République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
جامعة أبو بكر بلقايد– تلمسان
Université ABOUBEKR BELKAID – TLEMCEN
كلية علوم الطبيعة والحياة، وعلوم الأرض والكون
Faculté des Sciences de la Nature et de la Vie, et Sciences de la Terre et de l'Univers
Département de biologie

MEMOIRE

Présenté par

Benyahia Sarah

En vue de l'obtention du

Diplôme de MASTER

En Génétique

Thème

Retrieval of mammoth's (*Mammuthus primigenius*) metagenome assembled genome from environmental DNA andArctic animals and plants' abundance

Soutenu le 25/09/2022.

Devant le jury compose de :

Présidente	Brahami Nabila	MCA	Abou Bakr Belkaid Université de Tlemcen
Encadrant	Gaouar S.B.S	Professeur	Abou Bakr Belkaid Université de Tlemcen
Examinateur	Abdoallah sharaf	PhD	Czechia & Ain Shams University, Egypt

Année universitaire : 2021/2022

Acknowledgment

First, I would like to thank Allah, for letting me through all the difficulties. I have experienced his guidance day by day. None of this would have been possible without his reconcile.

I would like to acknowledge and give my warmest thanks to my supervisor Professor Gaouar Semir Bechir Suheil who made this work possible. His guidance and advice set the path of my project to carry on. I would also like to thank all of the committee members for letting my defense be an enjoyable moment, and for your brilliant comments and suggestions, thanks to you.

I would also like to give special thanks to my parents Abdelhakim and Nacera and my family as a whole Dounia, Lamia, Assia, and Fethallah for their continuous support and understanding when undertaking my research and writing my project. Your prayer for me was what sustained me this far.

Finally, I would like to thank my friends, as they were my biggest supporters and believers when needed.

Summary

List of figures…………………………………………………………………………I

List of tables…………………………………………………………...……………….III

Abbreviation list……………….……………………………………………………..V

Introduction……………………………………………………………………………..1

Chapter I: Mammoth

1. Presentation of the species……………………..………………………………………4
    1.1. History……………………………………………………………………………4
    1.2.Morphology……………………………………………………………………….4
2. Phylogenetic evolution…………………………………………………………………7
3. Distribution area……………………………………………………………………...9
4. Population dynamics………………..……………………………………………………10

4.1. Diet………………………………………………………………………………………11

4.2.Coexisting species……………………………………………………………………...12

5. Extinction…………………………………………………………………………………13
6. Available genomic data…………………………………………………………………..14

Chapter II: Bioinformatics

1. Bioinformatics and genetics……………………………………………………………17
2. Bioinformatics applications in genomics science………………………...……………17
3. Metagenomics……………………………………………………………………………18
    3.1. Metagenomics bioinformatics pipelines…………………………………...……19
    3.2.Used material……………………………………………………...……...19
        3.2.1. Hardware……………………………………………………………...19
        3.2.2. Software………………………………………………………………19
    3.3.Used Methods…………………………………………………………………20
        3.3.1. Quality control………………………………………………….…20
        3.3.2. Read counts…………………………………………………………...20
        3.3.3. Assembly……………………………………………………………...21
            3.3.3.1. Assessing assembly quality……………………………………..22
            3.3.3.2. The challenge of metagenomic assembly……………...……23
        3.3.4. Mapping reads………………………………………………………...23
        3.3.5. Binning………………………………………………………………..24

Chapter III: material and methods

1. Windows Subsystem for Linux……………………….……………………………27
2. Mammoth data selection and organization……………………………………..………27
3. Geographical regions and time interval…………………………………………………27
4. Work flow……………………………………………………………………………….31
5. Reads quality control……………………………………………………………………32
    a. FastQC……………………………………………………………………..32
    b. MultiQC……………………………………………………………………33
6. Megahit…………………………………………………………………………………..33
7. Bandage…………………………………………………………………………………34
8. MetaQuast………………………………………………………………………………..34
9. BBmap and Samtools..………………………………………………………………35

10. Anvio……………………………………………………………………………35
11. Bowtie 2………………………………………………………………………...36

Chapter IV: results

1. General statistics…………………………………………………………………38
2. Qulity control………………………………………………………………………40
2.1. Sequence quality…………………………………………………………...40
2.2. Per sequence quality………………………………………………………..40
2.3.Per base sequence content…………………………………………………..41
2.4. Per sequence GC Content…………………………………………………..41
2.5. Sequence length distribution………………………………………………41
2.6. Duplication levels…………………………………………………………..42
2.7. Adapter content……………………………………………………………42
3. Files Merge………………………………………………………………………43
4. Co-assembly………………………………………………………………………48
4.1. Bandage……………………………………………………………………..48
4.2. MetaQuast…………………………………………………………………..50
4.2.1. Contig length according to age interval……………………….…….50
4.2.2. Cumulative length according to Age interval……………………..…… 51
4.2.3. GC content according to age interval…………………………….…… 52
4.2.4. Report…………………………………………………………...…….53
5. Mapping…………………………………………………………………………..54
6. Anvi’o……………………………………………………………………………54
6.1. Databases Creating…………………………………………………………54
6.1.1. Open reading frames identification……………………………………..54
6.1.2. Assign taxonomy…………………………………...…………..55
6.2. Data profiling………………………………………………………………57
6.3. Hierarchical clustering visualization……………………………………..58
7. Taxonomic alignment………………………………...…………………..92

Discussion and Conclusion……………………………………………………..95

List of figures

Figure 1. A basic representation of the tasks in metagenomic research.	19
Figure 2. Metagenomic data possible workflows	20
Figure3. Sites distribution.	28
Figure 4. A diagram that describes the workflow followed to generate the MAGs	32
Figure 5. Mean quality scores	40
Figure 6. Per sequence quality scores	40
Figure 7. Per sequence GC content	41
Figure 8. Sequence length distribution.	42
Figure 9. Sequence duplication levels	42
Figure 10. Adapter content	43
Figure 11. Contig length according to age intervals in all four regions	51
Figure 12. Cumulative lengths according to age interval in all four regions	52
Figure 13. GC content according to age interval in all regions.	52
Figure 14. Hierarchical clustering tree of the contigs in Northwest and central Siberia 50+	59
Figure 15. Hierarchical clustering tree of the contigs in Northwest and central Siberia Early Holocene	60
Figure 16. Hierarchical clustering tree of the contigs in Northwest and central Siberia Late glacial	62
Figure 17. Hierarchical clustering tree of the contigs in Northwest and central Siberia LGM	64
Figure 18. Hierarchical clustering tree of the contigs in Northwest and central Siberia Mid-Holocene	66
Figure 19. Hierarchical clustering tree of the contigs in Northwest and central Siberia Pre-LGM	68
Figure 20. Hierarchical clustering tree of the contigs in Northwest and central Siberia Late Holocene	70
Figure 21. Hierarchical clustering tree of the contigs in North America 50+	72
Figure 22. Hierarchical clustering tree of the contigs in North America Early Holocene	74
Figure 23. Hierarchical clustering tree of the contigs in North America Late glacial	76

Figure 24. Hierarchical clustering tree of the contigs in North America LGM	78
Figure 25. Hierarchical clustering tree of the contigs in North America Pre-LGM	80
Figure 26. Hierarchical clustering tree of the contigs in Northeast Siberia Early Holocene	82
Figure 27. Hierarchical clustering tree of the contigs in Northeast Siberia Mid-Holocene	84
Figure 28. Hierarchical clustering tree of the contigs in Northeast Siberia Late glacial	85
Figure 29. Hierarchical clustering tree of the contigs in Northeast Siberia LGM	87
Figure 30. Hierarchical clustering tree of the contigs in Northeast Siberia Pre-LGM	89
Figure 31. Hierarchical clustering tree of the contigs in Northeast Siberia Pre-LGM	90

List of table

Table 1. Temporal distribution of the samples.	28
Table 2. Sites description	29
Table 3. General statistics resulting from the multiQC analysis of all of the 159 samples.	38
Table 4. samples merged to their region and age interval.	43
Table 5. Results of Megahit assemblies	48
Table 6. Bandage output based on the most suitable k-mer.	49
Table 7. summary report of the assemblies.	53
Table 8. number of genes indentified in each of the 19 databases using prodigal	54
Table 9. Matrix taxonomic reference table	55
Table 10. Taxonomically assigned bins abundance in Northwest and central Siberia 50+	59
Table 11. Bins general summary report Northwest and central Siberia 50+	60
Table 12. Bins general summary report Northwest and central Siberia Early Holocene	61
Table 13. Taxonomically assigned bins abundance in Northwest and central Siberia Late glacial	62
Table 14. Bins general summary report Northwest and central Siberia late glacial	63
Table 15. Taxonomically assigned bins abundance in Northwest and central Siberia LGM	64
Table 16. Bins general summary report Northwest and central Siberia LGM	65
Table 17. Taxonomically assigned bins abundance in Northwest and central Siberia Mid-Holocene	66
Table 18. Bins general summary report Northwest and central Siberia Mid-Holocene	67
Table 19. Taxonomically assigned bins abundance in Northwest and central Siberia Pre-LGM	68
Table 20. Bins general summary report Northwest and central Siberia Pre-LGM	69
Table 21. Taxonomically assigned bins abundance in Northwest and central Siberia Late Holocene	70
Table 22. Bins general summary report Northwest and central Siberia late Holocene	71
Table 23. Taxonomically assigned bins abundance North America 50+	72
Table 24. Bins general summary report North America 50+	73
Table 25. Taxonomically assigned bins abundance North America Early Holocene	74
Table 26. Bins general summary report North America Early Holocene	75
Table 27. Taxonomically assigned bins abundance North America Late glacial	76
Table 28. Bins general summary report North America Late glacial	77
Table 29. Taxonomically assigned bins abundance North America LGM	78
Table 30. Bins general summary report North America LGM	79
Table 31. Taxonomically assigned bins abundance North America Pre-LGM	80
Table 32. Bins general summary report North America Pre-LGM	81
Table 33. Taxonomically assigned bins abundance Northeast Siberia Early Holocene	82
Table 34. Bins general summary report Northeast Siberia Early Holocene	83

Table 35. Taxonomically assigned bins abundance Northeast Siberia Mid-Holocene	84
Table 36. Bins general summary report Northeast Siberia Mid-Holocene	84
Table 37. Bins general summary report Northeast Siberia late glacial	86
Table 38. Taxonomically assigned bins abundance Northeast Siberia LGM	87
Table 39. Bins general summary report Northeast Siberia LGM	88
Table 40. Bins general summary report Northeast Siberia Pre-LGM	89
Table 41. Taxonomically assigned bins abundance Northeast Siberia Pre-LGM 1	90
Table 42. Bins general summary report Northeast Siberia Pre-LGM	91
Table 43. Mammoth's abundance in the assigned bins	93

List of abbreviations

MSAs: multiple sequence alignment

MAGs: metagenomic assembled genomes

GNN: Graph neural network

MRCA: most recent common ancestor

eDNA: environmental DNA

aDNA: aciante DNA

INTRODUCTION

INTRODUCTION

Introduction

The woolly mammoths are thought to be the most well known of the enormous, amazing creatures that lived on Earth during the last ice age. The mammoths roamed many parts of the world until they became extinct. Around 2.5 million years ago, they spread from northern Eurasia into Western Europe, through Russia, and eventually into North America. However, the ancestors of mammoths originated in Africa about 5 million years ago**(Van der Valk and al, 2020)**. When they migrated from their original environments after evolving from their ancestors, the Elephantidae, they encountered a new set of environmental pressures. As a result, during the next two million years, they diverged into a variety of mammoth species. These species often overlapped in time and space, until, the steppe mammoth gave rise to the woolly mammoth (Mammuthus primigenius) around 700,000 years ago**(Baleka S, and al, 2021).**

The woolly mammoth's genome has proven to be environment-adaptive. Asian elephants are their closest surviving relatives, and when the genomes of these two species were compared, it was discovered that the woolly mammoth had more than 1600 gene alterations, many of which were connected to the adaption to the cold temperature, like hemoglobin structure, fat tissue, hair thickness, etc**(Chang, D and al, 2017)**. The steppe-tundra vegetation was also thought to have contributed to mammoths' ability to endure the harsh conditions in the Holocene**(Wang Y and al 2021)**.

Along with other arctic grazers, woolly mammoths created grasslands by preventing trees from dominating the plains and spreading massive amounts of nutrients over extremely long distances via their feces. These herds started to disappear at the end of the Pleistocene 11, 700 years ago, which caused the environment to change from a grasses dominant community to one with more shrubs**(Willerslev, E *et al,* 2014)**. The tundra ecosystem, which developed in their absence, is now impacted by and contributing to human-caused climate change. However, research has shown that even 10,000 years after they vanished, the tundra might become grasslands once more with the reintroduction of those grazers. The species may prevent the permafrost from thawing in the summer, thus reducing greenhouse gas emissions**(Novak BJ, 2018).**

The process is a new term called de-extinction, and today, reviving those grazers does not seem so impossible. With today's bioinformatics tools and the development of gene-editing technology, it is at least theoretically possible to alter some of the genes that give a mammoth its unique characteristics into an Asian elephant embryo to bring it back to life**(Novak BJ, 2018)**. Furthermore, it is highly likely to retrieve genomic information from specimens that are thousands of years old, which has helped us get insights on the Paleolithic period. However, the study of ancient DNA is still considered an obstacle. The degraded DNA of complex genomes of plants and animals presents a barrier due to the limiting bias in the bioinformatics classification of short fragments.

For a while now, metagenomics has been a technique to comprehend the genetic structure of a given organism and the living dynamics of extinct species. Even though this method of analysis is typically applied to microorganisms, and the majority of the software and tools utilized were created for that purpose, metagenomic has occasionally been applied to the eDNA of complex genomes. Wang Y and al 2021 presented work on the last quaternary dynamic of the arctic where they sequenced hundreds of metagenomic samples. In doing so they established the abundance and dynamics of the fauna and flora in 4 regions (North Atlantic, Northwest, and central Siberia, Northeast Siberia, and North America) and 8 age intervals (50+, Pre-LGM, LGM, Late Glacial, Early Holocene, Mid Holocene, Late Holocene, and Anthropocene) in the Arctic. However, no genome of the identified taxa was thoroughly analyzed due to the nature of the samples; being metagenomes. The interest of our work was to retrieve the mammoth's metagenome-assembled genomes by doing a co-assembly and a targeted binning on the 159 metagenomic samples where the mammoth's DNA was identified in previous studies. In addition, we aimed to get insights into the plants and animal abundance in the analyzed data as well.

CHAPTER I

CHAPTER I : MAMMOTH

Chapter 1: Mammoth

1. Presentation of the species

1.1.History

Mammoths (Mammuthus sp.) first appeared in Africa about five million years ago (Ma) and then spread throughout most of the Northern Hemisphere **(Roca AL 2015).** These animals haveamazed people across the world for ages, from cave paintings to the most recent Siberian discoveries. Their remains were among the earliest fossils to be identified, and were essential in the development of paleontology. Mammoths emerged in Africa during the late Miocene and migrated throughout Asia and Europe, and North America via Beringia, between the Middle Pliocene and the Early Pleistocene, according to fossil evidence **(Lister, A. 2005).** Mammuthus evolution during the Pleistocene is shown as a series of historically overlapping species, such as M. meridionalis (*southern mammoths*), M. trogontherii (*steppe mammoths*), M. columbi (*Columbian mammoths*), and M. primigenius (*woolly mammoths*) **(Chang, D. 2017).**

1.2.Morphology

The morphology of the mammoth is known better than other extinct arctic animals, and that is because of the discoveries made throughout the years of mammoth fossils and the preserved carcasses from the permafrost of different regions, like Siberia, Alaska, and ozocerite veins in Starunia **(Papageorgopoulou C, 2015).** These found fossils are single molars, skulls, skeletons, tusks, carcasses, and sculptures. Thus, the mammoth is considered an animal whose skeletal remains are known.

Not as everyone thinks, the woolly mammoth was actuly not a giant within the elephant family. This belief was generated because the early paleontologists, did not distinguish between different species of fossil elephants. Some of the bones of older fossil elephants (such as *Mammuthus trogontherii)* have been described to belong to *Mammuthus primigenius* **(Chang, D. 2017)**,when actually they were different species.

Body proportions

The mammoth was forced to switch to grass-eating as a way of survival. Thus, as a result of this needed adaptation, it had shorter limbs. However, its body was longer than current elephants. When the foot bears the body's weight, the cushion compresses and the phalanges expand laterally inside it, that is probably because they were shocked absorbent **(Mashchenko, E. 2013).** Those characteristics enabled the foot to expand, which allowed it to adapt to the ground's surface similarly to a minimally inflated tire. These large feet were designed to tolerate its massive body on the swampy and poorly drained ground as well. Mammoths' remains from Lyakhov Island, Sanga-Yurakh River, Beresovka River, Berelekh cemetery, and the Kirgilyakh River proved to have the best-preserved feet. It was described that in Berelekh, a left hind limb unearthed with four intact hooves about 30 mm was found **(Henryk Kubiak and al, 1982).**

The mammoth was known for its reddish-golden thick hair, which was about 30-40 mm long, and it covered the bottom half of the leg. The hair on the top of the head however was 20-25 mm long with a brown-reddish underwool underneath the surface. The tiny hoof size of many big animal grazers implied an adaptation to the hard dry substratum, which also implies less insulative plant cover and a deeper summer thaw in the Pleistocene Arctic **(Mashchenko, E and al 2013).**

The Head

The mammoth was known to have a bigger head than any other of its cousins. It was long, with curving tusks and a short neck. The mammoth head's enormous topknot was the result of large sinuses and likely masses of external hair and fat **(Mashchenko, E and al 2013).** The intermaxillary bones of the mammoth, which form the tusk's alveolae, were quite long and narrow in the middle; the bent tusks are what caused this restriction. The alveoli in modern elephants having straight tusks are almost parallel.

Hair and Coat of Fur

The mammoth coat is made up of an underwool, a longer, rougher, fluffier outer hair, and an even longer, sparser guard hair that is bristly but flexible. According to various sources, the shoulders had the longest hair (which was about 50 to 100 cm). The length of the hair on the forelimbs was 25–30 cm, whereas the length of the hair on the hindlimbs was 4–12 cm. The sides of the head's hair were 20 cm long (RYDER, 1974). Long, bristly, scratchy hairs were gathered in a clump at the tip of the tail. The lengths of the individual hairs were 94, 102, and 115 centimeters. On the sides of adult mammoths, these brownish, reddish, and occasionally yellowish hairs were most likely present. Since a single hair could support 400–700 grams, it had an average diameter of 250. Additionally, there were a few black, stiff, bristly hairs that were between 40 and 50 cm long and most likely from the neck and tail. The third type of hair was yellowish and 10–20 cm long underwool hair. Granules of a brownish-yellow pigment were discovered in the guard hairs' center but not in the underwool hair. Guard hairs had circular cross-sections, whereas underwool hairs had oval cross-sections. The lack of medullae in mammoth and rhinoceros hair, both modern and extinct, is typical of and typically unique to these species, while other animals such as bats, some sheep breeds, and some types of human hair do have hairs without medullae. The color of mammoth hair is described as being yellowish, reddish, brownish, and black. According to research (RYDER, 1974), the color of the hair from mammoth carcasses has not changed.

Skin and Fat Layer

Elephants' thick and scaly skin has earned them the nickname "pachyderms," which refers to thick, rough animals (thick skin). The mammoth's epidermis, which was the same as the epidermis of modern elephants, varied in thickness throughout the body. The average thickness was about 3 cm. The skin was thickest on the soles and thinnest on the head and legs (about 2 cm thick) (approximately 5-6 cm thick). The fat layer, however, was 8 to 9 cm thick beneath the skin. In-depth analyses of mammoth skin histology will be performed on the Kirgilyakh mammoth calf. Given that the thoracic vertebrae's neural spines were not longer in the mammoth than they are in contemporary elephants, some have hypothesized that the mammoth hump was a fat reservoir **(Mashchenko, E and al 2013).** This theory has existed for a very long time. In contrast, no evidence of a fat hump has been discovered in Arctic mammals. Such a hump is typical of dry-land animals like camels and zebu. The fat layers in large arctic mammals like musk oxen, reindeer, and polar bears are generally constant throughout the body. The animal doesn't lose body heat thanks to this uniform insulation, which is crucial in cold climates.

The Trunk

Most likely, the mammoth's trunk was even larger than modern elephant trunks. It was also covered with hair, which can still be seen on some of the mammoth trunks from Siberia's permafrost. According to (Boeskorov. 2016), the configuration of the stem's finger-like upper section and the bottom, flattened wide process, indicates outstanding adaptability to the grazing habit.

The Ears

Some researchers asserted that elephants rest in natural settings while making the recognizable punkah gesture with their ears. The fan-like veins on the medial surface of the elephant's head swell and lose heat to the atmosphere as the movement of the elephant's ears creates its local wind on either side of the elephant's head. It is obvious how much cooler the ear is than the other parts of the skin. This is most likely caused by the faster evaporation rate. Any injury or illness that makes it difficult for an elephant's ears to flap puts the animal at a significant disadvantage when it comes to controlling body temperature (Henryk Kubiak and al, 1982). The mammoth's small ears were on average 38 cm long and 29 cm wide. Like the rest of the mammoths, the ears were hair-covered. Only a few mammoth ears, including those of the Lena, Great Lyakhov Island, and Starunia mammoths, have been found in the permafrost of Siberia (Mashchenko, E and al 2013). The Asiatic elephant's ear, which is considerably smaller than the African elephant's ear, was 5–6 times smaller than the mammoth ear. Its shape was similar to that of species from Asia in general.

2. Phylogenetic evolution

According to genetic dating, the lineages that produced savanna elephants, forest elephants, Asian elephants, and woolly mammoths all started diverging from one another in Africa in the late Miocene, a period of increasing aridity on the continent. The relationship between the woolly mammoth (Mammuthus primigenius) and the modern elephant species is one question that ancient DNA research has definitively resolved. Morphological tests revealed that mammoths were more closely connected to Asian elephants than African elephants **(Enk J. et al 2016, Lister, 2001, Rohland, N 2010)**. Despite some studies placing them closer to African elephants and others placing them outside of an Asian-African elephant clade, it has been concluded that woolly mammoths are more closely linked to Asian elephants than African elephants. Early studies that relied on condensed nuclear DNA sequencing or short mtDNA sequences came up with contradictory results. The query was resolved by the creation of the entire mitochondrial genome sequences for elephants and mammoths **(Hauf J, et al 2000; Gilbert MT, et al 2007)**, were created, and the question was answered. This demonstrated that mammoth mitogenomes resembled Asian elephants more closely than African elephants. The four elephantid lineages started to split in the late Miocene, 6.8 million years ago. Mammoths and Asian elephants split up around 6.0 Mya, according to research. Loxodonta, Elephas, and Mammuthus all evolved in Africa **(Brandt AL, and al 2012),** where drier conditions may have had an impact on this **(Maglio VJ. 1973, Sanders WJ. 2010).** The relationship between the mammoth and living elephants was also investigated using nuclear sequences **(Rohland et al. 2010, Poinar et al. 2006, Miller et al. 2008),** as mitochondrial DNA may show a distinct evolutionary trend **(Roca AL. 2008).** The Asian elephant has once again been discovered to be the extinct mammoth's closest living relative (**Rohland N and al 2010**).

CHAPTER I : MAMMOTH

Between 2.6 Ma to 11.7 ka, the mammoth lineage underwent evolutionary modifications that gave rise to the southern mammoth (Mammuthus meridionalis) and steppe mammoth (Mammuthus trogontherii), which in turn produced the Columbian mammoth (Mammuthus columbi) and woolly mammoth (Mammuthus primigenius) **(Lister, A. 1996)**. Initially, it was believed that these species essentially descended from one another; however, the precise relationships between these taxa are uncertain. With several hybridization events and the potential for long-term cohabitation of distinct mammoth species, the evolutionary history suddenly seems more difficult.According to the dual source theory, the distribution of these features across time suggests two directed evolutionary sequences, such as meridionalis to columbi and trogontherii to primigenius. However, only one of the two basic mammoths could have produced the North American species given the premise of unidirectionality. Not only because meridionali's teeth cannot be reliably recognized in the North American record, but also because a few other characteristics, such as enamel thickness and the shape of the head and jaw, coincide with the basic characteristics of the molars, claim Lister and Sher (2015). These authors acknowledged that these assumptions could be problematic if unidirectionality does not hold and that despite their careful study, they were unable to rule out the possibility that M. meridionalis traveled to North America in an earlier, apparently dead-end incursion.

A more recent morphological hypothesis suggests that the steppe mammoth originated in Asia about 1.7 million years ago (Ma) from a population of southern mammoths. As early as 0.7 Ma, a second change took place in Asia when a population of steppe mammoths gave rise to the woolly mammoth **(Wei, G. 2003).** Then, these species began to colonize North America, Europe, and Asia. The fossil record suggests that the immigrants moved across Europe in waves, where they coexisted and possibly even interbred with the native populations of mammoths **(Lister, A. M and al 2005, 2001).** Until recently, it was believed that southern mammoths crossed the Atlantic and settled in North America around 1.5 million years ago when they gave rise to the Columbian, Jeffersonii, and Channel Islands pygmy mammoths (M. columbi) (M. jeffersonii). M. exilis **(Enk, J. 2016).** According to Lister and Sher **(Lister, A. M. and al. 2005),** all early North American mammoth fossils (1.5-1.3 Ma) that can be characterized morphologically are descended from steppe mammoth dispersals and those southern mammoths never migrated to North America. This hypothesis proposes that the steppe mammoths of northern Siberia were at various times the ancestors of both woolly and Columbian mammoths. This hypothesis was supported by the results of a recent study **(Enk,**

J. 2016), which examined complete mitogenomes from populations of North American mammoths). Additionally, this study found indications of hybridization and possibly male-mediated gene flow between pygmy mammoths, Columbian mammoths, Jefferson's mammoths, and North American woolly mammoths.

3. Distribution area

The earliest mammoths, which split off from other elephantids in the late Miocene, migrated from Africa to Eurasia and eventually made it to the Siberian subarctic by the Early Pleistocene. The primordial species Mammuthus meridionalis, sometimes known as the southern mammoth, crossed the Bering Land Bridge into North America during the early Irvingtonian North American Land Mammal Age (NALMA) (1.8–0.24 Ma). Mammuthus exilis, the Channel Islands pygmy mammoth, and Mammuthus Jefferson, Jefferson's mammoth, are two possible more specialized taxa that may have descended from the Columbian mammoth (Mammuthus columbi), a widespread species adapted to mid-continental parklands and grasslands. The woolly mammoth (Mammuthus primigenius), which first arrived in northern North America during the early Rancholabrean NALMA (0.125-0.011 Ma) in western Beringia, evolved from a different fundamental species, the cold-adapted steppe mammoth (Mammuthus trogontherii) (Chukotka). Then, in steppe habitats bordering the Laurentide ice sheet, woolly mammoths slowly migrated southward, eventually reaching the present-day Great Lakes region and Atlantic Coast. The Columbian mammoth (Mammuthuscolumbi), which coexisted with the periglacial woolly mammoth in Late Pleistocene North America, was larger physically and traveled to more temperate areas at that time **(Lister, A. M. 2001).**

The paleontological evidence indicates that southern mammoths predominated over the mid-latitudes of Eurasia during the later part of the MRCA (most recent common ancestor) (MRCA) of all sampled mammoths1. Around 1.7–1.6 Ma, in eastern Asia, the southern mammoth gave way to the steppe mammoth type (Wei, G. 2003). At 1.0–0.6 Ma (Lister, A. M. 2001), steppe mammoths displaced southern mammoths in Europe. As a result, we think that the paleontological split between southern and steppe mammoths contributed to the spread of clade 3 mammoths into Europe. Large samples of mammoth third molars underwent a detailed morphological analysis, and the results showed that there is insufficient proof that M. meridionalis existed in North America throughout the Pleistocene period (Lister and Sher, 2015). M. columbi, the single known predecessor, must have developed from M. trogontherii because it is widely acknowledged as an endemic to North America.

Although they could be found in almost every habitable area of Late Pleistocene North America, all mammoth species vanished by the beginning of the Holocene or quickly thereafter **(Haile et al., 2009).**

4. Population dynamics

To highlight demographic processes that took place before the advent of full-glacial conditions, ancient DNA investigations have focused on historical faunal migration, replacement, and population size variations (Binney, H. et al. 2017, Graham, R. W. et al. 2016). A mechanism of female philopatry combined with male-mediated gene flow has been proposed regarding Mammoth population structure. This mechanism is thought to be what drives the geographical pattern of morphological variation and offers a working hypothesis that can be tested in the future using mammoth nuclear genomes.

The morphological traits of late Pleistocene woolly mammoths can be seen in some North American and Eurasian mammoths. They nevertheless reflect three separate mitochondrial lineages and more than a million years of mitochondrial evolution, despite their similarities (Wang Y and al 2021). This suggests that either more extensive gene flow was occurring between geographic regions, but that dispersal was primarily, though not entirely, restricted to males, transmitting nuclear DNA and thus the majority of phenotypic characteristics, or that only limited gene flow was required to homogenize the populations across great distances.

The existence of Columbian and woolly mammoths in North America's mitochondrial clade 1, which was born and developed there, lends support to the idea of male-mediated gene flow. In this instance, roving male woolly mammoths and female Columbian mammoths probably crossed paths, giving rise to a relatively complex mitochondrial structure and phenotypic variety that included North American nominal mammoth species **(Lister & Sher 2015).** Therefore, it is likely that a comparable process accounts for both the Late Pleistocene mammoth complex in North America and the morphological change in Europe from the steppe mammoth to the woolly mammoth. Compared to those in Beringia, late Pleistocene mammoths in Europe exhibit a larger variety of variations, including both steppe and woolly mammoth morphotypes **(Pfeiffer T and al 2001, Spikings EC, and al 2007).** This is consistent with the woolly mammoths' predominantly male spread from Beringia to Europe after 200 ka, during which time they probably encountered and hybridized with clade 3 steppe mammoths.

African elephant sex differences in dispersal have been well-documented **(Tsai TS and al 2016, Hasseb A 2014).** While male-mediated dispersal is known to homogenize populations in terms of both morphological and nuclear genetic divergence, female philopatry results in a robust mitochondrial population structure in these elephants. It has been proposed that male-mediated gene flow maintains the morphological and nuclear similarities among savanna elephants, while female philopatry permits the persistence of the two profoundly divergent mitochondrial lineages **(Hasseb, A. 2014).** Several investigations looked into whether mammoths might potentially be affected by this process **(Brandt et al. 2012).** They claimed that, for a population without sex biases in dispersal, the effective population size estimated from mitochondrial DNA should be approximately 25% of the size estimated from the nuclear genome, but that, for a species that exhibits strong female philopatry, the ratio of coalescent time estimated from mitochondrial DNA to that of the nuclear loci should be higher than 0.25. They calculated the coalescent time using data from two mitochondrial genomes representing Clade 1 and Clade 2 **(Comstock KE, et al. 2002; Eggert LS, et al. 2006),** and they compared it to the coalescent time calculated using data from more than 300 nuclear loci. The concept of female philopatry and male-mediated gene flow in mammoths is supported by the fact that the ratio of the mitochondrial genome to nuclear dates was much higher than 0.25. Because giant savannah males outcompete smaller forest males, male-mediated gene flow from woolly mammoths into the steppe and Columbian mammoths exhibit a slightly different pattern than that observed from savannah elephants into forest elephants. Woolly mammoths, on the other hand, were smaller than both steppe and Columbian mammoths. However, if the steppe and Columbian mammoth morphologies resulting from the woolly mammoth combination gave a selective advantage, then sporadic interbreeding between the species would have been sufficient to propagate the woolly mammoth phenotype into those populations **(Wang Y and al 2021).**

4.1.Diet

The survival of mammoths into the Holocene in some regions is probably attributable to the persistence of the steppe–tundra vegetation of dry- and cold-adapted herbaceous plants that was present during the Pleistocene**(Wang Y and al 2021)**. This vegetation would have provided a suitable habitat for mammoths and possibly other dry land grazers such as horses, even thoughhorses were more restricted to a grassland environment and mammoth has hada greater dietary flexibility.

Likely, the persistence of the steppe-tundra flora of dry- and cold-adapted herbaceous plants that existed throughout the Pleistocene contributed to the survival of mammoths into the Holocene in some locations **(Wang Y and al 2021).** Even though horses were more confined to a grassland environment and mammoths had greater nutritional flexibility, this vegetation would have provided a good habitat for mammoths and perhaps other dry ground grazers like horses. During the Middle to Late Pleistocene (about 780 to 12 Kyr BP), Woolly Mammoths (Mammuthus primigenius) inhabited a huge area of steppe tundra that stretched from Western Europe to Asia and Beringia into North America **(Wang Y and al 2021).** Dramatic environmental change, such as the almost total disappearance of the cold, dry steppe-tundra (also known as the Mammoth steppe), as well as the extinction of cold-adapted animals like cave bears, cave hyenas, and woolly rhinoceroses, occurred toward the end of the Pleistocene. The Mammoth Steppe, a peculiar, now-extinct ecosystem dominated by huge mammal grazers, has been a source of debate for decades **(Wang Y and al 2021).**

Vereshchagin and Baryshnikov claim that the mammoth consumed willow, alder, and dwarf birch twigs in addition to sedges, grasses, and mosses during the summer. Its winter diet most likely included twigs from willow, birch, alder, larch, and pine as well as dry grasses. Gramineae made up 97.09% of the stomach contents of the Beresovka mammoth, according to pollen study results (Tikhomirov and Kupriyanova, 1954). According to pollen research, the Taimyr mammoth lived on a grassy steppe in an open area without any trees (Zaklinskaya, 1954). The stomach's flora is comparable to the deposits that surround frozen mammoth carcasses. The frozen cadavers' strong and healthy state shows that the mammoths survived well on this diet (Guil-Guerrero JL et al, 2014). The hypothesis states that an animal can tolerate more fiber the bigger it is (Guthrie). Perissodactyls and proboscidiens, for instance, may metabolize food more quickly than ruminants of equal size, allowing them to survive on a diet heavier in fiber and poorer in quality. Therefore, tiny ruminants, large ruminants, equids, and proboscidians are in order of increasing fiber tolerance. Where there is enough variation along this protein/fiber range, all of these grazers may coexist together. Overall, wide areas of tundra, steppe, prairie, parkland, or riparian habitat in grasslands were preferred by the mammoth (Martin).

4.2.Coexisting species

Among Arctic herbivores, coexistence was more prevalent than interspecies exclusion (Ritchie, M. 2002). It has been shown that some herbivores frequently appear together in both time and space. For instance, the presence of horse and mammoth eDNA is statistically strongly correlated with the prevalence of caribou, hare, and vole eDNA. In contrast, there was essentially no correlation between the spread of people across time and the occurrence of most herbivores (apart from hares). The idea of human overkill as the cause of Arctic megafaunal extinction is highly improbable given that several models purposefully overestimated the presence of humans, their largely independent distributions from megafauna, their sparseness in the high Arctic before 4 ka, and the scarcity of kill sites in archaeological records (Koch, P. L. and al 2006, Mann, D. H and al 2013). Savanna elephants and mammoths may have exhibited the highest level of male-male competition, according to the ratios of mtDNA to nuclear coalescence among different elephant species. However, field research is required to corroborate this theory for the living taxa **(Wang Y and al 2021)**.

5. Extinction

During the Middle to Late Pleistocene, Woolly Mammoths (Mammuthus primigenius) were among the species of megafauna with the greatest abundance of cold-adapted individuals (ca 780-12 Kyr BP). According to skeletal evidence, the mammoth persisted in continental Eurasia until around 10.7 ka and in Alaska until about 13.8 ka **(Mann, D. H., and al 2013)**. Recent eDNA findings, however, suggest that mammoths lived much longer than originally believed. These findings demonstrate that it persisted throughout the Early Holocene in present-day continental northeast Siberia until 7.3 0.2 ka and North America until 8.6 0.3 ka, with north-central Siberia particularly surviving as late as 3.9 0.2 ka. The longevity of the steppe-tundra vegetation in these areas is presumably what allowed mammoths to survive into the Holocene. On Wrangel Island, the last known Mammoth population can be found. Small populations were isolated by rising sea levels on St. Paul Island in the Bering Sea around 14,000 years ago and Wrangel Island in the Arctic Sea no later than 10,000 years ago **(Vartanyan et al., 2008),** both of which persisted into the Mid-Holocene.

