencing run (Figure 2). Note *fastq* files SRR363832 and SRR363833 had a mean quality of ≥ 25 and identical 'hiccup' profiles from cycles 70 – 80 which could not be removed or corrected using AfterQC (Figure 2).

*K*merGenie

KmerGenie (Chikhi & Medvedev, 2014) had putatively predicted the 'best' *k*-mers for De Bruijn graph assemblers using *fastq* files SRR530529 (*k*-mer = 27), SRR530530 (*k*-mer = 33), SRR530531 (*k*-mer = 31) and SRR530532 (*k*-mer =31) (Figure 3).

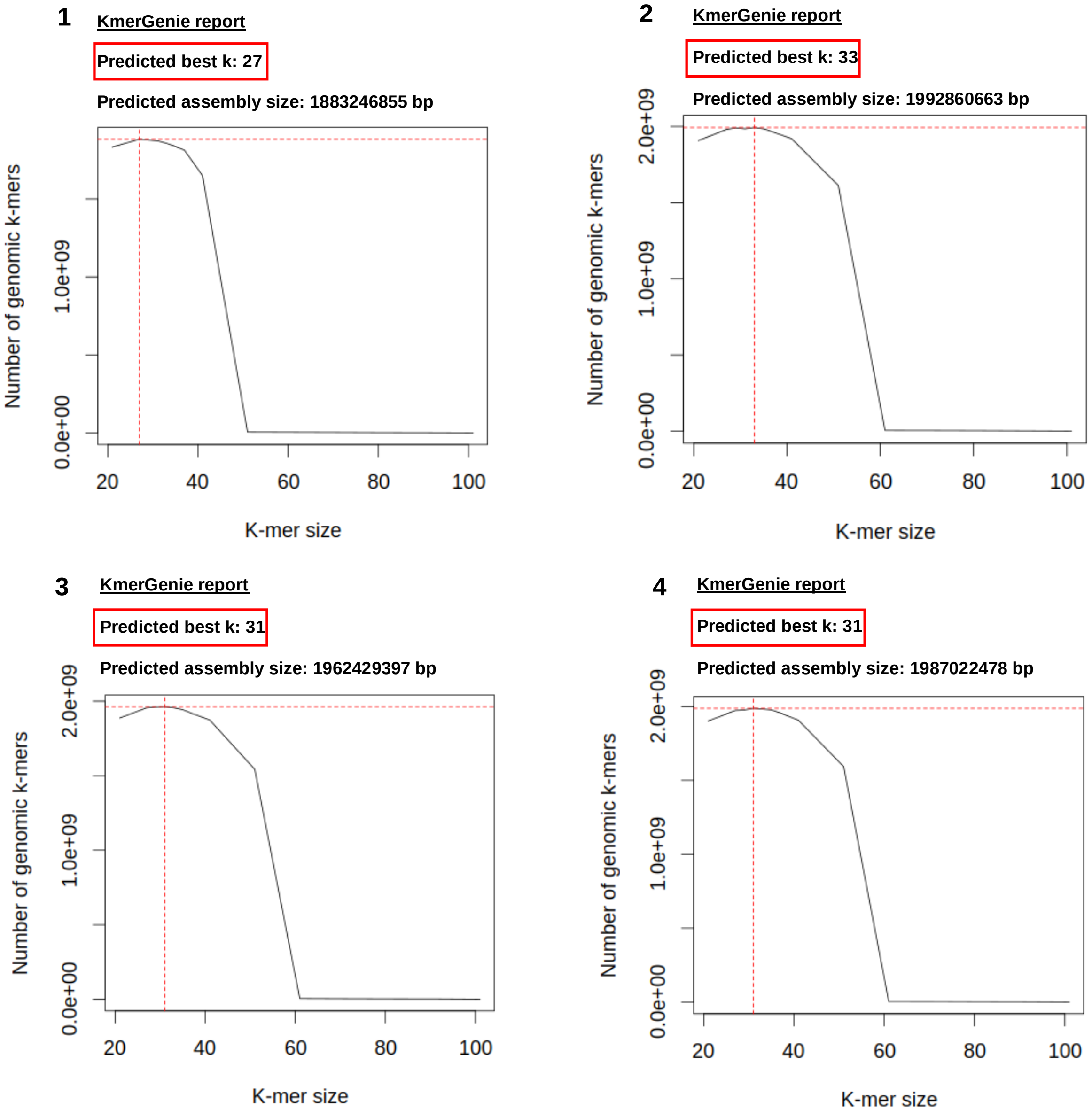


Figure 3. *K*-mer abundance histograms putatively predicting 'best' k-mers (red boxes) for De Bruijn graph (DBG) based genome assemblers using *fastq* files SRR530529 (1), SRR530530 (2), SRR530531 (3) and SRR530532 (4). X-axis represents predicted k-mer size, Y-axis represents genomic k-mers, also interpreted as the estimated genome size (in bp) when repeats are collapsed.

KmerGenie has predicted assembly sizes for *fastq* files SRR530529, SRR530530, SRR530531 and SRR530532 to be between 1.9Gbp (Giga base pairs) and 2.0Gbp (Figure 3).

Genome Assembly

Between six to seven split files (starting with 'xa'…), derived from large *fastq* data file SRR530529, SRR530531 and SRR530532 were assembled with De Bruijn graph (DBG) based *de novo* sequence assemblers 'ABySS' (2.0.2) and 'Velvet' (v1.2.09) with odd (integer) *k*-mer lengths, 27 and 31, based on KmerGenie (v1.7051) recommendations.

Table 8.

ABySS genome assembly contiguity statistics for six split *fastq* files (xaa, xab, xac, xad, xae, xaf) made from SRR530529.

SRR530529	File Size	Read-Type	*K*-mer	Assembler	N20	N50	N80	L50	Min	Max	Threshold size (bp)
xaa	681M	Single-end	27	ABySS 2.0.2	879	609	531	2257	500	5953	500
xab	682M	Single-end	27	ABySS 2.0.2	885	610	529	2210	500	7478	500
xac	681M	Single-end	27	ABySS 2.0.2	884	610	530	2227	500	5590	500
xad	670M	Single-end	27	ABySS 2.0.2	887	609	530	2061	500	4923	500
xae	680M	Single-end	27	ABySS 2.0.2	896	612	531	2185	500	5945	500
xaf	273M	Single-end	27	ABySS 2.0.2	1234	723	571	268	500	3027	500

Table 9.

Velvet genome assembly contiguity statistics for six split *fastq* files (xaa, xab, xac, xad, xae, xaf) made from SRR530529.

SRR530529	File Size	Read-Type	*K*-mer	Assembler	N20	N50	N80	L50	Min	Max	Threshold size (bp)
xaa	1.2G	Single-end	27	Velvet 1.2.09	665	565	520	6154	500	3076	500
xab	1.2G	Single-end	27	Velvet 1.2.09	665	566	521	6216	500	2413	500
xac	1.2G	Single-end	27	Velvet 1.2.09	665	565	520	6186	500	2273	500
xad	1.2G	Single-end	27	Velvet 1.2.09	666	566	520	5742	500	2438	500
xae	1.2G	Single-end	27	Velvet 1.2.09	667	564	520	5952	500	3142	500
xaf	572M	Single-end	27	Velvet 1.2.09	795	604	527	383	500	2408	500

Min – size of the smallest sequence. Max – size of the largest sequence. Threshold – number of sequences at least 500 base pairs.

Table 10.

<u>ABySS</u> genome assembly contiguity statistics for seven split *fastq* files (xaa, xab, xac, xad, xae, xaf, xag) made from SRR530531.

SRR530531	File Size	Read-Type	*K*-mer	Assembler	N20	N50	N80	L50	Min	Max	Threshold size (bp)
xaa	619M	Single-end	31	ABySS 2.0.2	987	633	535	1628	500	5961	500
xab	625M	Single-end	31	ABySS 2.0.2	968	628	534	1695	500	6096	500
xac	622M	Single-end	31	ABySS 2.0.2	969	638	536	1664	500	6163	500
xad	620M	Single-end	31	ABySS 2.0.2	965	636	537	1661	500	6096	500
xae	622M	Single-end	31	ABySS 2.0.2	950	633	534	1675	500	6473	500
xaf	630M	Single-end	31	ABySS 2.0.2	971	634	534	1749	500	6734	500
xag	19M	Single-end	31	ABySS 2.0.2	1803	1115	656	30	500	3065	500

Min – size of the smallest sequence. Max – size of the largest sequence. Threshold – number of sequences at least 500 base pairs.

