A genomic study of *Mycobacterium leprae* in medieval Northern Europe

Dissertation in fulfillment of the requirements for the degree “Dr. rer. nat.” of the Faculty of Mathematics and Natural Sciences at Christian-Albrechts-Universität zu Kiel

Dipl. -Ing. Marion Bonazzi

Kiel, 2017

First referee: Prof. Manuela Dittmar,

Faculty of Mathematics and Natural Sciences

Second referee: Prof. Almut Nebel,

Faculty of Medicine

Date of oral examination: 04/11/2016

Signed:

Prof. Wolfgang Duschl (Dean)

A la mémoire de Claire

Graduate School Human Development in Landscapes

Institute of Clinical Molecular Biology

List of figures

List of tables

List of abbreviations

µL	microliter
µm	micrometer
µM	micromolar (µmol per L)
AD	Anno Domini (Unless stated otherwise, all the dates in this dissertation are AD)
aDNA	ancient DNA
BC	before Christ
BCG	bacillus Calmette-Guérin
BL	borderline leprosy
CT	computer tomography
DEL	deletion
DLL	diffuse lepromatous leprosy
DNA	deoxyribonucleic acid
EB	extraction blank control
INS	insertion
LB	library blank control
LL	lepromatous leprosy
MTBC	*Mycobacterium tuberculosis* complex
Myr	million year
NC/nc	negative control
nd	no data
NGS	next-generation sequencing
nm	nanometer
nM	nanomolar (nmol per L)
PB	PCR blank control
PCR	polymerase chain reaction
SNP	single nucleotide polymorphism
SNV	single nucleotide variation
STR	single tandem repeat
TB	tuberculosis
TL	tuberculoid leprosy
und.	undetermin

1 Introduction

1.1 A short history of leprosy in Northern Europe

Leprosy is one of the oldest infectious human diseases (Donoghue 2013; Monot et al. 2005; Zias 2002). Leprosy sufferers often displayed serious disfiguration that led to strong social reactions such as marginalization and isolation of the diseased individual (Roberts & Manchester 1996). As a consequence, the depiction and discussion of the disease has been a part of ancient cultures for millennia (Singh & Cole 2011; Roberts & Manchester 1996).

Leprosy is thought to have originated in Africa before the Pleistocene era (Schuenemann et al. 2013; Roberts & Manchester 1996; Monot et al. 2005) and to have been brought into Europe when the armies of Alexander the Great returned from India in 327-326 BC (Dols 1979; Roberts & Manchester 1996). The first convincing evidence for leprosy in Northern Europe appears in written sources around 600 BC (Zias 2002; Kjellström 2012; Trautman 1984). However, on most occasions, the detail within historical records is insufficient to confirm the presence of the disease (Roberts & Manchester 1996). The first osteological evidence of leprosy in Europe dates from 5th century England, France and Sweden (Reader 1974; Kjellström 2012; Blondiaux 2002).

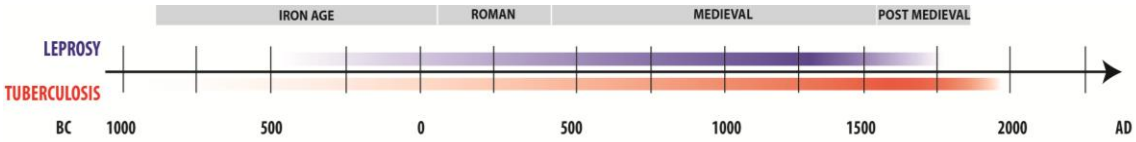


Figure 1: History of leprosy in Northern Europe
Intensity of the colour reflects the prevalence of the disease. As the first evidence of a disease does not necessarily relate to its actual first case and as there might also have been unreported cases which persisted after the official eradication date, the extremities of the coloured bars are blurred.

The 11th to 15th century medieval period saw the highest prevalence of the disease (Boldsen 2005a; Boldsen 2009b). In some communities in Denmark, it is thought that virtually 100% of the population was infected (Boldsen 2005b; Boldsen & Mollerup 2006; Boldsen 2009b). Leprosy remained endemic in Europe until it almost disappeared during the 16th century (Figure 1) (Lietman et al. 1997; Britton & Lockwood 2004; Bates & Stead 1993). Interestingly, between the 15th and the 16th centuries, the prevalence of leprosy decreased without the use of modern medication (Stone et al. 2009; Manchester 1984). At the same time in which leprosy disappeared, the number of tuberculosis (TB) cases skyrocketed (Aufderheide & Rodriguez-Martin 1998; Bates

& Stead 1993), despite the fact that TB was present in Europe long before the Middle-Ages (Singh & Cole 2011; Matheson et al. 2009; Stone et al. 2009; Nuorala 2004; Hulse 1972; Hershkovitz et al. 2008; Palfi et al. 1999; Mays & Taylor 2003; Faerman & Jankauskas 2000). Indeed, TB kept a low frequency until the 16th century (Trautman 1984), but rose to cause half of all human deaths during the following few centuries (Gagneux 2012).

The observed switch in disease prevalence around the 16th century is rather surprising. Indeed, the respective prevalence of leprosy and tuberculosis depend on the same environmental factors: population density and health status. Therefore, both should have kept rising during the late medieval period, as the proportion of people living in restrained low-income urban areas was steadily increasing (Hunter & Thomas 1984; Stone et al. 2009; Lenski & May 1994; Anderson & May 1982; Scheuer 1992). So far, the reasons behind the disappearance of leprosy from medieval Northern Europe are not fully understood, although several hypotheses have been proposed. Unfortunately, our understanding of the causes is still in a state at which none of the scenarios proposed seems more likely than any other.

Until present, the majority of researchers thought that the virulence of leprosy might have decreased over the time as it co-evolved with humans, as has been shown for other pathogens (Stone et al. 2009; Heesterbeek et al. 2015; Anderson & May 1982; May & Anderson 1983; Dubos 1980). Indeed, selection pressure would favour mildly virulent leprosy bacteria, as the infected host would live longer with the disease, thereby extending the possibilities of spreading the bacteria to new hosts. In contrast, highly virulent leprosy bacteria might kill their host before having the opportunity to spread (Anderson & May 1982; May & Anderson 1983; Lenski & May 1994). However, the recent sequencing of a medieval *M. leprae* genome has shown striking genetic conservation during the last 1000 years (Schuenemann et al. 2013; Han & Silva 2014). The fact that there is no major genetic variation between the ancient and modern seems to point towards a stability in the virulence of the pathogen, thereby suggesting another cause for the decrease in the prevalence of leprosy in the late Middle-Ages.

It has also been suggested that the early segregation of leprosy sufferers might have helped to limit the spread of the infection and participated in the decline of leprosy (Scheuer 1992; Boldsen 2007). However, recent studies have proven this hypothesis very unlikely, since leprosy displays a very long incubation period during which the affected individual is almost asymptomatic (see part 1.2.1) and is, nonetheless, highly contagious (G. A. Clark et al. 1987; Manchester 1984). Currently under investigation is the idea that a change in the human genome could have reduced susceptibility to leprosy (Fine 1982). Indeed, as with many infectious diseases, the fitness of the

individuals affected is reduced, a fact which eventually leads to the selection of more resistant individuals. In fact, some human genetic variations have been shown to be linked to leprosy resistance (Fine 1982; Fine 1981). It has also been suggested that the early segregation of leprosy sufferers might have artificially helped the genetic selection process (Scheuer 1992).

The most supported hypothesis is based on the observation that the modern-day bacillus Calmette-Guérin (BCG) vaccination against tuberculosis also provides partial protection against leprosy (Scheuer 1992; Chaussinand 1953; Fine 1984). Indeed, the main immune response to infection with either tuberculosis or leprosy is cell-mediated acquired immunity (Jopling & McDougall 1988; Ridley & Jopling 1966; Lurie 1955) in which the parasite species is recognised, ingested and digested by macrophages (Scheuer 1992). It has been shown that the cell-mediated response is not completely specific (Aufderheide & Rodriguez-Martin 1998; Scheuer 1992; Chaussinand 1953) and that *M. tuberculosis* hosts can be immune to other infections (Mackaness 1968). As a consequence, the growing exposure to TB at the end of the Middle-Ages could have provided increased resistance to leprosy (Scheuer 1992; G. A. Clark et al. 1987; Manchester 1984). Only very few ancient cases of co-infections have been documented, a fact which is surprising given that leprosy and TB were endemic in Northern Europe during the same period (Donoghue et al. 2005; Roberts & Manchester 1996). This suggests that individuals infected with TB could, indeed, have been partially immune to leprosy (Boldsen 2007; Chaussinand 1953; Lietman et al. 1997). In addition, TB is more contagious and virulent than leprosy (Gagneux 2012; Han & Silva 2014; Walther & Ewald 2004). Therefore, individuals infected with TB were more likely to die before infection by leprosy could develop into a chronic disease and lead to bone lesions (Stone et al. 2009). Finally, the late medieval period saw the rapid development of cattle breeding and trade in Northern Europe (Scheuer 1992; Boldsen 2009b). Cattle are hosts to *M. bovis*, a close relative of *M. tuberculosis*. The former has been proven to also infect humans (O'Reilly & Daborn 1995; Stone et al. 2009). The increased exposure to bovine TB through contact with cattle and the consumption of meat and dairy products from infected animals could have participated in the progressive immunisation of the population against leprosy (K. A. Clark et al. 1987; Monot et al. 2005). The hypotheses listed above are not mutually exclusive; it is likely that several or all of them played a role in the decrease of the prevalence of leprosy in favour of TB around the 16th century.

1.2 Aetiology of leprosy

1.2.1 The clinical symptoms of leprosy and its transmission mechanisms

The transmission of leprosy is thought to be principally airborne as transmitted via infected aerosols (Rodrigues & Lockwood 2011; Bryceson & Pfaltzgraff 1990; Sharma et al. 2015; Scheuer 1992; Jopling & McDougall 1988). The incubation time of leprosy ranges between 2-20 years depending on the infected person's immune status and the extent of his or her exposure to the pathogen (WHO 2016; Jopling & McDougall 1988; Fine 1982).

The facial peripheral nervous system is the first invaded by the bacteria (Jopling & McDougall 1988; WHO 2016; Scheuer 1992) an event which lead to the "rhino-maxillary syndrome" as well as the depigmented skin patches which have been so vividly described (Stieglmeier et al. 2014). The advent of the rhino-maxillary syndrome causes the anterior nasal spine and margins to disappear progressively, giving the nose a typical "sunken" appearance (Andersen & Manchester 1992; Andersen et al. 1994). In severe cases, the upper maxilla undergoes resorption and the front teeth are partially lost (Misch et al. 2010; Möller-Christensen 1978; Andersen et al. 1994; Boldsen 2009b). The infection can spread to the limbs (thereby causing permanent nerve damage) which is often accompanied by osteomyelitis under an overlying skin infection (Desikan & Job 1968). Because leprosy bacteria show a strong affinity to the Schwann cells around nerves (Han & Silva 2014; Stieglmeier et al. 2014; Kaplan & Cohn 1986; Scheuer 1992; Bloom & Godal 1983; Scollard et al. 2006), infected individuals experience a decrease in the sensation of pain in the extremities (Stieglmeier et al. 2014) which can, in turn, lead to a loss of phalanges due to injury and infection.

The spectrum of clinical symptoms related to leprosy depends greatly on how efficiently the host immune system restrains the pathogen (Ridley & Jopling 1966; Degang et al. 2014; Kaplan & Cohn 1986; Bloom & Mehra 1984; Scollard et al. 2006). Therefore, not all infected individuals will present the same symptomatic features. Consequently, the percentage of individuals who developed bone lesions in medieval times is estimated at only 5% (Roberts & Manchester 1996). Table 1 describes the main types of leprosy according to the Ridley-Jopling scale (Ridley & Jopling 1966; Degang et al. 2014) along with the newly-discovered diffuse lepromatous leprosy (DLL) (Han et al. 2009; Han et al. 2008) and their characteristics in terms of pathogen load (the number of bacteria living in the body) and host resistance (see Table 1). Most leprosy sufferers fall between LL and TL in the borderline leprosy type (BL) and demonstrate mixed characteristics. BL

is described as an unstable state that can develop into acute LL or chronic TT. In between those extremes are borderline conditions which display mixed features. These will not be detailed here, as they are unlikely to be properly diagnosable by osteological observation alone (see section 1.2.5 Comparison with *Mycobacterium tuberculosis*). For further reading on the subject, see section 8 References (Stieglmeier et al. 2014; Ridley & Jopling 1966).

The recently described DLL (Han et al. 2008; Han et al. 2009) represents an extreme form of leprosy. Until recently, it has only been observed in a small number of patients and has yet to be fully investigated. It is clinically characterised by systemic sub-cutaneous swelling as well as the formation of numerous large ulcerative skin lesions (Han et al. 2008).

Table 1: Different types of leprosy with regard to the immune status of the individual

	Diffuse lepromatous leprosy (DLL) (Han et al. 2008; Han et al. 2009; Han et al. 2015)	Lepromatous leprosy (LL)	Tuberculoid leprosy (TT)
Pathogen load (Britton & Lockwood 2004; Ridley & Jopling 1966; Jopling & McDougall 1988)	High	High	Low
WHO Classification	Multibacillary	Multibacillary	Paubacillary
Causative agent	*M. lepromatosis*	*M. leprae* *M. lepromatosis*	*M. leprae*
Bone changes	ND	Severe	Mild to none
Skin/nerve lesions (Stieglmeier et al. 2014)	Numerous	Numerous Symmetric	Rare
Skeletal involvement (Roberts & Manchester 1996; Blondiaux 2002)	ND	Usually symmetrical	Unilateral
Cell-mediated immunity (McMurray 1996; Britton & Lockwood 2004)	Low response	Low response	High response
Humoral antibody response (McMurray 1996; Stieglmeier et al. 2014; Britton & Lockwood 2004)	High	High	Low
Infectivity (Manchester 1984)	ND	High	Low
Immune evasion (Britton & Lockwood 2004; McMurray 1996)	No	No	High

Due to the very recent discovery of M. lepromatosis*, descriptions of the symptoms are still incomplete (ND).*

As living conditions and the health status of the population were poor in the Middle-Ages, the clinical presentation spectrum of leprosy is thought to have shifted towards the LL form (Andersen et al. 1994). This has been widely discussed (Boldsen 2001; Boldsen 2007; Stone et al. 2009; Roberts & Manchester 1996; Manchester 1984), as bone lesion frequency was higher in the past than it is today.

1.2.2 Causative agents of leprosy

To this day, two causative agents of leprosy are known. The first one, *Mycobacterium leprae (M. leprae)*, was first described at the end of the 19th century by the physician G. A. Hansen (Stieglmeier et al. 2014; Monot et al. 2005). More recently, a second causative bacterium, named *Mycobacterium lepromatosis (M. lepromatosis)* was discovered (Han et al. 2009). Both *M. leprae* and *M. lepromatosis* are human pathogens, although singular *M. leprae* infections have been described in animals such as chimpanzees (Alford et al. 1996; Valverde et al. 1998). In the southern United States, nine-banded armadillos have been shown to be naturally susceptible to *M. leprae*, forming the only known zoonotic reservoir for human leprosy (Sharma et al. 2015; Truman et al. 2011; Loughry et al. 2009).

M. leprae and *M. lepromatosis* are rod-shaped bacteria with exceptionally long generation times (Singh & Cole 2011; Han et al. 2008). Both species belong to the *Mycobacterium* genus as described by Lehman and Neumann in 1896 (Stieglmeier et al. 2014; McMurray 1996). Therefore, they display many of the biological features which are specific to *Mycobacterium* described below.

M. leprae and *M. lepromatosis* cells are non-motile rods 1-8µm long and 0.2-0.5µm in diameter that can form clusters (Stieglmeier et al. 2014; McMurray 1996; Scollard et al. 2006; Draper 1983; Cowdry 1978). Both species have a characteristic thick and highly hydrophobic cell wall rich in mycolic acids and peptidoglycans (Draper 1983; Daffe & Draper 1998; Draper et al. 1987). Although they are Gram positive bacteria, their cell wall prevents discoloration by acids during staining for observation (Scheuer 1992; Draper 1983) (Figure 2).

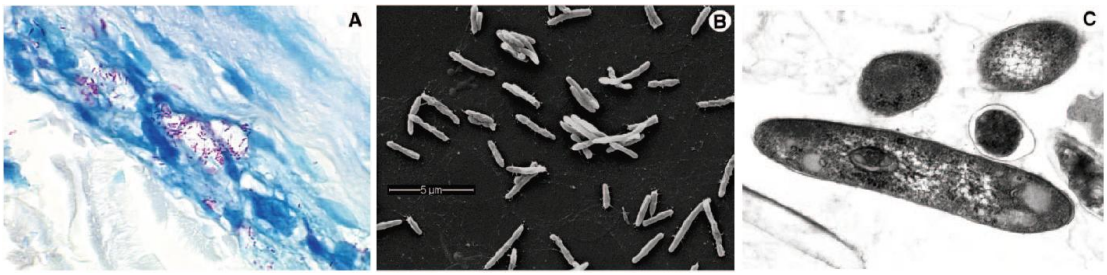


Figure 2: Morphology of M. leprae
A. M. leprae *stained via the Fite-Faraco method appears as red, rod-shaped organisms (magnification 800X). B.* M. leprae *under the scanning electron microscope, (magnification 12,000X). C. Internal features of* M. leprae *are observed in this ultrathin section of the bacilli under a transmission electron microscope (magnification 29,000X) (Scollard et al. 2006).*

Moreover, *M. leprae* and *M. lepromatosis* are obligate intracellular pathogens. Thus, they fail to grow on artificial media (Han & Silva 2014; Han et al. 2009; Stieglmeier et al. 2014; Scollard et al. 2006). Although they have recently been successfully cultivated in armadillos and the foot pads of mice (Singh et al. 2015; Singh & Cole 2011; Kirchheimer & Storrs 1971), the efficiency of the cultivation approach is strongly hindered by the extremely long doubling time required by the bacteria (Singh & Cole 2011; Han et al. 2008) (see Figure 2). Indeed, the division cycle of *M. leprae* lasts more than 20 days and is the longest amongst *Mycobacteria* (Britton & Lockwood 2004). For these reasons, phenotypic studies of the pathogens remain extremely rare. Consequently, the knowledge available today about the biology and drug-resistance of leprosy is mainly provided by genomic investigations performed directly on pathogen DNA isolated from human patients.

1.2.3 Basic infection mechanisms

Upon infection, *M. leprae* bacteria settle first in the intracellular milieu around Schwann cells and exhibit a particular preference for the extremities of the body due to its low optimal growth temperature (about 33°C) (Scollard et al. 2006; Britton & Lockwood 2004). There, it is quickly recognized and ingested by macrophages, its primary host cells. In tuberculoid leprosy (TT) infections, the macrophages digest the pathogen and present pathogen-specific antigens to T-cells. Subsequently, numerous T-cells are recruited to limit the spread of the infection, forming the characteristic granulomas observed in patients with TT (Scollard et al. 2006). As a consequence, only a few well-defined skin lesions are present, almost no antibodies are secreted and the pathogen load stays low (Britton & Lockwood 2004). On the other side, during the development of lepromatous leprosy (LL), the bacteria are not digested by the macrophages and spread extensively. Subsequently, the pathogen load is high and numerous skin patches appear.

Leprosy is the only bacterial disease which affects peripheral nerves (Scollard et al. 2006). Although this unique feature is likely related to the bacterium's capacity to bind to a protein of the basal lamina of Schwann cells, the molecular mechanisms leading to the invasion of the nerve are not fully understood (Britton & Lockwood 2004). On the *M. leprae* membrane, PGL-1 and a 21kDa protein have been shown to be involved in the uptake of the bacterium by Schwann cells when they interact with the host cell's laminin receptor (Britton & Lockwood 2004; Shimoji et al. 1999; Ng et al. 2000). How those protein interactions lead to the bacterium invading the host cell remains unclear (Britton & Lockwood 2004; Scollard et al. 2006). Afterwards, the bacterium slowly replicates inside the Schwann cell. Eventually, the host cell presents mycobacterial antigens to specific T-cells, triggering a strong chronic immune response (Spierings et al. 2001; Britton & Lockwood 2004). As a consequence, the nerve is damaged at two levels: the infected Schwann cells are specifically killed by CD4+ T-cells and the inflammation leads to further damage due to swelling inside the nerves.

1.2.4 Genetics and evolution of *Mycobacterium leprae* and *lepromatosis*

M. leprae and *M. lepromatosis* display the smallest of the Mycobacteriagenomes (See Table 2) (Han et al. 2015; Han et al. 2009; Singh & Cole 2011). This is thought to relate to their strict obligate parasite nature (Han et al. 2015; Han et al. 2009; Singh & Cole 2011). In addition, their GC content is significantly reduced compared to other Mycobacteria (Han et al. 2009; Clark-Curtiss et al. 1985; Imaeda et al. 1982; Gross & Wayne 1970; Wayne & Gross 1968). GC content in genomes has been shown to be dependent on the percentage of coding sequences (Pozzoli et al. 2008), with coding sequences being higher in guanine and cytosine than non-coding regions. In the case of *M. leprae* and *lepromatosis*, the reduced GC content is consistent with the reduced gene content (Table 2).

The *M. lepromatosis* genome has not yet been completely sequenced. The current *M. lepromatosis* genome drafts only differ from the *M. leprae* genome by 9.1%, placing it as a separate (yet very close) relative species (Han & Silva 2014; Han et al. 2008; Vera-Cabrera et al. 2011; Singh et al. 2015; Jessamine et al. 2012). Over the last few years, *M. lepromatosis* has been isolated from leprosy patients in Mexico, Canada and Singapore, suggesting that it might be as widely spread and as ancient as *M. leprae* (Han & Silva 2014; Monot et al. 2009; Han et al. 2009; Singh et al. 2015*)*. Genomic comparisons between the *M. lepromatosis* isolates and various

Mycobacterium genomes also support an ancient origin for the species (Han et al. 2009; Han & Silva 2014; Singh et al. 2015).

Almost half of the two bacteria genomes are composed of pseudogenes, indicating a strong reductive evolution resulting from one massive gene inactivation event within the last 20Myr (Gomez-Valero et al. 2007; Singh & Cole 2011) (see Figure 2). Indeed, during the adaptation to a parasitic lifestyle, genes non-essential to infection and parasitism are submitted to low selection pressure (Cole et al. 2001; Gomez-Valero et al. 2007). As a consequence, the accumulation of substitutions, insertions or deletions might deactivate complete groups of genes (Cole et al. 2001; Monot et al. 2005; Belda et al. 2010; Gil et al. 2008; Toh et al. 2006; Delmotte et al. 2006). Those deactivated genes would see their GC content drop, which would be consistent with the low GC percentage in *M. leprae* and *lepromatosis* genomes. Non-functional DNA was eventually lost, reducing the size of the genome (Gomez-Valero et al. 2007; Silva et al. 2001; Mira et al. 2001).

Table 2: Genomic and biological features of Mycobacterium leprae, lepromatosis *and two TB-causing relatives*

	M. leprae	*M. lepromatosis*	*M. tuberculosis*	*M. bovis*
References	(Monot et al. 2009; Singh & Cole 2011)	(Singh et al. 2015; Han & Silva 2014)	(Cole et al. 2001; Scollard et al. 2006)	(Michel et al. 2010; Wirth et al. 2008)
Genome size	3.27 Mb	3.21 Mb	4.41 Mb	4.35 Mb
G+C content	57.8 %	57.9 %	65.6 %	65.6 %
Genes	2,770	2,777	4,008	4,001
Pseudogenes	1,115 (41 %)	1,334 (47 %)	30 (<1 %)	33 (<1 %)
Growth on artificial media	No	No	Yes	Yes
Generation time	>20 days	>20 days	15-20 hours	16-20 hours
Main host	Human	Human	Human	Cattle
Pathogenicity in humans	Yes	Yes	Yes	Yes
Phenotype severity	High	Extreme	High	Mild
Bone lesions frequency (Holloway et al. 2011; Roberts & Manchester 1996)	5 %	ND	3-7 %	30-50 %

For comparison, data for the related species M. tuberculosis *and* M. bovis *are displayed on the right. Because* M. lepromatosis *has been discovered only recently, its information is incomplete.*

No *M. leprae*- or *M. lepromatosis*-specific plasmids have been described so far, although typical bacterial plasmids have been found in most Mycobacteria(Meyers 1995). Plasmids are commonly found in most bacterial pathogens, as they increase the capacity to adapt to changes in the environment and have been shown to often carry genes involved in pathogenicity or drug-resistance (Wang et al. 2016). During the reductive evolution process, *M. leprae* and *lepromatosis* might have lost their plasmids. Han et al. (Han & Silva 2014) constructed a phylogenetic tree of several *Mycobacterium* species using the sequences of the conserved rpoB protein gene. They showed that *M. leprae* and *M. lepromatosis* diverged from their last common ancestor ~10Myr ago, around the end of the reductive evolution process (Figure 3).

From those results (Anastasiou & Mitchell 2013; Han & Silva 2014), it can be assumed that the evolution of leprosy bacteria started when their last common ancestor (a parasite which targeted early human ancestors) underwent reductive evolution. It is thought that around the time that this genome reduction ended, *M. leprae* and *M. lepromatosis* became separate species by passing through an evolutionary bottleneck which lead to their high clonal stability.

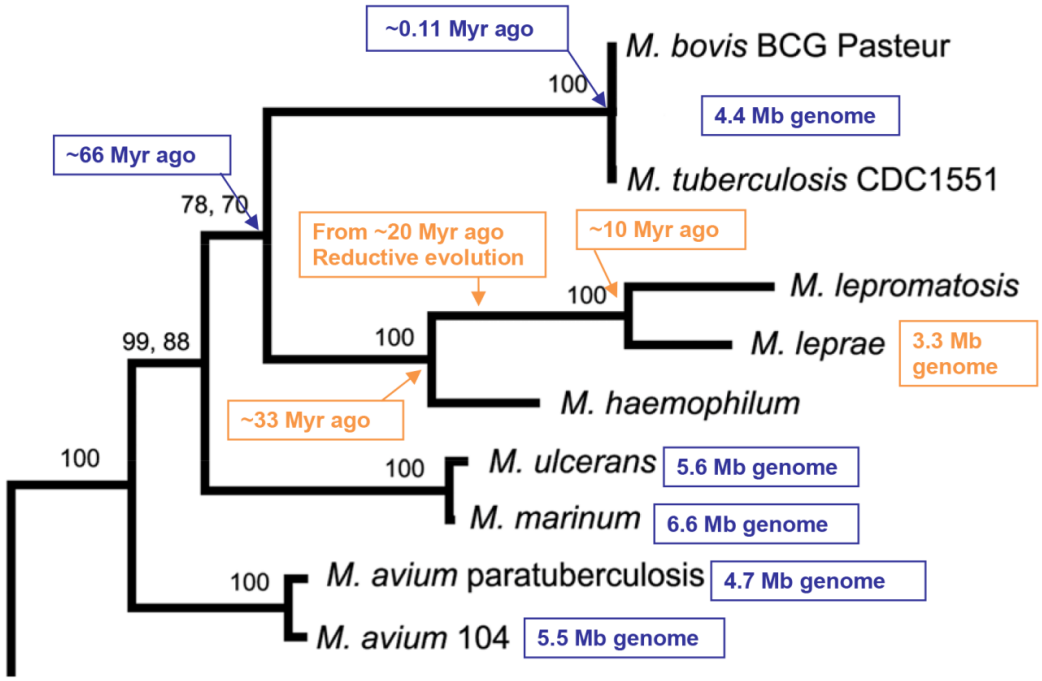


Figure 3: Phylogenetic tree of several Mycobacterium *species*
This Maximum-Likelihood tree is based on the amino acid sequences of rpoB proteins (Han & Silva 2014). The numbers to the left of each node are bootstrap values for 1000 replications. The length of the branches is proportional to the number of substitutions.

Monot and colleagues (Monot et al. 2005) have compared modern *M. leprae* genomes and grouped them into four genetic types defined by three single nucleotide polymorphisms (SNPs). More recently, the typing of leprosy strains has been refined to 16 genotypes (defined by 84 SNPs and InDels) which show more than 99% identity (Monot et al. 2009; Monot et al. 2005; Singh & Cole 2011). The study of the modern geographical distribution of those types and subtypes showed that the spread of leprosy followed human dispersals (Monot et al. 2009). Currently, *M. leprae* SNP types 2 and 3 are the only ones present in Europe. The geographical repartition of the various leprosy types suggests that leprosy was brought to Europe in successive events from an east African ancestor which resembled the present-day SNP type 2. From one of the earliest events, leprosy ancestral SNP type 2 evolved into SNP type 2 and 3 (Singh & Cole 2011; Monot et al. 2009) (Figure 4).