The circumpolar range of the woolly mammoth allowed it to survive thanks to several morphological modifications. These adaptations' underlying genes have been studied. A possible adaptation to the severe high-latitude temperatures of the Pleistocene glacial periods was found in the woolly mammoth hemoglobin (Hb) protein, which was shown to have lower heat of oxygenation than elephant Hb **(Noguchi H et al, 2012)**. The late-surviving Taimyr mammoths may have interacted and coexisted with humans throughout at least 20 kyr, contradicting the human overkill hypothesis that the mammoth extinction took place just a few centuries after the first human contact **(Koch, P. L. 2006).**

CHAPTER I : MAMMOTH

Genetic and paleontological evidence suggests at least two demographic losses in their once-large populations. After the first decline, populations recovered during the Middle Pleistocene 285 Kyr BP or Eemian interglacial 130-116 Kyr BP **(Palkopoulou et al., 2013)**, and a final decline occurred around the Pleistocene–Holocene transition **(Nyström et al., 2012; 2010; Palkopoulou et al., 2013; 2015; Thomas, 2012).** Additionally, there were two different phases of the extinction: a drop in continental populations at the end of the Pleistocene and their eventual extinction, mainly in Northern Siberia **(Nyström et al., 2012; 2010; Palkopoulou et al., 2013; 2015)**. Later, during the mid-Holocene, remnant populations on St. Paul Island **(Graham et al., 2016)** and Wrangel Island **(Vartanyan et al., 2008)** both went extinct.

Recent studies on the extinction of the mammoth fauna of northern latitudes during the Late Pleistocene and Holocene (circa 12 Ka to the present) have identified the final occurrence dates and geographic ranges for several Holarctic species **(Guthrie, R. D., 2003)** and highlighted the significance of environmental changes, particularly changes in vegetation, in understanding the extinction process. A decline in the number of bison starting around 30-35 Ka **(Drummond, A., 2005)** and the local extinction of brown bears and stilt-legged horses in Alaska at around 35 Ka **(Guthrie, R. D., 2003)** both provide evidence for the significance of early, pre-glacial events in determining the late-glacial and postglacial status of these surviving taxa. Contrary to other species that have been researched thus far, western-Beringian mammoths do not adhere to the complicated pattern of population size change and turnover. This underscores how crucial it is to identify the histories of many particular species to comprehend the Late Pleistocene extinctions. A founder effect after a post-marine isotope stage (MIS) 5 expansion and minimal or negative population growth during MIS 3 were two occurrences that would have caused a more gradual loss of genetic diversity for this group. It is still unclear to what extent this sample represents mammoths throughout their range and how demographic history and extinction are related.

6. Available genomic data

In the 1999 issue of the NCBI news, it was mentioned that there were ancient DNA sequences in GenBank. Many of the ancient DNA sequences that were then accessible came from old materials like human mummies and belonged to species that are still alive today. Others came from species that have long since gone extinct, including the saber-toothed cat, ground sloth, mammoth, and mastodon. Nevertheless, they were all brief fragments made up primarily of mitochondrial sequences, which are very different from the lengths of sequences required to gather comprehensive genomic data. Six years later, with the publishing of numerous genomic sequences including a full mitochondrial genome from the extinct woolly mammoth (Mammuthus primigenius), the potential of reconstructing the genomes of ancient extinct creatures has tantalizingly neared (Wang Y and al 2021).

The groups created the most comprehensive organelle genome from an extinct organism as well as the largest sequence data ever collected (28 million base pairs) from an extinct life. The woolly mammoth mitochondrial genome and genomic sequence have both been deposited to NCBI and are accessible for searches through the integrated Entrez system, BLAST services, and NCBI's trace archive. Woolly Mammoth [Organism] AND Woolly Mammoth [Organism] AND mitochondrion [Title], however, returned a surprisingly large number of woolly mammoth nucleotide sequences. The NCBI Taxonomy services provide quick access to information on all extinct taxa, including the woolly mammoth. To obtain the entire mitochondrial genomes from RefSeq (accession NC 007596) and GenBank (accessions DQ188829 and DQ316067).

CHAPTER II

CHAPTER II : BIOINFORMATICS

Chapter II: Bioinformatics

1. Bioinformatics and genetics

In addition to biology, which typically refers to genes, DNA, RNA, or proteins, bioinformatics also uses computer science and statistics to examine and analyze biological information. Consequently, it is expected of bioinformaticians to be able to use at least one programming language. Paulien H initially introduced the term "bioinformatics" in 1970; at the time, it denoted the study of information processing in biotic systems (Hogeweg, 2011). This interdisciplinary science creates methods for storing, retrieving, and analyzing biological data.

Gene expression and related genetic information serve as the bioinformatics field's raw materials. This data was previously investigated mostly by indirect methods like linkage analysis, karyotyping, etc., but because of modern sequencing technology, computer processing power, and bioinformatics tools, it is now being examined at a completely new level.

2. Bioinformatics applications in genomics science

Any level of complexity makes it difficult to make sense of the outcomes of a genetic experiment. However, the use of bioinformatics techniques enabled scientists to address issues that had perplexed them for years. Asking the correct questions, coming up with and testing ideas, and collecting and interpreting massive amounts of data to find biological phenomena are the main concerns of both genetics and bioinformatics. Furthermore, these two fields complement one another in a variety of ways. For instance, molecular biologists and geneticists have conducted extensive research on the about 25,000 genes on the approximately 3.2 billion nucleotides in human genomes. However, the bioinformatician who performs the study needs to be knowledgeable in both computer science and genetics in order to handle and compute the genetic patterns in computers **(Wang Y and al 2021).**

Genetic analysis and sequencing technology, which are becoming faster and more affordable daily, are employed in a variety of ways to improve life quality. Comparing genes or protein sequences inside or between animals, for example, can help to clarify the evolutionary links between organisms. The science of taxonomy and its applications can benefit from these kinds of analyses as well. Additionally, genetic analysis can be used to demonstrate an organism's existence in a certain habitat.

The following aspects of bioinformatics are transformed in the context of genetic research: knowledge management and expansion, data integration and mining, mastery of genes, genomes, and genetic variation data, design and analysis of genetic studies, and determination of the genetic and genomic interface.

3. Metagenomics

The process of applying sequencing to DNA that has been directly extracted from an environmental sample (eDNA) or a collection of related samples, resulting in at least 50 Mbp of randomly sampled sequence data, is known as metagenomics. In addition, metagenomics may help to establish ideas on interactions between community members since the samples are taken from communities rather than isolated populations (Sonia Dheur et al, 2019).

Metagenomes are catalogs of the genomic DNA sequences found in environmental materials. There are still many major bottlenecks in metagenomics, including the extraction and purification of high-quality DNA. This problem is made worse by the fact that there isn't a single extraction technique that works for all environmental samples. Small amounts of DNA are produced by low-biomass samples, which could not be enough for library creation. Small yields of ambient DNAs have been subjected to whole-genome amplification to produce microgram quantities for sequencing. This method has the significant benefit of being able to handle and maintain single-stranded DNA, which is crucial for viral samples.

This barrier has been broken because of developments in sequencing, computing capacity, and bioinformatic tools, which now make it possible to reconstruct species-level prokaryotic genomes from microbial communities (MAGs). The prokaryotic Tree of Life has been expanded and inhabited by uncultivated species as a result of MAGs, and the most recent investigations have revealed thousands of additional genomes, which provide fresh and pertinent functional and evolutionary insights (Mirdita M, and al 2021). However, there was a belief that this method could not be used for eukaryotes due to their bigger genomes, which are divided into numerous chromosomes, and the numerous non-coding and repetitive areas, which complicate the use of the available bioinformatics tools. Only a few instances of eukaryotic MAGs have been documented, frequently coming from groups with a low level of diversity.

CHAPTER II : BIOINFORMATICS

3.1.Metagenomics bioinformatics pipeline

Metagenomics data is used by scientists to examine the DNA of a community in a particular sample, for example, to determine its taxonomic composition, the genes present throughout the community or any indications of selection over a particular taxon. There are similar processes required to work with metagenomics data regardless of the specific analyses. (figure 1).

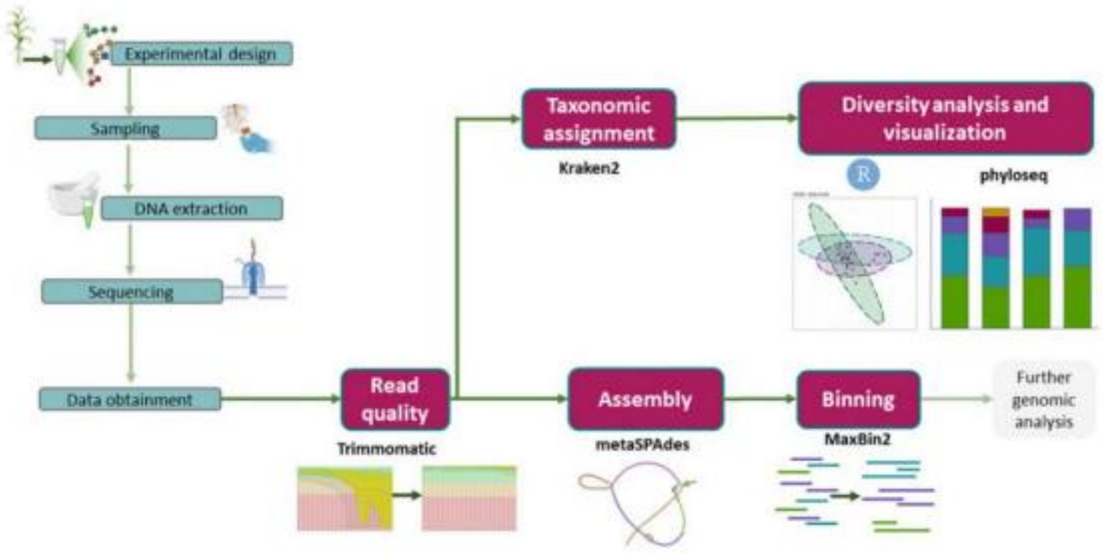


Figure 1. A basic representation of the tasks in metagenomic research. **(Garfias-Gallegos, D. *et al.* 2022).**

3.2.Used material

3.2.1. Hardware

To process the data utilized in the pipeline described here, a system with at least 64 GB of RAM, 300 GB of hard drive, and a multicore CPU (with at least six logic cores) are needed. Most laptops and PCs do not meet these criteria, hence it is typically advised to rent a virtual machine to get the necessary hardware **(Garfias-Gallegos, D. *et al.* 2022)**.

3.2.2. Software

A 64-bit Linux/Unix systems command line or Windows V10 with a Windows subsystem for Linux installed can execute the majority of pipelines.

For bash: bash provides the quick and flexible usage of algorithms made to evaluate metagenomic data, producing useful outputs for additional data analysis. SRA toolkit, FastQC, Trimmomatic, Kraken2, MaxBin2, metaSPAdes, megahit, Kraken-boom, and CheckM are a few examples of open-source applications that may be used on bash terminals.

For R: R is a popular data science program that enables users to analyze Shell outputs and produce publishable results by using community-designed packages. Phyloseq and ggplot2 are the most used R packages in the metagenomic pipeline.

3.3.used Methods

3.3.1. Quality control

The initial step in data processing must be to assess the quality of the data, followed by the removal of low-quality sequences from the sequenced reads. Users must also have an organized file system. Numerous factors could contribute to poor quality. It is expected that the first thousand reads (or more), regardless of the sequencing method employed, will be of relatively low quality and commonly contain "N"s. The Illumina software was unable to create a baseball for this base, as shown by an "N". As a result, views and attributes must govern the readings. Typically, Trimmomatic is used to filter low-quality reads and trim low-quality bases, whereas FastQC is used to display the quality. There are several tools to achieve quality control and trimming, but these two are the most used nowadays.

3.3.2. Read counts

The read count analysis is an alternative to metagenomic assembly and binning. It is possible to use the exact same raw data that the sequencing machine generates.

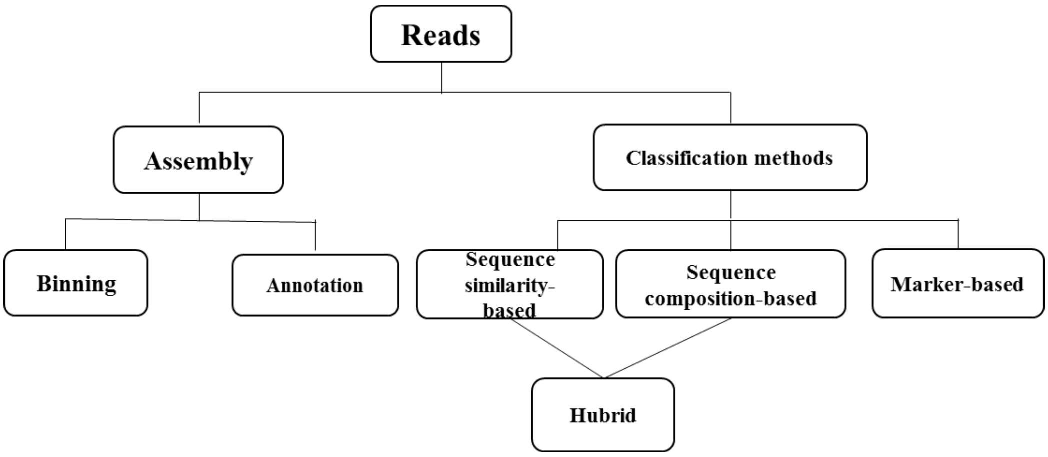


Figure 2. Metagenomic data possible workflows

3.3.3. Assembly

Sequence data can be analyzed without assembly, however, for the majority of investigations, longer, more contiguous sequences are preferable (contigs). Although next-generation (Illumina) sequencing is significantly less expensive, the information contained in a single read is constrained by the small read length (100–250 bases). Consequently, the assembly process that turns short reads into long sequences is frequently required (Tran Q and Phan V, 2020).

Based on read sequence similarity, assembly is the process of assembling sequence reads into continuous lengths of DNA, or contigs. Either the highest-quality nucleotide in each given read at each position, or the majority rule—the nucleotide that occurs most frequently at each position—is used to determine the consensus sequence for a contig. Depth or coverage is the measure of how many reads each consensus base has as a foundation. Even though compared to genomic assembly, metagenomic assembly is still a relatively young field of study, there has been an increase in interest in it over the past few years, which has led to the development of new tools. The assembly of metagenomic reads still faces unique difficulties, and algorithms must take these into account. In Overlap, Layout, and Consensus assembly—one of two methods used in genome assembly—read overlaps are identified, and an overlap graph is created (edges indicate overlapping reads). Then, based on the overlaps, reads are organized into contigs. To build a consensus sequence, the most likely sequence is picked. Or In dBg assembly, reads are divided into kmers by swiping a window of size k across the reads. The kmers are then converted into vertices in the dBg, with edges bridging any overlapped kmers. The number of times a kmer has been observed is kept track of and is displayed here above the kmers (Ryan Wick, 2016). By moving through the graph from edge nodes, contigs are created.

Long-read technologies, such as those from reputable Pacific Biosciences or the affordable, portable Oxford Nanopore Technologies MinION, are likewise becoming more and more popular with researchers. With individual reads spanning many genes, both may make assembly simpler or enable the creation of a much longer contiguous sequence. To assemble reads from these third-generation platforms, dBg techniques are abandoned and Overlap/Layout/Consensus models from the early days of Sanger sequencing are used instead. Although third-generation platform metagenome assembly methods have not yet been published, they can nevertheless produce excellent results when used with genome assembly tools like Canu or the extremely computationally efficient Minimap. As the cost comes down and the accuracy and yields improve, these new technologies are likely to seem increasingly attractive platforms for metagenomic experiments and there will likely be new metagenomic assembly tools devoted to them. There is many genomics assembly software that is trying to take on the challenge of metagenomic assembly. The most popular assemblers are: metaSPAdes, MEGAHIT, IDBA-UD, MetaRay, MetaVelvet

3.3.3.1.**Assessing assembly quality:**

With an ever-growing variety of metagenomic assemblers available, none of which offers a significant advantage over the others and each of which has its unique qualities, it is crucial to select the right tool for each application. An extensive overview of methods for evaluating genomes constructed out of metagenomes is given in a recent review paper **(Ayling M, and al, 2020).**

An assembly's quality is often inferred by using the N50, a popular statistic. The N50 is the smallest length of contigs that comprises 50% of the assembled bases if all the contigs in an assembly are ordered by length. An N50 of 10,000 bp, for instance, indicates that contigs of at least 10,000 bp are where 50% of the assembled bases are located. This statistic provides no measure of assembly correctness and just provides information on the contiguity of the assembled bases. It is also simple to manipulate (e.g., tools may decide to remove small contigs that they view as noise or chaff) **(Mikheenko A et al. 2015).** A new assembler might produce lengthy strings of random As, Cs, Gs, and Ts and achieve a high N50, but it wouldn't be accurate to the underlying genome; in fact, the N50 might even be greater than the biological genome. As a result, even though it is the most popular assembly statistic, it needs to be utilized with care and its significance recognized. For instance, well-known assembly techniques made for single genomes may result in metagenomic data set assemblies with high N50 values. However, this might have been done by consolidating inter-strain variation or deleting kmers representing lower coverage species, for example, by sacrificing complexity for contiguity **(Mikheenko A et al. 2015).**

To evaluate the quality of metagenome assembly, there are several techniques available:

MetaQUAST Quality ASsessment Tool is referred to as QUAST. The tool calculates numerous parameters to assess genomic assemblies. MUMmer, GeneMarkS, GeneMark-ES, GlimmerHMM, and GAGE are used by QUAST. Additionally, MetaQUAST makes use of the SILVA 16S rRNA database, Krona tools, MetaGeneMark, and BLAST. Contigs are BLAST-searched against a database of 16S rRNA genes using MetaQUAST, which then automatically downloads the top 50 references. Contigs that match these references are subsequently subjected to a quality assessment based on references. Such a method is only applicable to bacterial sequences **(Mikheenko A et al., 2015)**.

By measuring consistency between the input reads and the output assembly sequence using read quality, read pair orientation, read pair insert length, sequencing coverage, read alignment, and k-mer frequency, the Assembly Likelihood Evaluation framework (ALE) evaluates genomic and metagenomic assemblies (Clark SC and al, 2013). Two probabilities are defined by ALE using Bayesian statistics: I a probability distribution expressing the possibility of an assembly without any read information, and (ii) the probability of an assembly producing a specific collection of reads. To compare assemblies of the same genome, the probability that the assembly is correct—also known as the ALE score—combines these probabilities. The alignment of the reads onto the assembly, calculation of the placement sub-score (how well each read agrees with the assembly), insert sub-score (how well the distribution of paired-end read inserts agrees with the expected distribution), and depth sub-score is used to determine the probabilities (evenness of coverage). The tool also makes it easier to visualize alignments using a variety of genome browsers, which might be helpful for preliminary ad hoc analyses of assembly quality.

3.3.3.2.The challenge of metagenomic assembly

The quality and consistency of genome assemblies have significantly improved in recent years as a result of the usage of longer-range read data and improvements in assembly methods. These improvements have not been directly applied to metagenomic data sets, yet. After all, when assembling several genomes at various degrees of abundance, the assumptions established by the single genome assembly methods do not hold. Researchers had been using tools made for single genomes for many years before the creation of specialized assemblers for metagenomic data. In recent years, this has altered as a new class of tools founded on alternative concepts has emerged. Metagenomics assemblies' quality is far from perfect, nevertheless. The comun problems that are encounted when doing an assembly are: Unknown abundance and diversity, Related species, Memory and processing challenges, Initial classification of reads, Graph partitioning

3.3.4. Mapping reads

The contigs' characterization is a crucial step in any metagenomic investigation. There are numerous mapping methods and tools available. Any alignment tool, such as Bowtie, BLAST, bbmap, etc., might be used to match the contigs back to the original data in order to evaluate the reads' coverage. Other techniques for characterizing reads include the k-mer frequency, %GC content, and Condon utilization. The taxonomic assignment is also employed when a taxon is of interest.

3.3.5. Binning

The act of gathering reads or contigs and placing them on a specific genome is known as binning. It may be founded on compositional characteristics or the k-mer frequencies of a certain contig, which can likewise be utilized to differentiate between several species. Also known as differential abundance (or coverage) binning, abundance (coverage), or occasionally both (Luo Y and al, 2019). The primary objective of binning in metagenomic is to recover metagenome-assembled genomes (MAGs).

To achieve this, several binners were created. Deep learning-based techniques have more recently been utilized to enhance metagenomic binning. One of them is VAMB, a binner based on a variational auto-encoder proposed by Nissen et al. in 2021; it encodes composition and abundance characteristics into low-dimension embeddings. LR Binner is another deep learning method that adjusts variational auto-encoders for long reads. The semi-supervised Siamese neural network used by Semi Bin has must-link and cannot-link constraints that were discovered using reference genomes. However, the ones that are frequently employed are those that are based on compositional and abundance properties, such as:

One of the most effective binners is MetaBAT2 **(Zhang Z and colleagues, 2021).** It computes a pairwise distance matrix for all contig pairs using coverage and composition, with the composition feature based on an empirical posterior probability determined from a collection of reference genomes. The contigs are then connected based on their similarity scores, and a graph-based clustering technique is employed to bin the contigs based on their distances.

The MaxBin algorithm **(Wu YW and Singer SW, 2021)** divides the contigs or scaffolds of a metagenomic assembly into bins based on variations in coverage and tetranucleotide composition. A similar approach, MaxBin2, was reported by Wu et al. It uses an Expectation-Maximization algorithm to calculate the likelihood that a contig will fall into a specific bin and a single-copy marker gene to calculate the number of bins. Even though more composition and abundance approaches have been put forth, these two might be regarded as the most well-known and frequently applied.

Studies on long-read sequencing for metagenomics are still lacking, even though long-read sequencing technology is developing more rapidly. Most binning techniques have only been tested on short-read assemblies. The increased read length produces significantly better assemblies, produces more sparse assembly graphs, and allows for more reliable assessments of composition and coverage. One of the few binners made using long-read metagenomic data is GraphMB **(Lamurias A and colleagues, 2022).** Each contig's features and the contig-specific features of its neighboring contigs serve as the foundation for its graph-aware features. By learning representations of graph nodes based on node attributes and the network's structure, Graph Neural Networks (GNN), a form of deep learning model, are used to do this.

CHAPTER III

Material and Methods

CHAPTER III : MATERIAL AND METHODS

1. Windows Subsystem for Linux

The Windows Subsystem for Linux (WSL)was chosen as a work environment for many reasons. This feature of the Windows operating system allows users to run a Linux file system directly on Windows, alongside the traditional Windows desktop and apps, so you would get the best of both systems. In addition, WSL requires fewer resources (CPU, memory, and storage) than a full virtual machine and supports x64 and Arm CPUs. The Linux system could be run on any chosen Bash shell; in this case, it was Ubuntu 20.04, on windows 10 64x bits.

All the needed softwares and packages were then installedin the subsystem using conda version 4.12.0.Conda is an open-source package management system and environment management system. It helps to quickly install, run, and update packages and their dependencies. It easily creates, save, load, and switches between the created environments.The conda package and environment manager was included in the installed versions of Miniconda. Thus, each major analysis was runon a different environment to avoid conflict between packages. Otherwise, conda fixed or warned about any conflict.

2. Mammoth data selection and organization

The data was generated by Wang Y and al 2021. The dataset is publicly available on the European nucleotide archive (EMBL-ENA) and national center for biotechnology information (NCBI)databases,under project accession PRJEB43822. The 159 metgenomic samples where the mammoth's DNA was identified with abundance method were filtered, selected and downloaded from the original data set into the Windows Subsystem for Linux (WSL) Ubuntu 20.04 from EMBL-ENA in fastq format. Along with the nine supplementary data in the Excel format.

The fastq files were labeled with their run accession when downloaded, which was then matched with the sample ID in supplementary data 2 and 1 in order to get full information about each sample. Then all data was put together in an Excel file that contained Sample ID, Age (ka BP), Age interval, region, site ID, site abbreviation, sample type, age type, run accession, the read count, and a direct link to the fastq ftp of each sample.

3. Geographical regions and time interval

The 159 samples originated from fourdifferent regions; North Atlantic, Northwest and central Siberia, Northeast Siberia, and North America, 8 age intervals; 50+, Pre-LGM, LGM, Late

Glacial, Early Holocene, Mid Holocene, Late Holocene, and Anthropocene (table 1), and 41 sites across the Arctic. In addition, they came from two types of samples; Permafrost or Lake Sediment. The original site IDs were kept to ovoid confusions in future studies that may involve the whole data.

Table 1. Temporal distribution of the samples.

Interval	Age frame	Number of samples
50+	Older than 50 ka BP	17
Pre-LGM	50-26.5 ka BP	59
LGM	26.5-19 ka BP	37
Late Glacial	19-11.7 ka BP	23
Early Holocene	11.7-8.2 ka BP	16
Mid-Holocene	8.2-4.2 ka BP	6
Late Holocene	4.2 – 0 ka BP	1
Anthropocene	Younger than 1950 AD	0

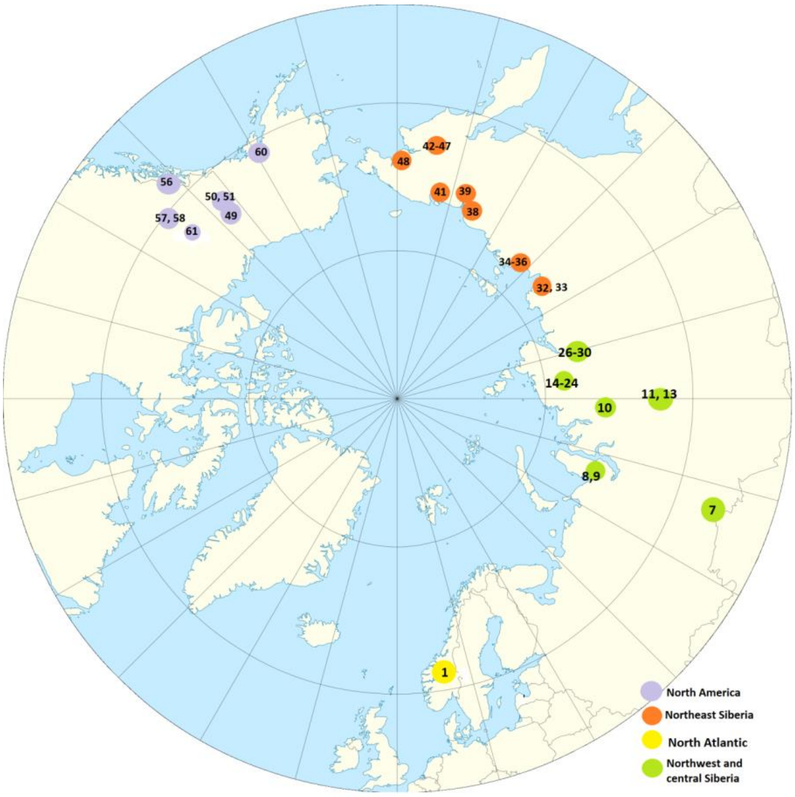


Figure3. Sites distribution. Samples (*n* = 159) from 41 sites were grouped into four geographical regions. The qite IDs are labelled on the map.

Table 2. Sites description

Site ID	Region	Site	Site abbreviation	Lat. (˚)	Long. (˚)
1	North Atlantic	Dovre 2006, Norway	06D1	62,3832	9,6742
7	Northwest and central Siberia	Yuribei River, Yugra, NW Russia	YUB	60,6009	71,9263
8	Northwest and central Siberia	Marita River 1, Yamal Peninsula, Russia	MarR1	68,6557	71,9225
9	Northwest and central Siberia	Marita River 2, Yamal Peninsula, Russia	MarR2	68,6565	71,9661
10	Northwest and central Siberia	Malay Kheta 2, Yenissei River, Russia	MK2	69,7397	84,8181
12	Northwest and central Siberia	Poloi 2, Yenisei River, Russia	PO2	66,7267	86,6391
13	Northwest and central Siberia	Ice Hill 4, Yenisei River, Russia	IH4	66,7580	86,6804
14	Northwest and central Siberia	Luktakh River, site 10, Taimyr Peninsula, Russia	LUR10	73,1565	93,4072
15	Northwest and central Siberia	Logata River, site 3, Taimyr Peninsula, Russia	LoR3	73,3504	96,9746
16	Northwest and central Siberia	Lake Taimyr, Taimyr River delta, Taimyr Peninsula, Russia	UTRD4	74,2664	99,8264
18	Northwest and central Siberia	Bol'shayaBalaknya River, site 1, Taimyr Peninsula, Russia	BBR1	72,5397	100,4312
21	Northwest and central Siberia	Lake Taimyr, Federov Island, Taimyr Peninsula, Russia	FI	74,6225	100,8280
22	Northwest and central Siberia	Bol'shayaBalaknya River, site 6, Taimyr Peninsula, Russia	BBR6	73,5262	101,0085

23	Northwest and central Siberia	Bol'shayaBalaknya River, site 7, Taimyr Peninsula, Russia	BBR7	73,5168	101,0089
24	Northwest and central Siberia	Baskura Peninsula, Lake Taimyr, Taimyr Peninsula, Russia	BAP	74,4936	101,2761
26	Northwest and central Siberia	Bol'shayaBalaknya River, site 9, Taimyr Peninsula, Russia	BBR9	73,6481	102,0177
28	Northwest and central Siberia	Bol'shayaBalaknya, site 5, Taimyr Peninsula, Russia	BBS5	73,6529	102,1207
29	Northwest and central Siberia	Bol'shayaBalaknya, site 6, Taimyr Peninsula, Russia	BBS6	73,6989	102,1969
30	Northwest and central Siberia	Khatanga site 1, Taimyr Peninsula, Russia	KS1	72,0967	102,3281
32	Northeast Siberia	DolgoeOzero, Lena River, Russia	DO	71,8667	127,0667
33	Northeast Siberia	Cape Bykovskii, Lena River, Russia	CAB	71,6667	129,5000
34	Northeast Siberia	Buor Kaya 1, Yana Bay, Russia	BK1	71,9062	132,7864
35	Northeast Siberia	Buor Kaya 2, Yana Bay, Russia	BK2	72,0028	132,8336
36	Northeast Siberia	Buor Kaya 3, Yana Bay, Rusia	BK3	71,9056	132,7853
38	Northeast Siberia	Kon’kovaya, Kolyma, Russia	KK	69,3833	158,4667
39	Northeast Siberia	DuvannyYar, Kolyma, Russia	DY	68,6667	159,0833
41	Northeast Siberia	Pleistocene Park, Kolyma, Russia	PP	68,4992	162,4068

42	Northeast Siberia	Maine River P1, Chukotka, Russia	MR1	64,2833	171,2500
43	Northeast Siberia	Maine River P2, Chukotka, Russia	MR2	64,2833	171,2500
44	Northeast Siberia	Maine River P3, Chukotka, Russia	MR3	64,2833	171,2500
45	Northeast Siberia	Maine River P4,Chukotka, Russia	MR4	64,2833	171,2500
47	Northeast Siberia	Maine River P6, Chukotka, Russia	MR6	64,2833	171,2500
48	Northeast Siberia	Anadyr Coast, Chukotka, Russia	AC	64,7352	177,3073
49	North America	Purgatory Site, Alaska, USA	PS	66,2333	-148,2667
50	North America	Stevens Village 1, Alaska, USA	SV1	65,9833	-148,9500
51	North America	Stevens Village 2, Alaska, USA	SV2	65,9833	-148,9500
56	North America	Quartz Creek, Yukon, USA	QC	60,9853	-130,5013
57	North America	Ross Mine, Yukon, Canada	RS	63,6900	-138,5800
58	North America	Christie Mine, Yukon, Canada	CM	63,6700	-138,6425
60	North America	Thistle Creek, Yukon, Canada	TC	63,0972	-139,5387
61	North America	Goldbottom Site, Yukon, Canada	GS	63,9333	-138,9667

4. Work flow

Several metagenomic pipelines that answers to different goals of analysis and fits different data frames are regularly generated. However, these pipelines generally require efficient laptops, and unfortunately, none of them preformed perfectly using WSL on Windows 10. Moreover, most metagenomic pipelines are made and fit for microorganisms only, and only few of them are fit to analyses eDNA of fauna and flora species; Hence, we developed a new work flow, presented in figure 4, where we put together all the memory efficient tools and software to analysis this type of data.

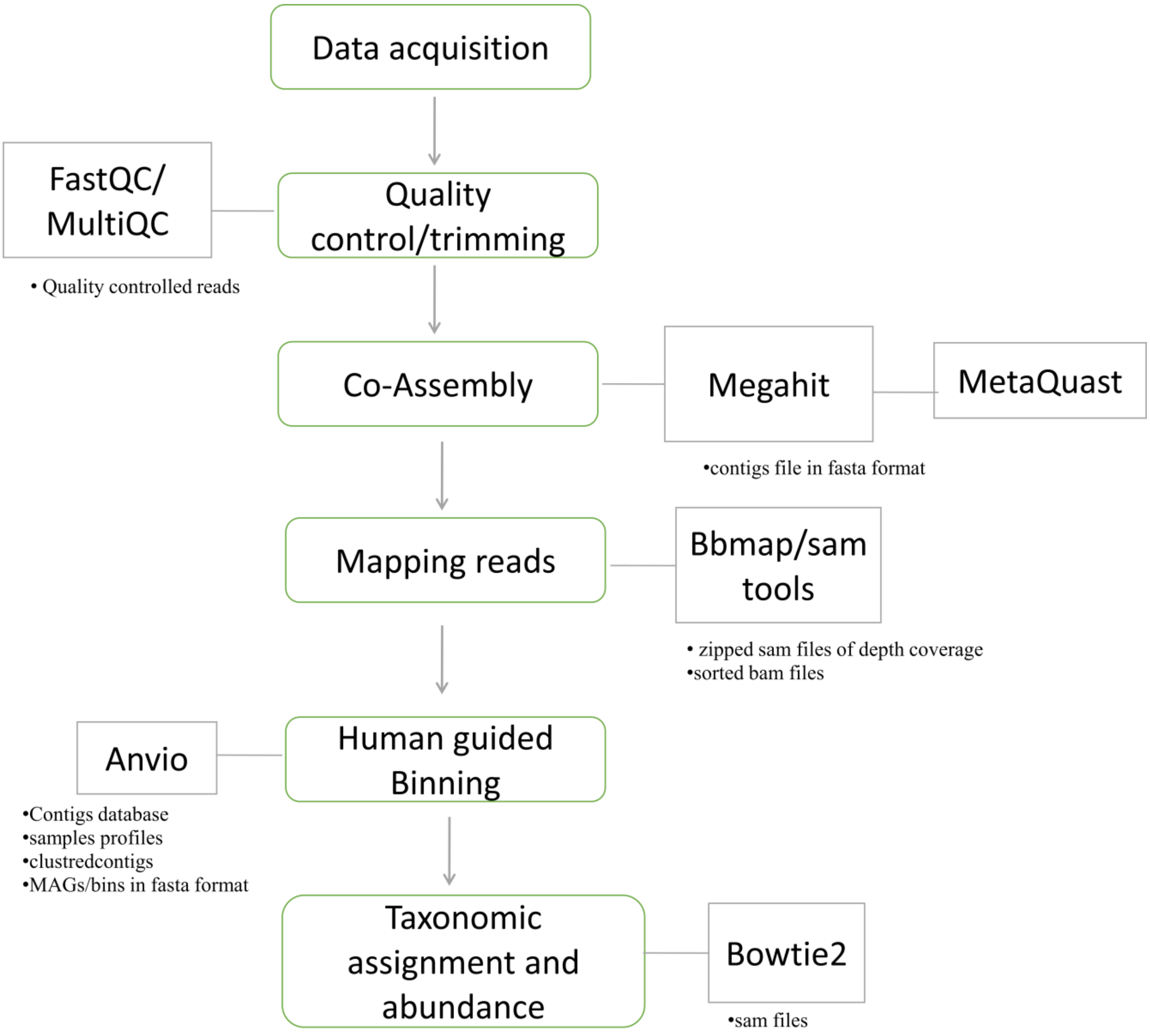


Figure 4. A diagram that describes the workflow followed to generate the MAGs

5. Reads quality controls

a. fastQC

Hundreds of millions of reads may be present in the raw sequencing reads that make up high throughput sequencing data, including PCR primers, adaptors, low-quality bases, duplicates, and other contaminants from the experimental methods. These might have an impact on the analysis's subsequent outcomes. As a result, this phase is crucial and necessary for any bioinformatic analysis. Simon Andrews from the Cambridge-based Babraham Institute created FastQC, a quality-control tool for high-throughput sequence data. The tool (version 0.11.9) was downloaded to the laptop's desktop, where all 159 samples were subjected to quality control. The resulting report includes some basic statistics as well as information on each base's quality, length, and GC and N content. It also includes information on sequence length distribution, sequence duplication levels, overrepresented sequences, and adaptor content. The output is a report in HTML format and a "zip" file that includes the text files fastqc data.txt and summary.txt.

b. MultiQC

Given the large number of samples, opening and analyzing each HTML file separately would take a lot of time. Phil Ewels created the MultiQC (v1.12) tool to aggregate and examine FastQC findings for numerous samples. Comparing your samples in detail is made feasible via visualization, which is not achievable when reading through reports one at a time. At the outset of the report, MultiQC gathers numerical statistics from each module to follow how the data reacts throughout the investigation. Additionally, it is designed with plugin hooks, a submodule system, and basic templating to facilitate simple expansion and customization. This analysis produces the following useful results: It is possible to observe if a portion of the sequences has consistently low-quality values, which is frequently the case since they are poorly imaged for a short sequence of reads, using general statistics, sequence counts, sequence quality histograms, and per sequence quality scores. The report also included information on each base sequence's content, its GC content, its per base N content, its length distribution, its level of duplication, and its overrepresented sequences.

