Next generation sequencing of DNA extracted from mummified tissue

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Rabab Khairat Ibrahim Abd Elhay

aus Tanta, Ägypten

Tübingen

2013

Tag der mündlichen Qualifikation:	06.08.2013
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. rer. nat. h. c. N. Blin
2. Berichterstatter:	Prof. Dr. rer. nat. J. Tomiuk

بسم الله الرحمن الرحيم

الْحَمْدُ لِلَّهِ الَّذِي هَدَانَا لِهَذَا وَمَا كُنَّا لِنَهْتَدِيَ لَوْلَا أَنْ هَدَانَا اللَّهُ

صدق الله العظيم

[الأعراف:43]

In the name of Allah, the Beneficent, the Merciful

The praise to Allah, Who hath guided us to this. We could not truly have been led aright if Allah had not guided us.

God Almighty has spoken the truth

[Surah Al-Araf: 43]

To

My beloved parents, my sweetheart sisters and their families

my beloved country
Egypt

Acknowledgment

First of all, I would like to thank God for his help and guidance to finish this work in satisfactory way, and for everything that i had. At the end, I would pray to him to guide and bless my way all the time. I am nothing without your guidance my Allah!!

I would like to express my deep gratitude to my supervisors Prof. Dr. Nikolaus Blin and PD. Dr. Carsten Pusch for their guidance, support and that they gave me such a great chance to be a part of their team and all the possibilities for evolving as a junior scientist.

A special gratitude i give to my supervisor PD. Dr Carsten Pusch, whose contribution in stimulating suggestions and encouragement, helped me to coordinate my project especially in writing this thesis. Without his encouragement and guidance this project would not have materialized.

I would like to show my greatest appreciation to my mentor Prof. Dr. Yehia Gad, Department of Medical Molecular Genetics, National Research Centre, Cairo, for his tremendous support, help and being with me even when he is far away. Thank you from all of my heart. God bless you and your beloved family.

All my appreciation to Prof. Dr. Jürgen Tomiuk for his useful comments, remarks, understanding and help whenever needed. Thank you very much!!

A special thanks goes to my team mate Markus Ball who help me a lot, gave me suggestions and surrounding me by his support, friendship, advices and patience. My thanks to my colleagues Chun-chi and Doris for kindly acceptance me in the group and their help and concern.

As well i would like to thank all the friends and the members in the Department of Medical Molecular Genetics, National Research Centre, Cairo, for their support and care.

I would like to acknowledge the Egyptian Ministry of Higher Education and Scientific Research as well as the DAAD “Deutscher Akademischer Austauch Dienst” for their financial support during my stay in Germany.

My sincere thanks to my friend Dina Fathalla for her support and being there all the time for me. My deep thanks to all my friends especially Hager, Yasmeen and Adel and special thanks to my dear uncle Prof. Dr. Abou El Asaad for his continuous encouragement.

Last but not least, i would like to express my thanks from the bottom of my heart to my parents for their endless love and support. I will be grateful forever for their love and support. I declare this work mainly and completely to my beloved parents and my dearest family. All my love and deep heart gratitude for them and for their support although the far distances between us. I should admit they are one of the main reasons of my success.

Table of contents

Summary

The application of next generation sequencing (NGS) has dramatically increased the amount of genetic information over the last few years. In the current study, DNA samples from eight Egyptian mummies were obtained and tested for the first time using the NGS technology. They were radiocarbon dated and placed within a time period between the Third Intermediate and the Graeco-Roman times (806 BC–124 AD). Initial characterization experiments using different established protocols for DNA extraction and polymerase chain reaction (PCR) and NGS showed variabilities in the retrievability and amplifiablitiy of the extracted DNA from the various Egyptian mummy samples. This can be due to the differences in the preservation status of the mummies or to the technical handling through the NGS multistep protocol. The inadequate storage environment within the mummy collection was the inducer of a bacterial bloom in the Egyptian mummy samples. This could be inferred from the reversal of the Eukaryota/Bacteria ratio in different samples taken from the same mummy after a lapse of time interval of 1.5-2.0 years.

We found that increasing the DNA concentration for NGS had a positive effect on the quality of the NGS library and the subsequent sequencing. As a result of a series of optimization experiments, the complete human mitochondrial genome of one mummy was recovered with an average coverage of 190 folds using the unique mapped reads. Using the identified mitochondrial SNPs of this mummy, haplogroup I2 was defined with a quality score of 97.7%. In addition, we address some of the SNPs that might be associated with certain human disease.

Since the results largely depend on the used bioinformatics tools, efforts were done to increase the findings' specificity and interpretation accuracy. From eight mummies, twenty-one data-sets were generated and six of them were analyzed in a metagenomic approach. It is known that the ancient Egyptians used a number of natural substances in the mummification procedure. A variety of Viridiplantae taxa were detected and some of them were documented as embalming materials in ancient texts. A number of these findings were confirmed by the conventional but sensitive and specific PCR method. According to stringent analyses, the pathogens *Plasmodium falciparum* and *Toxoplasma gondii* were detected in Egyptian mummies. Comparison was also done between the warm climate samples including Egyptian mummies and two Bolivian lowland skeletons on one hand, and the previously published cold climate NGS data-sets of the Saqqaq, the Denisova hominid and the Alpine Iceman. A unique bacterial fingerprint could be assigned

to the mummy group, comprising the Egyptian mummies and the Iceman mummy regardless of the different burial conditions. This showed that the metagenome of mummies is relatively unique, thus independent from the parameter temperature.

Zusammenfassung

Die Anwendung von “Next-generation sequencing” (NGS) hat die Genetik revolutioniert und konsequenterweise eine Masse an neuen Genomdaten geliefert. In meiner Dissertation habe ich Biopsien von von acht ägyptischen Mumien erhalten und habe sie mit Hilfe der neuen NGS Technologie untersucht. Alle Proben wurden zudem über die ^{14}C-Methode absolut datiert und stammen aus der Zeit von 806 v. Chr. bis 124 n. Chr.; dies deckt sich mit einer Epoche, die sich von der dritten Zwischenzeit bis hin zur griechisch-römischen Zeit erstreckt. Eine erste Charakterisierung aller Mumienproben über verschiedene DNA-Extraktionstechniken, konventionelle PCR Protokolle und NGS zeigte sehr schnell, dass es individuelle Unterschiede im Hinblick auf z.B. Inhibition, Extraktionsverhalten und Amplifizierbarkeit gab. Dies kann z.B. durch Unterschiede im Erhaltungszustand der Mumien oder aber auch durch das komplexe „Handling“ während des NGS Multistep-Protokolls erklärt werden. Unterschiede konnten auch an Mumien festgestellt werden, die unsachgemäß in den Sammlungen gelagert worden waren (Temperatur, Luftfeuchtigkeit, etc.). Proben von derselben Mumie, die aber in einem Abstand von anderthalb bis zwei Jahren genommen wurden, zeigten eindeutig das Ergebnis einer Bakterienblüte, weil sich das Verhältnis von Eukaryonten DNA zu Bakterien DNA umkehrte.

Ich konnte feststellen, dass das Anheben der DNA Konzentration für das NGS einen vorteilhaften Effekt sowohl auf die Qualität der zu erstellenden DNA Bibliotheken als auch auf das Sequenzieren im Speziellen hatte. Der Endpunkt einer Serie von Optimierungsstrategien war das Sequenzieren der DNA einer Mumie, bei der ich das komplette mitochondriale Erbgut mit einer 190-fachen Abdeckung erstellen konnte; hierbei nutze ich nur die über SAMtools als einzigartig deklarierten Fragmente, d.h. alle redundanten Sequenzen waren zuvor aussortiert worden. Die Auswertung der diagnostischen mitochondrialen SNPs ergab, dass es sich hierbei um die Haplogruppe I2 handelte (Qualitäts-Score von 97,7%). Zusätzlich habe ich SNPs bewertet, die nach heutiger Erkenntnis mit bestimmten Erkrankungen des Menschen assoziiert sein könnten.

Da die über NGS generierten Ergebnisse in hohem Maße von den benutzen bioinformatischen „Tools“ abhängen, habe ich großen Wert auf Parameter gelegt, die die Spezifität, Stringenz und Akkuratesse der Analysen erhöhen. Insgesamt wurden von acht Mumien einundzwanzig Datensätze generiert, wobei sechs davon genauer auf das Metagenom hin untersucht wurden. Es ist allgemein bekannt, dass die alten Ägypter eine

ganze Reihe von Naturstoffen im Rahmen der Mumifizierung benutzen. Eine ganze Reihe von Taxa aus der Gruppe Viridiplantae wurden identifiziert, und einige davon sind auch tatsächlich in den alten Schriften als Bestandteil des Balsamierungsrituals beschrieben. Die sowohl sensitive als auch spezifische Methode der konventionellen PCR konnte ich nutzen um einige der Pflanzen unabhängig zu bestätigen. Hochstringente Analysen ergaben, dass sich in den ägyptischen Mumien auch die DNA von Pathogenen wie *Plasmodium falciparum* oder *Toxoplasma gondii* finden ließ.

Schlussendlich strebte ich noch einen Vergleich zwischen Proben aus warmen Klimazonen (ägyptische Mumien, Tieflandskelette von Bolivien) und solchen aus Kaltklimaten an (Saqqaq Paläoeskimo, Denisovamensch, Tyroler Eismann). Ein hochspezifischer „bakterieller Fingerprint“ zeigt an, wenn es sich um Mumienmaterial handelt, wobei die Mumiengruppe hier aus den ägyptischen Mumien und der alpinen Gletschermumie bestand. Da sie aus unterschiedlichen Fundhorizonten und Klimaten stammen, belegt dies, dass das Metagenom von Mumien temperaturunabhängig überdauert.

1 Introduction

1.1 Ancient DNA

1.1.1 Definition and general aspects

Ancient DNA (aDNA) is DNA isolated from ancient specimens. Thus aDNA provides us an unique opportunity in the reconstruction of evolutionary tracks by combining morphological characteristics with genetic information of ancient species which can be compared with contemporary species (Willerslev and Cooper, 2005). Before the analysis of aDNA, taxonomy and evolutionary studies had been based on comparative studies of individuals sampled from recent species and populations. However, such studies permitted only to hypothesize about the evolutionary history of species whereas aDNA can be used to test these hypotheses (Avise, 2000; Willerslev and Cooper, 2005).

The isolation of DNA from ancient tissues (e.g. bone, tissue) was first reported with the implementation of PCR technology (Hagelberg et al., 1989; Horai et al., 1989; Hänni et al., 1990; Theusen and Engberg, 1990; Williams et al., 1990; Richards et al., 1995). However, the full verification of the authenticity of aDNA results has been hampered by the possibilities of contamination with contemporary DNA, DNA artifacts introduced during early cycles of PCR and the aDNA lesions secondary to its decay over time (Pusch et al. 2004; Pusch and Bachmann 2004). These technical challenges recalled the utilization of surrogate studies like that of the chemical structure of macromolecules, like proteins, in the ancient samples as an initial indicator of the preservation (Schmidt-Schultz and Schultz, 2007). Also, studies included the investigation of DNA damage in response to the effect of time, the temperature, and the surrounding burial conditions (Lindahl, 1993). Consequentially, several criteria for ascertainment of the authenticity of the aDNA results were postulated (Richards et al., 1995; Hebsgaard et al., 2005). Steadily, the aDNA studies have been influenced by the technological progress in the fields of chemistry and molecular biology. Recently, the invention of the next generation sequencing (NGS) technology initiated a new line in the aDNA research field where the whole genomes of a number of hominins and extinct mammalians could be established (Mardis, 2008; Millar et al., 2008; Green et al., 2010; Knapp and Hofreiter, 2010; Rasmussen et al., 2010; Reich et al., 2010; Keller et al., 2012; Fu et al., 2013).

1.1.2 Applications of ancient DNA studies

The aDNA studies have been involved in several successful applications including the understanding of many prehistoric genetic events pertaining to the evolution of the

human genome (Kirsanow and Burger, 2012; Schubert et al., 2012). First, aDNA studies has helped in demonstrating the genetic relationship between extinct hominins and modern humans and has improved the reconstruction of their phylogeny (Green et al., 2010; Rasmussen et al., 2010; Reich et al., 2010; Keller et al., 2012; Fu Q et al., 2013). Second, it theorized explanations for the effects of different environmental events on human migration trends and on the ecosystem in general (Willerslev et al., 2007; Stiller et al., 2010; Lorenzen et al., 2011). Third, aDNA aided in understanding and studying the interactions between different human populations and their putative ancestors (Fehren-Schmitz et al., 2010; 2011; Lawrence et al., 2010). Fourth, aDNA studies were instrumental towards understanding the prehistoric social dynamics using short tandem repeats (STR) and mitochondrial single nucleotide polymorphism (SNP) markers (Hummel and Schultes, 2000; Lalueza-Fox et al., 2011). Fifth, it gave insight into the association between various phenotypes and the selection over the populations' history using functional SNPs (Krause et al., 2007; Rasmussen et al., 2010; Gerbault et al., 2011; Keller et al., 2012). Sixth, it helped in understanding virulence mechanisms of several pathological agents which were common in specific areas or eras since ancient times (Nerlich et al., 2008; Hawass et al., 2010; Khairat et al., 2013; Lalremruata et al., 2013). Seventh, it assisted in answering some historical questions and identifying kinships of historical figures (Rogaev et al., 2009a, b; Coble et al. 2009; Hawass et al., 2010; 2012).

1.1.3 Post-mortem DNA decay and ancient DNA preservation in archeological samples

Taphonomy or diagenesis is the deterioration process of an organism after its death. Once an organism dies, a DNA degradation process is initiated by endogenous nucleases concurrently with the autolysis of the entire body. In the metabolically active cells, any error introduced into the DNA sequence rapidly initiates a DNA repair process to save the genetic information. The DNA structure has the advantage of the deoxyribose/ sugar backbone, which provides this molecule with chemical stability in comparison to ribonucleic acid (RNA) (Lindahl, 1993).

After death, DNA degradation occurs mainly through two reactions, hydrolysis and oxidation. The hydrolysis reaction mainly causes DNA backbone depurination, which with other hydrolytic processes will result in physical destruction and strand break formation within DNA molecules (Höss et al.,1996; Scholz et. 1998). The oxidation reaction targets many active spots in the chemical structure of DNA like nitrous bases and the sugar-

phosphate backbone of the DNA in the presence of oxygen (Fig. 1) (Lindahl, 1993; Richards et al., 1995; Pusch et al., 2004). Those two main processes hinder the retrieval of integral aDNA sequences and limit the correct information content obtained from their analysis. For example, hydantoins, which are the oxidative products of cytosine and thymidine, block DNA polymerases within the PCR. Additionally, the deamination products of cytosine, which are common in aDNA and modern DNA, causes incorrect bases to be inserted during the PCR (Pääbo, 1989; Höss et al.,1996; Hofreiter et al., 2001, Pusch et al., 2004; Pusch and Bachmann 2004). Moreover the fragmentation may cause nicks within one DNA strand which following the denaturation PCR step will result in a number of short fragments that are not suitable for the PCR reaction to proceed (Pusch et al., 1998).

Gradually, the DNA damage accumulates and becomes so extensive that no molecules remain to be effectively utilized for analysis. Pinpointing the aDNA survival time is still a subject of intensive debates. It is believed that it is mainly affected by the burial conditions like salt content, exposure to radiation, and availability of oxygen and free water. Moreover, the temperature and environment pH were considered in many studies as the most important affecting factors (Lindahl, 1993; Pusch et al. 2003; Zink and Nerlich, 2003; 2005; Campos et al., 2012). Theoretical studies postulated that under moderate conditions like physiological salt concentrations, neutral pH and a temperature of 15°C, it would take about 100,000 years for hydrolytic damage to destroy all DNA, and that the lower temperatures would extend this time limit (Lindahl, 1993). In some special conditions, such as rapid desiccation, low temperatures or high salt concentrations, nucleases can be destroyed or inactivated before all nucleic acids are reduced to mononucleotides (Lindahl, 1993). Thus, for example, the degradation rate of DNA due to hydration can be reduced 5-10 times by increasing the ionic strength (Lindahl, 1993). Moreover, the adsorption of DNA to hydroxyapatite can decrease the rate of DNA depurination two-fold, which will increase the chance to retrieve DNA sequences from old bones. However, it is argued that if this is the case, the DNA damage will be just slower, but not stopped completely (Lindahl T, 1993). Other studies underline that despite the temperature being a crucial factor for DNA survival, other factors like humidity may be a main factor and that the theoretical estimation can not be applicable to all aDNA samples (Poinar et al., 2003; Zink and Nerlich, 2003; 2005; Khairat et al. 2013). Recently, with technological progress and the use of whole genome sequencing facilities, aDNA studies tried to define the DNA damage degrees and types in the different archeological samples

(Briggs et al., 2007; 2010; Brotherton et al., 2007; Heyn et al., 2010; Keller et al. 2012). However, according to Rasmussen M et al. (2010), the deamination of the cytosine, as the most common DNA damage in the aDNA, showed a little impact on the final NGS results (Rasmussen et al., 2010; Keller et al., 2012).

The mechanisms of DNA degradation in archeological samples were a hot research topic at the beginning of aDNA studies, in particular the Egyptian mummies (Lindahl and Andersson, 1972; Lindahl and Nyberg, 1972; Pääbo, 1985; Lindahl, 1993; Haynes et al., 2002). One of the main concerns was the mechanism of DNA reactions with different organic and inorganic materials in archeological samples, e.g. in bones (Bell et al., 1996; Colson et al., 1997; Haynes et al., 2002; Campos et al., 2012). The degradation of DNA is a complex phenomenon where many chemical processes can act, resulting in the cross-linking or fragmentation of the DNA backbone or change of the nucleotide/base structure. The preservation degree of archeological samples has been considered the only indicator of the successful DNA retrieval (Lindahl, 1993; Pusch et al., 2003; Zink and Nerlich, 2003, 2005; Campos et al., 2012). It was noted that the DNA recovery rate was highly variable even within the samples from the same environment and even the same area, i.e. there was no simple correlation between the sample age and the DNA preservation (Pusch et al. 2003; Campos et al., 2012). Other studies suggested different ways to assess the preservation of DNA and the DNA damage extent in aDNA samples. Hydantoins are the oxidized form of pyrimidine and have been used to asses the DNA preservation using Gas chromatography/mass spectrometry (GC/MS) (Höss et al., 1996). Evidently the macromolecules, as well as DNA, will be affected with burial conditions and can be used as an indicator of the degree of DNA preservation and recovery (Poinar et al., 1996; Poinar and Stankiewicz, 1999). However, the latter methods were not that easy to apply routinely, or failed after a time to be a proper way to assess DNA preservation in all the ancient samples types (Collins et al., 2009).

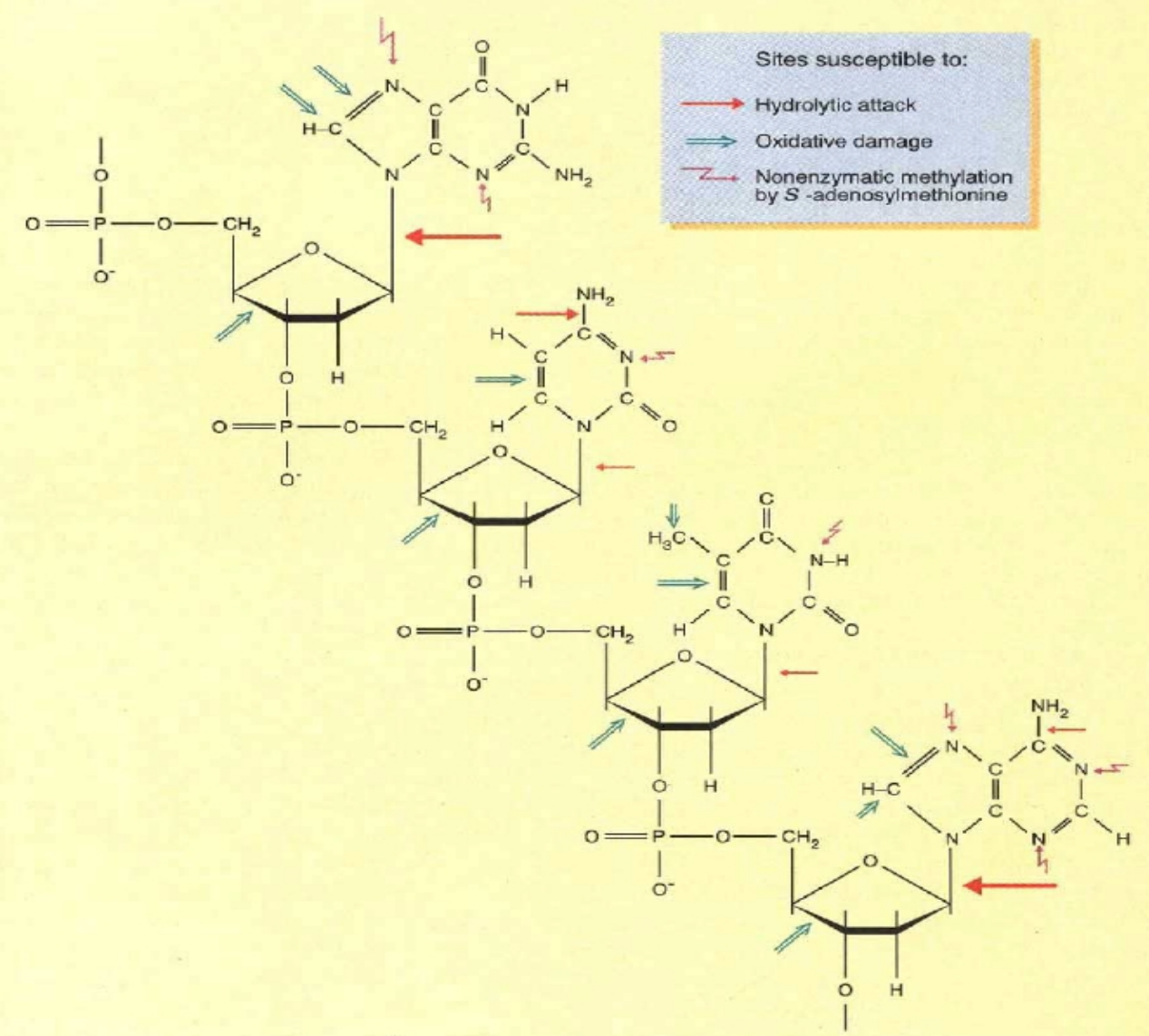


Figure 1: Target sites for DNA decay. Short segment of one strand of the DNA double-helix is shown with the four common bases (from top: guanine, cytosine, thymine, adenine). Sites susceptible to hydrolytic attack are indicated by solid red arrows, oxidative damage by open green arrows, and non-enzymatic methylation by S-adenosylmethionine as zig-zagged purple arrows. Major sites of damage are indicated by the large arrows (adapted from Lindahl, 1993).

1.1.4 Guidelines for ancient DNA studies

The retrieval of genetic information from paleontological samples has been complicated by many difficulties, which were mainly due to the DNA damage subsequently to postmortem degradation. These may entail miscoding lesions or physical destruction, as well as blocking lesions like inter-strand cross-links. Moreover, the modern DNA can be more favorable to be retrieved than the aDNA particularly with the use of

PCR technology. Therefore, guidelines and standards were established in order to guarantee authenticity of aDNA results and summed up in several publications (Richards et al., 1995; Gilbert at al., 2005b; Mitchell et al., 2005; Hansen et al., 2006; Roberts and Ingham, 2008; Heyn et al., 2010; Keller et al. 2012). Some of the guidelines have fizzled over time and the progress of the technology, while others are still valid. The most essential criteria include separation of the working areas, having negative control extractions and amplifications, reproducibility with different extractions and replication of experiments in another lab with the same results. Moreover, having inverted correlation between amplicon length and amplification efficiency, the aDNA should generate only amplified short fragments. Also, the exclusion of nuclear insertions of mitochondrial DNA by using overlapping primers for the same DNA segment can lead to more reliable results. The real challenge facing the aDNA studies is to approach the results cognitively and to consider the results on a case-by-case basis to satisfy the authenticity

1.2 Mummies

1.2.1 Mummification and the embalming materials

By definition, mummies are human and animal remains with preservation of non-bony tissues (David, 1997; Lynnerup, 2009). The word mummy is more likely derived from the Persian word, mumeia or the Arabic word “mumia”, which both means bitumen or pitch. For many people, the word mummy immediately calls the Egyptian mummies image, while in fact, and scientifically, the term refers to all naturally or artificially preserved bodies using desiccation to prevent putrefaction. Mummification was practiced in Egypt for more than 4,000 years, and perhaps developed as early as ca. 4,500 B.C. (David, 1997, 2008). The mummification process passed through many stages and started after the pre-Dynastic period by burying the dead bodies in simple graves in the desert sand. This type is called the natural mummification. Probably in the Dynastic period, the ancient Egyptians started the artificial mummification, which greatly affected the preservation status of their mummies. The Egyptians developed artificial mummification with the aim to keep the deceased body everlasting and to preserve the in-vivo physical appearance as intact as possible (David, 1997; 2008; Wisseman, 2001, Lynnerup, 2007). The intentional mummification was invented mainly due to religious beliefs, which stated that a good preserved corpse was needed to help the spirit to return back to its body in the afterlife (David, 1997; 2008; Jeziorska, 2008).

According to Book II of Herodotus's History, there was a number of mummification

protocols that were performed and common in his time (fifth century BC). The Diodorus records have also been considered as a main source to study the mummification techniques. Mummification did not remain static throughout Egyptian history and developed through time (Ikram, 2003). The mummification protocols in the 21th Dynasty [(1069-945 BC); Third Intermediate period] were considered the top of the long time mummification recipe modification (Wisseman, 2001).

The most effective mummification protocol was by removing the gastrointestinal tract containing the greatest source of bacteria and desiccation of the body. The desiccation was done by surrounding the body with naturally existing Egyptian Wadi natron salt, which is a mixture of sodium carbonate and traces of sodium bicarbonate, sodium chloride and sodium sulphate. The natron is considered one of the key materials for the mummification process, and its main function is to desiccate water from the tissues as well as to increase the pH to neutral or mildly alkaline. Consequently, this will stop or slow the enzymatic activity of both, the body's own enzymes and those from invading microorganisms. Moreover, the increase in the pH to the neutral or the mildly alkaline side could be an important factor of the DNA preservation, where the acidic environment is believed to be unfavorable for the nucleic acid survival and may cause nicks formation in DNA strands (Pusch et al., 1998; 2003; Ikram, 2003; Zink and Nerlich, 2003).

Other materials, such as myrrh, other resins and spices and oils were constituents of the embalming recipes throughout different dynasties (Ikram, 2003). Herodotus' accounts mention myrrh, cassia, palm wine, "cedar oil" and "gum" as main plant components of the embalming "recipes". Previous research suggested that, even if natron salt was widely used as a desiccant, due to the warm environmental conditions inside the tombs, the bodies would have decomposed without the application of specific organic substances (Buckley and Evershed 2001).

The embalming resin contained the sap of fir and pine trees, which the ancient Egyptians imported from Syria and Lebanon, while the frankincense and myrrh came from Southern Arabia and East Africa. Chemical analysis carried out on 13 Egyptian mummies dating from the mid-Dynastic (ca. 1,900 years BC) to the late Roman Periods (395 AD) suggested that unsaturated plant oils and animal fats were key components in mummification. Subsequently, more exotic substances were mixed up and applied either on the bodies or on the bandages (Buckley and Evershed; 2001). The peculiar properties of the unsaturated oils and fats allowed them to polymerize spontaneously. According to Buckley and Evershed (2001), the polymerization would have, in turn, produced a "*highly*

cross-linked aliphatic network, *which would have stabilized otherwise fragile tissues and/or wrappings against degradation by producing a physio-chemical barrier that impedes* the activities of *microorganisms"* (Buckley and Evershed, 2001).

Furthermore, while acting as body perfumes and neutralizing the bad smell of the putrefaction, oils, such as juniper and cedar and sometimes lettuce and castor, also had fungicial and bacteriostatic properties and decrease the body stiffness after the desiccation step and before the wrapping by linen (Ikram, 2003; Zink and Nerlich, 2003; Jeziorska, 2008). Buckley and Evershed (2001) showed that beeswax and coniferous resins were also used in the embalming procedures and that their use increased overtime. Those components were found both on the bodies and the wrappings. Components diagnostic for Pistachio resin were also found in a female Ptolemaic mummy (Buckley and Evershed, 2001). In a subsequent study, some components diagnostic for Pistachio exudate were also chemically characterized in three further but undated mummies (Nicholson et al., 2011).

1.2.2 Preservation of Egyptian mummies

The dry environmental conditions and the body desiccation are considered favorable conditions to preserve the remains, as well, the freezing conditions within cold environments. Both conditions hinder the microorganismic and cellular proteolytic enzyme effects (Lynnerup, 2007; Jeziorska, 2008). The earlier histological studies on Egyptian mummies showed the preservative effect of the mummification process, which was clear in the preservation of cellular components in comparison to the fresh sample (Lewin, 1968, Rabino-Massa and Chiarelli, 1972; Barraco, 1975; Barraco et al., 1977). Immunohistochemistry studies on the Egyptian mummies proved the presence of fine cellular components as neurochemicals (Cockburn and Cockburn, 1980; Holyle et al., 1997), as well as with the mummy models (Metcalfe and Freemont, 2012). Furthermore, biochemical studies evaluated the effects of the mummification protocol and the natron usage on the macromolecule preservation as well as that of the mummies, in general. From the comparisons of the natural and artificial mummies it was concluded that, in certain circumstances, natural mummification preserves tissues as effectively as the artificial process (Barraco et al., 1977; Lynnerup, 2007; Jeziorska, 2008). Comprehensively, each mummy status is unique and its preservation status has been presumed to depend on many factors including the mummification protocol, the time that elapsed between the death and the start of mummification process, the burial location

and condition till the excavation has been done and the storage of the samples after the excavation (Zink and Nerlich, 2003, 2005; Jeziorska, 2008; Khairat et al., 2013).

1.2.3 Molecular studies on Egyptian mummies

Studying of mummies could be at the expense of damaging them, however they are irreplaceable. Therefore, the molecular studies may be a way to retrieve more information with minimal damage (Wisseman, 2001). A lot of successful molecular and paleopathological studies have been done using Egyptian mummy tissue (Nerlich et al., 1997; Zink et al., 2000, 2001, 2003, 2006; Nerlich et al., 2008; Donoghue et al., 2010; Hawass et al.; 2010; 2012; Woide et al., 2010; Hekkala et al., 2011; Kurushima et al., 2012; Khairat et al., 2013). However, it was postulated that the mummies and other remains of a similar age can exhibit a wide range of variability with respect to DNA preservation (Poinar et al., 2003; Zink and Nerlich, 2003, 2005; Jeziorska, 2008). Doubts were cast by some on such studies on the assumption that the environmental conditions of Egypt might not be suitable for the conservation of DNA for long time-spans (Höss et al., 1996; Poinar et al, 1996; Marota et al, 2002). On the other hand, others stated that the sophisticated mummification procedure has a preservative effect, and may be equal to the effect of cold temperature (Zink and Nerlich, 2003, 2005; David, 2008; Jeziorska, 2008; Khairat et al. 2013). No one can exclude that particularly favorable burial conditions might allow human DNA to be preserved for a particularly long time-span, even in the warm climate of Egypt (Marota et al, 2002; Zink and Nerlich, 2003, 2005; Gilbert et al., 2005a; Hawass et al.; 2010, 2012; Hekkala et al., 2011; Kurushima et al., 2012; Khairat et al., 2013). Thus, the temperature may not be the sole factor to influence the preservation status and there are many factors equally affecting the accessibility to amplifiable DNA in ancient samples (Zink AR and Nerlich A, 2003, 2005). Additionally, human DNA was identified in a number of Egyptian mummies (Krings et al., 1999; Hawass et al.; 2010, 2012; Woide et al., 2010; Khairat et al., 2013), and a 2,000 years old Nubian sample collection (Fox, 1997) and Pompeii samples (Cipollaro et al., 1998, 1999; Guarino et al., 2000; Di Bernardo et al; 2002; Poinar et al., 2003). Consequently, patterns of bone diagenesis and aDNA damage may be different or present to various extents (Pusch et al., 2000; Zink and Nerlich, 2003, 2005; Cipollaro et al., 2005; Ottoni, et al.; 2009).

1.3 Next generation sequencing

1.3.1 General remarks

Demands to decrypt the DNA molecule was the main stimulus for advancing DNA sequencing technologies in the last twenty years. These efforts have been crowned in the last few years by the novel next generation sequencing (NGS) technology. The NGS was first introduced commercially in 2004, and resulted in an increase of digital genome DNA data deriving from modern and ancient specimens. The NGS technology is a massively parallel DNA sequencing systems that can simultaneously determine the sequences of a huge number of different DNAs. This allows collecting billions of reads (75 to 700 bp) in contiguity. The main advantage of NGS over the conventional Sanger sequencing technology is the production of large volumes of data in a short time with less manpower with still a relatively high to medium cost that progressively decrease with time (Liu et al., 2012). The preparation protocol of NGS libraries is mostly similar in the different NGS platforms. Thus, the genomic DNA is sheared to small segments, followed by repairing, blunting and ligating the ends with specific adaptors, which facilitate the amplification of the libraries without the need to clone them into plasmid vectors. Meanwhile, there are a number of NGS platforms available with different chemistry, enzymology, high resolution optics, hardware, and software engineering. Recently, there has been a new wave of technological progress in the field involving what is referred to as third generation sequencing (Hert et al., 2008; Mardis, 2008; Shendure and Ji, 2008; Pareek et al., 2011; Zhang et al., 2011; Liu et al., 2012).

1.3.2 NGS platforms

The NGS technology mainly refers to the massively parallel sequencing with amplification. Meanwhile there are three main NGS platforms with different NGS technologies:

1. Pyrosequencing approach has been developed and presented by 454 and the first 454 release was then purchased by Roche (Roche Applied Science, Penzberg, Germany).
2. Fluorescently labeled sequencing by synthesis, which has been developed by Solexa and their first release was the Genome Analyzer, then the company was purchased by Illumina (San Diego, California, United States), (www.illumina.com).
3. Sequencing by hybridization and ligation, which has developed and improved by Agencourt, and released the first SOLiD machine in 2006. Then it was purchased by Applied Biosystems (Foster City, California, United States) (Shendure and Ji, 2008).

Most of those platforms can sequence DNA library fragments from one end (single read) or from two ends (paired end). There is another option which is called mate-pair and used for specific purposes like de-novo sequencing, finishing genomes and structural variation detection. In some cases, NGS sequencing is required for many samples but with low coverage (e.g. PCR products, exomes, transcriptomes or sequencing after enrichment of specific genomic regions, or small genome sequencing). Performing this separately would mean an increase in costs. Therefore, multiplex adaptors have been introduced in such cases through the library preparation to facilitate sequencing of a number of samples in the same lane in a process called barcoding, indexing, or tagged NGS sequencing. The used index sequence is 6 base pair in length and it is specific enough for the readout separation afterwards using appropriate bioinformatic tools (Binladen et al., 2007; Meyer et al., 2007; Cronn et al., 2008; Hamady et al., 2008; Smith et al., 2010; Boers et al., 2012).

1.3.3 Comparison of different NGS platforms

Each NGS platform has its chemistry, enzymology and properties which have advantages as well as disadvantages in terms of read length, accuracy, consumables, and manpower. Both the advantages and disadvantages have influence on the platform choice for the different applications and the study requirements. For instance, the Roche read length is suitable to the initial genome and transcriptome characterization because the rather long read length is easy to assemble. Another example, the SOLiD accuracy increases with increasing the coverage to 30-fold, which makes it suitable for resequencing studies to determine the single nucleotide polymorphisms (SNPs) and copy number variations (CNVs) as well as for targeted resequencing and transcriptome studies (Hert et al., 2008; Mardis, 2008; Shendure and Ji, 2008; Metzker, 2010; Glenn, 2011; Liu et al., 2012) (Table 1).

In addition, each platform has its own bias in terms of library preparation, amplification, sequencing and evenness of coverage. Finally the platform characteristics including the machine capacity, run time, read length, and error profile have been rapidly changing over the last few years (Flicek and Birney, 2009; Glenn, 2011).

Table 1: Comparison between the three main NGS platforms in terms of advantages, disadvantages, biological applications, run time, instrument price and cost per million bases (Metzker, 2010; Glenn, 2011; Liu et al., 2012).

	454 Roche	Illumina	SOLiD
Advantages	Longer reads improve mapping in repetitive regions; fast run times	Currently the most widely used platform in the field	Two-base encoding provides inherent error correction
Disadvantages	High reagent cost; high error rates in homopolymer repeats	Low multiplexing capability of samples	Long run times
Biological applications	Bacterial and insect genome *de novo* assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics
Error rate (%)	1	≥ 0.1	>0.06 (SOLiD 4), >0.01 (SOLiD-5500 ,SOLiD-5500xl)
Primary errors	Indel	substitution	A-T bias
Time/run	24 Hours	3-10 Days	7 Days for SE 14 Days for PE
Read length	450-700 bp	100-150 bp (paired or single reads)	Up to 75 bp paired end up to (75x35 bp) mate paired up to (60x60bp)

1.3.4 NGS bioinformatics tools and analysis

Rapid progress in sequencing technologies and the availability of different NGS platforms have required suitable bioinformatic tools and algorithms for the straightforward analysis of these huge NGS data-sets. There are two differences between the NGS and capillary sequencing which have entailed more considerations in terms of downstream bioinformatic analysis. The first one entails, the short read length and the large data size which require an algorithm to increase the speed and the memory usage. The second difference is the platform error profile where each tends to have a characteristic error. For example, the Roche 454 platform tends to have deletions and insertion (so-called indels) in homopolymer segments while the error probabilities increase at the end of SOLiD- and Solexa-generated reads. Moreover, SOLiD data is in color space form, literally in four colors, each represents one of the four different nucleotide combinations. Thus, each platform has certain characteristics which entail a different tool to filter the raw data, and a different alignment algorithms and individual pipelines for analysis. Handling of NGS data starts with data filtering directly after sequencing in order to remove the most likely errors, i.e. specially those occurring at the strand ends in the case of Illumina or SOLiD. After

filtration the data are readily available for further analysis as outlined below (Mardis, 2008; Scheibye-Alsing et al., 2009).

The NGS platform output is usually the FASTQ file, which is defined as *a common file format for sharing sequencing read data combining both the sequence and an associated per base quality score* (Cock et al., 2009). The FASTQ file is an extension of FASTA format with the ability to store the quality score of each base, and because of its simplicity, it became a widely used format for interchanging NGS data (Fig. 2).There are different forms of the FASTQ file according to the used quality score, but the conversion between them is possible (Table 2) (Cock et al., 2009, Blankenberg et al., 2010).