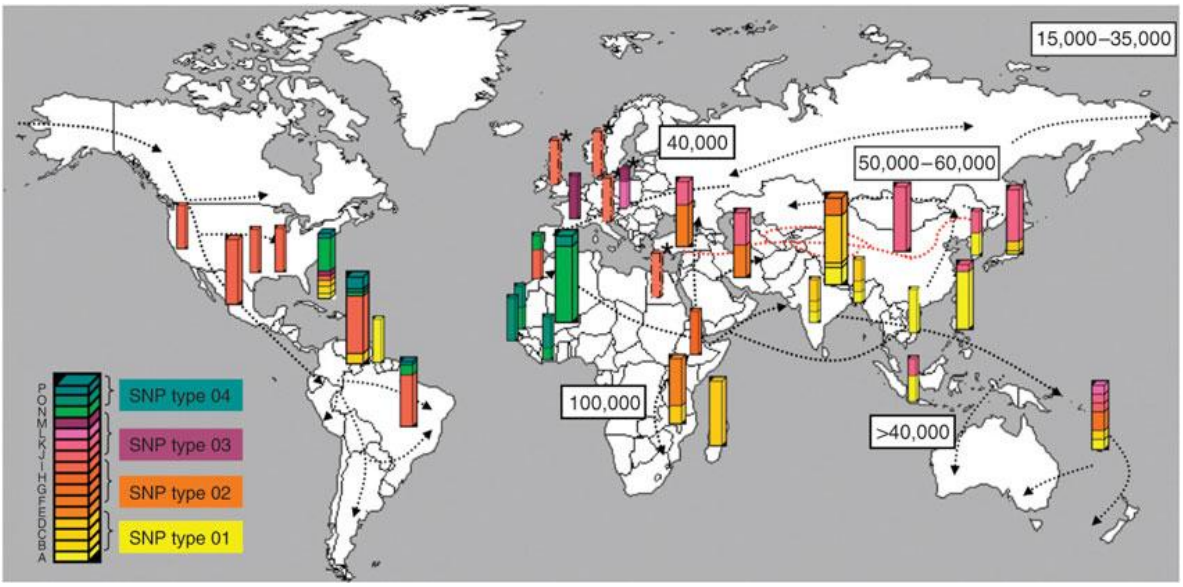


Figure 4: Geographical distribution of M. leprae *SNP types*
Pillars are located on the country of origin of the M. leprae *sample and are colour coded according to the scheme for the 16 SNP subtypes (Monot et al. 2009). The thickness of the pillar corresponds to the number of samples (1-5, thin; 6-29, intermediate; >30, broad). The grey arrows indicate human migration routes, with the estimated time of migration shown in years. The red dots indicate the location of the Silk Road in the first century, and * denotes results obtained from aDNA.*

1.2.5 Comparison with *Mycobacterium tuberculosis*

Bacteria of the Mycobacterium Tuberculosis Complex (MTBC) are close cousin species of *M. leprae* and *M. lepromatosis (see* Figure 3*)*. MTBC members cause tuberculosis (TB) in various animals (Stone et al. 2009; Manchester 1984). The two most common causes of TB in humans are *M. tuberculosis* and *M. bovis* (O'Reilly & Daborn 1995; Donoghue 2009; Fine 1984). Bacteria of the MTBC also display most of the *Mycobacterium*-specific biological features, as do leprosy-causing bacteria. One major discrepancy between leprosy-causative agents and MTBC members lies in their generation time. Indeed, MTBC species' duplication time is counted in hours, whereas *M. leprae* and *M. lepromatosis* need up to 20 days to duplicate (Monot et al. 2009; Scollard et al. 2006). Moreover, TB agents display larger genomes with almost no pseudogenes (see Table 2). This is considered consistent with their broader host range as well as their capacity to grow on artificial media, since the genes necessary to their own function were likely preserved during evolution. Only one *M. tuberculosis* plasmid (pTYGi9) has been published so far (NCBI, accession number NC_025025.1). Genetically, *M. leprae* and *M. tuberculosis* are about 50% identical, despite the fact that the *M. tuberculosis* genome is more than 1Mb bigger (Cole et al. 2001; Scollard et al. 2006; Imaeda et al. 1982; Gross & Wayne 1970). In general, *Mycobacteria* species show a high degree of genetic similarity (Gross & Wayne 1970; Imaeda et al. 1982; Stone et al. 2009) which can make them difficult to differentiate through molecular biology.

1.2.6 Osteological diagnosis of leprosy in ancient remains

The bone lesions specific to leprosy which develop during the course of the disease are often the only evidence which is preserved by which ancient leprosy sufferers can be osteologically identified (Boldsen 2007; Andersen et al. 1994; Andersen & Manchester 1992). The main parts of the skeleton involved in the development of leprosy are the facial bones as well as the small bones of the hands and feet (See Figure 5), (Boldsen 2007; Andersen & Manchester 1992; Andersen et al. 1994). This manifests on the anterior cranium in the rounding of the edges of the nasal aperture and the complete or partial disappearance of the nasal spine and vomer (Boldsen 2007). The upper maxilla shows evidence of bone resorption which is sometimes accompanied by the absence of the upper front teeth (Boldsen 2007). The hands and feet often display shortened small bones as well as terminal phalanges which are often partially absent (Boldsen 2007). Due to the restriction of bone lesions (especially facial deformations) chiefly to the LL form of leprosy, it is especially difficult to accurately diagnose other forms of the disease. Although TT leprosy can

sometimes be successfully diagnosed if the limb bones are present, borderline forms are extremely difficult to identify.

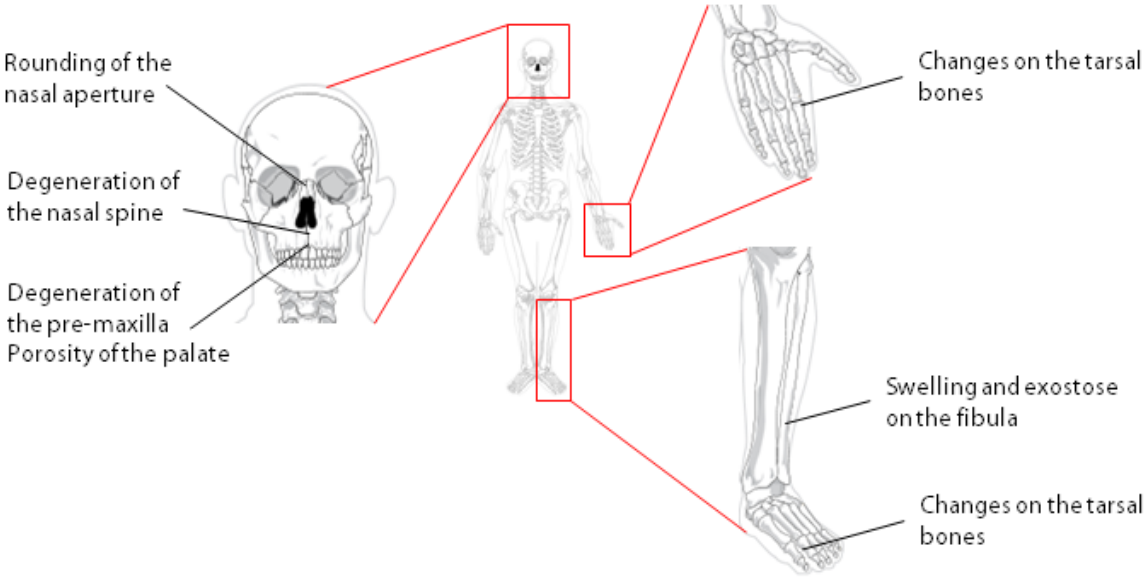


Figure 5: Diagnostic bone lesions for leprosy described in the literature
Extensive description of the lesions can be found in the work of J. Boldsen (Boldsen 2007) as well as photographic examples. This figure only shows the six features used during the diagnosis of the samples collected for this study.

An osteological diagnosis has several drawbacks. The bone features are evaluated and graded by eye and the observations are then used as a whole to establish a diagnosis (Boldsen 2007). Since the beginnings of paleopathology, experts have tried to establish standard gradation scales for each type of lesion according to its location and severity. Nevertheless, the reliability of the diagnosis is heavily dependent on the observer's experience. Moreover, the preservation state of the remains highly influences the diagnosis' quality, as any missing bone reduces the amount of information available. Another disadvantage originates from the aetiology of leprosy itself, since the percentage of affected individuals who develop bone lesions is relatively low. Although the development of bone lesions might have been more frequent in medieval times (Boldsen 2009a; Boldsen 2001), the percentage of medieval remains showing bone lesions is bound to not be an accurate representation of the actual number of persons who suffered from leprosy during that time period. Therefore, palaeopathological observations are still mainly used to identify the remains of leprosy sufferers and to roughly estimate the variations in leprosy prevalence over time rather than as a means to declare individuals free of the disease.

1.3 Ancient DNA studies of mycobacterial infections

1.3.1 The limitations of studies based on modern-day leprosy

Research on modern leprosy has yielded numerous publications about the epidemiology, immunology and aetiology of the disease (Scollard et al. 2006; Britton & Lockwood 2004). Those studies, however, are all based on either *M. leprae* strains that faced (and survived) modern medication for the last 70 years or patients who had benefitted from modern-day diagnosis methods and health care. Although they have helped unravel the history and the evolution of the disease since it first came in contact with humans (Singh et al. 2015; Singh & Cole 2011; Monot et al. 2005; Monot et al. 2009; Cole et al. 2001), those studies can only be partially used to understand the disease and its impact on human societies before the medical progress of the early 1900s. Indeed, differences in the development of the disease have been noticed (Boldsen 2001; Stone et al. 2009) and suggest that host-pathogen interactions in the past were more complex than those described by modern-day data alone.

1.3.2 Preservation of mycobacterial DNA molecules in ancient remains

Ancient remains of living organisms can still contain the endogenous DNA (Hagelberg & Clegg 1991; Pääbo 1989; Higuchi et al. 1984) of the deceased organism as well as that of any microorganism which lived in the animal at its time of death (Pääbo et al. 2004). The preservation of ancient DNA (aDNA) molecules in a deceased organism depends on many environmental factors, including temperature, humidity and oxygen levels (Poinar et al. 2008; Pääbo et al. 2004).

Since the mid-1980s, methods for the study of DNA have been progressively adapted to the characteristics of aDNA (Pääbo 1989; Higuchi et al. 1984; Pääbo 1985; Pääbo et al. 2004; Shapiro & Hofreiter 2012). Under certain conditions, it is possible to extract and analyze endogenous DNA from long-dead organisms as well as the DNA of the pathogens that might have infected them (Cooper & Waynet 1998). When preserved, aDNA is generally present in minimal amounts (Pääbo et al. 2004) and is accompanied by much larger amounts of exogenous microbial DNA (Hoss et al. 1996; Noonan 2005). Moreover, DNA molecules spontaneously degrade in the absence of repair mechanisms inside a living cell (Lindahl 1993). In consequence, aDNA is chemically modified compared to modern DNA. In fact, aDNA molecules are typically short (rarely longer than 150bp) and display an increased occurrence of Guanine (G) and Thymine (T) at the 3'-end and 5'-end of

the fragments, respectively, due to deamination of the purine bases (Lamers et al. 2009; Overballe-Petersen et al. 2012; Briggs et al. 2007; Brotherton et al. 2007).

Mycobacterium species present biological characteristics that enhance the preservation of their DNA in ancient remains. They have a tropism for bone structure cells and display a robust lipid-rich cell wall. The strong hydrophobic nature of this cell wall protects the *Mycobacterium* DNA from hydrolytic damage (Haas et al. 2000; Donoghue et al. 2004; Donoghue 2011; Donoghue et al. 2015). In addition, *Mycobacterium* genomes display high G+C contents, a feature that has been shown to increase DNA stability over time (Donoghue 2013). As a result, in aDNA extracts from human remains, *Mycobacterium* DNA fragments can sometimes be longer and less degraded than human DNA (Schuenemann et al. 2013; Mendum et al. 2014).

1.3.3 PCR investigations of ancient leprosy cases

The Polymerase Chain Reaction (PCR) was invented in 1984 and allows for the amplification of a specific DNA target fragment of known sequence. It was almost immediately adapted for the aDNA research field, leading to the first amplification of aDNA molecules which was reported the very same year (Higuchi et al. 1984; Pääbo 1985). The first analysis of ancient *Mycobacterium* DNA was performed in 1993 (Spigelman & Lemma 1993) and consisted of the PCR amplification of a *M. tuberculosis* 123bp-long fragment from the IS6110 insertion element. The first DNA amplification from ancient leprosy sequences was reported soon thereafter (Rafi et al. 1994). Sanger sequencing technology enables researchers to read the sequences produced during the PCR by using the PCR primers as sequencing primers. Since then, PCR has become a powerful tool for detecting and studying the aDNA of pathogens in archaeological samples (Matheson et al. 2009; Donoghue et al. 2001; Maricic et al. 2010; Fletcher et al. 2003).

Recently, PCR methods have allowed for the identification of cases of co-infection between leprosy and tuberculosis (Donoghue et al. 2005). In addition, PCR amplifications are commonly used by aDNA laboratories to genotype the isolated strains and to establish their place in the phylogeny of the *Mycobacterium* genus (Mendum et al. 2014; Taylor et al. 2013). However, the PCR approach is hindered by the possibility of contamination as the close similarity between *Mycobacterium* species can lead to non-specific amplifications (Stone et al. 2009). Several published results have been debated *a posteriori* due to missing negative controls or insufficient description of the measures taken to avoid contamination (Lindahl 1997; Hofreiter 2008; Barnes

& Thomas 2006; Cooper & Waynet 1998; Stone et al. 2009; Willerslev & Cooper 2005; Wilbur et al. 2009). Indeed, the very principle of PCR increases the risk of contamination as the molecular degradation of ancient DNA tends to reduce the amplification yield of authentic PCR products compared to non-degraded contaminating DNA fragments (Willerslev & Cooper 2005; Cooper & Waynet 1998; Hofreiter 2008).

1.3.4 Next-generation sequencing studies of ancient *Mycobacterium* genomes

Over the last decades, the development of Next-Generation Sequencing (NGS) has considerably increased the throughput of DNA sequencing while also allowing for the discovery of new ways of handling aDNA molecular specificities. NGS involves the ligation of all DNA molecules from an extract to known artificial adapters to create a DNA library (see Methods 3.2.2). That library can then easily be amplified and sequenced as a whole, avoiding the need to target known DNA sequences. Creating libraries also increases the life expectancy of DNA fragments (by protecting them at each end, excluding any interaction with a host genome and removing the need of clone cultivation) and enables the fragments to be re-amplified virtually endlessly (Kircher et al. 2012; Meyer & Kircher 2010). The library also preserves the misincorporated bases in the aDNA fragments, thereby establishing a new way to rule out modern contamination (Sawyer et al. 2012; Overballe-Petersen et al. 2012). Moreover, libraries can be identified using molecular barcodes to multiplex DNA sequencing reactions to reduce sequencing costs and to decrease the risks of contamination. To compensate for aDNA degradation, libraries can be prepared using an enzyme that replaces misincorporated Uraciles bases by Cytosines (Hofreiter et al. 2001), a great advantage in the field of aDNA. This “UDG treatment” greatly reduces the number of wrongly-sequenced Thymines and, consequently, the rate of false C->T polymorphisms observed in comparison with the reference sequence (Hofreiter et al. 2001; Shapiro & Hofreiter 2012; Lamers et al. 2009).

The basic “shotgun” approach consists of sequencing an aDNA library prepared directly from a whole aDNA extract. It is often used in the initial evaluation of the aDNA extract contents in terms of diversity and the proportions of species present. For the same reason, it is also a common metagenomics method for ancient samples. As the shotgun approach relies on the breadth of the reference genome databases available to match the DNA fragments, a large majority of the aDNA reads obtained cannot be attributed to any specific species and are classified as unknown (Krause 2010; Schuenemann et al. 2013). Typically, only 1-2 % of the reads actually belong to a species of interest.

Using NGS, new cases of *M. tuberculosis* and *M. leprae* infections have been confirmed (Mendum et al. 2014; Schuenemann et al. 2013; Müller et al. 2014; Kay et al. 2015). However, publications reporting the use of NGS to study ancient Mycobacteria infections are still scarce. Whole-genome drafts from ancient *M. leprae* strains were successfully assembled for the first time in 2013 (Schuenemann et al. 2013). Interestingly, in one sample, the ancient *M. leprae* DNA molecules were so well-preserved in the sample that no enrichment was necessary to perform *de-novo* assembly. To date, there is no record of successful ancient *Mycobacteria* Next-Generation Sequencing projects performed on more than a dozen individuals; most of them report the recovery of only a handful of partial ancient mycobacterial genomes.

1.3.5 Authenticity of the *Mycobacterium* genomes recovered

Numerous *Mycobacteria* species are free-living organisms that can subsist in the environment (Han & Silva 2014; Anastasiou & Mitchell 2013; Bouwman et al. 2012; Wilbur et al. 2009). Therefore, in an ancient DNA extract, the mycobacterial genetic material which is found can come either from mycobacterial infections which took place during the life of the person under investigation or mycobacterial colonization of the corpse following burial. Moreover, modern mycobacterial contamination can occur if the ancient remains are not kept and handled correctly (Yang & Watt 2005; Walther & Ewald 2004; Knapp et al. 2012; Krause 2010). In consequence, the authenticity of the ancient genomes recovered has to be carefully established. Contamination with modern DNA is commonly avoided by performing all aDNA pre-PCR work in dedicated facilities in which contamination levels are controlled by using blanks. Contamination of the endogenous DNA by the environment before excavation has to be assessed after sequencing by searching for DNA sequences specific to the endogenous or environmental species of interest. In addition, all settings have to be adjusted to minimise aspecific mapping during data processing. The precautions taken to avoid contamination and to check for the authenticity of the results in each part of this study will be detailed in the corresponding methods section (see part 4.1.1).

1.5 Research questions and objectives

So far, most ancient *M. leprae* DNA studies have been based on a few isolated cases from different periods and locations. The overall aim of this project is to investigate the genomic diversity of *M. leprae* from a defined geographic area and period (Germany and Denmark between 1000-1560 AD). The aims can be subdivided into several specific objectives:

a. **Collecting samples from human skeletal remains** which date from 11th – 16th-century Northern Europe which demonstrate bone lesions suggestive of leprosy. In addition, samples with tuberculosis lesions will be included as a means of investigating possible co-infections.

b. **Screening the remains for *M. leprae* and *M. tuberculosis* DNA using PCR** to identify samples with good *M. leprae* DNA preservation and to identify possible *M. tuberculosis* co-infection cases.

c. **Next-Generation Sequencing** of the confirmed leprosy cases from a and b: pooled shotgun sequencing will be performed to identify the best samples for re-sequencing.

d. **Confirming potential *M. tuberculosis* co-infection cases** using the Next-Generation Sequencing data obtained in c.

e. **Identifying *M. leprae* genetic variations to construct an *M. leprae* phylogeny**.

f. **Investigating the possible effects of the variations** using *in-silico* prediction tools.

g. **Interpreting the data obtained in the historical and biomedical context** to obtain a more detailed picture of the history and evolution of *M. leprae* in Northern Europe.

2 Material and Methods

2.1 Sample collection and dating

2.1.1 Description of the cemeteries and human remains

Figure 6: Geographical origin of the sample sets
In total, 140 samples were collected from three locations.

The first sample cohort was excavated in 1980/81 in St. Jørgen (Odense, Denmark) from an urban leprosarium cemetery that was in use between 1270 and 1560 (Boldsen 2007). Approximately 1544 graves were discovered in this cemetery (Boldsen 2005b). The diagnosis of the disease was confirmed by distinct lesions on the bones of the individuals concerned (Boldsen 2007; Boldsen & Mollerup 2006). From this cemetery, a subset of 34 samples was randomly selected for this study (see Supplementary Table 1:).

The second sample set comes from the Rathausmarkt cemetery (Schleswig, Germany) which was in use from 1000 to 1250 (Boldsen 2009a). The cemetery excavation in the 90s (Lüdtke 1997; Boldsen et al. 2013; Grupe 1997) uncovered 223 individuals. The anthropological evaluation was conducted by J. Boldsen, S. Weise and colleagues (Lüdtke 1997; Grupe 1997; Boldsen 2009a) and identified 120 individuals with leprosy- or tuberculosis-related lesions. Of those 120 individuals, 79 were further investigated within this study (see Supplementary Table 2).

The third sample collection was taken from material recovered from Ribe (Denmark). This cemetery was in use between 1250 and 1400 and was excavated in 1993. Although the analysis of the skeletal material is not yet complete (ongoing work is being undertaken by D. Pedersen and J. Boldsen, ADBOU, Denmark), the first 27 individuals diagnosed with either leprosy or tuberculosis were sampled for inclusion within this ancient DNA study (see Supplementary Table 3).

2.1.2 Radiocarbon dating

Radiocarbon dating was performed on those samples which showed promising results in the first sequencing phase. Indeed, the previously-cited date estimations for the cemeteries provided by the archaeologists were based on their observations of the changes in the arm position of the deceased over time. This is not sufficiently accurate for the purposes of phylogenetic analysis. The chosen samples were sent to the Curt-Engelhorn Archäometrie gGmbH laboratories where the collagen was extracted and prepared for radiocarbon dating on a mini-radiocarbon dating system.

2.2 Minimizing contamination by modern DNA

In fulfilment of the guidelines relative to contamination control in aDNA studies (Yang & Watt 2005; Pilli et al. 2013; Knapp et al. 2012), all surfaces and reusable utensils were extensively cleaned with bleach before and after work. In addition, UV lights were used to improve the decontamination process whenever possible. Moreover, negative controls were regularly used for each step involving non-indexed DNA molecules. Finally, the access of both persons and objects was strictly restricted in the laboratory to a one-direction route from the pre-PCR rooms to post-PCR rooms.

The pre-PCR facilities are located at the Universitätsklinikum Schleswig Holstein (UKSH, Kiel) campus and consist of three rooms used sequentially for sample preparation, DNA extraction and library preparation and PCR set-up. No modern DNA work or PCRs have ever been carried out in those rooms. In addition, this study is the first to extract and study *Mycobacterium* DNA in those rooms. Proper laboratory clothing was worn in the pre-PCR facilities, namely (and in the order of suit-up): a surgical mask, hairnet, sterile surgical latex gloves, hooded laboratory overall, latex boots and a second pair of gloves (which were changed as often as was necessary).

The PCR room was located in an independent building on the UKSH campus. As for the pre-PCR rooms, no leprosy studies (either modern or ancient) were conducted previous to this project. This room was used to run the screening and indexing PCRs (for which the reaction mixes had previously been prepared in the dedicated pre-PCR rooms) and to process the indexing PCR products (purification, amplification).

The sequencing facilities are located at the Zentrum für Molekulare Biowissenschaften (ZMB) on the main campus of Kiel University. Two sequencing platforms dedicated to Sanger sequencing and next-generation sequencing respectively took care of the technical work involved in DNA sequencing after thorough DNA preparation according to their requirements.

2.3 Selection of the remains and identification of putative leprosy cases

2.3.1 Pre-selection according to established archaeological data

Putative leprosy cases in each cemetery were carefully selected according to the work of expert palaeopathologists (Lüdtke 1997; Grupe 1997; Boldsen 2009a). For the leprosarium cemetery subset, samples with unspecific diagnoses were also added to the study (see Supplementary Table 1:). This choice was due to the fact that the prevalence of leprosy was high when the St. Jørgen leprosy hospital was open, in addition to the reliable diagnosis of leprosy in medieval times (Stone et al. 2009). Moreover, given the proximity between leprosy sufferers in the hospital, the poor understanding of the transmission of the disease prevalent at the time and the absence of modern medication or cleaning methods, it is likely that even a healthy person who spent time in the hospital would eventually have become infected.

Cemeteries were chosen according to location and time period. Indeed, the study was geographically restricted to Northern Germany and Denmark for organisational reasons. In order to effectuate a thorough investigation, it was also important to gather individuals from before, during and after the recorded decrease in the prevalence of leprosy in the region. Moreover, individuals showing bone lesions suggestive of tuberculosis were also sampled from each cemetery in order to gather information on *M. tuberculosis* which, again, is considered likely to have played a role in the disappearance of medieval leprosy from Europe.

2.3.2 Sample collection

Following the selection of the individuals to be sampled, one tooth from each skeleton was collected in person (in order to choose a non-broken tooth with the best preservation state possible). During the extraction of the teeth, the author was wearing gloves. Immediately thereafter, the teeth were stored in labelled individual DNA-free sample bags to prevent further contamination with modern human DNA. This method allowed the materials to be transported directly (rather than having to rely upon the post with the accordant danger that the material be exposed to X-rays). The documentation of the sampled teeth was conducted alongside the persons in charge of the various collections and was in accordance with their recommendations. The teeth were identified and pictures were taken whenever necessary.

2.3.3 Sample cleaning and pre-processing

All collected samples were given a unique lab ID (see Supplementary material for the lists) and were cleaned for 5min with 15mL of bleach before being rinsed three times with 50mL of bi-distillated DNA-free water. Whenever possible, samples were cut before cleaning and only about 500mg of material was prepared for extraction. After drying overnight at 37°C, the samples were ground for 20s at maximum speed using a ball mill homogenizer. Each sample was individually removed from its container and placed in a metal mill egg under a DNA-free hood. The metal eggs were thoroughly cleaned with soap and dried with pure ethanol between each sample. Under the hood, powder from each sample was recovered from its metal egg and weighed and stored in a 2mL DNA-free LoBind Eppendorf tube. Between 90 and 120mg of material was put into a new clean 2mL Eppendorf tube for DNA extraction. When not immediately in use, the powder tubes were stored at -20°C.

2.3.4 DNA isolation

The powder was decalcified and the collagen digested as follows (see Figure 7): Each aliquot of tooth powder was suspended in 500µL of EDTA (pH8, 0.5M) and 10µL of Proteinase K (0.25mg/mL). The tubes with the suspensions were hermetically sealed with Parafilm and left in the dark overnight at 37°C. A thermomixer was used to regulate the temperature while also preventing the powder from settling down. On the following day, another 20µL of Proteinase K

(0.25mg/mL) was added to each tube and the suspension was incubated on the thermomixer at 56°C for 120-150 min in the dark. Prior to DNA purification, the remaining solid particles were collected at the bottom of the tube by centrifugation for 3min at maximum speed. Two hundred microlitres (200µL) of the supernatant were then pipetted into a new clean tube for DNA isolation. The remaining material from the digestion step was stored at -20°C for possible re-extraction.

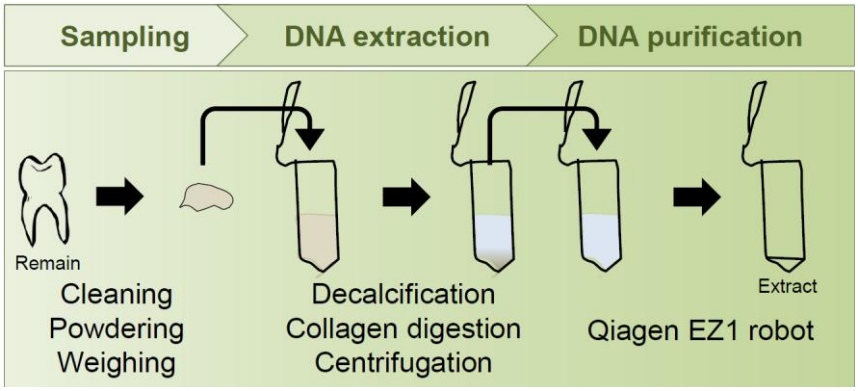


Figure 7: Schematic representation of the DNA extraction from ancient remains
The specificity of aDNA extraction from bones lies in the necessity of first dissolving the bone and collagen matrices.

The DNA purification was performed using a magnetic bead approach on an EZ1 DNA Investigator robot from Qiagen with a DNA Investigator Extraction Kit from the same supplier. Built-in options were set to "Trace" (minute amounts of DNA) with elution in a 50µL TE buffer. The extracts were stored at -20°C. One extraction blank (EB) for each run of the EZ1 robot was added prior to the digestion step, adding up to approximately one blank per five samples.

2.3.5 PCR screening for human and pathogen DNA

PCR screening was performed to identify samples containing *M. leprae* and *M. tuberculosis DNA*. In addition, a screening for human mitochondrial DNA (mtDNA) was conducted as a means to estimate the state of DNA preservation. Indeed, most human cells contain multiple copies of the mtDNA molecule compared to nuclear DNA. MtDNA is, thus, more likely to be present in ancient remains and its amplification can be a good way to discriminate samples with extremely degraded DNA from those samples with aDNA preservation sufficient for further analyses. Each PCR was repeated by a student assistant (Y. Burmeister) to ensure the reproducibility of the results. The PCR reactions were set up in 25µL using Immolase DNA polymerase (Bioline) according to the

supplier's recommendations. The master mix was composed of 1X Bioline Immobuffer (Bioline), 200µM 2.5mM dNTP mix (Bioline), 1.5mM 50mM MgCl2 (Bioline), 4% DMSO, 0.4µM of each primer and 2U Immolase DNA polymerase (Bioline). For the pathogen DNA PCRs, 5µL of template was used. For the mtDNA PCRs, the amount of template was reduced to 1µL. Several PCR blanks (PB) were added for each primer pair by replacing the template by the corresponding volume of the DNA-free water used in the master mix setup.