Table 11.

<u>Velvet</u> genome assembly contiguity statistics for seven split *fastq* files (xaa, xab, xac, xad, xae, xaf, xag) made from SRR530531.

SRR530531	File Size	Read-Type	*K*-mer	Assembler	N20	N50	N80	L50	Min	Max	Threshold size (bp)
xaa	1.1G	Single-end	31	Velvet 1.2.09	682	568	520	4218	500	3546	500
xab	1.1G	Single-end	31	Velvet 1.2.09	684	565	519	4226	500	2604	500
xac	1.1G	Single-end	31	Velvet 1.2.09	681	568	520	4215	500	2451	500
xad	1.1G	Single-end	31	Velvet 1.2.09	687	569	520	4187	500	4361	500
xae	1.1G	Single-end	31	Velvet 1.2.09	685	568	520	4269	500	4821	500
xaf	1.1 G	Single-end	31	Velvet 1.2.09	686	567	521	4476	500	3094	500
xag	41M	Single-end	31	Velvet 1.2.09	866	641	542	31	500	1832	500

Min – size of the smallest sequence. Max – size of the largest sequence. Threshold – number of sequences at least 500 base pairs.

Contiguity statistics for all genome assemblies in this study, regardless of *k* value, had been computed on sequences of threshold size or larger (i.e. sequences 500 bp or larger). Genome assemblies of split *fastq* data files SRR530529 (xaa, xab, xac, xad, xae, xaf), SRR530531 (xaa, xab, xac, xad, xae, xaf, xag) and SRR530532 (xaa, xab, xac, xad, xae,

Table 12.

<u>ABySS</u> genome assembly contiguity statistics for seven split *fastq* files (xaa, xab, xac, xad, xae, xaf, xag) made from SRR530532.

SRR530532	File Size	Read-Type	*K*-mer	Assembler	N20	N50	N80	L50	Min	Max	Threshold size (bp)
xaa	624M	Single-end	31	ABySS 2.0.2	941	631	535	1751	500	7630	500
xab	622M	Single-end	31	ABySS 2.0.2	978	633	534	1651	500	6119	500
xac	626M	Single-end	31	ABySS 2.0.2	960	631	535	1762	500	4011	500
xad	611M	Single-end	31	ABySS 2.0.2	967	631	537	1532	500	6493	500
xae	609M	Single-end	31	ABySS 2.0.2	983	634	535	1481	500	4828	500
xaf	629M	Single-end	31	ABySS 2.0.2	985	630	535	1735	500	6119	500
xag	105M	Single-end	31	ABySS 2.0.2	1279	792	581	97	500	2909	500

Min – size of the smallest sequence. Max – size of the largest sequence. Threshold – number of sequences at least 500 base pairs.

Table 13.

<u>Velvet</u> genome assembly contiguity statistics for seven split *fastq* files (xaa, xab, xac, xad, xae, xaf, xag) made from SRR530532.

SRR530532	File Size	Read-Type	*K*-mer	Assembler	N20	N50	N80	L50	Min	Max	Threshold size (bp)
xaa	1.1G	Single-end	31	Velvet 1.2.09	687	568	521	4285	500	2424	500
xab	1.1G	Single-end	31	Velvet 1.2.09	686	567	521	4243	500	2962	500
xac	1.1G	Single-end	31	Velvet 1.2.09	684	568	521	4338	500	3007	500
xad	1.1G	Single-end	31	Velvet 1.2.09	679	566	521	4180	500	2743	500
xae	1.1G	Single-end	31	Velvet 1.2.09	677	566	520	4035	500	2829	500
xaf	1.1G	Single-end	31	Velvet 1.2.09	691	569	520	4417	500	2711	500
xag	245M	Single-end	31	Velvet 1.2.09	930	657	539	107	500	1662	500

Min – size of the smallest sequence. Max – size of the largest sequence. Threshold – number of sequences at least 500 base pairs.

xaf, xag) were assembled with *k*-mer of either 27 or 31 using DBG assemblers ABySS and Velvet. ABySS assemblies for six split *fastq* files derived from SRR530529 had created output files (*fasta*) between sizes 273M (megabytes) and 682M (Table 8). N50 values for assemblies ranged between 609 and 723, L50 values ranged between 268 and 2257 (Table 8). Velvet assemblies for six split *fastq* files derived from SRR530529 created output assembly files (*fasta*) between sizes 1.2G (gigabyte) and 572M (Table 9). N50 values for ranged between 564 and 604, while L50 values ranged between 383 and 6216 (Table 9).

ABySS genome assemblies for seven split *fastq* files derived from SRR530531 created output assembly files (*fasta*) between sizes 19M and 630M (Table 10). N50 values for these split genome assemblies ranged between 628 and 1115, L50 values ranged between 30 and 1749 (Table 10). Velvet assemblies for seven *fastq* files derived from SRR530531 created output assembly files (*fasta*) between 41M and 1.1G (Table 11). N50 values for seven (xaa ~ xag) genome assemblies ranged between 565 and 641, L50 values ranged between 31 and 4476, and the longest sequence in the assembly was 4821 bp (Table 11).

ABySS assemblies for seven split *fastq* files deriving from SRR530532 had created output assembly *(fasta)* files with sizes ranging between 105M to 629M (Table 12). N50 values for the split genome assemblies (xaa ~ xag) ranged between 630 and 792, L50 values ranged between 97 and 1762, with the longest sequence in the assembly at 7630 base pairs (Table 12). Velvet genome assemblies for seven *fastq* files derived from SRR530532 created output assembly files (*fasta*) between the ranges of 245M and 1.1G (Table 13). N50 values for the seven split genome assemblies (xaa ~ xag) ranged between 566 and 657, L50 values ranged from 107 to 4417 with the longest sequence in the assembly at 3007 base pairs (Table 13).

Based on the N50 (Nx) metric results for all split data files (SRR530529, SRR530531, SRR530532) using the ABySS (v2.0.2) *de novo* assembler, achieved greater N50 values (see ‘N50’ for Table 8-13) than the Velvet (1.2.09), advocating ABySS as the ‘better’ genome assembler for next-generation sequencing, short-read, single-ended naked mole-rat *fastq* data. ABySS outperformed Velvet in generating larger sequences (see ‘Max’ values for Tables 8-13). ABySS greatly exceeded Velvet for L50 (see ‘L50’ for Tables 8-

13). ABySS created smaller assembly output (*fasta*) files than Velvet (see ‘File Size’ for Tables 8. This could have been due to ABySS creating ‘unitigs’ and Velvet creating ‘contigs’.

An objectives of this study were to perform quality cleaning/control (QC) steps with Illumina short-read NMR data files. We have done this using FastQC to initially see the base quality of all reads and then we have used AfterQC to automate filtering, error removal and base corrections. Another objective was to computationally perform genome assembly of cleaned data files with esoteric genome assemblers ABySS and Velvet. Based on widely used genome assembly assessment contiguity statistic, N50 (Nx), genome assembler ABySS noticeably outperformed Velvet in assembling most split files derived from larger data files.