6. Megahit

Megahit (v1.2.9) is an extremely quick and memory-efficient NGS assembler that is tailored for metagenomes, making it the assembler of choice. Conda was used to download the assembler in the Bioconda channel. A robust assembly that captures greater variety can be made possible by increasing read depth. This also makes it easier to compare samples and makes it easier to recover genomes from metagenomes due to uneven coverage. Instead of performing an independent assembly, where each input file would only contain reads from that particular sample, we opted to execute a co-assembly, in which the input files would be reads from many samples. The samples were split and merged following specific standards, due to the diversity of the data, which came from various regions and age interval (table 4). Excluding two samples (ERR6458978 from Northwest and Central Siberia Late Holocene and ERR6458853 from North Atlantic Early Holocene), which had to be studied and compiled as separate samples since they were located alone in the age range and region. In addition, one merged file, northeast Siberia Pre-LGM was simply too large to be co-assembled in a WSL, so it was randomly split into nsPLGM and nsPLMG1 for this process. To avoid confusion both the assembly and co-assembly processes will be referred to as assembly in this document.

Megahit was run on defaulted settings; including the used k-mers (21, 29, 39, 59, 79, 99,119, and 141). The quality of the assembly was evaluated with N50 and MetaQuast. N50 is the length cutoff for the longest contigs that contain 50% of the total genome length.

7. Bandage

In order to see de novo assembly graphs, a tool called Bandage (Bioinformatics Application for Navigating De novo Assembly Graphs) version 0.9.0was used to display connections that are not present in the contigs file. It was used to determine which other nodes have sequences that are contiguous with a node of interest, resolve ambiguities in the graph to improve or complete de novo assemblies, identify potentially problematic regions of an assembly, extract candidate sequences that extend through multiple nodes, and annotate graph images to illustrate assembly features. To visualize the assembly using Bandage and select the best K-mer, the fasta files of each K-mer were transformed into fastg files.

The size of the k-mer has a significant impact on how an assembly graph is structured. Large k-mers produce longer contigs with fewer connections, whereas small k-mers produce shorter contigs with a lot of connections. The optimum k-mer size is additionally influenced by read duration, read depth, and sequence complexity. As was previously said, Megahit performs assembly numerous times using various k-mers, therefore to select the best k-mer, we had to look at the FASTG files for each assembly using various k-mers in each sample to get the best graph for displaying in Bandage.

8. MetaQuast

The final step was to run MetaQuast (v2.2) on the contigs to finish the assembly analysis. For metagenomic datasets, MetaQUAST is an extension of Quast (QUality Assessment Tool). It was used to assess and contrast metagenome assemblies from various geographic locations and chronological epochs. The utility generates multi-genome tables and graphs, including Krona charts, and takes multiple assemblies. Except for the North Atlantic Early Holocene analysis, all analyses used the default settings. In this analysis, the —min-contig (-m) parameter was used to 100 rather than the default value of 500. Contigs shorter than 100 will not be taken into account since this parameter regulates the lower cutoff for a contig length (in bp). Additionally, contigs less than 500 won't be taken into account in the other studies that were performed with the default value of 500. Other than the N50, MetaQuast uses other tools that mitigate the shortcomings of the N50 to assess the quality of the assembly. Particularly the auN; the contiguity measurement The region beneath the Nx curves is called auN. Heng Li suggested and defended this metric. Since it takes into account the complete Nx curve and is more stable than N50, it is chosen over the latter. Although the Nx curve served as an inspiration for auN, sorting contigs by length is not necessary for the calculation of auN. Another important quality assembly assessment tool is the **L50;** the number of contigs equal to or longer than N50. In other words, L50 is the minimal number of contigs that cover half the assem.

9. BBmap and Samtools

After the assembly, mapping our reads for each sample to the assembly they built gives coverage information for each contig in each sample, which will help to recover metagenome-assembled genomes (MAGs). For that, we used bb map. BBMap is an open-source short-read aligner for DNA/RNAseq. It uses more memory than Burrows-Wheeler-based aligners, but the indexing speed is many times faster. Version 35.34 was installed on the subsystem using Conda. Then the unassembled reads were extracted using Samtools version 1.2. that was also installed using Conda.

An important step after the alignment was to convert all the Sam files into a sorted bam file to reduce the memory usage for binning. This was done using the faction Samtools sort.

10. Anvio

Anvio-7.1 was installed using conda in the WSL. However, as many metagenomic software and tools, anvio was build mainly for microorganisms analyses, thus many default parameters had to be reset. First thing after the installation, was to create an anvio contig database for each assembly. The database could then be filled with information like functional annotation, taxonomy assignment…etc. Creating a database for all 19 assemblies will calculate the tatranucleotide frequencies identifies, open reading frames using prodigal, and it will split the long contigs into segments of 20000 bp without breaking genes apart. More information may be added to the databases afterward.

Taxonomy assignment was done to help with the human guided binning. There are several ways to assign taxonomy for a metagenome sample. However, most tools and software offers a taxonomic database of microbial communities. Moreover, as mentioned above the arctic reference database of fauna and flora is quite poor. That’s why we first build a TAB-delimited matrix of taxonomic references. The species used for the assignment were chosen based on supplementary data 5 of plant abundance and supplementary data 6 of animal distribution and the check list of the article

Data profiling

After loading information into the contigs database, we provided information about each of the 159 samples so that each sample would have a profile database. The profile database of each sample contains information about the number of reads mapped to each contigs and where. Thereafter, all the profiles of all the samples of each assembly were merged into one or in this case 19 anvio profiles (each assembly alone).

Visualization the results was done using anvi-interactive, which allows to see the metagenome and how each samples’ reads were recruited to it. After running the command line, the terminal gives a link to the results that was opened using chrome navigator. Thereafter, 18 hierarchical trees were generated, displaying the coverage, GC content, length and taxonomy of each contig in each sample. The minimum contig length in this analysis was set to 1000 as advised from the anvio developers.

11. Bowtie 2

Bowtie 2 version 2.4.5 was downloaded and installed in a new environment in the terminal. This software is highly sensitive software used for index building and alignment. The first tool that was used was bowtie2 build to build an index. The index was built using one of the mammoth’s whole genome, which was downloaded in fasta format from ENA under the name acquisition ERR852028. The generated bins were than aligned against the reference index to identify the mammoths MAGs.

CHAPTERIV

Results

CHAPTER IV : RESULTS

1. General statistics

In the table , the general statistics of all 159 samples is shown. Including the duplicates percentage (Dups), the GC content, the reads length, the percentage of the failed reads, and the total sequences in millions (M seqs)

Table 3. General statistics resulting from the multiQC analysis of all of the 159 samples.

Sample Name	% Dups	% GC	Read Length	% Failed	M Seqs	Sample Name	% Dups	% GC	Read Length	% Failed	M Seqs
ERR6458624	0.0%	35%	43 bp	9%	0.2	**ERR6458603**	0.6%	39%	54 bp	0%	0.7
ERR6458546	1.7%	36%	48 bp	0%	0.2	**ERR6458615**	0.2%	39%	48 bp	0%	2.2
ERR6458549	1.8%	36%	71 bp	0%	1.1	**ERR6458619**	0.0%	39%	44 bp	9%	0.0
ERR6458553	3.6%	36%	66 bp	0%	1.1	**ERR6458714**	0.1%	39%	48 bp	0%	0.1
ERR6458571	0.4%	36%	48 bp	0%	1.5	**ERR6458719**	0.1%	39%	49 bp	0%	0.8
ERR6458601	0.2%	36%	53 bp	0%	0.5	**ERR6458735**	0.1%	39%	49 bp	0%	0.1
ERR6458608	0.4%	36%	55 bp	0%	1.5	**ERR6458779**	0.5%	39%	44 bp	0%	0.2
ERR6458622	0.3%	36%	52 bp	0%	6.5	**ERR6458816**	0.0%	39%	45 bp	0%	0.6
ERR6458623	0.1%	36%	49 bp	0%	0.4	**ERR6458870**	0.1%	39%	49 bp	0%	1.2
ERR6458626	0.0%	36%	45 bp	0%	0.9	**ERR6458871**	0.0%	39%	48 bp	0%	1.0
ERR6458747	0.2%	36%	59 bp	0%	0.2	**ERR6458878**	0.4%	39%	49 bp	0%	0.7
ERR6458780	0.0%	36%	42 bp	0%	2.2	**ERR6458969**	0.4%	39%	49 bp	18%	0.1
ERR6458518	0.2%	37%	46 bp	0%	1.3	**ERR6458970**	0.3%	39%	45 bp	18%	0.0
ERR6458547	2.1%	37%	65 bp	0%	0.4	**ERR6458994**	1.5%	39%	64 bp	9%	0.9
ERR6458552	1.2%	37%	51 bp	0%	0.4	**ERR6458998**	1.7%	39%	66 bp	9%	1.2
ERR6458572	0.5%	37%	48 bp	0%	0.2	**ERR6459038**	0.2%	39%	47 bp	0%	0.8
ERR6458594	0.3%	37%	53 bp	0%	1.3	**ERR6458510**	1.1%	40%	62 bp	0%	0.2
ERR6458595	0.2%	37%	53 bp	0%	1.0	**ERR6458513**	0.5%	40%	60 bp	0%	0.1
ERR6458596	0.1%	37%	52 bp	0%	1.0	**ERR6458533**	0.9%	40%	51 bp	0%	0.9
ERR6458602	0.0%	37%	47 bp	0%	0.3	**ERR6458534**	0.7%	40%	49 bp	0%	0.9
ERR6458605	0.4%	37%	54 bp	0%	0.8	**ERR6458542**	0.5%	40%	56 bp	0%	0.4
ERR6458610	0.2%	37%	53 bp	0%	1.0	**ERR6458543**	0.4%	40%	51 bp	0%	0.6
ERR6458617	0.3%	37%	54 bp	0%	0.3	**ERR6458630**	0.0%	40%	45 bp	0%	0.4
ERR6458621	0.1%	37%	48 bp	0%	1.2	**ERR6458631**	0.2%	40%	42 bp	9%	0.1
ERR6458628	0.1%	37%	46 bp	0%	0.2	**ERR6458717**	0.1%	40%	42 bp	0%	0.2
ERR6458684	0.1%	37%	43 bp	0%	0.5	**ERR6458753**	0.7%	40%	59 bp	0%	2.1
ERR6458700	0.1%	37%	52 bp	9%	0.4	**ERR6458778**	0.5%	40%	52 bp	0%	4.2
ERR6458703	0.2%	37%	53 bp	9%	1.5	**ERR6458782**	0.0%	40%	47 bp	0%	1.2
ERR6458707	0.2%	37%	44 bp	9%	0.3	**ERR6458822**	0.2%	40%	48 bp	0%	4.1
ERR6458754	0.2%	37%	55 bp	0%	1.2	**ERR6458853**	1.9%	40%	58 bp	0%	0.1
ERR6458785	0.0%	37%	47 bp	0%	1.0	**ERR6458864**	0.3%	40%	54 bp	0%	1.5
ERR6458789	0.0%	37%	44 bp	0%	1.9	**ERR6458865**	0.2%	40%	54 bp	0%	0.8

ERR6458819	0.0%	37%	44 bp	0%	1.0	**ERR6458895**	0.4%	40%	50 bp	0%	2.1
ERR6459003	1.6%	37%	64 bp	9%	0.6	**ERR6458946**	0.9%	40%	52 bp	0%	1.0
ERR6458489	0.2%	38%	42 bp	0%	0.7	**ERR6458974**	0.1%	40%	50 bp	0%	0.4
ERR6458499	0.6%	38%	53 bp	0%	3.3	**ERR6458498**	0.7%	41%	49 bp	0%	0.3
ERR6458509	0.7%	38%	54 bp	0%	3.2	**ERR6458526**	0.3%	41%	54 bp	0%	0.9
ERR6458512	0.6%	38%	47 bp	0%	1.8	**ERR6458528**	0.3%	41%	50 bp	0%	1.8
ERR6458520	0.1%	38%	44 bp	9%	0.1	**ERR6458538**	0.6%	41%	49 bp	0%	0.5
ERR6458522	0.1%	38%	52 bp	0%	0.7	**ERR6458587**	0.9%	41%	57 bp	0%	2.0
ERR6458525	0.3%	38%	49 bp	0%	4.6	**ERR6458607**	0.2%	41%	44 bp	0%	0.3
ERR6458532	0.7%	38%	57 bp	0%	2.3	**ERR6458649**	0.3%	41%	48 bp	0%	0.5
ERR6458566	0.2%	38%	56 bp	0%	0.5	**ERR6458652**	0.1%	41%	44 bp	0%	0.2
ERR6458604	0.7%	38%	54 bp	0%	0.8	**ERR6458775**	0.1%	41%	54 bp	0%	0.3
ERR6458606	0.4%	38%	47 bp	0%	0.3	**ERR6458863**	0.1%	41%	51 bp	0%	0.6
ERR6458611	0.1%	38%	45 bp	0%	0.4	**ERR6458866**	0.2%	41%	55 bp	0%	1.1
ERR6458612	0.4%	38%	50 bp	0%	2.0	**ERR6458867**	0.3%	41%	52 bp	0%	2.0
ERR6458627	0.3%	38%	52 bp	0%	1.5	**ERR6458868**	0.1%	41%	50 bp	0%	1.4
ERR6458691	0.1%	38%	49 bp	0%	0.5	**ERR6458869**	0.1%	41%	50 bp	0%	1.4
ERR6458697	0.5%	38%	52 bp	9%	3.7	**ERR6458954**	0.0%	41%	36 bp	9%	0.7
ERR6458698	0.1%	38%	47 bp	9%	0.5	**ERR6458973**	0.2%	41%	52 bp	0%	0.6
ERR6458702	0.2%	38%	50 bp	9%	0.3	**ERR6458531**	0.3%	42%	51 bp	0%	1.0
ERR6458705	0.1%	38%	48 bp	9%	1.2	**ERR6458544**	0.2%	42%	47 bp	0%	0.4
ERR6458706	0.1%	38%	49 bp	9%	0.2	**ERR6458609**	0.4%	42%	54 bp	0%	1.0
ERR6458712	0.1%	38%	44 bp	0%	0.9	**ERR6458625**	0.0%	42%	44 bp	0%	1.2
ERR6458715	0.2%	38%	52 bp	0%	1.3	**ERR6458654**	0.1%	42%	42 bp	0%	0.4
ERR6458718	0.1%	38%	49 bp	0%	1.2	**ERR6458723**	0.2%	42%	53 bp	0%	0.4
ERR6458732	0.2%	38%	53 bp	0%	0.3	**ERR6458745**	0.7%	42%	63 bp	0%	1.7
ERR6458736	0.1%	38%	46 bp	9%	0.1	**ERR6458751**	0.7%	42%	62 bp	0%	1.3
ERR6458748	0.1%	38%	50 bp	0%	0.1	**ERR6458770**	0.1%	42%	47 bp	0%	1.8
ERR6458784	0.0%	38%	46 bp	0%	0.8	**ERR6458772**	0.2%	42%	55 bp	0%	0.2
ERR6458786	0.0%	38%	43 bp	0%	1.1	**ERR6458781**	0.0%	42%	41 bp	9%	0.2
ERR6458790	0.3%	38%	48 bp	0%	1.0	**ERR6458791**	0.8%	42%	55 bp	0%	0.7
ERR6458890	0.1%	38%	46 bp	0%	0.4	**ERR6458795**	0.1%	42%	51 bp	0%	0.3
ERR6458960	1.0%	38%	51 bp	18%	8.6	**ERR6458799**	0.0%	42%	43 bp	0%	0.8
ERR6458961	0.4%	38%	49 bp	18%	2.7	**ERR6458887**	4.9%	42%	61 bp	9%	0.1
ERR6458964	0.3%	38%	51 bp	18%	2.0	**ERR6458530**	0.1%	43%	53 bp	0%	0.5
ERR6458971	0.0%	38%	40 bp	9%	8.4	**ERR6458579**	0.5%	43%	56 bp	0%	0.7
ERR6458978	0.2%	38%	46 bp	18%	1.4	**ERR6458586**	1.1%	43%	54 bp	0%	1.1
ERR6458980	0.1%	38%	49 bp	18%	0.5	**ERR6458597**	0.3%	43%	61 bp	0%	0.5
ERR6459036	0.3%	38%	46 bp	0%	0.3	**ERR6458752**	0.3%	43%	59 bp	0%	1.0
ERR6459039	0.6%	38%	47 bp	0%	0.5	**ERR6458491**	0.5%	44%	54 bp	0%	0.1
ERR6458507	0.7%	39%	53 bp	0%	1.0	**ERR6458616**	0.1%	44%	48 bp	0%	1.1

ERR6458508	0.6%	39%	54 bp	0%	0.6	**ERR6458653**	0.4%	44%	42 bp	0%	0.2
ERR6458514	0.6%	39%	54 bp	0%	0.5	**ERR6458716**	0.1%	44%	40 bp	0%	1.9
ERR6458524	0.5%	39%	56 bp	0%	0.7	**ERR6458769**	0.4%	44%	59 bp	0%	0.6
ERR6458527	0.2%	39%	46 bp	0%	0.7	**ERR6458893**	6.4%	44%	59 bp	0%	0.0
ERR6458536	0.6%	39%	49 bp	0%	0.9	**ERR6458727**	0.0%	45%	44 bp	9%	0.4
ERR6458545	0.8%	39%	56 bp	0%	1.2	**ERR6458761**	0.5%	45%	60 bp	0%	0.4
						ERR6458801	0.4%	46%	57 bp	0%	0.6

2. Quality control

2.1. Sequence quality

It represents the mean quality value across each base position in the read. All 159 samples were in the green area, indicating a good sequence quality.

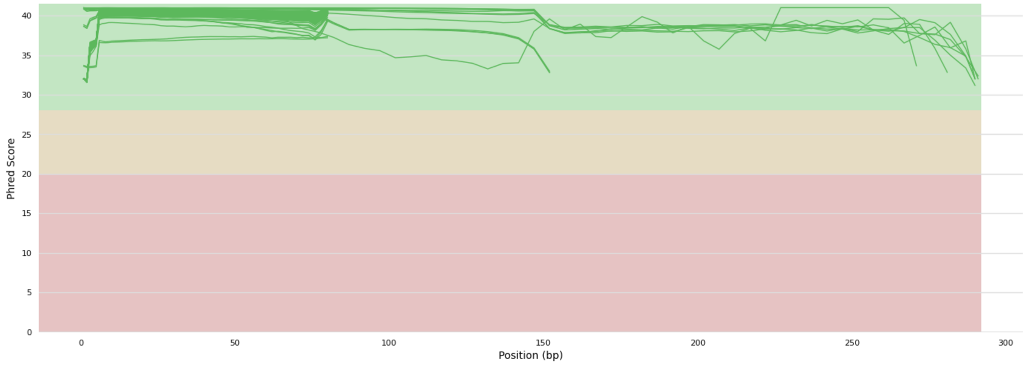


Figure 5. Mean quality scores

2.2. Per sequence quality scores

The figure represents the number of reads with average quality scores. It shows if a subset of reads has poor quality. All 159 samples scored a good quality.

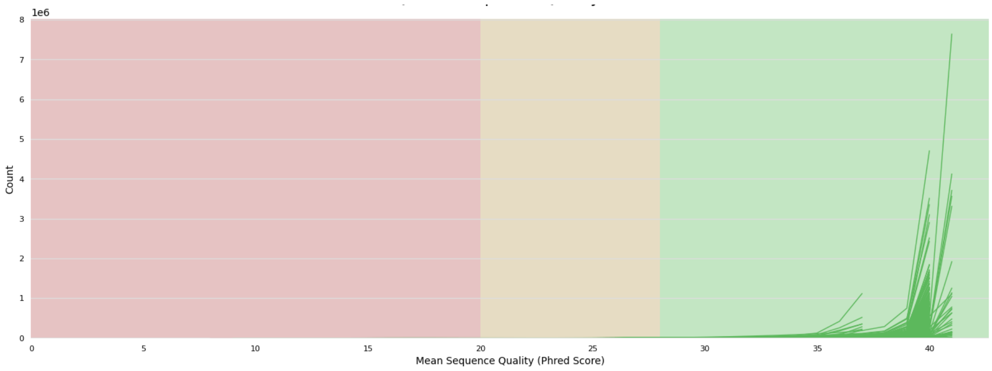


Figure 6. Per sequence quality scores

2.3. Per base sequence content

In the per base sequence output, 125 samples were represented in green, meaning they passed the analysis. 27 of the samples were represented in orange, which indicates that they passed with warning, 7 of them were represented in red meaning they failed. The warnings could be due to a difference between A and T bases, or the G and C that is bigger than 10% in any position. The failure however, is when those differences are greater than 20% in any position.

2.4. Per sequence GC Content

The average GC content of reads. Normal random library typically have a roughly normal distribution of GC content. 152 samples passed 7 samples with warning.

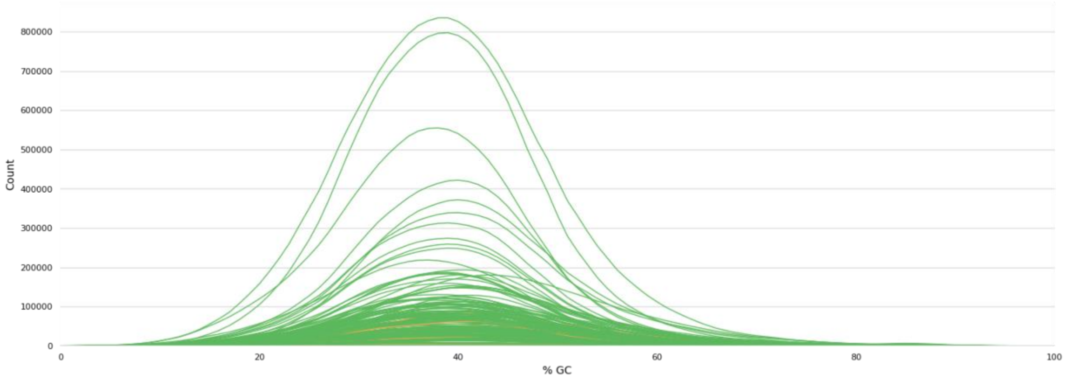


Figure 7. Per sequence GC content

2.5. Sequence length distribution

As expected, all 159 samples passed with warning. This module shows the distribution of fragment sizes in the analyzed files. In this case, the lengths of the sequences in the FastQ files were very different. Hence, it is entirely normal to encounte such warning in this type of data and thus the warnings here can be ignored.

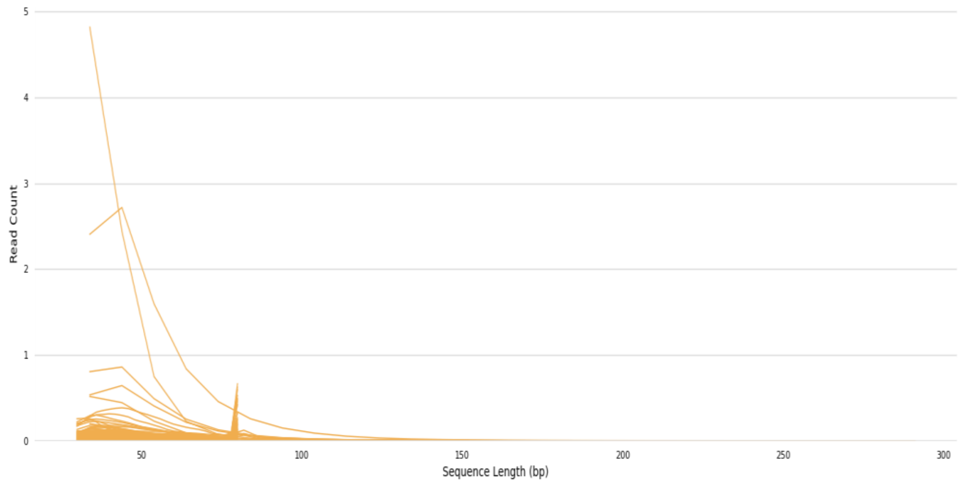


Figure 8. Sequence length distribution.

2.6. Sequence duplication levels

All 159 samples passed. Meaning non-unique sequences make up less than 20% of the total.

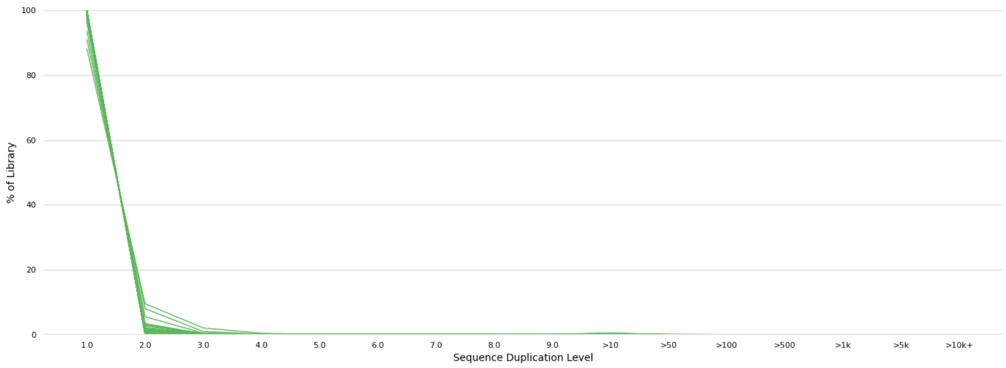


Figure 9. Sequence duplication levels

2.7. Adapter content

The graph proves that the adapter remove process was a success, as all 159 samples passed the adapter content analysis.

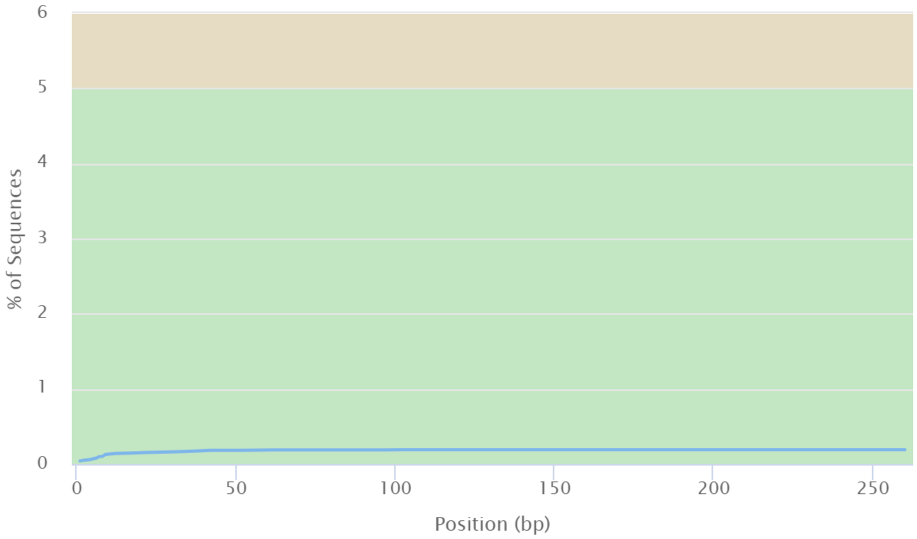


Figure 10. Adapter content

3. File merge

To increase coverage the fastq files were merged and renamed according to their regions and age interval. Therefore, each assembly was run on multiple samples (multiple fastq files), but the two samples, that of North West and central Siberia Late Holocene (NandSLH), and North Atlantic Early Holocene (NaEH) that were not merged. These files were taken as input as follow (table 4),

The merged samples from Northwest and central Siberia Per-LGM, labeled NandSPLMG, had the biggest number of samples merged (22 samples), followed by Northeast Siberia LGM labeled nsLGM with 19 samples. NaEH and NandSLH had only one sample. The largest number of reads was 21539521 bp and was that of Northeast Siberia LGM, and North Atlantic Early Holocene had the smallest number of read 77826.

Table 4. Samples merged to their region and age interval.

Region	**Age interval**	**run_accession**	**Compiled fastq ID**	**Number of reads (bp)**
Northwest and central Siberia	50+	ERR6458512	NandS50	12836175
		ERR6458518		
		ERR6458714		
		ERR6458715		
		ERR6458971		
	Pre-LGM	ERR6458489	NandSPLGM	17916421
		ERR6458491		

		ERR6458499		
		ERR6458507		
		ERR6458508		
		ERR6458509		
		ERR6458510		
		ERR6458527		
		ERR6458542		
		ERR6458543		
		ERR6458630		
		ERR6458631		
		ERR6458684		
		ERR6458702		
		ERR6458703		
		ERR6458705		
		ERR6458707		
		ERR6458717		
		ERR6458718		
		ERR6458719		
		ERR6458970		
		ERR6458980		
	LGM	ERR6458498	NandSLGM	7139209
		ERR6458544		
		ERR6458649		
		ERR6458652		
		ERR6458653		
		ERR6458654		
		ERR6458706		
		ERR6458712		
		ERR6458716		
		ERR6458994		
		ERR6458998		
	Late Glacial	ERR6458514	NandSLG	14114808
		ERR6458732		
		ERR6458735		
		ERR6458736		
		ERR6458946		
		ERR6458954		
		ERR6458960		
		ERR6458961		

		ERR6458969		
	Early Holocene	ERR6458513	NandSEH	2165284
		ERR6458964		
	Mid-Holocene	ERR6458973	NandSMH	1593409
		ERR6458974		
		ERR6459003		
	Late Holocene	ERR6458978	NandSLH	1369228
Northeast Siberia	Pre-LGM	ERR6458579	nsPLGM	11475340
		ERR6458604		
		ERR6458605		
		ERR6458611		
		ERR6458621		
		ERR6458624		
		ERR6458625		
		ERR6458626		
		ERR6458627		
		ERR6458723		
		ERR6458727		
		ERR6458761		
		ERR6458779		
		ERR6458780		
		ERR6458781		
		ERR6458782	nsPLGM1	20125898
		ERR6458784		
		ERR6458785		
		ERR6458786		
		ERR6458789		
		ERR6458799		
		ERR6458801		
		ERR6458816		
		ERR6458819		
		ERR6458822		
		ERR6458867		
		ERR6458868		
		ERR6458869		
		ERR6458870		
		ERR6458871		
	LGM	ERR6458586	nsLGM	21539521

		ERR6458587		
		ERR6458594		
		ERR6458595		
		ERR6458601		
		ERR6458602		
		ERR6458603		
		ERR6458606		
		ERR6458607		
		ERR6458608		
		ERR6458609		
		ERR6458610		
		ERR6458612		
		ERR6458623		
		ERR6458628		
		ERR6458770		
		ERR6458778		
		ERR6458790		
		ERR6458791		
	Late Glacial	ERR6458691	nsLG	2140823
		ERR6458769		
		ERR6458772		
		ERR6458775		
		ERR6458795		
		ERR6458999		
	Early Holocene	ERR6458520	nsEH	8846155
		ERR6458697		
		ERR6458698		
		ERR6458863		
		ERR6458864		
		ERR6458865		
		ERR6458866		
		ERR6458878		
	Mid-Holocene	ERR6458700	nsMH	459783
		ERR6458887		
North America	50+	ERR6458596	na50	18461289
		ERR6458616		
		ERR6458617		
		ERR6458619		
		ERR6458622		

		ERR6458745		
		ERR6458747		
		ERR6458751		
		ERR6458752		
		ERR6458753		
		ERR6458754		
		ERR6458895		
	Pre-LGM	ERR6458526	naPLGM	7445364
		ERR6458528		
		ERR6458530		
		ERR6458531		
		ERR6458532		
		ERR6458597		
		ERR6458890		
	LGM	ERR6458533	naLGM	4509546
		ERR6458748		
		ERR6458534		
		ERR6458536		
		ERR6458538		
		ERR6458545		
		ERR6458893		
	Late Glacial	ERR6458525	naLG	9595990
		ERR6458546		
		ERR6458547		
		ERR6458549		
		ERR6458552		
		ERR6458553		
		ERR6458571		
		ERR6458572		
	Early Holocene	ERR6458522	naEH	4125623
		ERR6458524		
		ERR6458566		
		ERR6458615		
North Atlantic	Early Holocene	ERR6458853	NaEH	77826

4. Co-assembly

CHAPTER IV : RESULTS

Megahit was used for the assembly, and it resulted in a fasta file that contains the contigs and a log file for each assembly. After the assembly, we first check the log file to take a look at some of the simple stats such as the size of the largest contigs, average contig length, N50, and to make sure things did not go wrong during the assembly (table). "na50" contained the biggest number of contigs 26128, where NaEH had a really small number (13). The largest contig was that of NandSEH with 44839 base pair, and the smallest contig was 200, which resulted in multiple samples (na50, naEH, naLGM, nsLG, and nsPLGM1). The N50 results are considered relatively small. The largest being that of naEH (661), and the smallest (293) resulted from the NaEH sample. These N50 results are expected for fragmented sequences of ancient DNA.

Table 5. Results of Megahit assemblies

Co assembled files	Contigs	Total (bp)	Min (bp)	Max (bp)	Avg (bp)	N50
na50	26128	1109271	200	19426	424	406
NaEH	13	3613	212	372	277	293
naEH	3283	1883395	200	14272	573	661
naLG	12690	5826413	202	22130	459	423
naLGM	5293	2208716	200	8285	417	391
NandS50	4016	2411340	270	10628	600	622
NandSEH	2338	1254923	275	16585	536	509
NandSLG	8815	5269115	269	44839	597	553
NandSLGM	4244	2694460	271	12510	634	635
NandSLH	403	235001	286	5191	583	581
NandSMH	759	426882	247	8853	562	500
NandSPLGM	10482	6568426	269	21820	626	636
naPLGM	10115	4112989	202	10044	406	389
nsEH	9120	4731473	202	16351	518	524
nsLG	2177	1033530	200	5792	474	455
nsLGM	25561	11667145	207	17394	456	444
nsMH	685	270530	236	4104	394	385
nsPLGM	11160	4774337	209	12822	427	409
nsPLGM1	19985	8757130	200	7061	438	431

4.1. Bandage

Bandage was used to view the assembly, get more information about it, and it helped us chose the best k-mer. For each k-mer (21, 29, 39, 59, 79, 99,119, and 141) a different output was generated. Hence, each fastq file converted to a fastg file had a different output depending on the k-mer.

Because the assembly was run on default parameters, some results were simply too complex to be interpreted or even taken into consideration. For example, na50 had more than 1 million nodes and edges for k21 and thus was impossible to be drawn in a graph. That is why; to jump over the complexity we will only mention the k-mers that had reasonable outputs. Some graphs consisted of many separate disconnected subgraphs; NandSPLGM k79, which indicates that the k-mer size is too large. Other graphs were too connected and very dense and tangled; naEH k21, which indicates that the k-mer size is too small. The most suitable k-mer would take into consideration the nature of the input (metagenome) the depth of coverage, and number of contigs. NaEH had 100% dead ends and a small number of nodes; it had a small number of contigs and not enough coverage, this could cause issues in the upcoming analyses.

Table 6. Bandage output based on the most suitable k-mer.