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Fig. 2: Example of the NGS read sequence and its associated information in the FASTQ file. Each read consists of four lines, the first shows the sequence ID, the second the sequence bases, the third is +, and the fourth is the quality score in ASCII format (adopted and changed according to http://en.wikipedia.org/wiki/FASTQ_format).

Table 2: Description of the different forms of the FASTQ file with the quality score in ASCII characters range and offset as well as the encoded original quality score type and range (adopted from Cock et al., 2009).

Description, OBF name	ASCII characters		Quality score	
	Range	Offset	Type	Range
Sanger standard `fastq-sanger`	33–126	33	PHRED	0 to 93
Solexa/early Illumina `fastq-solexa`	59–126	64	Solexa	−5 to 62
Illumina 1.3+ `fastq-illumina`	64–126	64	PHRED	0 to 62

The quality of the NGS platform outputs is checked and controlled by different tools which can be stand-alone programs or online tools in a cloud web-site. These quality control steps include converting between different FASTQ formats and removing the primer, the index and the adapter sequences, as well as those with low quality following the quality check. Also, careful homopolymer trimming is done in the case of 454 Roche

platform as a remedy to one of the essential problematic sequences (Martínez-Alcántara et al., 2009; Blankenberg et al., 2010; Cox et al., 2010; Schmieder et al., 2010; Patel and Jain, 2012). After the quality control steps, the analysis can start directly by alignment with specific reference or in the case of de-nevo sequencing by assembling. Each of the processes is considered assembling but with different algorithms.

1.3.4.1 Alignment

The alignment can be defined as the process of finding specific genomic sequences and to know precisely which species the sequences come from. It sounds like the BLAST analysis but there are some differences. First, it finds the matching sequences within specific reference genome sequence unlike the BLAST that search within a larger database (nucleotide collection database). Second, most of the alignment allows a few mismatches like SNPs or sequencing errors while the BLAST allows higher number of mismatches. Third unlike the case in BLAST analysis, most of the alignment softwares use the reads quality value to assess and increase the alignment accuracy. Therefore, due to these special requirements, the alignment algorithms are different from that of the BLAST (Flicek and Birney, 2009; Magi et al., 2010).

Generally, the alignment programs have a common multistep work-flow, starting by identifying and searching for the most likely places in the reference sequence to map using heuristic techniques. This is followed by more accurate and slower alignment within the pre-identified places in the first step using other accurate alignment algorithms. It is to be noted that it is computationally hard to start with the slow accurate alignment. There is a growing number of alignment methods for the short reads. Two commonly used methods are hash table-based implementations and the second is the Burrows Wheeler transform (BWT). In the hash table-based method, hash is mostly created from the input file or the reference, while in the Burrows-Wheeler transform method, index is mostly created from the reference to help the rapid searching using low memory. Both of those methods suit all the next generation platforms but the programs should be designed and modified to the different sequencing techniques (color-space like SOLiD or base-space like Roche and Solexa). One of the main rules controlling the alignment programs, is the inverse relation between the accuracy and the speed of the alignment. The accurate alignment usually takes a longer time, especially in the presence of polymorphisms and sequencing errors, in comparison with the heuristic techniques (Flicek and Birney, 2009).

The resulting alignment file of all existing alignment softwares is a Sequence

Alignment Map file (SAM), which is defined as "*generic alignment format for storing read alignments against reference sequences, support short and long reads*" (Li et al., 2009). In other words, the SAM file is a text format for storing sequence alignment information in tab delimited ASCII columns. It is considered as being a generic alignment output file from all the platforms, making it easier for the downstream applications after the alignment like indexing, variant calling, alignment viewing and other applications. Moreover, it is compact and small in size in comparison with the whole alignment file and thus does not load the usage memory. It is designed to suite an alignment of 10^{11} base pairs, which is enough for deep resequencing of the human genome. There is a compressed, indexed and binary version of SAM file called Binary Alignment Map file (BAM), which is usually accompanied with an index file (*.bai). The BAM indexing helps in fast recall of alignments overlapping a specified region without going through the whole alignments. The user can manipulate the SAM and BAM files using SAMtools which can convert SAM to BAM and vice versa, as well as execute sorting, merging and indexing, and thus allowing to retrieve reads in any regions easily (Flicek and Birney, 2009; Li et al., 2009; http://samtools.sourceforge.net/samtools.shtml).

There is a number of freely available alignment softwares (Table 3), which can be used either in stand-alone form or through a computation cloud web site. The stand-alone form of the alignment softwares may generally need computational and informatics knowledge, which makes it difficult for the life scientists, who have key knowledge and experimental experience for new discoveries. Therefore the demands was urging for informatic infrastructure which can provide an easy analysis interface and with certainty of the results. This has been fulfilled with the cloud websites which can act as service websites like genomequest (http://www.genomequest.com/) or DNAnexus (http://dnanexus.com/) but the leading one is the free Galaxy server (http://galaxyproject.org) (Giardine et al., 2005; Blankenberg et al., 2010; Goecks J et al., 2010). The Galaxy server provides the scientists with tools for the NGS reads handling, filtering, statistics, alignment, mapping and blast, as well as SAMtools or other tools to handle the mapping results files. Moreover, now any lab can set up their own Galaxy server and provide it with tailored tools for specific tasks using the help and screen cast page at (http://galaxycast.org) (Taylor et al., 2007).

Table 3: A number of free alignment softwares (adapted from Magi et al., 2010).

Program	Reference	Website	Platform	Aligned Gbp per CPU day
Maq	Li et al. 2008	http://maq.sourceforge.net/	Illumina, SOLiD (partial)	~0.2
Bowtie	Langmead et al. 2009	http://bowtiebio.sourceforge.net/index.shtml	Illumina	~7
SSAHA2	Ning et al., 2001	http://www.sanger.ac.uk/resources/software/ssaha2/	Illumina, SOLiD, 454	~0.5
BWA	Li and Durbin, 2010	http://biobwa.sourceforge.net/bwa.shtml	Illumina, SOLiD, 454	~7
SOAP2	Li et al., 2009	http://www.sanger.ac.uk/resources/software/ssaha2/	Illumina	~7

1.3.4.2 *De-novo* assembly

The sequencing of organism genomes has been started using the shotgun library with cosmids or clones. In this case, when the reference sequences is not known, the assembly was the solution, which is the computational reconstruction of the original DNA sequences by searching for the overlapping segments between the read sequences. Assembly was started by reconstructing long contigs to use it for genomes building. It was feasible with the old sequencing technologies using 800 bp long sequences. The short reads assembly algorithms have the same strategy of the long reads ones by searching for overlapping sequences between the reads. A modification, achieved with de Bruijn formulation, made the analysis suitable for the short length reads and the repeats. There is a number of available short reads assemblers like EDNA (Hernandez, 2008), VELVET (Zerbino and Birney, 2008), and EULER-SR (Chaisson and Pevzner, 2008; Flicek and Birney, 2009; Magi et al., 2010).

1.3.4.3 SNP detection

Generally, one of the main applications of DNA sequencing is the SNP determination which can be useful for genetic disease research, expression level estimation through RNA sequencing and population genetics studies. The low coverage studies are more often than the high coverage ones (with > 20x average coverage) due to the cost and the need of many samples in some studies. The SNP calling of the NGS data is particularly complicated with a coverage of <5x per site, which increases the uncertainty of the SNP and genotype calling due to the high error rates resulting from

alignment or base calling errors (Magi et al., 2010; Nielsen et al., 2011). The base calling algorithm determines first, the sequenced base from the fluorescence intensity data and assigns a measure of uncertainty or quality score to each base call. The base calling depends upon the used platforms, where every platform has its own error pattern. Certainly, increasing the accuracy of the base calling and decreasing the error rate will positively affect the SNP calling and all the downstream analysis processes (Magi et al., 2010; Nielsen et al., 2011).

Additionally, the alignment accuracy has a great effect on the SNP and genotype calling, where the incorrectly aligned reads will certainly increase the ratio of mis-called SNPs. The alignment accuracy depends upon the alignment algorithms, which should be qualified to differentiate between the real polymorphism and the machine errors. Furthermore, the number of the sequence identities between the reads sequences and the reference one will be different between organisms. For example the number of allowed mismatches in mapping with the human reference is different from that in mapping with the more variable organisms like *Drosophila melanogaster* (Magi et al., 2010; Nielsen et al., 2011). Following the mapping, evaluation and filtration of the mapped reads are a crucial step. The accuracy of SNP calling can be accomplished by using a cut off of a high sequencing depth of >20-folds. For moderate or low sequencing depth a probabilistic framework is suitable for the SNP calling, where the cutoff rule would not be suitable, as it would cause under estimating of the heterozygous SNP. The probabilistic framework relies on statistical calculations of the uncertainty of genotype likelihoods using the quality score and sometimes with incorporation of additional information like the allele frequencies and patterns of linkage disequilibrium (LD) (Magi et al., 2010; Nielsen et al., 2011).

1.3.4.4 Assembly/Alignment viewer

The huge data-sets produced by the NGS platforms required bioinformatics tools to visualize and browse the resulting alignment in a user-friendly interface. Therefore a number of visualization programs have been invented like the Text Alignment Viewer of SAMtools (Li et al., 2009), Maq View (Li et al., 2008), Tablet (Milne et al., 2010) and Integrative Genomics Viewer (IGV) (Robinson et al., 2011; Thorvaldsdottir et al., 2012). The NGS reads visualization software should quickly and efficiently process the huge NGS reads, providing the user with enough information and be feasible with the different assembly formats (Magi et al., 2010).

1.3.5 Ancient DNA studies and NGS

The invention of NGS technology had an impact on the aDNA studies and started a new era in paleogenetics. Being frequently shorter than 200 bp, the fragmented DNA exploits the advantages of the NGS technology. The utilization of NGS technology in the aDNA studies synchronized with the first release and availability of NGS platforms. Because of the complicated features of aDNA, earlier and heretofore many studies were concerned with the optimization of the aDNA library preparation and the analysis of the resulting data sets (Noonan et al., 2005; Green et al., 2006; Poinar et al., 2006; Maricic and Pääbo, 2009; Prüfer et al., 2010). Meanwhile, a number of whole genome sequences of extinct and aDNA samples were published (Miller et al., 2008; Green et al., 2010; Rasmussen et al., 2010; Reich et al., 2010; Keller et al., 2012).

The first NGS with an aDNA study was done using woolly mammoth (*Mammuthus primigenius*) sample from Siberia (~ 5000 years old), which yielded 28 million base pair used for genetic and metagenomics investigation (Poinar et al., 2006). This was followed by retrieving more of the mammoth nuclear genome from two different mammoth hair samples and using it for phylogenetic studying of the mammoth in comparison with the African elephant. The retrieved genetic material covered around 0.7 of the mammoth genome. The phylogenetic results suggested estimated times of the divergence between the mammoth and the African elephant and between the two mammoth used in the study (Miller et al., 2008). The NGS technology was used in a further comparative study on another mammoth sample to reveal more about its phylogenetic correlation with other elephants species (Rohland et al., 2010).

Most of the NGS studies on the ancient samples passed through specific steps which started by characterization studies of the ancient sample and were accompanied with optimization of the technical protocols to increase the yield of endogenous DNA (Binladen et al., 2007; Meyer et al., 2008; Burbano et al., 2010). Generally, those characterization studies are portraying a general picture about the sample contents and the endogenous DNA extent in comparison with the microbial DNA genetic material to set the proper way for further studies. For example, the NGS studies on Neanderthal, (40,000 years old) (Disotell, 2012), started by analysis of million base pairs of nuclear DNA sequences using the 454 sequencing platform (Green et al., 2006) and a characterization study using a metagenomic approach (Noonan et al., 2006), then followed by the Neanderthal mitochondrial genome (Green et al., 2008). Due to excessive presence of microbial DNA (95-99.8%) (Burbano et al., 2010, Green et al., 2010), targeted

investigation has been done using array-based enrichment to catch the Neanderthal genome sequences and enrich it (Burbano et al., 2010). Finally, this approach yielded the draft sequence of Neanderthal genome composed of 4 billion nucleotides with average coverage of 1.3-fold. By comparison with 5 human genomes from different world regions, the Neanderthal showed shared genomic variants with those Eurasian more than that with the sub-Sharan African which may suggest that the gene flow from Neanderthal to the non-Africans happened before the divergence of the Eurasian (Green et al., 2010).

The same conclusion has been deduced from another comparison study, which was done using a finger bone that was found in Denisova Cave in southern Siberia (dated to 50,000-30,000 BP). The results were supported by comparison with DNA evidence from a molar that was found in the same cave (Reich et al., 2010). In this study, the genomic sequence has been recovered from the finger bone with coverage of 1.9 folds of -what is believed to be- an archaic hominin who was living in East Asia while the Neanderthal lived in Europe and Western Asia. From the analysis and comparison with Neanderthal and modern human genomic data, it was concluded that the Denisova hominin shared a common ancestor with Neanderthal but had a distinct population history and shared about 4-6% with the genome of the Melanesian modern humans and not with all the Eurasians like Neanderthal (Krause et al., 2010; Reich et al., 2010; Meyer et al., 2012).

A genome sequence of an ancient human extinct paleo-Eskimo has been recovered using the NGS technology from four human tufts which were found in permafrozen sediments in northern Alaska. They belong to the Saqqaq culture and have been radiocarbon dated to around 4000 years before present (BP). The analysis of 20-fold coverage of extinct Saqqaq human genome suggested evidence for a migration from Siberia to the New World since ~5,500 years ago but this was an independent event from the migration of the Native Americans and Inuit (Rasmussen et al., 2010).

Keller et al (2012) published the complete nuclear and mitochondrial genome of the Tyolean Iceman, which has been dated to about 5,300 years BP. The complete genome showed a concordance with the previously published mitochondrial genome (Ermini et al., 2008) and by analysis it showed a genetic relation to the modern population of the Tyrrhenian Sea. Moreover, about 60% *of the Borrelia burgdorfferi* has been recovered from the Alpine Iceman NGS data-set. This is strong evidence that he was infected with the pathogen causing Lyme disease. In other words, he is the first case reported in human evolution to carry this disease. By genomic analysis, the genetic

material of the Iceman showed that he had brown eye, blood group zero, and was lactose-intolerant. In addition, available disease SNP data showed a high risk for coronary heart disease which was consistent with the radiological evidence of vascular calcification (Keller et al., 2012).

Recently, the oldest full genome have been published from permafrost preserved horse bone (560-780 Kyr BP) as well as others five genomes of five domestic horse breeds (*Equus ferus caballus*), a Przewalski's horse (*E. f. przewalskii)* anda donkey (*E. asinus*) (Orlando et al., 2013). Based on the study analysis, the divergence from the Equus genus to all modern horses, donkey and zebras was 4.0-4.5 million years ago before present. Additionally, the horse population size affected by the sever change in the climate and varied over the last 2 million years. The divergence between the Przewalski's and the domestic horses was estimated to be 38-72 kyr ago and the genetic variation among them was the same. This highlights the genetic viability of the Przewalski's horse (Orlando et al., 2013).

The progress in the field of aDNA studies has been synchronized with the invention of ad hoc bioinformatic tools as well as bioinformatic and statistical analysis studies either for the aDNA NGS data-sets or the post mortem DNA damage patterns (Vives et al., 2008; Briggs et al., 2009; Krause et al., 2010; Ginolhac et al., 2011). In extension, some studies used the huge genomic information to investigate their functional prospective as done with Ice-man data-sets (Keller et al., 2012).

The NGS studies of the aDNA samples have a high impact on the human evolution and migration studies (Reich et al., 2010). Recently, further studies presented the human genome of early modern humans accompanied with phylogenetic correlation with the modern human populations and the archaic groups like Neanderthal, for example the 30,000 years old from Kostenki Russia (Krause et al., 2010) and the 40,000 years old Tianyuan Cave, China (Fu et al., 2013).

1.3.6 Metagenomics and NGS ancient studies

The Metagenomics is defined *as "a discipline that enables the genomic study of uncultured microorganisms"* (Wooley et al., 2010). It was introduced for the first time in 1998. Unlike the other microbial genomic sequencing projects, metagenomic studies concentrate on a number of genes within environmental samples and study their biochemical functions and determine their interactions. The NGS causes a huge progress in the metagenomic studies in the last few years and facilitated the genomic studies with

low experimental price and lab work (Wooley and Ye, 2009, Wooley et al., 2010).

The metagenomic studies needed a different set of computational and statistical tools to deal with various challenging applications. Generally, these applications include, first, the assembly of the new genomes and the prediction of their genes, second, characterizing microbial diversity in the environmental samples and communities, third, the functional predication of the functionally unknown protein families within the metagenomic data-sets of various microbial communities, and finally the comparative metagenomics, which is simply the comparison between genomes in different microbial communities. It concerns the fingerprint determination of different environmental communities to gain more information about the characteristics of the sampled environments. Comparative metagenomic analysis became one of the main approaches in recent aDNA studies, where each ancient sample can be considered a distinct microbial community. This information had a great effect on the analysis of the aDNA NGS data-sets and enlightened the researchers about the preservation and the suitability of aDNA samples (Wooley and Ye, 2009). Analysis of comparative metagenomics results needs a computational tool like MEGAN, the first software to be used for the aDNA NGS metagenome data-sets. The preprocessing step of the data-sets is to align them to nucleotide database using BLAST or any comparison software. Then, the MEGAN computes and represents taxonomical content of the data-set using the NCBI taxonomy according to the BLAST top hits. The taxonomical assigning used the lowest common ancestor algorithm (LCA) which is dependent on the sequence conservation level (Fig. 3). MEGAN presents the data in a statistical and visually user-friendly way suitable for laptops. At first, MEGAN was used to study the metagenomics in single sample and afterwards, MEGAN 2.0 provided the facilities for functional comparative analysis. MEGAN 3.0 was developed for the statistical analysis of the pairwise comparison. Moreover, MEGAN provides the ability to compare between the different metagenomes in different samples and gives a statistical presentation of the assigned reads (Huson et al., 2007, 2009, 2011; Miller et al., 2008; Keller et al., 2012; Khairat et al., 2013).

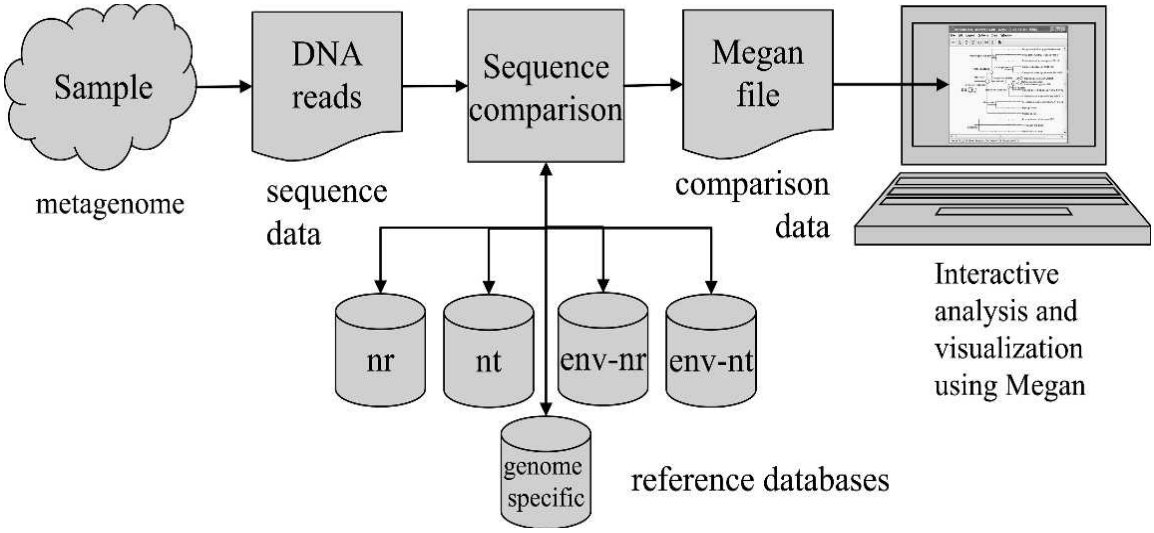


Figure 3: The work flow of the MEGAN starting by the comparison of the sequenced reads with the references database using any comparison software and followed by MEGAN processing of the BLAST top hits using LCA and NCBI taxonomy to assign the sequences (adapted from Huson et al., 2007).

1.4 The aim of the study

Within the last few years, the molecular approach using the PCR technology answered a number of historical questions regarding the Egyptian history. Moreover, authentic aDNA has been recovered from human and animal mummies (Hawass et al., 2010, 2012; Hekkala et al., 2011; Kurushima et al., 2012). Thus portending an end to a long term debate about the ability to recover authentic DNA from the Egyptian mummies. The aim of this study was the determination of the metagenomic pattern harboured in Egyptian mummies.

To address this question, we employed the novel technology "Next Generation Sequencing" (NGS), which was never used on DNA of Egyptian mummies before. Since embalmed bodies represent a typical metagenomic sample, we wanted to characterize the DNA of different sources: DNA of the individual itself, DNA of the materials used for the mummification process, DNA from the environmental sources and DNA that might point to pathogens or parasites that are hosted within the body. This approach is highly technical and requires a lot of bioinformatics work, first, we started the research with many optimization strategies to improve the DNA extraction, the library construction, and the NGS analysis parameters. The endpoint of this study will be the comparison of NGS data-sets from warm and cold climate aDNA samples. This will hopefully show the differences in metagenomic content, preservation parameters and microbial composition.

2 Material and Methods

2.1 Material

2.1.1 Devices and instruments

machine	company
Nanodrop	Thermo Fisher SCIENTIFIC, Waltham, Massachusetts, USA
MagNA Pure DNA Compact System	Roche Applied Science, Penzberg, Germany
GeneAmp® PCR System 9700 Thermocycler	Applied Biosystems, Foster City, California, USA
3130 Genetic Analyzer	Applied Biosystems, Foster City, California, USA
Genome Analyzer IIx (GAIIx)	Illumina San Diego, California, USA
Agilent 2100 bioanalyzer	Agilent technologies, California, USA
SOLiD 3 Plus System	Applied Biosystems, Foster City, California, USA

2.1.2 Chemicals

chemical	company
chloroform	Carl ROTH, Karlsruhe, Germany
Isopropanol	Merck KgaA, Darmstadt, Germany
ethanol	Merck KgaA, Darmstadt, Germany
nuclease free water	Carl ROTH, Karlsruhe, Germany
Proteinase K	Qiagen , Hilden, Germany
phenol	Carl ROTH, Karlsruhe, Germany
D-Sucrose	Fluka, Sigma-Aldrich, Buchs, Switzerland
Triton X-100	Sigma-Aldrich, St. Louis, Missouri, USA
EDTA	Merck KgaA (Darmstadt, Germany)
Tris-HCL	Sigma-Aldrich, St. Louis, Missouri, USA
Tryptone	Becton, Dickison and Company, New Jersey, USA
Yeast Extract	Becton, Dickison and Company, New Jersey, USA
NaCl	Merck KgaA, Darmstadt, Germany
Agar	Becton, Dickison and Company (New Jersey,U.S.)
Sephadex-G50 Fine	Sigma-Aldrich, St. Louis, Missouri, USA

2.1.3 Enzymes and kits

Enzyme and kit	company
FastStart PCR Master Mix	Roche Applied Science, Penzberg, Germany
CloneJET™ PCR Cloning Kit	Fermentas, Thermo Fisher Scientific, Waltham, Massachusetts, USA
Shot® TOP10 Chemically Competent E. coli	Invitrogen™ Life technologies, Carlsbad ,California, USA
HotStarTaq Plus DNA Polymerase	Qiagen , Hilden, Germany
ExoSAP-IT	USB Corporation, Cleveland, OH, USA
BigDye Terminator v3.1 chemistry	Applied Biosystems, Foster City, California, USA
New England Biolabs NEBNext™ DNA sample prep	New England Biolabs GmbH, Ipswich, Massachusetts, USA
MinElute kit	Qiagen , Hilden, Germany
Agencourt Ampure XP	Beckman Coulter Genomics, Brea, California, USA
Phusion HF Master Mix	Finnzyme, Thermo Fisher Scientific, Waltham, Massachusetts, USA
GeneRead qPCR SYBR Green Mastermix kit	Qiagen, Hilden, Germany
TrueSeq PE Cluster Kit v2	Illumina, Inc, San Diego, California,USA
TruSeq SBS Kit v5 GA	Illumina, Inc, San Diego, California,USA
SOLiD® Fragment Library Construction Kit	Invitrogen™ Life technologies, Carlsbad ,California, USA
SOLiD™ Library Column Purification Kit	Applied Biosystems, Foster City, California, USA
Platinum HiFi PCR Amplification Mix	Invitrogen™ Life technologies, Carlsbad ,California, USA
PureLink™ PCR Purification Kit	Invitrogen™ Life technologies, Carlsbad ,California, USA

2.1.4 Consumables

consumable	company
Nalgene® syringe filter (0.25 µm)	Thermo Fisher Scientific, Waltham, Massachusetts, USA

2.1.5 Buffers and media

buffer	chemicals	concentration
AEB buffer	D-Sucrose	8.00%
	Triton X-100	5.00%
	EDTA	50 mM
	Tris-HCL	50 mM

Medium	chemicals	concentration
LB- agar plates	Tryptone	1.00%
	Yeast Extract	0,5 %
	NaCl	1.00%
	Agar	1,5 %
	Ampicillin	50 mg/l

2.1.6 The primers sequences

2.1.6.1 The cloning primers

Primers name	Primer orientation	Primer sequence 5'-3'
PJET1.2 F	Forward	5' CGACTCACTATAGGGAGAGCGGC 3'
PJET1.2 R	Reverse	5' AAGAACATCGATTTTCCATGGCAG 3'

2.1.6.2 Illumina Library preparation

Name	Primer sequence 5'-3'
Adaptor 1	5' P-GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG 3'
Adaptor 2	5' ACACTCTTTCCCTACACGACGCTCTTCCGATCT 3'.
P1	5'AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGA TCT 3'
P2	P2: 5' GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT 3'

2.1.6.3 SOLiD library preparation

Name	Primer sequence 5'-3'
P1	5-CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGAT-3
P2	5-AGAGAATGAGGAACCCGGGGCAGTT-3
Primer 1	5-CCACTACGCCTCCGCTTTCCTCTCTATG-3
Primer 2	5-CTGCCCCGGGTTCCTCATTCT-3

2.1.7 Study samples

Table 4: The study specimens. The names, habitus, tissue types, excavation locations, carbon dating estimates, the corresponding dynasties and environmental average temperatures are given. AT= average temperture, SA= South America. H= hard tissue, S=soft tissue.

Specimen	Habitus	Tissue type	Excavation location	Dating	Dynasty/ period	AT 15°C
DMG1	mummy	Mummified tissue (H, S)	Tomb, Egypt	806-784 BC	3rd Intermediate Period	>
DMG2	mummy	Mummified tissue (S)	Tomb, Egypt	382-234 BC	Late Period/ Hellenistic-Ptolemaic Period	>
DMG3	mummy	Mummified tissue (S)	Tomb, Egypt	54-124 AD	Roman Period	>
DMG4	mummy	Mummified tissue (S)	Tomb, Egypt	358-204 BC	Late Period/ Hellenistic-Ptolemaic Period	>
DMG5	mummy	Mummified tissue (H, S)	Tomb, Egypt	402-385 BC	Late Period	>
DMG6	mummy	Mummified tissue (H, S)	Tomb, Egypt	801-777 BC	3rd Intermediate Period	>
DMG7	mummy	Mummified tissue (S)	Tomb, Egypt	ND	3rd Intermediate to Roman Period	>
DMG8	mummy	Mummified tissue	Tomb, Egypt	ND	3rd Intermediate to Roman Period	>
DMGS-1000	skeleton	H	Mineral soil burial, SA	400-1300 AD		>
DMGS-2000	skeleton	H	Mineral soil burial, S A	ca. 717 AD	-	>

This study included soft and hard tissue biopsies taken from eight human heads belonging to the collection of the Institute of Pre- and Protohistory, Department of Early Prehistory and Quaternary Ecology, Division of Paleoanthropology (Tübingen, Germany). The study mummies were recovered from the necropolis of Abusir el Meleq in the Fayum Valley (Lower Egypt) at the end of the 19th century Radiocarbon dating performed with INTCAL04 and CALIB5 protocol (Reimer et al., 2004) placed the mummies between (806 BC- 124 AD) which corresponds to a time span ranging from the Third Intermediate Period to the Roman Period (Table 4) (Reimer et al., 2004).

For comparison purpose, we included two bone samples taken from two mineral soil buried human skeletons (South America) as an example of aDNA samples from another warm climate environment.

2.2 Ancient DNA lab guidlines

DNA extraction work was conducted in a dedicated facility, physically isolated from the PCR technology facilities, the library preparation steps and the areas for the post-PCR work. Work surfaces were frequently cleaned with DNAase and irradiated with UV light. Disposable plastic items were used whenever possible. Non-disposal items were baked at 200°C, washed with DNase and irradiated with UV light. The solutions and buffers were prepared on a clean bench, autoclaved and sterilized by filtration using Nalgene syringe filter (0.25 µm). All work was conducted while wearing appropriate protective garments. Contamination was monitored by the use of negative and blank extraction controls, which were used along with each sample. Mitochondrial DNA typing of all lab working members was performed and used for comparsion with the results (Richards et al, 1995; Roberts et al., 2008; Hawass et al.; 2010; Keller et al., 2012; Khairat et al., 2013).

2.3 Study workflow and methods

The study workflow is summarized in figure 4. A detailed description of the adopted protocols and procedures follows.

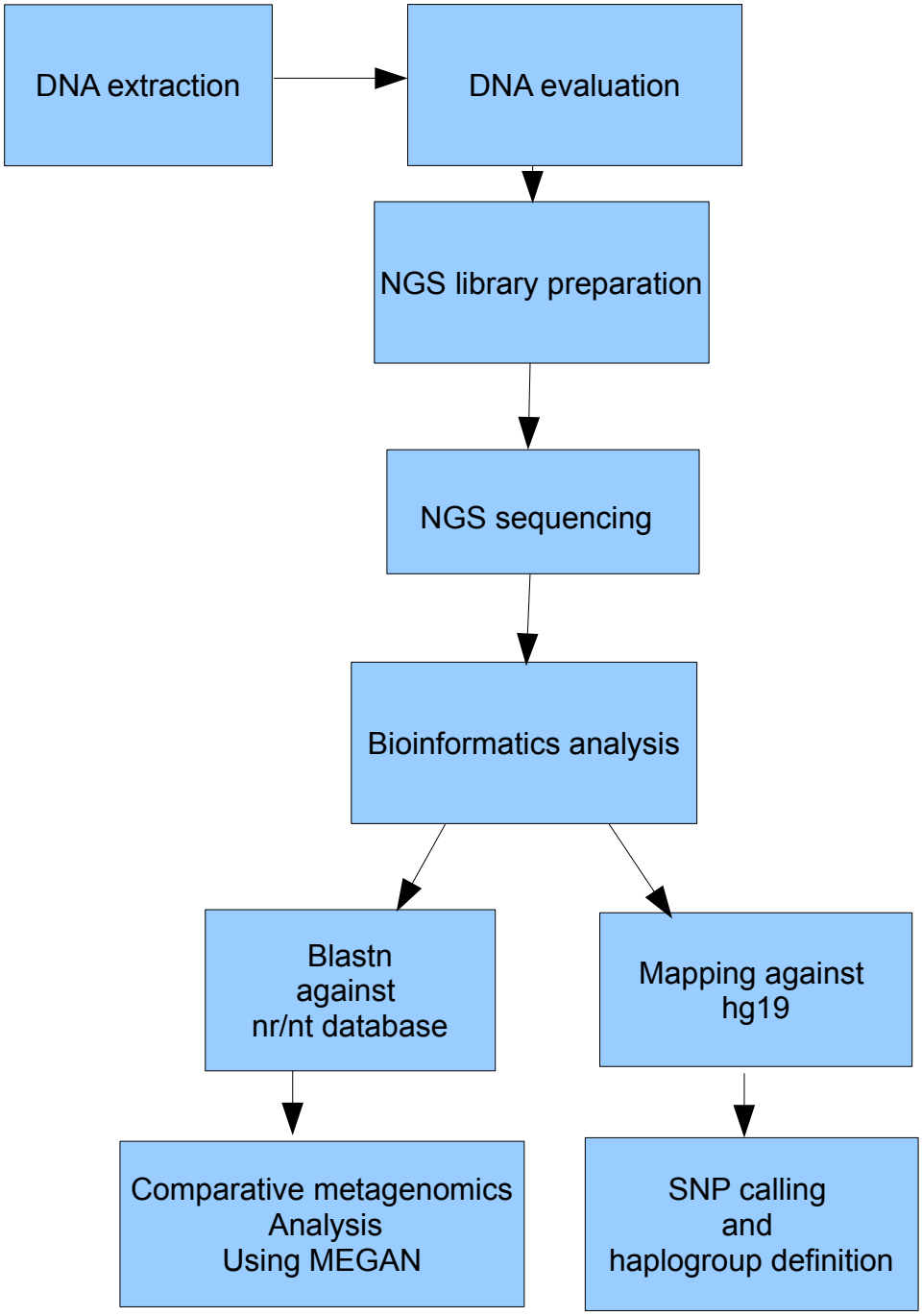


Figure 4: The study workflow and outlines.

2.3.1 DNA extraction

First, different DNA extraction protocols were applied, before the generated samples were tested and used for NGS library preparation. Initially, sampling of tissues and DNA extraction and purification were performed according to a protocol given by Scholz and Pusch (1997). Approximately 40 mg of tissue powder were mixed with 800 µl

of AEB buffer and 800 µl phenol in a 2 ml Eppendorf tube.

This mix was incubated at room temperture for 3 days on a shaker at 200 rpm. After incubation, the sample was centrifuged for 15 minutes at 13,000 rpm and the aqueous upper phase was carefully transferred to a new Eppendorf tube. To get rid of the phenol traces in the aqueous phase, 800 µl chloroform was added and mixed well for 2 minutes using a Vortex. After another ten-minute centrifugation at 13,000 rpm, the aqueous supernatant was transferred to a new Eppendorf tube and an equal volume of cold 100% Isopropanol was added. Then the tube was incubated for at least 2 hours at -20° C. Subsequently the mix was centrifuged at 4 °C for 15 min at 13,000 rpm, wherein a pellet formed at the bottom of the Eppendorf tube. The supernatant was decanted and the pellet was carefully washed with 400 µl of 70% ethanol to remove any salt residues. The supernatent was discarded and the DNA pellet was left to dry at room temperature. Afterwards, the pellet was dissolved in 50 ul nuclease free water and stored at -20° C for further downstream experiments.

Slightly modified version of the aformentioned extraction protocol was also tested and applied for increments in the DNA yield. Following the previously mentioned phenol-chloroform extraction protocol, the aqueous phase was applied to the MagNA Pure DNA Compact System for automated purification. The extracted DNA was viewed using gel electrophoresis, whereas 5 µl were loaded on 2% agarose gel to examine the DNA fragment size. Using the gel documentation system, photo was taken of each gel before and after the staining with ethidium bromide under the UV light to examine the extent of the inhibitors' presence. The extracted DNA was quantified using 1 µl DNA aliquot using nanodrop.

2.3.2 Polymerase chain reaction (PCR)

To evaluate the extracted DNA, a spiking reaction (Pusch and Bachmann, 2004) was done using an aliquot one µl of the extracted DNA and added in a total volume of 25 µl PCR. The PCR mix included 12.5 µl of 1x HotStarTaq Plus DNA Polymerase, 20 pmol of each pJET1.2 primers (forward and reverse), one µl of pJET1.2 vector, 10 mM each dNTP. Using a GeneAmp® PCR System 9700 Thermocycler, the cycling conditions of plasmid amplification were as follow:

step	temperature	time
1	94°C	5 min
2	94°C	30 sec
3	57°C	30 sec
4	72°C	30 sec
5	Repetition of steps 2 to 4 for 45 times	
6	72°C	10 min
7	10°C	hold

2.3.3 Cloning

Cloning of the PCR products was performed with the CloneJET™ PCR Cloning Kit. According to the manufacturer's protocol, the cloning of a sticky end PCR product started by setting a blunting reaction on ice including 10 µl of 2X Reaction Buffer, 1 µl of non-purified or 0.15 pmol purified PCR product, 1 µl of DNA Blunting Enzyme and the mixture completed to a total volume of 18 µl with nuclease-free water. The mixture was vortexed briefly, centrifuged for 3-5 s and incubated at 70°C for 5 min then finally chilled on ice.

The ligation reaction was set up on ice by adding 1 µl of pJET1.2/blunt Cloning Vector (50 ng/µl) and 1 µl of T4 DNA Ligase to the blunting reaction mixture. The ligation reaction components were mixed briefly, centrifuged for 3-5 s to collect drops and incubated at room temperature (22°C) for 30 min. Following the ligation reaction, the transformation reaction was done as recommended by the maufacturer's protocol using One Shot® TOP10 Chemically Competent *E. coli*. Two µl of the ligation reaction were added into a vial of One Shot® TOP10 Chemically Competent *E. coli*, which was gently mixed and incubated on ice for 30 minutes. After the cold incubation the cells were subjected to heat-shock for 45 seconds at 42°C without shaking and immediately transferred back to ice. Following the addition of 250 µl of S.O.C. Medium, the *E.coli* vials were incubated at 37°C for 1 hour with horizontal shaking at 200 rpm. The LB-agar plates were incubated at 37°C, and following the transformation, different aliquots were spread over them and reincubated at 37°C overnight. On the next day, the LB-agar plates were stored in -4°C refrigerator.