DISCUSSION

In this study, we wanted to create optimised *de novo* genome assemblies for the naked mole-rat (NMR), using previously published Illumina HiSeq 2000 female NMR reads, in order to produce an adequate genome tool that could be dissected by researches or indeed anyone who wanted to study and better understand the underlying genetics behind the fascinating phenotypic traits, presented by the NMR. This study had obtained short-read naked mole-rat (NMR) sequencing data, from the NCBI SRA website, quality checked them via FastQC (Brown et al, 2017) and AfterQC (Chen et al, 2017), putatively predicted 'best' *k*-mer lengths and assembly sizes via KmerGenie (Chikhi & Medvedev, 2014) and assembled cleaned *fastq* files, in order to compared the genome completeness and performances of two De Bruijn graph (DBG) algorithm based *de novo* assemblers 'ABySS' (Simpson *et al*, 2009) version 2.0.2 and 'Velvet' (Zerbino, 2008) version 1.2.09. Completeness and performance assessment of *de novo* genome assembly is measured by its N50 (Mäkinen et al, 2012). We discovered that the ABySS genome assembler, overall performed better, for N20, N50, N80 and L50 (Nx metrics) than the Velvet assembler for single-ended Illumina whole genome sequenced (WGS) female NMR reads SRR530529, SRR530531 and SRR530532. However further ABySS and Velvet *de novo* genome assemblies of NMR reads/runs from NCBI accession GCF_000247695.1 (HetGla_female_1.0), in particular paired-end reads optimised for parameter *k*, is needed to substantiate this claim. During this study there were some issues we faced when attempting to clean and assemble NMR *fastq* files for *de novo* genome assembly. One such issue, during the quality control (QC) phase of the study was the presence of 'hiccups' in *fastq* reads 'SRR363832' and 'SRR363833'. We previously mentioned that the hiccups could have potentially been caused by signal decay, phased sequencing, overclustering of flow cells or instrumental breakdown. Initially we tried to remove the 'hiccup' portions manually (Appendix A) but decided to 'automate' the process via. AfterQC (Chen et al, 2017), a programme designed for automatic filtering, trimming, error removing and quality control for *fastq* data. Unfortunately, the hiccups were not filtered/trimmed/removed with AfterQC, so it was naively decided to omit paired-end *fastq* files SRR363832 and SRR363833 from the study. In hindsight the solution to removing hiccups, would be to pad *fastq* files with N's ("gaps"). Another issue was concerning *fastq* file 'SRR530530', the programme KmerGenie (Chikhi & Medvedev, 2014) was used to putatively predicted best *k*-mer length for *de novo* De Bruijn graph assemblers, for that particular file, a *k*-mer of 33 was 'best' (Table 3, Figure 3). However, genome assembler Velvet (Zerbino, 2008) version 1.2.09 could not 'handle' *k*-mer values of more than 31, so naively, fastq file SRR530530 were not assembled with either the ABySS or Velvet sequence assembler. After some research, the solution would be to re-compile the Velvet assembler allowing *k*-mers of up to 127 (i.e. $ make 'MAXKMERLENGTH=127'. Otherwise default Velvet (Zerbino, 2008) maximum *k*-mer is 31. Another issue was regarding 're-combining' all smaller split *fastq* files for assembled files SRR530529 (xaa, xab, xac, xad, xae, xaf), SRR530531 (xaa, xab, xac, xad, xae, xaf, xag) and SRR530532 (xaa, xab, xac, xad, xae, xaf, xag). Initially, after genome assembly, with either ABySS or Velvet, *fasta* output files were re-combined using the Linux/UNIX command 'cat *.fa' this concatenated all files that had a file extension '.fa'. However, this would result in contigs/sequences that were not in the correct orientation. Thus, it was decided to not re-combine these smaller split files after having assembled them individually. Finally, a limitation with this study, and indeed the way genome assemblies are measured today, was the use of assembly quality quantification metric, N50 (Nx) to assess genome assemblies derived from *de novo* sequence assemblers ABySS and Velvet for *fastq* data files SRR530529, SRR530531 and SRR530532. This metric has been frequently employed to describe scale and contiguity of a genome assembly. N10 to N90 essentially describe the overall variation in contigs or scaffold lengths. But there is no clear association between these numbers and whether an assembly has any useful genetic information. Instead we can only obtain supplementary metrics to assess genome assemblies, like NG50 (Earl *et al*, 2011). NG50 (G stands for Genome) is the same as N50, however rather than compare total assembly sizes, comparisons are made against raw 'estimations' of genome sizes, which normalizes differences between differently sized assemblies (Bradnam *et al,* 2013). More recently, a study explored the use of machine learning (ML) and its potential for the production of autonomous auto-assemblers, capable of dealing with complex genomic datasets, however genome assemblers currently using ML algorithms are rare (Souza *et al*, 2019).

The current published NMR genomes, 'HetGla_1.0' and 'HetGla_female_1.0' mainly consists of next generation sequencing (NGS) short-reads, generated from Illumina HiSeq platforms, using a whole genome shotgun approach (Kim *et al,* 2011; Keane *et al*, 2014). However, problems exist with this type of technology which has resulted in fragmented assemblies with high percentages of unoccupied gaps, impeding proper analysis of NMR gene expression and function (Lewis et al, 2016). Using NGS short-reads for *de novo* genome assembly has led to four major challenges, the first one is high sequencing errors which can introduce artifacts into genome assemblies and complicate De Bruijn graphs (Liao *et al,* 2019). The next challenge is Illumina related sequencing bias caused by favouring

GC balanced regions, which may lead to unbalanced sequencing depths across a genome (Liao *et al,* 2019). The penultimate challenge of *de novo* assembly from short-reads is dealing with highly repetitive regions in a genome which is often the cause of gaps and mis-arrangements in a genome assembly. Finally, the last challenge of *de novo* assembly with short-reads is the huge computational resources needed when assembling mammalian sized genomes, *de novo* assembly demands huge computational power, significant random access memory (RAM), and large storage disk drives in order to operate. The invention of gigabase-sized genomes (< 3 Gbp) using short-read assemblers (ABySS, ALLPATHS-LG and SOAPdenovo), 512 GB of RAM and terabytes (TB) of disk space are known to take days or weeks and often require clusters/servers to run successfully (Schatz et al, 2012). Challenges in NGS assembly using short-reads can be overcome with third generation technologies such as single-molecule real-time sequencing (SMRT), generated longer reads from platforms like Pacific Biosciences (Pac Bio) and Oxford Nanopore (Roberts *et al,* 2013). SMRT sequencing takes away any of the DNA amplification step shared by older technologies (first generation/NGS) and is able to read nucleotide sequences at the single molecule level (Heather & Chain, 2016). An essential component of Pac Bio SMS machines is the zero-mode waveguide (ZMW), a tube-shaped cavity in which the DNA/polymerase complex is immobilised (Levene et al, 2003). Using DNA polymerase, fluorescent deoxyribonucleotide triphosphates (dNTPs) and the DNA library of interest, the expansion of DNA chains can be watched in real-time by monitoring the fluorescent signal produced at the base of the ZMW minus the interference from dNTPs (Heather & Chain, 2016). Alternative single molecule sequencing (SMS) technologies like Oxford Nanopore (GridION, MinION etc) work by denaturing double-stranded DNA which then ratchet through the nanopore preventing ionic flow which allows the sequence of biological bases to be implied by examining the current at each channel (Heather & Chain, 2016). There are three main contributions SMS long reads offer NGS technologies, the first one is guidance for solving highly repetitive regions, SMRT sequencing technologies produce read lengths of approximately 3000 bp to 20,000bp, roughly 30 to 200 times longer than NGS technologies (Roberts *et al,* 2013). These large read lengths can span some repetitive elements in *de novo* genome assembly thus assembling more contiguous sequences (Liao *et al,* 2019). The second contribution SMS long reads provides is assistance in the scaffolding process, long reads can scaffold through large repetitive structures and tend to have less composition bias. The last major benefit of SMS long reads, can be found in the gap closure process, using NGS reads for gap closure can result in mis-assembly rates of 20 to 500 times greater than using long reads for gap closure (Kosugi *et al,* 2015). The portability of Oxford Nanopore (ONP) MinION, allows for sequencing in the field (van Dijk *et al,* 2018). Ultra-long reads produced by ONP will one day allow complete, gapless assembly of mammalian genomes, boosting genetic research (van Dijk *et al,* 2018). But improvements are needed with this type of technology, a weakness with ONP and SMRT sequencing is their high error rate (~3%) and lower per read accuracy compared to short-read sequencing (Amarasinghe *et al,* 2020).

A New NMR Genome

On 04-AUG-2020 a new naked mole-rat (NMR) genome assembly, ‘Heter_glaber.v1.7_hic_pac’ was submitted onto the NCBI SRA database by the Institute of Zoology (IOZ), Chinese Academy of Sciences (CAS) which claimed to be ‘chromosome-level’ genome assembly (Zhou et al, 2020). The new genome assembly for the *Heterocephalus glaber* was created from pooled male and female embryonic fibroblast cells. The following technologies were used to generate this 'long-range' assembly: *in situ* Hi-C linkage information for generation of chromosome length scaffolds (Rao *et al*, 2014), PacBio (Pacific Biosciences) long reads for gap filling and published short-reads for polishing. This new genome assembly of the NMR revealed 375,829 new amino acid substitutions at disease-causing sites when compared against orthologs across 12 rodent assemblies (Zhou *et al*, 2020). The new NMR genome has provided a rich genomic source that can be data mined by researchers to uncover new mechanisms of aging and determinants of longevity in this model organism.

Table 14.

Genome assembly statistics from the Hi-C/PacBio NMR sequencing project: 'Heter_glaber.v1.7_hic_pac from Zhou *et al* (2020).