Table 3: PCR oligonucleotides, specificities and expected PCR product lengths

Target organism	Target region	Primer pair	Annealing temperature (°C)	Product length (bp)
Human mtDNA	15975-16158	mt1	60	183
	16106-16256	mt2	60	150
M. leprae	RLEP repetitive element	LP1 - LP2	52	130
	176239-176336	LP11 - LP12	52	98
M. tuberculosis	IS6110 Insertion Sequence	IS3 - IS4	58	92
	IS6110 Insertion Sequence	Ins1 - Ins2	68	264
	IS6100 Insertion Sequence	P1-P2	68	123
	IS6110 Insertion Sequence	Tb-A - Tb-B	61	123
	IS6110 Insertion Sequence	Tb-C - Tb-D	60	95
	IS1081 Insertion sequence	F-F2 - R-R3	60	113
	IS1081 Insertion sequence	IS1081-F2 - IS1081-F3	58	135

The oligonucleotide sequences and publication references are available in Supplementary Table 4.

The PCR runs consisted of a 10min activation steps at 94°C, followed by 40 cycles of DNA denaturation (94°C for 30s), primer annealing (for which the temperatures are described in Table 3) for 30s and elongation (72°C for 30s). A final elongation step at 72°C for 10min completed the run. The PCR products were kept at 4°C if used within the following 24 hours. Otherwise, they were stored at -20°C.

The results of the PCRs were visualized at the Sanger Sequencing platform at the IKMB by electrophoresis using the Qiaxcel DNA electrophoresis robot together with the QX DNA Size Marker 25–500bp v2.0. PCR products which presented a single clear band of the expected size were labelled as positive (+), products with no clear band were labelled as negative (-) and products which presented many bands were labelled as non-specific (nsp). All PCR products displaying bands were sequenced by the Sanger Sequencing platform of the IKMB and the corresponding chromatograms were compared with the expected target sequence. Target sequences for each of the primer pairs can be found in Supplementary Table 5. In case the

sequenced DNA did not match the expected sequence, a nucleotide BLAST was performed to identify possible contamination sources (see section 3.5.1 The mapping of PCR screening sequences against their genomic references).

2.4 Next-Generation Sequencing

2.4.1 A short description of the Next-Generation Sequencing workflow

The library preparation protocol is adapted from published methods (Kircher et al. 2012; Meyer & Kircher 2010). Briefly, all aDNA molecules from an extract are ligated to known artificial adapters (IS1, IS2 and IS3 from Figure 8) to create a library. That library is then identified using a unique combination of two molecular barcodes and amplified with primers which target the adapters. For the full list of kits, reagents, instruments and suppliers, see 12.1 Supplementary material. Two different libraries were prepared for each sample: one treated to remove Uraciles (UDG-treated) and one without treatment (NT) to preserve the damage patterns.

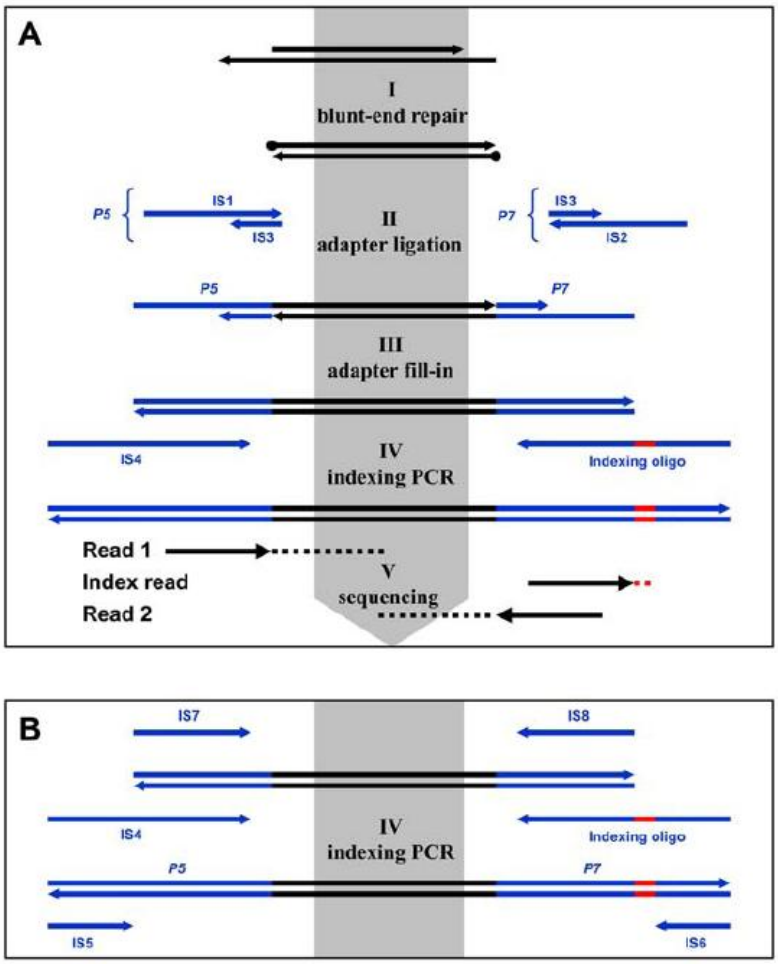


Figure 8: Overview of Next-Generation Sequencing library preparation
Although it would later be improved (Kircher et al. 2012), this diagram as well as the first description of the method dates from 2010 (Meyer & Kircher 2010).

To check for possible contamination, a minimum of one library negative control (LNC) per run was included for each set of libraries prepared. Moreover, the EB from the extraction step were identically processed and converted into libraries (LEB). Finally, all libraries (including LNCs and LEBs) were first shotgun sequenced in pools before a re-sequencing of those samples which showed promising results. The Next-Generation Sequencing platform performed the quality controls of the indexed libraries after amplification and the sequencing and the demultiplexing according to the sample barcodes.

2.4.2 Illumina sequencing library preparation

Ancient DNA extracts were converted into DNA libraries as follows: Overhanging 5'- and 3' -ends were removed to allow the ligation of P5 and P7 adapters to both ends of the ancient DNA fragments. Then, the adapters were filled in using Bsm Polymerase.

The blunt-end repair for non-UDG treated libraries was performed in a 50µL reaction mix containing 20µL of aDNA extract, 1X NEB Buffer 2, 300mM dNTPs mix, 0.8mg/mL BSA, 1mM ATP, 20U T4 Polynucleotide Kinase and 1U T4 Polymerase. The solution was incubated for 30min at room temperature (RT). For UDG-treated libraries, the reaction mix contained (aside from the aDNA extract) 1X NEB Buffer 2, 300mM dNTPs mix, 0.1mg/mL BSA, 1mM ATP, 20U T4 Polynucleotide Kinase and 3U USER Enzyme. A first three hour incubation step at 37°C was performed before adding 1U T4 Polymerase to the mix and following up with the normal 30min incubation at RT. MinElute purification into 18µL 1x TE Buffer was immediately conducted as described in the manufacturer's procedures (See 12.2 Supplementary methods). The library Adapter Mix was prepared according to the published protocol (Meyer & Kircher 2010). The ligation of the adapters was conducted by adding 21µL of a mix containing 1X Quick Ligase Buffer and 2.5µM Adapter Mix to the 18µL purified blunt-ended DNA molecules. 1µL of Quick Ligase was added at the end to prevent the formation of adapter chimeras. The reactions were incubated for 20min at RT and MinElute purification into 20µL 1x TE Buffer was immediately conducted, as described in the manufacturer's procedures. Finally, the fill-in step was conducted in a 40µL reaction containing the 20µL purified ligated DNA, 1X Thermopol Buffer, 125nM dNTPs mix and 16U Bsm Polymerase. Libraries were incubated for 20min at 37°C for the fill-in reaction to take place, followed by a 20min deactivation step at 80°C. The libraries were then stored at -20°C until the indexing PCR.

2.4.3 Library indexing, amplification and quality assessment

The indexing PCR of the libraries consisted of 5'- and 3'- tailed primers containing both a 6nt-long molecular barcode and general Illumina sequencing adapters (Kircher et al. 2012; Meyer & Kircher 2010). Specifically, 10µL of each library was added to a 40 µL mix containing 1X AccuPrime Reaction Mix, 0.3µM of each primer and 2U of AccuPrime Pfx Polymerase. For each library, a unique combination of two indexing primers was chosen and the reaction was performed four times to index the complete volume of library while avoiding possible PCR inhibition. The PCR

reaction itself consisted of a 10min activation step at 95°C, followed by nine cycles of 15s denaturation at 95°C, 30s annealing at 68°C and 32min extension at 68°C. The number of cycles was kept low to prevent the synthesis of adaptor chimeras when reaching the PCR plateau phase. MinElute purification of all 4 PCR products into a single 50µL 1x TE Buffer aliquot was immediately conducted as described in the manufacturer's procedure (see 12.2 Supplementary methods). When not amplified on the same day, indexed libraries were stored at -20°C.

Amplification of the indexed libraries took place in four parallel 50µL PCR reactions containing 5µL of template indexed library DNA, 1X AccuPrime Reaction Mix, 0.4µL of standard Illumina primers IS5 and IS6 and 20U/µL AccuPrime Pfx Polymerase. The IS5 and IS6 oligonucleotides target the common part of the indexing primers and, thus, are used to amplify all indexed library DNA fragments as a whole. MinElute purification of all four PCR products into a single 50µL 1x TE Buffer aliquot was immediately conducted as described in the manufacturer's procedures. When not quantified on the same day, amplified indexed libraries were stored at -20°C.

Quantification and quality control of the amplified indexed libraries, LNCs and LEBs was performed at the ZMB on Agilent DNA1000 bioanalyzer or Qiagen TapeStation according to the manufacturers' protocol. The DNA concentration in the amplified indexed library was estimated and library quality was checked in two different ways. First, as aDNA fragments are usually very short, this results in the presence of DNA constructs which are shorter than 300bp in the indexed library. Any concentration of longer molecules, therefore, is likely to be a result of contamination. The second verification concerns the presence of adapter chimeras. Narrow peaks of concentration for DNA fragments below 170bp usually originate from library and/or indexing adapter chimeras and are not related to contamination (Kircher et al. 2012; Meyer & Kircher 2010). Adapter chimeras are artificial DNA fragments produced during the library preparation and indexing when the relative proportion of adapters to insert DNA is too high in a reaction mix (Kircher et al. 2012; Meyer & Kircher 2010). Chimeras present little risk for the sequencing of blanks on a Illumina HiSeq, because the cluster generation step requires DNA fragments to be longer in order to properly form bridges (Kircher et al. 2012; Meyer & Kircher 2010). Amplified indexed libraries showing low concentrations were consequently re-amplified using the same PCR protocol, in order to obtain enough material for subsequent sequencing.

2.4.4 Pooled shotgun sequencing

For sequencing, pools of indexed libraries were designed to contain eight samples, each identified with a unique pair of sequencing indices. In addition, the forward and reverse indices were present only once in each pool to minimize the risks of cross-talk during sequencing and the percentages of each base in the indices were checked to avoid over-representation of one of the bases and the saturation of the sequencer's camera. The technicians in the ZMB then prepared the pools with equimolar amounts of the eight samples. Non-UDG treated libraries and UDG-treated libraries were sequenced separately on different Illumina Hi-Seq 2000 sequencing lanes in Single Flow Cell mode with 2x100bp read length. One flow cell under those conditions can yield up to 2 billion single-end reads. Consequently, one lane is expected to yield up to 250 million paired-end reads. Therefore, each of the eight samples sequenced on a lane is expected to yield up to about 25 million paired-end reads.

2.4.5 Re-sequencing

After a first analysis of the pooled sequencing run datasets, some samples showed high library complexity rates and only partial coverage for the *M. leprae* genome, suggesting that the re-sequencing of larger amounts of sequencing libraries would yield a larger number of informative reads mapping to the *M. leprae* genomes. For those samples (see section 4.4), the UDG-treated libraries were re-sequenced to increase the coverage and read depth on the reference genomes and to allow for a more reliable analysis of the recovered pathogen genome. Indeed, deeper sequencing decreases the proportion of artifactual sequence variants and the number and size of gaps in the draft genomes. The re-sequencing was run with virtually one Illumina Hi-Seq 2000 lane per sample. In reality, to maintain the cluster generation and sequencing quality, each sample was sequenced four times on $1/4^{th}$ of a lane. The four datasets generated were then pooled together after demultiplexing.

2.4.6 Additional datasets

Several samples originally collected for this study were later included in other projects and sequenced using different strategies. Most frequently, the same libraries generated for this study were enriched for human DNA (by L. Möbus in the context of her Master thesis and by L. Böhme as part of her PhD) As enrichment approaches are rarely fully specific and selective, numerous sequencing reads from those enriched libraries did not match the target species but could still be used to study other organisms. Consequently, sequencing datasets generated for side-projects were merged with the ones obtained during this study to maximise the number of informative reads which mapped to the genomes of the pathogens of interest.

2.5 Preliminary data analysis

2.5.1 The mapping of PCR screening sequences against their genomic references

In the *M. leprae* TN and the *M. tuberculosis* H37Rv references genomes, the sequence of each expected PCR product was extracted (Supplementary Table 5). All PCR products displaying bands were Sanger sequenced at the IKMB. DNA fragments were sequenced from both ends; forward and reverse reads were merged using the Geneious software and aligned to the corresponding target sequence. All chromatograms and alignments were manually checked to identify base calling or alignment errors. In case the sequenced DNA did not match the expected sequence, a nucleotide BLAST was performed to identify possible contamination sources (Altschul et al. 1990). Settings were set to "Optimize for more dissimilar sequences (discontiguous megablast)". All other settings were left as default.

2.5.2 Pathogen reference genomes and Next-Generation Sequencing target sequences

One reference genome per genome was downloaded from the NCBI databank. The full list of reference genomes with accession numbers is available in Table 4. In those cases in which several reference genomes were available for a species, the most commonly used or most recent assemblies were chosen. In addition to *M. leprae TN*, *M. lepromatosis FJ924*, *M. tuberculosis H37Rv* and *M. bovis AF2122/97,* all other mycobacterial reference genomes were included in the

dataset, along with the more distant relative *Y. pestis (C092)*. Indeed, *Y. pestis* is the human pathogen which caused the Black Death. It shares some limited genetic similarity with mycobacterial genomes and is archaeologically unlikely to have been present in the samples under study. It was chosen, therefore, as an outgroup to evaluate the baseline cross-species' non-specific mapping.

Table 4: List of reference sequences, strains and NCBI accession numbers

Species	Reference strain	RefSeq accession number (NCBI)
Mycobacterium abscessus	ATCC 19977	NC_010397.1
Mycobacterium africanum	GM041182	NC_015758.1
Mycobacterium avium	paratuberculosis K-10	NC_002944.2
Mycobacterium bovis	AF2122/97	NC_002945.3
Mycobacterium canettii	CIPT 140010059	NC_015848.1
Mycobacterium leprae	TN	NC_002677.1
Mycobacterium lepromatosis	FJ924	NZ_LAWX01000032.1
Mycobacterium marinum	M	NC_010612.1
Mycobacterium smegmatis	MC2 155	NC_008596.1
Mycobacterium tuberculosis	H37Rv	NC_000962.3
Yersinia pestis	C092	NC_003143.1

Due to the high levels of genetic conservation between the members of the *Mycobacteria* genus, a large number of reads could be expected to map to any *Mycobacteria* genome and prevent the accurate determination of the infective species in the samples under study. To reduce the percentage of non-species-specific reads mapping, “target” genomic regions characteristic of a species were extracted from each reference genome to facilitate the species identification step. To this end, the bacterial reference sequences were aligned to each other using the Progressive MAUVE algorithm with default settings. The genomic multiple alignment was used to evaluate the sequence conservation levels between the mycobacterial reference species and their distant *Y. pestis* relative.

Finally, regions present in only one of the genomes were consequently identified and extracted from the alignment backbone file using AWK. Those species-specific regions were then ordered to identify the two longest ones for *M. leprae*, *M. lepromatosis*, *M. bovis* and *M. tuberculosis.*

2.5.3 Pre-processing of reads

The demultiplexing of the reads according the adapter barcodes was performed by the NGS sequencing platform (M. Schilhabel). The resulting raw files were saved in the FastQ format and were subsequently copied through FTP to the aDNA group working directories on the Rechenzentrum clusters. The raw reads were processed using the EAGER (Efficient Algorithms for Ancient Human Genome Reconstruction (Peltzer et al. 2016)) pipeline which incorporates published bioinformatic softwares into a user-friendly framework optimized for aDNA analysis. The softwares are implemented in the form of modules that can be called and set up independently through a graphical user interface (GUI). The pipeline allows each module to be performed one after the other by automatically converting or preparing the output files from one step to fit input files requirements for the next step. The pipeline is described in more detail in 12.2 Supplementary methods. To start with, all sequencing results were analysed using FastQC to evaluate the quality of the sequencing run and to identify possible problems during library preparation. Module 2 from the EAGER pipeline was used to call for the FastQC software. The quality controls were performed with default settings used for the evaluation of modern DNA sequencing runs. As a result, warnings or errors were necessarily interpreted with regards to aDNA molecular specificities.

Raw sequencing reads needed to be prepared for mapping against the chosen reference genomes. All Illumina sequencing runs were performed as paired-end, creating two raw read files for each sample: a forward FastQ file and reverse FastQ file. Therefore, forward and reverse read files had to be merged to improve the mapping step. Moreover, since aDNA fragments are usually short, it is common to find adapter sequences at the end of the sequenced reads. If not removed, those adapter sequences might impair the rest of the analysis. The clipping of adaptors, the merging of the forward and reverse reads as well as a subsequent quality filtering step were performed by Module 3 from EAGER.

2.5.4 Species identification

For each UDG-treated library, the sequencing reads from the pooled samples were mapped against four mycobacterial reference genomes (*M. leprae*, *M. lepromatosis*, *M. tuberculosis* and *M. bovis*) as well as against the *Y. pestis* reference. The reads which mapped to each reference with low quality were then filtered out and PCR duplicates were removed. The analysis was

performed on the two libraries available for each sample (normal and UDG treated). For the non-UDG treated mapping reads, the coverage statistics were calculated for the whole reference as well as for the reference-specific regions. Indeed, if the reads are authentic DNA fragments from the organism, the coverage and read depth on the whole genome and on the target region should be consistent or increased. If the target region shows a significantly lower coverage, this would suggest that a large part of the reads actually originated from another (related) organism. For comparison, the sequencing reads from the same datasets were mapped against the human mtDNA reference to provide an estimate for the recovery of human aDNA.

2.5.5 Ancient DNA authentication using damage patterns

Ancient DNA typically displays short fragment lengths and increased occurrence of G and T towards the 3'-end and 5'-end of the DNA strand, respectively (Briggs et al. 2007; Pääbo 1989; Overballe-Petersen et al. 2012; Brotherton et al. 2007; Lamers et al. 2009). Those characteristic modifications depend not only on the age of the DNA molecules but also on numerous other environmental factors, such as climate or burial conditions. Therefore; the study of the DNA fragments' molecular characteristics (damage patterns) can differentiate between a young DNA sample (below 20yo) and an ancient DNA sample, although the molecular age of the sample cannot be used as a proxy for the estimation of the actual age of a sample (Sawyer et al. 2012; Overballe-Petersen et al. 2012; Seguin-Orlando et al. 2015).

Damage patterns are computed directly from NGS data obtained from non-UDG treated sequencing libraries (Molak & Ho 2011; Krause 2010; Ginolhac et al. 2011; Sawyer et al. 2012). The length distribution of the mapping reads as well as the frequencies of misincorporated bases can be calculated for a specific reference sequence (Briggs et al. 2007; Ginolhac et al. 2011; Jonsson et al. 2013). In this study, Module 9 from the EAGER pipeline (mapDamage2.0, see 12.2 Supplementary methods) (Ginolhac et al. 2011; Jonsson et al. 2013) was used to determine the damage patterns for all sequenced samples. Damage patterns were analyzed for each sample with every reference used for mapping. The frequencies of misincorporation over the first (3'-end) and last (5'-end) base of the fragments for each reference were compared to published data to determine the authenticity of the results and the reliability of the validation method. Fragment length distributions were used as complementary information to estimate the preservation of the molecules of interest.

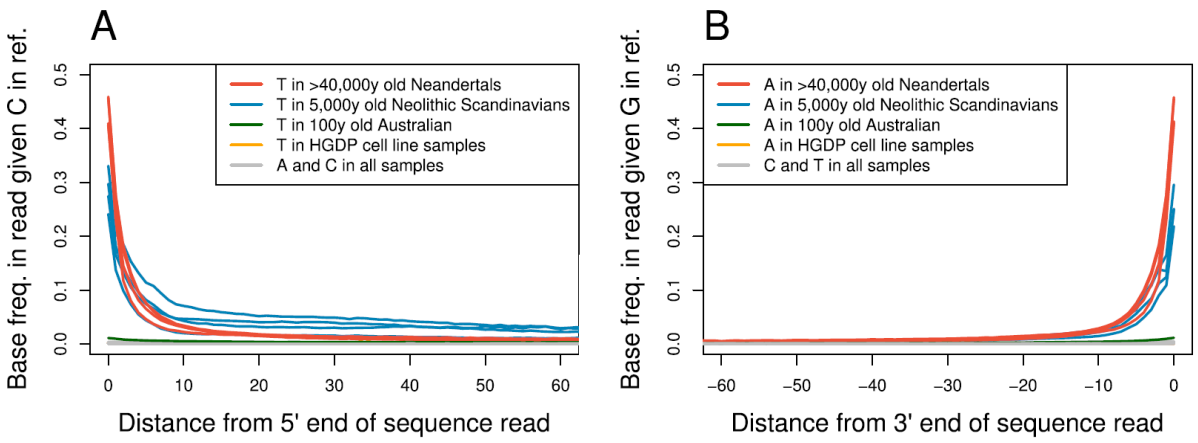


Figure 9: Example of misincorporation patterns in ancient DNA
A) C to T misincorporation frequency in samples from various ages. B) G to A misincorporation frequency in samples from various ages. HGDP cell line DNA was used as a modern control (Skoglund et al. 2014). MapDamage results possess a similar layout; a sample highly contaminated with modern sequences will show no increase in misincorporation frequency toward the ends of the reads.

2.5.6 Statistical tests

Wherever possible, basic statistical tests were performed on the results to guage the significance of the variations observed in the data distribution. Therefore, the null hypothesis is the identity between distributions. All tests described below and the graphical representations of the results were performed using Microsoft Excel as well as the software R (R Core Team 2008). Depending on the results, the proper Student/ Wilcoxon test was chosen to test for distribution variation significance.

2.6 Mycobacterial plasmid analyses

All the UDG-treated reads were mapped against the reference sequences of major mycobacterial plasmids. This was performed to see if a mycobacterial plasmid could be found associated with *M. leprae* in the medieval DNA samples. The list of reference sequences is available in Table 5. The coverage statistics were then compared between the genome of the modern species carrying the plasmid, the plasmid itself and the ancient *M. leprae*. The mapping of the sequencing reads to the reference sequences was performed using the Bowtie pipeline.

Table 5: Names and accession numbers of the plasmid reference sequences

Species	Plasmid name	Accession number (NCBI)
M. tuberculosis	pTYGi9	NC_025025.1
M. avium	pMAH135	NZ_AP012556.1
	pVT2	NC_005016.1
M. chubuense	pMYCCH01	NC_018022.1
	pMYCCH02	NC_018023.1
M. celatum	pCLP	NC_004963.1
M. kansasii	pMK12478	NC_022654.1
M. smegmatis	pMYCSM01	NC_019957.1
	pMYCSM02	NC_019958.1
	pMYCSM03	NC_019959.1
M. marinum	pMM23	NC_010604.1
	pMUM003	NC_019018.1
	pRAW	NZ_HG917973.1
M. yongonense	pMyong1	NC_020275.1
	pMyong2	NC_020276.1
M. gilvum	pMFLV01	NC_009339.1
	pMFLV02	NC_009340.1
	pMFLV03	NC_009341.1
	pMSPYR101	NC_014811.1
	pMSPYR102	NC_014812.1
M. lilfandi	pMUM002	NC_011355.1
M. ulcerans	pMUM001	NC_005916.1

Only mycobacterial plasmids clearly documented as part of a Mycobacterium *genome were taken into account. Several plasmids from unknown mycobacterial origin were not selected for this study, as it would have been virtually impossible to rule out contamination.*

2.7 Human DNA analyses

All the merged sequencing reads were mapped against the human GRCh38 mitochondrial DNA (mtDNA) sequence (NC_012920.1), X and Y chromosomes (NC_000023.11 and NC_000024.10, respectively) using the EAGER pipeline as well as modules alpha (α) to delta (δ) (see 12.2 Supplementary methods).

Those sequencing reads mapping to the human genome were mainly used to confirm the age of the remains through the use of damage patterns (see section 2.5). In addition, the number of reads mapping to chromosomes X and Y were analysed to confirm the anthropological sexing of the remains.

A female individual displays two copies of the X chromosome, while a male individual one copy of the X and one copy of the Y chromosome. Therefore, comparing the coverage statistics on the X and Y chromosomes should provide a good proxy for sex determination. The mapping statistics against the X and Y chromosomes were compared and the X/Y coverage ratio was used as a proxy for sex typing (see 12.2 Supplementary methods). The results were compared with the sex-typing results as completed by osteological observation (data from the collaboration partners, see 12.11 List of samples) or forensic STR typing (data from L. Böhme as part of her PhD).

2.8 *Mycobacterium leprae* data analysis

2.8.1 Assembly of the ancient genomes

All the reads available for each UDG-treated library originating from separate sequencing runs were pooled in order to maximise the recovery of the targeted ancient sequences. The final genome assemblies for the ancient *Mycobacteria* genomes were performed for each of the collected datasets by Dr. M. Nutsua. To improve the mapping results, the raw reads were mapped against the *M. leprae* TN reference genome using an in-house approach similar to EAGER with the caveat that BWA was replaced by Bowtie (Dr. M. Nutsua, see 12.1 Supplementary material). Indeed, BWA as available through EAGER only allows for a fixed number of mismatches between the read and the reference. Therefore, it is difficult to set a stringent mismatch cut-off for datasets with reads of various lengths. Bowtie allows the number of mismatches to be set as a percentage of the read length, which optimises the mapping of long reads while decreasing the risk of aspecific mapping of small reads. The mapping assemblies were generated from the read alignment files as described in Schuenemann et al., 2013. The Genome Analysis Toolkit (GATK) (McKenna et al. 2010; Van der Auwera et al. 2013) was used to call reference bases and variants from the mapping (module UnifiedGenotyper) according to three possibilities:

1) Call quality ≥ 50 & read depth ≥ 5 & fraction of mapping reads with the variant ≥90%: the SNP was called.
2) Call quality ≥ 50 & read depth ≥ 5 but fraction of mapping reads with the variant <90%: the reference base was called.
3) Neither 1) or 2) criteria are reached: the base was called as unknown (“N”).

2.8.2 Single nucleotide polymorphism typing of ancient *M. leprae* genomes

GATK UnifiedGenotyper was used to call all *M. leprae* genomic positions confidently covered (EMIT_ALL_CONFIDENT_SITES option) after removing the PCR duplicates from the mapped reads BAM file. All positions not listed in the output VCF file were subsequently considered as not covered (NC). Finally, a filter was applied to remove all positions with quality scores under 50. Therefore, the final variant list discriminates between positions with a variant, positions without variant (reference base) and positions which were not covered (or of insufficient quality).

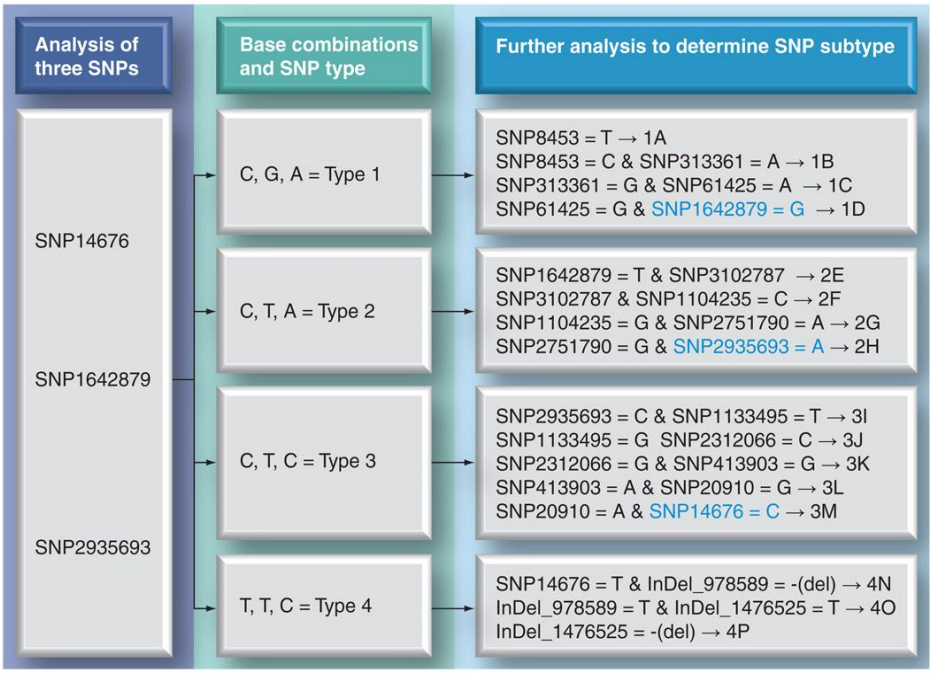


Figure 10: Single nucleotide polymorphism-typing for Mycobacterium leprae
Using the first three SNPs shown on the chart, the a strain is first typed into one of the four SNP types (1–4) and then subtyped using the three or four markers shown at right/SNP type to give 16 subtypes (A–P). SNPs shown in blue are the same as those used for typing into types 1, 2, 3 & 4 (at left). InDel: Insertion-deletion; SNP: Single nucleotide polymorphism. From (Singh & Cole 2011).