Co assembled files	**k-mer**	**Number of nodes**	**Number of edges**	**Edge overlaps (bp)**	**Percentage dead ends (%)**	**Median depth**
na50	29	431171	128342	29	78.06	2,65x
NaEH	21	13	0	n/a	100	3,00x
naEH	29	40084	11555	29	78.56	3,00x
naLG	29	260883	40248	29	88.60	2,00x
naLGM	29	75674	37383	29	63.63	3,06x
NandS50	29	95408	25549	29	80.25	3,00x
NandSEH	29	23289	7198	29	77.16	3,00x
NandSLG	29	290778	48182	29	87.76	2,00x
NandSLGM	29	108685	39172	29	73.43	2,93x
NandSLH	21	37702	30851	21	43.47	2,51x
NandSMH	29	28704	4690	29	87.73	2,00x
NandSPLGM	29	296537	80006	29	80.10	2,53x
naPLGM	29	126063	61428	29	64.10	3,00x
nsEH	29	120955	44104	29	73.09	3,00x
nsLG	29	37281	16657	29	67,06	3,00x
nsLGM	29	406661	147574	29	73.28	2,85x
nsMH	21	29655	19295	21	55.67	2,37x
nsPLGM	29	157086	57936	29	72.79	3,00x
nsPLGM1	29	265431	117362	29	67.48	3,00x

4.2. Metaquast

QUAST output contained several important information in different formats to evaluate the assemblies. The report text contained the assessment summary in plain text format. The report.tsv is a tab-separated version of the summary that was used for spreadsheets (Google Docs, Excel, etc). And the summary and plots in HTML and PDF formats.

4.2.1. Contig length according to age interval

Contig length plots show the contigs' length kbp. Nx varies from 0 to 100% and it is defined as the length for which the collection of all contigs of that length or longer covers x% of the assembly. The value of x is set at 90 by default. The region North America grouped 5 age intervals (+50, Per-LGM, LGM, Late Glacial, and Early Holocene). They all followed the same length distribution, however, Late glacial had clearly the longest contigs. North Atlantic (NaEH) had a very irregular distribution, which could be the result of the small number of contig and the dead ends. Northeast Siberia on the other hand grouped 5 age intervals (Per-LGM, LGM, late glacial, Early Holocene, mid Holocene) one of which was split to two for computational reasons (nsPLGM and nsPLGM1). The plot represented a very homologues distribution where the LGM AND Per-LGM had the longest contigs. Northwest and central Siberia had all of the seven age intervals where Late glacial represented the largest contig length.

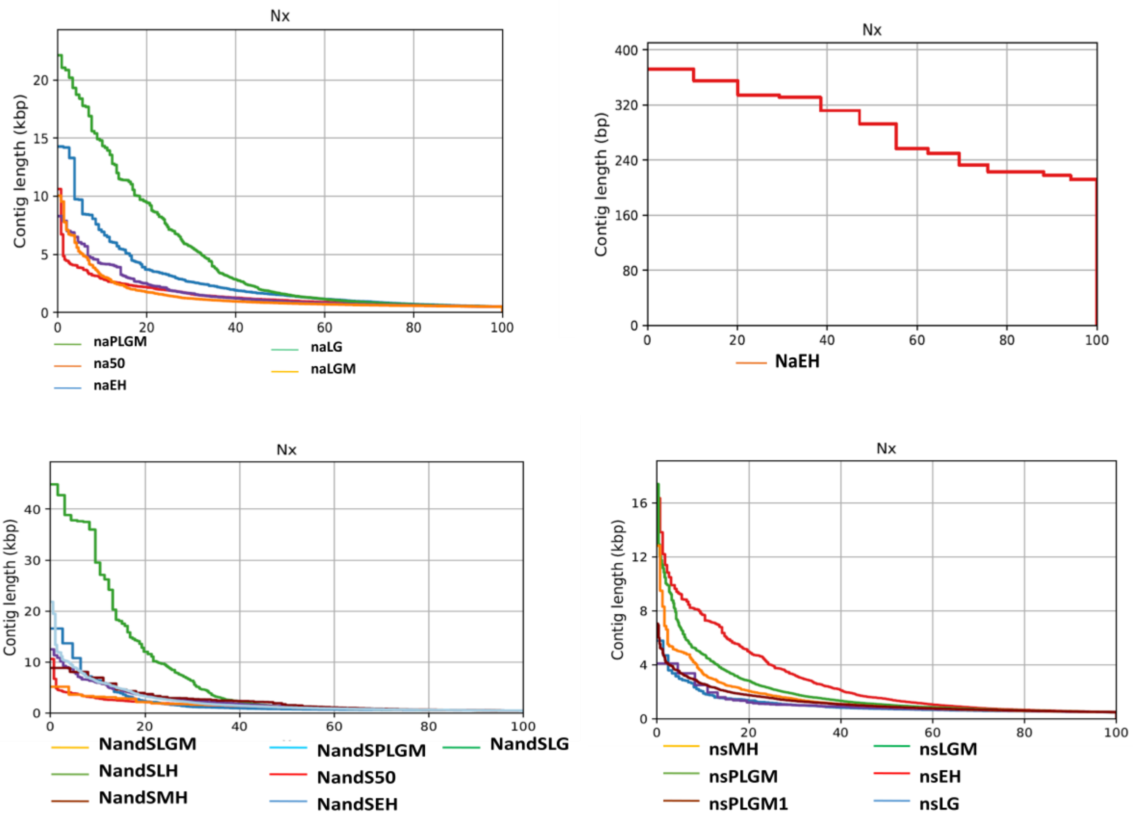


Figure 11. Contig length according to age intervals in all four regions

4.2.2. Cumulative length according to Age interval

In the cumulative length plots we look at the growth of contig lengths. On the x-axis, contigs are ordered from the largest to smallest, and the y-axis gives the size of the x largest contigs in the assembly. The cumulative length plot in the North America region had distinguished curve in naLG, with a length going up to 2500 kbp. North Atlantic's cumulative length was measured in base pairs and reached the 3200 bp. Northwest and central Siberia represented a somewhat diverse lengths, where Pre-LGM was the longest reaching 4000 kbp and Late Holocene was the shortest going less than 100 kbp. Northeast Siberia however represented the longest cumulative length measured in Mbp, where LGM surpassed 5Mbp.

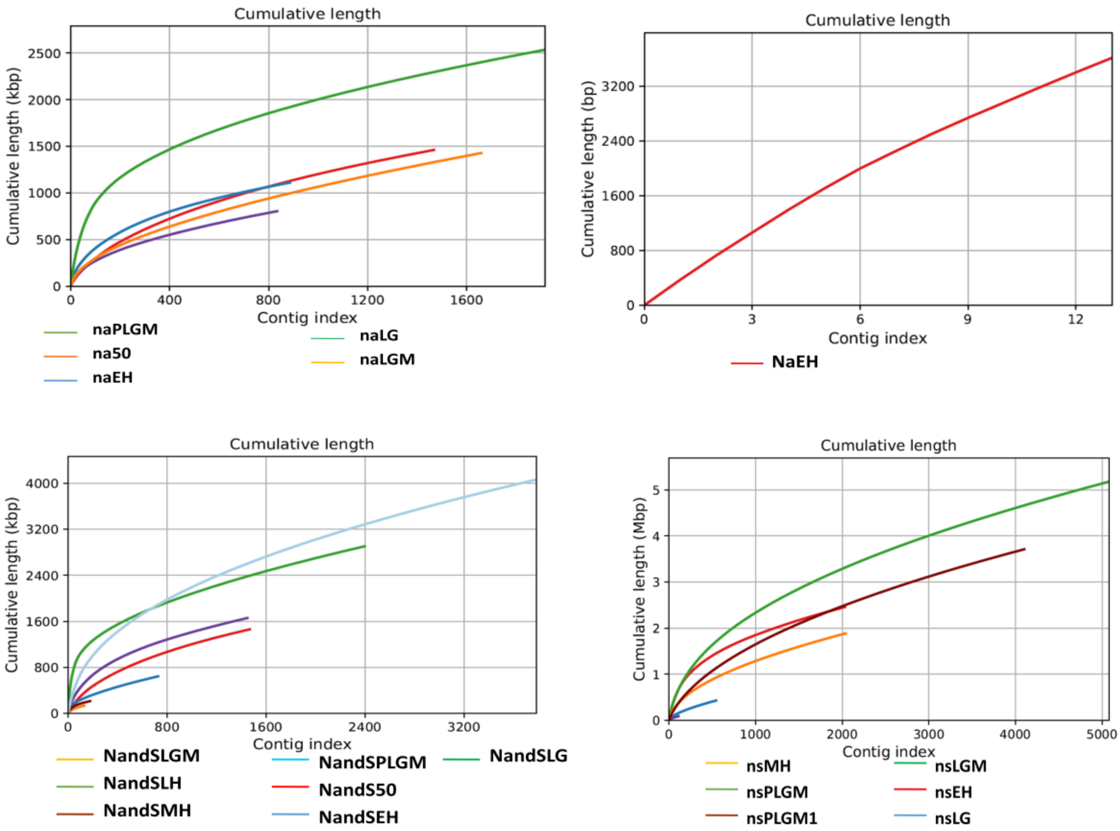
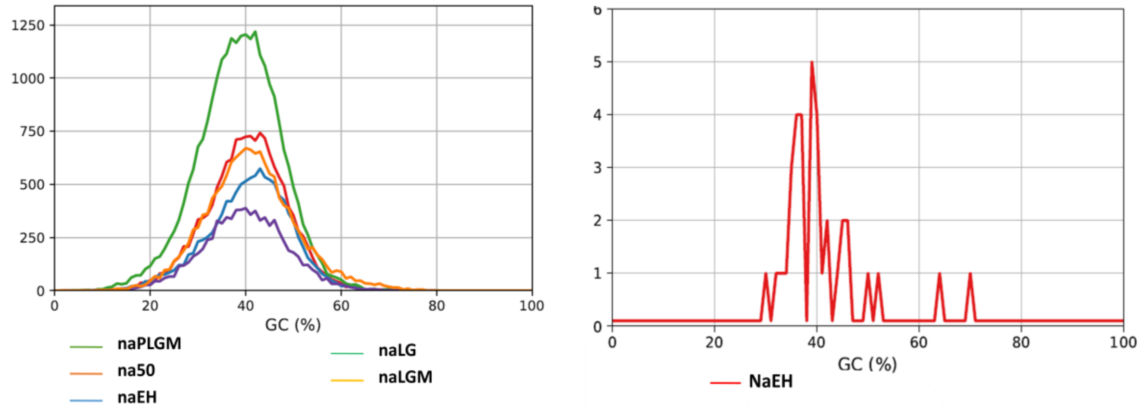


Figure 12. Cumulative lengths according to age interval in all four regions

4.2.3. GC content according to Age interval

GC content plot shows the distribution of GC content in the contigs. The x value is the GC percentage (0 to 100 %), and the y value is the number of non-overlapping 100 bp windows which GC content equals x %. The distribution is typically Gaussian, which is the case of all of the North America, northwest and central Siberia, and northeast Siberia regions. However, north Atlantic distribution appears to have multiple peaks. This could be because of the results represented above, indicating the small number of contigs and the dead ends.



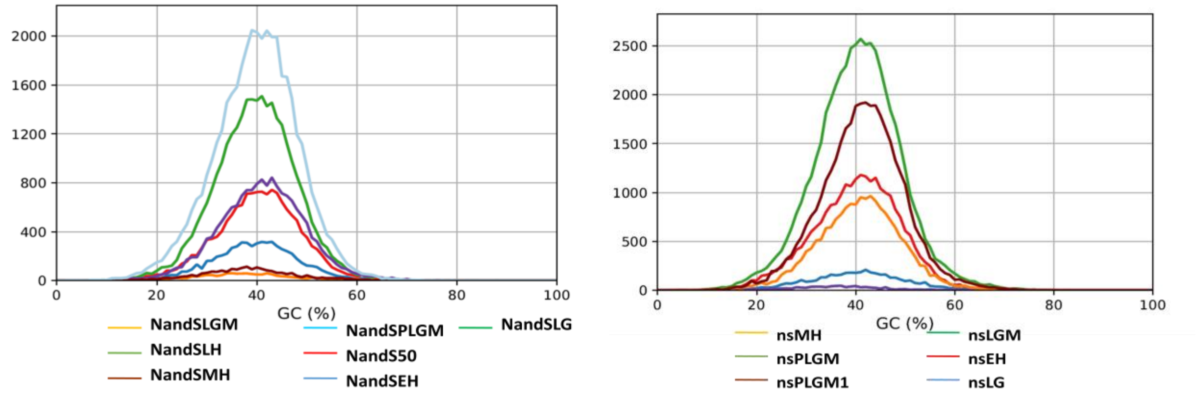


Figure 13. GC content according to age interval in all regions.

4.2.4. Report

The summary report clarifies important results that give details about the assemblies. The first 3 columns shows the number of contigs larger than 1000, 10000, and 25000 base pair respectively. nsLGM has the biggest number of contigs larger than 1000 (1319). However, only 12 of them were larger than 10000 and none of them exceeded 25000bp. The assemblies that had small number of contigs grater than 1000; NandSLH, nsMH, and NaEH, could cause problems in upcoming analyses. Especially for NaEH that had 0 contig grater than 1000. Also, NandSLG was the only assembly that contained contigs bigger than 25000 (10 contigs), where its largest was 44839 bp. auN is less affected by big jumps in contig lengths such is this case and it considers the entire Nx curve. A better N*x* curve is higher, or has a larger area under the curve, which is the case in most assemblies. NandSLG that had the biggest contig represents the biggest auN as well. However, NaEH had a relatively low auN. L50, the number of contigs equal to or longer than N50 of nsLGM was the largest at (1231), and that of NaEH was naturally the lowest.

Table 7. summary report of the assemblies.

Co assembled files	contigs (>= 1000 bp)	contigs (>= 10000 bp)	contigs (>= 25000 bp)	Largest contig	auN	L50	N's per 100 kbp
na50	447	1	0	10618	1490.3	406	0.00
NaEH	0	0	0	372	288.8	6	0.00
naEH	327	3	0	14272	2733.6	184	0.00
naLG	550	32	0	22130	4804.2	258	0.00
naLGM	198	0	0	8285	1728.2	213	0.00
NandS50	447	1	0	10618	1490.3	406	0.00
NandSEH	108	3	0	16585	2260.0	197	0.00

NandSLG	532	31	10	44839	7646.6	322	0.00
NandSLGM	461	4	0	12510	2348.2	303	0.00
NandSLH	43	0	0	5191	1547.0	35	0.00
NandSMH	58	0	0	8853	2520.9	33	0.00
NandSPLGM	1030	12	0	21820	2416.5	848	0.00
naPLGM	289	1	0	10044	1463.6	488	0.00
nsEH	609	6	0	16351	2875.8	366	0.00
nsLG	85	0	0	5792	1089.7	180	0.00
nsLGM	1319	9	0	17394	2000.8	1231	0.00
nsMH	20	0	0	4104	1127.9	37	0.00
nsPLGM	453	1	0	12822	1566.5	561	0.00
nsPLGM1	992	0	0	7061	1284.8	1218	0.00

5. Mapping

The output of the mapping was a sam file and some basic statistics like the number of reads used, mapped, unambiguous, ambiguous…etc. Using a bbmap command pileup, we generated a cov.txt file for each aligned reads that contained, amoug others, the average coverage, percent scaffolds with any coverage, and percent of reference bases covered, where in this case the reference were the contigs files. All of the sam files were converted into sorted bam files and then to a sorted bam file.

6. Anvio

6.1.Databases creating

6.1.1. Open reading frames identification

First thing when creating a contig database, anvio uses prodigal v2.6.3 to identify genes (table 8).

Table 8. number of genes indentified in each of the 19 databases using prodigal

Co assembled files	Number of genes indentified
na50	20650
NaEH	10
naEH	2746
naLG	10083

naLGM	4232
NandS50	3468
NandSEH	1826
NandSLG	8056
NandSLGM	3881
NandSLH	339
NandSMH	619
NandSPLGM	9351
naPLGM	8089
nsEH	7633
nsLG	1820
nsLGM	20840
nsMH	483
nsPLGM	8845
nsPLGM1	15912

6.1.2. Assign taxonomy

The main propos of this taxonomic assignment is to help with the human guided binning. The file text was converted to a TAB-delimited file to conduct a default_matrix assignment. Table 9 represents the taxonomic classification of the reference matrix and their NCBI taxonomic IDs.

Table 9. Matrix taxonomic reference table

gene_callers_id	t_domain	t_phylum	t_class	t_order	t_family	t_genus	t_species	Taxonomic ID in NCBI
1	Eukaryota	Chordata	Mammalia	Proboscidea	Elephantidae	Mammuthus	Mammuthus primigenius	37349
2	Eukaryota	Chordata	Mammalia	Artiodactyla	Bovidae	Bison	Bison bison	9901
3	Eukaryota	Chordata	Mammalia	Artiodactyla	Cervidae	Rangifer	Rangifer tarandus	9870
5	Eukaryota	Chordata	Mammalia	Perissodactyla	Rhinocerotidae	Coelodonta	Coelodonta antiquitatis	222863
7	Eukaryota	Chordata	Mammalia	Carnivora	Ursidae	Ursus	Ursus thibetanus	9642

8	Eukaryota	Chordata	Mammalia	Carnivora	Canidae	vulpes	Vulpes vulpes	9627
9	Eukaryota	Chordata	Mammalia	Perissodactyla	Lagomorpha	Leporidae	Leporidae arcticus	9979
10	Eukaryota	Chordata	Mammalia	Perissodactyla	Equidae	Equus	Equus caballus	9796
11	Eukaryota	Chordata	Mammalia	Artiodactyla	Bovidae	Ovis	Ovis aries	9940
12	Eukaryota	Chordata	Mammalia	Carnivora	Canidae	Canis	Canis lupus	9612
13	Eukaryota	Streptophyta	Magnoliopsida	Rosales	Rosaceae	Potentilla	Potentilla discolor	648872
14	Eukaryota	Streptophyta	Pinopsida	Pinales	Pinaceae	Larix	Larix kaempferi	4800
15	Eukaryota	Streptophyta	Polypodiopsida	Equisetales	Equisetaceae	Equisetum	Equisetum arvense	3258
16	Eukaryota	Streptophyta	Magnoliopsida	Ranunculales	Papaveraceae	Papaver	Papaver somniferum	3469
17	Eukaryota	Streptophyta	Magnoliopsida	Fagales	Betulaceae	Betula	Betula platyphylla	78630
18	Eukaryota	Streptophyta	Magnoliopsida	Malpighiales	Salicaceae	Salix	Salix brachista	2182728
19	Eukaryota	Streptophyta	Magnoliopsida	Asterales	Asteraceae	Artemisia	Artemisia annua	35608
20	Eukaryota	Streptophyta	Magnoliopsida	Rosales	Rosaceae	Dryas	Dryas octopetala	57948
21	Eukaryota	Streptophyta	Magnoliopsida	Brassicales	Brassicaceae	Draba	Draba maguirei	171821
22	Eukaryota	Streptophyta	Magnoliopsida	Alismatales	Potamogetonaceae	Potamogeton	Potamogeton perfoliatus	55320
23	Eukaryota	Streptophyta	Magnoliopsida	Poales	Juncaceae	Juncus	Juncus effusus	13579
24	Eukaryota	Streptophyta	Magnoliopsida	Myrtales	Onagraceae	Epilobium	Epilobium hirsutum	210355
25	Eukaryota	Streptophyta	Magnoliopsida	Lamiales	Plantaginaceae	Callitriche	Callitriche palustris	50469
26	Eukaryota	Streptophyta	Magnoliopsida	Poales	Poaceae	Puccinellia	Puccinellia distans	13649
27	Eukaryot	Streptophyta	Magnoliopsida	Ericale	Ericaceae	Pyrola	Pyrola rotundifolia	13651

	a			s				
28	Eukaryota	Streptophyta	Magnoliopsida	Lamiales	Lentibulariaceae	Utricularia	Utricularia tenuicaulis	262112
29	Eukaryota	Streptophyta	Magnoliopsida	Ericales	Ericaceae	Vaccinium	Vaccinium corymbosum	69266
30	Eukaryota	Streptophyta	Magnoliopsida	Malpighiales	Violaceae	Viola	Viola purpurea	97447
31	Eukaryota	Streptophyta	Magnoliopsida	Caryophyllales	Polygonaceae	Bistorta	Bistorta officinalis	125587
32	Eukaryota	Streptophyta	Magnoliopsida	Asterales	Asteraceae	Tephroseris	Tephroseris	1534674
33	Eukaryota	Streptophyta	Magnoliopsida	Poales	Poaceae	Agrostis	Agrostis stolonifera	63632
34	Eukaryota	Streptophyta	Magnoliopsida	Poales	Poaceae	Alopecurus	Alopecurus myosuroides	81473
35	Eukaryota	Streptophyta	Magnoliopsida	Poales	Poaceae	Anthoxanthum	Anthoxanthum odoratum	29661
36	Eukaryota	Streptophyta	Magnoliopsida	Poales	Poaceae	Calamagrostis	Calamagrostis breviligulata	286491
37	Eukaryota	Streptophyta	Magnoliopsida	Poales	Poaceae	Deschampsia	Deschampsia antarctica	159298
38	Eukaryota	Streptophyta	Magnoliopsida	Poales	Poaceae	Elymus	Elymus caninus	129741
39	Eukaryota	Streptophyta	Magnoliopsida	Poales	Poaceae	Nardus	Nardus stricta	29687
40	Eukaryota	Streptophyta	Magnoliopsida	Poales	Poaceae	Phleum	Phleum pratense	15957
41	Eukaryota	Streptophyta	Magnoliopsida	Lamiales	Plantaginaceae	Lagotis	Lagotis integra	1310058

6.2. Data profiling

Some reads in the bam files did not end up being used by anvio. This could be because those reads mapped but had no defined sequence, or they had a sequence but did not map. The minimum contig length had to be set to 1000 base pairs as advised from the anvio developers. Thus, contigs shorter than that were not taken into consideration. Unfortunately, as concluded above, NaEH had contigs shorter than 1000bp, so its profiling was impossible and meaningless.

6.3.Hierarchical clustering visualization

All of the generated trees have the same general layout; a hierarchical clustering of the contigs from the assembly at the center, where each leaf represents a contig or a fragment of it for the few cases were the contig exceeded 20000 bps. The layers radiating out of the clustering are layer of information that were added to the contig database. Each layer displays the added information for each corresponding contigs. Among the layers are the samples, which display the read coverage to each contig and helps in human guided binning, as well as a length layer, GC content, and taxonomy assignment.

Northwest and central Siberia 50+

The bins that were identified in the merged profile database and stored in the database as NandS50 bins collection, describe 4 bins accounting for 237,754 nucleotides, which represent 9.86% of all nucleotides stored in the contigs database, and 30.74% of nucleotides stored in the profile database. Out of the four bins, 2 (Bin 1 and Bin 2) were taxonomically assigned. Bin 1 was assigned to Salix plant, with the most abundance in sample ERR6458971. Bin 2 was assigned to Draba, with the most abundance in sample ERR6458971 as well (table). The merged profile database that was generated with the minimum contig length of 1000 contained 447 contigs, which correspond to 11% of all contigs, and 32% of all nucleotides found in the contigs database.

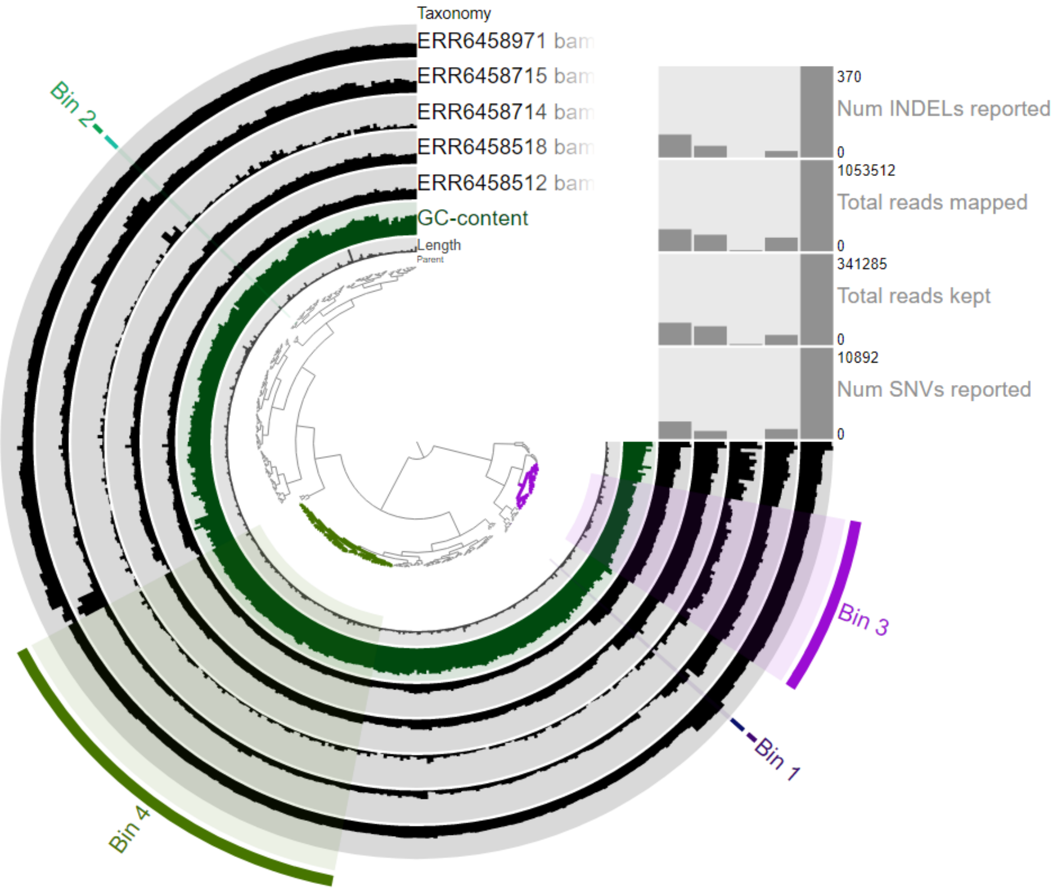


Figure 14. Hierarchical clustering tree of the contigs in Northwest and central Siberia 50+

Table 10. Taxonomically assigned bins abundancein Northwest and central Siberia 50+

	Bin 1	Bin 2
Length	1146	2786
GC_content	0.37	0.36
ERR6458512_bam_sorted	12.25	3.79
ERR6458518_bam_sorted	5.86	3.29
ERR6458714_bam_sorted	0.10	0.12
ERR6458715_bam_sorted	2.83	1.89
ERR6458971_bam_sorted	65.88	11.78
Taxonomy	Salix	Draba

A summary of each bin is presented in table. Bin 1 and 2 contained only one contig, which were the contigs assigned taxonomically. Bin 4 was the largest in total length and number of contigs. N50 of all bins is relatively high, and thus good.

Table 11. Bins general summary reportNorthwest and central Siberia 50+

Bins	total_length	num_contigs	N50	GC_content
Bin_1	1146	1	1146	37.347294938917976
Bin_2	2786	1	2786	36.46805455850682
Bin_3	78884	38	2269	33.37828494602917
Bin_4	154938	85	1910	44.64811446738392

Northwest and central Siberia Early Holocene

Bins that were identified in the merged profile database of Northwest and central Siberia Early Holocene and stored in the database as bin NandSEH collection, describe 7 bins accounting for 241,735 nucleotides, which represent 19.26% of all nucleotides stored in the contigs database, and 98.71% of nucleotides stored in the profile database. None of the 7 bins where taxonomically assigned, hence the binning depended mainly on the coverage. The merged profile database that was generated with the minimum contig length of 1000 contained 108 contigs, which correspond to 4% of all contigs, and 19% of all nucleotides found in the contigs database.

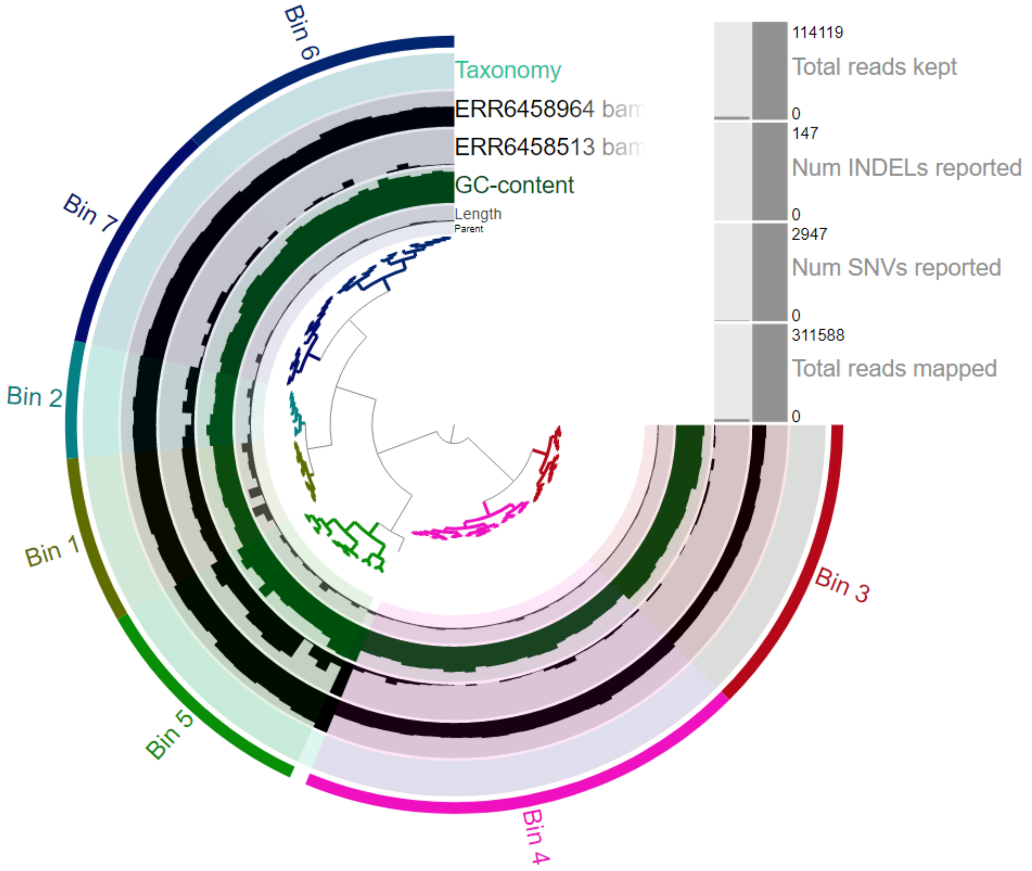


Figure 15. Hierarchical clustering tree of the contigs in Northwest and central Siberia Early Holocene

Table represent a summary of each bin. Tha total legth of bin 1 was the largest and Bin 3 had the biggest number of contigs. N50 of all bins is relatively high, and thus good.

Table 12 . Summary of the 7 bins in Northwest and central Siberia Early Holocene

Bins	total_length	num_contigs	N50	GC_content
Bin_1	71248	10	10799	33.89676411392948
Bin_2	20977	7	2892	35.35725051626265
Bin_3	21574	18	1211	42.053489428563296
Bin_4	40966	27	1579	36.088343473596076
Bin_5	38282	14	3620	44.511554383149324
Bin_6	24994	17	1540	48.21150012940316
Bin_7	23694	14	2130	35.828264891164565

Northwest and central Siberia Late glacial

Bins that were identified in the merged profile database Northwest and central Siberia Late glacial and stored in the database as "NandSLG" collection, describe 9 bins accounting for 1,112,529 nucleotides, which represent 21.11% of all nucleotides stored in the contigs database, and 65.88% of nucleotides stored in the profile database. 5 of the bins were taxonomically assigned to 5 different plant species. Bin 1 was assigned to Alopecurus. Bin 2 was assigned to Salix. Bin 3 was assigned to Utricularia. Bin 4 to Ursus, and Bin 5 was assigned to Epilobium (table). The merged profile database that was generated with the minimum contig length of 1,000 contained 532 contigs, which correspond to 6% of all contigs, and 32% of all nucleotides found in the contigs database.

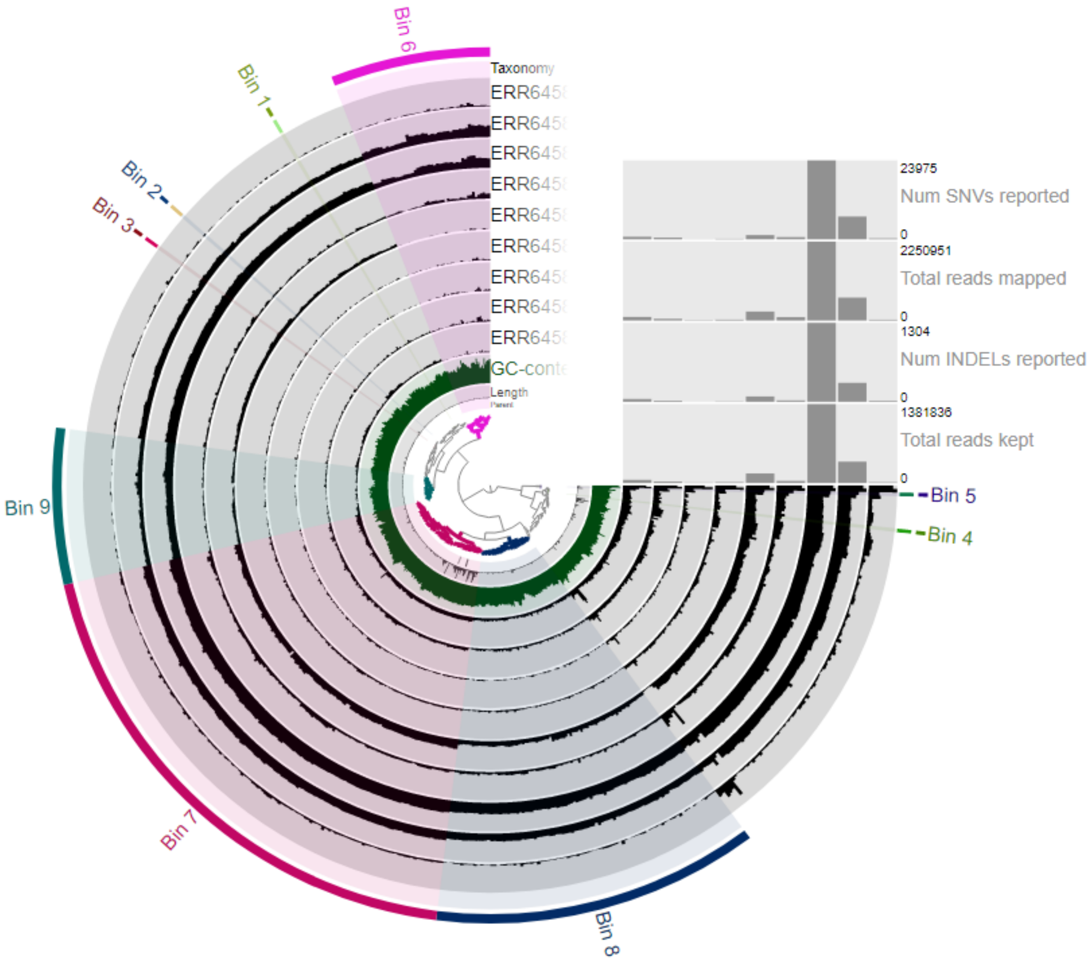


Figure 16.Hierarchical clustering tree of the contigs in Northwest and central Siberia Late glacial

Table 13. Taxonomically assigned bins abundance in Northwest and central Siberia Late glacial

	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5
length	1096	1019	1206	1123	2445
gc_content	0.36	0.34	0.42	0.38	0.42
ERR6458514_bam_sorted	0.17	0.72	0.24	3.87	16.61
ERR6458732_bam_sorted	0.11	0.12	0.17	1.24	5.33
ERR6458735_bam_sorted	0.07	0	0	0.20	1.17
ERR6458736_bam_sorted	0	0.05	0.17	0.48	2.21
ERR6458946_bam_sorted	0.73	0.63	1.91	1.91	42.99
ERR6458954_bam_sorted	0.06	0.24	0.33	1.44	8.35
ERR6458960_bam_sorted	8.58	10.97	10.46	105.75	459.63
ERR6458961_bam_sorted	2.80	3.09	3.57	25.06	115.74

ERR6458969_bam_sorted	0	0.16	0.03	0.42	2.02
Taxonomy	Alopecurus	Salix	Utricularia	Ursus	Epilobium

A summary of each bin is presented in table 14. The first 5 bin contained only 1 contig that was taxonomically assigned. Bin 7 had the biggest number of contigs, and the largest total length. N50 of all bins is relatively high, and thus good.