Twenty colonies were picked out of each LB-agar plates and used for the 25 µl PCR reaction containing 1x HotStarTaq Plus DNA Polymerase, 20 pmol of each pJET1.2 primers, 10 mM each dNTP. Using a GeneAmp® PCR System 9700 Thermocycler, the cycling conditions of clone amplification were as follow:

step	temperature	time
1	94°C	5 min
2	94°C	30 sec
3	57°C	30 sec
4	72°C	30 sec
5	Repetition of steps 2 to 4 for 45 times	
6	72°C	10 min
7	10°C	hold

2.3.4 Sanger sequencing

All the successful PCR products and the colony PCR products were cleaned by ExoSAP-IT by adding 2 µl ExoSAP-IT to 5 µl PCR products and incubating them at 37°C for 15 min followed by inactivation step at 80°C for 15 min. Following the ExoSAP-IT purification, the PCR products were used for Sanger cycle-sequencing reactions with BigDye Terminator v3.1 chemistry. For the cycle sequencing reaction, 1 µl of the purified PCR product was used with 2 µl of 5X BigDye Terminator v3.1 buffer, 3.2 pmol of each used amplification primer. The volume was filled to 10 µl total volume with free nuclease water. The cycling conditions using a GeneAmp® PCR System 9700 Thermocycler were as follow:

step	temperature	time
1	94°C	3 min
2	94°C	10 sec
3	50°C	5 sec
4	60°C	4 min
5	Repetition of steps 2 to 4 for 25 times	
6	4°C	hold

Sephadex-G50 Fine was used for preparation of purification columns by dissolving 1 g Sephadex-G50 in 15 ml bi-distilled water and mixed using a shaker for 1 hour. Then, a clean column was filled by 300 µl Sephadex-G50 suspension and centrifuged for 10 min at 3000 rpm. The total cycle sequencing reaction volume was applied to the prepared Sephedex-G50 columns and centrifuged for 5 min at 3000 rpm. The purified cycle sequencing products were run on a 3130 Genetic Analyzer.

2.3.5 NGS technologies and library preparation

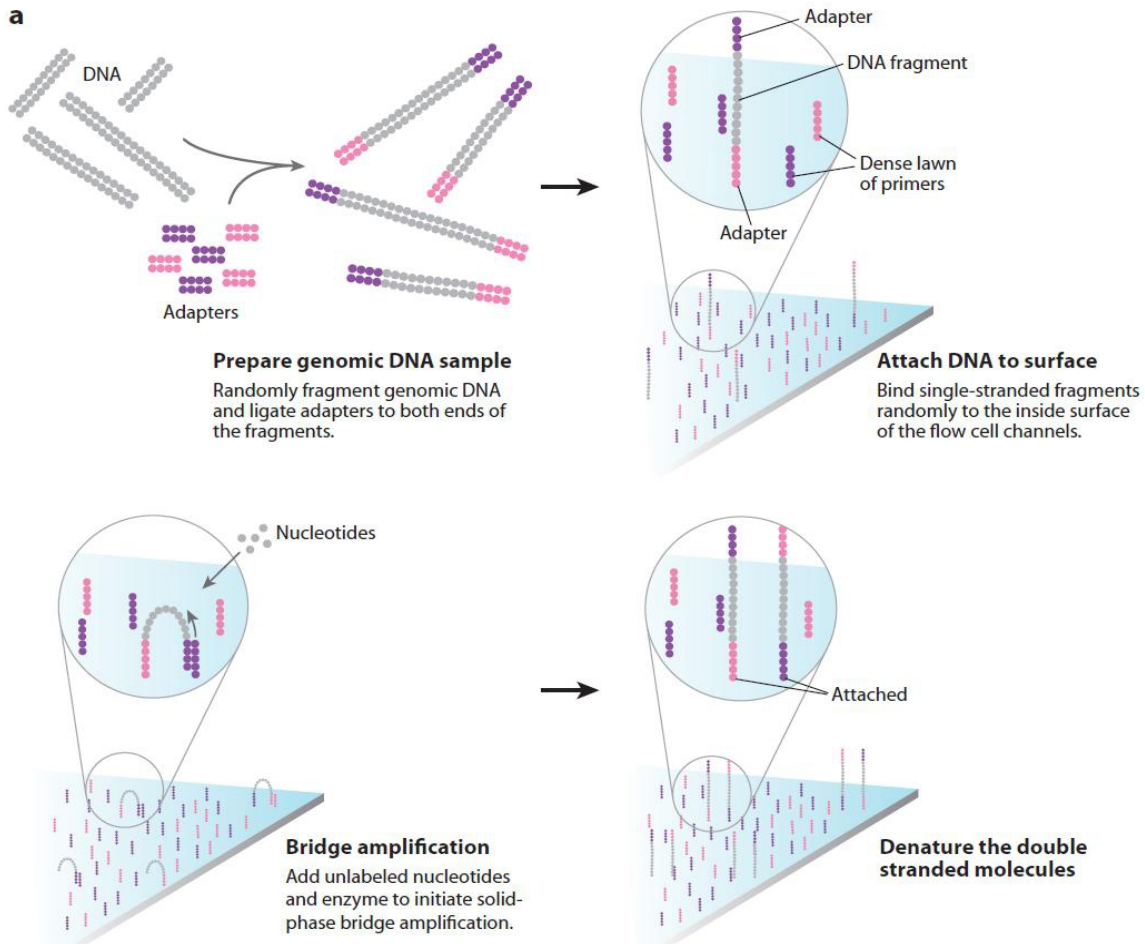
2.3.5.1 Fluorescently labeled sequencing by synthesis (Illumina)

2.3.5.1.1 Technology overview

Belongings to this technology, there are a number of machines with different specifications and in turn with different applications (the Genome Analyzer IIx and HiSeq machines for deep sequencing; MiSeq for small genomes Exome and PCR amplicons) (Table 5). The genomic DNA library is amplified on the surface of a flow cell which is an 8-channel sealed glass micro-fabricated device attached to its surface the Solexa adaptors, complementary-ready for the library amplification. By adding the polymerase, each DNA fragment amplifies in million copy clusters in a process called the bridge amplification (Fig. 5a). The sequencing reagents are added to the flow cell, including the sequencing primer, 3'-OH blocked fluorescently labeled dNTPs, and the polymerase. After the incorporation of the first nucleotide, the unused reagents are washed away from the flow cell and the scan reagent is added to facilitate the image to be taken by the optical system. The fluorescently labeled nucleotides are excited by laser and the images are taken in different units called tiles. Once the imaging is done, other chemicals are added to cleave the blockage at 3'-OH group and permit further nucleotide incorporation. The nucleotide incorporation cycles repeat in number depending on the read length and according to the machine specifications (Fig. 5b) (Hert et al., 2008; Mardis, 2008; Shendure and Ji, 2008; Liu et al., 2012).

Table 5: Different Illumina machines specifications' (adapted from www.Illumina.com).

	HiSeq 2500-1500	HiSeq 2000-1000	HiScanSQ	Genome analyzer IIx	Miseq
Output	600 Gb	300 Gb	150 Gb	95 Gb	7.8-8.5 Gb
Single Reads	3 Billion total	1.5 Billion	750 Million total	320 Million	15-17 Million
Paired-end Reads	6 Billion	3 Billion	1.5 Billion	640 Million	30-34 Million
Required input	50 ng -1 µg	50 ng –1 µg	50 ng –1 µg	50 ng – 1 µg	50 ng – 1 µg
Read length	2 × 100 bp	2 × 100 bp	2 × 100 bp	2 × 150 bp	2 × 150 bp
Percentage of Bases > Q30	> 85% (2 x 50 bp) > 80% (2 x 100 bp)	> 85% (2 x 50 bp) > 80% (2 x 100 bp)	> 85% (2 x 50 bp) > 80% (2 x 100 bp)	> 85% (2 x 50 bp) > 80% (2 x 100 bp)	> 85% (2 x 50 bp) > 70% (2 x 100 bp)


a
DNA
Adapters
Prepare genomic DNA sample
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.
Adapter
DNA fragment
Dense lawn of primers
Adapter
Attach DNA to surface
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.
Nucleotides
Bridge amplification
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
Attached
Denature the double stranded molecules

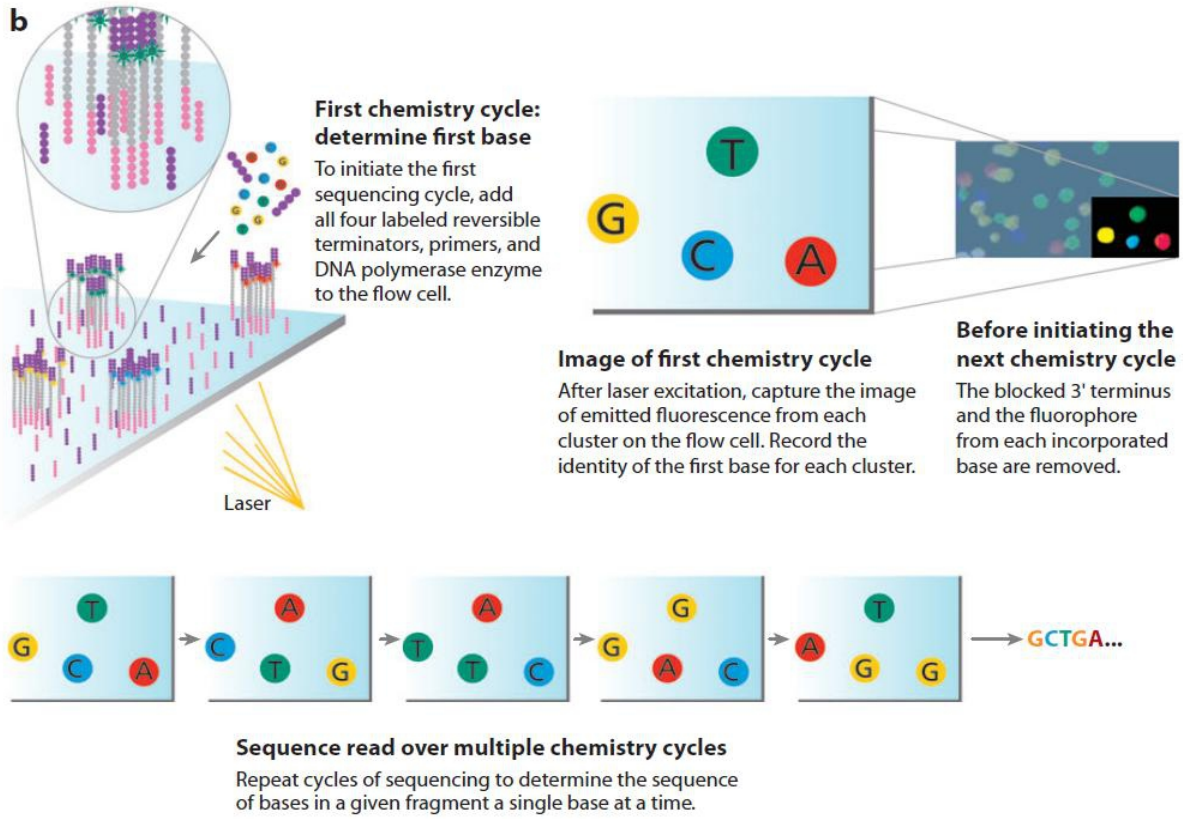


Figure 5: The sequencing by synthesis approach using fluorescently labeled nucleotide. a) Cluster formation after the bridge amplification process. b) The sequencing of clusters using the fluorescently labeled nucleotides following with excitation by the laser and imaging the flow cell (adapted from Mardis., 2008).

2.3.5.1.2 library preparation

2.3.5.1.2.1 End repair

The recommended protocol of New England Biolabs NEBNext™ DNA sample prep was used for whole genome library preparation and sequencing by the Genome Analyzer IIx (GAIIx). Up to 600 ng of genomic DNA was end-repaired using 1 µl DNA Poly I (Klenow LF) and 5 µl T4 DNA Polymerase, 4 µl dNTP mix (10 mM), 10µl of 10x end-polishing buffer in a total volume of 100 µl and incubation at 20°C for 1 hour.

2.3.5.1.2.2 Purification of end-repaired DNA using MinElute

To collect all the DNA fragments, the end-repaired DNA was purified using the MinElute kit as the range of recovered double strand DNA fragments spans in between 70bp-4kb. Five hundred µl of PB Buffer were added to 100 µl of the end-repair reaction product, mixed well and applied to the MinElute column. To increase the binding the DNA to the

MinElute membrane, the columns were incubated for 5 min at room temperature and then centrifuged for 3 min. After discarding the flow out, the MinElute column was washed using 750 µl of PE Buffer and centrifuged for 3 min. To get rid of all the traces of ethanol, additional centrifugation was done. The MinElute column was placed in a clean 1.5 ml microcentrifuge tube to elute the DNA in 35 µl free nuclease water. To increase the DNA yield the column was incubated for another 10 min at room temperature and then centrifuged for 3 min.

2.3.5.1.2.3 dA Tailing

The ligation of Illumina adaptor is mediated through an A tail. Therefore the purified and end-repaired DNA was tailed with dATPs using 5µl of 10X reaction buffer and 3µl of Klenow Fragment (3´→5´ exo) and incubated for 1 hour at 37°C. As previously described in 2.4.5.1.2.2, the DNA was purified using the MiniElute kit in 35 µl of nuclease-free water.

2.3.5.1.2.4 Ligation of adapters to DNA Fragments

A-tailed and purified DNA was ligated to Solexa adaptors (50 µM). In a total volume of 50µl, the adaptors were ligated to purified DNA using 10 µl of 5x T4 ligase buffer, 5µl Quick T4 ligase by incubation at room temperature for 15 minutes.

2.3.5.1.2.5 Purification using Agencourt Ampure XP

Following the ligation and to get rid of the excess of adaptor dimers, purification was done using Agencourt Ampure XP, thus removing all the DNA fragments with sizes less than 200 bp. In a 1:1 volume ratio, 50 µl of Ampure XP were added to the total volume of ligation reaction product and incubated for 5 min at room temperature. Then the mix was placed on a magnetic stand to separate beads from the supernatant. After the solution became clear (in about 5 minutes), the supernatant was discarded. Afterwards, the beads were washed by adding 300 µl of freshly prepared 80% ethanol and resuspension. The tube was placed again on the magnetic stand to separate the beads from the ethanol and the supernatant was carefully removed and discarded. The tube was spun and put back on the magnetic stand, to completely remove the residual ethanol. The beads were air dried by leaving the lid open for 10 minutes. The DNA fragments were eluted from the beads with 30 µl of nuclease-free water through their suspension using a vortex mixer and putting the tube on the magnetic stand until the solution was clear. Finally, the purified ligated DNA was transferred to a clean PCR tube.

2.3.5.1.2.6 Enrichment of the adapter-modified DNA fragments by PCR

The DNA fragments with adapter molecules on both ends were enriched using the PCR to amplify the amount of DNA in the library using the primers P1 and P2. In a total volume of 75 µl, 37.5 ul of 2x Phusion HF Master Mix, 1 µl each of both library PCR primers 1 and 2 (P1 and P2) (10 µM) and 30 µl of adapters ligated DNA were added and completed with water to 75 µl. The enrichment mixture was subjected to the following PCR cycling and the number of PCR cycles was minimized to avoid the library bias

step	temperature	time
1	98°C	30 sec
2	98°C	10 sec
3	65°C	30 sec
4	72°C	30 sec
5	Repetition of steps 2 to 4 for 6 times	
6	72°C	5 min
7	10°C	hold

The enrichment product was purified with Agencourt Ampure XP as previously described in 2.4.5.1.2.5 and eluted using 25 µl nuclease-free water.

2.3.5.1.2.7 Quantification of NGS libraries using qPCR

Before the last indexing and amplification step, qPCR was used to roughly estimate the concentration of each library and additionally to determine the appropriate number of the last amplification PCR cycles. In a 25 µl total volume, qPCR was done using 1 µl of the enriched library, 12.5 µl of 2x Phusion HF Master Mix and 1 µl each of both library primer P1 and the index (10 µM) as well as 1 µl of each SYBRGreen dye and fluorescein of GeneRead qPCR SYBR Green Mastermix kit. The mix was subjected to the following cycling conditions:

step	temperature	time
1	94°C	4 min
2	95°C	30 sec
3	57°C	30 sec
4	72°C	30 sec
5	Repetition from step 2 to step 4 for 20 times	
6	72°C	3 min
7	10°C	hold

2.3.5.1.2.8 Final libraries indexing and amplification

Using the information from the quantification step 2.4.5.1.2.7, the last amplification step of the library was done. Each library was amplified in a total volume of 50 µl using the 1 µl (10 µM) of each P1 and one of the illumina indexes, 25 µl of 2x Phusion HF Master Mix and 23 µl of the purified enriched library. The PCR cycle number depended on the quantification library step, varied from one library to another and was estimated from the qPCR curves. The used PCR cycling was as following:

step	temperature	time
1	94°C	4 min
2	95°C	30 sec
3	57°C	30 sec
4	72°C	30 sec
5	Repetition from step 2 to step 4 (different from library to another)	
6	72 °C	3 min
7	10 °C	Hold

The amplified and indexed libraries were purified again using Agencourt Ampure XP (as previously shown in 2.4.5.1.2.5) and the same 1:1 ratio and eluted in 25 µl nuclease-free water to be ready for NGS illumina sequencers.

2.3.5.1.2.9 Library evaluation and titration

A library aliquot was loaded on 2% agarose gel to check the final library DNA fragment pattern and sizes. Additionally, the library concentration in ng/µl was precisely estimated and quantified using a library aliquot (1 µl) by the Agilent 2100 bioanalyzer.

2.3.5.1.2.10 Cluster generation and NGS sequencing

Solexa libraries were diluted to a final concentration of 6 pM to be ready for Cluster amplification using TrueSeq PE Cluster Kit v2 and according to the illumina protocol. Afterwards, the Solexa samples were sequenced according to the manufacturer's recommendations using TruSeq SBS Kit v5 GA.

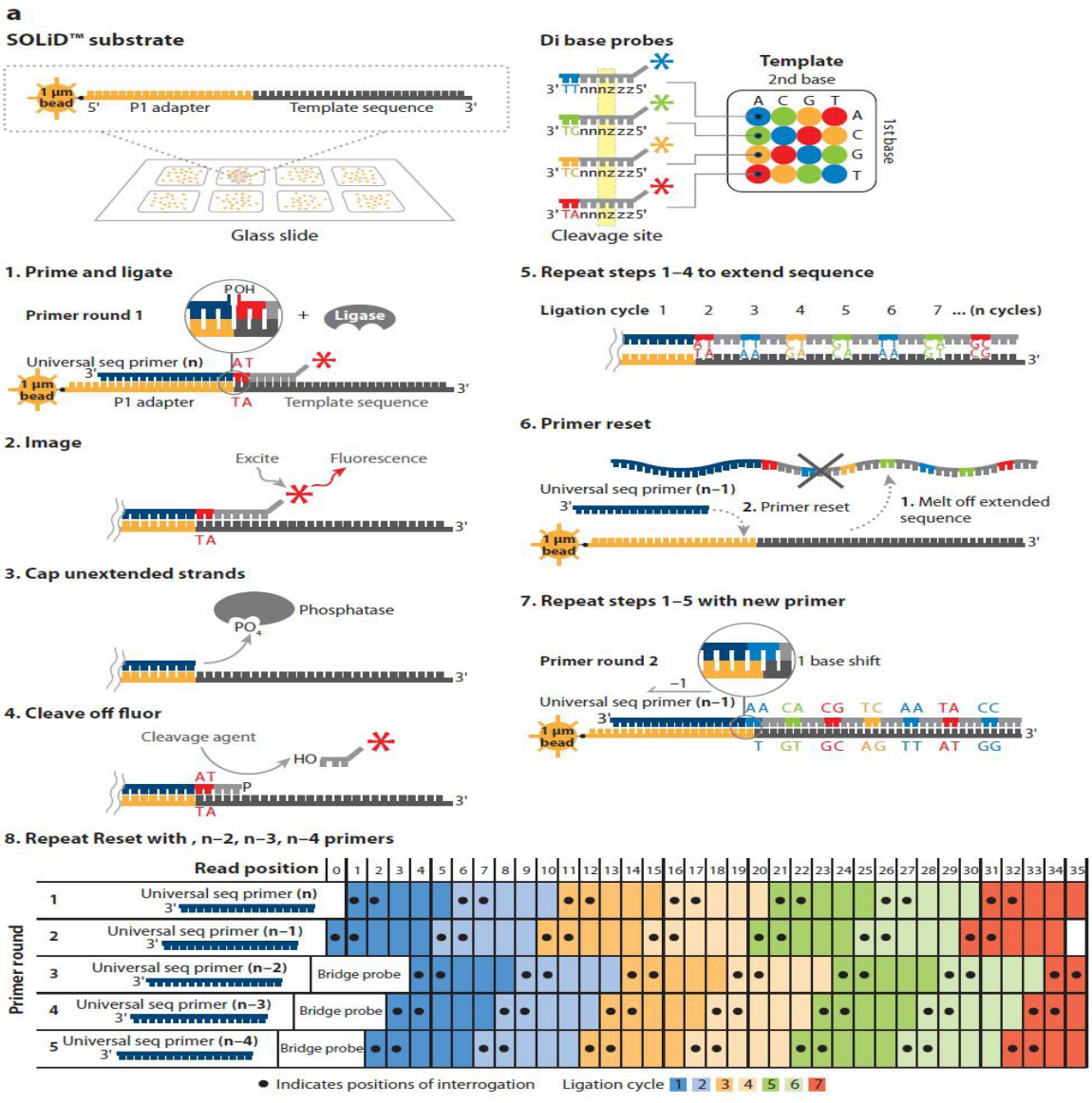
2.3.5.2 Sequencing by hybridization and ligation (SOLiD)

2.3.5.2.1 Technology ovierview

The most recent versions of SOLiD in 2013 are 5500 W, 5500xl W Genetic Analyzers (www.appliedbiosystems.com) which are more superior to the previous SOLiD™ 4 system in terms of efficiency and productivity. In the old SOLiD machines before 5500 W systems, the genomic DNA library was amplified using the emulsion PCR process before applying the amplified library beads to the flow cell, as known from the Roche/454 protocol. With the 5500W the emulsion PCR step is eliminated and the amplification is done directly in the flow cell.

The sequencing by hybridization and ligation approach is based on adding probes to hybridize with DNA library fragments and ligation using the ligase enzyme in a number of cycles. The probe consists of 8 nucleotides and with a different combination possibilities of the four nucleotide bases (Fig. 6a). Following the probes addition to the flow cell, the fully hybridized probe is ligated using the ligase and the fluorescent signal can be detected by the detection system. The non-hybridized fragments are blocked by a specific cap. After the signal detection, the last three bases are cleaved off to permit another hybridization process to be repeated for 7 cycles (Fig. 6a).

After these cycles, another round of hybridization ensues using a new primer whose start point is closer to the bead surface (n-1) and the ligation is repeated for 7 cycles. Thus, each nucleotide will be sequenced twice and the sequencing results will be evaluated and decoded with the mapping to the respective reference genome (Fig 6b) (Hert et al., 2008; Mardis, 2008; Shendure and Ji, 2008; Liu et al., 2012).


a
SOLiD™ substrate
1 µm bead
5'
P1 adapter
Template sequence
3'
Glass slide
Di base probes
Template
2nd base
A C G T
1st base
3' TTnnnzzz5'
3' TGnnnzzz5'
3' TCnnnzzz5'
3' TAnnnzzz5'
Cleavage site
1. Prime and ligate
Primer round 1
POH
Ligase
Universal seq primer (n)
AT
TA
2. Image
Excite
Fluorescence
3. Cap unextended strands
Phosphatase
PO4
4. Cleave off fluor
Cleavage agent
HO
P
5. Repeat steps 1–4 to extend sequence
Ligation cycle 1 2 3 4 5 6 7 ... (n cycles)
6. Primer reset
Universal seq primer (n–1)
2. Primer reset
1. Melt off extended sequence
7. Repeat steps 1–5 with new primer
Primer round 2
1 base shift
–1
Universal seq primer (n–1)
AA CA CG TC AA TA CC
T GT GC AG TT AT GG
8. Repeat Reset with , n–2, n–3, n–4 primers
Read position
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35
Primer round
1 Universal seq primer (n)
2 Universal seq primer (n–1)
3 Universal seq primer (n–2)
4 Universal seq primer (n–3)
5 Universal seq primer (n–4)
Bridge probe
● Indicates positions of interrogation
Ligation cycle 1 2 3 4 5 6 7

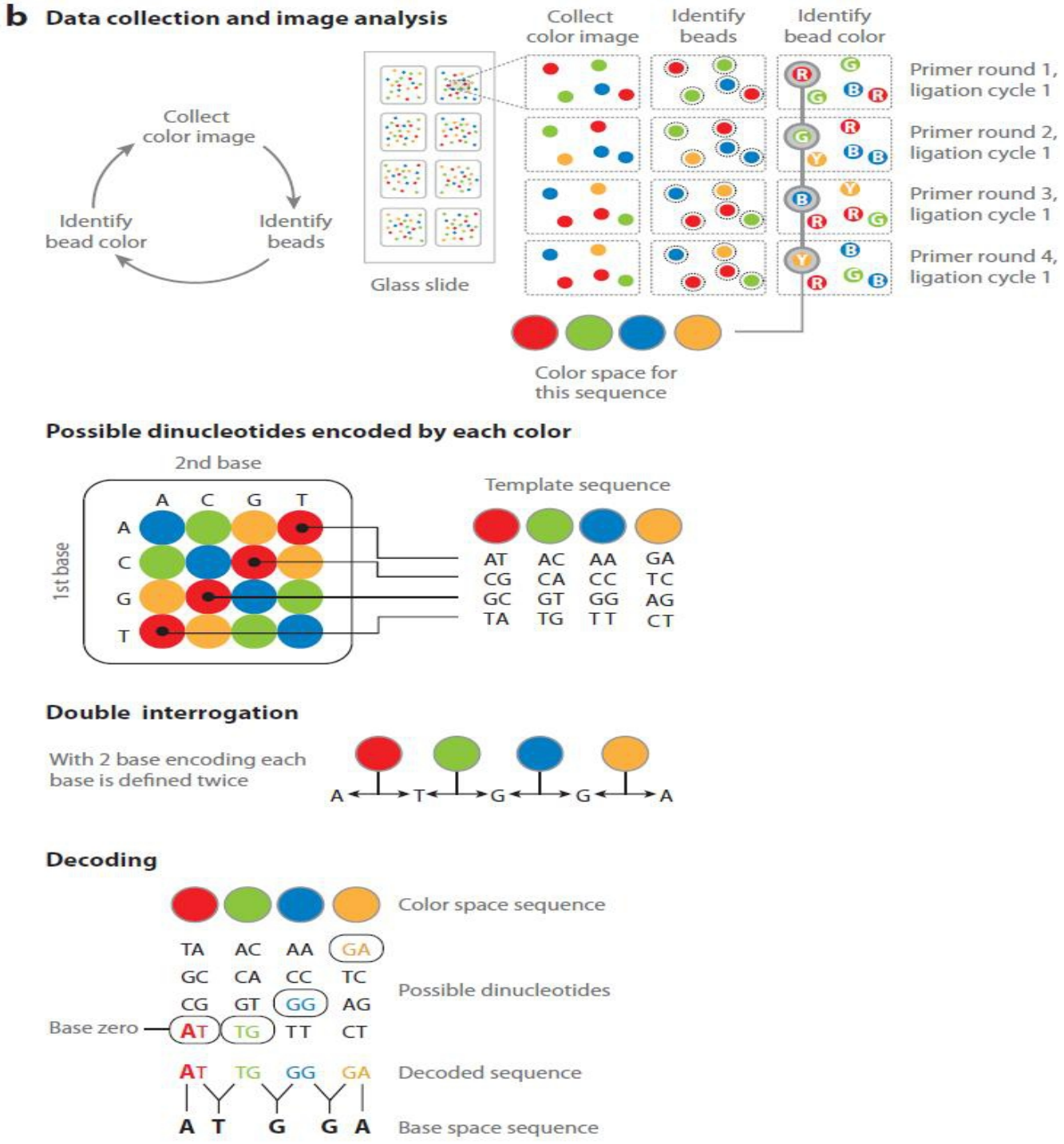


Figure 6: The sequencing by hybridization and ligation a) The hybridization and ligation steps b) The data collection and analysis of dinucleotide-encoded results (adapted from Mardis, 2008).

2.3.5.2.2 Library preparation

2.3.5.2.2.1 End repair

Using SOLiD® Fragment Library Construction Kit, genomic DNA was end-repaired using 1 µl of the endpolishing enzyme 1 (10 U/µl) and 2 µl of the endpolishing enzyme 2 (10 U/µl), 4 µl dNTP mix (10 mM), as well as 20 µl 5x end-polishing buffer, in a total volume of 100 µl. Following incubation at room temperature for 30 min, DNA was purified using the SOLiD™ Library Column Purification Kit.

2.3.5.2.2.2 Ligation

The SOLiD adaptors P1 and P2 (2.5 µM) were ligated to purified DNA with 40 µl of 5x T4 ligase buffer and 10 µl T4 ligase (5 U/µl), in a total volume of 200 µl at room temperature for 15 min.

2.3.5.2.2.3 Enrichment

DNA was then incubated with 380 µl Platinum HiFi PCR Amplification Mix and 10 µl each of both library PCR primers primer 1 and 2. In order to repair the gap in the double-stranded DNA molecules introduced during adaptor ligation. The following conditions for thermal cycling were applied:

step	temperature	time
1	72°C	20 min
2	95°C	5min
3	95°C	15 sec
4	62°C	15 sec
5	70°C	1 min
6	Repetition from step 3 to step 5 for 2 times	
7	70°C	5 min
8	4°C	hold

2.3.5.2.2.4 Purification of the amplified library

The amplified library was purified using the PureLink™ PCR Purification Kit. Four volumes of binding buffer (B2) was added to 1 volume of the amplified library (100 µl) and mixed well. The former mix was added to PureLink® spin column and centrifuged at room temperature at 10,000 × g for 1 minute. After discarding the flow through, the spin column was washed using 650 µl of wash buffer and centrifuged at room temperature at 10,000 ×

g for 1 minute. The flow through was discarded and the column was centrifuged at maximum speed at room temperature for 2–3 minutes to remove any residual wash buffer. The cleaned library was eluted in a clean collecting tube using 50 µl free nuclease water.

2.3.5.2.2.5 Final amplification

The purified library was amplified in total volume of 200 µl which included 100 µl of 2× Phusion HF Master Mix and 8 µl each of both library PCR primers 1 and 2 and cycled using the following conditions:

step	temperature	time
1	95°C	15 sec
2	62°C	15 sec
3	70°C	1 min
4	Repetition from step 1 to step 3 for 12 times	
5	70°C	5 min
6	4°C	hold

All libraries were purified as described above 2.4.5.2.2.4 sequenced using the SOLiD™ 3 Plus System.

2.3.6 Bioinformatic analyses:

The resulting NGS data-sets were sorted by the used index and any reads without the proper index were discarded. The remaining reads were filtered according to quality and the passed reads from the quality control were analyzed in two different approaches. First, mapping with the human genome reference (hg19) and second, metagenomic analysis using BLASTn algorithm and MEGAN software (4.70.4).

2.3.6.1 Mapping with the hg19:

The resulting NGS data-sets were uploaded to open server Galaxy (Giardine et al., 2005; Blankenberg et al., 2010; Goecks et al., 2010) (http://main.g2.bx.psu.edu/root). The data-set quality was checked using the FASTX-Tool kit. Then the reads in fastqsanger format were mapped against (hg19) using the Bowtie and BWA algorithms. The resulting SAM (Sequence Alignment/ Map) file was used to calculate the human content percentage and to determine the number of unique human reads. Additionally, the file was

used to filter the mitochondrial reads and calculate its percentage. After filtration of the unmapped reads, the SAM file was converted to binary format (BAM) file for browsing against the hg19 using the integrative genome viewer (IGV) (version 2.3).

2.3.6.2 SNP calling and haplogroup determination:

The BAM file was used for pileup file generation. It is a file format which describes the information of every mapped base through the genome. the file had a variable format but mainly describes the consensus base pair in this site, number of reads covering it and the quality of those base pairs (Fig. 7) (http://samtools.sourceforge.net/pileup.shtml).

```
seq1 272 T 24  ,.$.....,,,,,,,,,,,,,,,^+. <<<+;<<<<<<<<<<=<;<;7<&
seq1 273 T 23  ,.....,,,,,,,,,,,,,,,,,A <<<;<<<<<<<<3<=<<<;<<+
seq1 274 T 23  ,.$.....,,,,,,,,,,,,,,,,,    7<7;<;<<<<<<<<<=<;<;<<6
seq1 275 A 23  ,$.....,,,,,,,,,,,,,,,,,^l.  <+;9*<<<<<<<<<=<<:;<<<<
seq1 276 G 22  ...T,,,,,,,,,,,,,,,,,,  33;+<<7=7<<7<&<<1;<<6<
seq1 277 T 22  ....,,,,,,.C.,,,,,,,G.  +7<;<<<<<<<&<=<<:;<<&<
seq1 278 G 23  .....,,,,,,,,,,,,,,,,,,,^k.   %38*<<;<7<<7<=<<<;<<<<<
seq1 279 C 23  A..T,,,,,,,,,,,,,,,,,,, ;75&<<<<<<<<<=<<<9<<:<<
```

Figure 7: Example of the pileup file format with information of each mapped base pair (adapted from http://samtools.sourceforge.net/pileup.shtml)

The pileup file was used for SNP determination using the SNPs calling software VarScan 2 (Koboldt et al., 2012) (http://varscan.sourceforge.net/). To precisely define the SNPs with high quality, SNP calling parameters were adjusted as recommended by the software manual and set to *p*-value threshold of 0.05, with a minimum coverage of 8 reads, a minimum average base quality of 15 and a minimum allele variant frequency of 0.2.

The determined SNPs were used for the haplogroup determination using HaploGrep. It is a web-based application for the haplogroup determination (http://haplogrep.uibk.ac.at/) using the updated phylogenetic tree of global human mitochondrial DNA variation the Phylotree (http://www.phylotree.org/tree/main.htm/) (van Oven and Kayser, 2009; Kloss-Brandstaetter et al., 2010; Kloss-Brandstätter et al., 2011). The software was fed with the SNPs combinations and the proposed haplogroups were ordered by quality score percentage according to the used SNPs and its informativeness.

2.3.6.3 Metagenomic analysis

To study the metagenomics pattern of the Egyptian mummies, the mummy NGS

data-sets were aligned against NCBI nucleotide collection database (nr/nt) using the BLASTn algorithm (Altschul et al., 1990; Poinar et al., 2006; Huson et al., 2007). Generally, alignment was done using either BLASTn or Mega BLAST algorithms against (nr/nt) database. The analysis was done on the Tübingen university Galaxy server (https://galaxy.informatik.uni-tuebingen.de/galaxy-local/) and the open Galaxy server (https://galaxy.wur.nl) (hosted by Wageningen University and Research centre. The used word size was 28, 30, or 42 and the other used parameters were mainly the default of the galaxy server like the expectation value threshould of 0.001 to increase the likehood of the nearly exact matching hits. As well low complexity sequences were out-filtered using DUST algorithm to avoid the bias which could be introduced by the repetitive sequences.

The resulting hits were analyzed and browsed using MEGAN software (4.70.4) and different thresholds combinations were used to examine the specificity of the taxa representation and assignments (Huson et al., 2007; 2011). To increase the specificity of the MEGAN assignments, a strict set of MEGAN LCA parameters was used first. The strict set composed of a minimum bit score threshold of 55 and the win score of 80 in combination with minimum threshold of the required reads to support the taxa definition of 10 and top perecent threshold of 10 as well as the minimum complexity to 0.3 as proposed by the MEGAN manual to catch most of the low complexity short reads. This was followed by releasing the thresholds of the MEGAN LCA parameters and adopting another setting. The minimum number of the required reads to support the taxa definition was set to 5, the minimum bit score threshold to 35 and the win score to 50. Comparison of the results that were obtained from the application of the two thresholds settings was to evaluate the MEGAN results and findings.

Two pairs of Egyptian mummies DNA samples were dedicated for metagenomic analysis. They were pooled into two pairs DMG M1 (DMG2-I and DMG3-I) and DMG M2 (DMG1-I and DMG6-I). Both pools were used for library preparation according to the recommended Illumina protocol and subjected to sequencing using the Solexa platform. After mapping of the data-sets (DMG M 1, 2) against the hg19 and removing all the human reads, the unmapped reads were aligned against (nr/nt) database with word length of 28. The resulting tabular file was used for more analysis using MEGAN software. To reach the most accurate taxa assignment, the strict set of LCA thresholds as well as the less strict one were used and the resulting MEGAN files were compared.

The DNA samples from the mummies DMG 1, 2, were also used NGS by the SOLiD 3 technology resulting in DMG1-II and DMG2-II data-sets, respectively. Others

samples from the same mummies DMG 1 and 2 were taken at another time interval (after 1.5-2 years) and called DMG1-V and DMG2-III, respectively. The DNA samples DMG1-V and DMG2-III were sequenced using Illumina sequencing and resulted in the data-sets of DMG1-V and DMG2-III. One million reads of DMG1-II, -V and DMG2 -II and -III data-sets, were used for metagenomic analysis after alignment with BLASTn algorithm and comparison using MEGAN.

Another BLASTn alignment against nr/nt database was done using 1 milion reads of the data-set DMG 5. The alignment was done against (nr/nt) database using the recommended default settings by Galaxy server. Since the word size may affect the sensitivity and taxa prediction, different word lengths (30 and 42) were used and the resulting BLAST files were compared using MEGAN.

Finally, the bacterial and Viridiplantae taxa were compared using one million reads from each of the data-sets of the mummies (DMG 1, 2, 4, 5), the South-American skeleton samples (DMGS-1000 and DMGS-2000) and three previously published from cold climate specimens (the Saqqaq (Rasmussen et al., 2010), the Alpine Iceman (Keller et al., 2012), the Denisova hominin (Reich et al., 2010)) (Table 6). One million reads from each data-set was used for aligning against (nr/nt) database using either BLASTn or Mega BLAST algorithm. The resulting tabular files were further analyzed with MEGAN software. Low thresholds of the MEGAN LCA parameters were adopted in this analysis. The minimum number of the required reads to support the taxa definition was set to 1, the minimum bit score threshold to 35, the win score to 0. Comparsion of the results that were obtained from the application of the the forementioned thresholds setting was done .

Table 6: The previously published data-sets which were used for the metagenomics comparison with the study samples. H= hard tissue, AT= average temperature.

Specimen	Habitus	Tissue type	Excavation location	Dating	Dynasty/ period	A T 15°C	Reference
Saqqaq	hair	Hair	Cryotic soil, Western Greenland	4,044±31 BP	-	<<	Rasmussen et al. 2010
Denisova SL3003	digit	Bone	Cave soil, Russia	50,000-30,000 BP	Earlier Middle Palaeolithic	<<	Reich et al. 2010
Iceman	mummy	Bone	Glaciers, Italy	5,300 BP	Old Copper age	<<	Keller et al. 2012

3 Results

3.1 DNA extraction

Different biopsies from eight Egyptian mummy heads were used for the DNA extraction. The samples were either bone and/or muscle tissues. To optimize the concentration of extracted DNA from the different tissues, the DNA extraction protocol was modified and applied to the samples obtained from various Egyptian mummies. The phenol/ chloroform protocol for DNA extraction (Scholz and Pusch, 1997) was used with or without applying the MagNA Pure compact system. The DNA extracts were checked using agrose gel electrophoresis. Under UV-illumination, the gels were visualized before and after applying ethidium bromide. This first analysis step helped to reveal, the presence of co-extracted substances, which could seriously affect the subsequent DNA analysis (Fig.8).