Genomic positions involved in *M. leprae* genotyping were isolated from the filtered VCF files using a grep-based script prepared by the author for that specific purpose (Monot et al. 2005; Monot et al. 2009; Singh & Cole 2011). SNP types were inferred via two different approaches. First, the number of variants consistent with each genotype was calculated. The type with the highest number of consistent variants was established as the most likely SNP type (Monot et al, 2005). The second approach resembles an inference method (Sing et al, 2011). Only discriminating

positions were sought after; the SNP type was established without confirming the alleles of the other positions (see Figure 10).

2.8.3 Phylogenetic analyses

Phylogenetic trees were built as described in previous publications (Schuenemann et al. 2013; Mendum et al. 2014) with the help and support of Dr. M. Nutsua. The approach utilized made use of regions that were covered in all the genomes used in order to identify positions where variations could be found and to calculate the position of each genome on a phylogenetic tree. A multiple genome alignment of the six high-coverage ancient *M. leprae* sequences recovered in this study was constructed. In addition, the sequences from the study by V. Schünemann and colleagues were added to further refine the resolution of the trees (four *M. leprae* reference genomes, seven modern strains and five ancient strains). One of those previously published strains originated from the St. Jørgen cemetery, a site which was also sampled as part of the present study. Two whole genome multiple alignment softwares were used for comparison purposes: the progressiveMauve algorithm (Darling et al. 2010) and the MAFFT program (Katoh & Standley 2013).

From the multiple alignments, Mauve's SNP calling function was used to generate a table of all alignment columns where at least one of the strains contained a SNP. The concatenation of these alignment columns was then used as input for MEGA5 (Tamura et al. 2011) to construct Maximum Parsimony, Neighbour-Joining (Saitou & Nei 1987) and Maximum Likelihood trees (Schuenemann et al. 2013). To obtain rooted trees for all three reconstruction methods, *M. avium* 104 (NC_008595.1) was included as the out-group. Bootstrap values (Felsenstein 1985) were inferred from 500 replicates. The MP tree was obtained using the Close-Neighbour-Interchange algorithm (Nei & Kumar 2000) as described in Schuenemann et al., 2013. To assess the best model to be used for the evolutionary distances used for the Neighbour-Joining tree and the evolutionary model used for the Maximum Likelihood tree, the author and M. Nutsua used Mega's model test and chose the results with the highest BIC value. For both alignments (with and without the outgroup sequence), the Tamura 3-parameter model (Tamura 1992) and uniform rate proved the best model for all sites. All positions with less than 90% site coverage were eliminated (see Schuenemann et al. 2013 for more details about the method). The MAFFT alignment as well as the tree constructions were realized by Dr. M. Nutsua.

Regardless of the alignment software used, the inherent inconvenience of the method is that low-coverage samples must be discarded in order to maintain a sufficient number of informative SNPs within the multiple alignment of the genomes. With ancient DNA, this inconvenience becomes a major hurdle, especially as DNA preservation is often not good enough to permit the recovery of genomes with high coverage. To avoid losing too much information by discarding all the low-coverage genomes, a method was developed to allow for the placement of as many of the low-coverage genomes into the previously-constructed phylogenetic trees as is possible. First, progressiveMAUVE was used to generate a multiple alignment of all the ancient genomes recovered in this study. Then, based on the genetic distances between the *M. leprae* isolates and the pre-determined location of the high-coverage genomes within the tree topology, the low-coverage genomes were assigned to the branch of the two closest high-coverage genomes. The assignment was considered significant when identical between the two high-coverage anchors used. After assigning a low-coverage genome to a branch, all ancient and modern strains belonging to that branch were re-aligned in order to produce a branch-specific list of variants. Finally, this list was used as described above to construct branch-specific trees and to place the low-coverage genomes.

2.8.4 Evaluation of the effects of the variations observed

All SNPs occurring in at least one strain as identified with GATK (see section 2.8.3 Phylogenetic analyses) were analysed with respect to their effect on annotated genes. The SNPS were annotated using the software snpEff (Schuenemann et al. 2013; Cingolani et al. 2012). An annotation database of the *M. leprae* TN genome was constructed from the NCBI genomic reference. The up/ downstream region size parameter for reporting SNPs that are located upstream or downstream of protein-coding genes was set to 100 nts. Default parameters were used for all other settings. The results were compiled into a table containing information for each SNP regarding its effect on the genes in the strains in which the SNP occurs. The SNP distribution was plotted to evaluate the density of SNPs along the genome and to identify possible variations hotspots or conserved regions. Moreover, the types, effects and annotations of the SNPs were analysed in correlation with each other to highlight possible in-depth SNP analysis approaches.

Due to the impossibility of growing *M. leprae* as a culture, *in vivo* phenotypic studies confirming the *in silico* annotations are rare. Therefore, most of the predicted gene functions are based on comparisons with *M. tuberculosis* and numerous genes remain annotated as hypothetical

proteins whose functions are unknown. Variants for which there is enough information available to perform *in silico* evaluation of the variant effects were selected by filtering out the effects with the following annotations:

- Intergenic, Upstream, Downstream: The understanding of *M. leprae* intergenic regions is insufficient at present to estimate the effects of the intergenic variants on the expression of other genes.
- Hypothetical protein & pseudogene: The information available, if any, does not provide sufficient background knowledge of the transcripts.
- Stable RNAs: The variants in annotated stable RNAs were removed so that the study could focus on protein variants.
- Variants present on only one genome: This filtering step was performed to focus the study on common variants.

2.8.5 Case-study: detailed polymorphism effect estimation

Three genes were selected for in-depth analysis of the possible effects of the observed variants on their function. The selection was performed as follows. First, the functional profile of each gene was completed with GO annotations, the evolutionary rate amongst *Mycobacteria* and the number of InterPro domains. Then, genes with synonymous variants or unknown GO cellular components were filtered out. Finally, the variants were sorted according to protein length, number of InterPro Domains and evolutionary rates in order to select the longest protein gene, the gene with the highest number of InterPro domains and the best conserved gene, respectively.

Several aspects of the protein properties and functions were studied and the possible effects of their variants analysed by comparison with the reference sequence. For the purpose of this comparison, the protein with the amino acid reference was called the reference type (RT). First, the proteins' general chemical properties (pI, mW, aliphatic index, average hydrophobicty) were calculated using several online tools available through the SIB bioinformatics resource portal ExPasy (see list in 12.2.4 Bioinformatic pipeline descriptions). Then, the ProtScale tool of the same platform was used to compute the polarity, hydrophobicity and buried residue profiles of each protein, RT and variants to investigate the local impact of the amino acid change. In addition, the secondary structure of the protein was predicted using three different prediction tools (see list in 12.2 Supplementary methods) with and without amino acid change to evaluate the influence of the variant on the formation of helices and sheets. Finally, the amino acid change *loci* were identified within the protein-annotated regions using the UniprotKB and NCBI protein database to

investigate the active regions that might be influenced by the variants. Using the UniprotKB built-in Clustal Omega protein alignment server, the reference protein sequence was aligned to all the other available proteins with 100% similarity from UniRef (Bateman et al. 2015; Suzek et al. 2015; Sievers et al. 2011). In cases when the 100% cluster only contained the reference sequence, the similarity threshold was decreased to 90% (dnA and glcB). Then the protein sequence with variants was added to the alignment in order to investigate the influence of the variant on conserved patterns.

3 Results

3.1 PCR screening for human and pathogen DNA

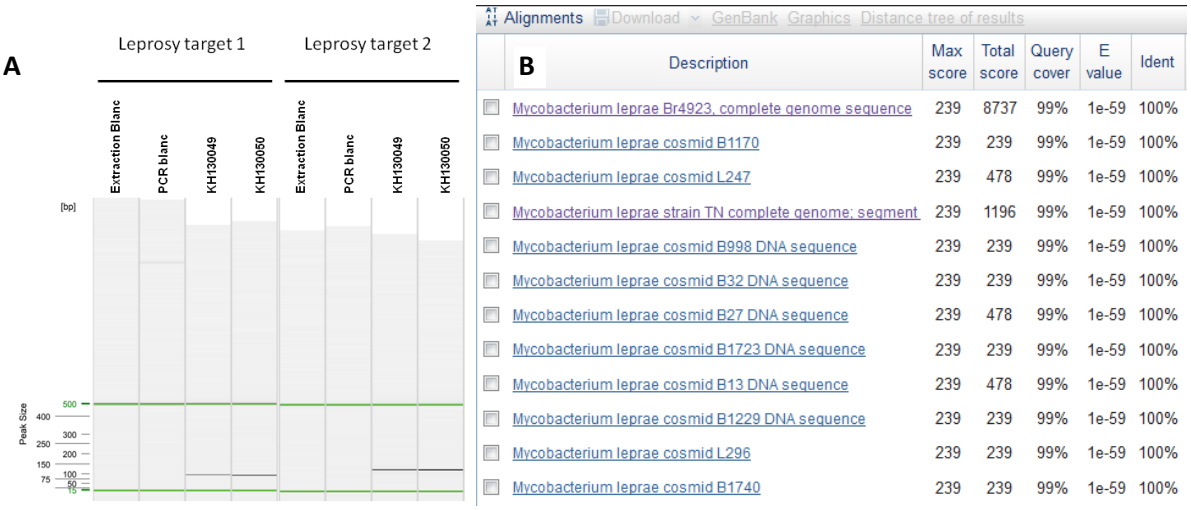
All extraction and PCR negative controls gave no PCR products for the primer pairs used. Most cemeteries showed quite good human mtDNA preservation, with at least 30% of the samples yielding the expected PCR products. In comparison, no sample yielded positive results for *M. tuberculosis* DNA, a result which will be discussed in more detail below. The number of samples which showed positive results for *M. leprae* DNA was null for all but a single cemetery (St. Jørgen, Denmark). The preservation level for *M. leprae* at St. Jørgen cemetery was as high as for human mtDNA; almost all samples which were positive for mtDNA also yielded *M. leprae* DNA (see 12.3 Supplementary results for further detail).

Table 6: PCR screening results summarized by cemetery

		PCR screening results		
Site	Number of samples	Number of leprosy positives	Number of tuberculosis positives	Number of human mtDNA positives
St. Jørgen	34	16	0?	32
Dagmargården	27	0	0?	20
Rathausmarkt	79	0	0	25

A sample is considered positive when the PCR products of the expected length and sequence were obtained with at least two different primer pairs. Detailed results per sample and primer pairs as well as the sequences obtained with each PCR are available in Supplementary Table 5.

The alignments of the sequenced *M. leprae* and human mtDNA PCR products to the target reference sequences showed results consistent with the electrophoresis observations (see Figure 11). For *M. tuberculosis*, several aspecific amplifications yielded multiple products of highly various lengths (see Figure 12). A total of seven different primer pairs were tested, each with two annealing temperatures. The same phenomenon was observed with all different primer pairs and with no consistent correlation to the sample, PCR setup or reagent batch used. Sequencing and BLAST search of the sequences showed hits in numerous bacterial species, consequently preventing the identification of common contamination sources. Although no sample could be clearly identified as tuberculosis positive, three samples showed consistent PCR products in two of the PCRs tested.



C

Target 1

Reference TGTCGGCGTGGTCAATGTGGCCGCACCTGAACAGGCACGTCCCCGTGCACGGTATAACTATTCGCACCTGATGTTATCCCTTGCACCAT
KH130050 TGTCGGCGTGGTCAATGTGGCCGCACCTGAACAGGCACGTCCCCGTGCACGGTATAACTATTCGCACCTGATGTTATCCCTTGCACCA

Target 2

Reference TAGAACAAATAGGGTGGTTCTGCTTCTATTGCACCGACCAACAGTAGGAATGGTCTGA
KH130050 TAGAACAAATAGGGTGGKTCTGCTTCTATTGCACCGACCAACAGTAGGAATGGTCTGA

Figure 11: Example of leprosy PCR result
A) Gel electrophoresis of the PCR products: the length is consistent with the expected product. B) BLAST results when using the product sequence as query to search the complete genomes database: the first hit is M. leprae. *C) Pairwise alignment of the PCR product sequence and the reference genome target.*

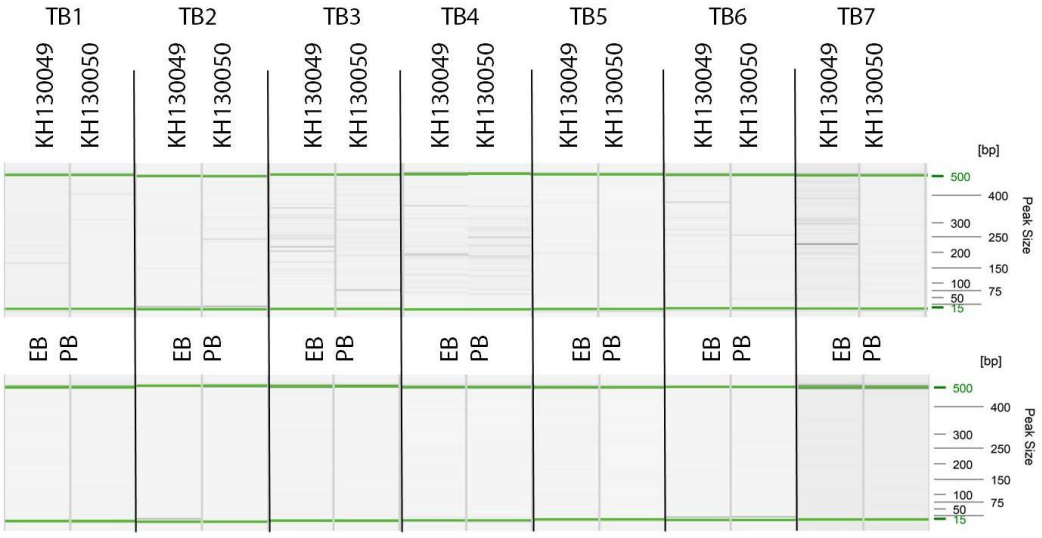


Figure 12: Example of tuberculosis PCR result
Gel electrophoresis of the PCR products. For ease of visibility, pictures were cut and blanks displayed on a lower line. Many aspecific products are observed, while no band shows a size consistent with the expected product.

3.2 Illumina sequencing library concentration and quality

All the indexed libraries were quantified on the Agilent bioanalyzer or the Qiagen TapeStation at the ZMB to evaluate not only the DNA quantity but also the general library quality before Illumina sequencing. Library blanks and extraction blanks were also quantified and showed no evidence of contamination. In addition, the concentration plots were checked for length distribution and the peaks which would indicate the presence of adapter chimeras. These results are summarized in Supplementary Table 7. The figures below are examples of the quantification plots for a sample and a blank. The peak in concentration for the sample is clearly below 300bp-fragments. The sample shows only few adapter chimeras (the peak is around 100bp). The blank shows no DNA apart from the lower and upper ladder bands and some adapter chimeras.

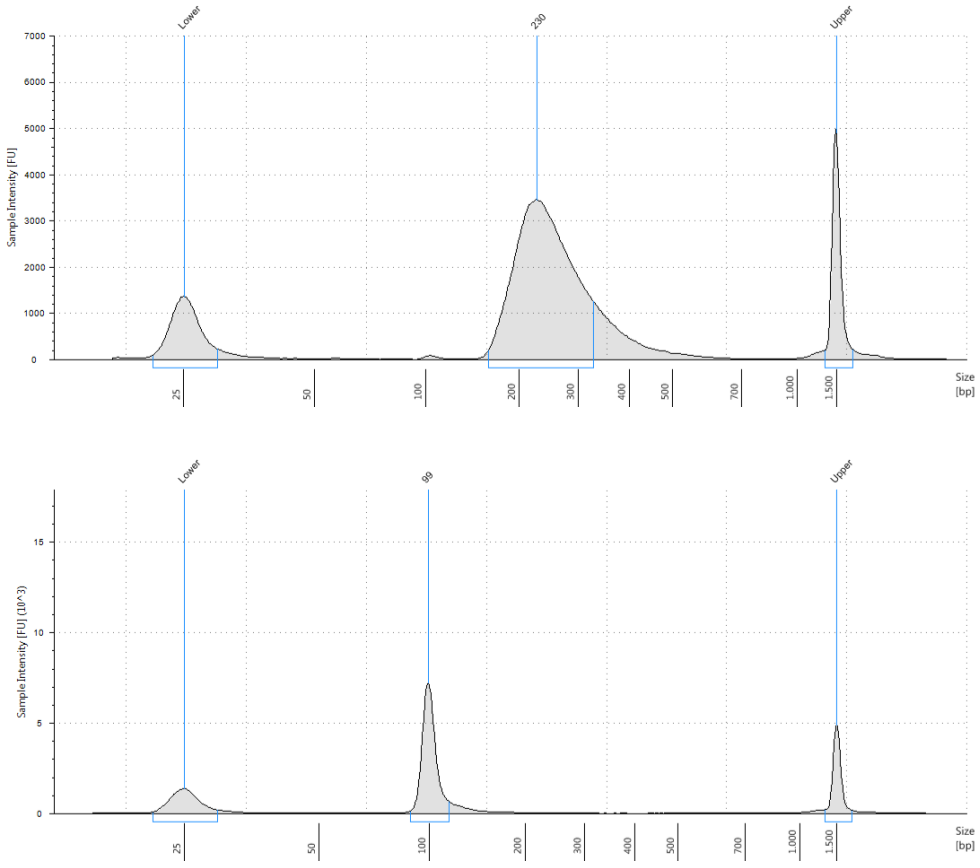


Figure 13: Example of quantification plots for a sample and a blank
UP: Sample SJG978 UDG-treated library. DOWN: UDG-treated library from extraction blank.

3.3 Preliminary analyses findings

3.3.1 Comparison of the reference genomes

The results of the Mauve multiple alignments of the *Mycobacteria* genomes were used to produce the genetic conservation distance matrix shown in Table 7.

Table 7: Genome conservation distance matrix between the Mycobacterium *genomes*

	M. lepromatosis	*M. leprae*	*M. bovis*	*M. africanum*	*M. tuberculosis*	*M. canettii*	*M. avium*	*M. abscessus*	*M. marinum*	*M. smegmatis*	Genome sizes (Mb)
M. lepromatosis		0.70	0.43	0.43	0.43	0.43	0.42	0.39	0.43	0.40	3.22
M. leprae			0.44	0.44	0.44	0.44	0.43	0.38	0.44	0.41	3.27
M. bovis				0.99	0.98	0.96	0.70	0.57	0.68	0.65	4.35
M. africanum					0.99	0.96	0.70	0.57	0.68	0.65	4.39
M. tuberculosis						0.96	0.70	0.57	0.68	0.65	4.41
M. canettii							0.69	0.57	0.67	0.64	4.48
M. avium								0.62	0.72	0.74	4.83
M. abscessus									0.60	0.64	5.07
M. marinum										0.69	6.64
M. smegmatis											6.99
Y. pestis	0.16	0.16	0.17	0.17	0.17	0.17	0.16	0.09	0.20	0.18	4.65

This matrix was obtained after multiple alignment of the genomes using Mauve. The colour formatting indicates the percentage of genomic conservation: from yellow (lowest) to green (highest).

The similarity between the *Mycobacteria* genomes under study is high, especially between *M. leprae* and *M. lepromatosis* and between the members of the Mycobacterium tuberculosis complex. *M. leprae* showed at least 38% genetic similarity with all the other species as well as 16% with the completely unrelated *Y. pestis* bacterium. While this result does not take genome sizes into account nor represent the actual relatedness of the species, it highlights the need to identify species-specific genomic regions that can be used to rule out artifactual *M. leprae* identification due to the presence of other *Mycobacteria* species.

Supplementary Figure 2-5 show the synteny in parts of the genomes for which species-specific regions (targets) were found using the multiple alignment file. No contradiction between the alignment backbone and the display was observed, apart from the *M. lepromatosis* genomic targets. Indeed, because this genome is still formed of non-ordered contigs, the target positions had to be first recalculated as positions per contig.

Table 8: Chosen genomic regions used as specific target for each species

	Genomic target 1			Genomic target 2		
	Start	End	Length (bp)	Start	End	Length (bp)
M. lepromatosis	3: 1	3:4950	4949	39:10452	39:14651	4199
M. leprae	1306215	1313716	7501	202661	209677	7016
M. bovis	1766079	1773879	7800	1764633	1765779	1146
M. tuberculosis	2969989	2980970	10981	1779267	1787933	8666
Y. pestis	1277307	1976310	699003	4371904	4611816	239912

3.3.2 Quality control of the sequencing data with FastQC

An extensive list of the FastQC results is provided in Supplementary Table 8. The most commonly observed failure report originated from the K-mer content control. Specifically, K-mer enrichment plots show several sharp peaks at the beginning of the sequences, suggesting the presence of a small amount of short and over-amplified DNA fragments. The second most common failure report was issued by the adapter content control, which indicated that part of the Illumina adapters were sequenced due to the fact that a large part of the sequencing reads were shorter than the maximal read length.

Most of the sequencing runs also triggered warnings during the control for GC percentage per sequence. The distribution of the average GC% amongst all sequences showed deviations from a normal distribution mainly on the left flank (the presence of a low, broad peak around 35-40bp) and at the ~60% maxima (the number of reads with ~60% GC is higher than the expected mono-modal distribution). The warning concerning the per-base sequence content was often reported as well. The G and C frequencies are in disequilibrium compared to the A and T frequencies (~30% and ~20%, respectively). Interestingly, UDG-treated libraries and non-UDG-treated libraries trigger the warning in a similar manner. Almost half of the runs triggered warnings for the per tile sequence quality. This warning is issued when a tile on the flowcell sequences a given read position whose quality is significantly lower than average. All the heatmaps provided by FastQC

showed variations limited to only a couple of cycles or a few number of basepairs in the reads. This variation was observed regularly for the last cycles at the end of a sequencing run. In addition, FastQC files originating from one sequencing lane displayed similar heatmaps patterns.

Some samples triggered the warning for the sequence quality per base and the corresponding quality whisker boxplots systematically showed a decreased quality very early in the read (decrease happening around 25-35bp). A few samples issued warnings for over-represented sequences. All of those samples showed an over-representation of sequences that could be attributed to the primers used during the indexing or amplification PCRs. Finally, one sample triggered the sequence duplication levels warning, indicating low levels of sequence diversity in the corresponding library (KH130040 UDG-treated).

3.3.3 Ancient DNA authentication

During a study, the first indication of contamination danger often comes when the blanks show traces of DNA. In this project, there were several blanks used at various steps in the analyses. The first blank used was to control against possible contamination during the extraction process. It was carried along into the later assays and did not show signs of contamination. When used during the PCR screening the extraction blanks were clean, as were the proper PCR blanks (see section 3.1 PCR screening for human and pathogen DNA). When converted to sequencing libraries and sequenced, the extraction blanks stayed clear of contamination, as did the library blanks (see section 3.2 Illumina sequencing library concentration and quality). All told, these results indicate that the sequences were not contaminated with modern DNA during the laboratory procedures and that the DNA recovered did, in fact, come from the ancient remains.

However, the use of blanks is not sufficient to authenticate DNA sequences as ancient, as remains could have been contaminated prior to their collection for a project. One characteristic of ancient DNA is the accumulation of deaminated bases at the end of the DNA fragments (see section 1.4.2). The frequency of the deamination of Cytosines and Guanines at the ends of the sequenced fragments can be calculated from the non-UDG treated datasets and used to rule out contamination with modern DNA. This analysis was performed for the human mitochondrial DNA as well as for the *M. leprae* DNA using the mapDamage2.0 software (Jonsson et al. 2013; Ginolhac et al. 2011) as implemented in EAGER (Peltzer et al. 2016). An example of the mapDamage output is provided in Figure 14.

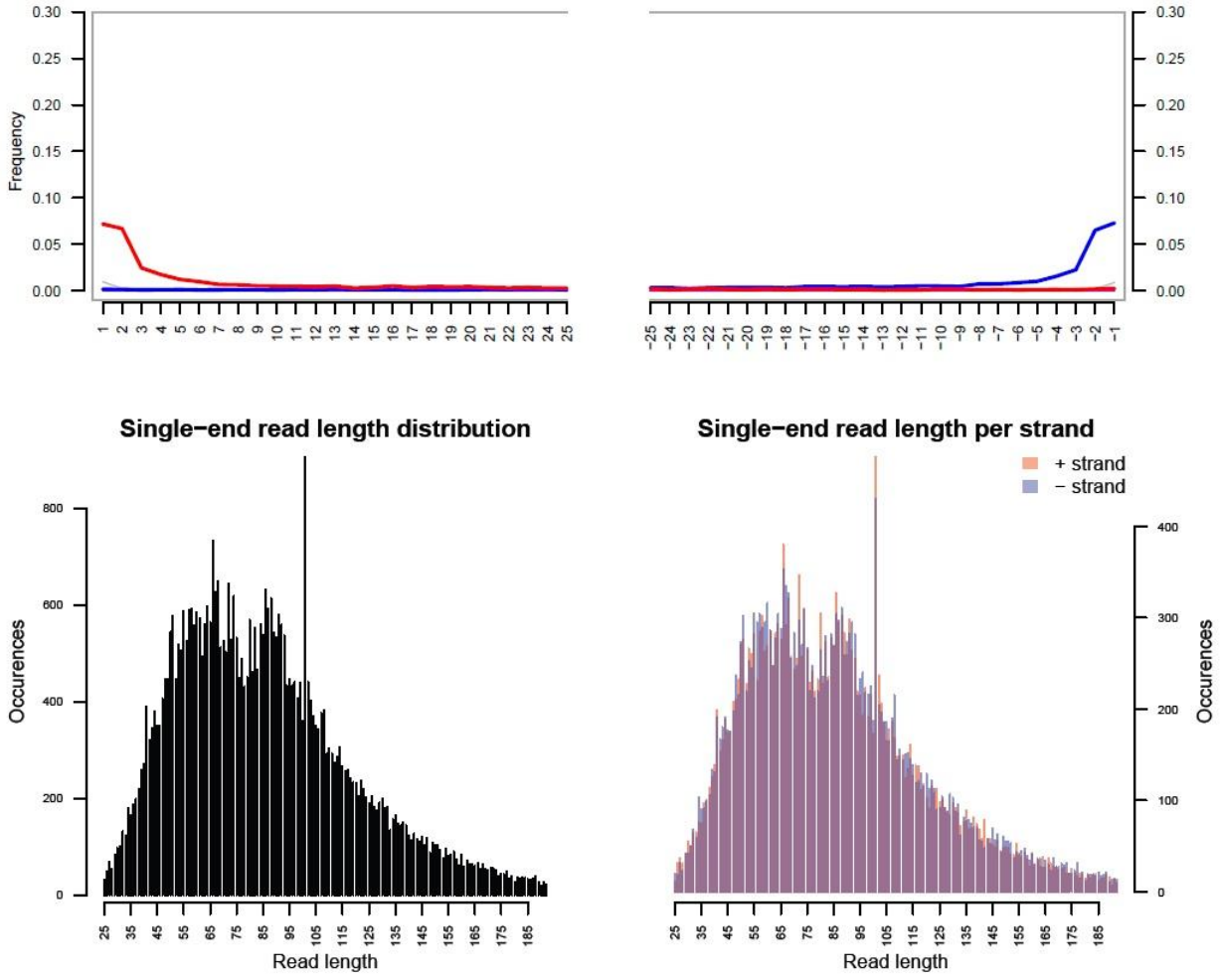
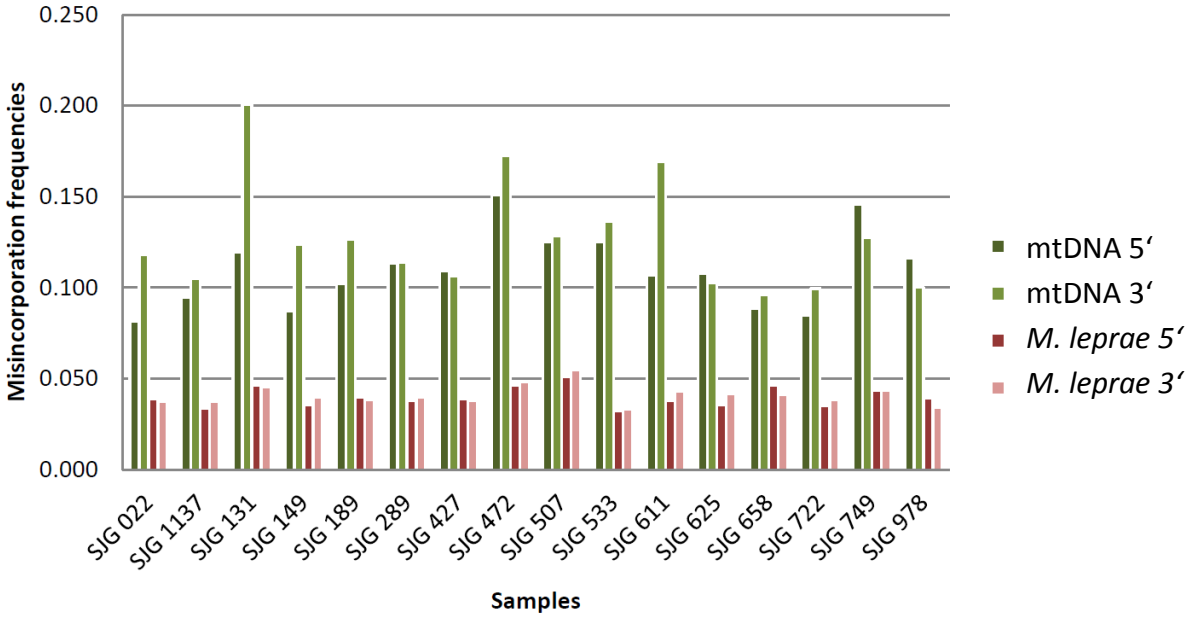


Figure 14: Example of mapDamage output
Sample SJG 507, UP: Misincorporation frequencies on the 5'-end (left) and 3'-end (right), DOWN: Read length distribution

The Figure 15 gathers the 1st base deamination frequency at the 5' and 3' end of the DNA fragments for each sample. The 5'-end C to T and 3'-end G to A misincorporation frequencies calculated were significantly higher for the reads which mapped to human mtDNA than for the reads which mapped to the *M. leprae* reference genome, suggesting that ancient *M. leprae* DNA is better preserved than its human DNA counterpart. In addition, the human mtDNA reads showed a lower 5'-end C to T misincorporation frequency for almost all the samples. This discrepancy between the two ends of the sequenced reads was not observed for the *M. leprae* reads. For both *M. leprae* and human reads, the increased misincorporation frequencies at the end of the DNA fragments was sufficient to rule out being in the presence of modern human or mycobacterial DNA. Nevertheless, in the case of *M. leprae* reads, a thorough study of the

literature was necessary to explain the lower degradation of the DNA and to authenticate the results (see discussion below).