Table 14. Bins general summary report Northwest and central Siberia late glacial

Bins	total_length	num_contigs	N50	GC_content
Bin_1	1096	1	1096	36.31386861313868
Bin_2	1019	1	1019	34.15112855740923
Bin_3	1206	1	1206	41.7910447761194
Bin_4	1123	1	1123	38.201246660730185
Bin_5	2445	1	2445	41.96319018404908
Bin_6	57924	42	1261	43.48334120188449
Bin_7	823439	136	13574	40.22585289490916
Bin_8	173027	86	2142	37.26903714318965
Bin_9	51250	41	1177	36.18866200774062

Northwest and central Siberia LGM

Bins that were identified in the merged profile database and stored in the database as "NandSLGM " collection, describe 9 bins accounting for 638,790 nucleotides, which represent 23.71% of all nucleotides stored in the contigs database, and 63.27% of nucleotides stored in the profile database. 5 of the nine bins (from bin 1 to 5) were taxonomically assigned to Anthoxanthum, Deschampsia, Equus, Viola, and Coelodonta respectively. The merged profile database that was generated with the minimum contig length of 1,000 contained 461 contigs, which correspond to 10% of all contigs, and 37% of all nucleotides found in the contigs database.

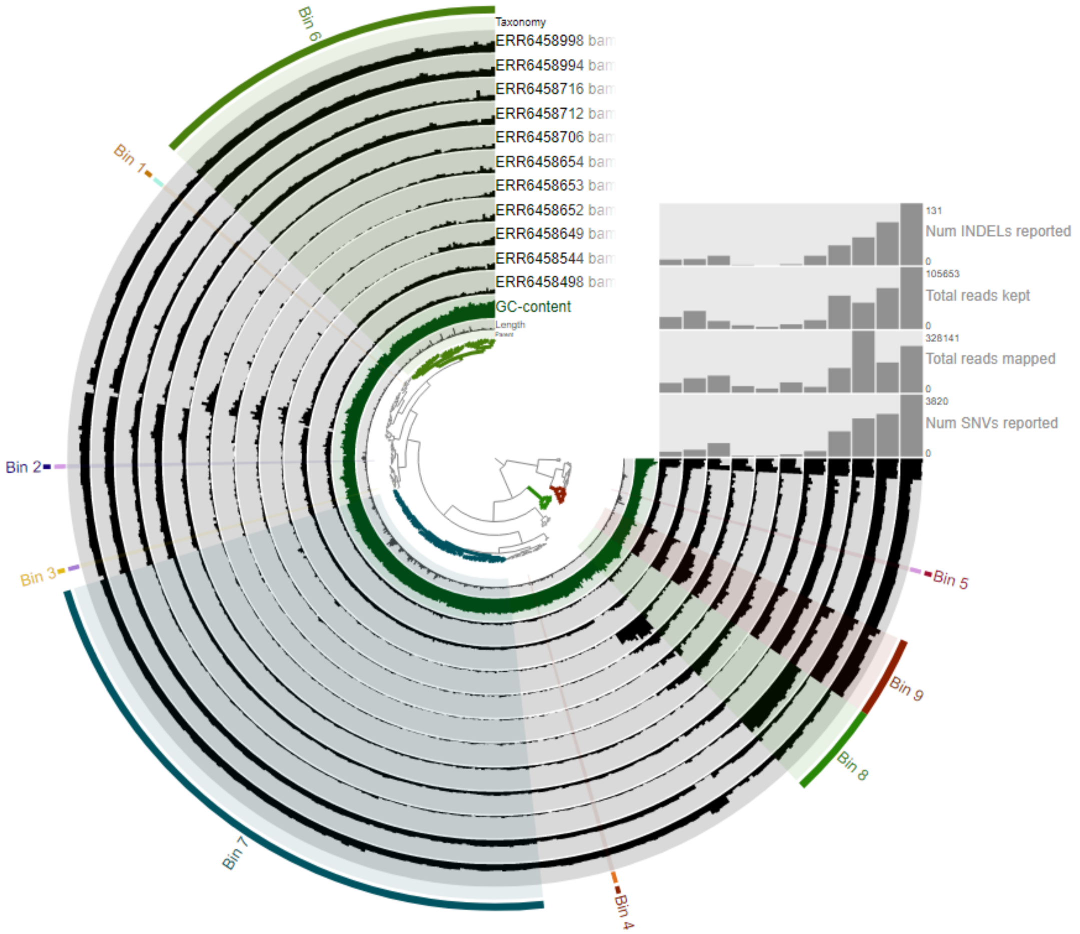


Figure 17.Hierarchical clustering tree of the contigs in Northwest and central Siberia LGM

Table 15. Taxonomically assigned bins abundance in Northwest and central Siberia LGM

	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5
length	1266	1111	2125	1037	1365
gc_content	0.37	0.37	0.39	0.41	0.36
ERR6458498_bam_sorted	0.43	1.87	1.39	0.37	2.26
ERR6458544_bam_sorted	0.38	1.48	1.27	0.63	2.87
ERR6458649_bam_sorted	0.05	0.08	0.18	0	1.65
ERR6458652_bam_sorted	0	0.17	0.31	0	0.7
ERR6458653_bam_sorted	0.09	0.03	0.17	0.04	0.4
ERR6458654_bam_sorted	0.04	0.12	0.24	0.13	0.56
ERR6458706_bam_sorted	0.36	0.66	1.12	0.40	1.78
ERR6458712_bam_sorted	1.30	2.81	4.11	1	6.46
ERR6458716_bam_sorted	0.71	1.92	2.20	0.08	1.29
ERR6458994_bam_sorted	1.82	5.20	4.33	2.09	12.13

ERR6458998_bam_sorted	2.38	7.55	5.07	4.76	18.84
Taxonomy	Anthoxanthum	Deschampsia	Equus	Viola	Coelodonta

A summary of each bin is presented in table. The first 5 bins had only one contig that was taxonomclly assigned. Bin 7 was the largest and had the longest total length. N50 of all bins is relatively high, and thus good.

Table 16. Bins general summary report Northwest and central Siberia LGM

Bins	total_length	num_contigs	N50	GC_content
Bin_1	1266	1	1266	37.203791469194314
Bin_2	1111	1	1111	37.263726372637265
Bin_3	2125	1	2125	38.682352941176475
Bin_4	1037	1	1037	41.46576663452266
Bin_5	1365	1	1365	36.556776556776555
Bin_6	179250	79	2506	39.23492735420048
Bin_7	393634	135	3387	44.85069998372705
Bin_8	27222	21	1288	48.63081935631369
Bin_9	31780	18	1788	40.32072630722199

Northwest and central Siberia Mid-Holocene

Bins that were identified in the merged profile database and stored in the database as "NandSMH" collection, describe 6 bins accounting for 102,889 nucleotides, which represent 24.10% of all nucleotides stored in the contigs database, and 74.54% of nucleotides stored in the profile database. Two of the five bins were taxonomically assigned. Bin 1 was assigned to Tephroseris and bin 2 was assigned to Elymus. The merged profile database that was generated with the minimum contig length of 1,000 contained 58 contigs, which correspond to 7% of all contigs, and 32% of all nucleotides found in the contigs database.

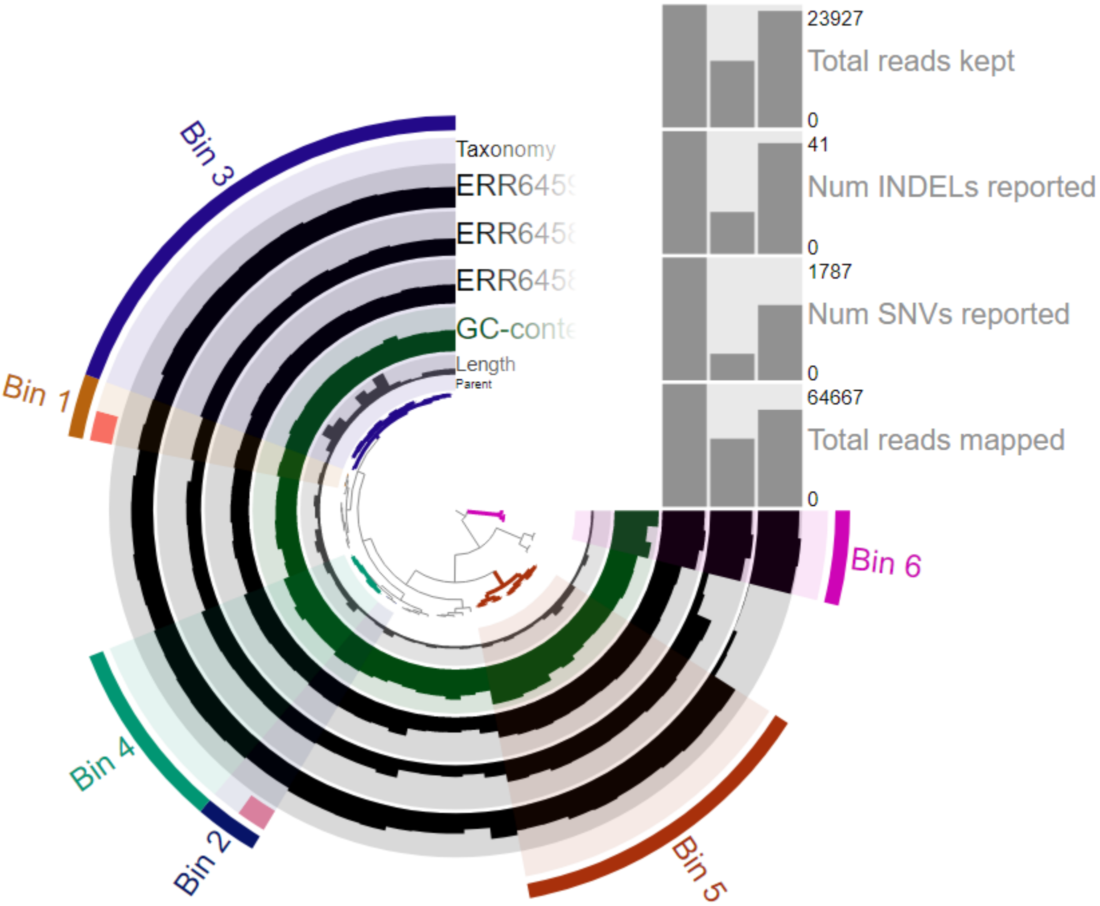


Figure 18. Hierarchical clustering tree of the contigs in Northwest and central Siberia Mid-Holocene

Table 17. Taxonomically assigned bins abundance in Northwest and central Siberia Mid-Holocene

	Bin 1	Bin 2
Length	2182	1549
gc_content	0.32	0.41
ERR6458973_bam_sorted	6.51	8.95
ERR6458974_bam_sorted	2.94	4.20
ERR6459003_bam_sorted	8.14	8.23
Taxonomy	Tephroseris	Elymus

A summary of each bin is presented in table. The contigs that were taxonomically assigned were binned with the other neighboring contigs according to the coverage information. Bin 3 had the biggest number of contigs and the largest total length. N50 of all bins is relatively high, and thus good.

Table 18. Bins general summary report Northwest and central Siberia Mid-Holocene

Bins	total_length	num_contigs	N50	GC_content
Bin_1	3919	2	2182	32.30173392286394
Bin_2	2670	2	1549	41.56205062228286
Bin_3	56203	15	5803	33.27757036642776
Bin_4	15017	6	2430	38.48215391621412
Bin_5	21954	10	2211	45.46571257054988
Bin_6	3126	3	1026	56.06400259909032

Northwest and central Siberia Pre-LGM

Bins that were identified in the merged profile database and stored in the database as "NandSPLGM" collection, described 6 bins accounting for 1,713,997 nucleotides, which represent 26.09% of all nucleotides stored in the contigs database, and 77.01% of nucleotides stored in the profile database. Only one contig was taxonomically assigned to Dryas and binned in bin 1. The merged profile database that was generated with the minimum contig length of 1,000 contained 1,030 contigs, which correspond to 9% of all contigs, and 33% of all nucleotides found in the contigs database.

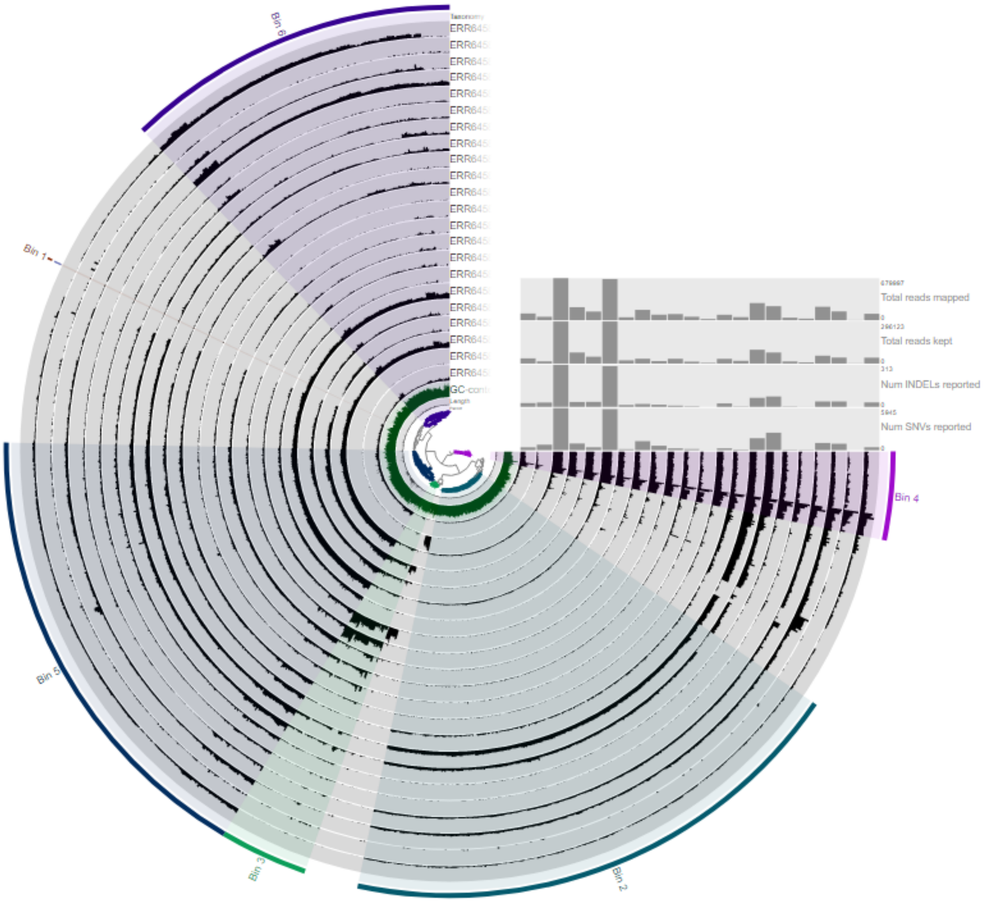


Figure 19.Hierarchical clustering tree of the contigs in Northwest and central Siberia Pre-LGM

Table 19. Taxonomically assigned bins abundance in Northwest and central Siberia Pre-LGM

Item Name	bin 1
ERR6458631_bam_sorted	0.12
ERR6458684_bam_sorted	0.06
ERR6458702_bam_sorted	0.23
ERR6458703_bam_sorted	0.43
ERR6458705_bam_sorted	0.20
ERR6458707_bam_sorted	0.06
ERR6458717_bam_sorted	0.07
ERR6458718_bam_sorted	0.06
ERR6458719_bam_sorted	0.35
ERR6458970_bam_sorted	0.44
ERR6458980_bam_sorted	0.03
Taxonomy	Dryas

A summary of each bin is presented in table. Bin 2 had the biggest number of contigs, and Bin one only contained one contig that was taxonomically assigned. Bin 5 had the largest total length. N50 of all bins is relatively high, and thus good.

Table 20. Bins general summary report Northwest and central Siberia Pre-LGM

Bins	total_length	num_contigs	N50	GC_content
Bin_1	1079	1	1079	28.915662650602407
Bin_2	391466	257	1516	40.608380026032656
Bin_3	60812	44	1304	39.548867413048924
Bin_4	82690	44	2018	35.195553496372185
Bin_5	829129	231	5711	40.58711597445279
Bin_6	348821	167	2347	42.05091091627324

Northwest and central Siberia Late Holocene

Bins that were identified in the single profile database ERR6458978 bam sorted and stored in the database as "NandSLH" collection, describe 9 bins accounting for 76,057 nucleotides, which represent 32.36% of all nucleotides stored in the contigs database, and 100.00% of nucleotides stored in the profile database. 2 contigs were taxonomically assigned and binned as bin 1 and Bin 2. Bin 1 was taxonomically assigned to Bison and Bin 2 to the plant Tephroseris. The single profile database that was generated with the minimum contig length of 1,000 contained 43 contigs, which correspond to 10% of all contigs, and 32% of all nucleotides found in the contigs database.

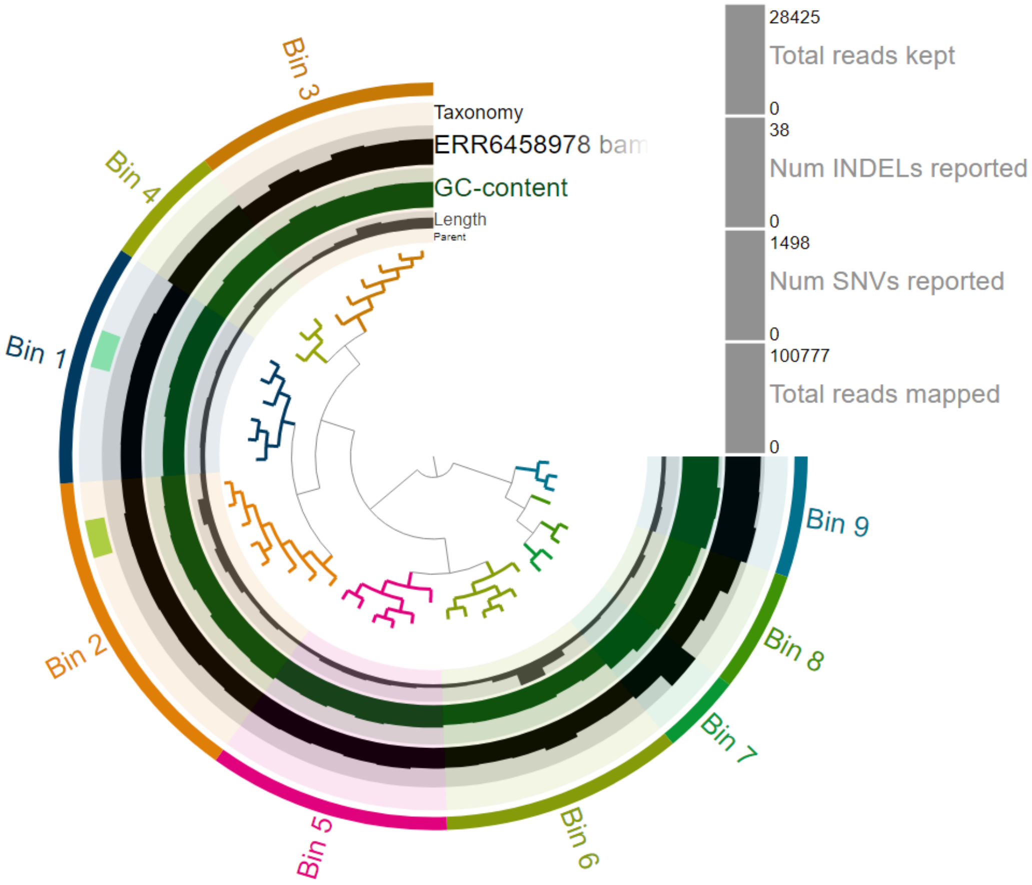


Figure 20.Hierarchical clustering tree of the contigs in Northwest and central Siberia Late Holocene

Table 21. Taxonomically assigned bins abundance in Northwest and central Siberia Late Holocene

	Bin 1	Bin 2
length	1314	3360
gc_content	0.36	0.37
ERR6458978_bam_sorted	11.02	13.09
Taxonomy	Bison	Tephroseris

A summary of each bin is presented in table. Bin 2 that was taxonomically assigned had the biggest number of contigs and largest total length. N50 of all bins is relatively high, and thus good.

Table 22. Bins general summary report Northwest and centralSiberia late Holocene

Bins	total_length	num_contigs	N50	GC_content
Bin_1	8640	6	1519	34.56858276297456
Bin_2	16137	8	2184	37.13526594378133
Bin_3	14289	6	3119	39.21239029677896
Bin_4	4083	3	1234	36.56069105811056
Bin_5	8595	6	1554	34.47378130151148
Bin_6	13382	6	2901	30.7655622262206
Bin_7	2393	2	1353	46.24907612712491
Bin_8	3879	3	1045	50.80860892697444
Bin_9	4659	3	1469	57.11114832943741

North America 50+

Bins that were identified in the merged profile database 'merged profile' and stored in the database as "na50" collection, describe 8 bins accounting for 629,853 nucleotides, which represent 26.12% of all nucleotides stored in the contigs database, and 81.42% of nucleotides stored in the profile database. Four of the bins were taxonomically assigned (bin1-bin4) to Alopecurus, canis, Equisetum, and papaver respectively. The merged profile database that was generated with the minimum contig length of 1,000 contained 447 contigs, which correspond to 11% of all contigs, and 32% of all nucleotides found in the contigs database.

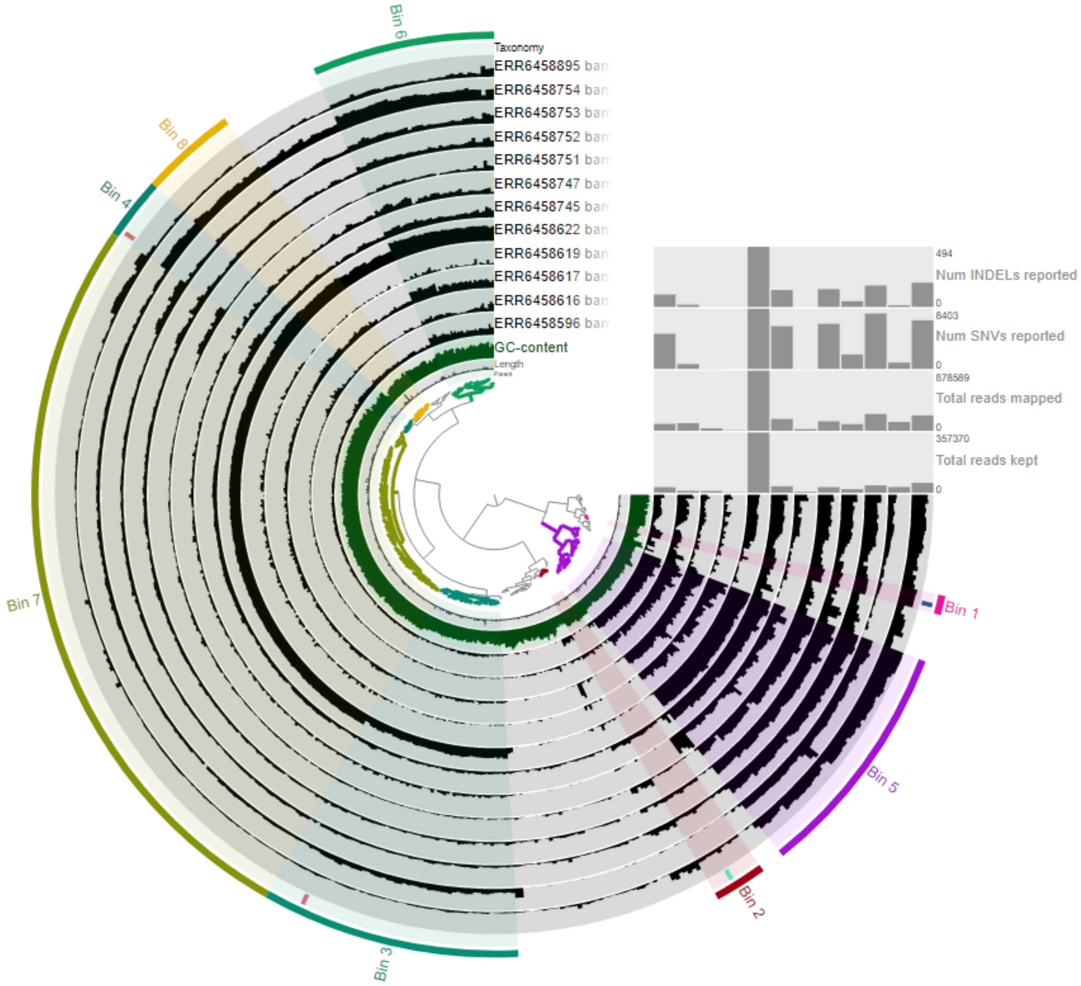


Figure 21. Hierarchical clustering tree of the contigs in North America 50+

Table 23. Taxonomically assigned bins abundance North America 50+

	Bin 1	Bin 2	Bin 3	Bin 4
length	1416	1146	2786	1120
gc_content	0.31	0.37	0.36	0.40
ERR6458596_bam_sorted,	6.98	0	0.22	1.69
ERR6458616_bam_sorted	2.17	0.48	0.20	0.27
ERR6458617_bam_sorted	0.58	0	0.26	0.95
ERR6458619_bam_sorted	0.18	0	0.03	0
ERR6458622_bam_sorted	6.36	0	8.01	22.36
ERR6458745_bam_sorted	6.84	1.53	1.13	1.38
ERR6458747_bam_sorted	0.21	0	0.45	0.82
ERR6458751_bam_sorted	12.70	0.46	0.49	0.29
ERR6458752_bam_sorted	2.54	0.79	0.62	1.02
ERR6458754_bam_sorted	12.48	1.08	1.05	0.77

ERR6458754_bam_sorted	0.69	0	1.42	4.86
ERR6458895_bam_sorted	23.36	0.87	0.71	1.23
Taxonomy	Alopecurus	Canis	Equisetum	Papaver

A summary of each bin is presented in table. Bin 6 contained the largest number of contigs and the biggest total length value. The bins assigned taxonomically were relatively small due to the depth of coverage. N50 of all bins is relatively high, and thus good.

Table 24. Bins general summary report North America 50+

Bins	total_length	num_contigs	N50	GC_content
Bin_1	5355	4	1359	32.26316370299117
Bin_2	21329	11	2200	40.24128598490689
Bin_3	101249	54	2004	36.14306301933084
Bin_4	20288	13	1367	38.17018336987229
Bin_5	102435	50	2206	35.01272626334031
Bin_6	64157	38	1882	38.0581511595216
Bin_7	267491	157	1827	44.408928901408714
Bin_8	47549	20	2468	35.19375641169047

North America Early Holocene

Bins that were identified in the merged profile database and stored in the database as "naEH" collection, describe 6 bins accounting for 407,766 nucleotides, which represent 21.65% of all nucleotides stored in the contigs database, and 55.74% of nucleotides stored in the profile database. Two of the six bins were taxonomically assigned to Phleum and Puccinellia respectively. The merged profile database that was generated with the minimum contig length of 1,000 contained 327 contigs, which correspond to 9% of all contigs, and 38% of all nucleotides found in the contigs database.

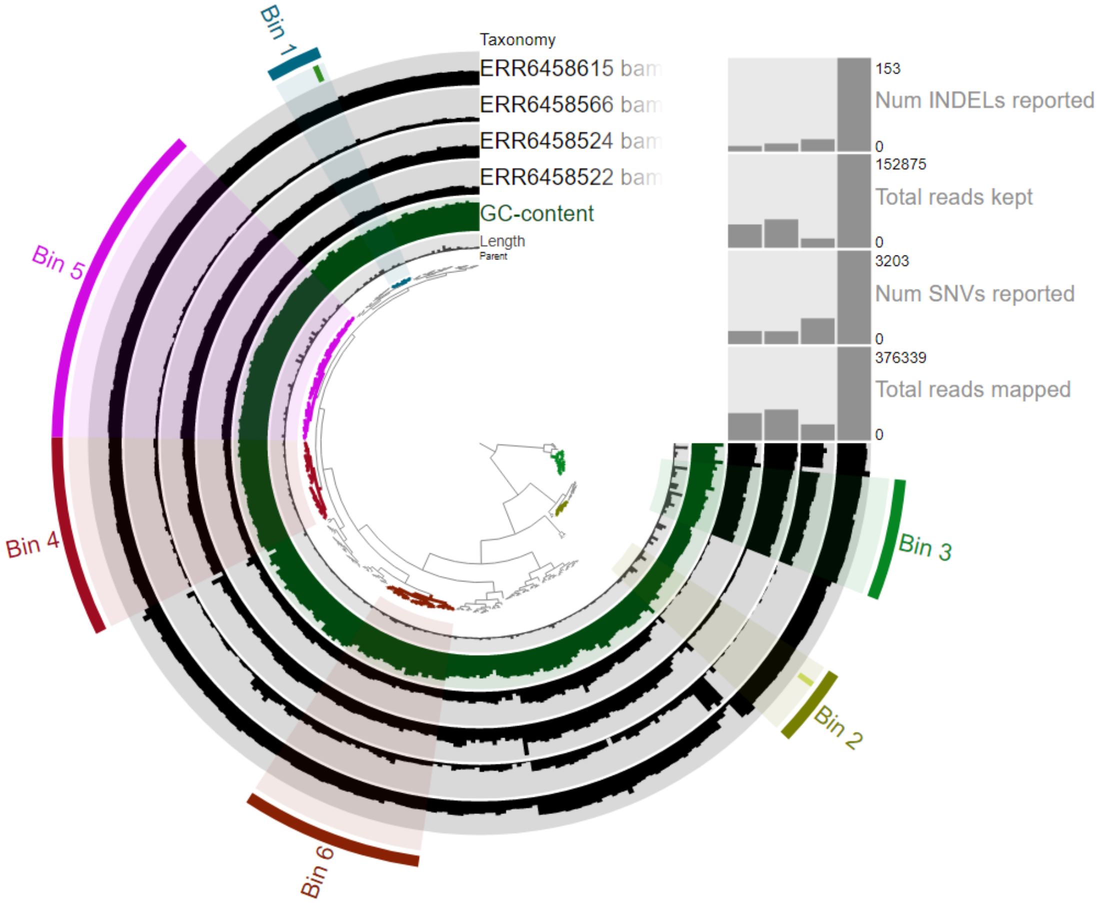


Figure 22. Hierarchical clustering tree of the contigs in North America Early Holocene

Table 25. Taxonomically assigned bins abundance North America Early Holocene

	Bin 1	Bin 2
length	1323	3335
gc_content	0.42	0.30
ERR6458522_bam_sorted,	1.76	0.06
ERR6458524_bam_sorted	2.14	0.50
ERR6458566_bam_sorted	0.32	0.03
ERR6458615_bam_sorted	5.04	13.30
Taxonomy	Phleum	Puccinellia

A summary of each bin is presented in table. Bin 5 contained the largest number of contigs and biggest number of total length value. N50 of all bins is relatively high, and thus good.

Table 26. Bins general summary report North America Early Holocene

Bins	total_length	num_contigs	N50	GC_content

Bin_1	18542	9	2638	42.866347637293984
Bin_2	47807	13	3609	32.69474429565328
Bin_3	101528	20	7589	35.18092122078389
Bin_4	53277	33	1573	45.11031049034663
Bin_5	129074	54	2668	45.42056441073029
Bin_6	57538	30	2073	44.00726784972712

North America Late glacial

Bins that were identified in the merged profile database 'merged profile' and stored in the database as "naLG" collection, describe 7 bins accounting for 1,153,697 nucleotides, which represent 19.80% of all nucleotides stored in the contigs database, and 70.59% of nucleotides stored in the profile database. Out of the even bins, one was taxonomically assigned to Epilobium (bin 1). The merged profile database that was generated with the minimum contig length of 1,000 contained 550 contigs, which correspond to 4% of all contigs, and 28% of all nucleotides found in the contigs database.

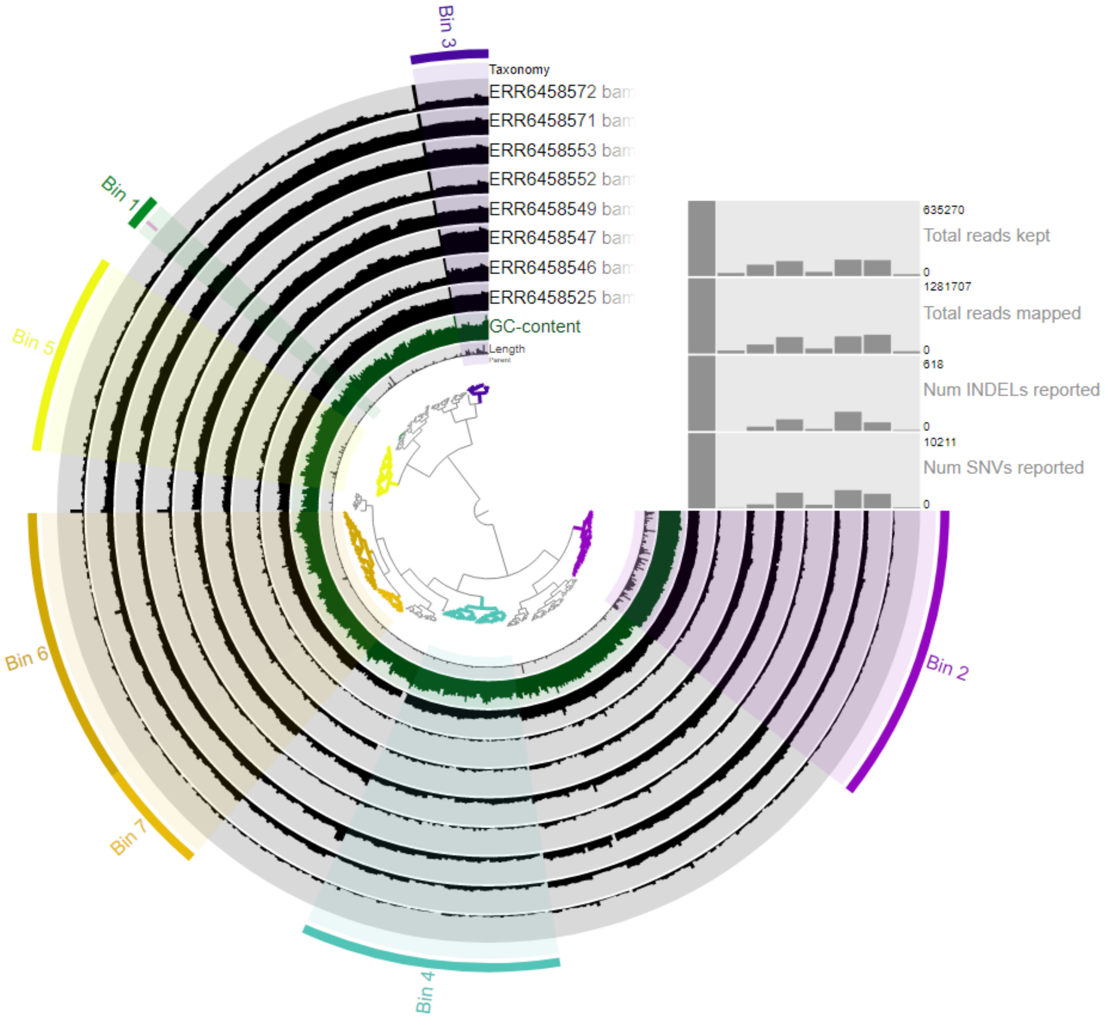


Figure 23. Hierarchical clustering tree of the contigs in North America Late glacial

Table 27. Taxonomically assigned bins abundance North America Late glacial

	Bin 1
Length	1583
gc_content	0.36
ERR6458525_bam_sorted,	26.59
ERR6458546_bam_sorted	0.75
ERR6458547_bam_sorted	2.38
ERR6458549_bam_sorted	9.83
ERR6458552_bam_sorted	1.26
ERR6458553_bam_sorted	9.67
ERR6458571_bam_sorted	5.96
ERR6458572_bam_sorted	0.60
Taxonomy	Epilobium

A summary of each bin is presented in table. Bin 1 contained 9 contigs that were assigned following the depth coverage of the one contig that was taxonomically assigned. Bin 2 had the largest number of contigs and total length value. N50 of all bins is relatively high, and thus good.

Table 28. Bins general summary report North America Late glacial

Bins	total_length	num_contigs	N50	GC_content
Bin_1	17376	9	2184	37.42865376816666
Bin_2	610584	77	11388	44.89020720634895
Bin_3	128127	20	8774	35.280885719409426
Bin_4	98673	67	1368	37.71805577923173
Bin_5	135086	52	3653	38.20233823687496
Bin_6	117001	71	1508	37.0691148648451
Bin_7	46850	29	1626	34.19213786492754

North America LGM

Bins that were identified in the merged profile database and stored in the database as "naLGM" collection, describe 9 bins accounting for 210,306 nucleotides, which represent 9.52% of all nucleotides stored in the contigs database, and 54.16% of nucleotides stored in the profile database. Out of the nine bins, one was taxonomically assigned to the plant Vulpes. The merged profile database that was generated with the minimum contig length of 1,000 contained 198 contigs, which correspond to 3% of all contigs, and 17% of all nucleotides found in the contigs database.