	Heter_glaber.v1.7_hic_pac
Depositor:	Chinese Academy of Sciences, CHN
NCBI Accession:	GCA_014060925.1
Assembly Level:	Scaffold
Assembler:	3D-DNA v. 180922
Sequencing Technology:	BGISEQ-500
Assembly Size:	2.89G
Coverage:	X77
Contig N50:	61.5kb
Scaffold N50:	94.5Mb
Contig L50:	13,891
Scaffold L50:	13
Total Contigs:	99,381
Total Scaffolds:	9,262
Total Gene:	29,195

The assembly size of the 'Heter_glaber.v1.7_hic_pac' was 2.89G (Gigabase), N50 for contigs and scaffolds were 61.5kb (kilobase) and 94.5Mb (Megabase) and scaffold L50 was 13 (Table 14). N50 values for the Hi-C/PacBio NMR genome assembly are higher than short-read Illumina HiSeq generated assemblies from Kim *et al* (2011) and Keane *et al* (2014) with contig N50s of 19.3kb and 47.8kb and scaffold N50s of 1.6Mb and 20.5Mb. Furthermore, scaffold L50s for Hi-

C/PacBio assembly (scaffold L50 = 13) seem more improved than Kim *et al* (2011) and Keane *et al* (2014) assemblies with L50s of 502 and 42. This assembly utilised next generation sequencing platform BGISEQ-500 (Zhu *et al,* 2018) with sequence assembling programme 3D-DNA v.180992 based on a 'novel' algorithm (Dudchenko *et al,* 2017) to generate chromosome-length scaffolds. Although this assembly has potential, it is not perfect. The perfect mammalian genome assembly does not exist, yet efforts are continuously being made which chip away at this goal.

Outstanding Questions?

1. **Can SMRT (single-molecule real-time) Oxford Nanopore sequencing improve/enhance *de novo* genome assembly for the naked mole rat?**

2. **How can we integrate machine learning (ML) to help tackle genome assembly problems of the future?**

3. **Would overlap layout consensus (OLC) based genome assemblers fair better or worse for Illumina short-read assemblies?**

4. **Is it time to move away from short-reads for genome assembly and transition into a long-read only era?**

5. **With more time, could I construct a 'truly optimised' genome assembly for the NMR, utilizing more reads, different *k-mers,* scaffolding tools?**

REFERENCES

Alkan, C., Sajjadian, S. and Eichler, E. E. (2011) Limitations of next-generation genome sequence assembly. *Nat Methods* 8 (1), 61-5.

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E. and Gouil, Q. (2020) Opportunities and challenges in long-read sequencing data analysis. *Genome Biology* 21 (1), 30.

Benevenuto, J., Ferrão, L. F. V., Amadeu, R. R. and Munoz, P. (2019) How can a high-quality genome assembly help plant breeders? *GigaScience* 8 (6).

Bennett, N. C. and Jarvis, J. U. M. (1995) Coefficients of digestibility and nutritional values of geophytes and tubers eaten by southern African mole-rats (Rodentia: Bathyergidae). *Journal of Zoology* 236 (2), 189-198.

Bens, M., Szafranski, K., Holtze, S., Sahm, A., Groth, M., Kestler, H. A., Hildebrandt, T. B. and Platzer, M. (2018) Naked mole-rat transcriptome signatures of socially suppressed sexual maturation and links of reproduction to aging. *BMC Biol* 16 (1), 77.

Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., Hastie, A. R., Chan, S., Lee, J., Lam, E. T., Liachko, I., Sullivan, S. T., Burton, J. N., Huson, H. J., Nystrom, J. C., Kelley, C. M., Hutchison, J. L., Zhou, Y., Sun, J., Crisà, A., Ponce de León, F. A., Schwartz, J. C., Hammond, J. A., Waldbieser, G. C., Schroeder, S. G., Liu, G. E., Dunham, M. J., Shendure, J., Sonstegard, T. S., Phillippy, A. M., Van Tassell, C. P. and Smith, T. P. L. (2017) Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature Genetics* 49 (4), 643-650.

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J. A., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W.-C., Corbeil, J., Del Fabbro, C., Docking, T. R., Durbin, R., Earl, D., Emrich, S., Fedotov, P., Fonseca, N. A., Ganapathy, G., Gibbs, R. A., Gnerre, S., Godzaridis, É., Goldstein, S., Haimel, M., Hall, G., Haussler, D., Hiatt, J. B., Ho, I. Y., Howard, J., Hunt, M., Jackman, S. D., Jaffe, D. B., Jarvis, E. D., Jiang, H., Kazakov, S., Kersey, P. J., Kitzman, J. O., Knight, J. R., Koren, S., Lam, T.-W., Lavenier, D., Laviolette, F., Li, Y., Li, Z., Liu, B., Liu, Y., Luo, R., MacCallum, I., MacManes, M. D., Maillet, N., Melnikov, S., Naquin, D., Ning, Z., Otto, T. D., Paten, B., Paulo, O. S., Phillippy, A. M., Pina-Martins, F., Place, M., Przybylski, D., Qin, X., Qu, C., Ribeiro, F. J., Richards, S., Rokhsar, D. S., Ruby, J. G., Scalabrin, S., Schatz, M. C., Schwartz, D. C., Sergushichev, A., Sharpe, T., Shaw, T. I., Shendure, J., Shi, Y., Simpson, J. T., Song, H., Tsarev, F., Vezzi, F., Vicedomini, R., Vieira, B. M., Wang, J., Worley, K. C., Yin, S., Yiu, S.-M., Yuan, J., Zhang, G., Zhang, H., Zhou, S. and Korf, I. F. (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience* 2 (1).

Brekke, T. D., Steele, K. A. and Mulley, J. F. (2018) Inbred or Outbred? Genetic Diversity in Laboratory Rodent Colonies. *G3 (Bethesda)* 8 (2), 679-686.

Browe, B. M., Vice, E. N. and Park, T. J. (2020) Naked Mole-Rats: Blind, Naked, and Feeling No Pain. *The Anatomical Record* 303 (1), 77-88.

Brown, J., Pirrung, M. and McCue, L. A. (2017) FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics* 33 (19), 3137-3139.

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.

Buffenstein, R. (2005) The naked mole-rat: a new long-living model for human aging research. *J Gerontol A Biol Sci Med Sci* 60 (11), 1369-77.

Buffenstein, R., Woodley, R., Thomadakis, C., Daly, T. J. M. and Gray, D. A. (2001) Cold-induced changes in thyroid function in a poikilothermic mammal, the naked mole-rat. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* 280 (1), R149-R155.

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I. A., Belmonte, M. K., Lander, E. S., Nusbaum, C. and Jaffe, D. B. (2008) ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res* 18 (5), 810-20.

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.

Chen, S., Huang, T., Zhou, Y., Han, Y., Xu, M. and Gu, J. (2017) AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. *BMC Bioinformatics* 18 (3), 80.

Chikhi, R. and Medvedev, P. (2014) Informed and automated k-mer size selection for genome assembly. *Bioinformatics* 30 (1), 31-37.

Ciszek, D. (2000) New colony formation in the “highly inbred” eusocial naked mole-rat: outbreeding is preferred. *Behavioral Ecology* 11 (1), 1-6.

Clarke, F. M. and Faulkes, C. G. (1999) Kin discrimination and female mate choice in the naked mole-rat Heterocephalus glaber. *Proc Biol Sci* 266 (1432), 1995-2002.

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.

Delaney, M. A., Nagy, L., Kinsel, M. J. and Treuting, P. M. (2013) Spontaneous histologic lesions of the adult naked mole rat (Heterocephalus glaber): a retrospective survey of lesions in a zoo population. *Vet Pathol* 50 (4), 607-21.

Delaney, M. A., Ward, J. M., Walsh, T. F., Chinnadurai, S. K., Kerns, K., Kinsel, M. J. and Treuting, P. M. (2016) Initial Case Reports of Cancer in Naked Mole-rats (Heterocephalus glaber). *Vet Pathol* 53 (3), 691-6.

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., Shamim, M. S., Machol, I., Lander, E. S., Aiden, A. P. and Aiden, E. L. (2017) De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* 356 (6333), 92-95.