Figure 15: 1st base deamination frequency at the 5' (C to T) and 3' (G to A) ends



Frequencies for SJG 404 are not reported because the number of non-UDG reads which mapped to each species was too low for mapDamage to perform reliably.

The fragment length distribution for the reads mapping to the human mtDNA reference revealed that more than 99% of the reads were shorter than 120bp (see Figure 16). In comparison, more than 99% of the reads were shorter than 90bp for the reads mapping to the *M. leprae* reference (see Figure 17). This would seem to indicate that, this time, the ancient human mtDNA was better preserved than its *M. leprae* counterpart. Both maximal read lengths stayed in the range of what could be expected of ancient DNA fragments, suggesting, once again, that contamination did not bias the sequencing results. The apparent inconsistency between the less damaged patterns but shorter reads in *M. leprae* probably only reflects the duality of DNA molecular degradation processes, as is discussed below.

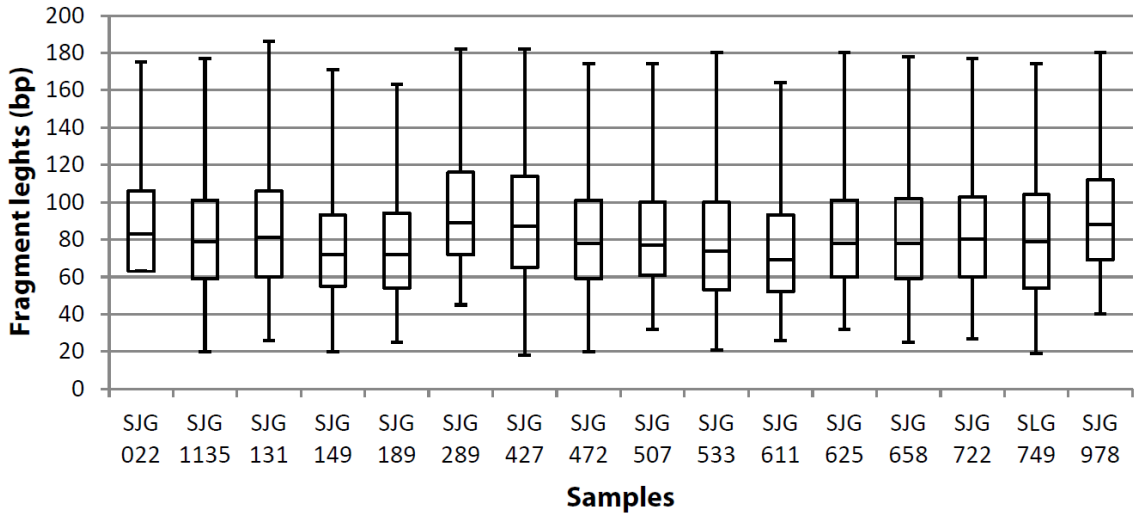


Figure 16: Fragment length distribution for the ancient human mtDNA reads
The length distribution for SJG 404 was not reported because the number of non-UDG reads which mapped to each species was too low for mapDamage to perform reliably.

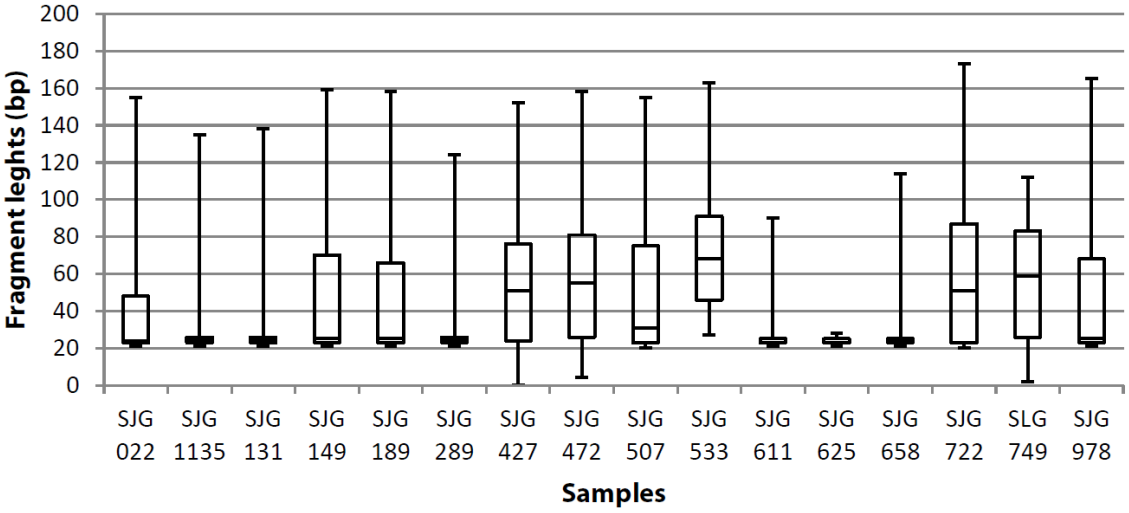


Figure 17: Fragment length distribution for the M. leprae *aDNA reads*
The length distribution for SJG 404 was not reported because the number of non-UDG reads which mapped to each species was too low for mapDamage to perform reliably.

3.3.4 Species identification

Species identification was performed on the datasets with eight samples per sequencing lane. The reads were first mapped against the human mtDNA. The results of this mapping step for the UDG-treated libraries are displayed in Table 9. The results for the non-UDG treated libraries are available in Supplementary Table 11.

Table 9: Human mtDNA reference coverage and read depth

Sample	Number of raw reads (millions)	Reads mapping (%)	Reference 1X covered (%)	Reference 4X covered (%)
SJG 022	7.1	0.02	98.42	73.09
SJG 131	17.6	0.01	99.56	82.27
SJG 149	8.0	0.01	97.46	71.52
SJG 189	6.4	0.03	69.99	71.52
SJG 289	17.1	0.01	99.79	92.32
SJG 404	21.6	0.01	99.23	80.99
SJG 427	28.9	0.01	99.92	99.19
SJG 472	9.7	0.02	99.44	92.14
SJG 507	19.6	0.01	99.96	99.95
SJG 533	28.4	0.03	99.97	99.94
SJG 611	27.2	0.01	99.78	95.18
SJG 658	38.8	0.01	99.79	92.32
SJG 722	25.0	0.02	99.91	99.79
SJG 749	22.8	0.04	99.99	99.91
SJG 978	22.3	0.01	99.57	88.30
SJG 1137	9.6	0.01	98.32	66.32

Overall, the human mtDNA was well-preserved in all samples sequenced, with the percentage of the mtDNA reference genome covered ranging from 69.99% to 99.99%. The UDG-treated reads were then mapped against five pathogen reference genomes to identify the pathogens that were likely to be present in the remains (see Table 10). The same procedure was applied to the non-UDG-treated sequencing reads, the results of which are displayed in 12.3 Supplementary results. The *M. leprae* reference genome showed the percentage of mapping reads with the highest coverage (with up to 99.99% of the reference covered). All three other *Mycobacteria* species also showed up to 55%, 5.80% and 5.74% for *M. lepromatosis*, *M. bovis* and *M. tuberculosis*, respectively. The coverage percentages for the distant cousin *Y. pestis* were significantly lower (no more than 0.57% of the reference genome was covered).

Table 10: Results of mapping against the pathogen genomes

Samples	*M. leprae*			*M. lepromatosis*			*M. bovis*			*M. tuberculosis*			*Y. pestis*		
	Reads mapping (%)	Reference covered (%)		Reads mapping (%)	Reference covered (%)		Reads mapping (%)	Reference covered (%)		Reads mapping (%)	Reference covered (%)		Reads mapping (%)	Reference covered (%)	
		1X	4X		1X	4X		1X	4X		1X	4X		1X	4X
SJG 022	0.10	16.07	0.05	0.02	2.77	0.04	0.01	0.46	0.04	0.01	0.45	0.04	<0.01	0.19	0.04
SJG 131	0.05	15.70	0.09	0.02	3.27	0.10	0.02	1.05	0.11	0.02	1.03	0.11	0.08	0.41	0.13
SJG 149	0.02	2.03	0.05	0.01	0.59	0.05	0.01	0.49	0.05	0.01	0.48	0.05	<0.01	0.23	0.05
SJG 189	0.22	28.50	0.08	0.05	5.07	0.03	0.01	0.44	0.03	0.01	0.44	0.03	<0.01	0.28	0.03
SJG 289	0.03	6.08	0.13	0.02	1.70	0.13	0.03	1.31	0.18	0.03	1.30	0.17	0.01	0.34	0.11
SJG 404	0.26	73.81	5.01	0.07	16.27	0.47	0.03	1.92	0.25	0.03	1.89	0.25	0.01	0.34	0.13
SJG 427	0.64	98.63	65.40	0.15	32.48	7.07	0.04	4.17	0.40	0.04	4.13	0.39	<0.01	0.34	0.13
SJG 472	0.14	25.73	0.14	0.04	4.97	0.09	0.03	1.01	0.11	0.03	0.99	0.11	<0.01	0.21	0.05
SJG 507	12.63	99.99	98.99	2.41	54.72	41.60	0.11	5.80	1.69	0.11	5.74	1.67	<0.01	0.27	0.06
SJG 533	1.44	99.97	98.48	0.27	39.62	17.76	0.02	2.99	0.33	0.02	2.95	0.32	<0.01	0.34	0.16
SJG 611	0.03	11.36	0.13	0.01	2.71	0.13	0.02	1.30	0.17	0.02	1.29	0.16	<0.01	0.51	0.20
SJG 658	0.04	20.97	0.19	0.02	5.07	0.16	0.02	2.03	0.22	0.02	1.98	0.21	<0.01	0.44	0.18
SJG 722	0.33	84.24	13.36	0.08	20.04	1.01	0.04	2.34	0.32	0.04	2.30	0.32	0.01	0.41	0.15
SJG 749	7.16	99.99	99.99	1.34	52.26	36.66	0.07	5.25	1.16	0.07	5.19	1.15	0.01	0.43	0.17
SJG 978	0.11	35.45	0.41	0.04	3.76	0.11	0.05	3.60	0.47	0.05	3.55	0.46	0.01	0.57	0.20
SJG 1137	0.10	15.99	0.12	0.04	3.76	0.11	0.04	1.22	0.15	0.04	1.20	0.15	0.01	0.30	0.10

Green cells indicate a sample which yielded a coverage of at least 30% for the species.

As expected, there seems to be a certain proportion of reads which mapped aspecifically, as the samples showing the best *M. leprae* coverage were also the ones with highest coverage for all the other species (including *Y. pestis*). The coverage statistics calculated for the two species-specific target regions for each species are displayed in Table 11. The coverage on the *M. bovis*-specific targets was always lower compared to the whole genome, which agrees with the very low coverage observed for this species in Table 10 and suggests that the bacterium was not present in our samples.

The coverage on the *M. leprae*-specific target was either consistent or higher compared to the whole reference genome. The slight decrease observed in some samples might be caused by variations in the read repartition along the reference (some regions being spontaneously better covered than others) and does not seem to have been caused by the presence of reads from another related species. The issue will be discussed further below. The coverage statistics for *M. lepromatosis* have to be considered with caution as the reference genome is not yet complete and only consists of several non-ordered contigs. However, there is a clear trend towards a lower coverage on the target regions, suggesting that the reads seen in the whole-genome mapping are likely cross-species mapping from *M. leprae* reads.

Interestingly, although none of the sequenced samples were positive for *M. tuberculosis* DNA during the PCR screening phase of this project, five samples display significantly higher coverage on the target regions compared to the whole *M. tuberculosis* genome (Table 11). The coverage statistics for the distant relative of *Y. pestis* are shown in Supplementary Table 12. Some samples also showed slightly better coverage on the regions possessed by *Y. pestis* and not the mycobacterial species. Since the coverage stayed extremely low, it is likely that those reads are cross-mapping reads from closely related species. Indeed, the target regions were only designed to discriminate between reads which mapped to both mycobacterial genomes and the *Y. pestis* reference and could not rule out the presence of DNA fragments with a shared similarity with *Y. pestis* but not *Mycobacteria*.

Table 11: Read depth on whole genomes and species-specific targets

	M. leprae				*M. lepromatosis*				*M. bovis*				*M. tuberculosis*			
	Whole reference		Targets		Whole reference		Targets		Whole reference		Targets		Whole reference		Targets	
Samples	All	Covered bases	All	Covered bases	All	Covered bases	All	Covered bases	All	Covered bases	All	Covered bases	All	Covered bases	All	Covered bases
SJG 022	0.176	1.070	0.199	1.117	0.021	1.124	0.000	0.000	0.003	2.260	0.000	0.000	0.003	2.269	0.000	0.000
SJG 131	0.180	1.172	0.182	1.141	0.031	1.580	0.000	0.000	0.014	4.784	0.000	0.000	0.014	4.767	0.002	1.000
SJG 149	0.022	1.207	0.027	1.000	0.006	1.882	0.000	0.000	0.004	3.286	0.000	0.000	0.004	3.286	0.000	0.000
SJG 189	0.338	1.191	0.325	1.220	0.037	1.126	0.000	0.000	0.000	1.066	0.000	0.000	0.003	2.342	0.000	0.000
SJG 289	0.078	1.403	0.031	1.000	0.025	2.874	0.000	0.000	0.021	6.407	0.000	0.000	0.021	6.424	0.009	3.179
SJG 404	1.361	1.850	1.472	1.961	0.174	1.576	0.000	0.000	0.029	6.019	0.000	0.000	0.029	5.970	0.026	8.159
SJG 427	4.547	4.612	4.851	4.782	0.551	2.525	0.008	0.000	0.046	4.161	0.000	0.000	0.046	4.146	0.021	5.468
SJG 472	0.305	1.204	0.326	1.310	0.042	1.347	0.000	0.000	0.012	4.453	0.000	0.000	0.012	4.457	0.013	3.486
SJG 507	64.795	64.796	66.359	66.359	7.030	18.208	0.066	1.573	0.101	6.491	0.000	0.000	0.100	6.494	0.038	9.060
SJG 533	10.730	10.733	10.727	10.727	1.135	4.106	0.009	1.000	0.025	3.128	0.000	0.000	0.025	3.143	0.009	2.048
SJG 611	0.129	1.196	0.095	1.000	0.026	1.891	0.000	0.000	0.015	4.700	0.003	1.000	0.015	4.708	0.002	1.000
SJG 658	0.251	1.253	0.187	1.103	0.052	1.810	0.000	0.000	0.028	5.266	0.000	0.000	0.027	5.261	0.009	2.734
SJG 722	1.931	2.301	2.011	2.381	0.242	1.805	0.000	0.000	0.047	7.819	0.000	0.000	0.047	7.806	0.016	5.328
SJG 749	42.564	42.564	44.098	44.098	4.470	12.319	0.049	1.257	0.073	5.419	0.000	0.000	0.072	5.413	0.029	7.263
SJG 978	0.474	1.370	0.519	1.275	0.093	1.957	0.004	1.000	0.054	5.267	0.000	0.000	0.053	5.277	0.033	10.460
SJG 1137	0.185	1.188	0.163	1.070	0.035	1.600	0.000	0.000	0.016	4.845	0.000	0.000	0.016	0.016	0.007	0.007

The colour coding indicates whether the target coverage was more than 0.1X higher (orange) or lower (blue) compared to the coverage on the whole reference. Cells were kept white when the difference between whole genome and target was below 0.1X.

3.3.5 Radiocarbon dating

Table 12: Radiocarbon and calibrated age of the dated samples

Sample	MAMS lab ID	14C age (±24)	Calibrated age (INTCAL13 & SwissCal1.0)	Collagen content (%)
SJG 404	26162	780	cal AD 1226-1268	4.1
SJG 427	26163	835	cal AD 1177-1250	3.6
SJG 472	26164	756	cal AD 1252-1280	3.6
SJG 507	26165	856	cal AD 1170-1214	1.8
SJG 533	26166	892	cal AD 1051-1206	3.8
SJG 722	26167	722	cal AD 1269-1285	2.1
SJG 749	26168	763	cal AD 1230-1277	2.0
SJG 978	26169	672	cal AD 1283-1381	6.5
SJG 1137	26170	675	cal AD 1282-1380	6.5

3.3.6 Mycobacterial plasmid results

Of the 22 plasmid reference sequences tested, 17 yielded virtually no mapping reads (data not shown). The five plasmids with mapping sequencing reads (pTYGi9, pMYCCH01, pMYCSM01, pMYCSM03 and pMSPYR101) all showed a coverage and read depth which was too low for their presence to be clearly assessed in order to rule out unspecific mapping (data not shown).

3.3.7 Re-sequencing

Eight (8) samples were chosen for re-sequencing (SJG 404, SJG 427, SJG 472, SJG 507, SJG 658, SJG 722, SJG 978 and SJG 1137). The reads from the re-sequencing runs were analysed using the BOWTIE pipeline described above. The results are shown in Table 13. The eight (8) samples showed minimum 4X per base coverage on at least 32% of the *M. leprae* reference genomes. Although all the re-sequenced DNA samples also yielded virtually complete human mtDNA sequences, several samples with high mtDNA recovery yielded only fractions of the *M. leprae* genome.

Table 13: Human mtDNA and M. leprae *DNA reference coverage and read depth for the UDG-treated libraries*

Samples	human mtDNA			*M. leprae*		
	Read length (bp)	Reference covered		Read length (bp)	Reference covered	
		1X	4X		1X	4X
SJG 404	74 (57-100)	1.0000	0.9996	88 (61-102)	1.0000	0.9979
SJG 427	85 (63-105)	1.0000	0.9998	101 (80-115)	1.0000	1.0000
SJG 472	74 (58-97)	1.0000	0.9996	75 (55-101)	0.9966	0.9060
SJG 507	82 (60-103)	1.0000	1.0000	95 (66-106)	1.0000	1.0000
SJG 658	77 (58-102)	1.0000	1.0000	68 (48-98)	0.5462	0.1894
SJG 722	74 (55-100)	1.0000	0.9999	84 (58-121)	1.0000	0.9999
SJG 978	85 (65-104)	0.9998	0.9976	92 (65-101)	0.9867	0.8081
SJG 1137	81 (62-105)	1.0000	0.9999	77 (53-109)	0.8732	0.3175

3.4 Human DNA results

3.4.1 Combined datasets mapping statistics

The pooled reads from the first sequencing runs, the re-sequencing runs and the additional datasets were analyzed using the previously-mentioned BOWTIE pipeline. The results are shown in Table 14. All the high-coverage samples also yielded virtually complete human mtDNA sequences, however several samples with high mtDNA recovery yielded only fractions of the *M. leprae* genome. This will be discussed in greater detail below.

Table 14: Coverage and read depth statistics for the human mitochondrial DNA

Samples	human mtDNA			
	# Reads mapping	Read length (bp)	Reference covered	
			1X	4X
SJG 022	3228	83 (64-109)	0.9998	0.9858
SJG 131	2157	83 (63-105)	0.9976	0.9708
SJG 149	5397	76 (60-98)	0.9992	0.9937
SJG 189	13671	77 (60-99)	0.9996	0.9972
SJG 289	3186	86 (65-110)	0.9990	0.9935
SJG 404	11497	74 (57-100)	1.0000	0.9996
SJG 427	22126	85 (63-105)	1.0000	0.9998
SJG 472	29163	74 (58-97)	1.0000	0.9996
SJG 507	98655	82 (60-103)	1.0000	1.0000
SJG 533	19914	84 (64-110)	1.0000	0.9997
SJG 611	1938	67 (53-88)	0.9966	0.8982
SJG 658	42832	77 (58-102)	1.0000	1.0000
SJG 722	3449	74 (55-100)	1.0000	0.9999
SJG 749	15347	81 (59-107)	1.0000	0.9996
SJG 978	13755	85 (65-104)	0.9998	0.9976
SJG 1137	13331	81 (62-105)	1.0000	0.9999

3.4.2 Sex typing based on X and Y chromosome coverage

Table 15 shows the results obtained by the X and Y mapping statistics sex typing proxy. Some samples only showed low coverage on the X and Y chromosomes. As the Y/X ratio is highly coverage-dependent, those samples were not typed and are listed below as undetermined. Two samples showed a Y to X ratio inconsistent with their osteologically-determined sex. This will be discussed further below.

Table 15: Results of the human sex typing

Sample	Sex typing from NGS data	Osteological sex
SJG 022	Und.	F
SJG 131	F	F
SJG 149	Und.	F
SJG 189	Und.	F
SJG 289	M	Und.
SJG 404	M	M
SJG 427	F	F
SJG 472	F	F
SJG 507	F	F
SJG 533	Und.	F
SJG 611	M	M
SJG 658	M	M
SJG 722	F	M
SJG 749	F	F
SJG 978	M	F
SJG 1137	F	F

Those individuals described as "Und." had inconclusive results due to low coverage on the sexual chromosomes. Inconsistencies with the osteological sexing provided by the archaeologist collaborators are highlighted in red.

3.5 Mycobacterial plasmid results

The pooled reads from the first sequencing runs, the re-sequencing runs and the additional datasets were analysed using the BOWTIE pipeline as has already been described. The mapping was repeated following the same methodolgy as had been completed in the first run against the reference sequences of the five (5) plasmids which previously showed reads which had mapped (pTYGi9, pMYCCH01, pMYCSM01, pMYCSM03 and pMSPYR101). Although mapping coverage and read depth did improve after re-sequencing, the coverage statistics stayed too low to be conclusive regarding the presence of a plasmid. No plasmid showed a coverage higher than about 2% of the reference (data not shown).

3.6 *M. leprae* genome findings

3.6.1 Combined datasets' mapping statistics

The pooled reads from the first sequencing runs, the re-sequencing runs and the additional datasets were analyzed using the BOWTIE pipeline which has already been mentioned. The results are summarized in Table 16.

Table 16: Combined datasets' coverage statitstics for M. leprae

	M. leprae			
			Reference covered	
Samples	# Reads mapping	Read length (bp)	1X	4X
SJG 022	15870	91 (63-116)	0.3190	0.0086
SJG 131	10875	80 (56-102)	0.2104	0.0022
SJG 149	5349	79 (54-101)	0.0616	0.0070
SJG 189	77462	82 (59-111)	0.6970	0.2055
SJG 289	5233	65 (47-96)	0.0765	0.0008
SJG 404	566204	88 (61-102)	0.9999	0.9979
SJG 427	11623054	101 (80-115)	1.0000	1.0000
SJG 472	362963	75 (55-101)	0.9966	0.9060
SJG 507	35561167	95 (66-106)	1.0000	1.0000
SJG 533	933382	98 (68-125)	0.9999	0.9998
SJG 611	6258	63 (46-92)	0.1102	0.0007
SJG 658	103592	68 (48-98)	0.5462	0.1894
SJG 722	1135985	84 (58-121)	1.0000	0.9999
SJG 749	4111522	94 (65-119)	1.0000	1.0000
SJG 978	1308555	92 (65-101)	0.9867	0.8081
SJG 1137	125425	77 (53-109)	0.8732	0.3175

In addition, the mapping to *M. tuberculosis* (whole reference and target regions, see Supplementary Table 13) showed no increase in the coverage statistics on the target regions, thereby confirming the likely absence of any co-infection cases from the sequenced samples.

3.6.2 Single nucleotide polymorphism analysis of the ancient M. leprae genomes

Genomic positions involved in *M. leprae* genotyping are listed in 12.3 Supplementary results for all the recovered ancient *M. leprae* genomes. Table 17 shows the various strains observed in the ancient *M. leprae* isolates. The genotyping method based on all the variants published in Monot et al. seems less reliable and less sensitive than the inference-like approach described by Singh et al. Irrespective of the method, it was not possible to determine the SNP type for the low-coverage genomes. The reasons behind this issue will be discussed below.

Table 17: Leprosy genotypes observed in this study

Sample	Most likely SNP type according to Monot et al (2005)	SNP type according to Singh et al (2011)
SJG 022	Und.	nd
SJG 131	Und.	1d
SJG 149	Und.	nd
SJG 189	Und.	Und.
SJG 289	Und.	nd
SJG 404	Und.	3I
SJG 427	Und.	3I
SJG 472	3I	3I
SJG 507	3I	3K
SJG 533	Und.	Und.
SJG 611		
SJG 658	Und.	1d
SJG 722	3I	3I
SJG 749	2F	2F
SJG 978	Und.	2F
SJG 1137	Und.	3I

Genotyping was done according to Monot et al, 2005 and Sing et al, 2011. The lower-case letters indicate a genotype determined with low confidence. The sample highlighted in red showed different results depending on the method used.

3.6.3 Phylogenetic analyses

The phylogenetic trees calculated using several multiple alignment and construction methods showed a consistent grouping of the samples into the four SNP-types (see 12.3 Supplementary results). Figure 18 below shows a representative example of the tree obtained. All high-coverage ancient strains were found to cluster according to their SNP-types. Sample SJG 533 (which gave inconclusive SNP typing results) clustered with the 3I strains. Sample SJG 507 fell into the recently described Branch 0, which confirms the SNP typing results obtained via the Singh approach.

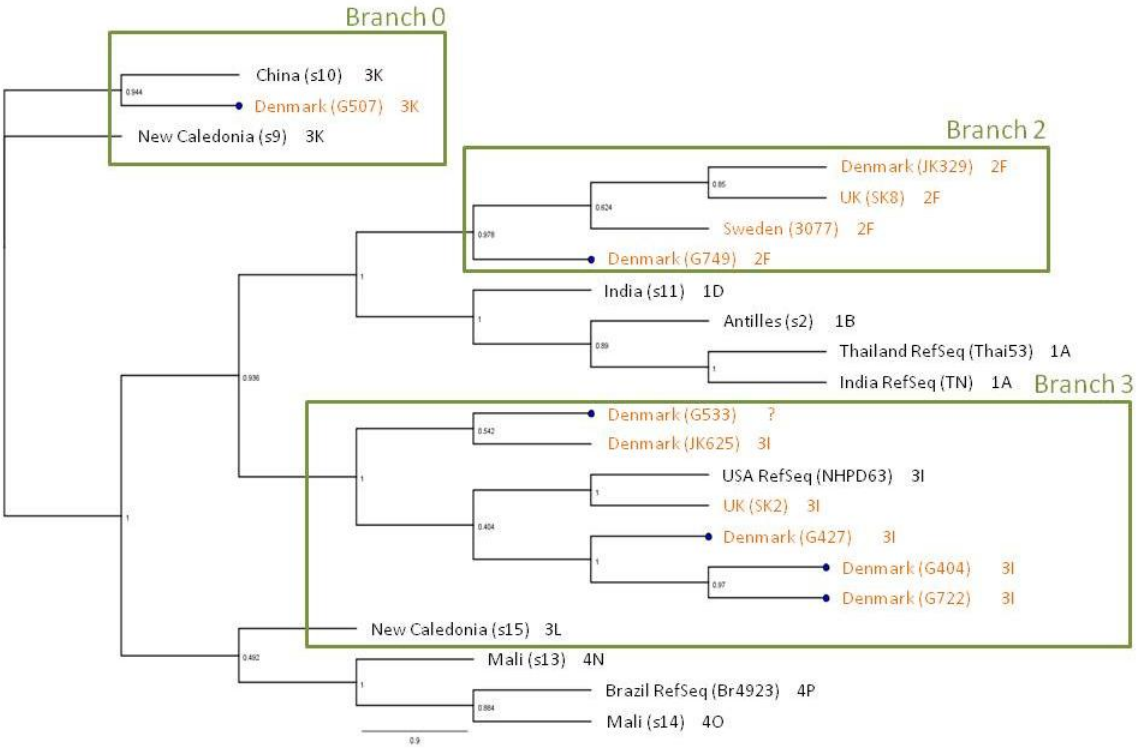


Figure 18: Phylogeny of M. leprae *including six new high-coverage genomes.*
Maximum Likelihood rooted tree constructed from the MAFFT multiple alignment. The node labels indicate the bootstrap values after 500 replications. The six new genomes are indicated with blue dots on the branch tips. Ancient genomes are shown in orange, modern ones in black. The name "RefSeq" is used to identify the four modern reference sequences. The scale represents the average substitution rate.