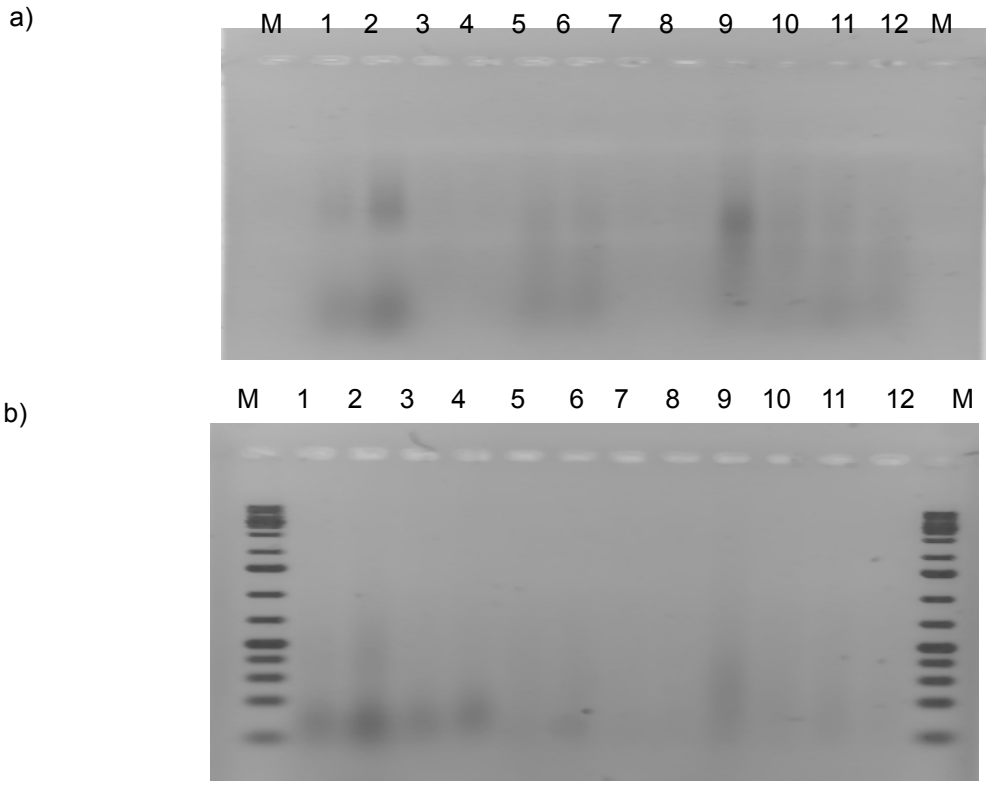


Figure 8: The DNA yields from twelve different Egyptian mummy samples extracted using the phenol/chloroform protocol a) before applying ethidium bromide and b) after applying ethidium bromide. Lanes (1-4) correspond to different DNA extracts obtained from DMG 2 biopsies. Lanes (5-8) correspond to different DNA extracts obtained from DMG 6 biopsies. Lanes (9-12) correspond to different DNA extracts obtained from DMG 1 biopsies. M is GeneRuler™ 1Kb plus DNA Ladder.

The combination of the phenol/ chloroform protocol with the MagNA Pure compact system had two notable effects on the resulting DNA extracts. First, the MagNA Pure system cleaned the resultant DNA from most of the co-extracted substances. Unfortunately, this was at the expense of the resulting DNA concentration in general. Second, the magnetic separation of the extracted DNA with the MagNA Pure caused a particular loss of DNA fragments less than 100 bp. On the other hand, bigger fragments up to 400 bp were frequently visualized by gel electrophoresis (Fig. 9).

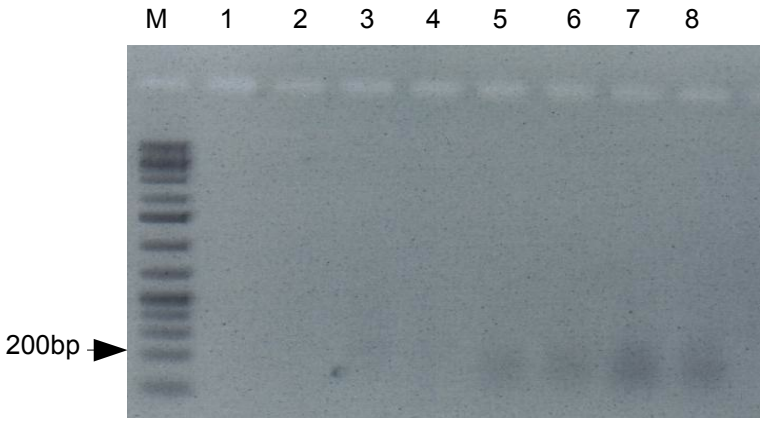


Figure 9: The DNA yields from various mummy samples that were extracted using the phenol/ chloroform protocol in combination with the MagNA Pure compact system. Lanes (1-4) correspond to different DNA extracts obtained from DMG 1 biopsies. Lanes (5-8) correspond to different DNA extracts obtained from DMG 2 biopsies. M is GeneRuler™ 1Kb DNA Ladder.

Aliquots of the purified DNA were used for the concentration measurement using the Nanodrop device. The concentrations of the mummy DNA samples were within a range of 3 to 25 ng/µl. To estimate the inhibitory effect of the co-extracted substances on the PCR reaction, a 1:10 diluted aliquot from each extract was used for a spiking reaction (Pusch and Bachmann, 2004). The inhibitory effect of the co-extracted materials was variable among the Egyptian mummy DNAs. Inhibition was more evident in the case of the DNA extracted using the phenol/chloroform protocol alone. Those extracted using the phenol/chloroform in combination with the MagNA Pure compact system showed no or negligible inhibition of the PCR reaction (Fig. 10).

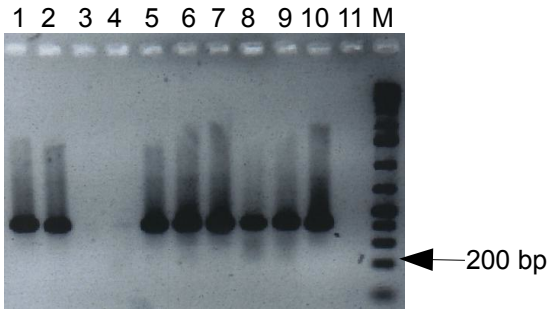


Figure 10: Spiking reaction using diluted aliquots of the Egyptian mummy DNA samples. Lanes 1-10 correspond to various 400 bp PCR products of a plasmid amplification reaction. They were generated in the presence of 1 µl of 1:10 diluted DNA aliquots from various Egyptian mummy DNA samples. Lanes (1-4) correspond to different DNA extracts obtained from DMG 6 biopsies. Lanes (5-8) correspond to different DNA extracts obtained from DMG 1 biopsies. Lane 9 corresponds to DMG 5 DNA. Lane 10 corresponds to DMG2 DNA. lane 11 is a negative control, M is GeneRuler™ 1Kb DNA Ladder.

Based on the the spiking PCR experiments and the DNA concentration estimation using Nanodrop, the extracted DNAs from four mummies (DMG 3, 6, 7 and 8) had either a low DNA concentration or exhibited co-extracted substances. Vice versa, the other four mummies (DMG 1, 2, 4, and 5) were higher in the estimated DNA concentration and showed a negligible PCR inhibition. Consequently, downstream NGS experiments were done on the latter extracts set. Since combined use of MagNA Pure compact system and phenol/chloroform protocol yielded the best results, it was adopted as the most optimal protocol for DNA extraction from mummies. Finally, by getting rid of the majority of inhibitors we obtained high quality DNA with sizes up to 400 bp, though in decreased amounts.

3.2 Characterization using NGS technology

Characterization by the NGS technology was employed to gather more information about the metagenomic pattern of the Egyptian DNA samples and its total human DNA content. Various DNA samples were characterized using small scale sequencing runs. DMG1 samples were available as soft and hard tissue biopsies. To assess the tissue type effect on the human content, the mummy soft tissue DMG1-III and hard tissue DMG1-IV were analyzed and compared. DNAs were sequenced using a small scale NGS run. The NGS libraries were prepared according to the recommended Illumina protocol by the using the Solexa adaptors and primers. For validation purposes, the libraries were cloned using the pJET 2.1 cloning kit and sequenced to examine the library DNA fragment

structure. The results showed the proper structure of the NGS library with adaptor ends and a short insert in between .

Following the indexing and the final amplification, the library concentrations and the fragment size ranges were examined using gel electrophoresis (Fig. 11). Using the Agilent 2100 bioanalyzer (Fig. 12), the fragment size ranges in the DMG1-III and DMG1-IV libraries were 260-586 bp and 265-338 bp, respectively. In both libraries, the fragment range from 265 bp to 338 bp exhibited the highest concentration (5.78-10.4 ng/ µl) (Fig. 12).

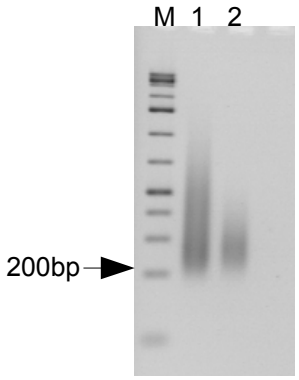
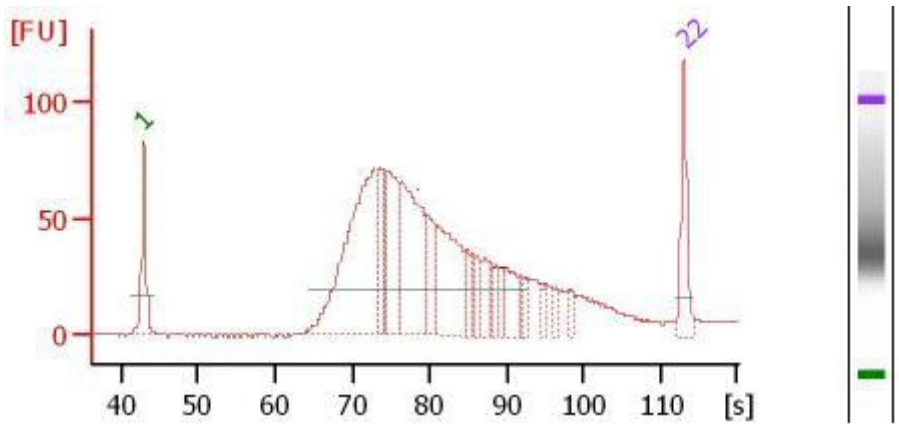


Figure 11: A gel image of the two Solexa libraries DMG1-III, DMG1-IV. Lanes 1 and 2 correspond to the DMG1-III and DMG1-IV libraries, respectively. Their DNA fragment size was within a range of 200-600bp. M is GeneRuler™ 1Kb DNA Ladder.

a)



b)

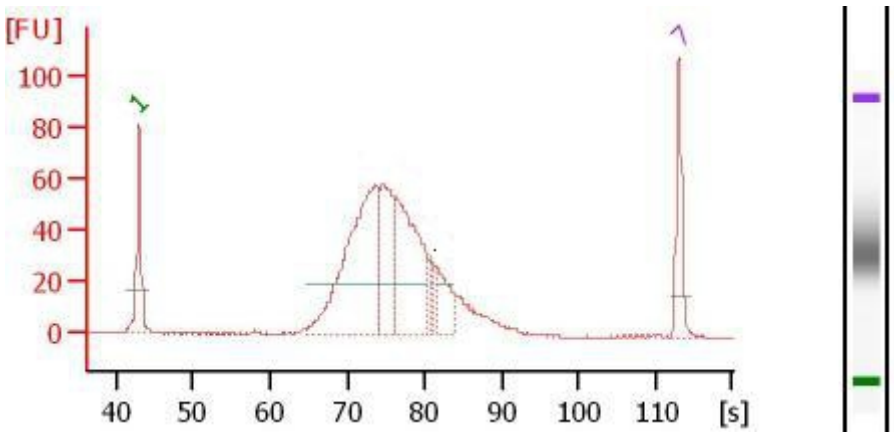


Figure 12: DNA concentrations and fragment size ranges of the DMG1-III and DMG1-IV Solexa libraries. They were determined using Agilent 2100 bioanalyzer. (a) DMG1-III (b) DMG1-IV, respectively. For each sample, there are two visual representations, an electropherogram (left) and a gel-like image (right). [s] is the integration time, [FU] is the height threshold. The fragments size ranges in DMG1-III and DMG1-IV were 260-586 bp and 265- 338 bp, respectively. The peaks a) 1 and 22 and b) 1 and 7 correspond to the size marker start and end, respectively.

The libraries were sequenced on the Genome Analyzer IIx. The resulting data-sets were about 31.9 million reads from DMG1-III and 29.8 millions from DMG1-IV (table 7). The data-set reads were aligned against the hg19 reference sequence using the mapping software Bowtie. The number of reads uniquely mapped to the human genome was higher from the DMG1-IV (17690) in comparison to those from the DMG1-III (11989). Neither DMG1-III nor DMG1-IV resulted in any mitochondrial reads. The complexity values, which referred to the percentage of the unique reads within the sequenced DMG1-III and DMG1-IV libraries, were 6.4% and 13.3% respectively (table 7).

Table 7: Results of sequencing and mapping of the DMG1-III and DMG1-IV libraries against the hg19 reference sequence.

sample	biopsy	total number of reads	number of unique reads	no. of mapped human reads	no. of unique mapped human reads	no. of mapped human mito reads	complexity
DMG1-III	soft tissue	31901965	2058002	243242	11989	0	6.40%
DMG1-IV	hard tissue	29784942	3962266	102007	17690	0	13.30%

For comparison purposes within different warm climate samples, DNA samples from two mineral soil buried skeletons from South America, DMGS-1000 and DMGS-2000, as well as from the Egyptian mummies DMG1, DMG2 and DMG4, were used for

another sequencing run on the Genome Analyzer IIx. In order to further improve the complexity of the preliminary run of DMG1-III and DMG1-IV, DNA concentrations were altered. Thus, the used DNA concentrations for the library preparations were increased to 116.25-637.5 ng in 35 µl. The bias introduced by PCR duplicates is one of the main obstacles affecting the quality of the generated NGS data. Prior to the final amplification and indexing of the libraries, a library aliquot was utilized in a qPCR reaction using the iCycler iQ™ Real Time PCR Detection System (BIO-RAD) (table 8). As established for the quantification of the NGS libraries (Buehler et al., 2010), the qPCR results were used to estimate the optimal number of library amplification cycles. The sequencing was done with paired end reads each of 76 bp length. The mapping was done against the hg19 reference sequence using the Bowtie applying default parameters for paired-end mapping (table 8). In the case of Egyptian mummy data-sets, as the DNA concentration increased, the number of the qPCR cycles number decreased and the complexity values of the resulting NGS data-sets rose. In contrast, the increase of the used DNA concentrations of DMGS-1000 and DMGS-2000 for the library preparation did not show the same effect (table 8).

Table 8: The DNA concentrations, the qPCR threshold cycle (C_t), the NGS sequencing and mapping results of various Egyptian mummies as well as DMGS-1000 and DMGS-2000. The alignment was carried out against the hg19 reference sequence. S= soft tissue, H= hard tissue

sample	sample type	C_t	total DNA [ng]	total no.of reads	unique reads	human reads	unique human reads	mito reads	unique mito reads	complexity %
DMG1-V	S	13.8	637.5	8748358	6477628	5668	4087	0	0	74.04%
DMG1-VI	H	16.9	116.25	9427780	4479801	36158	11114	36	12	47.51%
DMG2-III	S	15.1	536.25	7230450	5221935	58128	34091	96	58	72.22%
DMG4-I	S	16.7	255	7145286	4703167	40112	18142	30	10	65.80%
DMG-S1000	H	17.9	570	7221914	4263704	67475	22219	34	12	59.03%
DMG-S2000	H	12.2	120	28634834	25620197	10150	7267	6	3	89.47%

Using Bowtie, the number of paired-end reads mapped to the human genome was low in comparison to those single end ones of the forward or the reverse data-sets (table 9). For comparison purpose, the softwares Bowtie and BWA were used for paired-end and single-end mapping of the DMG1-V data-set. By mapping with the BWA software, the number of the paired-end mapped reads increased to 11,545 reads in comparison to

5,668 reads mapped using the alternative software Bowtie (table 9). The sum of results of the single-end mapping using BWA (10,696 reads) was comparable to the paired-end mapping (11,545 reads). On the other hand, this was not the case with the results of Bowtie mapping, where the sum of the results of single end mapping was 27,942 while the paired end reads were only 5,668 (table 9).

Table 9: The comparison of the paired-end and single-end mapping of DMG1-V using the two mapping softwares, Bowtie and BWA.

mapping software	forward reads	reverse reads	paired end mapped reads to human genome	forward reads mapped to human genome	reverse reads mapped to human genome
Bowtie	4374179	4374179	5668	13349	14593
BWA	4374179	4374179	11545	5673	5023

To assess the effect of the DNA extraction protocol on the NGS output, an experiment was conducted using the same biopsy DMG2-IV. The final aqueous extract according to the phenol/chloroform protocol was subdivided into two parts, one used for DNA precipitation using isopropanol alcohol, and the second additionally purified by the MagNA pure compact system. The resulting DNA extracts were used for the NGS library preparation and the resulting NGS data-sets were mapped against the hg19 reference sequence using BWA software. Table 10 shows the comparison between the mapping results from the two data-sets. The complexity was higher with the sole use of phenol/ chloroform protocol in comparison to that combined with MagNA pure application. The number of the mapped reads to the human genome was higher in the MagNA pure generated data-set (18,530 reads) than in the data-set generated with the phenol/ chloroform protocol alone (10,648 reads).

To investigate if different sampling from the same mummy had an effect on the generated NGS results, another biopsy DMG2-V DNA was used for NGS library preparation and sequencing. The DMG2-IV and DMG2-V DNAs were extracted using the same phenol/chloroform protocol and both were muscle specimens. Comparing the two data-sets of DMG2-IV and DMG2-V (table 10), it was shown that the number of mapped human reads in the latter was nearly half of that in the former.

Table 10: The effect of the DNA extraction protocol and biopsy type on the NGS output and the number of the mapped human reads from two DMG2 biopsies. S= soft tissue.

sample	biopsy	extraction method	total reads	total unique reads	complexity %	Human %	Human reads	unique human	Mito reads
DMG2-IV	S	phenol	8126534	7377048	90.77	0.12	10,648	8,974	0
DMG2-IV	S	MagNA Pure	5358046	3863177	72.1	0.32	18,530	12,313	0
DMG2-V	S	phenol	9802790	8393081	85.61	0.05	5,558	3,968	0

Since the initial experiments had exhausted the first biopsy batch, another set of biopsies was taken to be utilized for further NGS experiments. Three different samples were taken from the DMG1 mummy, one bone, DMG1-VII and two muscle biopsies DMG1-VIII and DMG1-IX. One bone sample was taken from the DMG5 mummy, DMG5-I. From the mummy DMG6, two different samples were taken, a bone biopsy, DMG6-II and a muscle specimen DMG6-III. Six new NGS libraries were prepared from the six DNA extracts of the three mummies (table 11).

Prior to the final indexing and amplification of the NGS libraries, aliquots were used for qPCR to estimate the optimal final amplification cycles for each library (Fig. 13). Utilizing the resulting information from the qPCR reaction, indexing of each library was done using a unique index and the suitable number of PCR cycles. Following the indexing, the libraries were examined using gel electrophoresis and the Agilent bioanalyzer 2100. The fragment size range was between 200 bp and 500 bp (Fig. 14). The libraries' concentrations were within the range of 4.15- 6.70 ng/ µl. The libraries were pooled and sequenced with paired end read of 100 bp of length.

The total reads number of the resulting NGS data-sets ranged between 6.8 to 11.6 million. The data-sets were mapped against the hg19 reference sequence using BWA mapping software. Compared to data-sets generated from the initial experiments, the complexity of the second set libraries increased and ranged between 83- 95%. The human content percentage ranged from 0.02 to 2.34%. The DMG5-I library showed the highest human content of 2.34% (table 11).

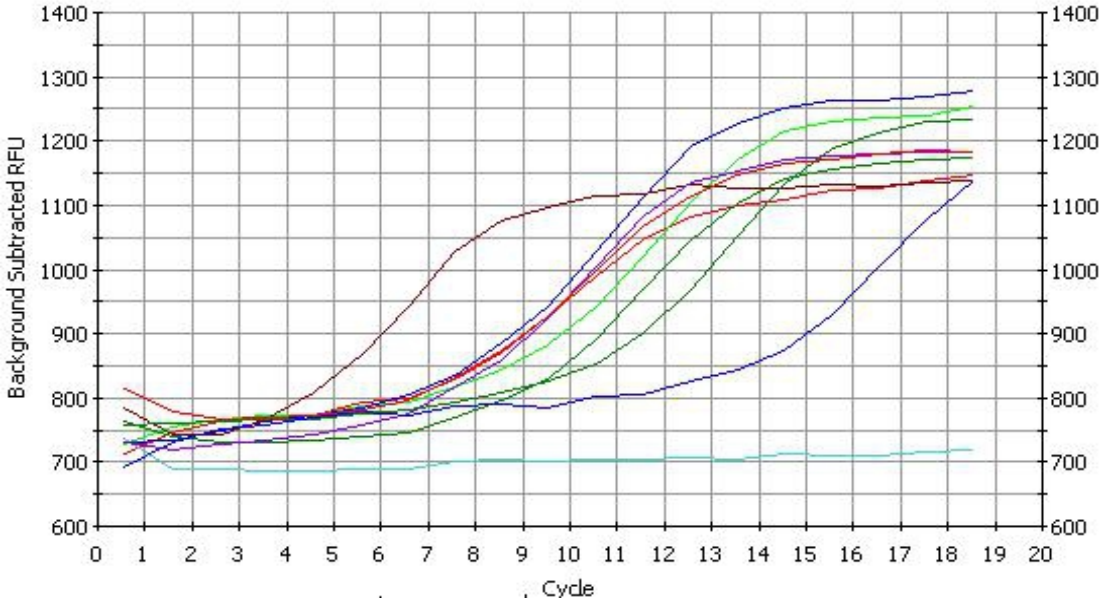


Figure 13: The result curve of iCycler iQ™ Real Time PCR Detection System, which represents the PCR amplification plot of different mummies NGS libraries and the negative control (light blue).

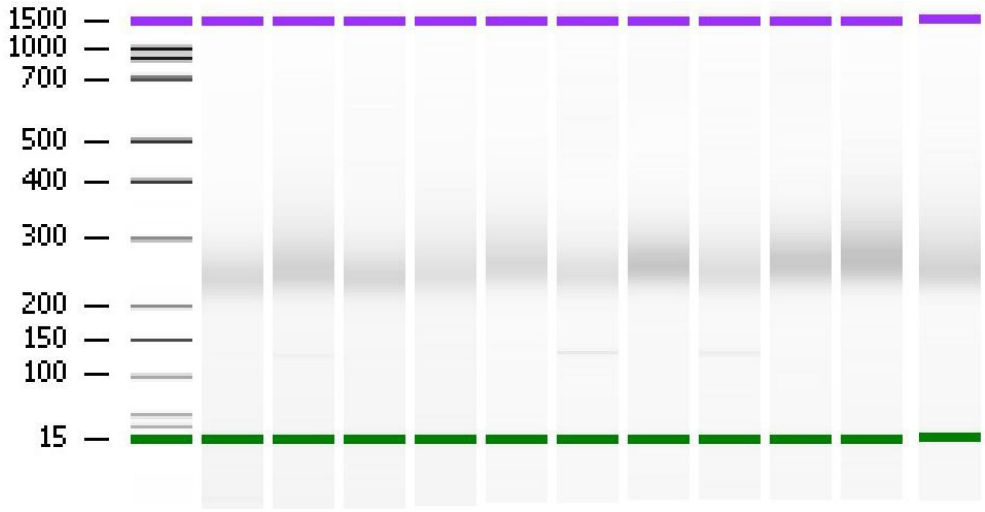


Figure 14: The gel-like image produced by the Agilent bioanalyzer of various NGS libraries from four different Egyptian mummy DNA samples. The first left lane is corresponding to the ladder with size range (15 to 1500 bp). The green and violet dashed lines are representative to the the first and last bands of the used ladder.

About 150,000 reads of the DMG5-I data-set were aligned to the hg19 reference sequence and about 1,900 reads mapped to the mitochondrial genome reference sequence (accession number NC_012920). The mapped reads to the mitochondrial genome reference were sufficient to cover the mitochondrial DNA sequence with an average coverage of 11.6 times (table 11). The good coverage of the mitochondrial

human genome facilitated the clear determination of a number of diagnostic SNPs (Fig. 15). By aligning the consensus mitochondrial sequence, the blast results gave a clear indication of haplogroup I2 (Terreros et al., 2011; Behar et al., 2012; Fernandes et al., 2012).

Table 11: The mapping results of six different libraries generated from three Egyptian mummies. The mapped human reads and the mapped reads to the mitochondrial genome reference (accession number NC_012920) and their coverage to it, are shown. S= soft tissue, H= hard tissue.

sample	biopsy	total reads	total unique reads	complexity %	Human %	Human reads	unique human	Mito reads	Mito coverage
DMG1-VII	H	7377982	6862661	93	0.16	12188	10885	14	0.08
DMG1-VIII	S	8782756	8366547	95.26	0.02	3042	1475	2	0.01
DMG1-IX	S	11675632	10101488	86.51	0.04	4694	3756	22	0.1
DMG5-I	H	6825330	6269541	91.85	2.34	154034	146600	1900	11.6
DMG6-II	H	6930836	6501315	93.8	0.03	1884	1820	2	0.01
DMG6-III	S	9010662	7507397	83.31	0.02	1236	1192	0	0

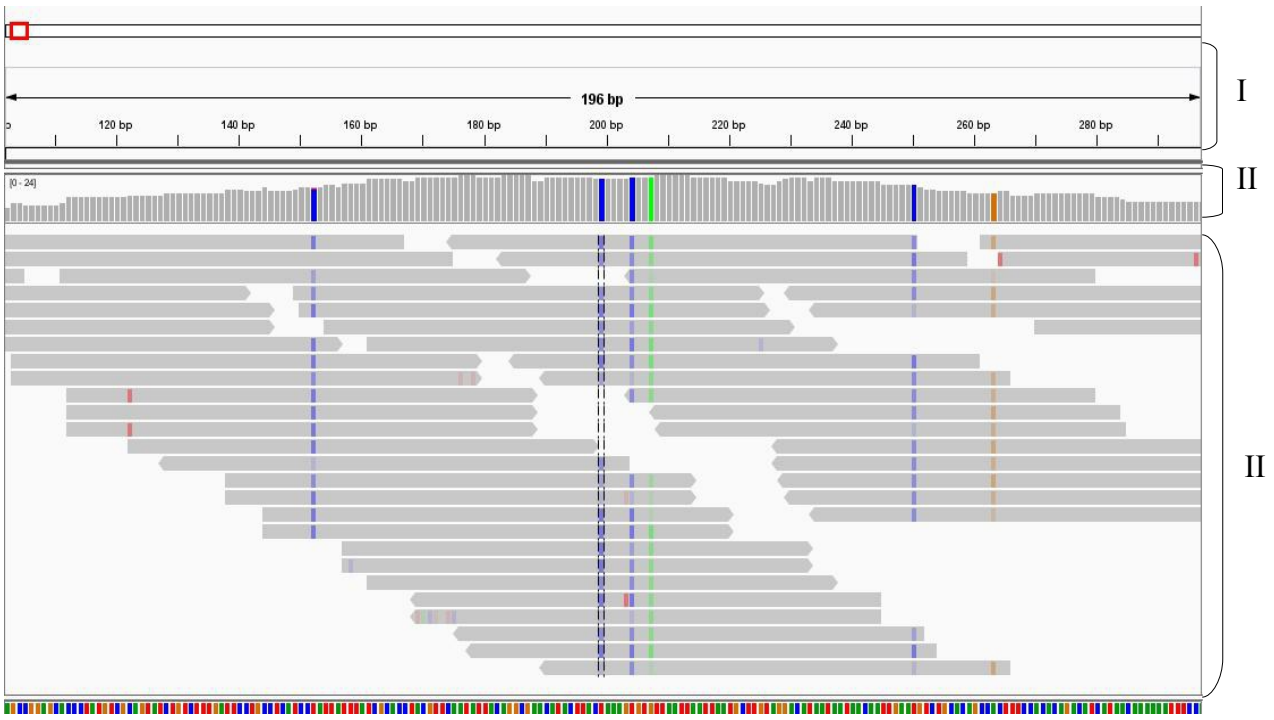


Figure 15: The integrative genomics viewer (IGV) software snapshot of a 196 bp mitochondrial HVII segment sequence of the DMG5 data-set, mapped against the hg19 sequence. Bar I is corresponding to the positions in the mitochondrial genome reference sequence. Bar II is corresponding to the contig coverage. Bar III is corresponding to the sequences that have been mapped to the mitochondrial genome reference sequence. The grey color bar is corresponding to the aligned reads. The blue colour corresponds to cytosine mutation in the aligned reads (C). The green colour corresponds to adenine mutation in the aligned reads (A). The brown colour corresponds to guanine mutation in the aligned reads (G). The red colour corresponds to thymine mutation in the aligned reads (T).

3.3 Large-scale sequencing

As a result of the initial NGS characterization runs, the DMG5-I showed the highest human content. Hence, the DMG5-I library was subjected to another paired end sequencing run with 100 bp read length for precise haplogroup determination and to retrieve more genetic information (table 12). The resulting data-set consisted of about 258 million paired end reads. The whole data-set was used for paired end mapping using BWA software against the hg19 reference sequence. The mapping resulted in about 6 million reads mapped to the human genome, constituting a human content percentage of 2.37% of the whole sequencing data-set. This value was nearly equal to the first small-scale trial of sequencing run (i.e. 2.34%). Over 19 thousand reads were uniquely mapped to the mitochondrial human genome reference (accession number NC_012920) covering it 117.9 times (table 12).

Table 12: Mapping results of the large-scale sequencing DMG5-I data-set. Mapping was done against the hg19 sequence.

total reads	258619448
unique reads	90409764
human reads	6153259
unique human reads	2433932
unique human %	2.37%
complexity %	35.00%
total mito reads	66425
average coverage of total mito reads	400.9 X
unique mito reads	19543
average coverage of unique mito reads	117.9 X

The complexity of the NGS library is an important factor which affects the data recovery from the NGS library and it is a direct indicator of the unique reads number. The complexity of the DMG5-I data-sets decreased by increasing the number of the sequenced reads (tables 11 and 12). To estimate the pattern of the complexity decrease obtained after massive sequencing throughput, read subsets from the forward DMG5-I deep sequencing data-set were examined. Correlating the number of reads versus the complexity percentage values showed that there was an inverse relationship between the two parameters. The increase in the reads number caused a decline in complexity values (table 13, Fig. 16). However, the decreasing pattern of complexity started to level off when it reached its half at a number of 50 millions reads (table 13, Fig. 16).

Table 13: Correlation between the read numbers of DMG5-I NGS data-set and the complexity percentage values

number of reads	unique reads	complexity
1000000	976177	97.60%
10000000	8239082	82.30%
25000000	16697827	66.80%
50000000	25719584	51.40%
75000000	31892032	42.50%
100000000	38629896	38.60%
129309724	45204882	35.00%

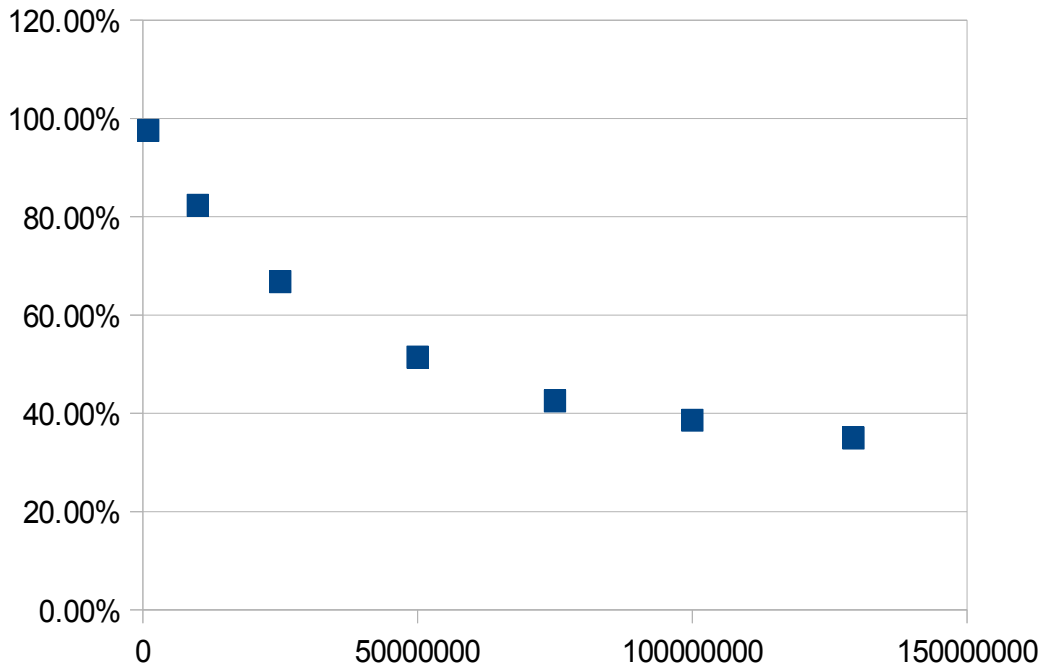


Figure 16: The relationship between the read numbers increase and the complexity percentage values. The Y axis represents the complexity values in percentage and the X axis represents the read numbers.

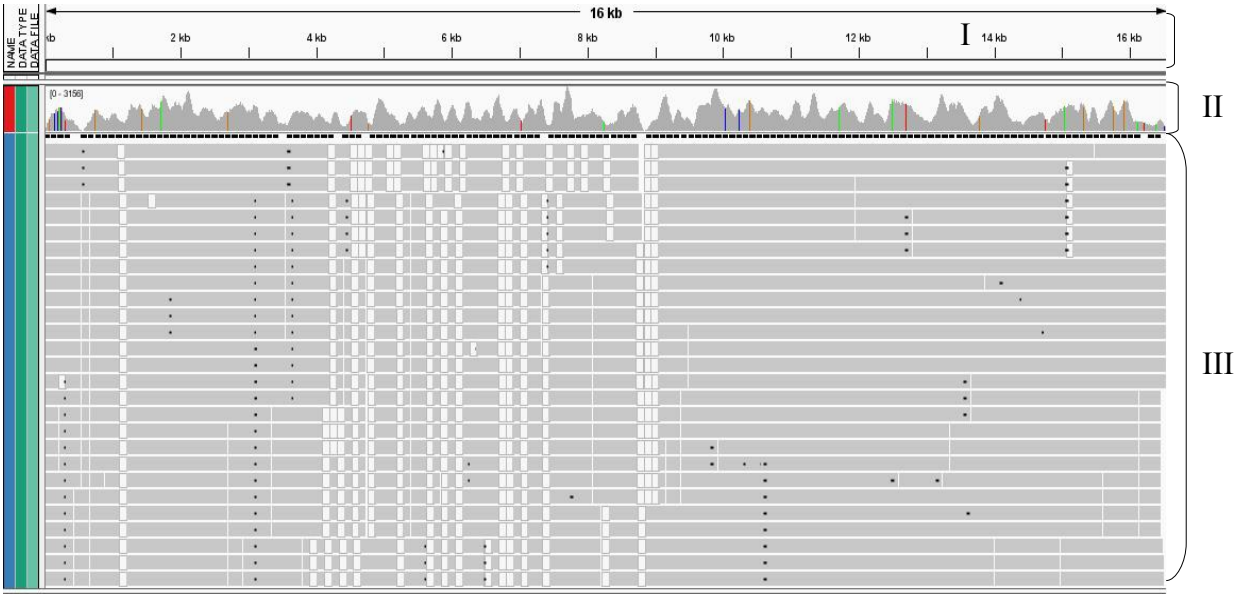
Continuing the in-depth sequencing of the DMG5 library, seven further single read sequencing data-sets were generated. Each of them yielded about 125 million reads that were mapped to the hg19 reference sequence using the mapping software BWA. The resulting binary alignment /map files (BAM) were merged together using the SAM tools and were used for further calculation. The human content was ~2% as estimated either separately with the previous runs in each data-set or collectively from the final calculation

from the merged BAM file. By merging the BAM together, the complexity decreased to 17.6% which affected the coverage of the human genome sequence in total and lowering it to 11%. On the other hand, out of 3.5 million unique human reads, about 31,625 unique ones were mapped to the mitochondrial genome, covering it 190.86 times (table 14, Fig. 17).