Earl, D., Bradnam, K., St John, J., Darling, A., Lin, D., Fass, J., Yu, H. O., Buffalo, V., Zerbino, D. R., Diekhans, M., Nguyen, N., Ariyaratne, P. N., Sung, W. K., Ning, Z., Haimel, M., Simpson, J. T., Fonseca, N. A., Birol, İ., Docking, T. R., Ho, I. Y., Rokhsar, D. S., Chikhi, R., Lavenier, D., Chapuis, G., Naquin, D., Maillet, N., Schatz, M. C., Kelley, D. R., Phillippy, A. M., Koren, S., Yang, S. P., Wu, W., Chou, W. C., Srivastava, A., Shaw, T. I., Ruby, J. G., Skewes-Cox, P., Betegon, M., Dimon, M. T., Solovyev, V., Seledtsov, I., Kosarev, P., Vorobyev, D., Ramirez-Gonzalez, R., Leggett, R., MacLean, D., Xia, F., Luo, R., Li, Z., Xie, Y., Liu, B., Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F. J., Yin, S., Sharpe, T., Hall, G., Kersey, P. J., Durbin, R., Jackman, S. D., Chapman, J. A., Huang, X., DeRisi, J. L., Caccamo, M., Li, Y., Jaffe, D. B., Green, R. E., Haussler, D., Korf, I. and Paten, B. (2011) Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res* 21 (12), 2224-41.

Edrey, Y. H., Hanes, M., Pinto, M., Mele, J. and Buffenstein, R. (2011) Successful aging and sustained good health in the naked mole rat: a long-lived mammalian model for biogerontology and biomedical research. *Ilar j* 52 (1), 41-53.

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.

Gompertz, B. (1825) On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies. *Philosophical Transactions of the Royal Society of London* 115, 513-583.

Grimm, D., Hagmann, J., Koenig, D., Weigel, D. and Borgwardt, K. (2013) Accurate indel prediction using paired-end short reads. *BMC Genomics* 14, 132.

Hahn, M. W., Zhang, S. V. and Moyle, L. C. (2014) Sequencing, assembling, and correcting draft genomes using recombinant populations. *G3 (Bethesda)* 4 (4), 669-79.

Hamilton, J. P. and Robin Buell, C. (2012) Advances in plant genome sequencing. *The Plant Journal* 70 (1), 177-190.

Heather, J. M. and Chain, B. (2016) The sequence of sequencers: The history of sequencing DNA. *Genomics* 107 (1), 1-8.

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.

Ilacqua, A., Kirby, A. and Pamenter, M. (2017) Behavioural responses of naked mole rats to acute hypoxia and anoxia. *Biology Letters* 13, 20170545.

Keane, M., Craig, T., Alföldi, J., Berlin, A. M., Johnson, J., Seluanov, A., Gorbunova, V., Di Palma, F., Lindblad-Toh, K., Church, G. M. and de Magalhães, J. P. (2014) The Naked Mole Rat Genome Resource: facilitating analyses of cancer and longevity-related adaptations. *Bioinformatics* 30 (24), 3558-60.

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.

Kim, E. B., Fang, X., Fushan, A. A., Huang, Z., Lobanov, A. V., Han, L., Marino, S. M., Sun, X., Turanov, A. A., Yang, P., Yim, S. H., Zhao, X., Kasaikina, M. V., Stoletzki, N., Peng, C., Polak, P., Xiong, Z., Kiezun, A., Zhu, Y., Chen, Y., Kryukov, G. V., Zhang, Q., Peshkin, L., Yang, L., Bronson, R. T., Buffenstein, R., Wang, B., Han, C., Li, Q., Chen, L., Zhao, W., Sunyaev, S. R., Park, T. J., Zhang, G., Wang, J. and Gladyshev, V. N. (2011) Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature* 479 (7372), 223-227.

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.

Kosugi, S., Hirakawa, H. and Tabata, S. (2015) GMcloser: closing gaps in assemblies accurately with a likelihood-based selection of contig or long-read alignments. *Bioinformatics* 31 (23), 3733-3741.

Lander, E. S. Linton, L. M. Birren, B. Nusbaum, C. Zody, M. C. Baldwin, J. Devon, K. Dewar, K. Doyle, M. FitzHugh, W. Funke, R. Gage, D. Harris, K. Heaford, A. Howland, J. Kann, L. Lehoczky, J. LeVine, R. McEwan, P. McKernan, K. Meldrim, J. Mesirov, J. P. Miranda, C. Morris, W. Naylor, J. Raymond, C. Rosetti, M. Santos, R. Sheridan, A. Sougnez, C. Stange-Thomann, N.

Stojanovic, N. Subramanian, A. Wyman, D. Rogers, J. Sulston, J. Ainscough, R. Beck, S. Bentley, D. Burton, J. Clee, C. Carter, N. Coulson, A. Deadman, R. Deloukas, P. Dunham, A. Dunham, I. Durbin, R. French, L. Grafham, D. Gregory, S. Hubbard, T. Humphray, S. Hunt, A. Jones, M. Lloyd, C. McMurray, A. Matthews, L. Mercer, S. Milne, S. Mullikin, J. C. Mungall, A. Plumb, R. Ross, M. Shownkeen, R. Sims, S. Waterston, R. H. Wilson, R. K. Hillier, L. W. McPherson, J. D. Marra, M. A. Mardis, E. R. Fulton, L. A. Chinwalla, A. T. Pepin, K. H. Gish, W. R. Chissoe, S. L. Wendl, M. C. Delehaunty, K. D. Miner, T. L. Delehaunty, A. Kramer, J. B. Cook, L. L. Fulton, R. S. Johnson, D. L. Minx, P. J. Clifton, S. W. Hawkins, T. Branscomb, E. Predki, P. Richardson, P. Wenning, S. Slezak, T. Doggett, N. Cheng, J.-F. Olsen, A. Lucas, S. Elkin, C. Uberbacher, E. Frazier, M. Gibbs, R. A. Muzny, D. M. Scherer, S. E. Bouck, J. B. Sodergren, E. J. Worley, K. C. Rives, C. M. Gorrell, J. H. Metzker, M. L. Naylor, S. L. Kucherlapati, R. S. Nelson, D. L. Weinstock, G. M. Sakaki, Y. Fujiyama, A. Hattori, M. Yada, T. Toyoda, A. Itoh, T. Kawagoe, C. Watanabe, H. Totoki, Y. Taylor, T. Weissenbach, J. Heilig, R. Saurin, W. Artiguenave, F. Brottier, P. Bruls, T. Pelletier, E. Robert, C. Wincker, P. Rosenthal, A. Platzer, M. Nyakatura, G. Taudien, S. Rump, A. Smith, D. R. Doucette-Stamm, L. Rubenfield, M. Weinstock, K. Lee, H. M. Dubois, J. Yang, H. Yu, J. Wang, J. Huang, G. Gu, J. Hood, L. Rowen, L. Madan, A. Qin, S. Davis, R. W. Federspiel, N. A. Abola, A. P. Proctor, M. J. Roe, B. A. Chen, F. Pan, H. Ramser, J. Lehrach, H. Reinhardt, R. McCombie, W. R. de la Bastide, M. Dedhia, N. Blöcker, H. Hornischer, K. Nordsiek, G. Agarwala, R. Aravind, L. Bailey, J. A. Bateman, A. Batzoglou, S. Birney, E. Bork, P. Brown, D. G. Burge, C. B. Cerutti, L. Chen, H.-C. Church, D. Clamp, M. Copley, R. R. Doerks, T. Eddy, S. R. Eichler, E. E. Furey, T. S. Galagan, J. Gilbert, J. G. R. Harmon, C. Hayashizaki, Y. Haussler, D. Hermjakob, H. Hokamp, K. Jang, W. Johnson, L. S. Jones, T. A. Kasif, S. Kaspryzk, A. Kennedy, S. Kent, W. J. Kitts, P. Koonin, E. V. Korf, I. Kulp, D. Lancet, D. Lowe, T. M. McLysaght, A. Mikkelsen, T. Moran, J. V. Mulder, N. Pollara, V. J. Ponting, C. P. Schuler, G. Schultz, J. Slater, G. Smit, A. F. A. Stupka, E. Szustakowki, J. Thierry-Mieg, D. Thierry-Mieg, J. Wagner, L. Wallis, J. Wheeler, R. Williams, A. Wolf, Y. I. Wolfe, K. H. Yang, S.-P. Yeh, R.-F. Collins, F. Guyer, M. S. Peterson, J. Felsenfeld, A. Wetterstrand, K. A. Myers, R. M. Schmutz, J. Dickson, M. Grimwood, J. Cox, D. R. Olson, M. V. Kaul, R. Raymond, C. Shimizu, N. Kawasaki, K. Minoshima, S. Evans, G. A. Athanasiou, M. Schultz, R. Patrinos, A. Morgan, M. J. International Human Genome Sequencing, C. Whitehead Institute for Biomedical Research, C. f. G. R. The Sanger, C. Washington University Genome Sequencing, C. Institute, U. D. J. G. Baylor College of Medicine Human Genome Sequencing, C. Center, R. G. S. Genoscope and, C. U. Department of Genome Analysis, I. o. M. B. Center, G. T. C. S. Beijing Genomics Institute/Human Genome, C. Multimegabase Sequencing Center, T. I. f. S. B. Stanford Genome Technology, C. University of Oklahoma's Advanced Center for Genome, T. Max Planck Institute for Molecular, G. Cold Spring Harbor Laboratory, L. A. H. G. C. Biotechnology, G. B. G. R. C. f. *Genome Analysis, G. Scientific management: National Human Genome Research Institute, U. S. N. I. o. H. Stanford Human Genome, C. University of Washington Genome, C. Department of Molecular Biology, K. U. S. o. M. University of Texas Southwestern Medical Center at, D. Office of Science, U. S. D. o. E. and The Wellcome, T. (2001) Initial sequencing and analysis of the human genome. *Nature* 409 (6822), 860-921.

Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G. and Webb, W.

W. (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* 299 (5607), 682-6.

Lewis, K. N., Rubinstein, N. D. and Buffenstein, R. (2018) A window into extreme longevity; the circulating metabolomic signature of the naked mole-rat, a mammal that shows negligible senescence. *GeroScience* 40 (2), 105-121.

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.

Liang, S., Mele, J., Wu, Y., Buffenstein, R. and Hornsby, P. J. (2010) Resistance to experimental tumorigenesis in cells of a long-lived mammal, the naked mole-rat (Heterocephalus glaber). *Aging Cell* 9 (4), 626-35.

Liao, X., Li, M., Zou, Y., Wu, F.-X., Yi, P. and Wang, J. (2019) Current challenges and solutions of de novo assembly. *Quantitative Biology* 7 (2), 90-109.

Lv, H., Wang, Y., Han, F., Ji, J., Fang, Z., Zhuang, M., Li, Z., Zhang, Y. and Yang, L. (2020) A high-quality reference genome for cabbage obtained with SMRT reveals novel genomic features and evolutionary characteristics. *Scientific Reports* 10 (1), 12394.

McNab, B. K. (1966) The Metabolism of Fossorial Rodents: A Study of Convergence. *Ecology* 47 (5), 712-733.

Miller, J. R., Koren, S. and Sutton, G. (2010) Assembly algorithms for next-generation sequencing data. *Genomics* 95 (6), 315-27.

Miyawaki, S., Kawamura, Y., Oiwa, Y., Shimizu, A., Hachiya, T., Bono, H., Koya, I., Okada, Y., Kimura, T., Tsuchiya, Y., Suzuki, S., Onishi, N., Kuzumaki, N., Matsuzaki, Y., Narita, M., Ikeda, E., Okanoya, K., Seino, K., Saya, H., Okano, H. and Miura, K. (2016) Tumour resistance in induced pluripotent stem cells derived from naked mole-rats. *Nat Commun* 7, 11471.

Munro, D., Baldy, C., Pamenter, M. E. and Treberg, J. R. (2019) The exceptional longevity of the naked mole-rat may be explained by mitochondrial antioxidant defenses. *Aging Cell* 18 (3), e12916.

Mäkinen, V., Salmela, L. and Ylinen, J. (2012) Normalized N50 assembly metric using gap-restricted co-linear chaining. *BMC Bioinformatics* 13, 255.

Omerbašić, D., Smith, E. S., Moroni, M., Homfeld, J., Eigenbrod, O., Bennett, N. C., Reznick, J., Faulkes, C. G., Selbach, M. and Lewin, G. R. (2016) Hypofunctional TrkA Accounts for the Absence of Pain Sensitization in the African Naked Mole-Rat. *Cell Rep* 17 (3), 748-758.

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.

Padovani de Souza, K., Setubal, J. C., Ponce de Leon F. de Carvalho, A. C., Oliveira, G., Chateau, A. and Alves, R. (2019) Machine learning meets genome assembly. *Briefings in Bioinformatics* 20 (6), 2116-2129.

Park, T. J., Lewin, G. R. and Buffenstein, R. (2010) Naked Mole Rats: Their Extraordinary Sensory World. In Breed, M. D. and Moore, J. (editors) *Encyclopedia of Animal Behavior.* Oxford: Academic Press. 505-512.

Park, T. J., Reznick, J., Peterson, B. L., Blass, G., Omerbašić, D., Bennett, N. C., Kuich, P. H. J. L., Zasada, C., Browe, B. M., Hamann, W., Applegate, D. T., Radke, M. H., Kosten, T., Lutermann, H., Gavaghan, V., Eigenbrod, O., Bégay, V., Amoroso, V. G., Govind, V., Minshall, R. D., Smith, E. S. J., Larson, J., Gotthardt, M., Kempa, S. and Lewin, G. R. (2017) Fructose-driven glycolysis supports anoxia resistance in the naked mole-rat. *Science* 356 (6335), 307.

Petruseva, I. O., Evdokimov, A. N. and Lavrik, O. I. (2017) Genome Stability Maintenance in Naked Mole-Rat. *Acta Naturae* 9 (4), 31-41.

Pop, M. (2004) Shotgun Sequence Assembly. *Advances in Computers* 60, 193-248.

Pop, M., Salzberg, S. L. and Shumway, M. (2002) Genome sequence assembly: algorithms and issues. *Computer* 35 (7), 47-54.

Port, A. M., Ruth, M. R. and Istfan, N. W. (2012) Fructose consumption and cancer: is there a connection? *Curr Opin Endocrinol Diabetes Obes* 19 (5), 367-74.

Pusey, A. and Wolf, M. (1996) Inbreeding avoidance in animals. *Trends in ecology & evolution (Amsterdam)* 11 (5), 201-206.

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S. and Aiden, E. L. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159 (7), 1665-80.

Roberts, R. J., Carneiro, M. O. and Schatz, M. C. (2013) The advantages of SMRT sequencing. *Genome Biology* 14 (6), 405.

Ross-Gillespie, A., O'Riain, M. J. and Keller, L. F. (2007) Viral epizootic reveals inbreeding depression in a habitually inbreeding mammal. *Evolution* 61 (9), 2268-73.

Russell, J. J., Theriot, J. A., Sood, P., Marshall, W. F., Landweber, L. F., Fritz-Laylin, L., Polka, J. K., Oliferenko, S., Gerbich, T., Gladfelter, A., Umen, J., Bezanilla, M., Lancaster, M. A., He, S., Gibson, M. C., Goldstein, B., Tanaka, E. M., Hu, C.-K. and Brunet, A. (2017) Non-model model organisms. *BMC Biology* 15 (1), 55.

Saldmann, F., Viltard, M., Leroy, C. and Friedlander, G. (2019) The Naked Mole Rat: A Unique Example of Positive Oxidative Stress. *Oxid Med Cell Longev* 2019, 4502819.

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.

Schatz, M. C., Witkowski, J. and McCombie, W. R. (2012) Current challenges in de novo plant genome sequencing and assembly. *Genome Biology* 13 (4), 243.

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.

Schuhmacher, L. N., Callejo, G., Srivats, S. and Smith, E. S. J. (2018) Naked mole-rat acid-sensing ion channel 3 forms nonfunctional homomers, but functional heteromers. *J Biol Chem* 293 (5), 1756-1766.

Seluanov, A., Hine, C., Azpurua, J., Feigenson, M., Bozzella, M., Mao, Z., Catania, K. C. and Gorbunova, V. (2009) Hypersensitivity to contact inhibition provides a clue to cancer resistance of naked mole-rat. *Proc Natl Acad Sci U S A* 106 (46), 19352-7.

Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. *Nat Biotechnol* 26 (10), 1135-45.

Sherman, P. W., Braude, S. and Jarvis, J. U. M. (1999) Litter Sizes and Mammary Numbers of Naked Mole-Rats: Breaking the One-Half Rule. *Journal of Mammalogy* 80 (3), 720-733.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. and Birol, I. (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res* 19 (6), 1117-23.

Thermal Hyperalgesia. (2013) In Gebhart, G. F. and Schmidt, R. F. (editors) Encyclopedia of Pain. Berlin, Heidelberg: Springer Berlin Heidelberg. 3948-3948.

Trapnell, C. and Salzberg, S. L. (2009) How to map billions of short reads onto genomes. *Nat Biotechnol* 27 (5), 455-7.

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. and Thermes, C. (2018) The Third Revolution in Sequencing Technology. *Trends Genet* 34 (9), 666-681.

Vezzi, F., Cattonaro, F. and Policriti, A. (2011) e-RGA: enhanced Reference Guided Assembly of Complex Genomes. *EMBnet.journal; Vol 17, No 1: Next Generation Sequencing Data Analysis*.