Amongst the ten low-coverage genomes, five were consistently assigned to Branch 3I (SJG 1137, SJG 189, SJG 472 and SJG 611) and Branch 2F (SJG 658). For the remaining five low-coverage genomes, it was not possible to confidently assign the sequence to one branch. Branch 3I and Branch 2F of the *M. leprae* partial tree constructions showed that the low-coverage genomes clustered within the branch (Figure 19). The clustering patterns observed were consistent with the global phylogeny which was previously constructed. Moreover, the genetic distance observed between members of the 3I Branch is higher than those between members of the 2F Branch, which is consistent with the trend observed on the global tree.

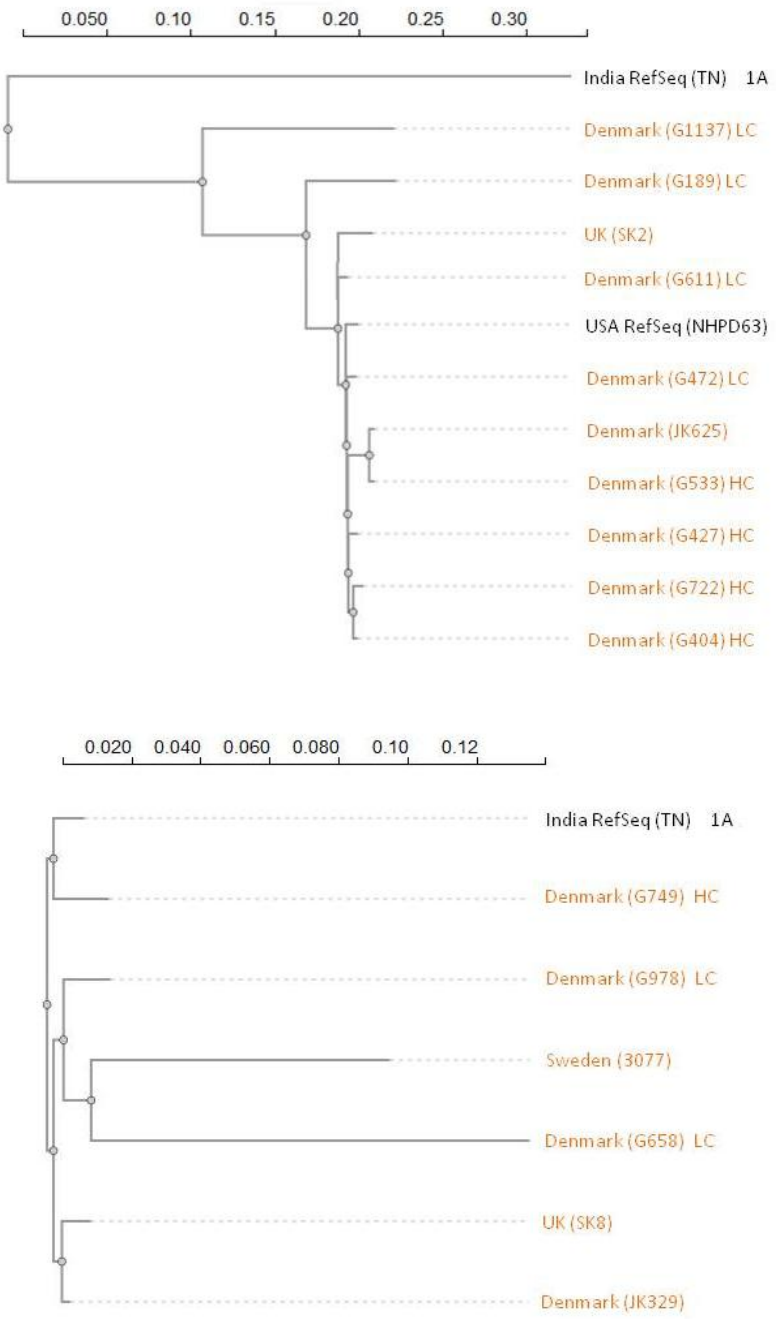


Figure 19: Partial genetic distance trees for Branch 3. UP: Branch 3I, DOWN: Branch 2F.
HC indicates the high-coverage genomes, LC the low coverage isolates. M. leprae *TN was used as an outgroup to root the tree.*

3.6.4 Variant annotation

The variant lists created with higher-stringency criteria were submitted to the SNPeff annotation software. Table 18 provides an overview of the number of variants per sample. The number of variants is highly variable between the samples, which is due to the differences in the coverage of *M. leprae (as has previously been observed*) and explains the range of variant rates displayed.

Table 18: SNPeff results overview statistics

Sample	Total number of effects found	Number of variants	Variant rate on the *M. leprae* reference
SJG 022	102	14	1/233,443
SJG 131	155	19	1/172,010
SJG 149	125	17	1/192,247
SJG 189	193	29	1/112,696
SJG 289	165	20	1/163,410
SJG 404	1,229	188	1/17,384
SJG 427	1,234	189	1/17,292
SJG 472	1,145	174	1/18,782
SJG 507	1,613	268	1/12,194
SJG 533	761	124	1/26,356
SJG 611	119	14	1/233,44
SJG 658	676	97	1/33,692
SJG 722	1,252	192	1/17,021
SJG 749	638	107	1/30,543
SJG 978	1,363	196	1/16,674
SJG 1137	775	114	1/28,668

In addition, SNPeff also provides an overview of the types of variants and their locations (Table 19). It is notable that SNPeff has only annotated SNPs. Moreover, most of the variants are located upstream or downstream from annotated genes; only a few are located inside exons, in a splice site or inside a non-gene transcript region. This is coherent with the expected location of variants in bacteria, since non-coding regions are submitted to a lowest selection pressure and, therefore, tend to accumulate mutations. The low number of SNPs found in intergenic regions might be related to the short length of those regions in a genome as small as *M. leprae*.

Table 19: M. leprae *variants per type and location*

Sample	Types of variants				Location of the effects					
	SNP	INS	DEL	Other	Downstr.	Exon	Intergenic	Splice site	Transcript	Upstr.
SJG 022	14	0	0	0	49	11	2	0	1	39
SJG 131	19	0	0	0	78	19	0	0	0	58
SJG 149	17	0	0	0	66	16	1	0	0	42
SJG 189	29	0	0	0	85	18	1	0	10	79
SJG 289	20	0	0	0	82	20	0	0	0	63
SJG 404	188	0	0	0	527	118	28	1	41	514
SJG 427	189	0	0	0	543	124	25	1	39	502
SJG 472	174	0	0	0	498	117	22	1	34	473
SJG 507	268	0	0	0	676	98	75	0	94	670
SJG 533	124	0	0	0	307	72	16	0	36	330
SJG 611	14	0	0	0	61	14	0	0	0	44
SJG 658	97	0	0	0	326	85	7	0	5	253
SJG 722	192	0	0	0	553	130	23	1	38	507
SJG 749	107	0	0	0	279	48	21	0	38	252
SJG 978	196	0	0	0	629	155	17	0	24	538
SJG 1137	114	0	0	0	358	87	8	1	18	303

The proportion of SNPs in each location is similar between all the samples, as is the genomic distribution of the variants (data not shown). The SNP distributions consistently showed a striking peak in variant frequency between 1,300,000 and 1,400,000 (see Figure 20). This region might be the location of a variable element, as is discussed below.

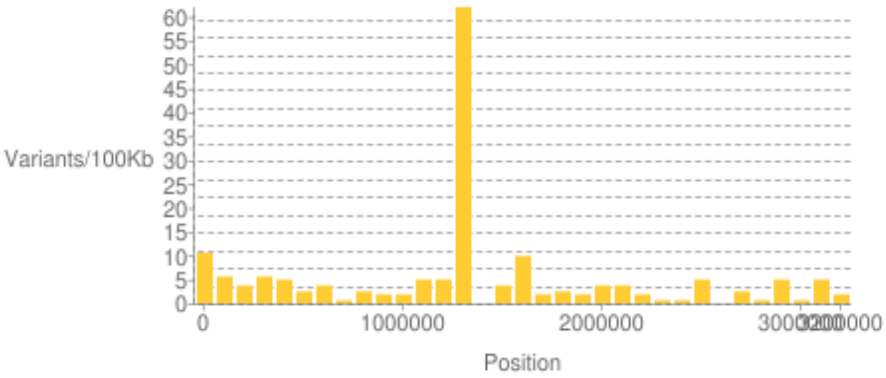


Figure 20: Example of SNP distribution along the M. leprae *genome (SJG 472)*

Finally, SNPeff provides an overview of the effects of the variants (Table 20). This overview is useful to estimate the proportion of SNPs, where an important effect might be expected for each sample.

Table 20: M. leprae *variants effects by functional class and impact*

	Functional class			Impact			
Sample	Missense	Silent	None	Low	Moderate	High	Modifier
SJG 022	1	2	99	2	1	0	99
SJG 131	1	0	154	0	1	0	154
SJG 149	3	0	122	0	3	0	122
SJG 189	3	3	187	3	3	0	187
SJG 289	1	0	164	0	1	0	164
SJG 404	46	24	1,159	24	46	0	1,159
SJG 427	44	24	1,166	24	44	0	1,166
SJG 472	35	27	1,083	27	35	0	1,083
SJG 507	65	32	1,516	32	64	1	1,516
SJG 533	40	22	699	22	40	0	699
SJG 611	0	0	119	0	0	0	119
SJG 658	14	7	655	7	14	0	655
SJG 722	38	29	1,185	29	38	0	1,185
SJG 749	28	17	593	17	28	0	593
SJG 978	36	23	1,304	23	36	0	1,304
SJG 1137	19	13	743	13	19	0	743

3.6.5 Possible effects of the variations observed

Once the results from all the samples were merged, the collected list of SNP effects showed 696 unique effects. Supplementary Figure 6 shows the distribution of the effects per type, impact, affected gene products and pathways before filtering. The filtering steps removed 648 effects as shown in Table 21. Most of the filtered variants were removed due to the fact that they were located in intergenic regions or because they were hypothetical protein genes or stable RNA genes.

Table 21: Number of annotated effects during and after filtering

	Number of effects
Before filtering	696
After removal of intergenic variants	522
After removal of hypothetical proteins and pseudogenes	304
After removal of the stable RNAs	127
After filtering of the variants observed in only one genome	48

Figure 21 below shows the distribution of the effects per type, impact, affected gene products and pathways.

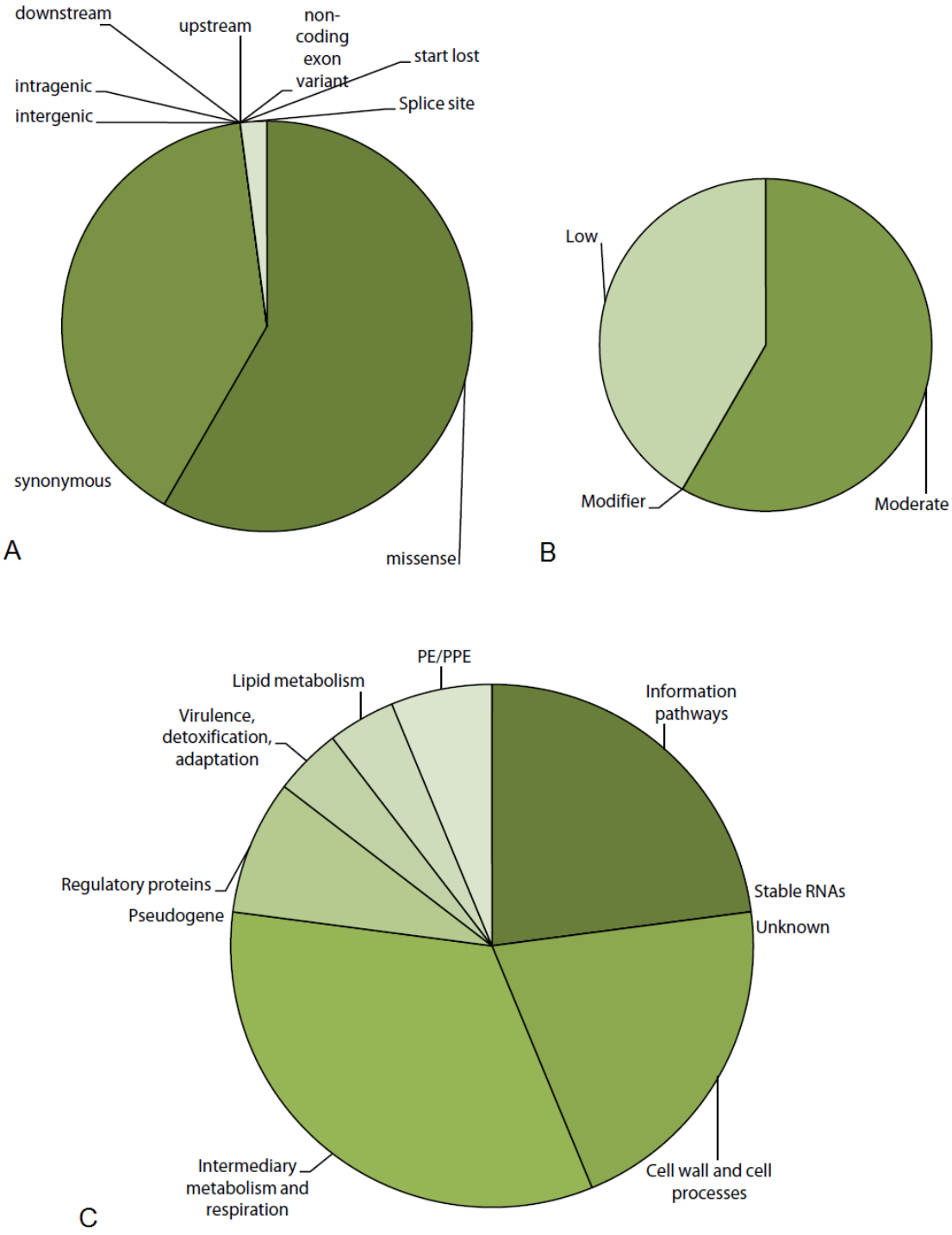


Figure 21: M. leprae *variant distribution after filtering*

A) SNP distribution per annotated effect, B) SNP distribution per impact, C) SNP distribution per product functional category.

Table 22 highlights the most frequent amino acid changes after filtering. Finally, Figure 22 displays the number of genes and variants in each functional category.

Table 22: Amino acid change observed after filtering

		New amino acid															
		Ala	Arg	Asn	Asp	Gln	Glu	Gly	Ile	Leu	Lys	Pro	Ser	Ter	Thr	Tyr	Val
Reference amino acid	Ala	3	0	0	0	0	0	0	0	0	0	0	0	0	1	0	3
	Arg	0	3	0	0	1	0	0	0	0	0	0	0	0	0	0	0
	Asn	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	Asp	0	0	2	1	0	1	0	0	0	0	0	0	0	0	0	0
	Cys	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	Gln	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
	Glu	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0
	Gly	0	0	0	1	0	0	2	0	0	0	0	1	0	0	0	0
	His	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	Leu	0	0	0	0	0	0	0	0	1	0	3	0	0	0	0	0
	Lys	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
	Met	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
	Phe	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	Pro	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
	Ser	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0
	Ter	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
	Thr	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
	Tyr	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
	Val	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2

The number of occurrences of each amino acid change is recorded in the corresponding cell. The colour intensity (from light to dark orange) highlights those changes which occur more often.

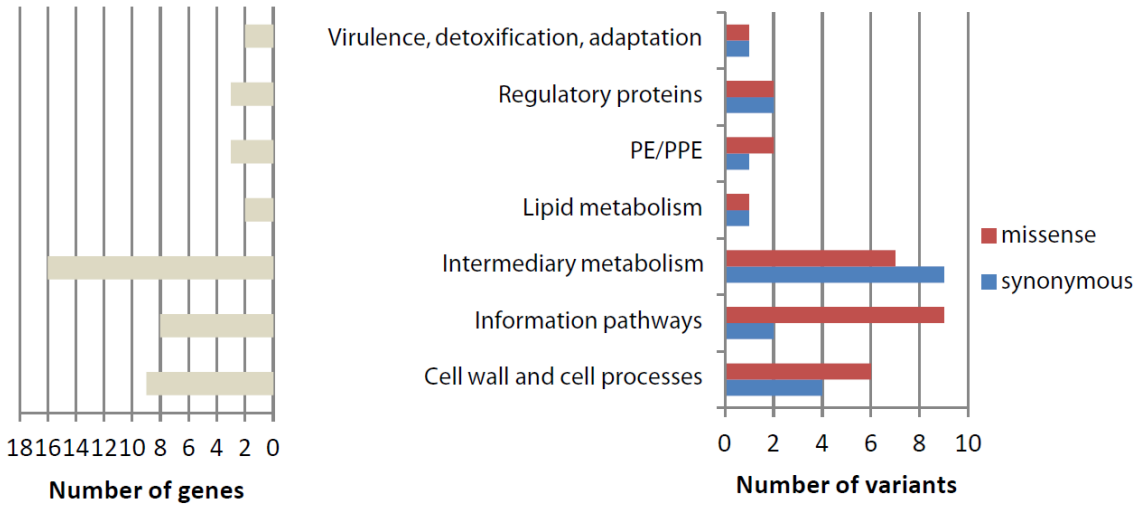


Figure 22: Number of genes and variants of each type per functional category

3.6.6 Selection of variants of interest and detailed SNP effect estimation

After the first filtering steps, only two possible variants of interest were found. One was a variant annotated as related to adaptation and virulence, the other a variant in a splice site. Unfortunately, both were missing critical protein contextual information and their effects could not be studied in detail.

The three proteins selected for in-depth analysis are listed in Table 23 along with their annotations and GO profiles. The dnaA protein showed two different possible amino acid changes. Those two changes can be observed both simultaneously and separately in this study's sample set. Therefore, the rest of the analyses were performed for the RT and the three possible variant combinations. The chemical properties of the proteins are shown in Table 24. None of the protein variants showed a pI, mW, hydrophobicity or aliphatic index variation suggestive of a major change in the protein solubility or activity. The same observation was made on the profiles computed using the protein sequences (Supplementary Figure 7-9). The secondary structure predictions are available in 12.3 Supplementary results. Although it is not possible to assess the effect of the variants on the proteins' activity levels, it is noticeable that the glcB variant does not seem to affect the secondary structure, while the secondary structure is affected by the other variants.

Figures 22-24 show the variant loci in their protein feature context. In dnaA and glcB, the variant loci were located in annotated conserved sites with known amino acid patterns. Similar results were obtained with the protein multiple alignment (Figures 25-26). The gyrA variant was outside of any conserved region (data not shown). The dnaA and glcB variants were located in conserved regions. The variants dnaA ser25gly and glcB leu591pro were also present in other species, while the dnaA gly295ser variant was only present in the sequences from the present study.

Table 23: Proteins selected for in-depth prediction of the effects of the SNPs

Mycobrowser			GO annotations			OrthoDB			SNPeff					
Gene	Protein	Functional category	Biological process	Cellular component	Molecular function	Evolutionnary rate	# of InterPro domains	Protein lenght	AA old	AA change locus	AA new	AA old	AA change locus	AA new
dnaA	Chromosomal replication initiator DnaA	Information pathways	Regulation of DNA replication	Cytoplasm	ATP binding DNA replication origin binding	0.87	8	521	Ser	25	Gly	Gly	295	Ser
gyrA	DNA gyrase subunit A	Information pathways	DNA topological change	Chromosome Cytoplasm	DNA binding DNA topoisomerase (ATP-hydrolyzing) activity endonuclease activity	0.87	7	1249	Leu	379	Pro			
glcB	Malate synthase G	Intermediary metabolism and respiration	Glyoxylate cycle tricarboxylic acid cycle	Cytoplasm	ATP binding Malate synthase activity Metal ion binding	0.86	4	731	Lys	591	Glu			

The colours indicate which criteria seleted which protein. Note that dnaA possesses two annotated variants.

Table 24: Chemical properties of the proteins

		pI		mW (kDa)		Hydrophobicity		Aliphatic Index
Gene	Variant	ExPASy Compute pI/mW	ExPASy ProtParam	ExPASy Compute pI/mW	ExPASy ProtParam	SOSUI	ExPASy ProtParam	ExPASy ProtParam
dnaA	RT	5.46	5.46	58.6	58.6	-0.353	-0.353	89.17
	ser25gly	5.46	5.46	58.6	58.6	-0.352	-0.352	89.17
	gly295ser	5.46	5.46	58.6	58.6	-0.354	-0.354	89.17
	ser25gly + gly295ser	5.46	5.46	58.6	58.6	-0.353	-0.353	89.17
gyrA	RT	5.65	5.65	138.5	138.5	-0.162	-0.162	95.65
	leu379pro	5.65	5.65	138.5	138.5	-0.167	-0.167	95.35
glcB	RT	5.09	5.09	80.1	80.1	-0.25	-0.25	98.43
	lys591glu	5.03	5.03	80.1	80.1	-0.25	-0.25	98.43

The different tools used to obtain each value are displayed in italics. Several tools were used for each value whenever possible. The references for each tool are available in 12.2 Supplementary methods.

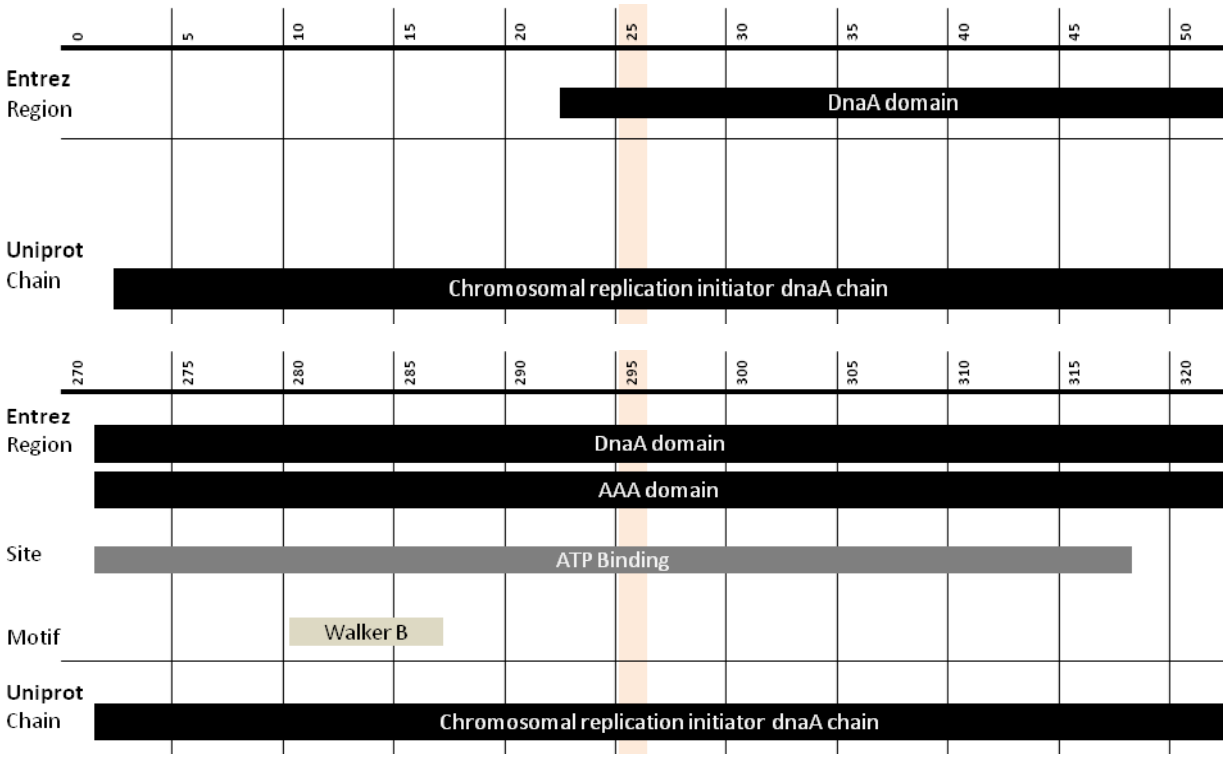


Figure 23 dnaA variant loci and protein features
UP: locus 1: ser25gly, DOWN: locus2: gly295ser

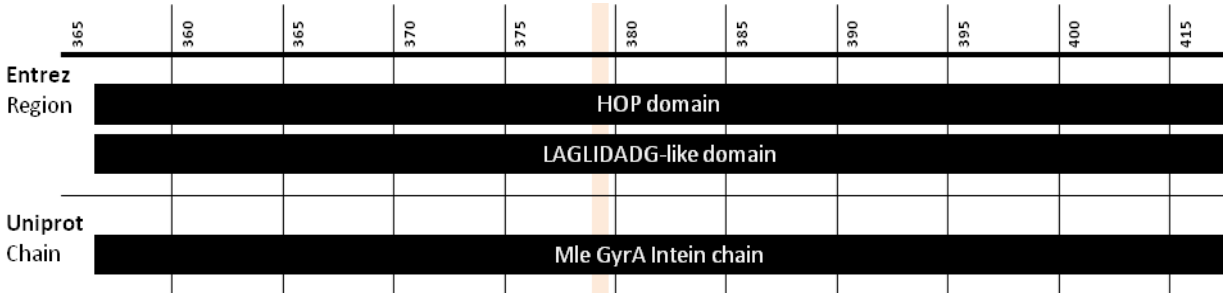


Figure 24: gyrA variant locus and protein features
Amino acid change: leu379pro

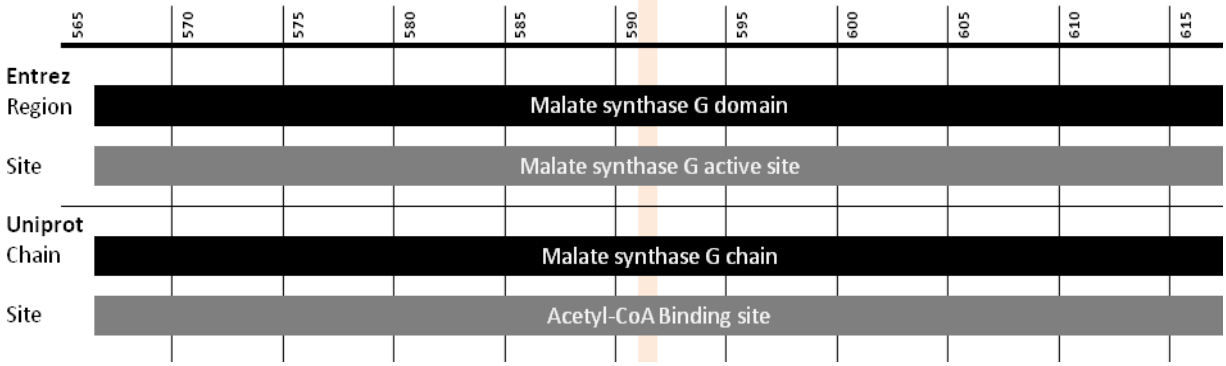


Figure 25: glcB variant locus and protein features
Amino acid change lys591glu

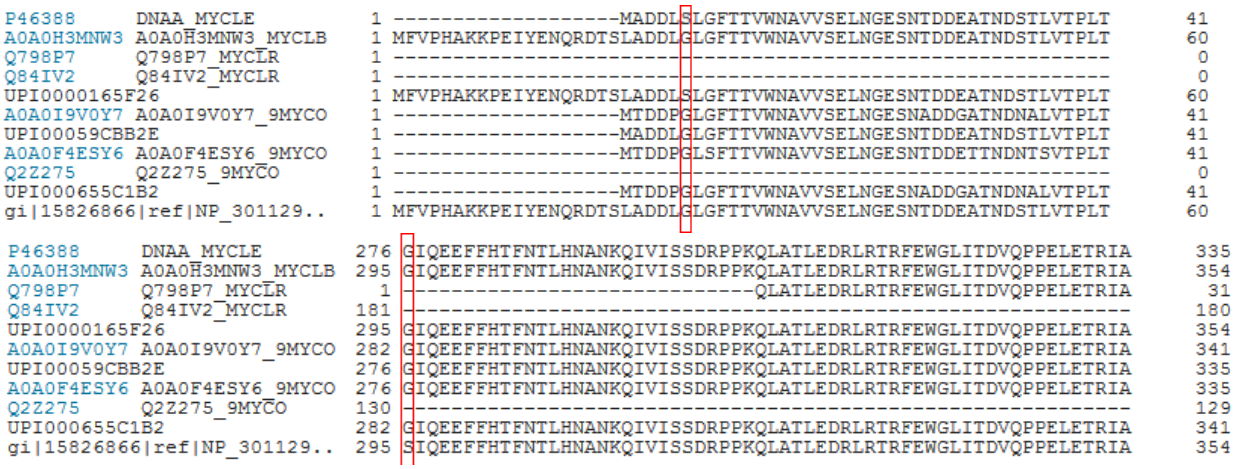


Figure 26: dnaA variant loci inside their conserved amino acid sequence

UP: locus 1, DOWN: locus 2

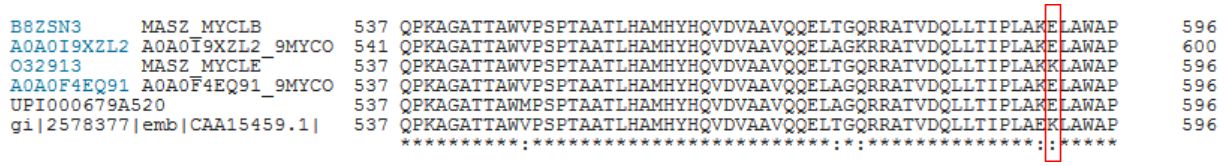


Figure 27: glcB variant locus inside its conserved amino acid sequence

Amino acid change lys591glu

4 Discussion

4.1 Authenticity of the results

4.1.1 aDNA good laboratory practices

Several reviews have discussed the necessity of setting up rigorous guidelines for good ancient DNA work in order to increase the quality of the results and avoid false findings (Shapiro & Hofreiter 2012; Cooper & Poinar 2000). The guidelines involve 1) the use of a physically-isolated area for pre-PCR work, 2) the systematic implementation of negative controls, 3) the control of PCR results with regards to the expected product length and sequence, 4) the setting up of biological and technological replicates, 5) the control of PCR-based variant detection using cloning, 6) the setting up of independent replication in other research groups when studying human DNA and 7) the testing of the recovered DNA for characteristic ancient DNA damage patterns using Next-Generation Sequencing.