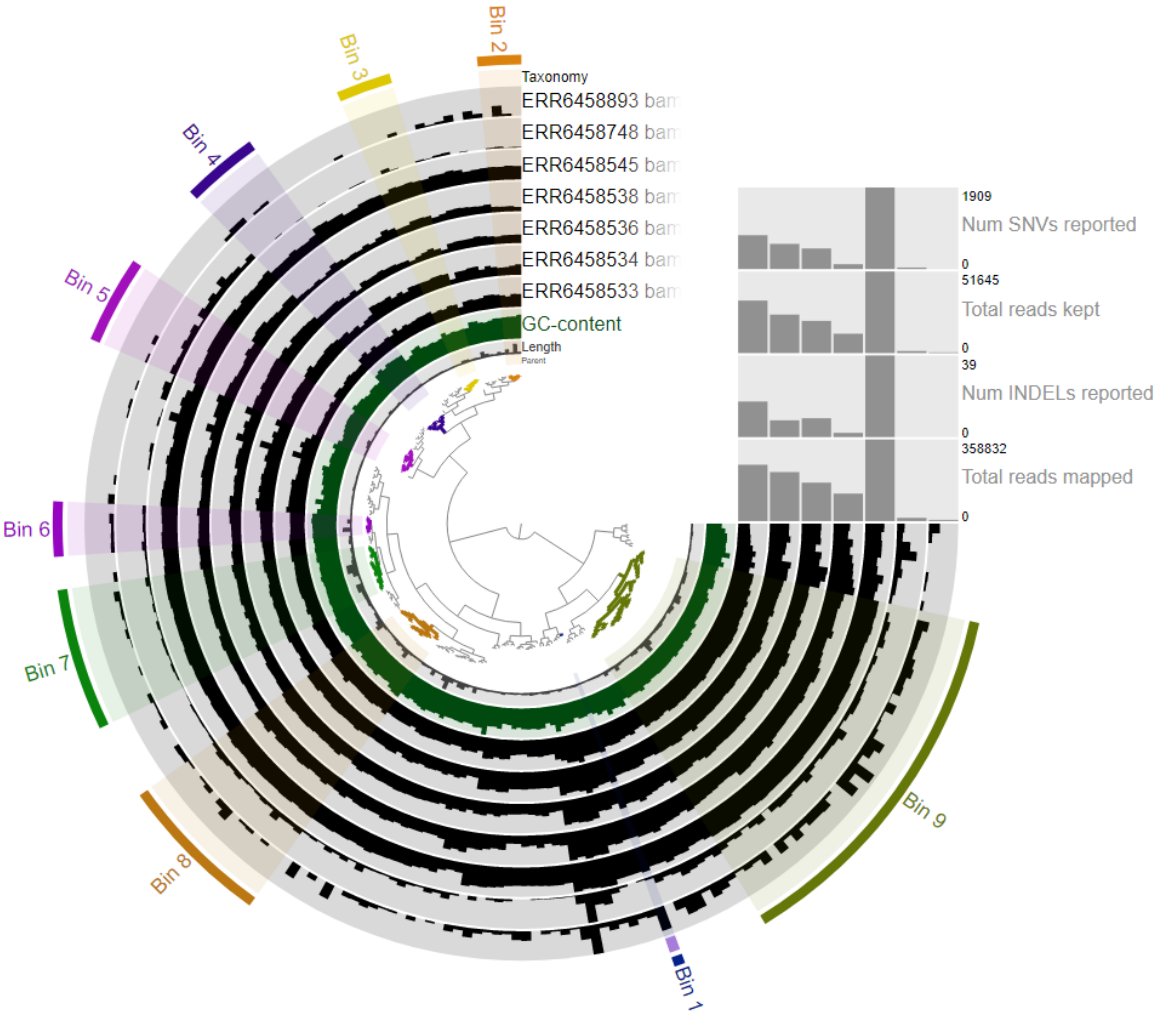


Figure 24.Hierarchical clustering tree of the contigs in North America LGM

Table 29. Taxonomically assigned bins abundance North America LGM

	Bin 1
length	1249
gc_content	0.33
ERR6458533_bam_sorted,	13.05
ERR6458534_bam_sorted,	11.35
ERR6458536_bam_sorted,	7.71
ERR6458538_bam_sorted,	3.43
ERR6458545_bam_sorted,	21.84
ERR6458748_bam_sorted,	0.10
ERR6458893_bam_sorted,	0.40
Taxonomy	Vulpes

A summary of each bin is presented in table. The contig assigned taxonomically was binned in bin 1 alone. Bin 9 was the largest in number of contigs and in total length value. N50 of all bins is relatively high, and thus good.

Table 30. Bins general summary report North America LGM

Bins	total_length	num_contigs	N50	GC_content
Bin_1	1249	1	1249	33.306645316253004
Bin_2	12322	4	6505	44.472877857836195
Bin_3	7548	5	1983	36.813244337824145
Bin_4	9393	7	1328	44.635073662931916
Bin_5	12869	8	1641	30.905889963534417
Bin_6	12185	5	2889	48.01245533640973
Bin_7	43984	13	4555	43.88596616529884
Bin_8	31473	14	2548	47.82686733036456
Bin_9	79283	34	2512	34.324520416438155

North America Pre-LGM

Bins that were identified in the merged profile database and stored in the database as "naPLGM" collection, describe 9 bins accounting for 319,397 nucleotides, which represent 7.77% of all nucleotides stored in the contigs database, and 59.54% of nucleotides stored in the profile database. Three of the bins, bin1 – bin 3, were taxonomically assigned to Canis, Salix, and Juncus respectively. The merged profile database that was generated with the minimum contig length of 1,000 contained 289 contigs, which correspond to 2% of all contigs, and 13% of all nucleotides found in the contigs database.

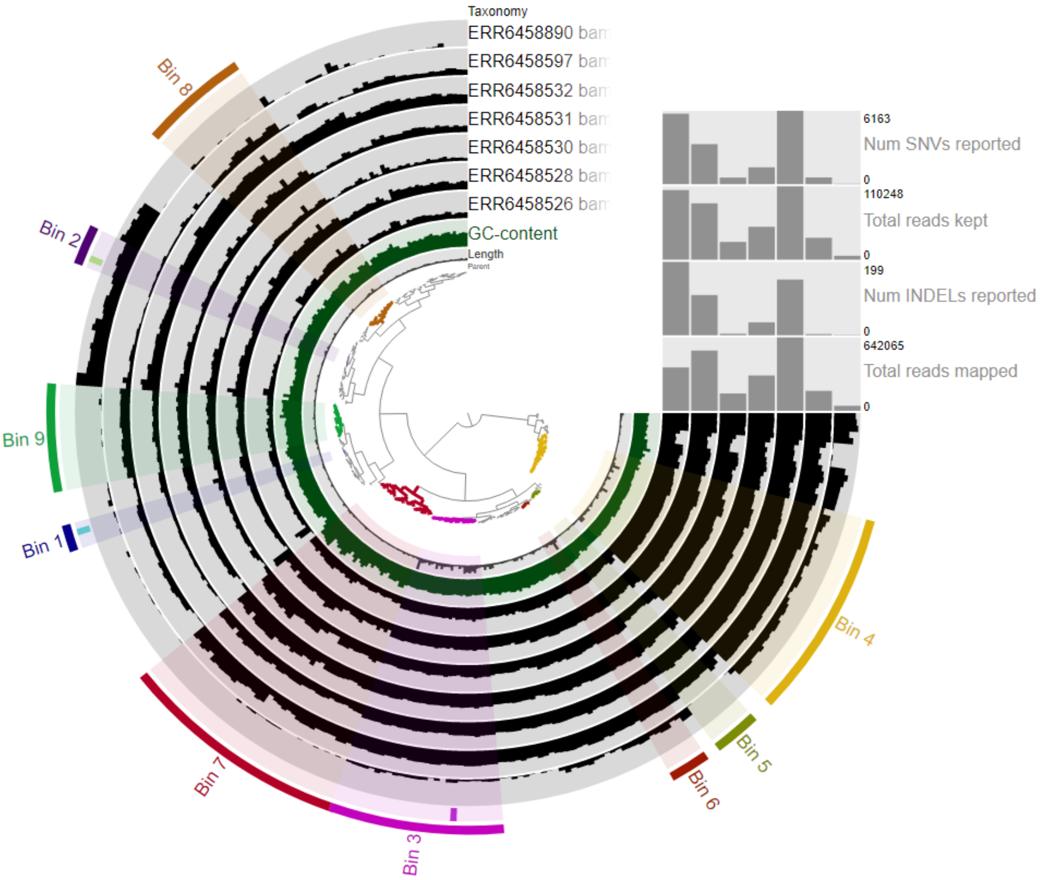


Figure 25. Hierarchical clustering tree of the contigs in North America Pre-LGM

Table 31. Taxonomically assigned bins abundance North America Pre-LGM

	Bin 1	Bin 2	Bin 3
length	1035	1913	1740
gc_content	0.42	0.42	0.45
ERR6458526_bam_sorted,	1.51	0.79	6.63
ERR6458528_bam_sorted,	2.41	3.69	6.47
ERR6458530_bam_sorted,	1.34	0.98	2.03
ERR6458531_bam_sorted,	1.21	1.29	4.67
ERR6458532_bam_sorted,	1.05	5.77	8.30
ERR6458597_bam_sorted,	1.28	0.29	4.13
ERR6458890_bam_sorted,	0.04	1.84	0.3
Taxonomy	Canis	Salix	Juncus

A summary of each bin is presented in table. Bin 7 had the biggest number of contigs, and bin 3 that was taxonomically assigned and binned according to the depth of coverage, had the largest total length value. N50 of all bins is relatively high, and thus good.

Table 32. Bins general summary report North America Pre-LGM

Bins	total_length	num_contigs	N50	GC_content
Bin_1	4599	4	1229	43.030741544069286
Bin_2	10226	6	1913	40.598667230311406
Bin_3	99340	26	5127	44.43701860530535
Bin_4	80504	31	2989	34.07308814719286
Bin_5	8422	7	1154	37.66438759472494
Bin_6	24115	6	3616	48.127130868358336
Bin_7	45310	34	1225	49.06142917993259
Bin_8	25513	16	1568	38.19011228088911
Bin_9	21368	16	1256	47.11323014506792

Northeast Siberia Early Holocene

Bins that were identified in the merged profile database and stored in the database as "nsEH" collection, describe 10 bins accounting for 1,070,227 nucleotides, which represent 22.62% of all nucleotides stored in the contigs database, and 70.28% of nucleotides stored in the profile database. Four of these bins were taxonomically assigned to Lagotis, Pyrola, Tephroseris, and Artemisia respectively. The merged profile database that was generated with the minimum contig length of 1000 contained 609 contigs, which correspond to 6% of all contigs, and 32% of all nucleotides found in the contigs database.

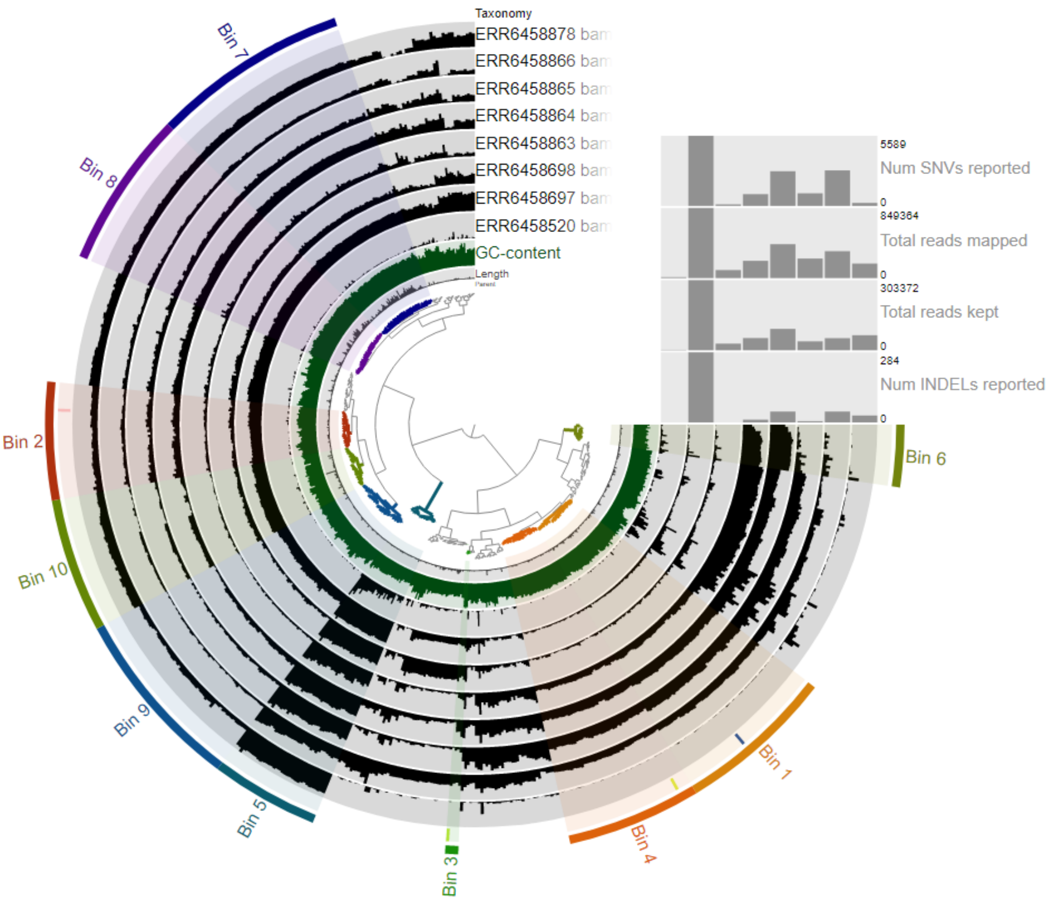


Figure 26. Hierarchical clustering tree of the contigs in Northeast Siberia Early Holocene

Table 33. Taxonomically assigned bins abundance Northeast Siberia Early Holocene

	Bin 1	Bin 2	Bin 3	Bin 4
length	3155	1366	1133	1679
gc_content	0.43	0.39	0.32	0.45
ERR6458520_bam_sorted,	0.02	0.05	0.03	0.03
ERR6458697_bam_sorted,	0.44	14.13	2.87	0.58
ERR6458698_bam_sorted,	0.15	0.53	4.19	0.13
ERR6458863_bam_sorted,	1.24	0.32	0.21	1.19
ERR6458864_bam_sorted,	3.55	1.85	0.43	3.17
ERR6458865_bam_sorted,	2.38	0.86	0.11	1.64
ERR6458866_bam_sorted	2.54	0.70	5.52	1.77
ERR6458878_bam_sorted	0	1.99	0.07	0.02
Taxonomy	Lagotis	Pyrola	Tephroseris	Artemisia

A summary of each bin is presented in table. Bin 8 contained the largest number of total length value and biggest number of contigs. N50 of all bins is relatively high, and thus good.

Table 34. Bins general summary report Northeast Siberia Early Holocene

Bins	total_length	num_contigs	N50	GC_content
Bin_1	108275	48	2592	43.327041720000885
Bin_2	124908	41	3268	35.26799567422481
Bin_3	66298	36	2265	38.003882679887816
Bin_4	5121	4	1226	47.440536516936675
Bin_5	65398	41	1591	43.64944047109401
Bin_6	55199	33	1645	31.885458667485572
Bin_7	27283	19	1261	31.39232234870308
Bin_8	310696	60	7131	44.32091728641376
Bin_9	216606	48	7143	45.651680519675836

Northeast Siberia Mid-Holocene

Bins that were identified in the merged profile database and stored in the database as "nsMH" collection, describe 10 bins accounting for 26,766 nucleotides, which represent 9.89% of all nucleotides stored in the contigs database, and 87.50% of nucleotides stored in the profile database. One of the 10 bins was taxonomically assigned to Elymus. The merged profile database that was generated with the minimum contig length of 1,000 contained 20 contigs, which correspond to 2% of all contigs, and 11% of all nucleotides found in the contigs database.

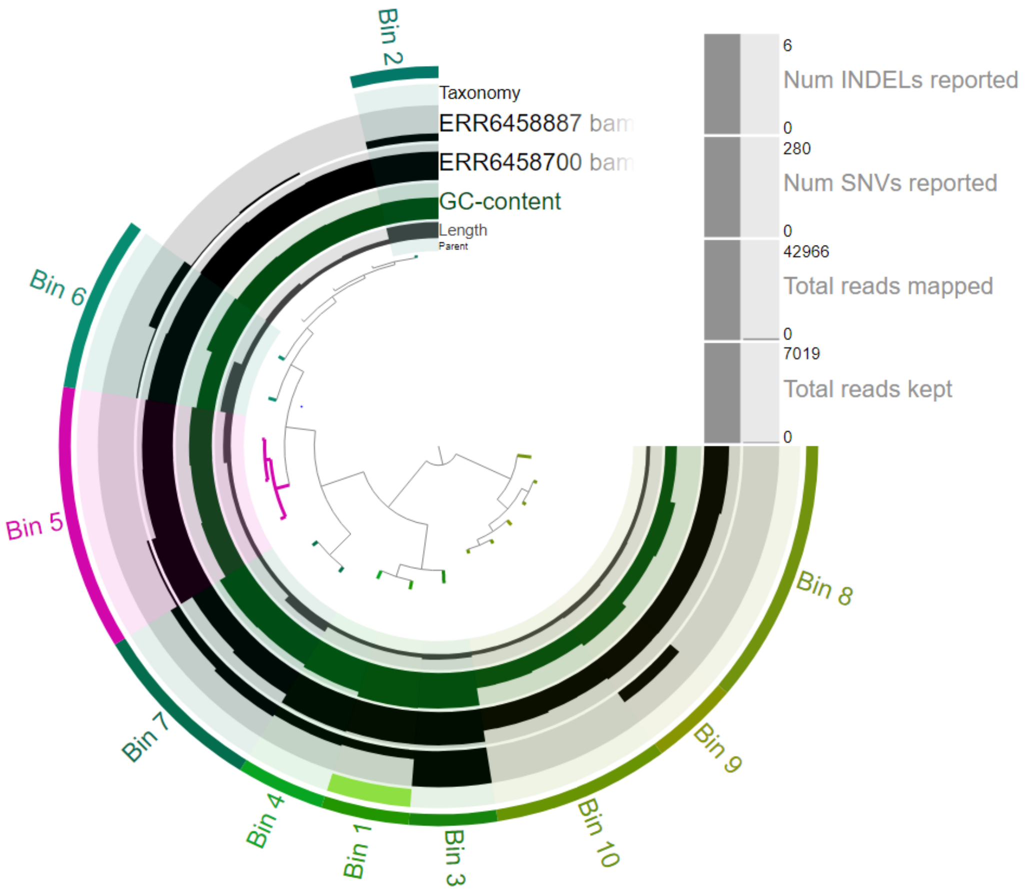


Figure 27.Hierarchical clustering tree of the contigs in Northeast Siberia Mid-Holocene

Table 35. Taxonomically assigned bins abundance Northeast Siberia Mid-Holocene

	Bin 1
length	1060
gc_content	0.61
ERR6458700_bam_sorted,	20.53
ERR6458887_bam_sorted,	0.16
Taxonomy	Elymus

A summary of each bin is presented in table. The number of contigs was extremely limited from 1 to 3 per bin. Bin 6 contained the biggest total length value. N50 of all bins is relatively high, and thus good.

Table 36. Bins general summary report Northeast Siberia Mid-Holocene

Bins	total_length	num_contigs	N50	GC_content
Bin_1	1060	1	1060	60.84905660377359
Bin_10	2080	2	1057	29.48924962383625

Bin_2	4104	1	4104	36.91520467836257
Bin_3	1446	1	1446	59.95850622406639
Bin_4	1071	1	1071	53.50140056022409
Bin_5	4180	3	1115	37.96331062527289
Bin_6	4759	2	3378	36.09174712723509
Bin_7	3551	2	2511	49.87200548356462
Bin_8	3170	3	1009	24.407104709768408
Bin_9	1345	1	1345	33.7546468401487

Northeast Siberia Late glacial

Bins that were identified in the merged profile database and stored in the database as "nsLG" collection, describe 8 bins accounting for 93,339 nucleotides, which represent 9.03% of all nucleotides stored in the contigs database, and 68.59% of nucleotides stored in the profile database. None of them was assigned to any of the input matrix. The merged profile database that was generated with the minimum contig length of 1,000 contained 85 contigs, which correspond to 3% of all contigs, and 13% of all nucleotides found in the contigs database.

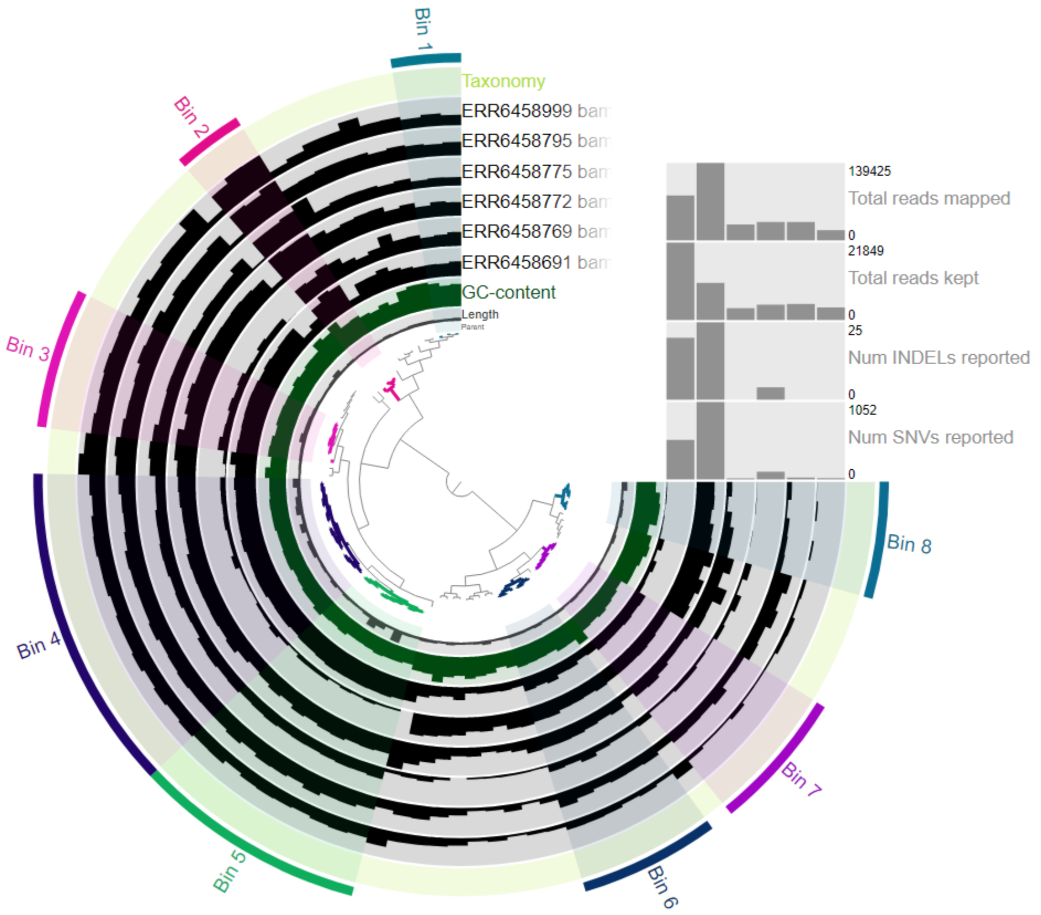


Figure 28. Hierarchical clustering tree of the contigs in Northeast Siberia Late glacial

A summary of each bin is presented in table. The biggest number of contigs was that of Bin 4, which had the biggest value of total length as well. N50 of all bins is relatively high, and thus good.

Table 37. Bins general summary report Northeast Siberia late glacial

Bins	total_length	num_contigs	N50	GC_content
Bin_1	3930	3	1251	43.85195902512673
Bin_2	3120	3	1043	43.44609159778015
Bin_3	11861	6	2050	35.3369586350565
Bin_4	28416	14	2336	32.057641712380175
Bin_5	22508	10	2871	29.683640174424426
Bin_6	7738	6	1352	36.26757509358
Bin_7	7760	6	1425	31.22383324700258
Bin_8	8006	5	1760	43.44091021318711

Northeast Siberia LGM

Bins that were identified in the merged profile database 'merged profile' and stored in the database as "nsLGM" collection, describe 13 bins accounting for 1,965,105 nucleotides, which represent 16.84% of all nucleotides stored in the contigs database, and 73.30% of nucleotides stored in the profile database. Out of the 13 bins only one contigs was assigned taxonomically to Puccinellia. The merged profile database that was generated with the minimum contig length of 1,000 contained 1,319 contigs, which correspond to 5% of all contigs, and 22% of all nucleotides found in the contigs database.

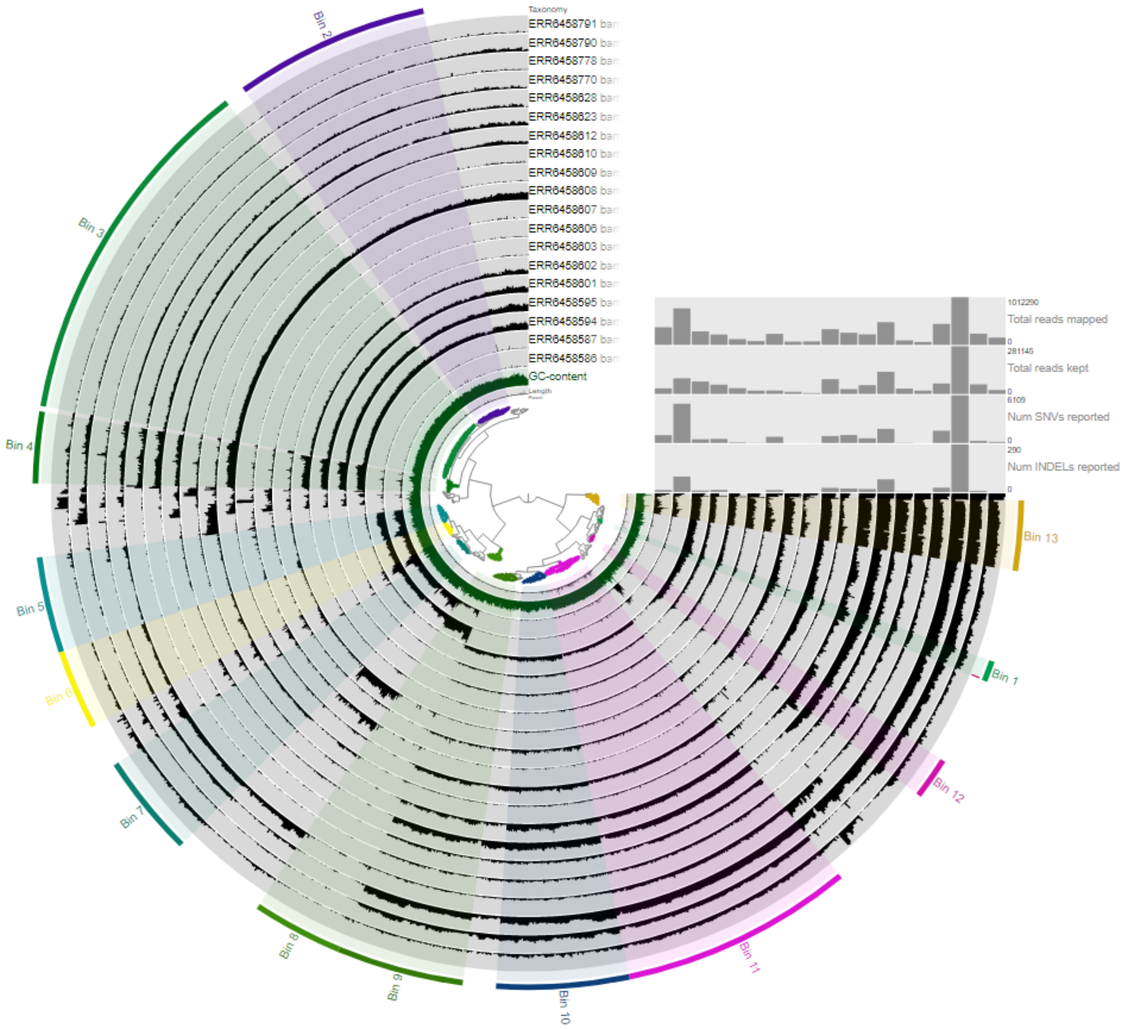


Figure 29. Hierarchical clustering tree of the contigs in Northeast Siberia LGM

Table 38. Taxonomically assigned bins abundance Northeast Siberia LGM

	Bin 1
ERR6458607_bam_sorted,	0
ERR6458608_bam_sorted,	4.39
ERR6458609_bam_sorted,	0.18
ERR6458610_bam_sorted,	0.82
ERR6458612_bam_sorted,	7.10
ERR6458623_bam_sorted,	0.54
ERR6458628_bam_sorted,	0.64
ERR6458770_bam_sorted,	0.18
ERR6458778_bam_sorted,	4.54
ERR6458790_bam_sorted,	2.10
ERR6458791_bam_sorted,	0.43

Taxonomy	Puccinellia

A summary of each bin is presented in table. This represents the biggest number of bin in all of the assemblies. Bin 3 had the biggest number of contigs, and bin 11 represented the largest total length value. N50 of all bins is relatively high, and thus good.

Table 39. Bins general summary report Northeast Siberia LGM

Bins	total_length	num_contigs	N50	GC_content
Bin_1	21108	12	1786	37.84581112661381
Bin_10	138775	76	1930	36.92994636179427
Bin_11	544995	134	5311	44.887135897415604
Bin_12	42745	24	1933	38.688978980364546
Bin_13	63797	40	1532	33.948029554993965
Bin_2	188748	112	1665	41.53685186296929
Bin_3	426516	206	2318	39.96523041817213
Bin_4	61141	41	1611	34.68779173465291
Bin_5	84669	55	1455	43.860500776831195
Bin_6	74040	46	1547	43.36695687992665
Bin_7	114768	59	2208	43.45884984898087
Bin_8	84015	52	1581	39.228931886926915
Bin_9	119788	73	1658	35.61284044490949

Northeast Siberia Pre-LGM

Bins that were identified in the merged profile database 'merged profile' and stored in the database as "nsPLGM" collection, describe 11 bins accounting for 598,304 nucleotides, which represent 12.53% of all nucleotides stored in the contigs database, and 71.08% of nucleotides stored in the profile database. None of the contigs was assigned taxonomically. The merged profile database that was generated with the minimum contig length of 1,000 contained 453 contigs, which correspond to 4% of all contigs, and 17% of all nucleotides found in the contigs database.

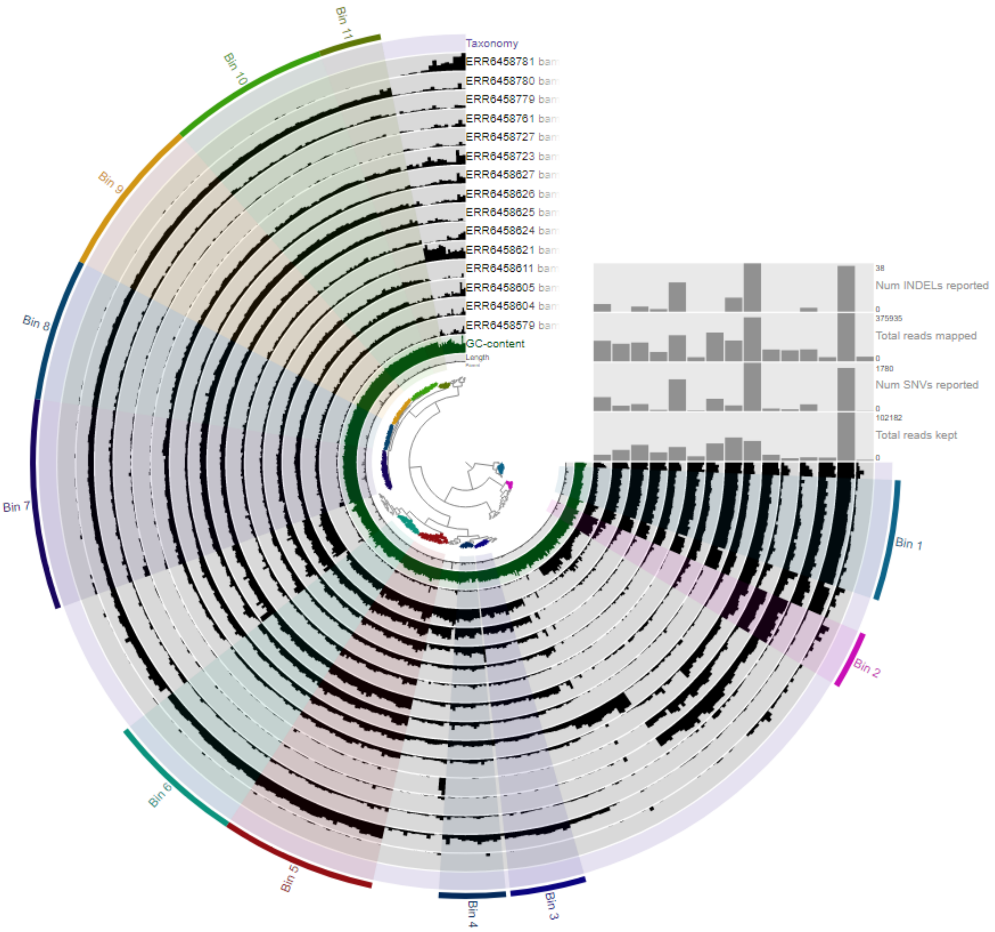


Figure 30Hierarchical clustering tree of the contigs in Northeast Siberia Pre-LGM

A summary of each bin is presented in table. Bin 7 had the longest total length and the largest number of contigs. N50 of all bins is relatively high, and thus good.

Table 40. Bins general summary report Northeast Siberia Pre-LGM

Bins	total_length	num_contigs	N50	GC_content
Bin_1	65971	27	2963	32.2157408368226
Bin_10	65598	36	1952	44.79625062712242
Bin_11	20423	14	1385	43.53113304203486
Bin_2	27850	13	2330	46.31019287296977
Bin_3	29737	17	1873	37.224744375548305
Bin_4	32224	15	2635	34.45635536685946
Bin_5	61544	35	1700	39.93347549496721
Bin_6	48747	31	1561	38.12776269494297
Bin_7	105952	47	2456	44.73896048265449
Bin_8	55066	32	1873	45.05829035015655
Bin_9	85192	36	2490	44.54120862653574

CHAPTER IV : RESULTS

Northeast Siberia Pre-LGM 1

Bins that were identified in the merged profile database and stored in the database as "nsPLGM1" collection, describe 9 bins accounting for 1,152,956 nucleotides, which represent 13.17% of all nucleotides stored in the contigs database, and 70.16% of nucleotides stored in the profile database. Out of them, three were assigned taxonomically to Epilobium, viola, and Agrostis respectively. The merged profile database that was generated with the minimum contig length of 1,000 contained 992 contigs, which correspond to 4% of all contigs, and 18% of all nucleotides found in the contigs database.