Williams, L. J., Tabbaa, D. G., Li, N., Berlin, A. M., Shea, T. P., Maccallum, I., Lawrence, M. S., Drier, Y., Getz, G., Young, S. K., Jaffe, D. B., Nusbaum, C. and Gnirke, A. (2012) Paired-end sequencing of Fosmid libraries by Illumina. *Genome Res* 22 (11), 2241-9.

Xiao, B., Wang, S., Yang, G., Sun, X., Zhao, S., Lin, L., Cheng, J., Yang, W., Cong, W., Sun, W., Kan, G. and Cui, S. (2017) HIF-1α contributes to hypoxia adaptation of the naked mole rat. *Oncotarget* 8 (66), 109941-109951.

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.

Zerbino, D. R. (2010) Using the Velvet de novo assembler for short-read sequencing technologies. *Curr Protoc Bioinformatics* Chapter 11, Unit 11.5.

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.

Zhang, X., Goodsell, J. and Norgren, R. B. (2012) Limitations of the rhesus macaque draft genome assembly and annotation. *BMC Genomics* 13 (1), 206.

Zhong, L., Zhang, K., Huang, X., Ni, P., Han, Y., Wang, K., Wang, J. and Li, S. (2003) A Statistical Approach Designed for Finding Mathematically Defined Repeats in Shotgun Data and Determining the Length Distribution of Clone-Inserts. *Genomics, Proteomics & Bioinformatics* 1 (1), 43-51.

Zhou, X., Dou, Q., Fan, G., Zhang, Q., Sanderford, M., Kaya, A., Johnson, J., Karlsson, E. K., Tian, X., Mikhalchenko, A., Kumar, S., Seluanov, A., Zhang, Z. D., Gorbunova, V., Liu, X. and Gladyshev, V. N. (2020) Beaver and Naked Mole

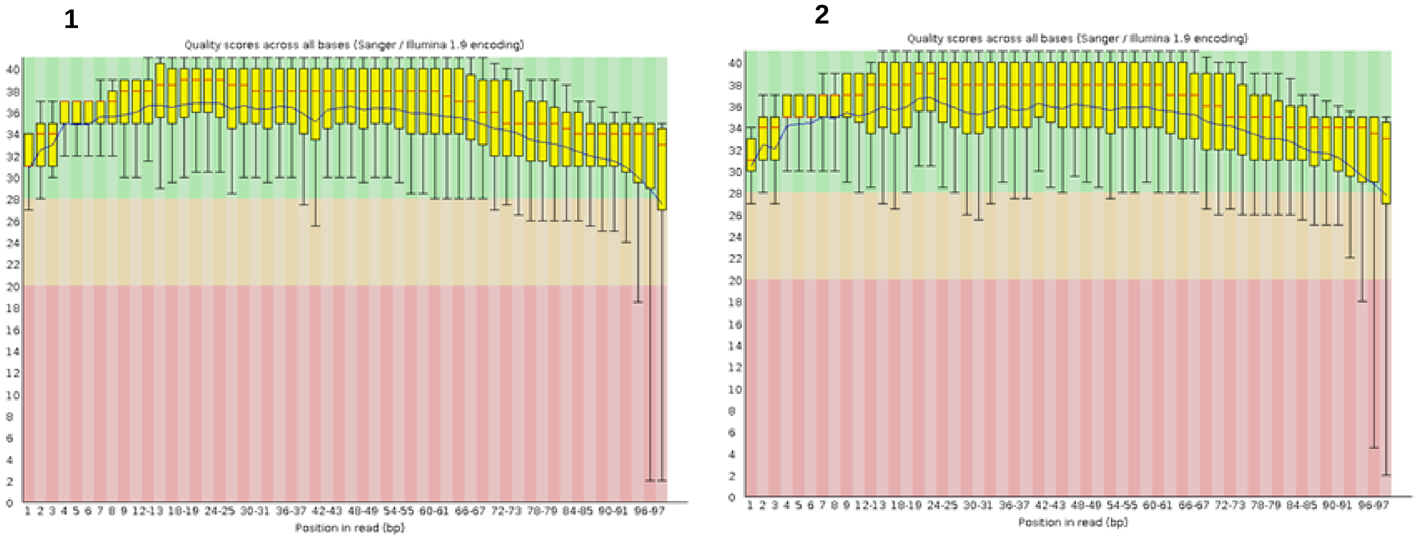
Rat Genomes Reveal Common Paths to Longevity. *Cell Reports* 32 (4), 107949.

Zhu, F.-Y., Chen, M.-X., Ye, N.-H., Qiao, W.-M., Gao, B., Law, W.-K., Tian, Y., Zhang, D., Zhang, D., Liu, T.-Y., Hu, Q.-J., Cao, Y.-Y., Su, Z.-Z., Zhang, J. and Liu, Y.-G. (2018) Comparative performance of the BGISEQ-500 and Illumina HiSeq4000 sequencing platforms for transcriptome analysis in plants. *Plant Methods* 14 (1), 69.

APPENDICES

Appendix A – Hiccup Trimming

Fastq files 'SRR363832.fastq' and 'SRR363833.fastq' contained 'hiccups' at position 94 to 110bp. Hiccup-containing paired-end reads SRR363832.fastq and SRR363833.fastq were split into two separate 'miniature' reads, hiccups were 'cut-out', using the FASTX-toolkit specifically the 'fastx-trimmer' (v.0.0.6) package, by slicing the first and last base on either sides and then re-combined using the Linux/UNIX concatenate (`cat`) command.



Appendix-Figure 1. Per base sequence quality (scores) for 'hiccup-free' *fastq* files SRR363832_filtered_trim_merged.fastq (1) and SRR363833_filtered_trim_merged.fastq (2). X-axis are the base positions (bp) in the read; Y-axis is the Phred quality scores (0-40). Median values are represented by the central red line, the yellow boxes represents the inter-quartile range (25%-75%), upper and lower whiskers represent 10% and 90% values and the mean quality is represented by the blue line.

A further FastQC analysis (Appendix-Figure 1) validates that both paired-end hiccup-containing *fastq* files 'SRR363832_filtered_trim_merged.fastq' and 'SRR363833_filtered_trim_merged.fastq' were 'hiccup-free' via the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Average base quality of paired-end reads is very good (≥30), with no apparent bias and an expected decrease in quality towards the end of the read due to signal decay or phasing during the sequencing run. But a fatal flaw exists using this method of trimming, cutting out the hiccups and stitching them back together resulted in sequences which are not ‘biologically adjacent’. Replacing the hiccups with gaps (N’s) would maintain sequence accuracy for the other bases in the read.