The details of the measures taken within this study in relation to these guidelines are fully described in the corresponding methods section. In general, pre-PCR steps were carried out in a dedicated aDNA room exempt of any modern DNA work and were thoroughly maintained using cleaning procedures specifically intended to remove any and all DNA traces. Three types of negative controls were systematically used to assess contamination levels during the extraction, the PCRs and the library preparation. A separate scientist replicated all PCR-based assays and Next-Generation Sequencing datasets were checked for coherence between datasets of the same DNA extract. Adhesion to the guidelines specific to PCR or Next-Generation Sequencing are discussed below. Although independent replication of the results by another laboratory was not performed, all the results were compared to previously-published data from samples of similar historical context. As the study focusses on mycobacterial DNA, the risk of artifactual results due to human contamination is negligible.

In addition to the guidelines, the choice of the reagents and PCR oligonucleotides were carefully thought through. Indeed, contamination can happen through the use of contaminated reagents and consumables. To avoid indirect contamination by the reagents, all the reagents were chosen for their DNA-free certifications. Moreover, all reagents able to sustain UV radiation and room temperatures were de-contaminated under UV light between each use.

Because it followed the guidelines for good ancient DNA practices, this study places itself in the best situation to yield reliable results.

4.1.2 PCR results

Because of the molecular degradation of aDNA fragments, the polymerases used to perform PCRs inherently show a higher amplification rate for non-degraded contaminant DNA fragments (Willerslev & Cooper 2005; Cooper & Waynet 1998; Hofreiter 2008). Consequently, PCR is strongly sensitive to contamination and many aDNA PCR results have been heavily debated after publication due to missing negative controls or insufficient description of the measures taken to avoid contamination (Lindahl 1997; Hofreiter 2008; Barnes & Thomas 2006; Cooper & Waynet 1998; Stone et al. 2009; Willerslev & Cooper 2005; Wilbur et al. 2009).

Three criteria are essential to authenticate ancient DNA PCR results: negative controls must show no amplification, the PCR product length must be consistent with the expected product size and the PCR product sequence must match the reference sequence.

In a typical PCR amplification from an aDNA extract, contamination might arise from the extraction step or from the PCR set-up step. Here, two types of blanks were used to monitor the contamination levels at the end of the analysis: Extraction blanks and PCR blanks. If contamination happens during the extraction procedure, only the extraction blank should return positive. If both blank types return positive, this would reveal contamination during PCR setup. A PCR blank that shows contamination while the extraction blank remains clear could be a sign for cross-contamination between the reaction tubes during the PCR procedure.

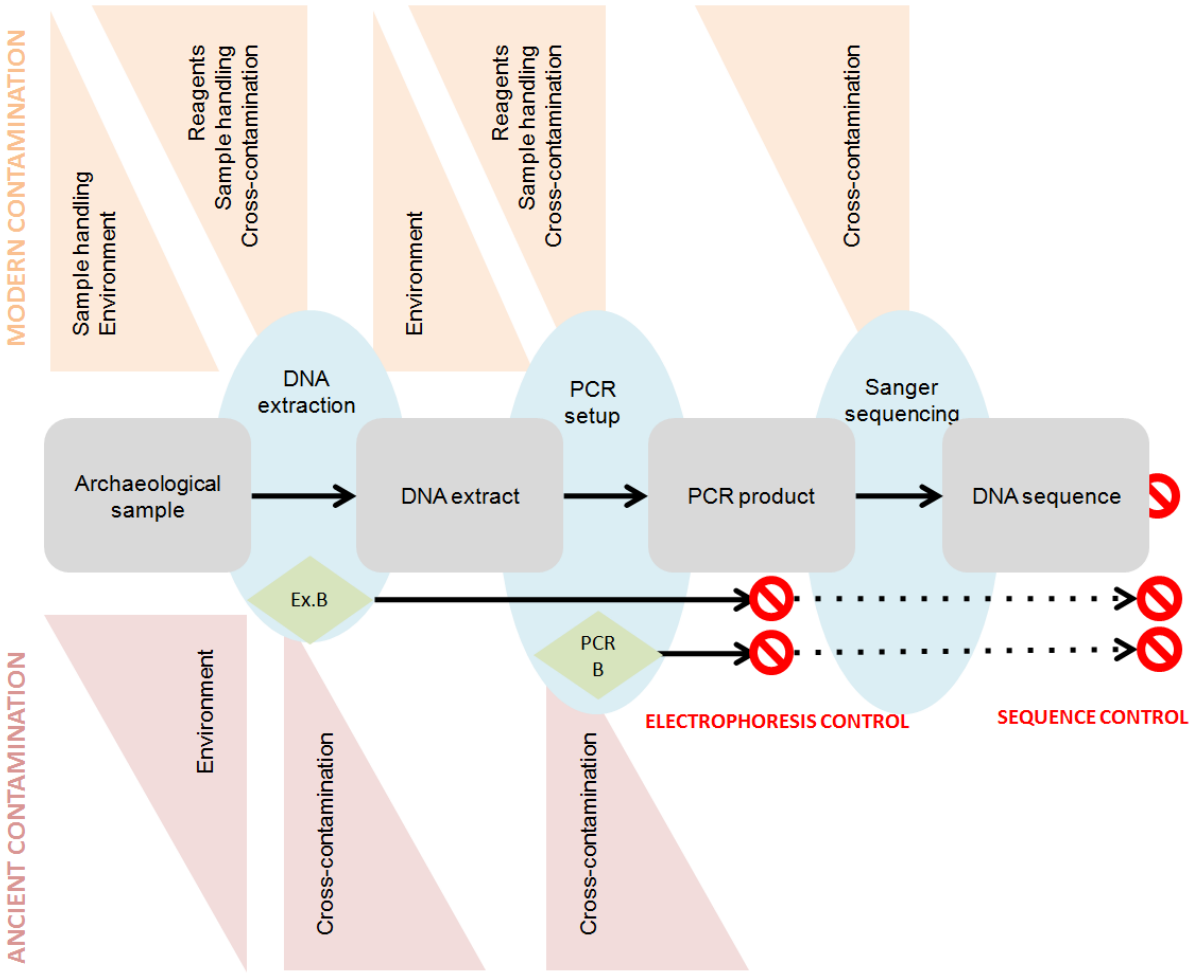


Figure 28: Sources of PCR contamination and measures taken to estimate its level
The extraction and PCR blank were created during the extraction and library blank, respectively, by replacing the volume of DNA solution with molecular-grade water. Blanks were then checked twice, once for the presence of amplification products, and (if necessary) a second time for the origin of the amplified sequence.

In this study, all extraction and PCR negative controls gave no PCR products for the primer pairs used. Moreover, apart from the case of *M. tuberculosis* (which will be discussed further below) all the PCR products obtained showed lengths and sequences consistent with the target sequences. Therefore, with regards to PCR contamination control procedures and checkpoints, this study fulfils the authenticity criteria discussed in the literature and the presented results are reliable.

4.1.3 Next-Generation Sequencing results

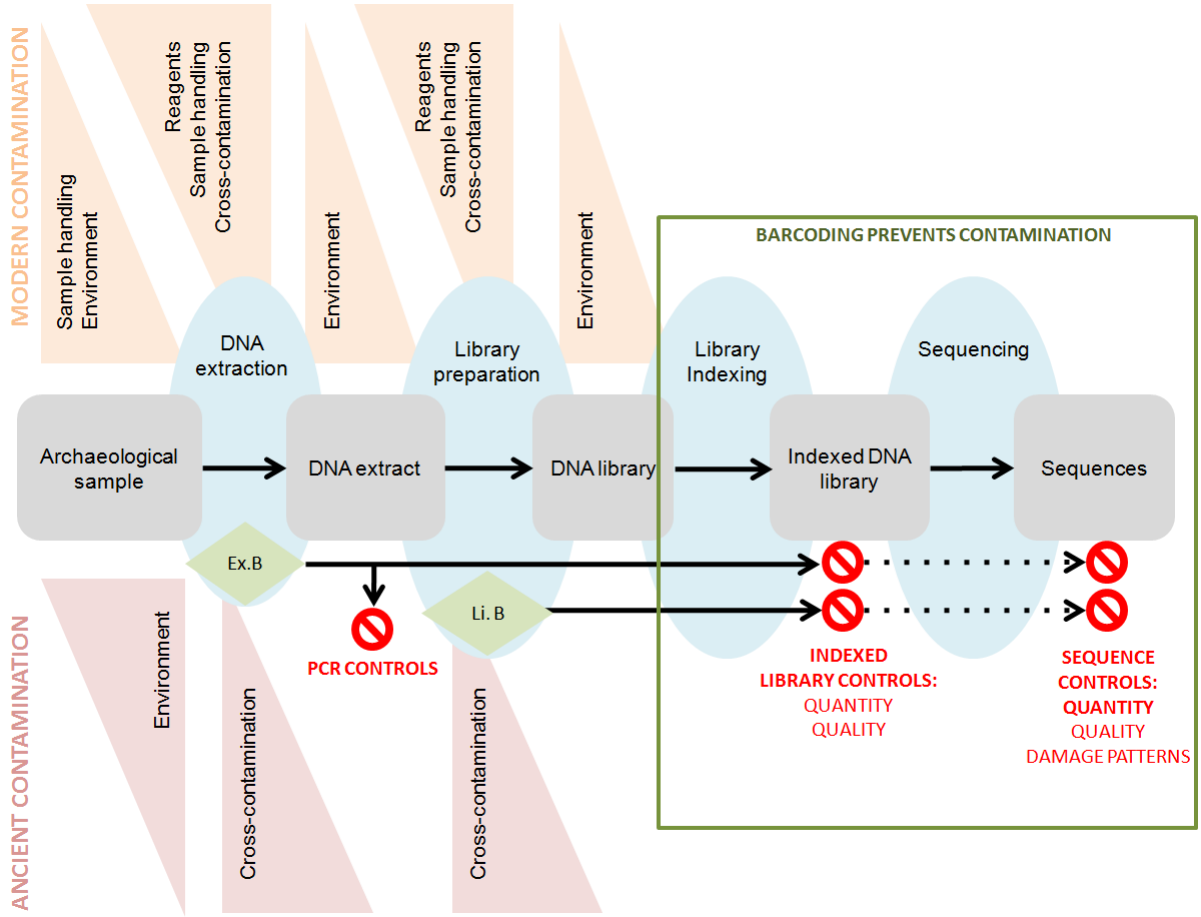


Figure 29: NGS sources of contamination and measures taken to avoid it
In addition to the PCR negative controls, several contamination control steps were taken during the NGS procedure. A library blank was added to each sample set. All the libraries were checked for concentration. Finally all the sequences were examined for degradation patterns characteristic of ancient DNA fragments.

The contamination or cross-contamination of samples is considered negligible after the indexing of the Illumina sequencing libraries because only the DNA sequences flagged with the expected barcode for a specific sample will be retrieved after sequencing (Kircher et al. 2012; Meyer & Kircher 2010; Cooper & Poinar 2000). Therefore, only contamination which happens before the indexing step presents a risk for the authenticity of the results. The contamination of the results might come from either the laboratory work or the sample itself (Figure 29). In addition, the DNA recovered for a species might not be ancient (even if present in the remains before extraction). These last three issues will each be addressed individually in the following.

The non-contamination of the DNA extracts by exogenous DNA was already assessed once during the PCR screening, as has been discussed previously. Library blanks help to assess the levels of

contamination during the library preparation and indexing procedures. The quality and quantity control of the extraction blank and library blanks after indexing of the Illumina sequencing libraries showed virtually empty libraries, sometimes presenting an adapter chimera peak around 100bp. In the case of sequencing blanks, chimeras are actually to be expected since the library preparation and indexing procedure is applied to a "DNA extract" composed only of water and buffers. Chimeras present little risk for the sequencing of blanks on a Illumina HiSeq, because the cluster generation step requires DNA fragments to be longer in order to properly form bridges (Kircher et al. 2012; Meyer & Kircher 2010). The absence of modern contamination is confirmed by the second control using the extraction and library blanks after sequencing, when the blanks showed no contaminating DNA sequences once again. Some blanks yielded small amount of small reads which were too short to map specifically anywhere. The origin of those small reads is still unclear, although they were likely brought in by one (or several) of the enzyme kits, considering that enzymes are mass-produced in modified cell lines such as *E. coli* and that their purification process is often not well described by the suppliers. All in all, the extraction blanks and library blanks showed no evidence of contamination with modern DNA after the sampling of the remains.

The endogenous DNA from the remains themselves might well carry contamination that dates from the time the skeleton spent buried or during its storage after excavation, or even from the osteological study of the remains. DNA molecules not processed and stored for molecular biological use will display the same features as "ancient DNA" after about 20 years (Mitchell et al. 2005; Brotherton et al. 2007; Overballe-Petersen et al. 2012; Sawyer et al. 2012). Therefore, skeletons that spent time in storage are likely to yield large numbers of ancient-looking DNA molecules from various organisms. To authenticate the results of any ancient DNA study, it is important to also discuss the issue of pre-sampling contamination. This type of contamination is rarely a risk for studies which rely on a mapping approach to identify the DNA fragments belonging to the species of interest. In this study, a mapping strategy was applied to discriminate *M. leprae* DNA fragments and human mtDNA fragments from the rest of the sequenced DNA molecules and strict mapping criteria were chosen to minimize the risk of aspecific mapping. In addition, species-specific genomic regions were chosen in order to evaluate the sequence coverage and read depth with a minimal background mapping level. Therefore, the results presented are unlikely to be artefacts due to the presence of close relative species within the sample.

The last step for the authentication of aDNA results is to show that the recovered DNA molecules for the species under study present the spontaneous molecular degradation characteristics which are commonly referred to as “damage patterns”. As was described in the introduction, after the death of an organism, DNA molecules start to accumulate molecular modifications. Amongst those modifications are the deamination of the purine bases (Lamers et al. 2009; Overballe-Petersen et al. 2012; Briggs et al. 2007; Brotherton et al. 2007) which leads to an increased occurrence of Guanine (G) and Thymine (T) at the 3’-end and 5’-end of the DNA molecule, respectively. This is accompanied by the apparition of single-strand breaks leading to a progressive fragmentation of the DNA molecule (Pääbo 1989; Hofreiter, Serre, et al. 2001; Willerslev & Cooper 2005). In consequence, the authentification of aDNA should not rely only on the damage patterns, but should instead be discussed in relation to fragment length distribution. In this study, two results were used to evaluate the length of the DNA fragments. The first of these was the quality control of the indexed sequencing libraries. The second was the insert length distribution provided by the software which also calculated the damage patterns. The first was only an approximation since the plot showed the length distribution of the insert plus both side library and indexing adapters. However, it also provided the first evaluation of the fragment lengths: given the length of the adapter constructs, the total length with aDNA insert should not exceed 300bp. The insert distribution length provided by the mapDamage 2.0 software gives a better estimate of the length of the fragments mapping the genome of the species under study. Moreover, as it is calculated from the exact same dataset as the one used to compute the misincorporation frequencies at the ends of the DNA fragments, both results can be directly compared. In all the datasets, the fragment length distribution observed was consistent with the ancient origin of the recovered genomes.

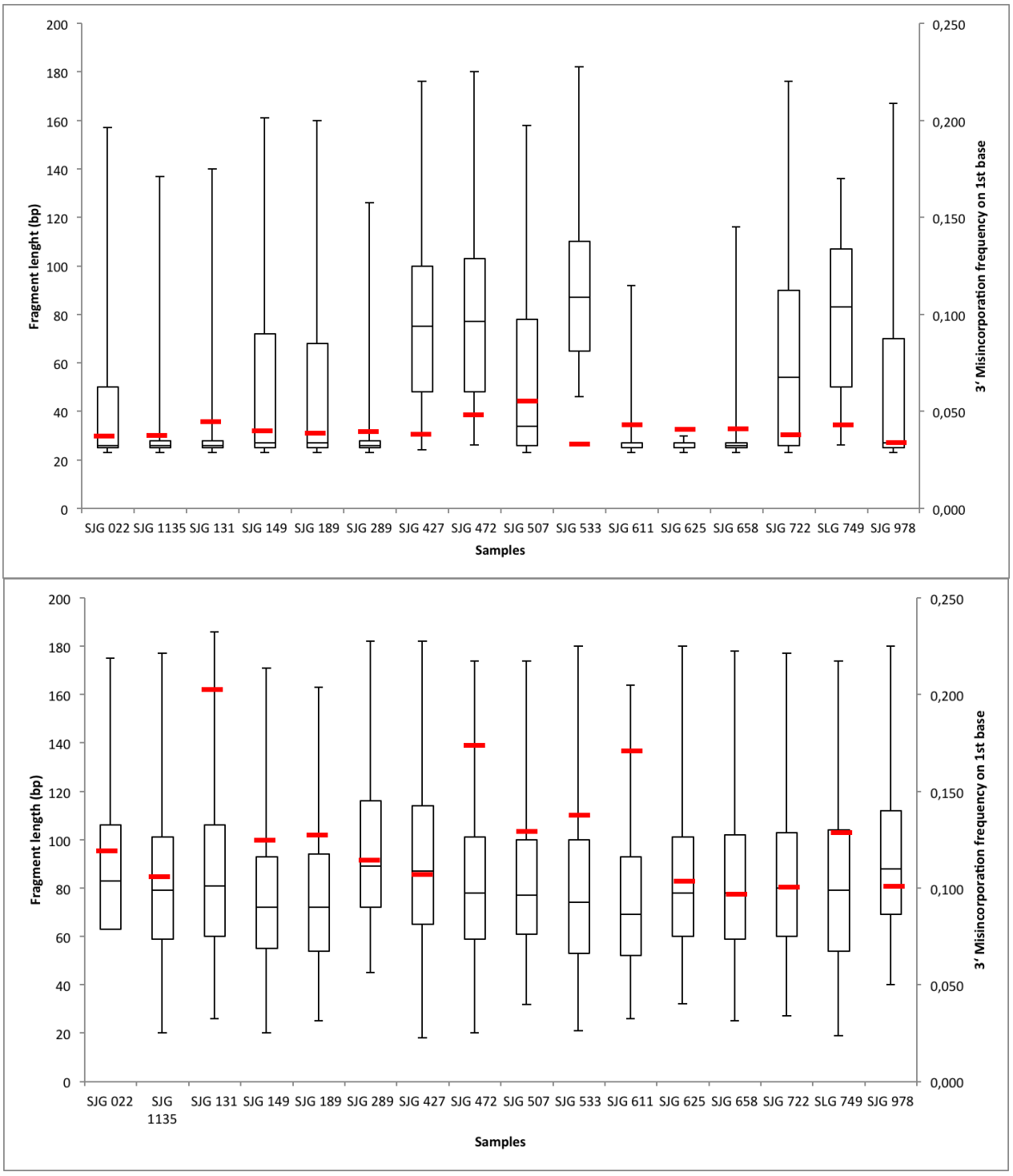


Figure 30: Correlation between aDNA length distribution and damage patterns
G to A frequency on the first 3' base of the fragments is displayed in red. UP: Data from the reads mapping to the M. leprae *genome. DOWN: Data from the reads mapping to the human mtDNA.*

The damage patterns observed for the human mtDNA fragments ranged from 0.076 to 0.150 on the 5' end (C->T) and 0.097 to 0.200 on the 3' end (G->A), numbers that are consistent with previously published data (Schuenemann et al. 2013). The damage patterns observed for the *M.*

leprae DNA fragments ranged from 0.033 to 0.051 on the 5' end and 0.030 to 0.052 on the 3' end. Those numbers are also consistent with several previous medieval *M. leprae* studies (Schuenemann et al. 2013). Interestingly, for the human mtDNA, the 5'-end of the molecules always displayed higher misincorporation frequencies than the 3'-end, a phenomenon which has not often been observed within *M. leprae* data. A recent study published after the laboratory work of this project showed that the use of Accuprime polymerase to amplify indexed libraries tended to reduce the misincorporation frequency on the 5'-end of the DNA molecules (Seguin-Orlando et al. 2015). The observation was made on ancient equid DNA and is consistent with the damage patterns shown by the human mtDNA sequences. It is curious that this bias seems almost negligible for the mycobacterial DNA. Several publications have suggested that *M. leprae's* thick and highly hydrophobic cell wall enhances the preservation of its DNA after death (Schuenemann et al. 2013; Mendum et al. 2014). The data presented here supports this hypothesis, as the damage patterns were generally higher for the human mtDNA than for *the* M. leprae DNA. It is likely that this differential DNA preservation is, at least partly, also involved in the different behaviour of the DNA libraries when amplified with Accuprime.

The mapDamage results showed an apparent conflict between the damage patterns and the length distributions. On one side, the damage pattern data suggests that *M. leprae* is better preserved than the human mtDNA (lower misincorporation frequencies at the end of the aDNA fragments). However, the distribution plots would instead suggest that the reverse were true: the human mtDNA was the best preserved as it presented longer DNA fragments. In fact, these results do not actually contradict each other. Indeed, although the DNA molecular degradation processes leading to misincorporations and strand breaks are independent, the apparition of a strand break in a DNA fragment creates new 3' and 5'- ends which display lower misincorporation frequencies, decreasing the average misincorporation frequencies over all 5' and 3' fragment ends. Therefore, from two samples of the same age and deamination rate over time, the sample with the highest strand break rate over time will show lower damage patterns at the end of the DNA fragments (see Figure 31).

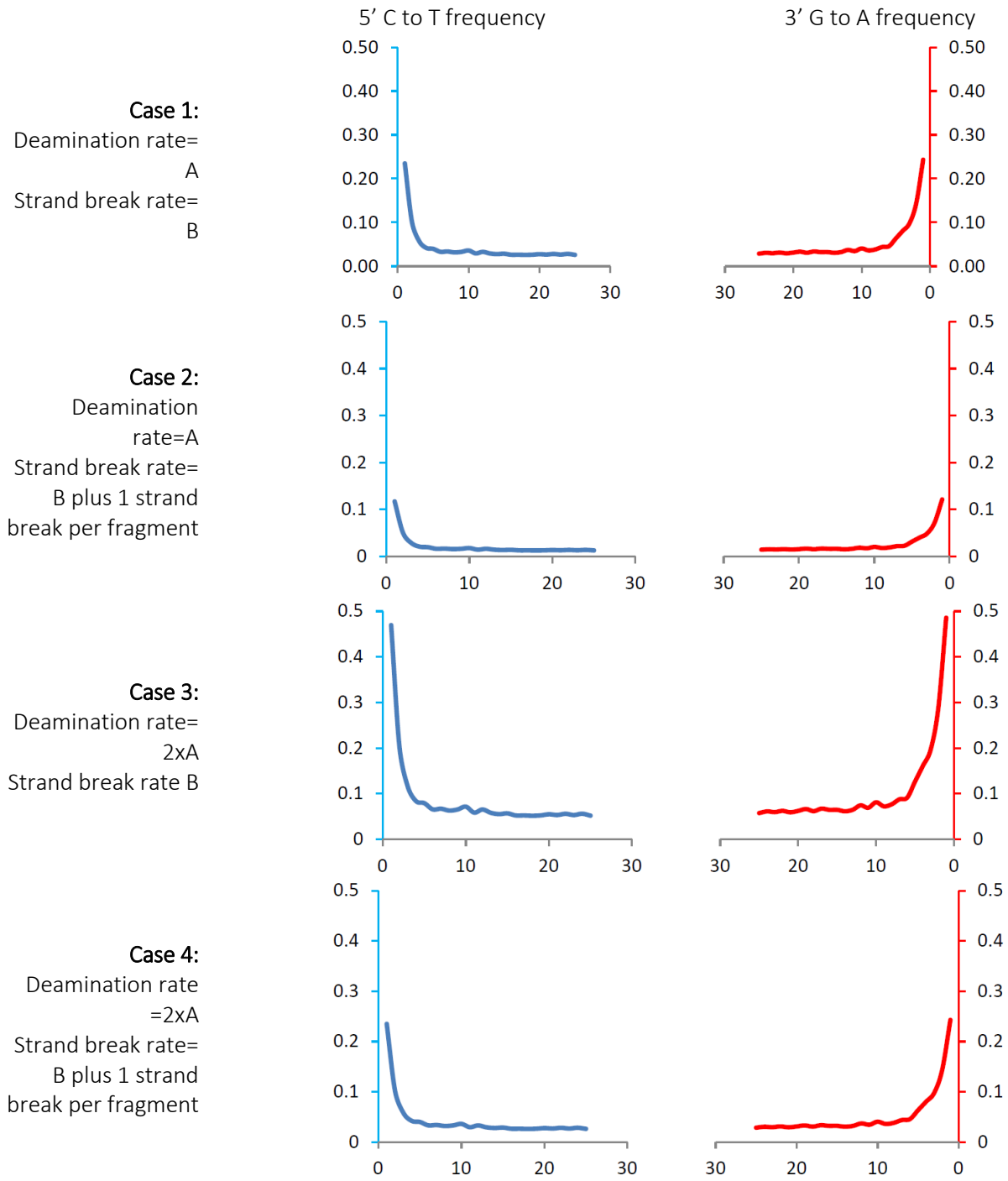


Figure 31: Influence of deamination and strand breaks on damage patterns

The dataset from one of the samples from this study was used to simulate simple cases when the rates of deamination and strand break over time vary. To simplify the situation, this simulation considers deamination and strand breaks as the only degradation mechanisms and assumes that strand breaks would occur in the middle of aDNA fragments.

In summary, the results showed clean blanks before and after sequencing and damage patterns consistent with the medieval origins of the remains. The possibility of exogenous contamination of the samples is considered negligible after the indexing of the Illumina sequencing libraries because only the DNA sequences flagged with the expected barcode for a specific sample would be retrieved after sequencing (Kircher et al. 2012; Meyer & Kircher 2010; Cooper & Poinar 2000). Moreover, as a mapping strategy is used to reconstruct the mycobacterial genomes, the likelihood of endogenous contamination is strictly limited to DNA from close relatives of the species under study. Therefore, the Next-Generation Sequencing results from this study can also be considered reliable.

4.1.4 Age of the samples

Table 25: Comparison between radiocarbon dating and burial period

Sample	Calibrated radiocarbon age	Burial period (arm position)
SJG 404	cal AD 1226-1268	1000-1375
SJG 427	cal AD 1177-1250	1000-1375
SJG 472	cal AD 1252-1280	1000-1375
SJG 507	cal AD 1170-1214	1000-1375
SJG 533	cal AD 1051-1206	1000-1375
SJG 722	cal AD 1269-1285	1000-1250
SJG 749	cal AD 1230-1277	1000-1375
SJG 978	cal AD 1283-1381	1000-1375
SJG 1137	cal AD 1282-1380	1000-1250

The burial period is an estimate of the sample age based on burial practice as determined by the archaeological collaborators within this project. Two of the samples which exhibited apparently inconsistent results are highlighted in orange.

The radiocarbon dating results were consistent with the archaeological determination based on burial practices in all but two samples. Sample SJG 722 also showed sexing results inconsistent with the archaeological data. These two problematic areas for this sample are most likely linked and will be discussed later on.