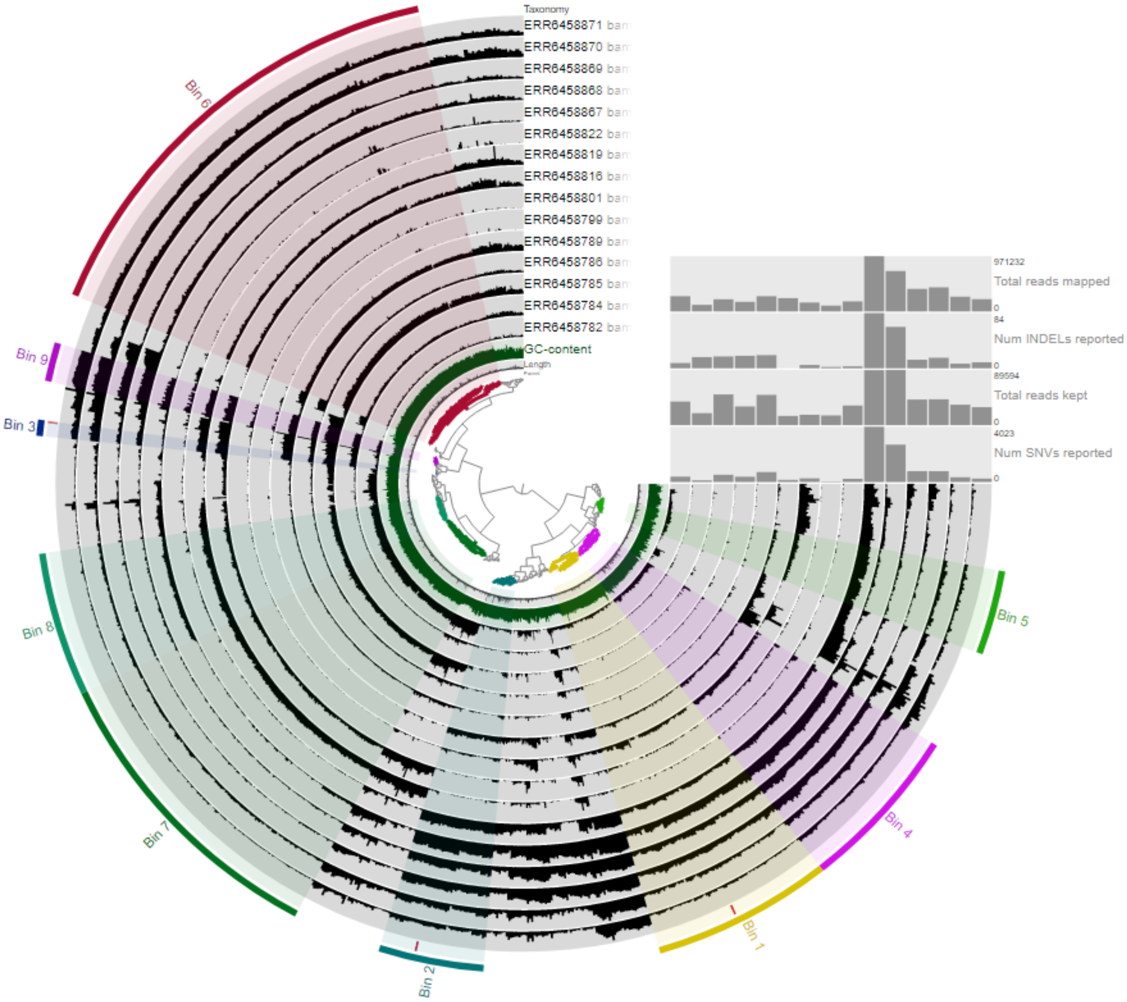


Figure 31.Hierarchical clustering tree of the contigs in Northeast Siberia Pre-LGM 1

Table 41.Taxonomically assigned bins abundance Northeast Siberia Pre-LGM 1

	Bin 1	Bin 2	Bin 3
length	1310	2078	1203
gc_content	0.42	0.43	0.33

ERR6458782_bam_sorted,	0.77	0.40	0.31
ERR6458784_bam_sorted,	0.15	0.16	2.61
ERR6458785_bam_sorted,	0.06	0.23	3.88
ERR6458786_bam_sorted,	0.09	0.27	2.21
ERR6458789_bam_sorted,	0.11	0.31	9.03
ERR6458799_bam_sorted,	0.43	0.84	0.03
ERR6458801_bam_sorted,	0.91	0.32	0.04
ERR6458816_bam_sorted,	0.04	0.09	2.63
ERR6458819_bam_sorted,	0.20	0.08	3.11
ERR6458822_bam_sorted,	0.86	6.07	0.05
ERR6458867_bam_sorted,	3.91	11.25	0.40
ERR6458868_bam_sorted,	1.83	2.70	1.29
ERR6458869_bam_sorted,	1.67	2.91	0.71
ERR6458870_bam_sorted,	0.78	1.05	2.99
ERR6458871_bam_sorted,	0.33	0.85	1.99
Taxonomy	Epilobium	Viola	Agrostis

A summary of each bin is presented in table. Bin 6 represented the biggest number of contigs and largest length out of all of the bins. N50 of all bins is relatively high, and thus good.

Table 42. Bins general summary report Northeast Siberia Pre-LGM

bins	total_length	num_contigs	N50	GC_content
Bin_1	118868	79	1543	43.73697359750038
Bin_2	112011	46	2993	44.93713132082871
Bin_3	8343	7	1164	40.53414296934607
Bin_4	159386	73	2596	44.05909121689478
Bin_5	61640	37	1692	40.11945625271005
Bin_6	357922	201	1879	44.52126374922106
Bin_7	205376	135	1475	40.72289356687146
Bin_8	105765	63	1729	40.97674263817697
Bin_9	23645	17	1315	29.73193386911831

7. Taxonomic alignment

Naturally, not all bins mapped to the reference index. In total, 149 bins were generated from anvio, 38 of them were assigned to plants and animals from the tab delimited matrix. And 97 of them mapped to the mammoth index after the bowtie alignment as shown in table. The abundance level deferred drastically across the fasta files (bins), which is due to the difference in number of contigs and reads length.

Table 43. Mammoth's abundance in the assigned bins

Co assembled files	Number of bins	Number of reads	Overall alignment (%)	Co assembled files	Number of bins	Number of reads	Overall alignment (%)	Co assembled files	Number of bins	Number of reads	Overall alignment (%)
na50	1	72	0.00	NandSLGM	1	17	0.00	nsLG	1	54	0.00
	2	284	0.00		2	15	0.00		2	44	2.27
	3	1337	0.00		3	28	0.00		3	158	1.27
	4	273	0.00		4	14	0.00		4	376	0.53
	5	1358	0.59		5	19	5.26		5	295	0.34
	6	857	0.82		6	2358	0.59		6	104	0.00
	7	3575	0.84		7	5125	0.62		7	106	1.89
	8	624	0.64		8	373	1.34		8	107	0.93
naEH	1	243	0.00		9	423	0.47	nsLGM	1	280	0.00
	2	616	0.00	NandSLH	1	116	0.00		2	2526	0.71
	3	1299	0.23		2	212	0.00		3	5640	0.60
	4	717	1.12		3	187	0.53		4	824	0.85
	5	1692	0.53		4	56	1.79		5	1140	1.14
	6	763	0.66		5	116	0.86		6	995	0.60
	7	3974	0.40		6	177	1.13		7	1523	0.59
naLG	1	231	0.00		7	32	0.00		8	1131	1.33
	2	7746	0.18		8	53	1.89		9	1599	0.50
	3	1633	0.24		9	63	0.00		10	1845	0.98
	4	1333	1.05	NandSMH	1	52	0.00		11	7011	0.31
	5	1764	0.40		2	37	0.00		12	573	1.05
	6	1573	1.21		3	724	0.14		13	854	0.59
	7	626	0.48		4	197	0.51	nsMH	1	15	0.00
naLGM	1	17	0.00		5	289	0.69		2	53	0.00
	2	160	0.00		6	43	0.00		3	20	5.00
	3	100	0.00	NandSPLGM	1	15	0.00		4	15	0.00
	4	128	2.34		2	5270	0.91		5	56	0.00
	5	173	0.00		3	826	0.73		6	63	0.00
	6	160	0.62		4	1100	0.82		7	47	0.00
	7	568	0.35		5	10701	0.39		8	44	0.00

	8	414	0.72		6	4605	0.72		9	18	0.00
	9	1045	0.86	naPLGM	1	63	0.00		10	29	0.00
NandS 50	1	16	0.00		2	137	0.00	**nsPLGM**	1	863	0.58
	2	36	0.00		3	1279	0.00		2	368	0.82
	3	1045	0.48		4	1052	0.86		3	398	0.25
	4	2063	0.78		5	117	1.71		4	423	0.24
NandS EH	1	904	0.22		6	311	0.96		5	819	0.49
	2	274	1.09		7	617	1.13		6	657	0.76
	3	298	2.01		8	342	0.58		7	1392	0.43
	4	552	0.72		9	292	1.37		8	738	1.36
	6	337	0.00	**nsEH**	1	1423	0.00		10	871	0.34
	7	319	0.94		2	877	0.00		11	276	1.09
NandS LG	1	15	0.00		3	71	0.00	**nsPLGM1**	1	1601	0.00
	2	14	0.00		4	876	0.00		2	1465	0.00
	3	16	0.00		5	739	0.81		3	115	0.00
	4	15	0.00		6	368	0.54		4	2101	0.67
	5	32	0.00		7	3958	0.00		5	825	0.97
	6	789	1.39		8	2779	0.29		6	4778	0.88
	7	10496	0.21		9	1215	1.07		7	2767	0.76
	8	2289	0.66		10	1623	0.62		8	1413	0.99
	9	699	0.72						9	322	0.93

Discussion and Conclusion

DISCUSSION AND CONCLUSION

Discussion

The aim of our study was to retrieve the mammoth metagenomic assembled genomes from a set of eDNA samples, and in doing so we got insights on the arctic dynamics of floral and fuana in the studied regions and age interval.The taxonomic assignment was mainly conducted to guide the human binning, and have some information about the abundance and distribution of some species. All of the assigned contigs were then binned into a fasta file for future studies.

The mammoth is thought to be extict in northeast Siberia by the start of Mid-Holocene (7.3 ka), in North America by the end of early Holocene (8.6ka), in north atlatic in late glacial, and survived into the late Holocene as late as 3.9Ka in North central Siberia**(Wang Y and al 2021)**.

The persistence of the steppe tundra vegetation, which consisted of Pleistocene herbaceous plants, was thought to have contributed to the mammoth's survival into the Late Holocene in Northwest and Central Siberia **(Willerslev, E et al, 2014).**In this region 12 plants and 4 animals were assigned in 6 age intervals. Among them was Tephroseris, the herbaceous flowering plant, was assigned in Late Holocene, further reinforcing the hypothesis. In the same clustering tree the Bison was assigned. According to **Stuart A, 2015**, skeletal evidence suggests that bison were widespread in the Arctic during the Pleistocene epoch but had gone extinct by the early Holocene.However, the Bison’s DNA was detected in Mid-Holocene in recente studies (6.4ka) in northeast Siberia **(Wang Y and al 2021)**. Thus, the contigs assigned to Bison indicate that it lived in Arctic Northwest and central Siberia into late Holocene.

In Northeast Siberia samples, 9 plants were assigned in 4 age intervals, where Mid-Holocene assisted the presence of Elymus, a Graminoid plant that was assigned in Northwest and central Siberia Mid-Holocene as well. Belonging to the same group (Graminoid), Phleum was assigned in North America Early Holocene. Graminoids are thought to be the dominant herbaceous plant group, and some results suggest that this group composed the mammoth steppe with other forbs and willow shrubs**(Bigelow, N. H. 2003).** One of the most abundant plant species in all regions and age intervals was Salix, it was assigned in Northwest and central Siberia 50+, late glacial, and North America Pre-LGM. Salix is a woody plant, which became abundant in these regions probably as the postglacial climate warmed up **(Mangerud, J, 2020).** Studies show that the abundance of woody taxa reached its highest point during the transition from the Pleistocene to the Holocene**(Rasmussen S, *et al* 2006).**It is wherepreviously abundant plants rapidly declined, and other plants appeared and became abundant. This vegetation turnover was attributed to climate heating and CO2 reaching the Holocene levels, which caused the transition from the cold-adapted steppe tundra species (like Artemisia) to a mosaic of herbaceous and woody plants**(Monteath AJ, 2021).** Artemisia was present in the hierarchical tree of Northeast Siberia Early Holocene with Lagotis, Pyrola, and Tephroseris. However, additional research has demonstrated that the Holocene's moisture caused the growth of aquatic plants, which resulted in the decline of Artemisia **(Mangerud, J, 2020).**

Out of the 12 animals, 6 were assigned to contigs. Equus (horse) and Coelodonta (rhino) were assigned to Northwest and central Siberia LGM alongside Anthoxanthum, Deschampsia, and Viola. Several articles support the co-occurrence and co-existing of arctic megafauna animals**(Stuart A, 2015; Ritchie M, 2002; Graham R, 2016)**. Animal coexistence is supported by skeletal evidence, which demonstrates that bison, equis, and coelodonta lived together for a long time**.**Vulpes was assigned in North America LGM. Ursus was assigned to Northwest and central Siberia Late glacial tree with some plants (Alopecurus, Salix, Utricularia, and Epilobium). Canis was assigned in North America Pre-LGM hierarchical tree with Salix and Juncus, and North America 50+ with Alopecurus, Equisetum, and Papaver.

After the human-guided binning, the 152 MAGs were retrieved in fasta format, 38 of them were already assigned taxonomically and guided us in the binning. The other 114 were aligned against the mammoth full DNA index to assign the mammoth’s DNA abundance and retrieve the mammoth MAGs.

The limiting bias in the bioinformatics classification of short fragments of degraded ancient DNA from the large complex genomes of plants and animals have presented an obvious obstacle. Nonetheless, seeing the nature of the samples (being ancient DNA and eDNA of complex organisms), the mapped contigs could be binned into metagenomic assembled genomes. The mapping levels ranged from 0.1 to 5 and were assigned to 97 bins, with a big difference in the length of the reads.In Northwest and centrak Siberia late Holocene the 5 bins out of the 9 were aligned to the index. Bin 3 had the biggest number of reads (187bp).

DISCUSSION AND CONCLUSION

Conclusion

The study was conducted on a set of arctic eDNA metagenome data, extracted and sequenced by Wang and al 2021. The dataset was downloaded from ENA EMBL into WSL ubuntu 20.04, where it was analyzed. The goal of the analyses was to extract metagenomic assembled genomes by doing a targeted binning of the assembled samples, to get more insight into the genetic structure and dynamics of the studied population.Despite the fact that numerous new algorithms are frequently developed to effectively assemble and bin short reads, few, if any, of them concentrate on the eDNA of complex genomes. More effective bioinformatics tools are needed to gain more understanding of the woolly mammoth's genomic structure as part of the effort to make this miraculous and potentially game-changing discovery. Reviving the mammoth may provide the solution to the global warming and climate change that the world has been facing for a few decades. Even though this work in a small step to understand genomic structure of a giant complex mammalian, we believe that metagenome sequencing of eDNA holds great potential to pangenomic and dynamics of extinct species. From the perspective of the generated data, we can undertake several analyses that may answer questions about the mammoth’s genomic structure and extinction. From the perspective of the generated data, we can undertake several analyses that may answer questions about the mammoth’s abundance and extinction. Using these MAGs and other available mammoth datasets, it is possible to do a pan genomic or environment distribution analysis.

DISCUSSION AND CONCLUSION

References

Adrian M. Lister and Andrei V. Sher. The Origin and Evolution of the Woolly Mammoth. Science 294, 1094 (2001); DOI: 10.1126/science.1056370

Astakhov, V. I. & Isayeva, L. L. The Ice Hill - an Example of Retarded Deglaciation in Siberia. Quaternary Sci Rev 7, 29-40, doi:Doi 10.1016/0277-3791(88)90091-1 (1988).

Ayling M, Clark MD, Leggett RM. New approaches for metagenome assembly with short reads. Brief Bioinform. 2020 Mar 23;21(2):584-594. doi: 10.1093/bib/bbz020. PMID: 30815668; PMCID: PMC7299287.

Baleka S, Varela L, Tambusso PS, Paijmans JLA, Mothé D, Stafford TW Jr, Fariña RA, Hofreiter M. Revisiting proboscidean phylogeny and evolution through total evidence and palaeogenetic analyses including Notiomastodon ancient DNA. iScience. 2021 Dec 4;25(1):103559. doi: 10.1016/j.isci.2021.103559. PMID: 34988402; PMCID: PMC8693454.

Barnes I, Shapiro B, Lister A, Kuznetsova T, Sher A, Guthrie D, Thomas MG. Genetic structure and extinction of the woolly mammoth, Mammuthus primigenius. Curr Biol. 2007 Jun 19;17(12):1072-5. doi: 10.1016/j.cub.2007.05.035. Epub 2007 Jun 7. PMID: 17555965.

Bigelow, N. H. Climate change and Arctic ecosystems: 1. Vegetation changes north of 55°N between the last glacial maximum, mid-Holocene, and present. J. Geophys. Res. 108, https://doi.org/10.1029/2002jd002558 (2003).

Binney, H. et al. Vegetation of Eurasia from the last glacial maximum to present: key biogeographic patterns. Quat. Sci. Rev. 157, 80–97 (2017).

Binney, H. et al. Vegetation of Eurasia from the last glacial maximum to present: key biogeographic patterns. Quat. Sci. Rev. 157, 80–97 (2017).

Boeskorov, G.G., Mashchenko, E.N., Plotnikov, V.V. et al. Adaptation of the woolly mammoth Mammuthus primigenius (Blumenbach, 1799) to habitat conditions in the glacial period. Contemp. Probl. Ecol. 9, 544–553 (2016). https://doi.org/10.1134/S1995425516050024

Brandt AL, Ishida Y, Georgiadis NJ, Roca AL. 2012. Forest elephant mitochondrial genomes reveal that elephantid diversification in Africa tracked climate transitions. Mol. Ecol. 21:1175–89

Brandt AL, Ishida Y, Georgiadis NJ, Roca AL. 2012. Forest elephant mitochondrial genomes reveal that elephantid diversification in Africa tracked climate transitions. Mol. Ecol. 21:1175–89

Castro CJ, Ng TFF. U50: A New Metric for Measuring Assembly Output Based on Non-Overlapping, Target-Specific Contigs. J Comput Biol. 2017 Nov;24(11):1071-1080. doi: 10.1089/cmb.2017.0013. Epub 2017 Apr 18. PMID: 28418726; PMCID: PMC5783553.

Chang D, Knapp M, Enk J, Lippold S, Kircher M, Lister A, MacPhee RD, Widga C, Czechowski P, Sommer R, Hodges E, Stümpel N, Barnes I, Dalén L, Derevianko A, Germonpré M, Hillebrand-Voiculescu A, Constantin S, Kuznetsova T, Mol D, Rathgeber T, Rosendahl W, Tikhonov AN, Willerslev E, Hannon G, Lalueza-Fox C, Joger U, Poinar H, Hofreiter M, Shapiro B. The evolutionary and phylogeographic history of woolly mammoths: a comprehensive mitogenomic analysis. Sci Rep. 2017 Mar 22;7:44585. doi: 10.1038/srep44585. PMID: 28327635; PMCID: PMC5361112.

Chang, D. et al. The evolutionary and phylogeographic history of woolly mammoths: a comprehensive mitogenomic analysis. Scientific Reports 7, 44585, doi:10.1038/srep44585 (2017).

Chang, D., Knapp, M., Enk, J. et al. The evolutionary and phylogeographic history of woolly mammoths: a comprehensive mitogenomic analysis. Sci Rep 7, 44585 (2017). https://doi.org/10.1038/srep44585

Clark SC, Egan R, Frazier PI, Wang Z. ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. Bioinformatics. 2013 Feb 15;29(4):435-43. doi: 10.1093/bioinformatics/bts723. Epub 2013 Jan 9. PMID: 23303509.

Clark, P. U. et al. The Last Glacial Maximum. Science 325, 710–714 (2009).

Comstock KE, Georgiadis N, Pecon-Slattery J, Roca AL, Ostrander EA, et al. 2002. Patterns of molecular genetic variation among African elephant populations. Mol. Ecol. 11:2489–98

Debruyne, R. et al. Out of America: Ancient DNA evidence for a new world origin of late Quaternary woolly mammoths. Curr. Biol. 18, 1320–1326, doi: 10.1016/j.cub.2008.07.061 (2008).

Dickinson A, Yeung KY, Donoghue J, Baker MJ, Kelly RD, McKenzie M et al. The regulation of mitochondrial DNA copy number in glioblastoma cells. Cell Death Differ 2013; 20: 1644–1653.

Drummond, A., Rambaut, A., Shapiro, B., and Pybus, O. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. 22, 1185–1192.

Eggert LS, Rasner CA, Woodruff DS. 2002. The evolution and phylogeography of the African elephant inferred from mitochondrial DNA sequence and nuclear microsatellite markers. Proc. R. Soc. Lond. B Biol. Sci. 269:1993–2006

Enk J, Devault A, Widga C, Saunders J, Szpak P, Southon J, Rouillard J-M, Shapiro B, Golding GB, Zazula G, Froese D, Fisher DC, MacPhee RDE and Poinar H (2016) Mammuthus Population Dynamics in Late Pleistocene North America: Divergence, Phylogeography, and Introgression. Front. Ecol. Evol. 4:42. doi: 10.3389/fevo.2016.00042

Enk, J. et al. Mammuthus population dynamics in Late Pleistocene North America: Divergence, phylogeography and introgression. Frontiers in Ecology and Evolution 4, 42, doi: 10.3389/evo.2016.00042 (2016).

Enk, J. et al. Mammuthus population dynamics in Late Pleistocene North America: Divergence, phylogeography and introgression. Frontiers in Ecology and Evolution 4, 42, doi: 10.3389/evo.2016.00042 (2016).

Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016 Oct 1;32(19):3047-8. doi: 10.1093/bioinformatics/btw354. Epub 2016 Jun 16. PMID: 27312411; PMCID: PMC5039924.

Fry E, Kim SK, Chigurapti S, Mika KM, Ratan A, Dammermann A, Mitchell BJ, Miller W, Lynch VJ. Functional Architecture of Deleterious Genetic Variants in the Genome of a Wrangel Island Mammoth. Genome Biol Evol. 2020 Mar 1;12(3):48-58. doi: 10.1093/gbe/evz279. PMID: 32031213; PMCID: PMC7094797.

Garfias-Gallegos, D. et al. (2022). Metagenomics Bioinformatic Pipeline. In: Pereira-Santana, A., Gamboa-Tuz, S.D., Rodríguez-Zapata, L.C. (eds) Plant Comparative Genomics. Methods in Molecular Biology, vol 2512. Humana, New York, NY. https://doi.org/10.1007/978-1-

0716-2429-6_10

Gilbert MT, Tomsho LP, Rendulic S, Packard M, Drautz DI, et al. 2007. Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. Science 317:1927–30

Graham, R. W. et al. Timing and causes of mid-Holocene mammoth extinction on St. Paul Island, Alaska. Proc. Natl Acad. Sci. USA 113, 9310–9314 (2016).

Graham, R. W. et al. Timing and causes of mid-Holocene mammoth extinction on St. Paul Island, Alaska. Proc. Natl Acad. Sci. USA 113, 9310–9314 (2016).

Graham, R.W., Belmecheri, S., Choy, K., Culleton, B.J., Davies, L.J., Froese, D., Heintzman, P.D., Hritz, C., Kapp, J.D., Newsom, L.A., et al. (2016). Timing and causes of mid-Holocene mammoth extinction on St. Paul Island, Alaska. Proc. Natl. Acad. Sci. U.S.a. 113, 9310–9314.

Guil-Guerrero JL, Tikhonov A, Rodríguez-García I, Protopopov A, Grigoriev S, Ramos-Bueno RP. The fat from frozen mammals reveals sources of essential fatty acids suitable for Palaeolithic and Neolithic humans. PLoS One. 2014 Jan 8;9(1):e84480. doi: 10.1371/journal.pone.0084480. PMID: 24416235; PMCID: PMC3885556.

Guthrie, R.D. (2003). Rapid body size decline in Alaskan Pleistocene horses before extinction. Nature 426, 169–171.

Hampton S. Bandage application. J Wound Care. 1998 Oct;7(9 Suppl):5-8. doi: 10.12968/jowc.1998.7.sup9.5. PMID: 9887730.

Haseeb A, Makki MS, Haqqi TM. Modulation of ten-eleven translocation 1 (TET1), Isocitrate Dehydrogenase (IDH) expression, alpha-Ketoglutarate (alpha-KG), and DNA hydroxymethylation levels by interleukin-1beta in primary human chondrocytes. J Biol Chem 2014; 289: 6877–6885.

Hauf J, Waddell PJ, Chalwatzis N, Joger U, Zimmermann FK. 2000. The complete mitochondrial genome sequence of the African elephant (Loxodonta africana), phylogenetic relationships of Proboscidea to other mammals and D-loop heteroplasmy. Zoology 102:184–95

Henryk Kubiak, DAVID M. HOPKINS, JOHN V. MATTHEWS, CHARLES E. SCHWEGER, STEVEN B. YOUNG. 1982 MORPHOLOGICAL CHARACTERS OF THE

MAMMOTH: AN ADAPTATION TO THE ARCTIC-STEPPE ENVIRONMENT, Paleoecology of Beringia, Academic Press, Pages 281-289, https://doi.org/10.1016/B978-0-12-355860-2.50028-4.

Ho, S. Y. W., Saarma, U., Barnett, R., Haile, J. & Shapiro, B. The effect of inappropriate calibration: Three case studies in molecular ecology. PLoS ONE 3, e1615, doi: 10.1371/journal.pone.0001615 (2008).

Hou Y, Zhang X, Zhou Q, Hong W, Wang Y. Hierarchical Microbial Functions Prediction by Graph Aggregated Embedding. Front Genet. 2021 Jan 18;11:608512. doi: 10.3389/fgene.2020.608512. PMID: 33584804; PMCID: PMC7874084.

Ivanov PL, Wadhams MJ, Roby RK, Holland MM, Weedn VW, Parsons TJ. Mitochondrial DNA sequence heteroplasmy in the Grand Duke of Russia Georgij Romanov establishes the authenticity of the remains of Tsar Nicholas II. Nat Genet 1996; 12: 417–420.

Kelly RD, Mahmud A, McKenzie M, Trounce IA, St John JC. Mitochondrial DNA copy number is regulated in a tissue specific manner by DNA methylation of the nuclear-encoded DNA polymerase gamma A. Nucleic Acids Res 2012; 40:10124–10138.

Kelly RD, Rodda AE, Dickinson A, Mahmud A, Nefzger CM, Lee W et al. Mitochondrial DNA haplotypes define gene expression patterns in pluripotent and differentiating embryonic stem cells. Stem Cells 2013; 31: 703–716.

Koch, P. L. & Barnosky, A. D. Late Quaternary extinctions: state of the debate. Ann. Rev. Ecol. Evol. Syst. 37, 215–250 (2006).

Koch, P. L. & Barnosky, A. D. Late Quaternary extinctions: state of the debate. Ann. Rev. Ecol. Evol. Syst. 37, 215–250 (2006).

Lamurias A, Sereika M, Albertsen M, Hose K, Nielsen TD. Metagenomic binning with assembly graph embeddings. Bioinformatics. 2022 Aug 16:btac557. doi: 10.1093/bioinformatics/btac557. Epub ahead of print. PMID: 35972375.

Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 2012, 9, 357.

Lee WT, Sun X, Tsai TS, Johnson JL, Gould JA, Garama DJ, Gough DJ, McKenzie M, Trounce IA, St John JC. Mitochondrial DNA haplotypes induce differential patterns of DNA methylation that result in differential chromosomal gene expression patterns. Cell Death

Discov. 2017 Sep 11;3:17062. doi: 10.1038/cddiscovery.2017.62. PMID: 28900542; PMCID: PMC5592988.

Lee WTY, Cain JE, Cuddihy A, Johnson J, Dickinson A, Yeung KY et al. Mitochondrial DNA plasticity is an essential inducer of tumorigenesis. Cell Death Discovery 2016; 2: 16016.

Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015 May 15;31(10):1674-6. doi: 10.1093/bioinformatics/btv033. Epub 2015 Jan 20. PMID: 25609793.

Liang KC, Sakakibara Y. MetaVelvet-DL: a MetaVelvet deep learning extension for de novo metagenome assembly. BMC Bioinformatics. 2021 Jun 2;22(Suppl 6):427. doi: 10.1186/s12859-020-03737-6. PMID: 34078257; PMCID: PMC8171044.

Lister, A. M. & Sher, A. V. The origin and evolution of the woolly mammoth. Science 294, 1094–1097, doi: 10.1126/science.1056370 (2001).

Lister, A. M. & Sher, A. V. The origin and evolution of the woolly mammoth. Science 294, 1094–1097, doi: 10.1126/science.1056370 (2001).

Lister, A. M. In The Proboscidea: Trends in Evolution and Paleoecology (eds Jeheskel, Shoshani & Pascal, Tassy) 203–213 (Oxford University Press, 1996).

Lister, A. M., Sher, A. V., van Essen, H. & Wei, G. The pattern and process of mammoth evolution in Eurasia. Quaternary International 126–128, 49–64, doi: 10.1016/j.quaint.2004.04.014 (2005).

Lister, A. M., Sher, A. V., van Essen, H. & Wei, G. The pattern and process of mammoth evolution in Eurasia. Quaternary International 126–128, 49–64, doi: 10.1016/j.quaint.2004.04.014 (2005).

Luo Y, Yu YW, Zeng J, Berger B, Peng J. Metagenomic binning through low-density hashing. Bioinformatics. 2019 Jan 15;35(2):219-226. doi: 10.1093/bioinformatics/bty611. PMID: 30010790; PMCID: PMC6330020.

Maglio VJ. 1973. Origin and evolution of the Elephantidae. Trans. Am. Phil. Soc. Phila. New Ser. 63:1–149

Mangerud, J. The discovery of the Younger Dryas, and comments on the current meaning and usage of the term. Boreas 50, 1–5 (2020).

Mangerud, J. The discovery of the Younger Dryas, and comments on the current meaning and usage of the term. Boreas 50, 1–5 (2020).

Mann, D. H., Groves, P., Kunz, M. L., Reanier, R. E. & Gaglioti, B. V. Ice-age megafauna in Arctic Alaska: extinction, invasion, survival. Quat. Sci. Rev. 70, 91–108 (2013).

Mann, D. H., Groves, P., Kunz, M. L., Reanier, R. E. & Gaglioti, B. V. Ice-age megafauna in Arctic Alaska: extinction, invasion, survival. Quat. Sci. Rev. 70, 91–108 (2013).

Mashchenko, E. Boeskorov, Gennady. Baranov, V. 2013. 438 Morphology of a mammoth calf (Mammuthus primigenius) from Ol’chan (Oimiakon, Yakutia) 47 10.1134/S0031030113040096

Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies. Bioinformatics. 2016 Apr 1;32(7):1088-90. doi: 10.1093/bioinformatics/btv697. Epub 2015 Nov 26. PMID: 26614127.

Miller W, Drautz DI, Ratan A, Pusey B, Qi J, et al. 2008. Sequencing the nuclear genome of the extinct woolly mammoth. Nature 456:387–90

Mirdita M, Steinegger M, Breitwieser F, Söding J, Levy Karin E. Fast and sensitive taxonomic assignment to metagenomic contigs. Bioinformatics. 2021 Mar 17;37(18):3029–31. doi: 10.1093/bioinformatics/btab184. Epub ahead of print. PMID: 33734313; PMCID: PMC8479651.

Monteath AJ, Gaglioti BV, Edwards ME, Froese D. Late Pleistocene shrub expansion preceded megafauna turnover and extinctions in eastern Beringia. Proc Natl Acad Sci U S A. 2021 Dec 28;118(52):e2107977118. doi: 10.1073/pnas.2107977118. PMID: 34930836; PMCID: PMC8719869.

Nikolskiy, P. A., Sulerzhitsky, L. D. & Pitulko, V. V. Last straw versus Blitzkrieg overkill: climate-driven changes in the Arctic Siberian mammoth population and the Late Pleistocene extinction problem. Quat. Sci. Rev. 30, 2309–2328 (2011).

Nissen JN, Johansen J, Allesøe RL, Sønderby CK, Armenteros JJA, Grønbech CH, Jensen LJ, Nielsen HB, Petersen TN, Winther O, Rasmussen S. Improved metagenome binning and assembly using deep variational autoencoders. Nat Biotechnol. 2021 May;39(5):555-560. doi: 10.1038/s41587-020-00777-4. Epub 2021 Jan 4. PMID: 33398153.

Noguchi H, Campbell KL, Ho C, Unzai S, Park SY, Tame JR. Structures of haemoglobin from woolly mammoth in liganded and unliganded states. Acta Crystallogr D Biol Crystallogr. 2012 Nov;68(Pt 11):1441-9. doi: 10.1107/S0907444912029459. Epub 2012 Oct 18. PMID: 23090393.

Novak BJ. De-Extinction. Genes (Basel). 2018 Nov 13;9(11):548. doi: 10.3390/genes9110548. PMID: 30428542; PMCID: PMC6265789.

Nyström, V., Dalén, L., Vartanyan, S., Lidén, K., Ryman, N., and Angerbjörn, A. (2010). Temporal genetic change in the last remaining population of woolly mammoth. Proc. Biol. Sci. 277, 2331–2337.

Nystrom,V.,Dalen,L.,Vartanyan,S.,Liden,K.,Ryman,N.,andAngerbjorn, A.(2010).Temporalgeneticchangeinthelastremainingpopulation ofwoollymammoth. Proc.R.Soc.BBiol.Sci. 277,2331–2337.doi: 10.1098/rspb.2010.0301

Nystrom,V.,Dalen,L.,Vartanyan,S.,Liden,K.,Ryman,N.,andAngerbjorn, A.(2010).Temporal genetic change in the last remaining population of woolly mammoth. Proc.R.Soc.BBiol.Sci. 277,2331–2337.doi: 10.1098/rspb.2010.0301

Palkopoulou,E.,Dalen,L.,Lister,A.M.,Vartanyan,S.,Sablin,M.,Sher,A., etal.(2013).Holarctic geneticstructureandrangedynamicsinthe woolly mammoth. Proc.Biol.Sci. 280:20131910.doi:10.1098/rspb.2013.1910

Papageorgopoulou C, Link K, Rühli FJ. Histology of a Woolly Mammoth (Mammuthus primigenius) Preserved in Permafrost, Yamal Peninsula, Northwest Siberia. Anat Rec (Hoboken). 2015 Jun;298(6):1059-71. doi: 10.1002/ar.23148. PMID: 25998640.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015 Jul;25(7):1043-55. doi: 10.1101/gr.186072.114. Epub 2015 May 14. PMID: 25977477; PMCID: PMC4484387.

Peng Y, Leung HC, Yiu SM, Chin FY. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012 Jun 1;28(11):1420-8. doi: 10.1093/bioinformatics/bts174. Epub 2012 Apr 11. PMID: 22495754.

Pfeiffer T, Schuster S, Bonhoeffer S. Cooperation and competition in the evolution of ATP-producing pathways. Science 2001; 292: 504–507.

Piro, V.C.; Lindner, M.S.; Renard, B.Y. DUDes: A top-down taxonomic profiler for metagenomics. Bioinformatics 2016, 32, 2272–2280.

Poinar HN, Schwarz C, Qi J, Shapiro B, Macphee RD, Buigues B, Tikhonov A, Huson DH, Tomsho LP, Auch A, Rampp M, Miller W, Schuster SC. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science. 2006 Jan 20;311(5759):392-4. doi: 10.1126/science.1123360. Epub 2005 Dec 20. PMID: 16368896.

Poinar HN, Schwarz C, Qi J, Shapiro B, MacPhee RD, et al. 2006. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science 311:392–94

Rasmussen, S. O. et al. A new Greenland ice core chronology for the last glacial termination. J. Geophys. Res. 111, https://doi.org/10.1029/2005jd006079 (2006).

Ritchie, M. in Competition and Coexistence (eds Sommer, U. & Worm, B.) 109–131 (Springer, 2002).

Ritchie, M. in Competition and Coexistence (eds Sommer, U. & Worm, B.) 109–131 Springer, 2002).

Roca AL, Ishida Y, Brandt AL, Benjamin NR, Zhao K, Georgiadis NJ. Elephant natural history: a genomic perspective. Annu Rev Anim Biosci. 2015;3:139-67. doi: 10.1146/annurev-animal-022114-110838. Epub 2014 Dec 8. PMID: 25493538.

Roca AL. 2008. The mastodon mitochondrial genome: a mammoth accomplishment. Trends Genet. 24:49–52

Rogers RL, SlatkinM (2017) Excess of genomic defects in a woolly mammoth on Wrangel island. PLoS Genet 13(3): e1006601. doi:10.1371/journal.pgen.1006601

Rohland N, Reich D, Mallick S, Meyer M, Green RE, et al. 2010. Genomic DNA sequences from mastodon and woolly mammoth reveal deep speciation of forest and savanna elephants. PLOS Biol. 8:e1000564

Rohland, N., Siedel, H. & Hofreiter, M. A rapid column-based ancient DNA extraction method for increased sample throughput. Mol. Ecol. Resources 10, 677–683, doi: 10.1111/j.1755-0998.2009.02824.x (2010).