Table 14: The sequencing and mapping results of the merged BAM DMG5-I file. Mapping was done against the hg19 reference sequence.

total reads	~840000000
human reads	19935018
unique human reads	3519138
human %	2.37%
complexity %	17.60%
average coverage of unique human reads	0.11 X
total mito reads	209781
average coverage of total mito reads	1266.1 X
unique mito reads	31625
average coverage of unique mito reads	190.86 X

a)



b)

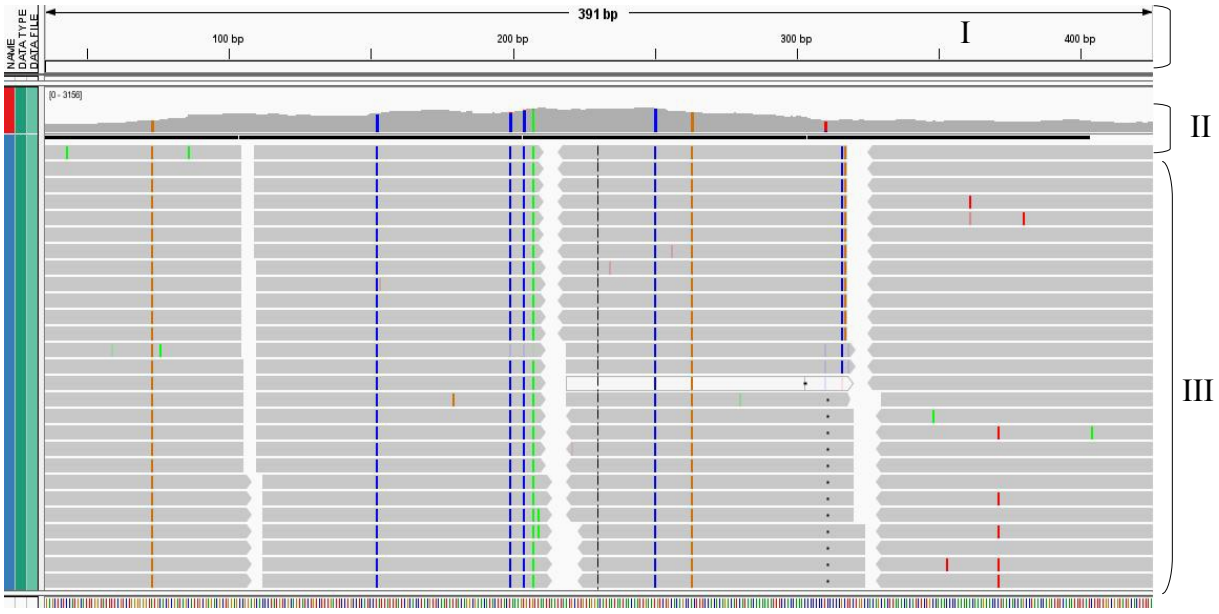


Figure 17: The IGV software snapshot of merged BAM file of DMG5-I data-set a) the coverage of reads mapped to the entire 16.5 kb mitochondrial human genome (accession number NC_012920). b) the coverage of a 391 bp segment of a mitochondrial HVII segment. Bar I is corresponding to the positions in the mitochondrial genome reference sequence. Bar II is corresponding to the contig coverage. Bar III is corresponding to the sequences that have been mapped to the mitochondrial genome reference sequence. The grey color bar is corresponding to the aligned reads. The blue colour corresponds to cytosine mutation in the aligned reads (C). the green colour corresponds to adenine mutation in the aligned reads (A). The brown colour corresponds to guanine mutation in the aligned reads (G). The red colour corresponds to thymine mutation in the aligned reads (T).

The high coverage of the mitochondrial genome facilitated a straightforward SNP determination. Using the SAM tools, the pileup file of the mapped reads to the mitochondrial genome was used for SNP determination. Using the VarScan software, the SNP determination resulted in 32 SNPs (table 15).

Table 15: The DMG5-I SNPs. The SNPs were identified using the VarScan software. Their positions, their corresponding reference base pairs, allele frequencies and population preponderance is given. [E= European, A= Asian, S= Subsaharan, C= Caucasian, DA= disease associated, PR= global private mutation, HSM= hot spot mutation, ND= not detected. The SNP allele and population frequencies are present according to available data deposited in the dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP). The used reference sequence is the mitochondrial human genome (accession number NC_012920).

position	73	152	199	204	207	250	263	750	1438	1719	2706	4529	4769	7028	8251	8860
reference	A	T	T	T	G	T	A	A	A	G	A	C	A	C	G	A
SNP	G	C	C	C	A	C	G	G	G	A	G	T	G	T	A	G
SNP allele frequency	1,00	1,00	ND	ND	ND	ND	ND	ND	0.8-.0.9	0.03 -0.05	0.5-1	ND	0.9-1	0.5-1	0.9 -1	ND
population with higher frequency	C	S	ND	ND	ND	ND	ND	ND	A	S	A, S	ND	A, S	A, S	S	ND
comment	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
predominance %	99.6	96.1	93.7	94.2	99.9	96.9	99.1	99.7	99.7	99.6	100	99.1	99.9	97.8	94.7	96.2

position	10034	10238	10398	11719	12501	12705	13780	14766	15043	15326	15758	15924	16129	16223	16391	16519
reference	T	T	A	G	A	C	A	C	G	A	A	A	G	C	G	T
SNP	C	C	G	A	G	T	G	T	A	G	G	G	A	T	A	C
SNP allele frequency	ND	0.9-1	0.07-1	ND	ND	0.05 0.99	ND	0.4-1.00	0.02-0.99	0.9-1	ND	0.018 -0.03	ND	ND	ND	ND
Population with higher frequency	ND	A	S	ND	ND	A, S	ND	A, S	S	A, S	ND	S	ND	ND	ND	ND
comment	-	-	DA		PR	-	-	-	DA	-	-	-	-	-	-	HSM/ DA
predominance %	96.7	95.9	98.8	96.3	99.9	97.9	95.9	96.5	94	99.8	98.9	98.8	99.8	98.9	96.9	93.3

Based on the defined SNPs (table 15), the mitochondrial haplogroup was determined using the phylotree-based web application, HaploGrep. It excluded two SNPs from its analysis. The first one was 12501G, which was recognized by the HaploGrep as a global private mutation (PR). The second one was 16519C and was known as a hot spot mutation (HSM). According to the HaploGrep report, the remaining 30 SNPs combination could be identified as a haplogroup I2 with high ranking and with a haplogroup quality score of 97.7%. Differentiation between the haplogroup I2 and its ancestor I2'3 relies

mainly on the presence of the diagnostic SNP 15758G (Fig. 18).

Figure 18: The VarScan report snapshot of the DMG5-I haplogroup determination. Defining SNPs are indicated. SNPs 152C and 207A are the defining SNPs of the haplgroup I2'3, while 15758G is the defining SNP of the haplogroup I2.

3.3.1 Defined SNPs and their reported association with diseases

The mitochondrial SNPs and their association with a number of human diseases was subject of various studies (Murakami et al., 2002; Otaegui et al., 2004; Ghezzi et al., 2005; Pyle et al., 2005; Kazuno et al., 2006; Bai et al., 2007; Liao et al., 2008; Pezzotti et al., 2009; Rollins et al., 2009; Juo et al., 2010; Ebner et al., 2011; Mosquera-Miguel et al., 2012; Sequeira et al., 2012). Using the DMG 5 identified SNPs, a search was done through the Mitomap database [i.e. genome database of polymorphisms and mutations of the human mitochondrial DNA (MITOMAP)] (http://www.mitomap.org/bin/view.pl/MITOMAP/PolymorphismsCoding) and the published records. Three of the identified SNPs were reported to show disease association (DA) (table 16). They are the SNPs A10398G, G15043A and T16519C. In the alignment, the SNPs were covered 2084, 1713 and 419 times, respectively. The predominance percentage of the three SNPs was 98.8, 94.0 and 93.3%, respectively.

Table 16: The identified SNPs in DMG5-I and their reported disease association

SNP	gene location	associated disease	reference
A10398G	NADH dehydrogenase subunit 3 gene (ND3)	Cancer, breast cancer	Kazuno et al., 2006; Pezzotti et al., 2009
		Metabolic syndrome	Juo et al., 2010
		Type-2 diabetes mellitus (T2DM)	Liao et al., 2008
		Parkinson's disease	Otaegui et al., 2004; Ghezzi et al., 2005; Pyle et al., 2005; Kazuno et al., 2006
		Alzheimer disease	Kazuno et al., 2006
		Bipolar disorder (BD)	Kazuno et al., 2006; Rollins et al., 2009

G15043A	Cytochrome b gene (Cytb)	Major depressive disorder (MDD)	Rollins et al., 2009
T16519C	Control region of the mitochondrial genome (D-loop)	Malignant Melanoma	Ebner et al., 2011
		Breast cancer	Bai et al., 2007
		Type-2 diabetes mellitus (T2DM)	Liao et al., 2008
		Metabolic disorders	Murakami et al., 2002
		Schizophernia (SZ)	Mosquera-Miguel et al., 2012; Sequeira et al., 2012
		Bipolar disorder (BD)	Sequeira et al., 2012

3.4 Metagenomic analyses

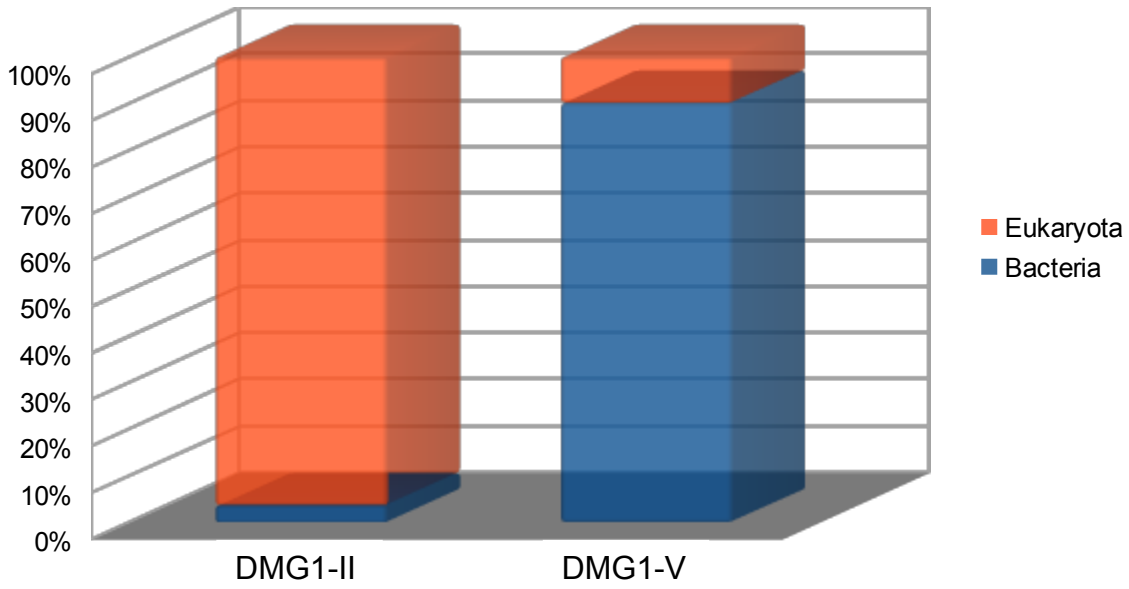
3.4.1 Metagenomes in the Egyptian mummy specimens

The metagenomic signature of Egyptian mummy tissue may originate from various sources. The DNA can be endogenous, introduced by the handling through the embalming process and from the mummification procedure itself. Of course it may also stem from environmental influences i.e. introduced at the burial or storage times. Thus, the bacterial and herbal taxa were the main target in our metagenomic analysis.

3.4.1.1 The effect of storage condition

Two different samples were taken from the same mummy within a time period of 1.5-2.0 years. DMG1-II and DMG2-II were taken from the mummies DMG 1 and 2 prior to the other two samples DMG1-V and DMG2-III. Since we knew about a bacterial bloom that had occurred in the Egyptian mummy repository within that time period, we wanted to examine its effect on the resultant metagenomic pattern. One million reads from the two pairs of data-sets were compared [the DMG1-II versus DMG1-V] and [DMG2-II versus DMG2-III]. The alignment of the mummy data-sets against the nr/nt database was done using the BLASTn algorithm. The BLAST results were analyzed using the MEGAN software. Since the Viruses and Archaea percentage were quite low, the comparison included mainly the Eukaryota and the Bacteria percentages. A comparison was done using the MEGAN software. A reversal in the Eukaryota:Bacteria DNA relative abundance was noted in both mummy sample pairs (Fig 19 a, b).

a)



b)

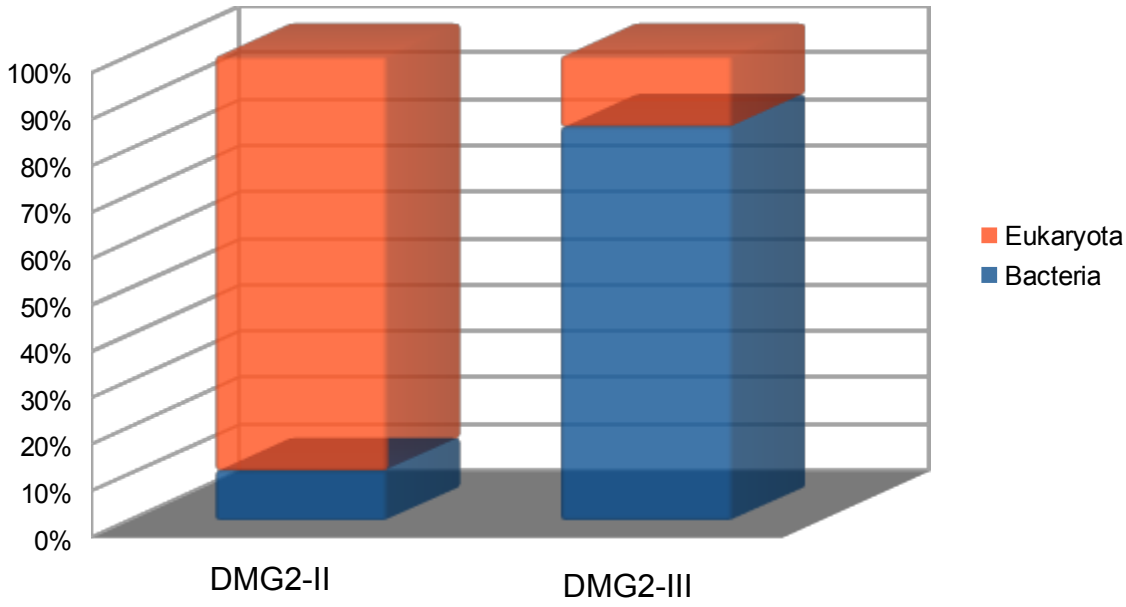


Figure 19: Relative abundance of eukaryotic and bacterial DNA in the samples (a) DMG1 and (b) DMG2 using one million reads. The presence of Eukaryota percentage is shown in red colour and the Bacteria is shown in blue colour.

3.4.1.2 Effect of the used word size on the BLASTn results

The effect of the used word size on the BLASTn alignment results and subsequently the taxa representation within the data-set was tested as follows. One million reads from the large scale sequencing data-set of mummy DMG5-I was aligned against the nr/nt database using the BLASTn algorithm. Different word sizes (30 and 42) were used for the BLASTn alignment and the result files were defined as DMG5-Ia and

DMG5-Ib, respectively. The result files were analyzed using the strict thresholds set of MEGAN LCA parameters. The different taxa representations were compared and analyzed. The metagenomic patterns of the two data-sets showed a slight difference in the percentage of the taxa representation. The percentage of unidentified sequences (no hits i.e. reads with no determined hits by the BLASTn) rose from 80.97% in DMG5-Ia to 88.24% in DMG5-Ib. The "not- assigned reads" are hits that could not be defined by MEGAN. After the out-filtration of the reads of "no hits", "not assigned" and "low complexity", the percentage of bacterial hits were 84.14 % and 76.22% in DMG5-Ia and DMG5-Ib, respectively. On the other hand, the Eukaryota hit percentage in DMG5-Ib slightly increased in comparison to the DMG5-Ia which reflects the increase in specificity. The values for Viruses and Archaea hits were negligible and did not change in the two data-sets (Fig. 20 a, b).

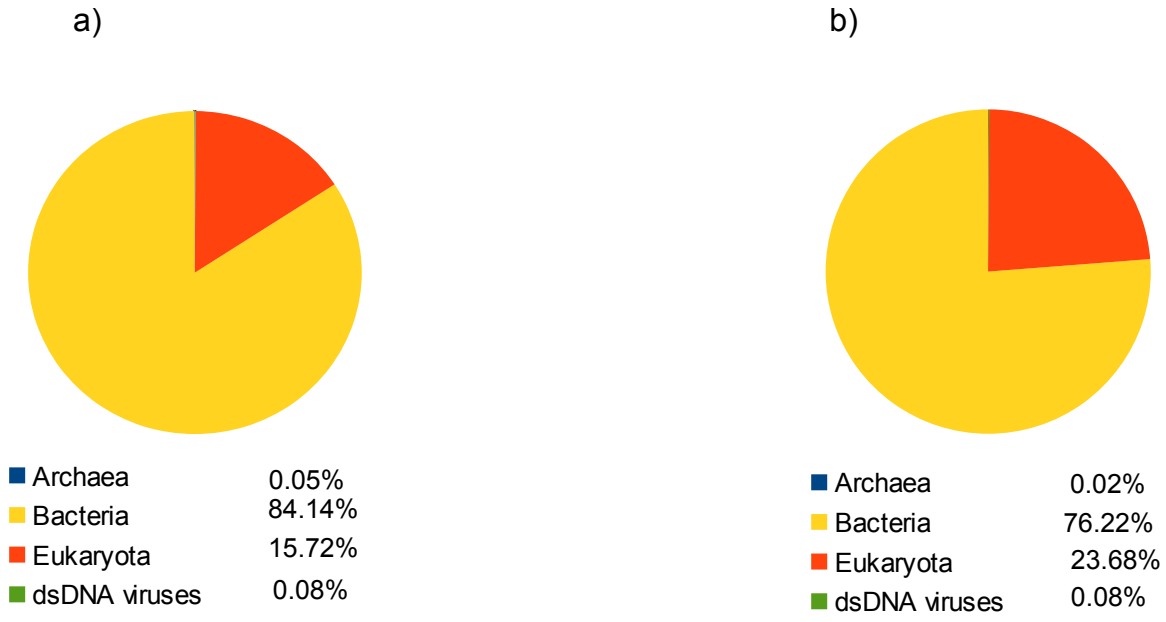


Figure 20: Comparison of two data-sets DMG5-Ia and DMG5-Ib after out-filtration of "no hits", "not assigned" and "low complexity" reads. a) DMG5-Ia and b) DMG5-Ib including relative percentage of Eukaryota, Bacteria, archaea, and viruses. Archaea= dark blue, Bacteria= yellow, Eukaryota= orange, and viruses= dark green.

The comparative analysis was extended to test the effect of the tested word size parameter on the bacterial taxa representation within the DMG 5 data-sets. The firstly dominant Firmicutes phylum showed an increase to 70.12% in the DMG5-Ia in comparison to 57.77% in DMG5-Ib, mostly on the expense of the Proteobacteria phylum percentage (Fig. 21 a, b). In addition to Firmicutes, there were other taxa in high

proportion i.e. the Proteobacteria. The Proteobacteria was the only taxon that showed an increase to 39.37% in DMG5-Ib in comparison to 26.07% in DMG5-Ia. The Actinobacteria were in low proportion and slightly similar in both data-sets, DMG5-Ia (2.33%) and DMG5-Ib (2.06%) (Fig. 21). The other bacterial taxa were present in low percentages. A number of the detected bacteria were documented to live in extreme hard environment like Firmicutes, Deinococcus-Thermus, and Aquificae. To assess this finding, search was done for those bacterial taxa in DMGS-1000 and DMGS-2000 and Iceman data-sets as further examples of warm and cold climate samples. Deinococcus-Thermus was represented in DMGS-1000 and DMGS-2000 plus the Egyptian mummies. Aquificae were represented only in the Egyptian mummies, while Firmicuties were represented in the Alpine Iceman, DMGS-1000 and DMGS-2000 and in high percentage in Egyptian mummy data-sets.

In-depth analysis at the phylum level showed minor differences in the taxa representations within the two data-sets. For the phylum Viridiplantae (Fig. 22), it is to be noted that taxa with only few reads did not reproducibly appear within the two data-sets. This resulted in some taxa that were uniquely represented in the DMG5-Ia, like the families Papilionoideae (*Fabeae, Lotus japonicus*, *Medicago truncatula),* Malpighiales *(Ricinus communis*, *Populus trichocarpa)*, Brassiceae, Vitis *(Vitis vinifera)*, Triticeae *(Triticum monococcum)*, Andropogoneae *(Zea mays)* and *Oryza sativa*. Vice versa, only one taxon, *Triticum urartu*, was uniquely represented in the DMG5-Ib (Fig. 22, 23). This highlights that elevating the BLASTn word size results in a higher specificity. There were three taxa from the family Poaceae (Fig. 25, 26), that were dominant in both data-sets i.e. *Hordeum vulgare*, *Sorghum bicolor* and *Triticum aestivum* (Fig. 23). To assess this finding, search was done for those bacterial taxa in DMGS-1000 and DMGS-2000 and Iceman data-sets as examples of warm and cold climate samples. None of these three taxa were represented in the cold or warm climate samples data-sets.

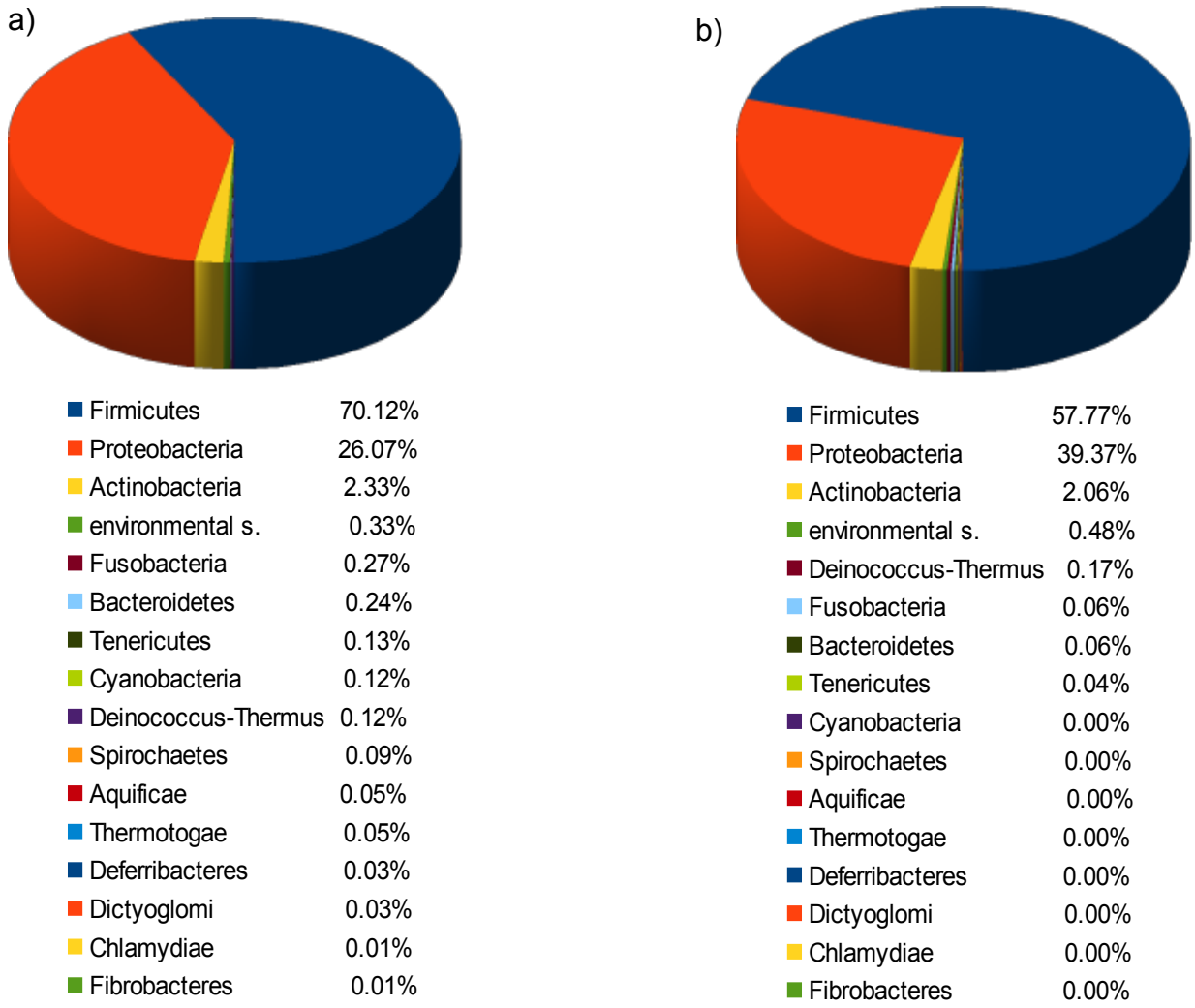


Figure 21: Representation of the various bacterial taxa in a) DMG5-Ia and b) DMG5-Ib data-sets.

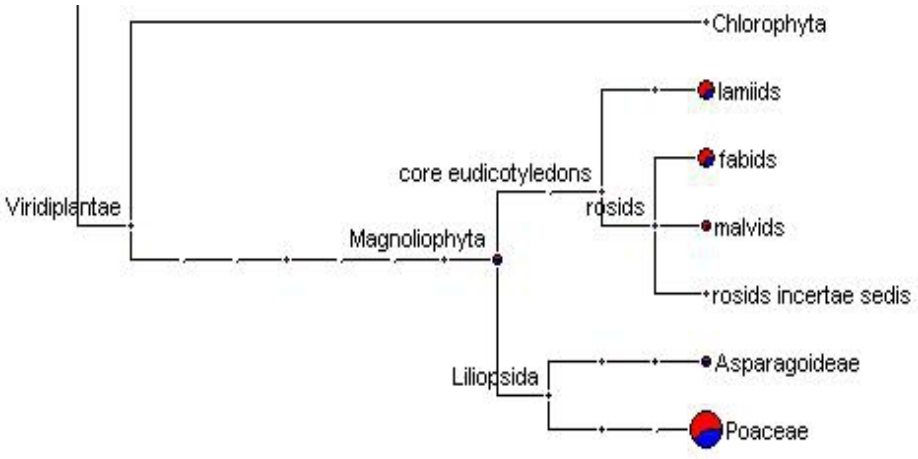


Figure 22: Detail of the MEGAN comparison of the Viridiplantae representation in the data-sets DMG5-Ia (red colour) and DMG5-Ib (blue colour). The taxa node diameter corresponds to the number of assigned reads.

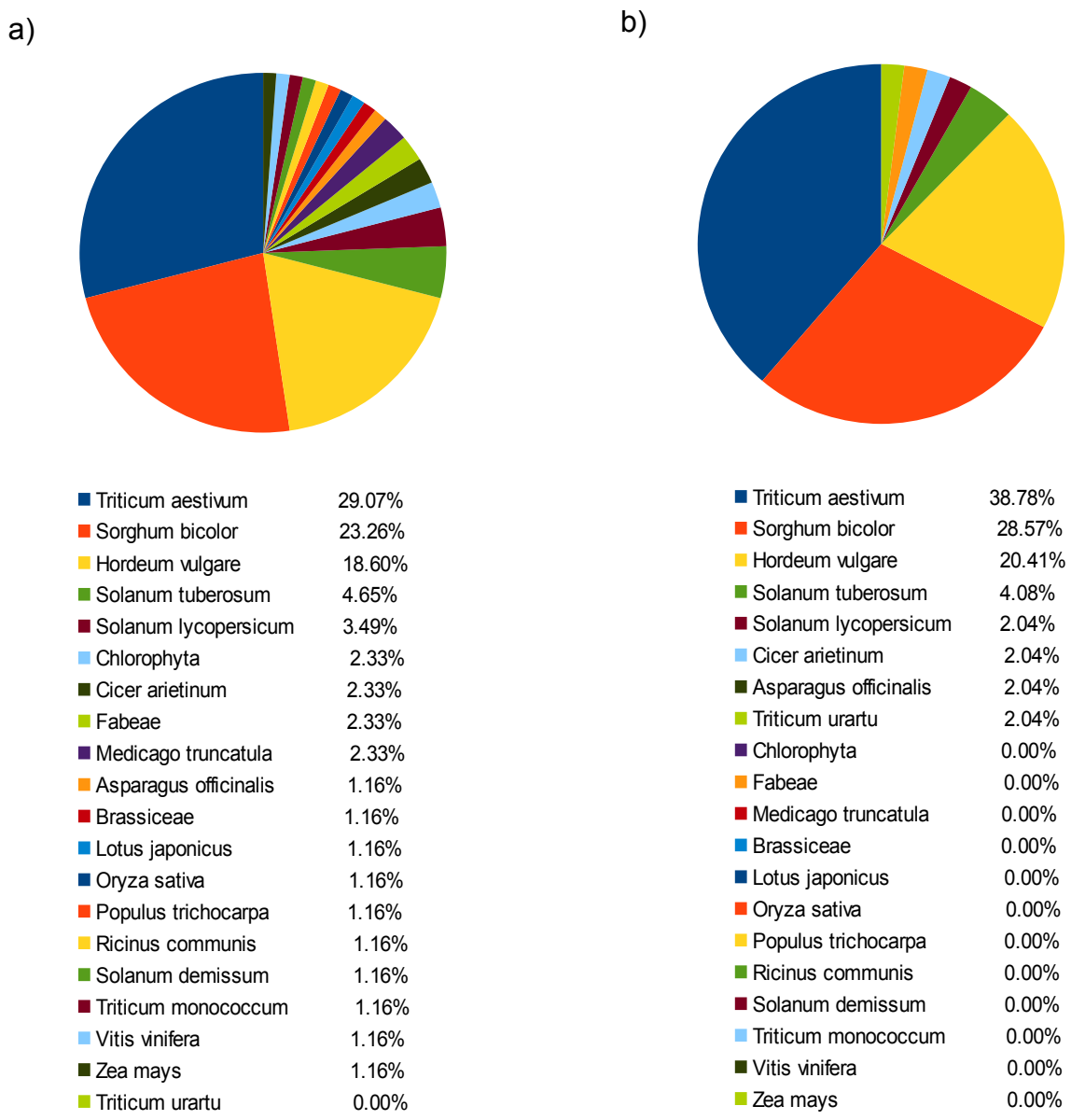


Figure 23: Comparison of the Viridiplantae taxa in two data-sets a) DMG 5-1 and b) DMG 5-2. The taxa were ordered by descending percentage values.

3.4.1.3 Metagenomics pattern in other mummies

To learn more about the metagenomic composition in mummies, two pairs of pooled mummy DNAs were used to generate new NGS data-sets (DMG M1, M2). They were mapped to the hg19 reference sequence and all the human reads were filtered out. The remaining reads were aligned against (nr/nt) using BLASTn algorithm. From about one million reads out of DMG M1 data-set, about 405,388 reads were assigned using the strict thresholds set of MEGAN LCA parameters.

Nearly 97% of the assigned reads were eukaryotic taxa while 2.85% were assigned to bacterial taxa. A negligible number of reads (≤ 0.01%) were assigned to

Viruses or environmental samples. The environmental samples can be referred to any material collected from an environmental source. Within the bacterial taxa, the Firmicutes were again the most dominant phylum (48.03%), followed by the Proteobacteria (29.31%) and the Actinobacteridae (13.19%). Around 90% of the Firmicutes were Clostridia while the rest were Bacilli. Approximately 99% of the Eukaryota were mainly Gnathostomata with a lower presence of Alveolata, fungi, and Viridiplantae (respectively 0.34%, 0.74%, and 0.22%).

The superphylum Alveolata encompassed Apicomplexa which were represented by two pathogens. Eleven reads were specific for the *Plasmodium* genus specifically subgenus *Vinckeia* (*Plasmodium subgenus,* mainly in mammalian) and 1270 reads were specific for *Toxoplasma gondii* (Fig. 24).

MEGAN LCA parameters have an effect on the resultant taxa shown in the MEGAN output. To assess our findings using the strict LCA thresholds, a number of MEGAN LCA parameter thresholds were released and changed. The minimum bit score threshold of 35 was adopted, the win score set to the default and the minimum threshold of the required reads to support the taxon definition was changed to 1. Changing the MEGAN LCA parameters, the assigned reads to the *Toxoplasma gondii* remained the same, while those assigned to the *Plasmodium* genus became 13 reads. A specific one was assigned to *Plasmodium falciparum* with a score of 77 and another one to *Plasmodium knowlesi* (primate malaria) with a hit score of 68. The rest were assigned to the subgenus *Vinckeia.* Moreover, the results outcome slightly varied within the ensuing time period due to the various NCBI taxonomy updates of the MEGAN software.

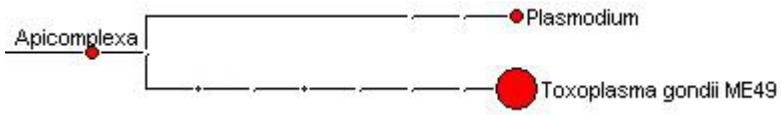
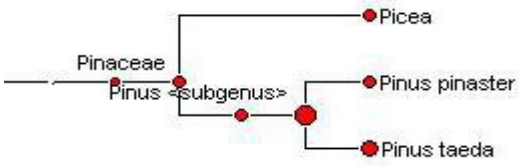


Figure 24: Detail of the MEGAN output of the Apicomplexa phylum within data-set DMG M1. The taxa node diameter corresponds to the number of assigned reads.

Nearly half of the Viridiplantae assigned reads (54.27%) were ascribed to *Nicotiana tabacum*. About 12% of the Viridiplantae assigned reads were attributed to the

family Triticum and 11.6% to the family Pinaceae. The Pinaceae were especially represented by 69 reads specific to the species Pinus (Fig. 25). By releasing the thresholds of the MEGAN parameters to minimum support to 5, the minimum bit score threshold to 35 and the win score to 50, most of the representations remained the same. By decreasing the minimum support threshold to 1 and the win score to the default, most of the representations were the same, except increasing their variability at the species level.

a)



b)

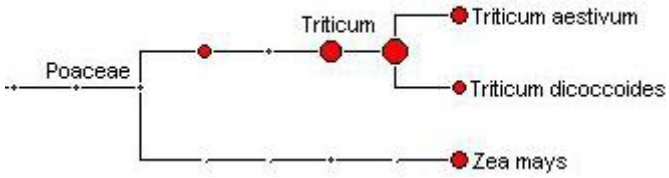


Figure 25: Details of the MEGAN output highlighting the most frequent Viridiplantaea taxa harboured in the DMG M1 dat-aset. a) Pinaceae family and b) Poaceae family. The taxa node diameter corresponds to the number of assigned reads.

The same analytical approach was applied to the data-set DMG M2. Around 445,557 reads from a total one million reads were assigned following to alignment against nr/nt. Using the most strict MEGAN LCA parameters, about 79% of the assigned reads had a hit in eukaryotic taxa, 20.7% to Bacteria and less than 0.01% to viruses or Archaea. Ninety-six percent of the eukaryotic were mainly Metazoa and especially Gnathostomata with a lower presence of Alveolata, Fungi, and Viridiplantae (respectively 0.26%, 0.035%, and 0.84%). From those assigned to the Apicomplexa, 44 reads were assigned to the genus Plasmodium. Twenty-six reads of them were hits specific to *Plasmodium falciparum* with a score of >75. Almost 94.7% of the reads assigned to Apicomplexa were identified as *Toxoplasma gondii* (843 reads from a total of 890 reads were assigned to Apicomplexa) (Fig. 26). By releasing the MEGAN LCA parameters, the assigned reads to the Apicomplexa were slightly increased to 901 reads and those assigned to the *Toxoplasma gondii* increased to 894. Those assigned to *Plasmodium* genus became 49

reads in total among which 29 assigned to *Plasmodium falciparum* and 16 were assigned to the subgenus *Vinckeia.*

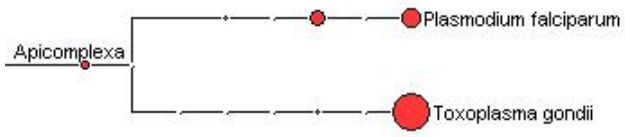


Figure 26: Details of the MEGAN output of the Apicomplexa phylum within the data-set DMG M2. The node diameter corresponds to the number of assigned reads.

Within the assigned reads to the bacterial taxa, the Proteobacteria was the most dominant phylum (64.56%) followed by the Actinobacteria (28.38%) and the Firmicutes (5.56%). The rest of bacterial taxa were in small percentages. Within the assigned reads to Actinobacteria, 52 reads were assigned to the Mycobacteria and specifically 11 to the *Mycobacterium avium*, 5 of which were specific with a bit score of ≥ 70.

The assigned reads to the Viridiplantae in DMG M2 showed a large spectrum which was consistent with that of the DMG M1 but with a broader spectrum. Around 12.3% of the assigned reads to the Viridiplantae was *Nicotiana tabacum* taxa. The Pinaceae family was represented by 3% of the assigned reads to the Viridiplantae (Fig 27) and the Triticeae by 5%.

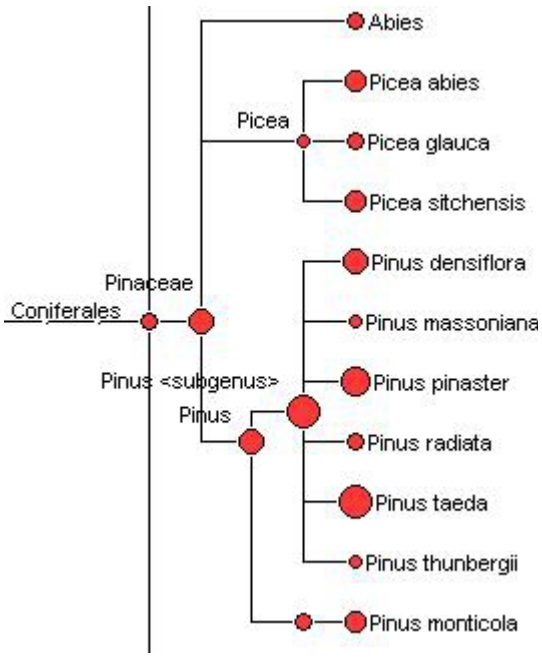


Figure 27: Details of the MEGAN output showing the Viridiplantaea clade in the DMG M2 data-set. The taxa node diameter corresponds to the number of assigned reads.

3.4.2 Metagenomic comparison between warm and cold climate samples

A metagenomic comparison was done using a reads subset from two groups, the warm and cold climate samples. The warm climate samples included the Egyptian mummy data-sets (DMG1-V, DMG2-III, DMG4-I, DMG5-Ib) as well as the two other data-sets of DMGS-1000 and DMGS-2000 originating from Bolivian lowland skeletons (Table 4). The cold climate group included the previously published aDNA NGS data-sets of the Saqqaq (Rasmussen et al., 2010), the Denisova hominid (Reich et al., 2010) and the Alpine Iceman (Keller et al., 2012) (Table 6). A comparison and in-depth analysis was done between the two groups with special regard to the Bacteria taxa (Fig. 28, 29) and Viridiplantae taxa using MEGAN software (Fig. 30).