SUPPLEMENTARY INFORMATION

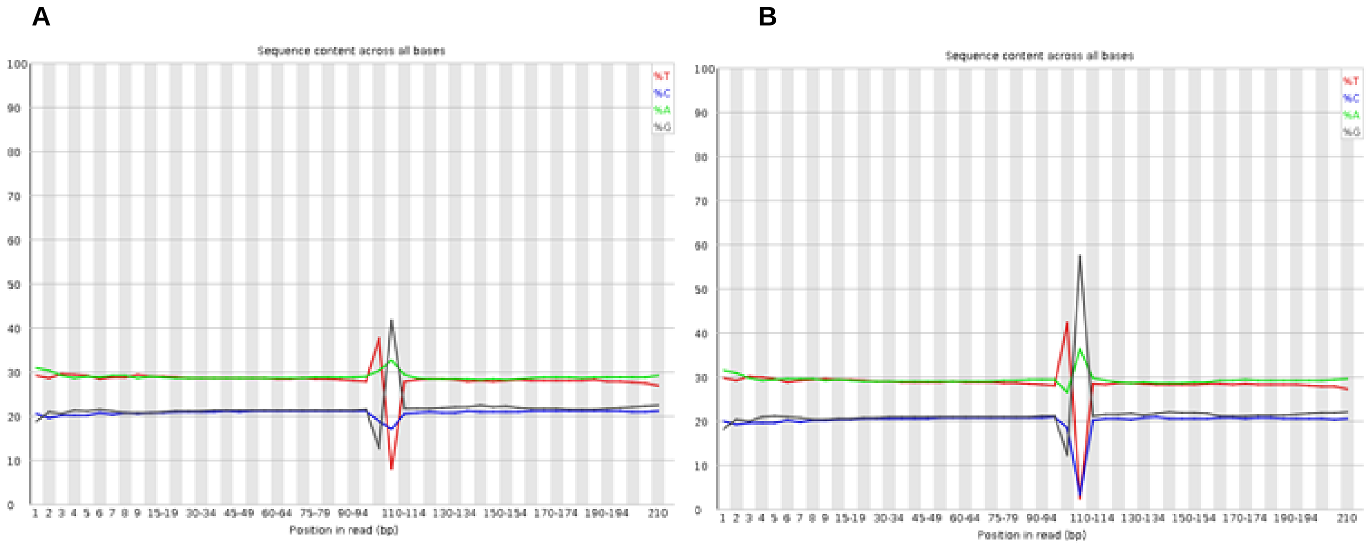
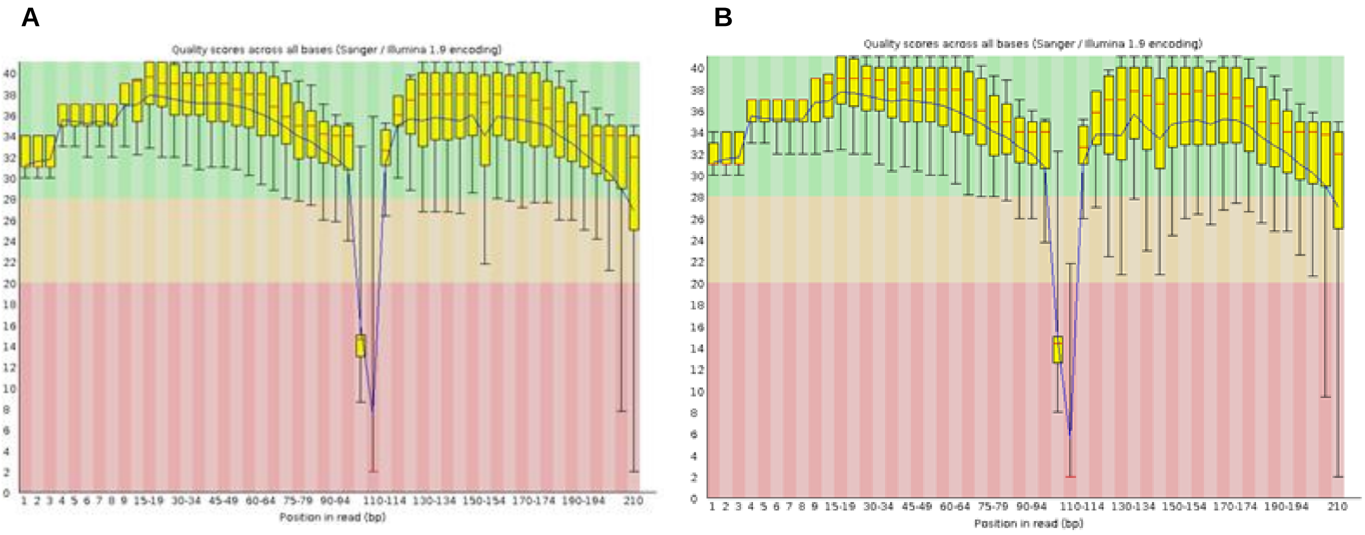


Figure S1. Per Base Sequence Content plot for *fastq* file SRR363832 (A) and SRR363833 (B). X axis are the positions in the read (bp) Y axis are the nucleotide base percentages.

(A) Out of proportion and overrepresented sequences. ‘Hiccup’ at position 94 to 112 bp.

(B) Out of proportion and overrepresented sequences. ‘Hiccup’ at position 94 to 112 bp.



<u>Figure S2. Per Base Sequence Quality plot for *fastq* file SRR363832 (A) and SRR363833 (B) after FASTX-Toolkit trimming. X axis are the positions in the read (bp) Y axis are the quality scores (0-40).</u>

(A) `fastq_quality_filter -v -Q 64 -q 20 -p 75 -i SRR363832.fastq -o SRR363832_filtered.fastq -Q33`.

(B) `fastq_quality_filter -v -Q 64 -q 20 -p 75 -i SRR363833.fastq -o SRR363833_filtered.fastq -Q33`

Key:

-v verbose report number of sequences.

-Q Determines the input quality ASCII offset (in this case 64).

-q minimum quality score to keep.

-p minimum % of bases that must have [-q] quality

-Q33 filter to a quality of (Phred+33).

= Using the fastq_quality_filter tool from the FASTX-Toolkit has resulted in an overly aggressive trim of raw reads.

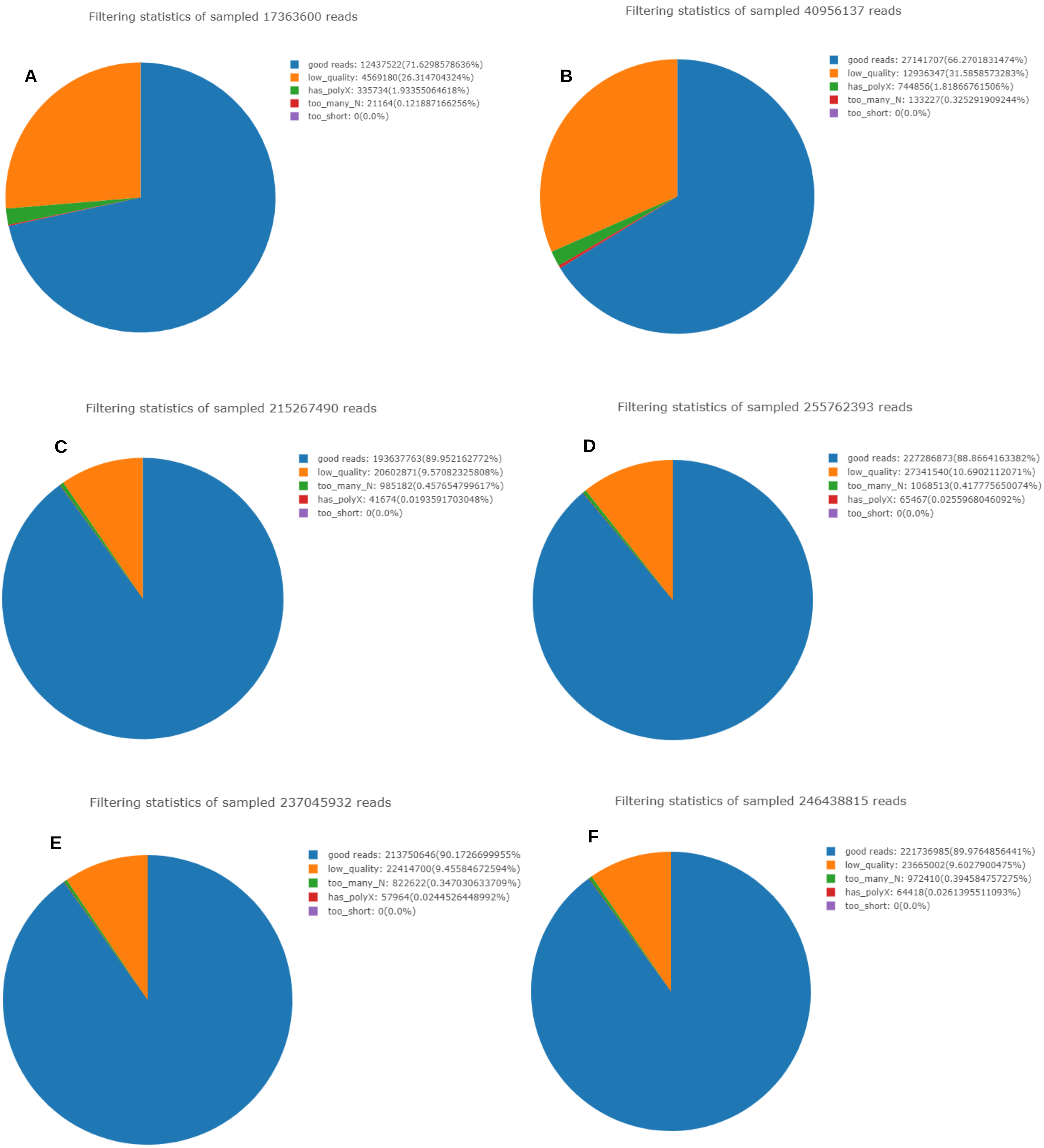


Figure S3. AfterQC filtered statistics for sample SRR363832 (A), SRR363833 (B), SRR530529 (C), SRR530530 (D), SRR530531 (E), SRR530532 (F). Good reads are coloured blue, low quality reads are coloured in orange, Ns (gap) content is coloured in green, polyX content is coloured in red (not visible) and too short reads are coloured in purple (not visible) .