RYDER, M. Hair of the mammoth. Nature 249, 190–192 (1974). https://doi.org/10.1038/249190a0

Sanders WJ, Gheerbrant E, Harris JM, Saegusa H, Delmer C. 2010. Proboscidea. In Cenozoic Mammals of Africa, eds. L Werdelin, WJ Sanders. Berkeley: Univ. Calif. Press

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015 Oct 1;31(19):3210-2. doi: 10.1093/bioinformatics/btv351. Epub 2015 Jun 9. PMID: 26059717.

Sonia Dheur, Sven J Saupe. Quand la biogéographie métagénomique reconfigure les spatialités. EspacesTemps.net, Association Espaces Temps.net, 2019, in Matthieu Noucher, Irène Hirt et Xavier Arnaud de Sartre, Métrologies critiques de l’espace, 10.26151/espacestemps.net-388a-dd19. halshs-02283892v2

Sønstebø, J. H. et al. Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate. Molecular Ecology Resources 10, 1009-1018, doi:10.1111/j.1755-0998.2010.02855.x (2010).

St John JC, Facucho-Oliveira J, Jiang Y, Kelly R, Salah R. Mitochondrial DNA transmission, replication and inheritance: a journey from the gamete through the embryo and into offspring and embryonic stem cells. Hum Reprod Update 2010; 16: 488–509.

St John JC, Facucho-Oliveira J, Jiang Y, Kelly R, Salah R. Mitochondrial DNA transmission, replication and inheritance: a journey from the gamete through the embryo and into offspring and embryonic stem cells. Hum Reprod Update 2010;16: 488–509.

Stuart, A. J. Late Quaternary megafaunal extinctions on the continents: a short review. Geol. J. 50, 338–363 (2015).

Thomas, M.G. (2012). The flickering genes of the last mammoths. Mol. Ecol. 21, 3379–3381.

Tikhomirov, B. A. “An Expedition That Never Was-Benkendorf’s Expedition to the River Indigirka.” The Geographical Journal, vol. 128, no. 4, 1962, pp. 443–46. JSTOR, https://doi.org/10.2307/1792040. Accessed 16 Sep. 2022.

Tran Q, Phan V. Assembling Reads Improves Taxonomic Classification of Species. Genes (Basel). 2020 Aug 17;11(8):946. doi: 10.3390/genes11080946. PMID: 32824429; PMCID: PMC7465921.

Tsai TS, Rajasekar S, St John JC. The relationship between mitochondrial DNA haplotype and the reproductive capacity of domestic pigs (Sus scrofa domesticus). BMC Genetics 2016; 17: 67.

van der Valk, T., Pečnerová, P., Díez-del-Molino, D. et al. Million-year-old DNA sheds light on the genomic history of mammoths. Nature 591, 265–269 (2021). https://doi.org/10.1038/s41586-021-03224-9

Vasil'Chuk, Y., Punning, J., & Vasil'Chuk, A. (1997). Radiocarbon Ages of Mammoths in Northern Eurasia: Implications for Population Development and Late Quaternary Environment. Radiocarbon, 39(1), 1-18. doi:10.1017/S0033822200040856

Wallace DC. Bioenergetics in human evolution and disease: implications for the origins of biological complexity and the missing genetic variation of common diseases. Philos Trans R Soc B Biol Sci 2013; 368: 20120267.

Wang, Y., Pedersen, M.W., Alsos, I.G. et al. Late Quaternary dynamics of Arctic biota from ancient environmental genomics. Nature 600, 86–92 (2021). https://doi.org/10.1038/s41586-021-04016-x

Wei, G., Taruno, H., Jin, C. & Xie, F. The earliest specimens of the steppe mammoth, Mammuthus trogontherii, from the Early Pleistocene Nihewan Formation, North China. Earth Science 57, 289–298, doi: 10.1007/s11430-010-4001-4 (2003).

Willerslev, E. et al. Fifty thousand years of Arctic vegetation and megafaunal diet. Nature 506, 47–51 (2014).

Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014 Mar 3;15(3):R46. doi: 10.1186/gb-2014-15-3-r46. PMID: 24580807; PMCID: PMC4053813.

Wu YW, Singer SW. Recovering Individual Genomes from Metagenomes Using MaxBin 2.0. Curr Protoc. 2021 May;1(5):e128. doi: 10.1002/cpz1.128. PMID: 33961733.

Zhang Z, Zhang L. METAMVGL: a multi-view graph-based metagenomic contig binning algorithm by integrating assembly and paired-end graphs. BMC Bioinformatics. 2021 Jul 22;22(Suppl 10):378. doi: 10.1186/s12859-021-04284-4. PMID: 34294039; PMCID: PMC8296540.

Wang, Y., Pedersen, M.W., Alsos, I.G. et al. Late Quaternary dynamics of Arctic biota from ancient environmental genomics. Nature 600, 86–92 (2021). https://doi.org/10.1038/s41586-021-04016-x

Stuart, A. J. Late Quaternary megafaunal extinctions on the continents: a short review. Geol. J. 50, 338–363 (2015).

Baleka S, Varela L, Tambusso PS, Paijmans JLA, Mothé D, Stafford TW Jr, Fariña RA, Hofreiter M. Revisiting proboscidean phylogeny and evolution through total evidence and palaeogenetic analyses including Notiomastodon ancient DNA. iScience. 2021 Dec 4;25(1):103559. doi: 10.1016/j.isci.2021.103559. PMID: 34988402; PMCID: PMC8693454.

Ritchie, M. in Competition and Coexistence (eds Sommer, U. & Worm, B.) 109–131 (Springer, 2002).

Rasmussen, S. O. et al. A new Greenland ice core chronology for the last glacial termination. J. Geophys. Res. 111, https://doi.org/10.1029/2005jd006079 (2006).

Mangerud, J. The discovery of the Younger Dryas, and comments on the current meaning and usage of the term. Boreas 50, 1–5 (2020).

Novak BJ. De-Extinction. Genes (Basel). 2018 Nov 13;9(11):548. doi: 10.3390/genes9110548. PMID: 30428542; PMCID: PMC6265789.

Mangerud, J. The discovery of the Younger Dryas, and comments on the current meaning and usage of the term. Boreas 50, 1–5 (2020).

Binney, H. et al. Vegetation of Eurasia from the last glacial maximum to present: key biogeographic patterns. Quat. Sci. Rev. 157, 80–97 (2017).

Clark, P. U. et al. The Last Glacial Maximum. Science 325, 710–714 (2009).

Roca AL, Ishida Y, Brandt AL, Benjamin NR, Zhao K, Georgiadis NJ. Elephant natural history: a genomic perspective. Annu Rev Anim Biosci. 2015;3:139-67. doi: 10.1146/annurev-animal-022114-110838. Epub 2014 Dec 8. PMID: 25493538.

Chang, D., Knapp, M., Enk, J. et al. The evolutionary and phylogeographic history of woolly mammoths: a comprehensive mitogenomic analysis. Sci Rep 7, 44585 (2017). https://doi.org/10.1038/srep44585

Lister, A. M., Sher, A. V., van Essen, H. & Wei, G. The pattern and process of mammoth evolution in Eurasia. Quaternary International 126–128, 49–64, doi: 10.1016/j.quaint.2004.04.014 (2005).

Monteath AJ, Gaglioti BV, Edwards ME, Froese D. Late Pleistocene shrub expansion preceded megafauna turnover and extinctions in eastern Beringia. Proc Natl Acad Sci U S A. 2021 Dec 28;118(52):e2107977118. doi: 10.1073/pnas.2107977118. PMID: 34930836; PMCID: PMC8719869.

Papageorgopoulou C, Link K, Rühli FJ. Histology of a Woolly Mammoth (Mammuthus primigenius) Preserved in Permafrost, Yamal Peninsula, Northwest Siberia. Anat Rec (Hoboken). 2015 Jun;298(6):1059-71. doi: 10.1002/ar.23148. PMID: 25998640.

Mashchenko, E. Boeskorov, Gennady. Baranov, V. 2013. 438 Morphology of a mammoth calf (Mammuthus primigenius) from Ol'chan (Oimiakon, Yakutia) 47 10.1134/S0031030113040096

Enk J, Devault A, Widga C, Saunders J, Szpak P, Southon J, Rouillard J-M, Shapiro B, Golding GB, Zazula G, Froese D, Fisher DC, MacPhee RDE and Poinar H (2016) Mammuthus Population Dynamics in Late Pleistocene North America: Divergence, Phylogeography, and Introgression. Front. Ecol. Evol. 4:42. doi: 10.3389/fevo.2016.00042

Poinar HN, Schwarz C, Qi J, Shapiro B, Macphee RD, Buigues B, Tikhonov A, Huson DH, Tomsho LP, Auch A, Rampp M, Miller W, Schuster SC. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science. 2006 Jan 20;311(5759):392-4. doi: 10.1126/science.1123360. Epub 2005 Dec 20. PMID: 16368896.

Barnes I, Shapiro B, Lister A, Kuznetsova T, Sher A, Guthrie D, Thomas MG. Genetic structure and extinction of the woolly mammoth, Mammuthus primigenius. Curr Biol. 2007 Jun 19;17(12):1072-5. doi: 10.1016/j.cub.2007.05.035. Epub 2007 Jun 7. PMID: 17555965.

Willerslev, E. et al. Fifty thousand years of Arctic vegetation and megafaunal diet. Nature 506, 47–51 (2014).

Fry E, Kim SK, Chigurapti S, Mika KM, Ratan A, Dammermann A, Mitchell BJ, Miller W, Lynch VJ. Functional Architecture of Deleterious Genetic Variants in the Genome of a Wrangel Island Mammoth. Genome Biol Evol. 2020 Mar 1;12(3):48-58. doi: 10.1093/gbe/evz279. PMID: 32031213; PMCID: PMC7094797.

Chang D, Knapp M, Enk J, Lippold S, Kircher M, Lister A, MacPhee RD, Widga C, Czechowski P, Sommer R, Hodges E, Stümpel N, Barnes I, Dalén L, Derevianko A, Germonpré M, Hillebrand-Voiculescu A, Constantin S, Kuznetsova T, Mol D, Rathgeber T, Rosendahl W, Tikhonov AN, Willerslev E, Hannon G, Lalueza-Fox C, Joger U, Poinar H, Hofreiter M, Shapiro B. The evolutionary and phylogeographic history of woolly mammoths: a comprehensive mitogenomic analysis. Sci Rep. 2017 Mar 22;7:44585. doi: 10.1038/srep44585. PMID: 28327635; PMCID: PMC5361112.

RYDER, M. Hair of the mammoth. Nature 249, 190–192 (1974). https://doi.org/10.1038/249190a0

Boeskorov, G.G., Mashchenko, E.N., Plotnikov, V.V. et al. Adaptation of the woolly mammoth Mammuthus primigenius (Blumenbach, 1799) to habitat conditions in the glacial period. Contemp. Probl. Ecol. 9, 544–553 (2016). https://doi.org/10.1134/S1995425516050024

Graham, R. W. et al. Timing and causes of mid-Holocene mammoth extinction on St. Paul Island, Alaska. Proc. Natl Acad. Sci. USA 113, 9310–9314 (2016).

Enk, J. et al. Mammuthus population dynamics in Late Pleistocene North America: Divergence, phylogeography and introgression. Frontiers in Ecology and Evolution 4, 42, doi: 10.3389/evo.2016.00042 (2016).

Lister, A. M. & Sher, A. V. The origin and evolution of the woolly mammoth. Science 294, 1094–1097, doi: 10.1126/science.1056370 (2001).

Rohland, N., Siedel, H. & Hofreiter, M. A rapid column-based ancient DNA extraction method for increased sample throughput. Mol. Ecol. Resources 10, 677–683, doi: 10.1111/j.1755-0998.2009.02824.x (2010).

Henryk Kubiak, DAVID M. HOPKINS, JOHN V. MATTHEWS, CHARLES E. SCHWEGER, STEVEN B. YOUNG. 1982 MORPHOLOGICAL CHARACTERS OF THE MAMMOTH: AN ADAPTATION TO THE ARCTIC-STEPPE ENVIRONMENT, Paleoecology of Beringia, Academic Press, Pages 281-289, https://doi.org/10.1016/B978-0-12-355860-2.50028-4.

Rogers RL, SlatkinM (2017) Excess of genomic defects in a woolly mammoth on Wrangel island. PLoS Genet 13(3): e1006601. doi:10.1371/journal.pgen.1006601

Adrian M. Lister and Andrei V. Sher. The Origin and Evolution of the Woolly Mammoth. Science 294, 1094 (2001); DOI: 10.1126/science.1056370

Hauf J, Waddell PJ, Chalwatzis N, Joger U, Zimmermann FK. 2000. The complete mitochondrial genome sequence of the African elephant (Loxodonta africana), phylogenetic relationships of Proboscidea to other mammals and D-loop heteroplasmy. Zoology 102:184–95

Gilbert MT, Tomsho LP, Rendulic S, Packard M, Drautz DI, et al. 2007. Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. Science 317:1927–30

Brandt AL, Ishida Y, Georgiadis NJ, Roca AL. 2012. Forest elephant mitochondrial genomes reveal that elephantid diversification in Africa tracked climate transitions. Mol. Ecol. 21:1175–89

Maglio VJ. 1973. Origin and evolution of the Elephantidae. Trans. Am. Phil. Soc. Phila. New Ser. 63:1–149

Sanders WJ, Gheerbrant E, Harris JM, Saegusa H, Delmer C. 2010. Proboscidea. In Cenozoic Mammals of Africa, eds. L Werdelin, WJ Sanders. Berkeley: Univ. Calif. Press

Rohland N, Reich D, Mallick S, Meyer M, Green RE, et al. 2010. Genomic DNA sequences from mastodon and woolly mammoth reveal deep speciation of forest and savanna elephants. PLOS Biol. 8:e1000564

Roca AL. 2008. The mastodon mitochondrial genome: a mammoth accomplishment. Trends Genet. 24:49–52

Bigelow, N. H. Climate change and Arctic ecosystems: 1. Vegetation changes north of 55°N between the last glacial maximum, mid-Holocene, and present. J. Geophys. Res. 108, https://doi.org/10.1029/2002jd002558 (2003).

Poinar HN, Schwarz C, Qi J, Shapiro B, MacPhee RD, et al. 2006. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science 311:392–94

Miller W, Drautz DI, Ratan A, Pusey B, Qi J, et al. 2008. Sequencing the nuclear genome of the extinct woolly mammoth. Nature 456:387–90

Lister, A. M. & Sher, A. V. The origin and evolution of the woolly mammoth. Science 294, 1094–1097, doi: 10.1126/science.1056370 (2001).

Debruyne, R. et al. Out of America: Ancient DNA evidence for a new world origin of late Quaternary woolly mammoths. Curr. Biol. 18, 1320–1326, doi: 10.1016/j.cub.2008.07.061 (2008).

Lister, A. M. In The Proboscidea: Trends in Evolution and Paleoecology (eds Jeheskel, Shoshani & Pascal, Tassy) 203–213 (Oxford University Press, 1996).

Wei, G., Taruno, H., Jin, C. & Xie, F. The earliest specimens of the steppe mammoth, Mammuthus trogontherii, from the Early Pleistocene Nihewan Formation, North China. Earth Science 57, 289–298, doi: 10.1007/s11430-010-4001-4 (2003).

Lister, A. M., Sher, A. V., van Essen, H. & Wei, G. The pattern and process of mammoth evolution in Eurasia. Quaternary International 126–128, 49–64, doi: 10.1016/j.quaint.2004.04.014 (2005).

Enk, J. et al. Mammuthus population dynamics in Late Pleistocene North America: Divergence, phylogeography and introgression. Frontiers in Ecology and Evolution 4, 42, doi: 10.3389/evo.2016.00042 (2016).

Ho, S. Y. W., Saarma, U., Barnett, R., Haile, J. & Shapiro, B. The effect of inappropriate calibration: Three case studies in molecular ecology. PLoS ONE 3, e1615, doi: 10.1371/journal.pone.0001615 (2008).

Lee WT, Sun X, Tsai TS, Johnson JL, Gould JA, Garama DJ, Gough DJ, McKenzie M, Trounce IA, St John JC. Mitochondrial DNA haplotypes induce differential patterns of DNA methylation that result in differential chromosomal gene expression patterns. Cell Death Discov. 2017 Sep 11;3:17062. doi: 10.1038/cddiscovery.2017.62. PMID: 28900542; PMCID: PMC5592988.

Ivanov PL, Wadhams MJ, Roby RK, Holland MM, Weedn VW, Parsons TJ. Mitochondrial DNA sequence heteroplasmy in the Grand Duke of Russia Georgij Romanov establishes the authenticity of the remains of Tsar Nicholas II. Nat Genet 1996; 12: 417–420.

St John JC, Facucho-Oliveira J, Jiang Y, Kelly R, Salah R. Mitochondrial DNA transmission, replication and inheritance: a journey from the gamete through the embryo and into offspring and embryonic stem cells. Hum Reprod Update 2010; 16: 488–509.

Chang, D. et al. The evolutionary and phylogeographic history of woolly mammoths: a comprehensive mitogenomic analysis. Scientific Reports 7, 44585, doi:10.1038/srep44585 (2017).

Astakhov, V. I. & Isayeva, L. L. The Ice Hill - an Example of Retarded Deglaciation in Siberia. Quaternary Sci Rev 7, 29-40, doi:Doi 10.1016/0277-3791(88)90091-1 (1988).

Sønstebø, J. H. et al. Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate. Molecular Ecology Resources 10, 1009-1018, doi:10.1111/j.1755-0998.2010.02855.x (2010).

Binney, H. et al. Vegetation of Eurasia from the last glacial maximum to present: key biogeographic patterns. Quat. Sci. Rev. 157, 80–97 (2017).

Graham, R. W. et al. Timing and causes of mid-Holocene mammoth extinction on St. Paul Island, Alaska. Proc. Natl Acad. Sci. USA 113, 9310–9314 (2016).

Pfeiffer T, Schuster S, Bonhoeffer S. Cooperation and competition in the evolution of ATP-producing pathways. Science 2001; 292: 504–507.

St John JC, Facucho-Oliveira J, Jiang Y, Kelly R, Salah R. Mitochondrial DNA transmission, replication and inheritance: a journey from the gamete through the embryo and into offspring and embryonic stem cells. Hum Reprod Update 2010;16: 488–509.

Tsai TS, Rajasekar S, St John JC. The relationship between mitochondrial DNA haplotype and the reproductive capacity of domestic pigs (Sus scrofa domesticus). BMC Genetics 2016; 17: 67.

Haseeb A, Makki MS, Haqqi TM. Modulation of ten-eleven translocation 1 (TET1), Isocitrate Dehydrogenase (IDH) expression, alpha-Ketoglutarate (alpha-KG), and DNA hydroxymethylation levels by interleukin-1beta in primary human chondrocytes. J Biol Chem 2014; 289: 6877–6885.

Brandt AL, Ishida Y, Georgiadis NJ, Roca AL. 2012. Forest elephant mitochondrial genomes reveal that elephantid diversification in Africa tracked climate transitions. Mol. Ecol. 21:1175–89

Comstock KE, Georgiadis N, Pecon-Slattery J, Roca AL, Ostrander EA, et al. 2002. Patterns of molecular genetic variation among African elephant populations. Mol. Ecol. 11:2489–98

Eggert LS, Rasner CA, Woodruff DS. 2002. The evolution and phylogeography of the African elephant inferred from mitochondrial DNA sequence and nuclear microsatellite markers. Proc. R. Soc. Lond. B Biol. Sci. 269:1993–2006

Dickinson A, Yeung KY, Donoghue J, Baker MJ, Kelly RD, McKenzie M et al. The regulation of mitochondrial DNA copy number in glioblastoma cells. Cell Death Differ 2013; 20: 1644–1653.

Kelly RD, Mahmud A, McKenzie M, Trounce IA, St John JC. Mitochondrial DNA copy number is regulated in a tissue specific manner by DNA methylation of the nuclear-encoded DNA polymerase gamma A. Nucleic Acids Res 2012; 40:10124–10138.

Lee WTY, Cain JE, Cuddihy A, Johnson J, Dickinson A, Yeung KY et al. Mitochondrial DNA plasticity is an essential inducer of tumorigenesis. Cell Death Discovery 2016; 2: 16016.

Kelly RD, Rodda AE, Dickinson A, Mahmud A, Nefzger CM, Lee W et al. Mitochondrial DNA haplotypes define gene expression patterns in pluripotent and differentiating embryonic stem cells. Stem Cells 2013; 31: 703–716.

Wallace DC. Bioenergetics in human evolution and disease: implications for the origins of biological complexity and the missing genetic variation of common diseases. Philos Trans R Soc B Biol Sci 2013; 368: 20120267.

Tikhomirov, B. A. “An Expedition That Never Was-Benkendorf’s Expedition to the River Indigirka.” The Geographical Journal, vol. 128, no. 4, 1962, pp. 443–46. JSTOR, https://doi.org/10.2307/1792040. Accessed 16 Sep. 2022.

Vasil'Chuk, Y., Punning, J., & Vasil'Chuk, A. (1997). Radiocarbon Ages of Mammoths in Northern Eurasia: Implications for Population Development and Late Quaternary Environment. Radiocarbon, 39(1), 1-18. doi:10.1017/S0033822200040856

Guil-Guerrero JL, Tikhonov A, Rodríguez-García I, Protopopov A, Grigoriev S, Ramos-Bueno RP. The fat from frozen mammals reveals sources of essential fatty acids suitable for Palaeolithic and Neolithic humans. PLoS One. 2014 Jan 8;9(1):e84480. doi: 10.1371/journal.pone.0084480. PMID: 24416235; PMCID: PMC3885556.

van der Valk, T., Pečnerová, P., Díez-del-Molino, D. et al. Million-year-old DNA sheds light on the genomic history of mammoths. Nature 591, 265–269 (2021). https://doi.org/10.1038/s41586-021-03224-9

Ritchie, M. in Competition and Coexistence (eds Sommer, U. & Worm, B.) 109–131 Springer, 2002).

Koch, P. L. & Barnosky, A. D. Late Quaternary extinctions: state of the debate. Ann. Rev. Ecol. Evol. Syst. 37, 215–250 (2006).

Mann, D. H., Groves, P., Kunz, M. L., Reanier, R. E. & Gaglioti, B. V. Ice-age megafauna in Arctic Alaska: extinction, invasion, survival. Quat. Sci. Rev. 70, 91–108 (2013).

Nikolskiy, P. A., Sulerzhitsky, L. D. & Pitulko, V. V. Last straw versus Blitzkrieg overkill: climate-driven changes in the Arctic Siberian mammoth population and the Late Pleistocene extinction problem. Quat. Sci. Rev. 30, 2309–2328 (2011).

Mann, D. H., Groves, P., Kunz, M. L., Reanier, R. E. & Gaglioti, B. V. Ice-age megafauna in Arctic Alaska: extinction, invasion, survival. Quat. Sci. Rev. 70, 91–108 (2013).

Nystrom,V.,Dalen,L.,Vartanyan,S.,Liden,K.,Ryman,N.,andAngerbjorn, A.(2010).Temporalgeneticchangeinthelastremainingpopulation ofwoollymammoth. Proc.R.Soc.BBiol.Sci. 277,2331–2337.doi: 10.1098/rspb.2010.0301

Noguchi H, Campbell KL, Ho C, Unzai S, Park SY, Tame JR. Structures of haemoglobin from woolly mammoth in liganded and unliganded states. Acta Crystallogr D Biol Crystallogr. 2012 Nov;68(Pt 11):1441-9. doi: 10.1107/S0907444912029459. Epub 2012 Oct 18. PMID: 23090393.

Koch, P. L. & Barnosky, A. D. Late Quaternary extinctions: state of the debate. Ann. Rev. Ecol. Evol. Syst. 37, 215–250 (2006).

Palkopoulou,E.,Dalen,L.,Lister,A.M.,Vartanyan,S.,Sablin,M.,Sher,A., etal.(2013).Holarctic geneticstructureandrangedynamicsinthe woolly mammoth. Proc.Biol.Sci. 280:20131910.doi:10.1098/rspb.2013.1910

Nystrom,V.,Dalen,L.,Vartanyan,S.,Liden,K.,Ryman,N.,andAngerbjorn, A.(2010).Temporalgeneticchangeinthelastremainingpopulation ofwoollymammoth. Proc.R.Soc.BBiol.Sci. 277,2331–2337.doi: 10.1098/rspb.2010.0301

Thomas, M.G. (2012). The flickering genes of the last mammoths. Mol. Ecol. 21, 3379–3381.

Graham, R.W., Belmecheri, S., Choy, K., Culleton, B.J., Davies, L.J., Froese, D., Heintzman, P.D., Hritz, C., Kapp, J.D., Newsom, L.A., et al. (2016). Timing and causes of mid-Holocene mammoth extinction on St. Paul Island, Alaska. Proc. Natl. Acad. Sci. U.S.a. 113, 9310–9314.

Nyström, V., Dalén, L., Vartanyan, S., Lidén, K., Ryman, N., and Angerbjörn, A. (2010). Temporal genetic change in the last remaining population of woolly mammoth. Proc. Biol. Sci. 277, 2331–2337.

Guthrie, R.D. (2003). Rapid body size decline in Alaskan Pleistocene horses before extinction. Nature 426, 169–171.

Drummond, A., Rambaut, A., Shapiro, B., and Pybus, O. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. 22, 1185–1192.

Sonia Dheur, Sven J Saupe. Quand la biogéographie métagénomique reconfigure les spatialités. EspacesTemps.net, Association Espaces Temps.net, 2019, in Matthieu Noucher, Irène Hirt et Xavier Arnaud de Sartre, Métrologies critiques de l’espace, 10.26151/espacestemps.net-388a-dd19. halshs-02283892v2

Mirdita M, Steinegger M, Breitwieser F, Söding J, Levy Karin E. Fast and sensitive taxonomic assignment to metagenomic contigs. Bioinformatics. 2021 Mar 17;37(18):3029–31. doi: 10.1093/bioinformatics/btab184. Epub ahead of print. PMID: 33734313; PMCID: PMC8479651.

Garfias-Gallegos, D. et al. (2022). Metagenomics Bioinformatic Pipeline. In: Pereira-Santana, A., Gamboa-Tuz, S.D., Rodríguez-Zapata, L.C. (eds) Plant Comparative Genomics. Methods in Molecular Biology, vol 2512. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-2429-6_10

Tran Q, Phan V. Assembling Reads Improves Taxonomic Classification of Species. Genes (Basel). 2020 Aug 17;11(8):946. doi: 10.3390/genes11080946. PMID: 32824429; PMCID: PMC7465921.

Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies. Bioinformatics. 2016 Apr 1;32(7):1088-90. doi: 10.1093/bioinformatics/btv697. Epub 2015 Nov 26. PMID: 26614127.

Ayling M, Clark MD, Leggett RM. New approaches for metagenome assembly with short reads. Brief Bioinform. 2020 Mar 23;21(2):584-594. doi: 10.1093/bib/bbz020. PMID: 30815668; PMCID: PMC7299287.

Castro CJ, Ng TFF. U50: A New Metric for Measuring Assembly Output Based on Non-Overlapping, Target-Specific Contigs. J Comput Biol. 2017 Nov;24(11):1071-1080. doi: 10.1089/cmb.2017.0013. Epub 2017 Apr 18. PMID: 28418726; PMCID: PMC5783553.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015 Oct 1;31(19):3210-2. doi: 10.1093/bioinformatics/btv351. Epub 2015 Jun 9. PMID: 26059717.

Clark SC, Egan R, Frazier PI, Wang Z. ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. Bioinformatics. 2013 Feb 15;29(4):435-43. doi: 10.1093/bioinformatics/bts723. Epub 2013 Jan 9. PMID: 23303509.

Piro, V.C.; Lindner, M.S.; Renard, B.Y. DUDes: A top-down taxonomic profiler for metagenomics. Bioinformatics 2016, 32, 2272–2280.

Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 2012, 9, 357.

Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015 May 15;31(10):1674-6. doi: 10.1093/bioinformatics/btv033. Epub 2015 Jan 20. PMID: 25609793.

Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014 Mar 3;15(3):R46. doi: 10.1186/gb-2014-15-3-r46. PMID: 24580807; PMCID: PMC4053813.

Peng Y, Leung HC, Yiu SM, Chin FY. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012 Jun 1;28(11):1420-8. doi: 10.1093/bioinformatics/bts174. Epub 2012 Apr 11. PMID: 22495754.

Liang KC, Sakakibara Y. MetaVelvet-DL: a MetaVelvet deep learning extension for de novo metagenome assembly. BMC Bioinformatics. 2021 Jun 2;22(Suppl 6):427. doi: 10.1186/s12859-020-03737-6. PMID: 34078257; PMCID: PMC8171044.

Luo Y, Yu YW, Zeng J, Berger B, Peng J. Metagenomic binning through low-density hashing. Bioinformatics. 2019 Jan 15;35(2):219-226. doi: 10.1093/bioinformatics/bty611. PMID: 30010790; PMCID: PMC6330020.

Nissen JN, Johansen J, Allesøe RL, Sønderby CK, Armenteros JJA, Grønbech CH, Jensen LJ, Nielsen HB, Petersen TN, Winther O, Rasmussen S. Improved metagenome binning and assembly using deep variational autoencoders. Nat Biotechnol. 2021 May;39(5):555-560. doi: 10.1038/s41587-020-00777-4. Epub 2021 Jan 4. PMID: 33398153.

Zhang Z, Zhang L. METAMVGL: a multi-view graph-based metagenomic contig binning algorithm by integrating assembly and paired-end graphs. BMC Bioinformatics. 2021 Jul 22;22(Suppl 10):378. doi: 10.1186/s12859-021-04284-4. PMID: 34294039; PMCID: PMC8296540.

Wu YW, Singer SW. Recovering Individual Genomes from Metagenomes Using MaxBin 2.0. Curr Protoc. 2021 May;1(5):e128. doi: 10.1002/cpz1.128. PMID: 33961733.

Hou Y, Zhang X, Zhou Q, Hong W, Wang Y. Hierarchical Microbial Functions Prediction by Graph Aggregated Embedding. Front Genet. 2021 Jan 18;11:608512. doi: 10.3389/fgene.2020.608512. PMID: 33584804; PMCID: PMC7874084.

Lamurias A, Sereika M, Albertsen M, Hose K, Nielsen TD. Metagenomic binning with assembly graph embeddings. Bioinformatics. 2022 Aug 16:btac557. doi: 10.1093/bioinformatics/btac557. Epub ahead of print. PMID: 35972375.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015 Jul;25(7):1043-55. doi: 10.1101/gr.186072.114. Epub 2015 May 14. PMID: 25977477; PMCID: PMC4484387.

Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016 Oct 1;32(19):3047-8. doi: 10.1093/bioinformatics/btw354. Epub 2016 Jun 16. PMID: 27312411; PMCID: PMC5039924.

Hampton S. Bandage application. J Wound Care. 1998 Oct;7(9 Suppl):5-8. doi: 10.12968/jowc.1998.7.sup9.5. PMID: 9887730.

Abstact

The most iconic giant animal that roamed the planet during the last ice age may have been the woolly mammoths. Woolly Mammoths (*Mammuthus primigenius*) were the engineers of grasslands, keeping trees from growing onto the plains and dispersing large amounts of nutrients over immense distances via their dung. At the end of the Pleistocene, these herds vanished leading to an ecosystem conversion away from abundant grasses toward a more shrub-dominated community, which is now affected by and contributing to human-driven climate change. Studies showed that the tundra can be converted back to grasslands with the introduction of grazers again, starting with the most well-known, the Woolly mammoth. Today, reviving those grazers doesn't seem so impossible. Alongside today's bioinformatics tools, the rise of gene-editing technology, and the recovery of ancient genomic data, it is at least theoretically possible that some of those changes could be edited into the embryo of its closest living relative. Metagenomic has occasionally been applied to the eDNA of complex genomes as a way of studying the genomic structure and dynamics of a given organism. Our goal was to retrieve the mammoth's metagenome-assembled genomes by doing a co-assembly and a targeted binning on the 159 metagenomic samples where the mammoth's DNA was identified in previous studies. In addition, we aimed to get insights into the plants and animal abundance in the analyzed data as well. After the human-guided binning, the 152 MAGs were retrieved in fasta format, 38 of them were already assigned taxonomically to 6 arctic animals and 20 plants, and guided us in the binning. The other 114 were aligned against the mammoth's full DNA index to assign the mammoth's DNA abundance, and 94 bins were retrieved as the mammoth's metagenomic assembled reads.

Résumé

L'animal géant le plus emblématique qui a parcouru la planète au cours de la dernière période glaciaire a peut-être été les mammouths laineux. Les mammouths laineux (Mammuthus primigenius) étaient les ingénieurs des prairies, empêchant les arbres de pousser dans les plaines et dispersant de grandes quantités de nutriments sur d'immenses distances via leurs excréments. À la fin du Pléistocène, ces troupeaux ont disparu, entraînant une conversion de l'écosystème des graminées abondantes vers une communauté dominée par les arbustes, qui est maintenant affectée et qui contribue au changement climatique d'origine humaine. Des études ont montré que la toundra peut être reconvertie en prairies avec la réintroduction de brouteurs, et on commençant par le plus connu, le mammouth laineux. Aujourd'hui, faire revivre ces brouteurs ne semble pas si impossible. Grace aux outils bioinformatiques d'aujourd'hui, à l'essor de la technologie d'édition de gènes et à la récupération d'anciennes données génomiques, il est au moins théoriquement possible que certains de ces changements puissent être modifiés dans l'embryon de son cousin vivant le plus proche. La métagénomique a parfois été appliquée à l'eDNA de génomes complexes comme moyen d'étudier la structure génomique et la dynamique d'un organisme donné. Notre objectif était de récupérer les génomes assemblés par métagénome (MAGs) du mammouth en faisant un co-assemblage et un binning ciblé sur les 159 échantillons métagénomiques où l'ADN du mammouth a été identifié dans des études précédentes. En outre, nous avons visé également à obtenir des informations sur l'abondance des plantes et des animaux dans les données analysées. Après le binning guidé par l'homme, les 152 MAG ont été récupérés au format fasta, 38 d'entre eux étaient déjà assignés taxonomiquement à 6 animaux et 20 plantes arctiques, et nous ont guidés dans le binning. Les 114 autres ont été alignés sur l'index ADN complet du mammouth pour attribuer l'abondance de l'ADN du mammouth, et 94 bins ont été récupérés.

تلخيص

ربما كان الحيوان العملاق الأكثر شهرة الذي جاب الكوكب خلال العصر الجليدي الأخير هو الماموث الصوفي. كان الماموث الصوفي (Mammuthus primigenius) مهندسي الأراضي العشبية ، مما منع الأشجار من النمو في السهول وتشتيت كميات كبيرة من العناصر الغذائية على مسافات شاسعة عبر فضلاتها. بحلول نهاية العصر الجليدي ، اختفت هذه القطعان ، مما أدى إلى تحويل النظام البيئي من أعشاب وفيرة إلى مجتمع تهيمن عليه الشجيرات ، والذي يتأثر الآن ويساهم في تغير المناخ الذي يسببه الإنسان. أظهرت الدراسات أنه يمكن تحويل التندرا مرة أخرى إلى الأراضي العشبية مع إعادة إدخال حيوانات الرعي ، بدءًا من الماموث الصوفي الأكثر شهرة. اليوم ، لا يبدو إحياء هؤلاء الرعاة مستحيلاً. بفضل أدوات المعلوماتية الحيوية اليوم ، وظهور تقنية تحرير الجينات ، واستعادة البيانات الجينية القديمة ، فمن الممكن نظريًا على الأقل أن تتغير بعض هذه التغييرات في جنين أقرب أقربائه على قيد الحياة. تم تطبيق الميتاجينوميات أحيانًا على eDNAللجينومات المعقدة كطريقة لدراسة التركيب الجينومي وديناميكيات كائن حي معين. كان هدفنا هو استعادة جينومات الماموث المجمعة (MAGs) عن طريق القيام بالتجميع المشترك والتجميع المستهدف على 159 عينة ميتاجينومية حيث تم تحديد الحمض النووي الضخم في الدراسات السابقة. بالإضافة إلى ذلك ، نهدف أيضًا إلى الحصول على معلومات حول وفرة النباتات والحيوانات في البيانات التي تم تحليلها. بعد التجميع الموجه بشريًا ، تم استرداد 152 MAGsبتنسيق فاستا ، تم تخصيص 38 منها بالفعل تصنيفيًا لستة حيوانات في القطب الشمالي و 20 نباتًا ، وقامت بتوجيهنا في سلة المهملات. تمت محاذاة الصناديق الـ 114 المتبقية مع مؤشر DNA الماموث الكامل لتعيين وفرة الحمض النووي العملاق ، وتم استرداد 94 حاوية.