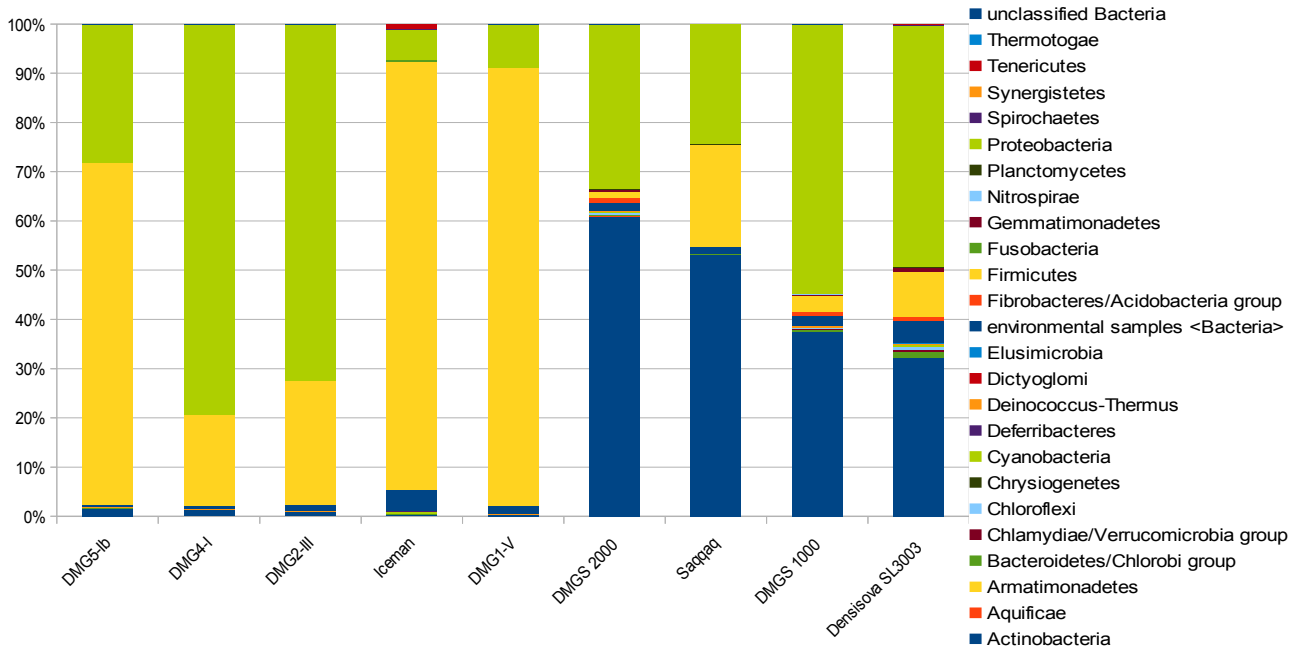


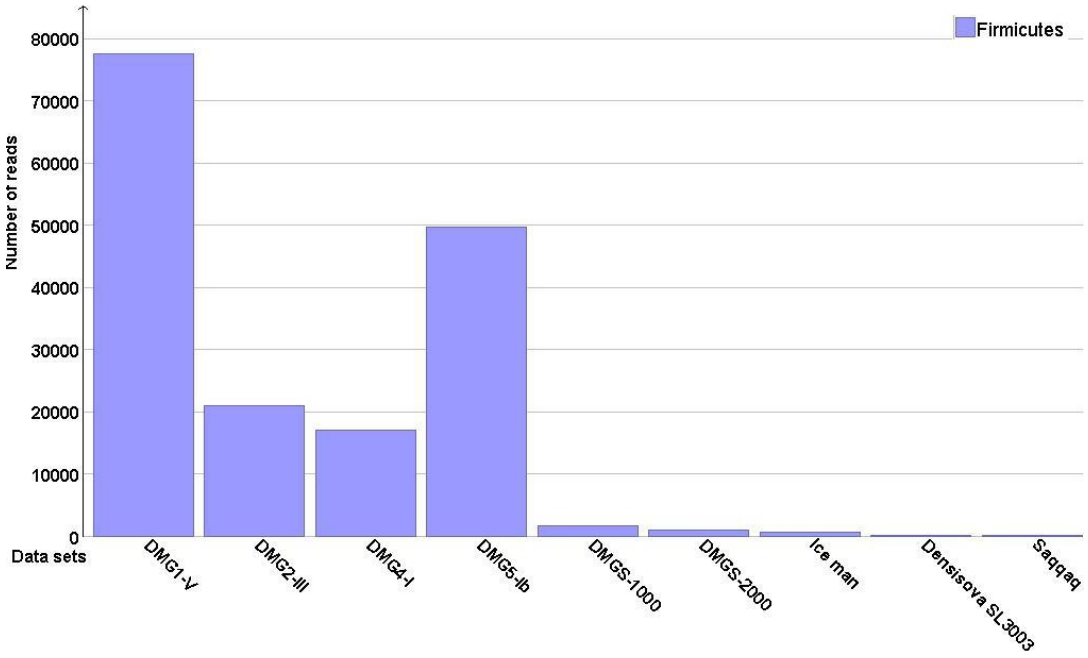
Fig. 28: Characterisation of Bacteria taxa (in percentage) in the NGS data-sets of the warm and cold climate samples. The warm climate group included the Egyptian mummies and Bolivian lowland samples DMGS-1000 and DMGS-2000. The cold climate samples included the previously published data-sets for the Saqqaq paleo-eskimo, the Denisova hominid and the Alpine Iceman (Rasmussen et al. 2010; Reich et al. 2010; Keller et al. 2012).

The relative comparison of bacterial taxa showed that the Egyptian mummies and Iceman data-sets had a similar pattern according to the Actinobacteria percentage (<10 %). While, within the Saqqaq, Denisova, DMGS-1000 and DMGS-2000 data-sets, the Actinobacteria percentage increased to ≥28 % (Fig. 28). The Proteobacteria showed

variable presence through all the samples; they ranged from low in the Iceman mummy (5.94 %) to high in the Egyptian mummy sample 4 (78.44 %) (Fig. 28).

There was an increased percentage of Firmicutes within the mummy group, with dominant representation in the Egyptian mummy data-sets (Fig. 29a). By performing an in-depth analysis of the Firmicutes, the clostridia family was highly represented in the Egyptian mummies data-sets in comparison to the other data-sets of either warm or cold climate samples (Fig. 29 b).

a)



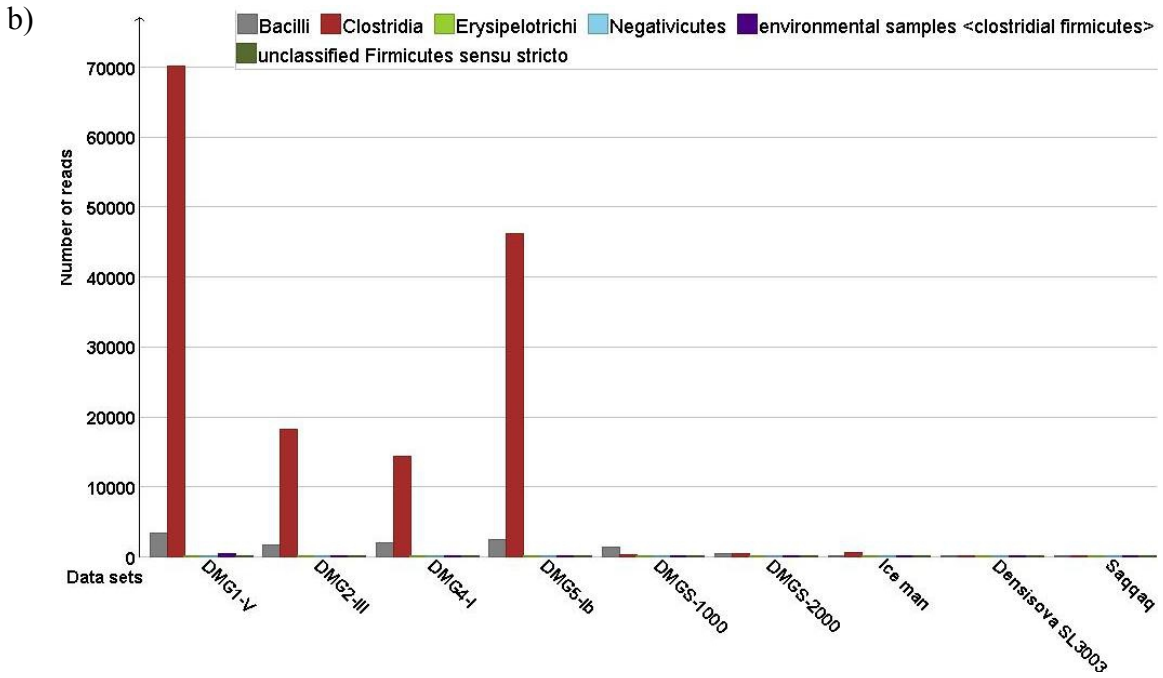
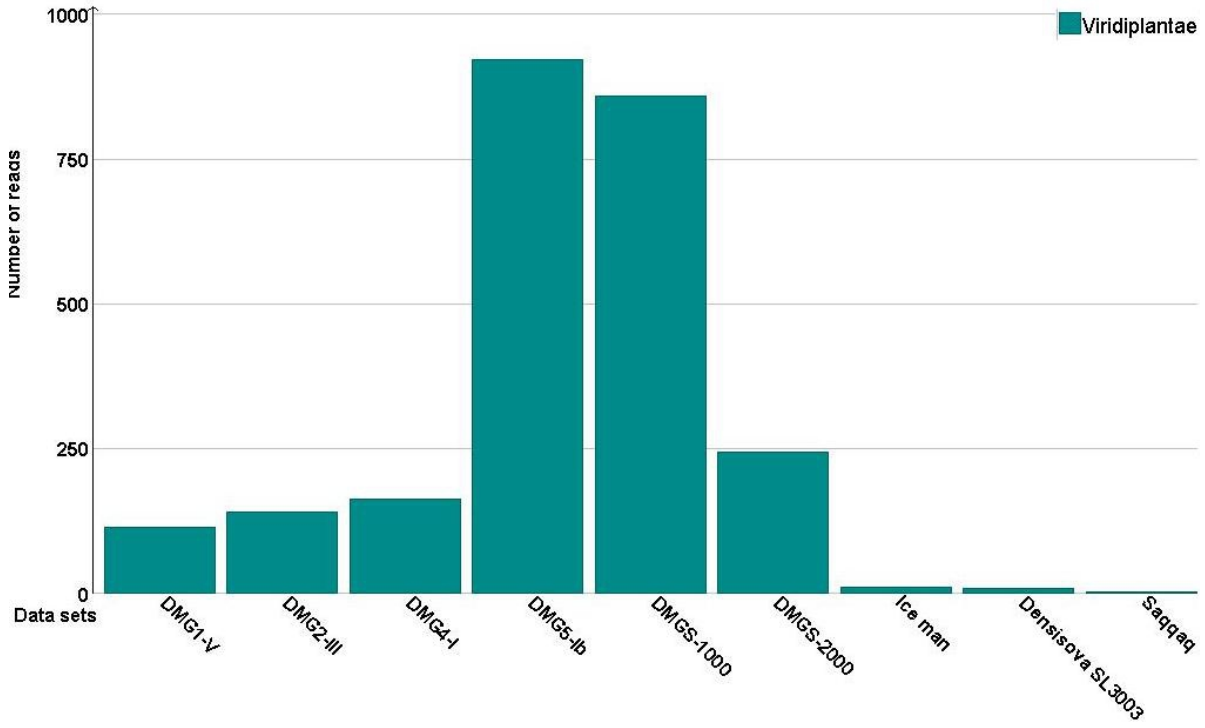


Figure 29: Comparison of the absolute read numbers of the Firmicutes family representation in the NGS data-sets of warm and cold climate groups. a) Total Firmicutes read numbers, Grey= Firmicutes. b) The detailed characterization of the phylum Firmicutes, Bacilli= gray, Clostridia= dark red, Erysipelotrici= green, Negativicutes= light blue, enviromental samples= dark blue, and unclassified Firmicutes= dark green. The warm climate samples included the Egyptian mummies and Bolivian lowland samples DMGS-1000 and DMGS-2000. The cold climate samples included the previously published cold climate samples for the Saqqaq paleo-eskimo, the Denisova hominid and the Alpine Iceman (Rasmussen et al. 2010; Reich et al. 2010; Keller et al. 2012).

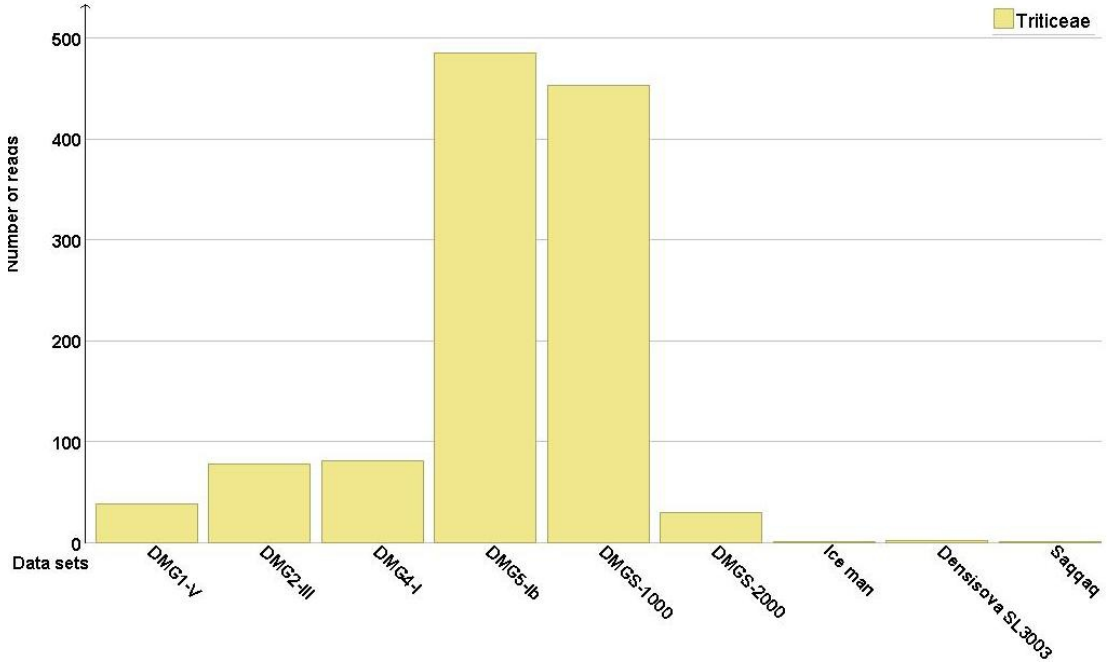
Since the ancient Egyptians used a number of plants and herbs in the mummification procedure, a search through the Viridiplantae taxa was done. A comparison of the Viridiplantae phylum representations was generated for the NGS data-sets of the warm and cold climate samples. The Egyptian mummies and the warm climate samples showed a higher Viridiplantae content than those observed in the cold climate ones (Fig. 30 a). The relative amount of Viridiplantea reads number in the warm climate samples was about 100 times more abundant than those in the cold climate samples. However, the total reads number in the cold climate group was higher in comparison to the warm climate group. The Triticeae family contents showed a similar pattern in the warm climate samples (Fig. 30 b), while some taxa were variably represented in the warm climate samples like the Pinaceae family. The read numbers of the Pinaceae family in data-sets DMG1-V, 2-III, 4-I, 5-Ib and DMGS-1000 was 3.5-24 times higher than the one established for the cold climate samples. However, in DMGS-2000, it was in an equal

range to those in the cold climate samples (Fig 30 c). Other taxa like the *Ricinus communis* and *Populus trichocarpa, Lotus japonicus, Cucumis sativus* and Fabaceae family were exclusively represented in the Egyptian mummies data sets, without any occurrence in the cold climate data-sets.

a)



b)



c)

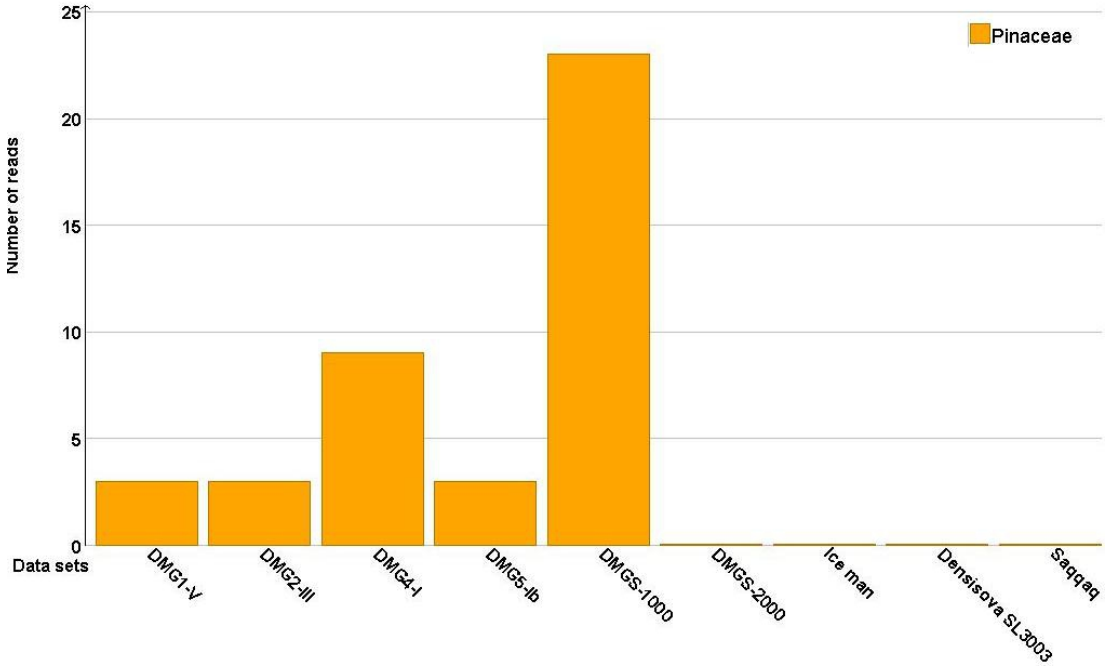


Figure 30: Comparison of the absolute read numbers of Viridiplantae taxa representation in the NGS data-sets of warm and cold climate groups. a) The total Viridiplantae representation, b) the family Triticeae representation and c) the family Pinaceae representation. The warm climate samples included the Egyptian mummies and Bolivian lowland samples DMGS-1000 and DMGS-2000. The cold climate samples included the previously published cold climate samples for the Saqqaq paleo-eskimo, the Denisova hominid and the Alpine Iceman (Rasmussen et al. 2010; Reich et al. 2010; Keller et al. 2012).

4 Discussion

Egyptian mummies are considered unique archeological samples, and they often show a high degree of preservation. This is highlighted in a series of anatomical, biochemical and molecular studies (Lewin, 1968, Rabino-Massa and Chiarelli, 1972; Barraco, 1975; Barraco et al., 1977; lynnerup, 2007; Jeziorska, 2008). This preservation status is a result of the process of mummification which includes the use of resin, oil and herbs. Studies on Egyptian mummies were confined by the technical limitations of the traditional PCR technology. In this study, we examine Egyptian mummies for the first time using the NGS technology with the aim to shed light upon the metagenomic paradigm and its DNA retrievability. Up till now, it is not known whether Egyptian mummies can act as a prolific template for the NGS technology. Likewise, it is not known whether Egyptian mummies have a metagenomic signature that can help in differentiating them from other types of human remains.

4.1 The use of the NGS technology on Egyptian mummy tissue

Ancient Egyptians used sophisticated mixtures of natural substances to embalm human bodies (Wisseman , 2001). The ancient Book II of “Herodotus's Histories” provides the most detailed account of the procedures performed by the Egyptian priests (fifth century BC). This information coupled with the records of Diodorus provided the basis for the study of mummification techniques (Wisseman, 2001; Ikram, 2003). Nevertheless, the remaining information about the embalming protocols is quite scarce. According to the ancient Egyptian beliefs, the purpose of the mummification was the preservation of the body after death. The process varied and diversified during the different periods of the Egyptian history (Wisseman 2001; Ikram 2003; David 2008; Jeziorska 2008). The preservation of the Egyptian mummies was the subject of many studies and even at the molecular level (Rabino-Massa and Chiarelli, 1972; Barraco et al., 1977, Wisseman, 2001; David, 2008; Jeziorska, 2008; Metcalfe and Freemont, 2012).

4.1.1 Preservation of Egyptian mummy tissue

The preservation status of the mummies has been presumed to depend on many factors including the used mummification protocol, the time that elapsed between the death and the start of mummification process, the burial location and condition till the excavation was done, and the storage of the samples after the excavation. Thus this makes each mummy a unique specimen (Zink and Nerlich, 2003; 2005; Jeziorska, 2008).

Mummies examined in this study were randomly chosen i.e. they did not exhibit a good preservation status and were not selected with regard to certain pathologies. However, the study samples, though were small, they provided ample genomic information. In the case of our study samples, inappropriate storage conditions in the Institute of Pre- and Protohistory and Medieval Archaeology at the Eberhard-Karls University of Tübingen, induced a bacterial bloom which had a notable effect on the NGS output. The effect of inadequate storage conditions is exemplified at its best when comparing the data-sets of DMG1-II and -V, and those of DMG2-II and -III. The 1.5-2.0 year time interval between taking the DMG1-II and DMG2-II samples on one hand and those of DMG1-V and DMG2-III had its consequences on the DNA composition of the generated data-sets (Fig. 19). In one million reads subset, the high Euokaryota to the bacterial content ratios within the DMG1-II and DMG2-II were replaced by reversed values in the case of DMG1-V and DMG2-III. The repletion of bacterial content in the DMG1-V and DMG2-III samples points to a bacterial bloom that occurred due to inadequate storage conditions. The overrepresented bacterial content within the test sample of one million reads largely reduced the number of human reads, in general

4.1.2 Technical challenges

The PCR technology was the dominant method to retrieve genetic information from the aDNA samples including tissues obtained from Egyptian mummies (Nerlich et al., 1997; 2008; Zink et al., 2000; 2001; 2003; 2006; Donoghue et al., 2010; Hawass et al.; 2010; 2012; Woide et al., 2010; Hekkala et al., 2011; Kurushima et al., 2012). Authentic aDNA has now been recovered from human and animal remains, thus ending a long term debate about the survivability of DNA within the Egyptian mummies (Nerlich et al., 1997; 2008; Zink et al., 2000; 2001; 2003; 2006; Donoghue et al., 2010; Hawass et al.; 2010; 2012; Woide et al., 2010; Hekkala et al., 2011; Kurushima et al., 2012). The molecular approach using the PCR technology answered a number of questions in the Egyptian history. Hawass et al., (2010) illustrated the familial relationships of the late 18th Dynasty and provided information on some infectious diseases and pathologies. Recently, another study collected data on the harem conspiracy that was described in the Judicial Papyrus of Turin (Hawass et al., 2012).

With the invention of NGS technology, the output of genetic studies has been accelerated. This also applied to the aDNA field, particularly that the NGS technology helped to overcome many of the shortcomings connected to aDNA research (Mardis,

2008; Kircher and Kelso, 2010; Shapiro and Hofreiter, 2010). Recently, several ancient genomes have been published providing more information about the human ancestors and migration patterns (Miller et al., 2008; Green et al., 2010; Krause et al., 2010; Rasmussen et al., 2010; Reich et al., 2010; Keller et al., 2012; Fu et al., 2013). The molecular Egyptology studies were hindered by the paucity and non-availability of the samples and the fear to cause damage to such precious archaeological remains (Wisseman, 2001). On the other hand, the use of the NGS technology on Egyptian mummy samples can generate a plethora of genetic information while causing minimal damage to the remains.

Through the current study, technical challenges were tackled in an attempt to maximize the experimental outputs, including the DNA retrievability. The presence of minute amounts of endogenous DNA is an obstacle in many aDNA studies (Willerslev and Cooper, 2005; Green et al., 2010). The current study shows that the used DNA concentration had a clear effect on the quality of the established NGS libraries. Another experimental line showed that the combined use of the phenol/chloroform extraction and the "MagNA Pure" purification eliminated inhibitory effects and subsequently improved the DNA quality. However, there were two accompanying adverse effects. First, the net DNA yield concentration decreased. Second, there was particular loss of DNA fragments smaller than 100 bp with preservation of larger fragments up to 400 bp.

Due to promising extraction quality and DNA yield of the tested mummies, four of them were used for further characterization using small scale NGS runs. The human content in the study samples ranged from 0.2-2.34%, reaching its highest in the mummy DMG 5. According to the evaluation of the NGS libraries by qPCR, the used DNA concentration increased the library efficiency and subsequently its complexity (table 8). In other words, the higher the DNA concentration, the lower the required qPCR cycles and the higher the complexity of the library. However, this was not the case with the South-American samples, where the DNA amount or concentration could not correlate with the qPCR results and the complexity (table 8). This highlights that the South-American samples as mineral soil embedded skeletons may had a different DNA taphonomy. The assessment of the used DNA extraction protocol on the NGS output showed that the purification using the MagNA pure system increased the probability of human DNA sequences. This may be due to the high cut-off of the magnetic separation. Since, most of bacterial genomes are smaller in size in comparison to the human genome size, the small fragments (<100 bp) are most probable from a bacterial source. Using the same DNA

extraction protocol and Illumina technology, different biopsies from the same mummy showed differences in the NGS run output (table 10). In some cases, the yield of human DNA reads almost doubled when the phenol/chloroform protocol was replaced by the new protocol that combines phenol/chloroform extraction with the purification by the MagNA Pure system. This can be due to sample differences or to the effect of the multistep NGS library preparation protocol on the final output. Generally, the refinement and the optimization of the DNA extraction and NGS library preparation protocols increased the library efficiency and the final output.

4.1.3 Bioinformatic challenges

In addition to the technical challenges, the bioinformatic analysis of the aDNA NGS data-sets is also challenging and was a concern in several publications (Briggs et al., 2007; Prüfer et al., 2010; Orlando et al., 2011; Schubert et al., 2012). Nevertheless, the NGS technology and its bioinformatics tools are continually evolving. It has been shown that the short aDNA fragments will potentially be more abundant in the NGS data sets. The bioinformatic analysis and the aDNA characteristics can result in the loss of a number of aDNA sequences. The misincorporations, which have been generated by the DNA damage over time as well as the NGS platform errors, may increase the mismatching opportunities (Schubert et al., 2012). One of the distinctive features of ancient NGS reads, the excess of C to T mismatches at 5' end and G to A at 3' end, a result of end repairing of overhanging aDNA ends (Briggs et al., 2007; Prüfer et al., 2010; Orlando et al., 2011; Schubert et al., 2012). Such excess of mismatches at the ends may result in rejection of the reads by the default mapping software parameters. The default parameters allow a maximum of three mismatches. Therefore, if the mismatches increase, as in the case of aDNA sequences, such reads will not be mapped to reference sequence. Subsequently, this will cause the loss of a number of precious aDNA reads, especially those of the extinct ones (Briggs et al., 2007; Prüfer et al., 2010; Kirsanow and Burger, 2012; Schubert et al., 2012).

Several studies suggested an ad hoc aligner for the aDNA which takes the postmortem misincorporations into consideration (Briggs et al., 2007; 2009; Prüfer et al., 2010; Rasmussen et al., 2010). Others suggested to choose carefully the alignment programs and to modify the default alignment parameters to suit the excess of misincorporations in the aDNA reads. Certainly, great care should be devoted to carry on with all those suggested solutions to not allow for spurious reads (Kirsanow and Burger,

2012; Schubert et al., 2012).

In the current study, two mapping software (BWA and Bowtie) were used and compared. The BWA was more suitable for the aDNA characteristics either with the default parameters or after changing the number of the permitted mismatches. The BWA is mainly designed for sequencing error rates below 2%. It shows more flexibility i.e. more mismatches can be allowed by the user and optionally it permits the mapping of insertions or deletions (indels) particularly in the read ends. A typical finding in Illumina data is the high error rate at the termini. For Illumina reads, BWA-short may optionally trim low-quality bases from the 3'-end before alignment and thus allows more reads to align (http://bio-bwa.sourceforge.net/) (Langmead and Salzberg, 2012; Schubert et al., 2012). Bowtie is recognized by a disadvantage in the paired-end mapping, as it is searching for the only ungapped and consonant reads. Thus, it may be the reason for the loss of a lot of reads after mapping (Langmead and Salzberg, 2012; Schubert et al., 2012). In the current study, we noticed a difference in the number of the human mapped reads after single-end and paired-end mapping using Bowtie and BWA. The paired-end mapped reads were increased using BWA (11,545 reads) in comparison to the same analysis performed by Bowtie (5,668 reads). Additionally, this was not compatible with the results of single reads mapping (table 9). In the current study, the BWA software using the default parameters improved the recovery of many aDNA reads in comparison with Bowtie. This is consistent with a recent study recommending the use of BWA software and modifying the default parameters to improve the mapping of the aDNA reads (Schubert et al., 2012).

4.1.4 Large scale sequencing and SNPs definition

The retrieval of authentic aDNA from modern human remains is problematic since it can be easily mixed with the contaminant modern human DNA. But the fulfillment of a few points helps in the authentication of the NGS results (Green et al., 2009; Krause et al., 2010). First, the guidelines for aDNA studies should be precisely followed (Richards et al., 1995; Gilbert at al., 2005b; Mitchell et al., 2005; Hansen et al., 2006; Roberts and Ingham, 2008; Heyn et al., 2010; Keller et al. 2012). Second, the employment of deep sequencing to increase the accuracy of the defined SNPs is recommended. Some authors noted that studying the overall degradation pattern such as fragment length, nucleotide misincorporation and the fragmentation pattern can be useful in determining the presence of aDNA (Green et al., 2009; Krause et al., 2010). However, other authors noted that the deamination based damage, as the most common DNA damage in the

aDNA, showed only a little impact on the final NGS results (Rasmussen et al., 2010; Keller et al., 2012).

Nielsen et al. (2011) stated that a high coverage of the defined SNPs (more than 20 fold) is required to increase its reliability. The high coverage of the defined SNPs may help in the determination of the authentic ones from any mixed sequences. In turn, this will ascertain the authenticity of the sequenced aDNA. According to the initial small-scale NGS sequencing runs, the mummy samples of DMG5-I exhibited a human content percentage of 2.34% of approximately seven million reads. About 150,000 reads were mapped to the hg19 human genome reference sequence including about 1900 reads assigned to the rCRS mitochondrial reference sequence. The number of the mapped reads to the mitochondrial genome covered almost the entire whole mitochondrial genome with an average coverage of 11.6-folds. The mitochondrial SNP pattern showed a unique signal with a high coverage and was used for preliminary haplogroup determination. The mitochondrial molecular signature of DMG 5 was distinct from those of the lab members. This was further confirmed by initiating and analysing further paired end run onto 7 Illumina lanes. Following the mapping about 20 million reads were mapped to the human genome reference hg19 from a total of about 840 million reads. By filtration of the duplicates about 3.5 million reads were uniquely mapped to the human reference genome, representing 11% of the human nuclear genome. This included 31625 reads uniquely mapped to the mitochondrial genome reference with an average coverage of 190.86 fold (table 14).

The ratio of the length of the nuclear versus the mitochondrial genomes [3 billion versus 16 thousand base pairs, respectively] is about 200,000:1 (Strachan and Read, 2007). However, according to the copy number, the ratio between them was estimated to be about 1:1000 in human cells but differing according to the tissue types (Green et al., 2008; Schwarz, 2009). Taking into account the genomes' lengths and copy number, the ratio of the nuclear to mitochondrial genomes in a diploid cell will theoretically range from 375-37.5, for mitochondrial number 1000-10000, respectively. By just calculating the copy number ratio for mitochondrial DNA (mtDNA) and nuclear DNA (nucDNA) in NGS studies, the ratio of the nuclear versus the mitochondrial genomes was found highly variable in studies evaluating aDNA from various specimens (Wandeler et al., 2003; Green et al., 2008; Schwarz et al., 2009; Burbano et al., 2010; Green et al., 2010; Rasmussen et al., 2010; Keller et al., 2012; Meyer et al., 2012). The ratio of the nuclear to mitochondrial sequences in the Tryolean Iceman genome was 1:152.6 (Keller et al., 2012). Further, it

was 1:190 in the extinct Paleo-Eskimo hominin study which was recovered from a permafrost-preserved hair sample (Rasmussen et al., 2010). After enrichment using the single-stranded library protocol, the archaic Denisovan hominin genome had an average coverage of 31-folds of the human genome with a copy number ratio of the nuclear to mitochondrial genomes of 1:132.2 (Meyer et al., 2012). In samples with low endogenous DNA content, like the Neanderthal (Green et al., 2010), its draft genome sequence has been constructed using three libraries from three different bone samples. After a targeted capture of specific regions of the nuclear genome and sequencing on Illumina GAII platform, total coverage of about 1.3-folds was recovered (Burbano et al., 2010) with a variable mitochondrial genome coverage of -35, -29, -72 folds using the Illumina sequencing technology (Green et al., 2010).

The copy number ratio in the mammoth samples varied, but the qPCR and the 454 data-sets showed some discordance. According to the qPCR results, nuclear to mitochondrial ratio ranged from 1: 245 to 1: 17,369, while those pertaining to the 454 results were from 1:317 to 1:3965 (Schwarz et al., 2009). Consequently, the authors concluded that the high mitochondrial content might not be a general indicator of the high nuclear DNA content and there is a wide range of nuclear/mitochondrial genome ratios in the ancient studies, sometimes reaching a 60-fold difference (Schwarz et al., 2009).

In the current study, we calculated the nuclear and mitochondrial average coverage using the unique mapped reads for comparison to other publications. The copy number ratios of the recovered nuclear to mitochondrial genomes in DMG5-1 was 1:1,727.3. Nevertheless, more research is required on further Egyptian mummies to assess this ratio. It must be noted that this NGS approach here, is the first study on the Egyptian mummies as a warm climate sample. Schwarz et al. (2009) proposed three explanations for the preferential preservation of the mtDNA in relation to the nuclear DNA through the diagenesis. First, the double membrane of the mitochondrion may protect the mtDNA from degradation (Schwarz et al., 2009). Second, the interaction between the chromosomal protein and the nuclear DNA may hinder the PCR reaction and may even cause its loss through the DNA extraction process (Binladen et al., 2006; Schwarz et al., 2009). The last one, which is less likely, proposes that both of the mtDNA and nuclear DNA may have been exposed to differential types of endonucleases (Schwarz et al., 2009). Another suggested element relates to the presence of the mitochondrial DNA pseudogenes in the nuclear DNA genome (numts) which might affect the mitochondrial DNA copy number estimation (Greenwood and Pääbo, 1999). In our opinion, many of

those hypotheses may not be applicable to NGS aDNA results. First, the nuclear membrane is also a double membrane. Second, there is no scientific study in the published records on the "chromosomal protein effect" on the PCR. Third, with the in-depth sequencing the effect of numts will be undermined, particularly in NGS studies with a low nuclear genome copy number. Also the numts, as scarse and short fragments, will not affect the overall genomes copy number representations. In the case of our results, the value of 1:1,727.3 is compatible to the estimated ratio in the human cells (Morin et al., 2007; Schwarz , 2009).

4.1.5 Haplogroup determination

The spatial dynamics of humans during their evolution have been always of major scientific interest. Within this context, the exploitation of genetic markers like short tandem repeats or SNPs, has been instrumental in phylogenetic studies. The SNPs variance is different in the nuclear and mitochondrial genomes and depends on its mutation rate. The mutation rate of the mitochondrial genome is higher in comparison to the nuclear one. Therefore, the SNP variance of the nuclear genome will be too slow to follow the recent human history. However, the mitochondrial genome characteristics provide a different paradigm to follow the human evolution, since it is maternally inherited and without recombination events. It affords a way to follow the maternal lineage through the human evolution (Horai et al., 1995; Jorde et al., 1998).

The total mapped reads of the DMG5-I to mitochondrial genome reference were calculated and it was 1,266.1 fold. The coverage reaches its maximum (3,156 fold) in the middle of the genome (around position 7700) and goes down to its minimum of 7 fold coverage close to the ends of mitochondrial genome reference. Additionally, there is a drop in coverage in-between the positions 8832 to 8874 bp (average coverage 55 fold). This high coverage facilitated the definition of 32 SNPs through the whole mitochondrial genome with p value threshold of 0.05, a minimum coverage of 8 reads, a minimum average base quality of 15 and a minimum allele variant frequency of 0.2 (table 15). Eleven SNPs were located over the control region, seven SNPs were in the HVII and four in the HVI. The remaining twenty-one SNPs were spread over the coding region of the mitochondrial human genome. The SNPs coverage was in range of 75 to1947 fold and the predominance percentage for the called SNPs was in the range of 90.8-100% (table 15).

The mitochondrial SNPs of the mummy DMG5 permitted an accurate haplogroup

determination. The HaploGrep, a website application based on the Phylotree database, was used for the haplogroup determination (http://haplogrep.uibk.ac.at/) (van Oven and Kayser, 2009; Kloss-Brandstaetter et al., 2010; Kloss-Brandstätter et al., 2011). Thirty SNPs from a total of 32 SNPs were accepted and used by the HaploGrep for the haplogroup identification. The combination of the used SNPs defined the I2 haplogroup with a quality score of 97.7%. The haplogroup I2 differs from its ancestor haplogroup based on the presence of the SNP 15758G (Behar et al., 2012) (http://www.phylotree.org/tree/subtree_N.htm). The estimated time of its origin is 6,387 ± 2,449 BP and its ancestor haplogroup is I2'3, a subclade of haplogroup I, which is in turn a subclade of the haplogroup N1e'1 (Behar et al., 2012; Fernandes et al., 2012). The most probable origin of the haplogroup I is in west Asia but due to a lack of data, the haplogroup I2 origin has not been defined yet (Derenko et al., 2007; Terreros et al., 2011; Behar et al., 2012; Fernandes et al., 2012).

The origin of the haplogroups and the migrations of humans across the world were subject to debate for decades. There are two main contrastive theories, the multiregional model and the most recent common ancestor or recent African model (Disotell, 2012). The multiregional model relies on the idea of hominin evolution over the continents to modern *Homo sapiens*. Differentially, the other model is based on the theory that all the humans originated from a common mitochondrial African ancestor, the mitochondrial Eve. And subsequently, the modern human replaced the hominin all over the world (Cann et al., 1987). As a result of this theory, the "out of Africa" theory was postulated according to the identification of different mitochondrial DNA sequences over all continents (Vigilant et al., 1991). There are two main scenarios for the "out of Africa" migration (Foley and Lahr, 1992). First, through the southern coastal route from eastern Africa via Bab el Mandeb to the Arabian Peninsula (Foley and Lahr, 1992; Rose , 2010). This theory was supported by several studies, some of them were genetic. They showed the presence of the haplogroups M and N in Ethiopia as well as the presence of eastern African mtDNA in the southern Arabian Peninsula. These results were supported by Y chromosome studies which showed the presence of Arabian haplogroups in the Ethiopian gene pool (Richards et al., 2003; Kivisild et al., 2004). The second scenario is the migration from Africa through northern Africa and the middle East towards Eurasia (Foley and Lahr, 1992; Beyin , 2011).

In the modern population, haplogroup I2 was found patchy distributed over the northern hemisphere i.e. Russia (Malyarchuk et al., 2010; Fernandes et al., 2012),

Ireland, England, Canada, Chechnya, Turkey, and the Czech Republic (Fernandes et al., 2012). Using 277 DNA samples, a study has been done on modern Egyptians from Alexandaria city (Saunier et al., 2009). According to the mitochondrial genomes analysis, 238 different mitochondrial haplogroups have been defined. This showed a high degree of mitochondrial DNA diversity, which was explained by the authors to be due to the distinctive bio-geographic position of Egypt and the sequential ruling by different populations (Saunier et al., 2009). According to this study, European haplogroups were the most dominant (67.5%), followed by African (20.6%) and Asian (11.9%). Interestingly, there was no occurrence of haplogroup I2. Even, its mitochondrial ancestor, haplogroup I, was at a low frequency of only 3.2% (Saunier et al., 2009) and with a low diversity (Fig. 31) (Fernandes et al., 2012). Haplogroup I is the most frequent clade within N1 in Europe (Fig. 32), has the highest frequency in the Gulf region and exhibits high diversity in Gulf, Anatolia, and southeast Europe. Therefore, it is believed to have originated in the Near east and/or Arabia (Fig. 31, 32) (Fernandes et al., 2012). Terreros et al., (2011) claimed that this haplogroup was affected by a genetic drift or a founder effect which could be concluded from many distinctive distributions of haplogroup I. First, its origin was postulated to be in the middle East. However, its prevalence is not as high in the modern populations of those regions (Terreros et al., 2011). Second, it is observed around the world at a low frequency of <3%, and it is scattered in Europe, Asia and Africa (Fig. 32) (Fernandes et al., 2012). It is relatively more represented in Europe in isolated areas like Krk (a Croatian island in the northern Adriatic Sea) (11.3%) and Lemko (one of several small ethnic sub-groups inhabiting the Carpathian Mountains in Slovakia) (11.3%) (Fig. 32) (Terreros et al., 2011). Third, differentially, it showed a high percentage in the Iranian population with a prevalence in the north (9.7%) in comparison to the south (1.7%). However, it is at a low frequency in other Arabian countries (1-2%) (Fig. 32). Curiously, haplogroup I was found at a high frequency in Danish Iron Age and Viking Age population samples with frequency (10-20%) (Melchior et al., 2008). This high frequency in the ancient Dane samples has been confirmed by another follow-up study (Melchior et al., 2010) with frequency (13%). This is inconsistent with the low frequency of haplogroup I in the Danish modern population (Fig. 32). This is explained by Melchior et al. (2010) that it was common in the ancient times in the Southern Scandinavian population then vanished by time due to a genetic drift or immigration events (Melchior et al., 2010).

The insufficient information about haplogroup I2 and its origin impeded any final conclusion about the origin of the defined haplotype. But we can claim that it can

resemble its ancestor origin and distribution over Eurasia. There were reported back and forth migration movements from Asia (Cruciani et al., 2002), Near east (Henn et al., 2012) and the Arabian Peninsula to and from Africa through the historical times (Fernandes et al., 2012). Those regions coincide with the proposed origins of haplogroup I. Since Egypt is considered a geographical corridor, it is likely to find haplogroup I and its subclades in the ancient Egyptian population. Nevertheless, this is the first complete mitochondrial genome recovered from an Egyptian mummy. Hence, more studies are needed to investigate the haplogroup I2 distribution in the ancient Egyptian populations.

By studying haplogroup I2 ancestors, the haplogroups N1 (including haplogroup I), N2 (including haplogroup W) and X are all very rare and present only in patchy distribution over Euroasia (Richards et al., 2000; Fernandes et al., 2012) (Fig. 31, 32). The estimated age of the N haplogroup is 55-65 ka which is very close to the estimated time of the L3 origin in eastern Africa, coinciding with the first migration wave out of Africa (60-70 ka ago). This supports that the origin of N haplogroup was the Arabian Peninsula subsequent to the migration out of Africa through the southern coastal route (Fernandes et al., 2012; Soares et al., 2012) (Fig. 33). Fernandes et al., (2012) claimed based on the presence of the haplogroups I, N1a and N1f in eastern Africa, that there was a back migration from the Arabian Peninsula to Africa 15-40 ka ago. The estimated age of lineages N1a2, N1f, and possibly also N1c, N1d, and N1e is around 15–55 ka ago and they are considered the most ancient non-African mtDNA lineages (Fernandes et al., 2012). The haplogroup N1e (the closest ancestor to the haplogroup I) is present mainly in the Arabian Peninsula and split from the I haplogroup about 30 ka ago (Fernandes et al., 2012). Thus, the estimated time of haplogroup I origin is 20,857 ± 3,594 before present (Fig. 33) (Behar et al., 2012)

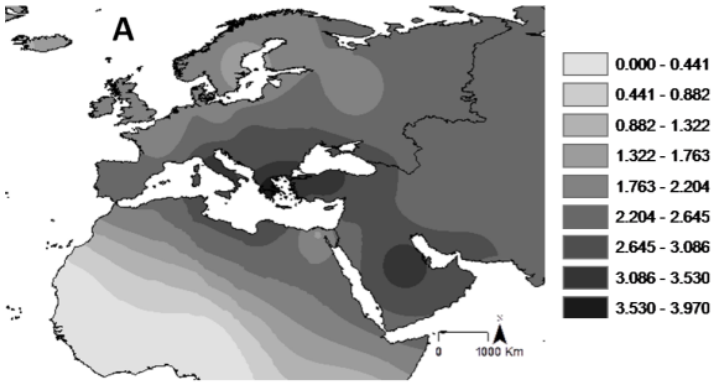


Figure 31:Distribution map for haplogroup I according to the nucleotide diversity measure Pi based on HVS-

I data (adapted from Fernandes et al., 2012).

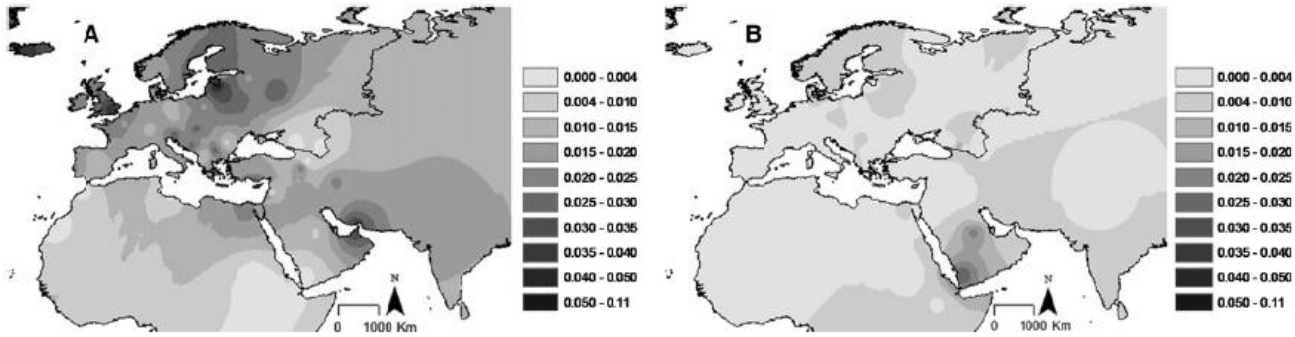


Figure 32: The distribution frequency based on HVS-1 data in the modern populations. A) haplogroup I and B) haplogroup N1a. The degree of the grey colour corresponds to the frequency extent (adapted from Fernandes et al., 2012).

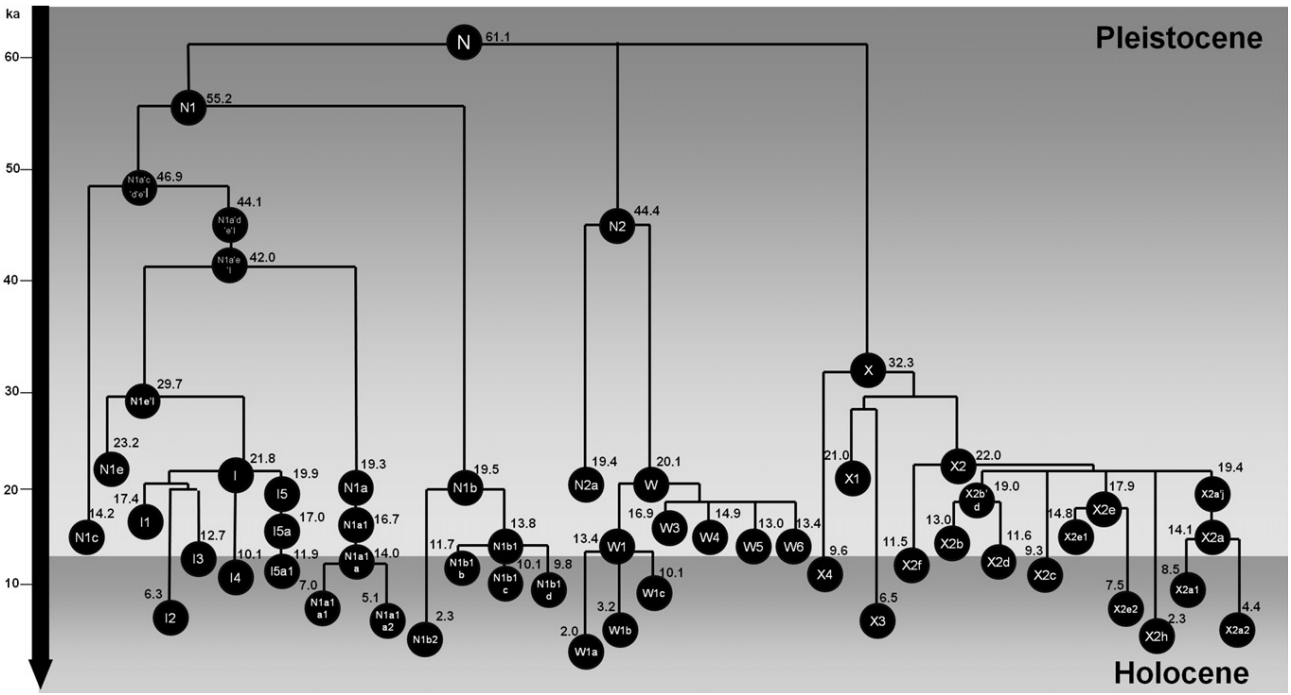


Figure 33: Haplogroup N tree without haplogroup R (NxR) and its clades, maximum likehood (ML) estimated age in ka (thousand years). Each black bubble corresponds to a haplogroup, above each bubble the estimated origin times are given in ka. Pleistocene lasted from about 2,588,000 to 11,700 years ago, while the period of Holocene started from 11,700 years ago (adapted from Fernandes et al., 2012).

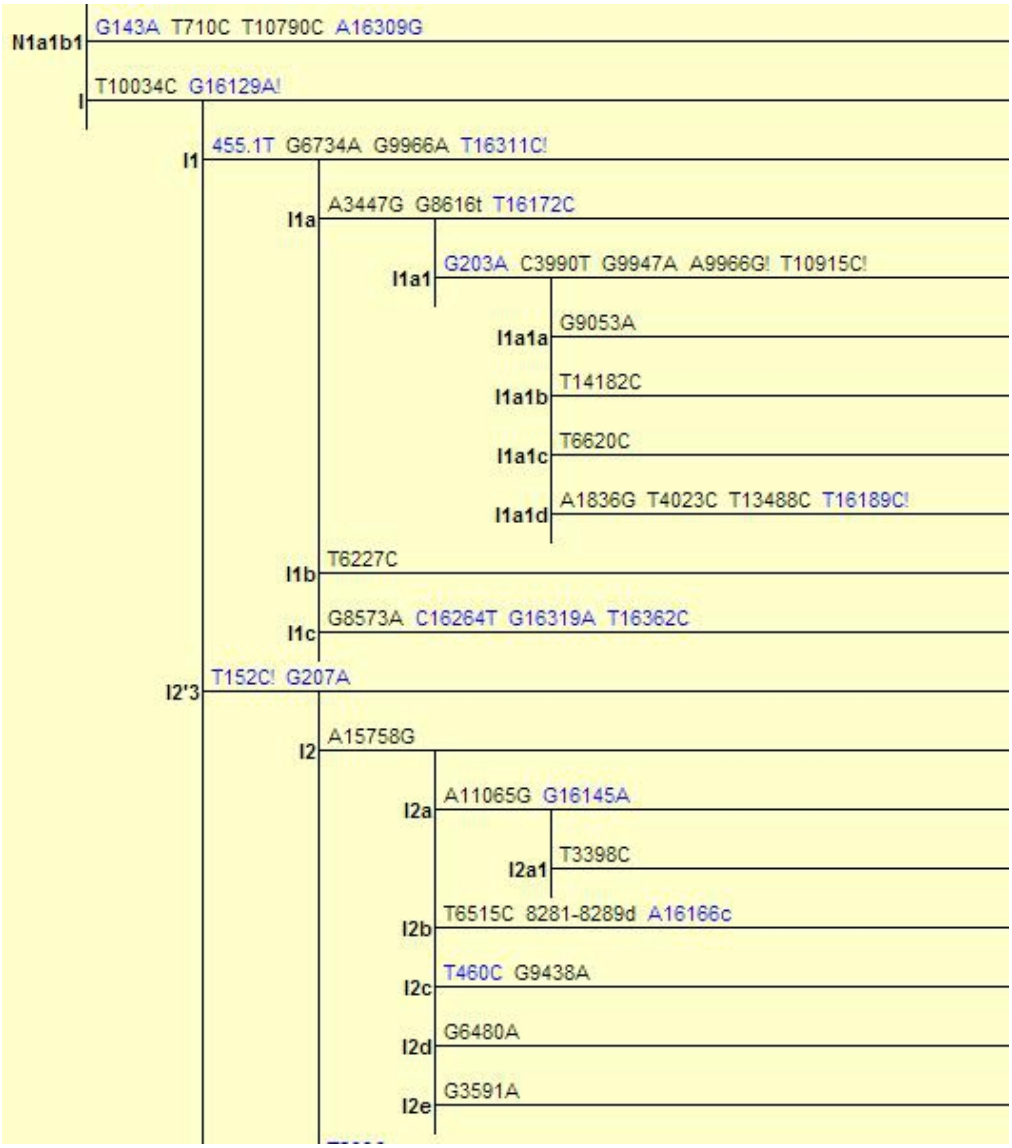


Figure 34: Screen shot of the Phylotree mtDNA subtree N. The diagnostic mutations written in blue are located to the non-coding region of the mitochondrial genome, while those written in black are located to the coding region of the mitochondrial genome. (http://www.phylotree.org/tree/subtree_N.htm).

4.1.6 Defined SNPs and disease association

While evaluating SNPs of the DMG5-I mitochondrion, we also focused on the identification of disease-relevant polymorphisms. Following a search through the published data in the Mitomap database (www.mitomap.org), three polymorphisms of the DMG 5 identified SNPs showed a reported association with different diseases (Table 16).

The first SNP is at a position A10398G and is a variation of the first nucleotide in the codon 114 of NADH dehydrogenase subunit 3 gene (ND3) (10059-10404). This

mutation causes an amino acid substitution from threonine to alanine. Although it is a reported marker of the haplogroup I, J, and K, it has been reported in many genetic association studies with disease susceptibility e.g. breast cancer (Pezzotti et al., 2009) and metabolic syndrome (Juo et al., 2010). Another study suggested this SNP to be a genetic factor against diabetes mellitus type-2 (T2DM) (Liao et al., 2008). It has been also propounded in combination with two other mutations to promote longevity in Finnish and Japanese populations (Niemi, 2005). Also, it has been proposed to be a protective factor against Parkinson's disease and to decrease its risk especially in the haplogroup cluster UKJT (Otaegui et al., 2004; Ghezzi,. et al., 2005; Pyle et al., 2005). Additionally, It has been reported to have a correlation with the lithium response in bipolar disorder (BD), and to increase the mitochondrial matrix pH. Therefore, this polymorphism may play a role in those complex diseases through its effect on the mitochondrial matrix pH since the decrease of pH in the mitochondrial matrix is reported in bipolar disorder (Kazuno et al., 2006; Rollins et al., 2009) (http://www.mitomap.org/bin/view.pl/MITOMAP/MutationsCodingControl).

The second polymorphism is G15043A and is a synonyms mutation in the third nucleotide in the codon 99 of cytochrome b gene (Cytb) (14747-15887). This polymorphism showed a strong association with the major depressive disorder (Rollins et al., 2009).

The third polymorphism is T16519C. It is non-coding and present in the control region of the mitochondrial genome (D-loop). Although, this polymorphism has the highest world-wide mutation rates (Soares et al., 2009), It has been reported to be associated with metastasis in malignant melanoma within Middle European Caucasians (Ebner et al., 2011), and to increase the risk of breast cancer (Bai et al., 2007). It has also been reported to be in higher frequency within type-2 diabetes mellitus (T2DM) Chinese Han cases in comparison to controls ($p < 0.0001$) (Liao et al., 2008) and associated with metabolic disorders, such as exercise oxygen consumption (Murakami et al., 2002). Moreover, a recent report showed an association of T16519C in schizophrenia cases (Mosquera-Miguel et al., 2012). However, it has been recently reported that both mtDNA 16519T and 16519C variants showed associations in psychiatric disorders of schizophrenia and bipolar disorder after using postmortem brain samples for DNA extraction and mitochondrial SNP analysis (Sequeira et al., 2012). It must be stressed that such information collected on association studies were done on modern populations. Unfortunately, no published records are available regarding those SNPs and their

association in the modern or ancient Egyptian populations.

4.2 Metagenomics

4.2.1 Identification of metagenomic features in Egyptian mummies

In addition to the host genome, the aDNA samples contained a large amount of sequences from microbial and environmental sources. Therefore, they are considered metagenomic samples (Huson et al., 2007). The metagenomic analysis approach is a typical topic in most of the aDNA NGS studies, like those on the mammoth, neanderthal, and the Alpine Iceman (Poinar et al., 2006; Reich et al. 2010; Keller et al., 2012). The metagenomic representations of the previously published ancient samples differed from sample to sample according to the burial conditions, its temperature and the preservation status (Poinar et al., 2006; Huson et al., 2007; Wooley and Ye, 2009; Keller et al., 2012). The mammoth study was an example of permafrost conservation and showed a high degree of molecular preservation. Thus, 45.43% of the total reads aligned to the mammoth genome using blastz algorithm with an identity threshold reaching 90%. After the exclusion of all mammoth reads, the alignment to the NCBI nt/nr database showed 18.44% of unidentified sources and 5.76% of bacteria taxa which were mainly Proteobacteria (1.75%) and Actinobacteria (0.91%). The percentages of the Archaea and viruses taxa showed very low values of 0.24% and 0.09%, respectively. The percentage of all the other Eukaryota (others than Gnathostomata) including the plants were represented by 4.15%, while the percentage of the reads aligned to environmental sources was 14.15% (Poinar et al., 2006).

Another example of the cold climate samples is the Tyrolean Iceman (Keller et al., 2012). The metagenomic analysis of the Iceman data-set was done by aligning the reads to the NCBI nt/nr database using blastn algorithm. Using the MEGAN software, the results showed a high human content of 77.6% and a small proportion of bacterial taxa (0.84%). The other Eukaryota were represented by 18.05% within the whole data-set (Keller et al., 2012). The bacteria taxa were mainly Clostridia *(*72%), and interestingly, 0.16% of the bacterial hits assigned to Borrelia. The Borrelia finding was confirmed by an alignment of one million reads to the *B. burgdorferi* genome (60% coverage of the Borrelia genome) (Keller et al., 2012). Another example of cold climate samples is the paleo-eskimo Saqqaq. About 84.2% of the total reads assigned to human, while 0.403 % and 0.4% assigned to bacteria and other Eukaryota taxa, respectively. About 15% were unidentified sequences and a negligible percentage assigned to viruses and Archaea taxa (Rasmussen et al., 2010).

An opposite example with a low endogenous DNA and a high bacterial content is the neanderthal (95%-99% of the extracted DNA derived from microbial genomes) (Green et al., 2010). Metagenomic analysis of one million reads of the neanderthal data-set showed that about 79% of the data-set was unidentified. From 254,933 unique reads, there were two groups with high significance. The first taxonomic order is Actinomycetales with 6.8% and the second is the order primates with 6.2% (Green et al., 2006).

For first insights on the Egyptian mummy metagenomics pattern, two pooled mummy datasets DMG M1 and DMG M2 were dedicated for the analysis using the MEGAN software. To increase the specificity of the MEGAN taxa assignment, a stringent threshold set of the LCA algorithm parameters was used. Although this analytical approach might decrease the total number of hits, it increased the accuracy and the probability of the resulting findings. As a result, highly specific hits of the species *Plasmodium falciparum* were detected in the dataset DMG M2 in contrast to the DMG M1. This presence of *P. falciparum* was confirmed using the PCR technology in the DMG M2 samples. There are a number blood-parasite species of the genus Plasmodium which are responsible for different forms of human malaria. *P. falciparum* causes the most dangerous and severe form of malaria. The presence of *P. falciparum* malaria were reported in the Egyptian mummies (Nerlich et al., 2008; Bianucci et al., 2008; Hawass et al., 2010). The same approach detected the protozoan *Toxoplasma gondii*, which was supported by 1,270 and 843 specific reads in DMG M1 and DMG M2, respectively. Additionally, a recent genetic study showed that the ancient Egyptians reared domesticated cats, Felis silvestris catus, and that the taming of cats occurred prior to or during Predynastic and Early Dynastic periods. Since cats are the definitive host of *T. gondii,* this likely supports the presence of Toxoplasmosis in the Egyptian mummies (Kurushima et al. 2012).

A number of Viridiplantaea taxa like Pinaceae family was detected in the two data-sets DMG M1 and M2. The occurrence of *Pinus sp.* (pine) was highlighted in the DMG M2 and confirmed by PCR. Pinus was documented in many published records and ancient texts as a main component of embalming resins (Serpico and White, 2000; Germer, 2002; Ikram, 2003). Likewise, half of the Viridiplanteae reads were specific hits to *Nicotiana tabacum* which was repeatedly detected in the data-sets DMG M1and M2. Nicotine was chemically detected in hair samples of Egyptian mummies (Balabanova et al., 1992; Parsche et al., 1993; Catemell and Weems, 2001; David, 2008). It was present in small quantities which is consistent with a trace dietary source like *Withania somnifera*

(Ashwagandha or winter cherry) or *Apium graveolens* (Celery) (David, 2008). However, Catemell and Weems (2001) claimed that the occurrence of nicotine in their samples might due to other reasons. Thus, the plant material and gums were applied internally and externally through the embalming process. Additionally, it can also be introduced by other plant goods in the tomb or the storage environment. Interestingly, traces of tobacco were found in the tombs of King Tutankhamun and Ramses II (Buckland and Panagiotakopulu, 2001; Germer, 2002). Here we show DNA evidence of the tobacco plant which might confirm the earlier findings.

4.2.2 Bioinformatic analysis challenges

The MEGAN analysis requires a pre-step of BLASTn alignment against the nt/nr database. In this study, we focused on MEGAN parameters that might have influence on the probability and specificity of the MEGAN results. Additionally, there is a trade-off between the computation time (and subsequently the speed) and the word size of the BLASTn. The BLASTn default word size is 11 and optionally it can be changed up to 64. Increasing the word size resembles the Megablast which is known to be more specific and more suitable for intraspecies comparison (http://www.ncbi.nlm.nih.gov/books/NBK1762/). We investigated the effect of the used word size on the BLASTn results and subsequently on the MEGAN output. One million reads of the dataset DMG 5 were aligned against the nt/nr databse using different word sizes of 30 and 42, resulting in two tabular files, DMG5-Ia and DMG5-Ib, respectively. However, the use of word size 42 decreased the hits number in comparison to that of 30. On the other hand, the specificity of the blast results particularly on the species level increased. This was exemplified by the decrease of variance in Bacteria and Viridiplantae taxa in DMG5-Ib. Thus, a number of taxa with low presence disappeared in DMG5-Ib in comparison to DMG5-Ia. By using the strict MEGAN LCA thresholds, only the assigned reads with the highest probability and specificity were assigned. Interestingly, a number of the bacterial taxa in the two data-sets are documented to live in hard environment i.e. Firmicutes, Deinococcus-Thermus and Aquificae (Fig. 21). Accordingly, Firmicutes and Deinococcus-Thermus were also represented in DMGS-1000, DMGS-1000 and the Iceman data-sets. The embalming protocol and especially the desiccation process were considered antimicrobial which would affect the bacteria growth. Subsequently, we can postulate that the mummification can be also selective towards specific category of taxa i.e. the taxa that could survive in harsh environments.

Additionally, the most dominant Viridiplantae taxa in both data-sets were *Hordeum vulgare*, *Sorghum bicolor* and *Triticum aestivum* (Fig. 23). The three taxa originate from north Africa and some of them were the oldest crops in the world and archaeologically recorded in Egypt from the Neolithic era (Badr et al., 2000; Dixon, 2007; Brenchley et al., 2012). We may speculate, that they might be added by the application of honey in the poorly understood embalming process (i.e. as a source of pollen materials). Alternatively, it may be introduced to the mummy by the tomb environment since the ancient Egyptians used to bury a number of food materials and flowers with the mummies (Serpico and White 2000; Germer, 2002; Ikram, 2003). Or it can be due to the scattered pollen in the burial environment. More studies on Egyptian mummies are needed to reach a conclusion about the source of these findings.

4.2.3 Comparison with other warm and cold climate samples

One million reads from NGS data-sets of different mummy DMGS-1000 and DMGS-2000 as well as those of the previously published Saqqaq, Denisova hominid and the Alpine Iceman (Rasmussen et al. 2010; Reich et al. 2010; Keller et al. 2012) were subjected to the alignment against the nr/nt database. The BLAST results were analyzed and compared using MEGAN. The comparison of the bacterial representations showed an obvious difference which permitted the grouping of the samples in two groups, i.e. the Egyptian mummies and the Iceman metagenome as one group, and another one that included the Saqqaq, Denisova, S1000 and S2000. The main difference between the two groups was the Actinobacteria percentage. Based on the taxon, the two groups could be easily defined as a mummy group and a non-mummy group (Actinobacteria content: <10 % vs ≥28 %, respectively), regardless of the burial condition temperature (Fig. 28). Actinobacteria are very common in soil where they decompose organic materials. According to this a bacterial fingerprint can be concluded for all the mummified tissues regardless of the burial environment and temperature. Moreover, within the mummy group, an increased percentage of Firmicutes was also noticed and was more dominant in Egyptian mummy data-sets. Firmicutes often build endospores and can survive in extreme conditions for a long time. Within the Firmicutes family, the clostridia family was uniquely over-represented in the Egyptian mummies datasets in comparison to the others of either warm or cold climate samples (Fig. 29 a, b).

Viridiplantae representations were compared among the warm and cold climate data-sets (Fig 30 a, b, c). It is suspected that some taxa identified here may have been

utilized in the protocol used for mummification by the ancient Egyptian priests. The data-sets deriving from the warm climate samples showed a higher proportion of the Virdiplanteae even with using the strict MEGAN LCA thresholds. Some of them were uniquely represented in the Egyptian mummies data-sets like *Lotus japonicus, Cucumis sativus* and Fabaceae family. Others like Pinaceae family and the taxa *Ricinus communis* and *Populus trichocarpa* were represented in the warm climate samples generally. A number of the assigned viridiplantae taxa like Castor (*Ricinus communis*), linseed (*Linum usitatissimum*), olive (*Olea europaea L.*), almond (*Prunus dulcis*), populus (*Populus euphratica*), garlic (*Allium sativum*), lotus (*Nymphaea lotus*), fir (*Abies cilicica*) and pine (*Pinus sp.*) were documented to be used in the embalming recipe or for decoration purposes (Serpico and White, 2000; Germer, 2002). Interestingly, they were repeatedly detected in NGS datasets of different Egyptian mummies either in traces or to a higher extent. Additionally, the presence of *Pinus* species was confirmed by PCR technology. More studies are needed to figure out if they are authentic and belong to the embalming materials or if they are present as a result of scattered pollen.

4.3 Conclusion

In summary, the NGS technology is a potential tool to discover more information about the ancient Egyptian civilization. Through the extensive investigation of further Egyptian mummies, we can learn more about their history, their origin and the common pathological and genetic diseases that were present in the ancient Egyptian time. By further studying the Egyptian mummies' metagenomes, more genetic information about the traces of embalming materials can be derived. This information may provide a clear picture about the effect of the embalming materials on the preservation of the Egyptian mummies. Additionally, more can be concluded about the taphonomic process and the microbial degradation process as well as about infectious diseases prevalent in ancient Egypt.

5 References:

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403-410.

Avise JC (2000) Phylogeography: the history and formation of species. Cambridge, MA: Harvard University Press.

Badr A, Müller K, Schäfer-Pregl R, El Rabey H, Effgen S, Ibrahim HH, Pozzi C, Rohde W, Salamini F (2000) On the origin and domestication history of Barley (Hordeum vulgare). Mol Biol Evol 7:499-510.

Bai RK, Leal SM, Covarrubias D, Liu A, Wong LJ (2007) Mitochondrial genetic background modifies breast cancer risk. Cancer Res 67:4687–4694.

Balabanova S, Parsche F, Pirsig W (1992) First identification of drugs in Egyptian mummies. Naturwissenschaften 8: 358.

Barraco RA (1975) Preservation of proteins in a mummified tissue. Paleopathol Newsl 11: 8.

Barraco RA, Reyman TA, Cockburn TA (1977) Paleobiochemical analysis of an Egyptian mummy. J Hum Evol 6:533-546.

Behar DM, van Oven M, Rosset, Metspalu M, Loogväli E-L, Silva NM, Kivisild T, Torroni A, Villems R (2012) A "Copernican" reassessment of the human mitochondrial DNA tree from its root. Am J Hum Genet 90:675-684.

Bell LS, Skinner MF, Jones SJ (1996) The speed of post mortem change to the human skeleton and its taphonomic significance. Forensic Sci Int 82:129–140.

Beyin A (2011) Upper Pleistocene Human Dispersals out of Africa: A Review of the Current State of the Debate. Int J Evol Biol 2011:615094.

Bianucci R, Mattutino G, Lallo R, Charlier P, Jouin-Spriet H, Peluso A, Higham T, Torre C, Massa ER (2008) Immunological evidence of Plasmodium falciparum infection in an Egyptian child mummy from the Early Dynastic Period. J Archaeol Sci 35:1880–1885.

Binladen J, Wiuf C, Gilbert MTP, Bunce M, Barnett R, Larson G, Greenwood AD, Haile J, Ho SYW, Hansen AJ, Willerslev E (2006) Assessing the fidelity of ancient DNA sequences amplified from nuclear genes. Genetics 172:733–741.

Binladen J, Gilbert MT, Bollback JP, Panitz F, Bendixen C, Nielsen R, Willerslev E (2007) The use of coded PCR primers enables high-throughput sequencing of multiple

homolog amplification products by 454 parallel sequencing. PLoS One 2:e197.

Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A; Galaxy Team (2010) Manipulation of FASTQ data with Galaxy. Bioinformatics. 26(14):1783-5.

Brenchley R, Spannagl M, Pfeifer M, Barker GL, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D, Kay S, Waite D, Trick M, Bancroft I, Gu Y, Huo N, Luo MC, Sehgal S, Gill B, Kianian S, Anderson O, Kersey P, Dvorak J, McCombie WR, Hall A, Mayer KF, Edwards KJ, Bevan MW, Hall N (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature 491: 705–710

Briggs AW, Stenzel U, Johnson PL, Green RE, Kelso J, Prüfer K, Meyer M, Krause J, Ronan MT, Lachmann M, Pääbo S (2007) Patterns of damage in genomic DNA sequences from a Neandertal. Proc Natl Acad Sci USA 104:14616-14621.

Briggs AW, Good JM, Green RE, Krause J, Maricic T, Stenzel U, Lalueza-Fox C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Schmitz R, Doronichev VB, Golovanova LV, de la Rasilla M, Fortea J, Rosas A, Pääbo S (2009) Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science 325:318-321.

Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Pääbo S (2010) Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Res 38:e87.

Brotherton P, Endicott P, Sanchez JJ, Beaumont M, Barnett R, Austin J, Cooper A (2007) Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. Nucleic Acids Res 35:5717-5728.

Buckland PC and Panagiotakopulu E (2001) Rameses II and the tobacco beetle. Antiquity 75: 549-556.

Buckley SA and Evershed RP (2001) Organic chemistry of embalming agents in Pharaonic and Graeco-Roman mummies. Nature 413:837-841.

Buehler B, Hogrefe HH, Scott G, Ravi H, Pabón-Peña C, O'Brien S, Formosa R, Happe S (2010) Rapid quantification of DNA libraries for next-generation sequencing. Methods 50:S15-8.

Burbano HA, Hodges E, Green RE, Briggs AW, Krause J, Meyer M, Good JM, Maricic T, Johnson PL, Xuan Z, Rooks M, Bhattacharjee A, Brizuela L, Albert FW, de la Rasilla M, Fortea J, Rosas A, Lachmann M, Hannon GJ, Pääbo S (2010) Targeted

investigation of the Neandertal genome by array-based sequence capture. Science 328:723-725.

Campos PF, Craig OE, Turner-Walker G, Peacock E, Willerslev E, Gilbert MT (2012) DNA in ancient bone - where is it located and how should we extract it? Ann Anat 194:7-16.

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325:31-36.

Cartmell, Larry W, Weems, Cheryl (2001) OVERVIEW OF HAIR ANALYSIS: A REPORT OF HAIR ANALYSIS FROM DAKHLEH OASIS, EGYPT. Chungara Revista de Antropología Chilena, Julio-Sin mes 289-292.

Chaisson MJ and Pevzner PA (2008) Short read fragment assembly of bacterial genomes. Genome Res 18:324-330.

Cipollaro M, Di Bernardo G, Galano G, Galderisi U, Guarino F, Angelini F, Cascino A (1998) Ancient DNA in human bone remains from Pompeii archaeological site. Biochem Biophys Res Commun 247:901-904.

Cipollaro M, Galderisi U, Di Bernardo G (2005) Ancient DNA as a multidisciplinary experience. J Cell Physiol 202:315-22.

Coble MD, Loreille OM, Wadhams MJ, Edson SM, Maynard K, Meyer CE, Niederstätter H, Berger C, Berger B, Falsetti AB, Gill P, Parson W, Finelli LN (2009) Mystery solved: the identification of the two missing Romanov children using DNA analysis. PLoS One 4:e4838.

Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38:1767-1771.

Collins MJ, Penkman KE, Rohland N, Shapiro B, Dobberstein RC, Ritz-Timme S, Hofreiter M (2009) is amino acid racemization a useful tool for screening for ancient DNA in bone? Proc Biol Sci 276:2971-2977.

Colson IB, Bailey JF, Vercauteren M, Sykes BC, Hedges REM (1997) The preservation of ancient DNA and bone diagenesis, Ancient Biomolecules 12:109–117.

Cox MP, Peterson DA, Biggs PJ. SolexaQA (2010) At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11:485.

Cruciani F, Santolamazza P, Shen P, Macaulay V, Moral P, Olckers A, Modiano D, Holmes S, Destro-Bisol G, Coia V, Wallace DC, Oefner PJ, Torroni A, Cavalli-Sforza LL, Scozzari R, Underhill PA (2002) A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. Am J Hum Genet 70:1197-1214.

David AR (1997) Disease in Egyptian mummies: the contribution of new technologies. Lancet 349:1760–1763.

David AR (2008) Egyptian mummies: an overview. In: David AR (ed) Egyptian mummies and modern science, 1st edn. Cambridge University press, New York 10-18.

Derenko M, Malyarchuk B, Grzybowski T, Denisova G, Dambueva I, et al. (2007) Phylogeographic analysis of mitochondrial DNA in northern Asian populations. Am J Hum Genet 81: 1025–1041.

Di Bernardo G, Del Gaudio S, Cammarota M, Galderisi U, Cascino A, Cipollaro M (2002) Enzymatic repair of selected cross-linked homoduplex molecules enhances nuclear gene rescue from Pompeii and Herculaneum remains. Nucleic Acids Res 30:e16.

Disotell TR (2012) Archaic human genomics. Am J Phys Anthropol. 149 Suppl 55:24-39.

Dixon DM (2007) A note on cereals in ancient Egypt. In: The Domestication and Exploitation of Plants and Animals. Ucko PJ and Dimbleby GW, eds. Aldine Pub. Co. I, Chicago, USA.

Donoghue HD, Lee OY, Minnikin DE, Besra GS, Taylor JH, Spigelman M (2010) Tuberculosis in Dr Granville's mummy: a molecular re-examination of the earliest known Egyptian mummy to be scientifically examined and given a medical diagnosis. Proc Biol Sci 277:51-56.

Ebner S, Lang R, Mueller EE, Eder W, Oeller M, Moser A, Koller J, Paulweber B, Mayr JA, Sperl W, Kofler B (2011) Mitochondrial haplogroups, control region polymorphisms and malignant melanoma: a study in middle European Caucasians. PLoS One 6:e27192.

Ermini L, Olivieri C, Rizzi E, Corti G, Bonnal R, Soares P, Luciani S, Marota I, De Bellis G, Richards MB, Rollo F (2008) Complete mitochondrial genome sequence of the Tyrolean Iceman. Curr Biol 18:1687-1693.

Fehren-Schmitz L, Reindel M, Cagigao ET, Hummel S, Herrmann B (2010) Pre-

Columbian population dynamics in coastal southern Peru: A diachronic investigation of mtDNA patterns in the Palpa region by ancient DNA analysis. Am J Phys Anthropol 141:208-221.

Fehren-Schmitz L, Warnberg O, Reindel M, Seidenberg V, Tomasto-Cagigao E, Isla-Cuadrado J, Hummel S, Herrmann B (2011) Diachronic investigations of mitochondrial and Y-chromosomal genetic markers in pre-Columbian andean highlanders from South Peru. Ann Hum Genet 75:266-283.

Fernandes V, Alshamali F, Alves M, Costa MD, Pereira JB, Silva NM, Cherni L, Harich N, Cerny V, Soares P, Richards MB, Pereira L (2012) The Arabian Cradle: Mitochondrial Relicts of the First Steps along the Southern Route out of Africa. The American Journal of Human Genetics 90:347.

Flicek P, Birney E (2009) Sense from sequence reads: methods or alignment and assembly. Nat Methods 6:S6-S12.

Foley RA and Lahr MM (1992) Beyond out of Africa: Reassessing the origins of Homo sapiens. J. Hum. Evol 22:523–529.

Fox CL (1997) mtDNA analysis in ancient Nubians supports the existence of gene flow between sub-Sahara and North Africa in the Nile Valley. Ann Hum Biol 24:217-227.

Fu Q, Meyer M, Gao X, Stenzel U, Burbano HA, Kelso J, Pääbo S (2013) DNA analysis of an early modern human from Tianyuan Cave, China. Proc Natl Acad Sci USA 110:2223-2227.

Gerbault P, Liebert A, Itan Y, Powell A, Currat M, Burger J, Swallow DM, Thomas MG (2011) Evolution of lactase persistence: an example of human niche construction. Philos Trans R Soc Lond B Biol Sci 366:863-877.

Germer R (2002) Die Heilpflanzen de Ägypter. Düsseldorf/Zürich, Artemis und Winkler.

Ghezzi D, Marelli C, Achilli A, Goldwurm S, Pezzoli G, Barone P, Pellecchia MT, Stanzione P, Brusa L, Bentivoglio AR, Bonucceli UU, Petrozzi L, Abbruzzese G, Marchese R, Cortelli P, Grimaldi D, Matinelli P, Ferrarese C, Garavaglia B, Sangiorgi S, Carelli V, Torroni A, Albanese A, Zeviani M (2005) Mitochondrial DNA haplogroup K is associated with a lower risk of Parkinson's disease in Italians. Eur J Hum Genet 13:748-752.

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A (2005) Galaxy: a

platform for interactive large-scale genome analysis. Genome Res 15:1451-1455.

Gilbert MT, Barnes I, Collins MJ, Smith C, Eklund J, Goudsmit J, Poinar H, Cooper A (2005) Long-term survival of ancient DNA in Egypt: response to Zink and Nerlich (2003). Am J Phys Anthropol 128:110-114; discussion 115-118.

Gilbert MT, Bandelt HJ, Hofreiter M, Barnes I (2005) Assessing ancient DNA studies. Trends Ecol Evol 20:541-544.

Ginolhac A, Rasmussen M, Gilbert MT, Willerslev E, Orlando L (2011) mapDamage: testing for damage patterns in ancient DNA sequences Bioinformatics 27:2153-2155.

Glenn TC (2011) Field guide to next-generation DNA sequencers. Mol Ecol Resour11:759-769.

Goecks J, Nekrutenko A, Taylor J; Galaxy Team (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11:R86.

Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, Du L, Egholm M, Rothberg JM, Paunovic M, Pääbo S (2006) Analysis of one million base pairs of Neanderthal DNA. Nature 444:330-336.

Green RE, Malaspinas AS, Krause J, Briggs AW, Johnson PL, Uhler C, Meyer M, Good JM, Maricic T, Stenzel U, Prüfer K, Siebauer M, Burbano HA, Ronan M, Rothberg JM, Egholm M, Rudan P, Brajković D, Kućan Z, Gusić I, Wikström M, Laakkonen L, Kelso J, Slatkin M, Pääbo S (2008) A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. Cell 134:416-426.

Green RE, Briggs AW, Krause J, Prüfer K, Burbano HA, Siebauer M, Lachmann M, Pääbo S (2009) The Neandertal genome and ancient DNA authenticity. EMBO J 28:2494-2502.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, Alkan C, Prüfer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Höber B, Höffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz RW,

Johnson PL, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Pääbo S (2010) A draft sequence of the Neandertal genome. Science 328:710-722.

Greenwood AD, Pääbo S (1999). Nuclear insertion sequences of mitochondrial DNA predominate in hair but not in blood of elephants. Mol. Ecol. 8:133–137.

Guarino FM, Angelini F, Odierna G, Bianco MR, Di Bernardo G, Forte A, Cascino A, Cipollaro M. Detection of DNA in ancient bones using histochemical methods. Biotech Histochem 75:110-7.

Hagelberg E, Sykes B, Hedges R (1989) Ancient bone DNA amplified. Nature 342:485.

Hänni C, Laudet V, Sakka M, Begue A, et Stehelin D (1990). [Amplification of mitochondrial DNA fragments from ancient human teeth and bones]. C R Acad Sci III 310: 365-370.

Hansen AJ, Mitchell DL, Wiuf C, Paniker L, Brand TB, Binladen J, Gilichinsky DA, Rønn R, Willerslev E (2006) Crosslinks rather than strand breaks determine access to ancient DNA sequences from frozen sediments. Genetics 173:1175-1179.

Hawass Z, Gad YZ, Ismail S, Khairat R, Fathalla D, Hasan N, Ahmed A, Elleithy H, Ball M, Gaballah F, Wasef S, Fateen M, Amer H, Gostner P, Selim A, Zink A, Pusch CM (2010) Ancestry and pathology in King Tutankhamun's family. JAMA 303:638-647.

Hawass Z, Ismail S, Selim A, Saleem SN, Fathalla D, Wasef S, Gad AZ, Saad R, Fares S, Amer H, Gostner P, Gad YZ, Pusch CM, Zink AR (2012) Revisiting the harem conspiracy and death of Ramesses III: anthropological, forensic, radiological, and genetic study. BMJ 345:e8268.

Haynes S, Searle JB, Bretman A, Dobney KM (2002) Bone preservation and ancient DNA: the application of screening methods for predicting DNA survival. Journal of Archaeological Science 29: 585-592.

Hebsgaard MB, Phillips MJ, Willerslev E (2005) Geologically ancient DNA: fact or artefact? Trends Microbiol 13:212-220.

Hekkala E, Shirley MH, Amato G, Austin JD, Charter S, Thorbjarnarson J, Vliet KA, Houck ML, Desalle R, Blum MJ (2011) An ancient icon reveals new mysteries: mummy DNA resurrects a cryptic species within the Nile crocodile. Mol Ecol 20:4199–4215.

Henn BM, Botigué LR, Gravel S, Wang W, Brisbin A, Byrnes JK, Fadhlaoui-Zid K, Zalloua PA, Moreno-Estrada A, Bertranpetit J, Bustamante CD, Comas D (2012) Genomic ancestry of North Africans supports back-to-Africa migrations. PLoS Genet 8:e1002397.

Hernandez D, François P, Farinelli L, Osterås M, Schrenzel J (2008) De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res 18:802-809.

Hert DG, Fredlake CP, Barron AE (2008) Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. Electrophoresis 29:4618-4626.

Heyn P, Stenzel U, Briggs AW, Kircher M, Hofreiter M, Meyer M (2010) Road blocks on paleogenomes--polymerase extension profiling reveals the frequency of blocking lesions in ancient DNA. Nucleic Acids Res 38:e161.

Hofreiter M, Serre D, Poinar HN, Kuch M, Pääbo S. Ancient DNA (2001) Nat Rev Genet 2:353-359.

Horai S, Hayasaka K, Murayama K, Wate N, Koike H, Nakai N (1989) DNA amplification from ancient human skeletal remains and their sequence analysis. Proc Jpn Acad Ser B 65:229-233.

Horai S, Hayasaka K, Kondo R, Tsugane K, Takahata N (1995) Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc Natl Acad Sci USA 92:532-536.

Höss M, Jaruga P, Zastawny TH, Dizdaroglu M, Pääbo S (1996) DNA damage and DNA sequence retrieval from ancient tissues. Nucleic Acids Res 24:1304-1307.

Hoyle CH, Thomas PK, Burnstock G, Appenzeller O (1997) Immunohistochemical localisation of neuropeptides and nitric oxide synthase in sural nerves from Egyptian mummies. J Auton Nerv Syst 67:105-108.

Hummel S and Schultes T (2000) from skeletons to finger prints - STR typing of ancient DNA. Ancient Biomol 3:103:116.

Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Res 17:377-86

Huson DH, Mitra S, Ruscheweyh HJ, Weber N, Schuster SC (2011) Integrative analysis of environmental sequences using MEGAN4. Genome Res 21:1552-1560.

Huson DH, Richter DC, Mitra S, Auch AF, Schuster SC (2009) Methods for comparative metagenomics. BMC Bioinformatics 10 Suppl 1:S12.

Ikram S (2003) Death and burial in ancient Egypt. Longman, Harlow.

Jeziorska M (2008) Palaeopathology at the beginning of the new millennium: a review of the literature. In: David AR (ed) Egyptian mummies and modern science, 1st edn. Cambridge University Press, New York, pp 83–98.

Jorde LB, Bamshad M, Rogers AR (1998) Using mitochondrial and nuclear DNA markers to reconstruct human evolution. Bioessays 20:126-136.

Juo SH, Lu MY, Bai RK, Liao YC, Trieu RB, Yu ML, Wong LJ (2010) A common mitochondrial polymorphism 10398A>G is associated metabolic syndrome in a Chinese population. Mitochondrion 10:294–299.

Kazuno A, Munakata K, Nagai T, Satoshi Shimozono S, Tanaka M, Yoneda M, Kato N, Miyawaki A, Kato T (2006) Identification of Mitochondrial DNA Polymorphisms That Alter Mitochondrial Matrix pH and Intracellular Calcium Dynamics. PLoS Genet 2: e128.

Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, Leidinger P, Backes C, Khairat R, Forster M, Stade B, Franke A, Mayer J, Spangler J, McLaughlin S, Shah M, Lee C, Harkins TT, Sartori A, Moreno-Estrada A, Henn B, Sikora M, Semino O, Chiaroni J, Rootsi S, Myres NM, Cabrera VM, Underhill PA, Bustamante CD, Vigl EE, Samadelli M, Cipollini G, Haas J, Katus H, O'Connor BD, Carlson MR, Meder B, Blin N, Meese E, Pusch CM, Zink A (2012) New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. Nat Commun 3:698.

Khairat R, Ball M, Chang CC, Bianucci R, Nerlich AG, Trautmann M, Ismail S, Shanab GM, Karim AM, Gad YZ, Pusch CM (2013) First insights into the metagenome of Egyptian mummies using next-generation sequencing. J Appl Genet (in press).

Kircher M, Kelso J (2010) High-throughput DNA sequencing - concepts and limitations. Bioessays 32:524-536.

Kirsanow K, Burger J (2012) Ancient human DNA. Ann Anat 194:121-132.

Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, Pennarun E, Parik J, Geberhiwot T, Usanga E, Villems R. (2004) Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. Am J Hum Genet 75:752-770.

Kloss-Brandstaetter A, Pacher D, Schoenherr S, Weissensteiner H, Binna R, Specht G, Kronenberg F (2010) HaploGrep: a fast and reliable algorithm for automatic

classification of mitochondrial DNA haplogroups.http://www.haplogrep.uibk.ac.at.

Kloss-Brandstätter A, Pacher D, Schönherr S, Weissensteiner H, Binna R, Specht G, Kronenberg F (2011) HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. Hum Mutat 32:25-32.

Knapp M and Hofreiter M (2010) Next Generation Sequencing of Ancient DNA: Requirements, Strategies and Perspectives. Genes 1:227-243.

Koboldt D, Zhang Q, Larson D, Shen D, McLellan M, Lin L, Miller C, Mardis E, Ding L, Wilson R (2012) VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research. 22(3):568-76.

Krause J, Lalueza-Fox C, Orlando L, Enard W, Green RE, Burbano HA, Hublin JJ, Hänni C, Fortea J, de la Rasilla M, Bertranpetit J, Rosas A, Pääbo S (2007) The derived FOXP2 variant of modern humans was shared with Neandertals. Curr Biol 17:1908-1912.

Krause J, Briggs AW, Kircher M, Maricic T, Zwyns N, Derevianko A, Pääbo S (2010) A complete mtDNA genome of an early modern human from Kostenki, Russia. Curr Biol 20:231-236.

Krings M, Salem AE, Bauer K, Geisert H, Malek AK, Chaix L, Simon C, Welsby D, Di Rienzo A, Utermann G, Sajantila A, Pääbo S, Stoneking M (1999) mtDNA analysis of Nile River Valley populations: A genetic corridor or a barrier to migration? Am J Hum Genet 64:1166-1176.

Kurushima JD, Ikram S, Knudsen J, Bleiberg E, Grahn RA, Lyons LA (2012) Cats of the Pharaohs: Genetic Comparison of Egyptian Cat Mummies to their Feline Contemporaries. J Archaeol Sci 39:3217-3223.

Lalueza-Fox C, Rosas A, Estalrrich A, Gigli E, Campos PF, García-Tabernero A, García-Vargas S, Sánchez-Quinto F, Ramírez O, Civit S, Bastir M, Huguet R, Santamaría D, Gilbert MT, Willerslev E, de la Rasilla M (2011) Genetic evidence for patrilocal mating behavior among Neandertal groups. Proc Natl Acad Sci USA 108:250-253.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature Methods 9:357-359.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.

Lawrence DM, Kemp BM, Eshleman J, Jantz RL, Snow M, George D, Smith DG

(2010) Mitochondrial DNA of protohistoric remains of an Arikara population from South Dakota: implications for themacro-Siouan language hypothesis. Hum Biol 82:157-178.

Lewin PK (1968) The ultrastructure of mummified skin cells. Can Med Assoc J 98:1011-1012.

Li H and Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589-595.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25:2078.

Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18:1851-1858.

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25:1966-1967.

Liao WQ, Pang Y, Yu CA, Wen JY, Zhang YG, Li XH (2008) Novel mutations of mitochondrial DNA associated with type 2 diabetes in Chinese Han population. Tohoku J Exp Med 215:377-384.

Lindahl T (1993) Instability and decay of the primary structure of DNA. Nature 362:709-715.

Lindahl T and Andersson A (1972) Rate of chain breakage at apurinic sites in double-stranded deoxyribonucleic acid. Biochemistry 11:3618-3623.

Lindahl T and Nyberg B (1972) Rate of depurination of native deoxyribonucleic acid. Biochemistry 11:3610-3618.

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M (2012) Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012:251364.

Lorenzen ED, Nogués-Bravo D, Orlando L, Weinstock J, Binladen J, Marske KA, Ugan A, Borregaard MK, Gilbert MT, Nielsen R, Ho SY, Goebel T, Graf KE, Byers D, Stenderup JT, Rasmussen M, Campos PF, Leonard JA, Koepfli KP, Froese D, Zazula G, Stafford TW Jr, Aaris-Sørensen K, Batra P, Haywood AM, Singarayer JS, Valdes PJ, Boeskorov G, Burns JA, Davydov SP, Haile J, Jenkins DL, Kosintsev P, Kuznetsova T, Lai X, Martin LD, McDonald HG, Mol D, Meldgaard M, Munch K, Stephan E, Sablin M, Sommer RS, Sipko T, Scott E, Suchard MA, Tikhonov A, Willerslev R, Wayne RK, Cooper

A, Hofreiter M, Sher A, Shapiro B, Rahbek C, Willerslev E (2011) Species-specific responses of Late Quaternary megafauna to climate and humans. Nature 479:359-364.

Lynnerup N (2007) Mummies. Am J Phys Anthropol 134:162-190.

Lynnerup N (2009) Methods in mummy research. Anthropol Anz 67:357-384.

Magi A, Benelli M, Gozzini A, Girolami F, Torricelli F, Brandi ML (2010) Bioinformatics for Next Generation Sequencing Data. Genes. 1(2), 294-307.

Malyarchuk B, Derenko M, Denisova G, Kravtsova O (2010) "Mitogenomic diversity in Tatars from the Volga-Ural region of Russia". Mol Biol Evol 27: 2220–2226.

Mardis ER. Next-generation DNA sequencing methods (2008) Annu Rev Genomics Hum Genet 9:387-402.

Maricic T, Pääbo S (2009) Optimization of 454 sequencing library preparation from small amounts of DNA permits sequence determination of both DNA strands. Biotechniques 46:51-2, 54-7.

Marota I, Basile C, Ubaldi M, Rollo F (2002) DNA decay rate in papyri and human remains from Egyptian archaeological sites. Am J Phys Anthropol 117:310-318.

Martínez-Alcántara A, Ballesteros E, Feng C, Rojas M, Koshinsky H, Fofanov VY, Havlak P, Fofanov Y (2009) PIQA: pipeline for Illumina G1 genome analyzer data quality assessment. Bioinformatics 25:2438-2439.

Melchior L, Kivisild T, Lynnerup N, Dissing J (2008). Evidence of Authentic DNA from Danish Viking Age Skeletons Untouched by Humans for 1,000 Years. PLoS ONE 3:e2214.

Metcalfe R, Freemont T (2012) Variations in immunohistochemical preservation of proteins in a mummification model. J Anat 220:112-115.

Metzker ML (2010) Sequencing technologies - the next generation.Nat Rev Genet 1:31-46.

Meyer M, Briggs AW, Maricic T, Höber B, Höffner B, Krause J, Weihmann A, Pääbo S, Hofreiter M (2008) From micrograms to picograms: quantitative PCR reduces the material demands of high-throughput sequencing. Nucleic Acids Res 36:e5.

Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, de Filippo C, Sudmant PH, Alkan C, Fu Q, Do R, Rohland N, Tandon A, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, Dabney J, Shendure J, Kitzman J,

Hammer MF, Shunkov MV, Derevianko AP, Patterson N, Andrés AM, Eichler EE, Slatkin M, Reich D, Kelso J, Pääbo S (2012) A high-coverage genome sequence from an archaic Denisovan individual. Science 338:222-226.

Millar CD, Huynen L, Subramanian S, Mohandesan E, Lambert DM (2008) New developments in ancient genomics. Trends Ecol Evol 23:386-93.

Miller W, Drautz DI, Ratan A, Pusey B, Qi J, Lesk AM, Tomsho LP, Packard MD, Zhao F, Sher A, Tikhonov A, Raney B, Patterson N, Lindblad-Toh K, Lander ES, Knight JR, Irzyk GP, Fredrikson KM, Harkins TT, Sheridan S, Pringle T, Schuster SC (2008) Sequencing the nuclear genome of the extinct woolly mammoth. Nature 456:387-390.

Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D (2010) Tablet--next generation sequence assembly visualization. Bioinformatics 26:401-402.

Mitchell D, Willerslev E, Hansen A (2005) Damage and repair of ancient DNA. Mutat Res 571:265-276.

Mosquera-Miguel A, Torrell H, Abasolo N, Arrojo M, Paz E, Ramos-Rios R, Agra S, Paramo M, Brenlla J, Martinez S, Vilella E, Valero J, Gutierrez-Zotes A, Martorell L, Costas J, Salas A (2012) No evidence that major mtDNA European haplogroups confer risk to schizophrenia. Am. J. Med. Genet. B, Neuropsychiatri. Genet 159B:414–421.

Murakami H, Ota A, Simojo H, Okada M, Ajisaka R, Kuno S (2002) Polymorphisms in control region of mtDNA relates to individual differences in endurance capacity or trainability. Jpn J Physiol 52:247-256.

Nerlich AG, Haas CJ, Zink A, Szeimies U, Hagedorn HG (1997) Molecular evidence for tuberculosis in an ancient Egyptian mummy. Lancet 350:1404.

Nerlich AG, Schraut B, Dittrich S, Jelinek T, Zink AR (2008) Plasmodium falciparum in ancient Egypt. Emerg Infect Dis 14:1317-1319.

Nicholson TM, Gradl M, Welte B, Metzger M, Pusch CM, Albert K (2011) Enlightening the past: analytical proof for the use of Pistacia exudates in ancient Egyptian embalming resins. J Sep Sci 34:3364–3371.

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12:443-451.

Niemi AK, Moilanen JS, Tanaka M, Hervonen A, Hurme M, Lehtimäki T, Arai Y, Hirose N, Majamaa K (2005) A combination of three common inherited mitochondrial DNA

polymorphisms promotes longevity in Finnish and Japanese subjects. Eur J Hum Genet 13:166-170.

Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: a fast search method for large DNA databases. Genome Res 11:1725-1729.

Noonan JP, Hofreiter M, Smith D, Priest JR, Rohland N, Rabeder G, Krause J, Detter JC, Pääbo S, Rubin EM (2005) Genomic sequencing of Pleistocene cave bears. Science 309:597-599.

Orlando L, Ginolhac A, Raghavan M, Vilstrup J, Rasmussen M, Magnussen K, Steinmann KE, Kapranov P, Thompson JF, Zazula G, Froese D, Moltke I, Shapiro B, Hofreiter M, Al-Rasheid KA, Gilbert MT, Willerslev E (2011) True single-molecule DNA sequencing of a pleistocene horse bone. Genome Res 21:1705-1719.

Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, Schubert M, Cappellini E, Petersen B, Moltke I, Johnson PL, Fumagalli M, Vilstrup JT, Raghavan M, Korneliussen T, Malaspinas AS, Vogt J, Szklarczyk D, Kelstrup CD, Vinther J, Dolocan A, Stenderup J, Velazquez AM, Cahill J, Rasmussen M, Wang X, Min J, Zazula GD, Seguin-Orlando A, Mortensen C, Magnussen K, Thompson JF, Weinstock J, Gregersen K, Røed KH, Eisenmann V, Rubin CJ, Miller DC, Antczak DF, Bertelsen MF, Brunak S, Al-Rasheid KA, Ryder O, Andersson L, Mundy J, Krogh A, Gilbert MT, Kjær K, Sicheritz-Ponten T, Jensen LJ, Olsen JV, Hofreiter M, Nielsen R, Shapiro B, Wang J, Willerslev E (2013) Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature 499:74-78.

Otaegui D, Paisán C, Sáenz A, Martí I, Ribate M, Martí-Massó JF, Pérez-Tur J, López de Munain A (2004). Mitochondrial polymporphisms in Parkinson's Disease. Neurosci Lett 370:171-174.

Ottoni C, Koon HE, Collins MJ, Penkman KE, Rickards O, Craig OE (2009) Preservation of ancient DNA in thermally damaged archaeological bone. Naturwissenschaften 96:267-78.

Pääbo S (1985) Molecular cloning of Ancient Egyptian mummy DNA. Nature 314:644-645.

Pääbo S, Higuchi RG, Wilson AC (1989) Ancient DNA and the polymerase chain reaction. The emerging field of molecular archaeology. J Biol Chem 264:9709-9712.

Pareek CS, Smoczynski R, Tretyn A (2011) Sequencing technologies and genome

sequencing. J Appl Genet 52:413-35.

Parsche F, Balabanova S, Pirsig W (1993) Drugs in ancient populations. Lancet 341:503.

Patel RK, Jain M (2012) NGS QC Toolkit: a toolkit for quality control of next generation sequencing data 7:e30619.

Pezzotti A, Kraft P, Hankinson SE, Hunter DJ, Buring J, Cox DG (2009) The Mitochondrial A10398G Polymorphism, Interaction with Alcohol Consumption, and Breast Cancer Risk. PLoS ONE 4: e5356.

Poinar HN, Höss M, Bada JL, Pääbo S (1996) Amino acid racemization and the preservation of ancient DNA. Science 272:864-866.

Poinar HN and Stankiewicz BA (1999) Protein preservation and DNA retrieval from ancient tissues. Proc Natl Acad Sci USA 96:8426-8431.

Poinar H, Kuch M, McDonald G, Martin P, Pääbo S (2003) Nuclear gene sequences from a late pleistocene sloth coprolite. Curr Biol 13:1150-1152.

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 (2006) Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science 311:392-394.

Prüfer K, Stenzel U, Hofreiter M, Pääbo S, Kelso J, Green RE (2010) Computational challenges in the analysis of ancient DNA. Genome Biol 11:R47.

Pusch CM, Giddings I, Scholz M (1998) Repair of degraded duplex DNA from prehistoric samples using Escherichia coli DNA polymerase I and T4 DNA ligase. Nucleic Acids Res 26:857-859.

Pusch CM, Blin N, Broghammer M, Nicholson GJ, Scholz M (2000) Adaptor-mediated amplification of minute amounts of severely fragmented ancient nucleic acids. J Appl Genet 41:303-315.

Pusch CM, Broghammer M, Blin N (2003) Molecular phylogenetics employing modern and ancient DNA. J Appl Genet 44:269-290.

Pusch CM and Bachmann L (2004) Spiking of contemporary human template DNA with ancient DNA extracts induces mutations under PCR and generates nonauthentic mitochondrial sequences. Mol Biol Evol 21:957-964.

Pusch CM, Broghammer M, Nicholson GJ, Nerlich AG, Zink A, Kennerknecht I,

Bachmann L, Blin N (2004) PCR-induced sequence alterations hamper the typing of prehistoric bone samples for diagnostic achondroplasia mutations. Mol Biol Evol 21:2005-2011.

Pyle A, Foltynie T, Tiangyou W, Lambert C, Keers SM, Allcock LM, Davison J, Lewis SJ, Perry RH, Barker R, Burn DJ, Chinnery PF (2005) Mitochondrial DNA haplogroup cluster UKJT reduces the risk of PD. Ann Neurol 57:564-567.

Rabino-Massa E and Chiarelli B (1972) The histology of naturally desiccated and mummified bodies. J Hum Evol 1:259-262.

Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, Metspalu M, Metspalu E, Kivisild T, Gupta R, Bertalan M, Nielsen K, Gilbert MT, Wang Y, Raghavan M, Campos PF, Kamp HM, Wilson AS, Gledhill A, Tridico S, Bunce M, Lorenzen ED, Binladen J, Guo X, Zhao J, Zhang X, Zhang H, Li Z, Chen M, Orlando L, Kristiansen K, Bak M, Tommerup N, Bendixen C, Pierre TL, Grønnow B, Meldgaard M, Andreasen C, Fedorova SA, Osipova LP, Higham TF, Ramsey CB, Hansen TV, Nielsen FC, Crawford MH, Brunak S, Sicheritz-Pontén T, Villems R, Nielsen R, Krogh A, Wang J, Willerslev E (2010) Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature 463:757-762.

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PL, Maricic T, Good JM, Marques-Bonet T, Alkan C, Fu Q, Mallick S, Li H, Meyer M, Eichler EE, Stoneking M, Richards M, Talamo S, Shunkov MV, Derevianko AP, Hublin JJ, Kelso J, Slatkin M, Pääbo S (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468:1053–1060.

Reimer PJ, Baillie MGL, Bard E, Bayliss A, Beck JW, Bertrand CJH, Blackwell PG, Buck CE, Burr GS, Cutler KB, Damon PE, Edwards RL, Fairbanks RG, Friedrich M, Guilderson TP, Hogg AG, Hughen KA, Kromer B, McCormac FG, Manning SW, Ramsey CB, Reimer RW, Remmele S, Southon JR, Stuiver M, Talamo S, Taylor FW, van der Plicht J, Weyhenmeyer CE (2004) IntCal04 Terrestrial radiocarbon age calibration, 0–26 cal kyr BP. Radiocarbon 46:1029–1058.

Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T, Villems R, Thomas M, Rychkov S, Rychkov O, Rychkov Y, Gölge M, Dimitrov D, Hill E, Bradley D, Romano V, Calì F, Vona G, Demaine A, Papiha S, Triantaphyllidis C, Stefanescu G, Hatina J, Belledi M, Di Rienzo A, Novelletto A, Oppenheim A, Nørby S, Al-Zaheri N, Santachiara-Benerecetti S, Scozari R, Torroni A,

Bandelt H-J (2000) Tracing European Founder Lineages in the Near Eastern mtDNA Pool. Am J Hum Genet 67:1251–1276.

Richards M, Rengo C, Cruciani F, Gratrix F, Wilson JF, Scozzari R, Macaulay V, Torroni A (2003) Extensive female-mediated gene flow from sub-Saharan Africa into Near Eastern Arab populations. Am J Hum Genet 72:1058–1064.

Richards MB., Sykes BC.,Hedges REM (1995) Authenticating DNA Extracted From Ancient Skeletal Remains. Journal of Archaeological Science 22:291–299.

Roberts C and Ingham S (2008) Using ancient DNA analysis in palaeopathology: a critical analysis of published papers, with recommendations for future work. Int J Osteoarchaeol 18:600-613.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative genomics viewer. Nat Biotechnol 29:24-26.

Rohland N, Reich D, Mallick S, Meyer M, Green RE, Georgiadis NJ, Roca AL, Hofreiter M (2010) genomic DNA sequences from mastodon and woolly mammoth reveal deep speciation of forest and savanna elephants. PLoS Biol 8:e1000564.

Rogaev EI, Grigorenko AP, Faskhutdinova G, Kittler EL, Moliaka YK (2009) Genotype analysis identifies the cause of the "royal disease". Science 326:817.

Rogaev EI, Grigorenko AP, Moliaka YK, Faskhutdinova G, Goltsov A, Lahti A, Hildebrandt C, Kittler EL, Morozova I (2009) Genomic identification in the historical case of the Nicholas II royal family. Proc Natl Acad Sci USA 106:5258-5263.

Rollins B, Martin MV, Sequeira PA, Moon EA, Morgan LZ, Watson SJ, Schatzberg A, Akil H, Myers RM, Jones EG, Wallace DC, Bunney WE, Vawter MP (2009) Mitochondrial Variants in Schizophrenia, Bipolar Disorder, and Major Depressive Disorder. PLoS ONE 4:e4913.

Rose JI (2010). New light on human prehistory in the Arabo-Persian Gulf Oasis. Curr Anthropol 51:849–883.

Saunier JL, Irwin JA, Strouss KM, Ragab H, Sturk KA, Parsons TJ (2009) Mitochondrial control region sequences from an Egyptian population sample. Forensic Sci Int Genet 3:e97-e103.

Scheibye-Alsing K, Hoffmann S, Frankel A, Jensen P, Stadler PF, Mang Y, Tommerup N, Gilchrist MJ, Nygård AB, Cirera S, Jørgensen CB, Fredholm M,Gorodkin J (2009) Sequence assembly. Comput Biol Chem 33:121-136.

Schmieder R, Lim YW, Rohwer F, Edwards R (2010) TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. BMC Bioinformatics 11:341.

Scholz M and Pusch CM (1997) An efficient isolation method for high quality DNA from ancient bones. Technical Tips Online 2:61–64.

Scholz M, Giddings I, Pusch CM (1998) A polymerase chain reaction inhibitor of ancient hard and soft tissue DNA extracts is determined as human collagen type I. Anal Biochem 259:283-286.

Schubert M, Ginolhac A, Lindgreen S, Thompson JF, Al-Rasheid KA, Willerslev E, Krogh A, Orlando L (2012) Improving ancient DNA read mapping against modern reference genomes. BMC Genomics. 13:178.

Schmidt-Schultz TH, Schultz M (2007) Well preserved non-collagenous extracellular matrix proteins in ancient human bone and teeth. Int J Osteoarchaeol 17:91-99.

Schwarz C, Debruyne R, Kuch M, McNally E, Schwarcz H, Aubrey AD, Bada J, Poinar H (2009) New insights from old bones: DNA preservation and degradation in permafrost preserved mammoth remains. Nucleic Acids Res 37:3215-3229.

Sequeira A, Martin MV, Rollins B, Moon EA, Bunney WE, Macciardi F, Lupoli S, Smith EN, Kelsoe J, Magnan CN, van Oven M, Baldi P, Wallace DC, Vawter MP (2012) Mitochondrial Mutations and Polymorphisms in Psychiatric Disorders. Front Genet 3:103.

Serpico M and White R (2000). Oil, fat and wax. In: Paul T. Nicholson PT and Shaw I. Ancient Egyptian materials and technology. Cambridge Univ. Press.

Shapiro B, Hofreiter M (2010) Analysis of ancient human genomes. Bioessays 32:388-391.

Shendure J and Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135-1145.

Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, Salas A, Oppenheimer S, Macaulay V, Richards MB (2009) Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet 84:740-759.

Soares P, Alshamali F, Pereira JB, Fernandes V, Silva NM, Afonso C, Costa MD, Musilova E, Macaulay V, Richards MB, Cerny V, Pereira L (2012). The expansion of mtDNA haplogroup L3 within and out of Africa. Mol. Biol. Evol 29:915-927.

Stiller M, Baryshnikov G, Bocherens H, Grandal d'Anglade A, Hilpert B, Münzel SC,

Pinhasi R, Rabeder G, Rosendahl W, Trinkaus E, Hofreiter M, Knapp M (2010) Withering away--25,000 years of genetic decline preceded cave bear extinction. Mol Biol Evol 27:975-8.

Strachan T and Read A (2010) Human molecular genetics. 4th edition. Garland science, Taylor & Francis group, New York, USA.

Taylor J, Schenck I, Blankenberg D, Nekrutenko A (2007) Using galaxy to perform large-scale interactive data analyses. Curr Protoc Bioinformatics Chapter 10:Unit 10.5.

Terreros MC, Rowold DJ, Mirabal S, Herrera RJ (2011) Mitochondrial DNA and Y-chromosomal stratification in Iran: relationship between Iran and the Arabian Peninsula. J Hum Genet 56:235–246.

Theuson I. and Engberg J (1990). Recovery and analysis of human genetic material from mummified tissue and bone. Journal of Archaeological Science 17:679–689.

Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14:178-192.

van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30:E386-E394.

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. Science 255:737.

Vives S, Gilbert MT, Arenas C, Gigli E, Lao O, Lalueza-Fox C (2008) Statistical analysis of post mortem DNA damage-derived miscoding lesions in Neandertal mitochondrial DNA. BMC Res Notes 1:40.

Wandeler P, Smith S, Morin PA, Pettifor RA, Funk SM (2003) Patterns of nuclear DNA degeneration over time: A case study in historic teeth samples. Mol Ecol 12:1087–1093.

Willerslev E, Cooper A (2005). Ancient DNA. Proc Biol Sci. Jan 7;272(1558):3-16.

Willerslev E, Cappellini E, Boomsma W, Nielsen R, Hebsgaard MB, Brand TB, Hofreiter M, Bunce M, Poinar HN, Dahl-Jensen D, Johnsen S, Steffensen JP, Bennike O, Schwenninger JL, Nathan R, Armitage S, de Hoog CJ, Alfimov V, Christl M, Beer J, Muscheler R, Barker J, Sharp M, Penkman KE, Haile J, Taberlet P, Gilbert MT, Casoli A, Campani E, Collins MJ (2007) Ancient biomolecules from deep ice cores reveal a forested southern Greenland. Science 317:111-114.

Williams SR, Longmire JL, Beck LA (1990). Human DNA recovery from ancient bone. J Phys Anthropol 81:318.

Wisseman S (2001) Preserved for the afterlife. Nature 413:783-784.

Woide D, Zink A, Thalhammer S (2010) Technical note: PCR analysis of minimum target amount of ancient DNA. Am J Phys Anthropol 142:321-327.

Wooley JC and Ye Y (2009) Metagenomics: Facts and Artifacts, and Computational Challenges* J Comput Sci Technol 25:71-81.

Wooley JC, Godzik A, Friedberg I (2010) A primer on metagenomics. PLoS Comput Biol 6:e1000667.

Zerbino DR and Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821-829.

Zhang J, Chiodini R, Badr A, Zhang G (2011) The impact of next-generation sequencing on genomics. J Genet Genomics 38:95-109.

Zink A, Reischl U, Wolf H, Nerlich AG (2000) Molecular evidence of bacteremia by gastrointestinal pathogenic bacteria in an infant mummy from ancient Egypt. Arch Pathol Lab Med 124:1614-1618.

Zink A, Reischl U, Wolf H, Nerlich AG, Miller R (2001) Corynebacterium in ancient Egypt. Med Hist 45:267-272.

Zink AR, Grabner W, Reischl U, Wolf H, Nerlich AG (2003) Molecular study on human tuberculosis in three geographically distinct and time delineated populations from ancient Egypt. Epidemiol Infect 130:239-249.

Zink A, Nerlich AG (2003) Molecular analyses of the "Pharaos:" Feasibility of molecular studies in ancient Egyptian material. Am J Phys Anthropol 121:109-111.

Zink A and Nerlich AG (2005) Long-term survival of ancient DNA in Egypt: Reply to Gilbert et al. American Journal of Physical Anthropology 128:115-118.

Zink AR, Spigelman M, Schraut B, Greenblatt CL, Nerlich AG, Donoghue HD (2006) Leishmaniasis in ancient Egypt and Upper nubia. Emerg Infect Dis 12:1616-1617.

The used URLs

www.Illumina.com

http://en.wikipedia.org/wiki/FASTQ_format

http://samtools.sourceforge.net/samtools.shtml

http://www.genomequest.com/

http://dnanexus.com/

http://galaxyproject.org

http://galaxycast.org

http://maq.sourceforge.net/

http://bowtiebio.sourceforge.net/index.shtml

http://www.sanger.ac.uk/resources/software/ssaha2/

http://bio-bwa.sourceforge.net/

www.appliedbiosystems.com

http://main.g2.bx.psu.edu/root

http://samtools.sourceforge.net/pileup.

http://varscan.sourceforge.net/

http://www.haplogrep.uibk.ac.at.

http://www.phylotree.org/tree/main.htm/

https://galaxy.informatik.uni-tuebingen.de/galaxy-local/

https://galaxy.wur.nl

http://www.ncbi.nlm.nih.gov/projects/SNP

http://www.mitomap.org/bin/view.pl/MITOMAP/PolymorphismsCoding

http://www.phylotree.org/tree/subtree_N.html

www.mitomap.org

http://www.ncbi.nlm.nih.gov/books/NBK1762/

The used samples/ data-sets Glossary

Sample/ data-set code	Type	Description
DMG1-I	soft tissue	Solexa library and its data-set
DMG1-II	soft tissue	SOLiD library and its data-set
DMG1-III	soft tissue	Solexa library and its data-set
DMG1-IV	hard tissue	Solexa library and its data-set
DMG1-V	soft tissue	Solexa library and its data-set
DMG1-VI	hard tissue	Solexa library and its data-set
DMG1-VII	hard tissue	Solexa library and its data-set
DMG1-VIII	soft tissue	Solexa library and its data-set
DMG1-IX	soft tissue	Solexa library and its data-set
DMG2-I	soft tissue	Solexa library and its data-set
DMG2-II	soft tissue	SOLiD library and its data-set
DMG2-III	soft tissue	Solexa library and its data-set
DMG2-IV	soft tissue	Solexa library and its data-set
DMG2-V	soft tissue	Solexa library and its data-set
DMG3-I	soft tissue	Solexa library and its data-set
DMG4-I	soft tissue	Solexa library and its data-set
DMG5-I	hard tissue	Solexa library and its data-set
DMG5-Ia	data-set	BLASTn file using a word size 30
DMG5-Ib	data-set	BLASTn file using a word size 42
DMG6-I	soft tissue	Solexa library and its data-set
DMG6-II	soft tissue	Solexa library and its data-set
DMG6-III	soft tissue	Solexa library and its data-set
DMG-M1	DNA	Pooled DNA samples of DMG2-I and DMG3-I
DMG-M2	DNA	Pooled DNA samples of DMG1-I and DMG 6-I
Saqqaq	data-set	Solexa published data-set (Rasmussen et al. 2010)
Denisova hominid	data-set	Solexa published data-set (Reich et al. 2010)
Iceman	data-set	SOLiD published data-set (Keller et al. 2012)