4.2 DNA preservation in the various cemeteries

4.2.1 Human mitochondrial DNA preservation

The human mtDNA PCRs were performed in order to estimate the human DNA preservation levels in the samples and to provide a baseline from which to interpret the results of the pathogen DNA PCRs. The St. Jørgen cemetery showed the highest number of human mtDNA PCR positives, suggesting that it is the best-preserved cemetery of this study. With almost 50% of the samples yielding mtDNA PCR products, it counts among the best-preserved of the cemeteries that have been used in ancient DNA studies to date. The two last large sample sets from Ribe and the Rathausmarkt showed lower human DNA preservation. The preservation of human mtDNA in the remains was apparently unrelated with the sex, age at death or historical age of the remains.

4.2.2 *M. tuberculosis* DNA preservation

The PCR results for *M. tuberculosis* yielded no clear tuberculosis positive. Several other attempts were performed as well as PCR optimisation and extensive troubleshooting, all without results. The PCR optimisation assays included the testing of two new annealing temperatures (+2°C and -2°C compared to the published annealing temperatures) and the testing of two new template DNA input volumes (2µL and 10µL). The PCR troubleshooting included the testing of new oligonucleotides, new reagents and a different PCR machine. As the PCR setup was also tested successfully with modern *M. tuberculosis* DNA, the absence of results for the ancient samples suggests that either the bacterial DNA was not present in the samples (i.e. the individuals were not infected with tuberculosis), or that the *M. tuberculosis* DNA was present, albeit in amounts that were too small or that were too degraded to be successfully recovered with the protocols used in this study.

For the three samples which showed partial *M. tuberculosis* DNA preservation, the presence of the pathogen seems uncorrelated with any of the archaeological information which is available.

4.2.*3* *M. leprae* DNA preservation

The number of samples which showed positive results for *M. leprae* DNA was null for all but one cemetery (St. Jørgen). The two other large sample sets yielded no positives for *M. leprae* DNA, which was rather surprising. Indeed, the samples were collected from individuals whose remains demonstrated leprosy lesions and who were collected from cemeteries whose use lives coincided with the period during which leprosy was highly prevalent in the region. However, DNA preservation is highly dependent on burial conditions and varies greatly not only from one cemetery to another but also from one species to the other. Therefore, the absence of a PCR positive for *M. leprae* DNA in the Ribe and Rathaus Markt cemeteries might simply be due to poor DNA preservation as a whole. This interpretation would be compatible with the lower preservation levels of human mtDNA within those two cemeteries as compared with that at St. Jørgen.

The St. Jørgen cemetery preservation level for *M. leprae* was extremely high compared to that of other cemeteries previously studied (Schuenemann et al. 2013; Taylor et al. 2013). Indeed, almost 50% of the samples showed positive *M. leprae* fragments. This number was not completely unexpected, as the individuals buried in the cemetery either were diagnosed with leprosy during their lifetime, or lived in close contact with infected people for a prolonged time. Moreover, a study published in 2013 had already showed very high *M. leprae* preservation in one St. Jørgen sample (Schuenemann et al. 2013), thereby suggesting that the burial environment at the site was favourable to the recovery of medieval mycobacterial DNA. Nevertheless, the number of samples which showed *M. leprae* DNA preservation goes far beyond current achievements in terms of positive leprosy molecular identifications from the same geographical origin.

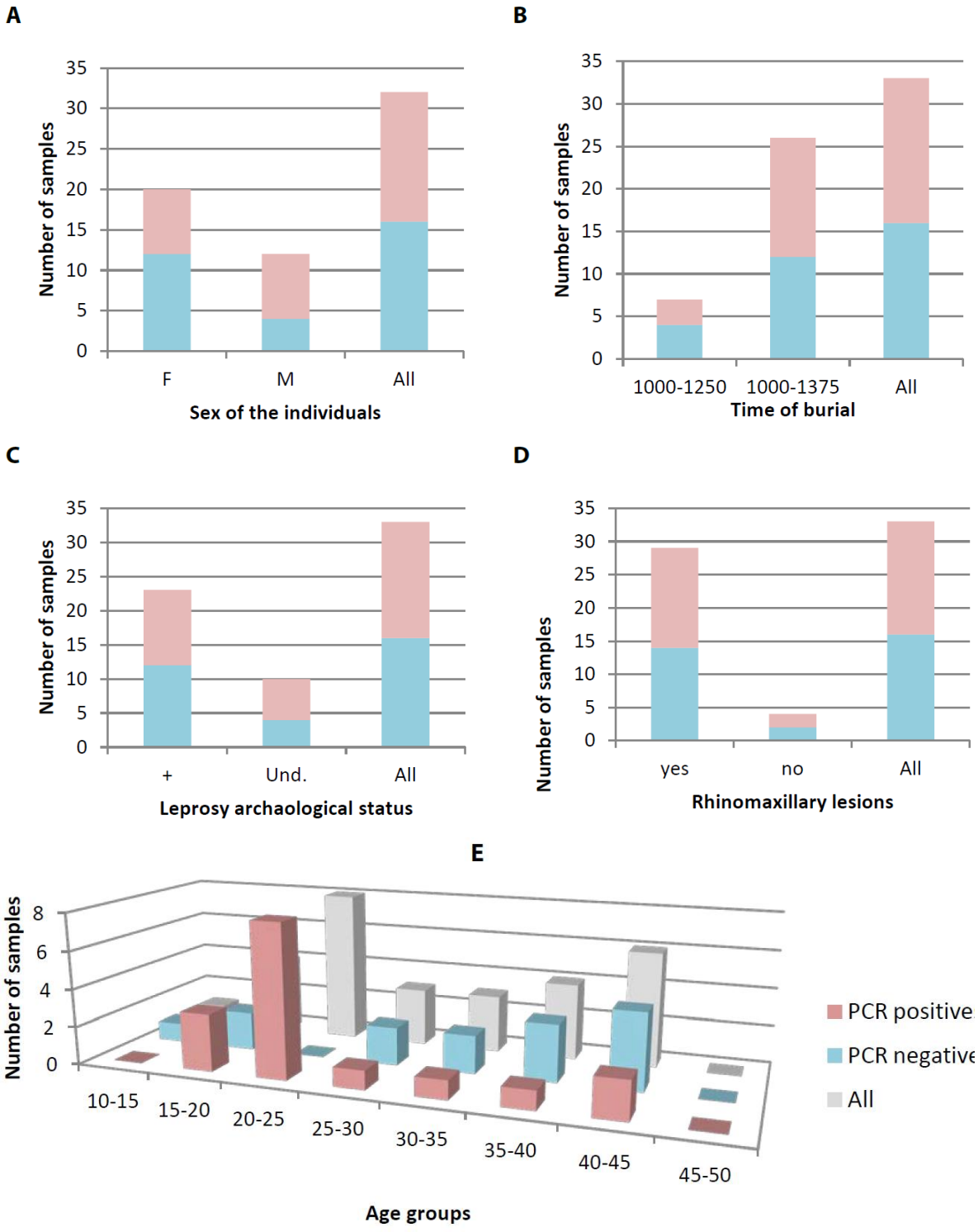


Figure 32: PCR results in relation with the archaeological features.
A) Proportion of PCR positives according to the archaeological sex of the samples. B) Proportion of PCR positive according to burial time period. C) Proportion of PCR positives according to palaeopathological status. D) Proportion of PCR positives according to whether or not the individuals presented rhinomaxillary lesions. E) Age at death distribution amongst the PCR positives and negatives.

The number of samples yielding PCR positive results seems to be independent of the age at death of the sample and the sex of the individuals. In addition, the number of PCR positives does not seem to be correlated with the presence of leprosy-specific bone lesions (with or without rhinomaxillary lesions). This might be due to a sampling bias, as all the remains presented here originated from a leprosarium hospital cemetery. Although the number of samples does not allow for the quantification of the correlation between age at death and the proportion of PCR positives, there is a clear trend for individuals between 20 and 25 to yield PCR positive results for *M. leprae*. This cannot be explained only by a DNA preservation difference and would suggest a higher number of *M. leprae* cells in the organisms of the 20-25 year old individuals before their death. The development of leprosy towards its multibacillary form is linked to the type of immune response of the affected individual and is dependent on his or her life quality. Therefore, if confirmed, the higher pathogen load in medieval 20-25 year olds might indicate that individuals around that age tended to respond to the infection through antibody-mediated immunity, which has been shown to promote the immune evasion of *M. leprae*.

4.3 Next-Generation Sequencing reads pre-processing

4.3.1 Data quality

The FastQC sequence quality controls returned numerous warnings. After careful examination, all the warnings likely originated from ancient DNA molecular characteristics (see Table 26). Therefore, it can be assumed that no major technical failure took place during the library preparation, indexing or sequencing. In addition to the preliminary FastQC control, several bioinformatics steps are designed to optimise the quality of the datasets for each sample before mapping the reads to the reference genomes. The first step is the trimming of the adapter sequences which are likely to be sequenced due to the short length of aDNA insert molecules. Those adapter sequences are removed to prevent errors during the mapping step and increase general mapping quality. After adapter trimming, the forward and reverse sequencing reads are mapped. This increases the length of the DNA sequences that will be used for the mapping step, therefore decreasing the risks of aspecific mapping. Moreover, merging raises the per base sequencing quality and decreases the number of sequencing errors by selecting the base with the best quality score in case of conflict between forward and reverse sequences. Finally, the merged

reads are trimmed if the sequencing quality drops towards the end of the read and filtered to remove the reads which are too short to be accurately mapped.

Table 26: FastQC observed warnings and possible aDNA cause

Step	Control for	Possible explanation
2 3	Per base sequence quality Per tile sequence quality	Ancient DNA short reads are likely to enhance the per base sequence quality drop towards the ends of the reads
5	Per base sequence content	The formation of the aDNA library and indexing are likely to lead to over-representation of adapter sequences within the sample
6	Per base GC content	The shotgun sequencing approach used means that the DNA sequence likely originates from several organisms with various GC contents.
9	Sequence duplication levels	aDNA library diversity is commonly reduced compared to modern DNA libraries; therefore, this warning is likely to be triggered by the presence of numerous PCR duplicates.
10	Overrepresented sequences	aDNA is not homogenously preserved in a genome; therefore, the reads from regions with high coverage are likely to seem over-represented
11	Adapter content	Ancient DNA inserts are usually shorter than the read lengths; therefore, the sequencing adapters are likely to be sequenced
12	K-mer content	aDNA is not homogenously preserved in a genome; therefore, the reads from regions with high coverage are likely to appear over-represented

A rather large variability was observed between the raw numbers of reads available for each sample after sequencing, although the protocols were designed to input the samples in equimolar amounts. The causes of this variability are difficult to investigate. Pipetting uncertainty is likely involved, as well the cluster generation method which might be uneven when fragments of various lengths and GC contents are processed in parallel.

4.3.2 Identification and analysis of human DNA reads

Human mitochondrial DNA reads were identified by mapping the sequencing reads to the GRCh38 mitochondrial DNA (mtDNA) reference sequence (NC_012920.1). The percentage of reads mapping to the human mtDNA also fits in the frame of previous studies involving samples from the same geographical and historical contexts (Schuenemann et al. 2013). The coverage and read depth over the human mtDNA were also in agreement with previous studies.

The sex typing proxy developed seems to have been reliable; eight out of 16 samples showed results which matched osteological sex. However, two samples showed an Y/X ratio inconsistent with their osteologically-determined sex. To control the results obtained with the Y/X ratio proxy, STR typing was performed (by Lisa Böhme for her PhD). For the two samples showing Y/X ratios incompatible with the osteological sexing, the STR analysis confirmed the Y/X ratio results (see Table 27). An investigation was undertaken with the archaeologists in charge of the sample collections in Odense to understand the origin of the contradiction. It was discovered that for those two samples, a loose tooth was collected from the numerous loose teeth labelled as belonging to this individual. In the case of very close graves, teeth might have been exchanged between individuals during the excavation, as it is very difficult to properly assess the origin of a tooth when the maxillae are not well preserved. Therefore, it is possible that the material used in samples SJG 722 and SJG 978 actually came from other individuals. For the remainder of this text, these two samples will be referred to as SJG 722* and SJG 978*. In the case of sample SJG 722, this differential between intended sample grave and actual sample grave also explains the discrepancy between the radiocarbon date and the archaeological estimate of the burial period which was discussed above.

Table 27: Results and comparison of the various human aDNA analyses performed

Sample	Sex typing		
	NGS data	STR typing*	Osteology*
SJG 022	Und.	Und.	F
SJG 131	F	Und.	F
SJG 149	Und.	F	F
SJG 189	Und.	F	F
SJG 289	M	M	Und.
SJG 404	M	M	M
SJG 427	F	F	F
SJG 472	F	F	F
SJG 507	F	F	F
SJG 533	Und.	F	F
SJG 611	M	M	M
SJG 658	M	M	M
SJG 722	F	F	M
SJG 749	F	Und.	F
SJG 978	M	M	F
SJG 1137	F	F	F

*Columns marked with * indicate results not obtained personally. The label "Und." denotes inconclusive results. Red cell shading highlights inconsistencies.*

4.3.3 Identification of *M. leprae* DNA reads

Mycobacterium leprae DNA reads were identified by mapping the sequencing reads to the TN strain reference sequence. The percentage of raw sequencing reads mapping to the reference ranged from 0.01% to about 12%. Due to the close genetic similarity between the mycobacterial species, the presence of *M. leprae* was confirmed using the results of the mapping onto the two *M. leprae*-specific genomic targets. The overall coverage and read depth statistics were interpreted while taking into account the 38-44% of genetic identity observed during the preliminary analyses. The other references used as mapping comparisons all yielded similar low coverage and read depth. In addition, there was a clear positive correlation between the coverage of *M. leprae* and the coverage of the other species. In addition, the coverage statistic on the *M. leprae*-specific targets were similar or higher compared to the whole-genome data. Those results suggest that the reads which mapped to the other species were likely *M. leprae* DNA fragments which mapped aspecifically due to sequence similarity.

The Next-Generation Sequencing results confirmed the presence of *M. leprae* in all the samples selected using the PCR screening. However, the levels of DNA recovery are highly variable. Some samples yielded only a low coverage over the *M. leprae* reference with a number of raw reads and a proportion of PCR duplicates which suggest low library complexity. Therefore, only seven samples were selected for re-sequencing.

4.3.4 Testing for the presence of other infectious pathogens

The mapping-to-targets approach was repeated with three other mycobacterial reference genomes *(M. lepromatosis, M. tuberculosis* and *M. bovis*) and *Y. pestis*. All four species might have been present during the time period studied. They represent, therefore, good co-infection candidates as well as good aspecific read mapping controls. The results of the mapping to the reference genomes all showed read mapping. The statistics of the mapping to the species-specific targets ruled out the presence of *M. bovis* or *M. lepromatosis*, as the coverage of the target was systematically significantly lower than the coverage of the whole reference genome. The mapping statistics for the *Y. pestis* target regions are slightly more difficult to interpret, as the target was only chosen to be present in *Y. pestis* but not in the mycobacterial genomes. Given the low coverage obtained, it is likely that the reads originated from another *Yersinia* species.

No evidence of *M. tuberculosis* infection was found after sequencing. This result is consistent with the absence of positive PCR for *M. tuberculosis* DNA in the first phase of this study. The coverage statistics for the *M. tuberculosis* reference showed increased coverage on the targets for five samples, which might be explained by the extremely high coverage of *M. leprae*. Indeed, after re-sequencing, the *M. tuberculosis* coverage statistics on the targets no longer showed the increase which was previously observed. The absence of *M. tuberculosis* is rather surprising, since the three cemeteries studied were in use during the rise of tuberculosis in Northern Europe. It is an arduous task to determine whether the methods used were sufficiently sensitive. It has been suggested that *M. tuberculosis* DNA is more difficult to recover than DNA from its cousin *M. leprae* (Harkins et al. 2015; Taylor et al. 2013). Indeed, *M. leprae* possesses a thicker cell wall than *M. tuberculosis* (Donoghue 2013) with different mycolic acids. Therefore, *M. tuberculosis* DNA might be more susceptible to environmental factors promoting DNA degradation (such as humidity). In addition, the affinity of *M. leprae* for teeth is rather well documented, whereas there is no data on the affinity of *M. tuberculosis* for the oral cavity (Boldsen 2007). As a consequence, teeth might not be the best starting material to recover *M. tuberculosis* DNA from ancient tuberculosis sufferers.

4.3.5 Relative abundance of human and mycobacterial DNA

The percentage of reads obtained for the human mtDNA is consistent with the reference size as well as with previously-published aDNA results. However, the fraction of reads mapping to the *M. leprae* genome is extremely high for some samples. Only one previous study reports such a proportion of reads mapping to the *M. leprae* reference genome after shotgun sequencing and without enrichment (Schuenemann et al. 2013). In fact, the sample described in this previous study comes from the same cemetery, suggesting that this cemetery's environment favorises high *M. leprae* DNA preservation.

The results from the current study concur with the previous observations which place *M. leprae* as the organism with the best *post-mortem* DNA preservation in teeth DNA extracts. The *M. leprae* DNA fragments recovered are indeed present in high amounts and with lower signs of degradation than the mitochondrial DNA of the human host. The reasons for the difference in preservation are not yet fully understood, but there is an agreement around the likely involvement of *M. leprae*'s thick waxy coat composed of highly hydrophobic mycolic acids (Schuenemann et al. 2013).

4.3.6 Mycobacterial plasmids

Only a few mycobacterial plasmid references were found to show sequencing read mapping (pTYGi9, pMYCCH01, pMYCSM01, pMYCSM03 and pMSPYR101). It was not possible, however, to rule out unspecific mapping or contamination. Indeed, not more than 2% of each reference was covered. Moreover, the number of mapping reads was too low to perform damage pattern computation. If the DNA fragments are of authentic ancient origin, the low number of reads might be a side-effect of the small length of the plasmid genomes. However, the methods used in this study are insufficient for further analysis. The results only suggest that the preservation of mycobacterial plasmid sequences seems possible. The sequencing of ancient mycobacterial plasmids has not yet been performed and plasmids are known to greatly influence virulence in bacterial pathogen species (Wang et al. 2016). Therefore, given the trends observed in this study, it seems both possible and pertinent that more resources should be brought to bear on the recovery of ancient mycobacterial plasmid sequences. Methods such as targeted enrichment captures might help to overcome the hurdle presented by the low number of mapping reads and confirm the results using damage patterns as well as phylogenetic downstream studies.

4.4 Reconstruction and analysis of medieval *M. leprae* genomes

4.4.1 Coverage and read depth of the *M. leprae* genome drafts

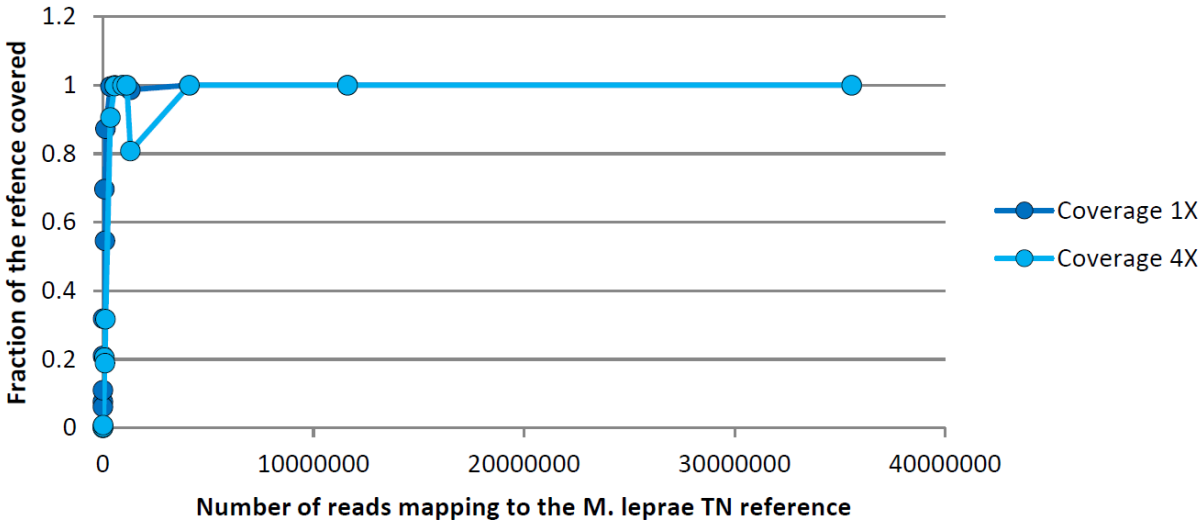


Figure 33: M. leprae *TN reference covered depending on the number of mapping reads*

While the number of data points is not sufficient for quantitative analysis, the minimum number of reads mapping to recover a near-complete ancient *M. leprae* genome from samples from this cemetery would be about 4 million. Considering that only roughly 0.05-13 % of the raw reads map to the *M. leprae* reference in this dataset, the sequencing reads should be set up to provide 350-400 million reads per sample. For most of the samples, however, this represents two sequencing lanes, which is still too expensive for most research groups.

Eight (8) out of the 16 samples sequenced yielded very highly-covered genomes with up to a read depth of 600 X. While this is rather unexpected for aDNA studies, in the literature there is an example of a sample from this cemetery which yielded high amounts of *M. leprae* DNA (Schuenemann et al. 2013). The DNA preservation in those eight samples was sufficient to also perform *de-novo* assemblies of the ancient mycobacterial genomes (by J. Suzat for his master thesis, data not shown).

4.4.2 Variant calling and effect prediction

The effect distribution along the reference genome consistently showed a frequency peak between positions 1,300,000 and 1,400,000. The *M. leprae* genomic annotations showed the presence of three rRNA genes in this region (Figure 34). As rRNA sequences are highly variable regions, the observed peak in variant frequency is likely a consequence of the presence of those genes and is not an indication of a virulence-related variation hotspot.

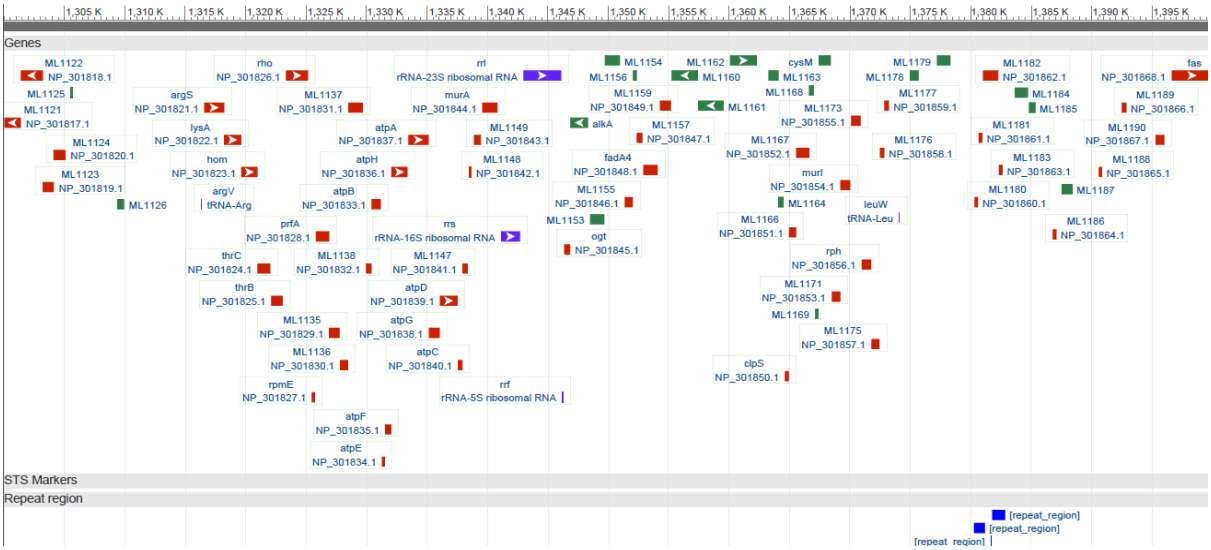


Figure 34: M. leprae *annotation context between positions 1,300,000 and 1,400,000*
This region shows the highest frequency of variants, most likely because of the presence of the sequences coding for the rRNA 5S, 16S and 32S that are highly variable.

After the first filtering steps, only two variants of possible interest were found. One was a variant annotated as related to adaptation and virulence, the other a variant in a splice site. Unfortunately, both were missing critical protein contextual information and their effects could not be studied in detail. The three proteins eventually selected were cytoplasmic proteins. dnA and gyrA are proteins which interact with DNA and glcB is involved in malate metabolic pathways.

M. leprae dnaA (ML0001) is a chromosomal replication initiator protein whose function and features have been determined through sequence homology (Cole et al. 2001). It plays an important role in the initiation and regulation of chromosomal replication (Bateman et al. 2015; Suzek et al. 2015) and possesses binding sites for DNA (dnaA-box), ATP and acidic phospholipids. In the sample set for this study, two variants were found in the protein sequence (ser25gly and gly295ser). No modification of the overall protein chemical properties could be noticed when one or both variants were present. However, the in-depth study of the variant loci and context showed that both are located in regions conserved amongst members of the dnaA protein family and are predicted to induce a change in the secondary protein structure. The second variant in particular is located in the ATP binding site and is predicted to locally modify an α-helix near one of the three binding site's highly conserved amino acid motifs (Walker B).

M. leprae gyrA (ML0006) is an ATP-dependent DNA gyrase subunit. The GyrA chain is responsible for the separation and rejoining of the two strands of DNA during the replication process. Its features and functions have been predicted using its homology with *M. tuberculosis* gyrA. It has been shown to possess a HOP domain as well as an Intein chain. In the samples analyzed for this study, one variant was found in the HOP domain (leu379pro) and is predicted to locally change the secondary structure of the protein by disrupting a long alpha-helix. It was, however, located outside of the main conserved motifs and is, therefore, considered unlikely to have major consequences on protein activity.

M. leprae glcB (ML2029) is predicted to be homologous to *M. tuberculosis* malate synthase G and to take part in intermediary metabolism processes. One variant was found in the conserved region of glcB acetyl-coA binding site (lys591glu). This variant was predicted to have very little effect on the secondary structure of the protein.

In all three studied cases, the amino acid change was predicted to have a local effect on the secondary structure of the protein and, in most of the cases, the variants were found inside conserved protein regions. However, our knowledge of the protein expression and activity in *M. leprae* is very limited (due to the impossibility of cultivating the bacterium). The only information available comes from *in-silico* predictions and comparisions with *M. tuberculosis*. Little is known

about the specificity of each protein's cellular location, biochemical properties, structure and importance for *M. leprae* fitness or virulence. Therefore, it is impossible go any further than the protein's primary and secondary structure when attempting to predict the effects of a variant. While it is impossible to clearly quantify the effect of the variants on the protein's function and to assess their influence on *M. leprae's* overall fitness and virulence, it was still possible to detail the variant effects up to secondary structure. Such an evaluation has never been previously attempted on variants originating from ancient mycobacterial genomes and could easily be scaled up to all the variants found in well-documented proteins. Moreover, while it was not the topic of this study, it is likely that protein modeling by homology could detail the consequences of the variants on the activity levels of the protein. However, to determine whether or not a variant might influence pathogen fitness and virulence, more biological background information is needed.

4.4.3 Strain typing

Table 28: SNP types of the recovered genomes

Sample	Most likely SNP type according to Monot et al (2005)	SNP type according to Singh et al (2011)
SJG 022	Und.	Nd
SJG 131	Und.	1d
SJG 149	Und.	Nd
SJG 189	Und.	Und.
SJG 289	Und.	Nd
SJG 404	Und.	3I
SJG 427	Und.	3I
SJG 472	3I	3I
SJG 507	3I	3K
SJG 533	Und.	Und.
SJG 611		
SJG 658	Und.	1d
SJG 722	3I	3I
SJG 749	2F	2F
SJG 978	Und.	2F
SJG 1137	Und.	3I

The results from the various typing methods were consistent for most samples. In the few cases where there were discrepancies, the typing results from the Singh et al. approach were used, as this method is considered the most accurate (Singh & Cole 2011). Inconsistencies from the other

methods are likely due to the low coverage of some recovered genomes or to the high genetic proximity between strains. No association could be observed between the *M. leprae* strains, the age of the samples and the osteological features. This observation, however, must be considered in relation to the low number of samples. Five (5) samples were typed as 3I, two (2) as 2F and one (1) as 3K. The implications of the presence of those strains is discussed below.

4.5 Phylogenetic placement of *M.leprae* strains

4.5.1 Consistency between the phylogenetic trees

Three types of trees (Maximum-Likelihood, Maximum-Parsimony and Neighbour Joining) were constructed from two different whole-genome multiple alignments of the six new, highly-covered genomes and the previously-published ancient and modern genomes. All three types of tree showed consistent results insofar as each new sample clustered with previously-published strains of the same type. The overall phylogeny is also consistent with previously-published trees, with four major branches corresponding to groups 1 to 4. In addition, sample SJG 507 clusters in every tree with modern 3K strains, confirming the existence of the recently-described fourth branch (0).

Only one inconsistency was noticed between the SNP typing results and the branching of the new train in the phylogeny (sample SGJ 533). Indeed, the *M. leprae* strain from SJG 533 clusters within Branch 3 with its closest neighbour being an ancient 3I strain from Denmark. Its strain typing, however, yielded inconsistent results insofar as the first SNP trio was different from the four possible combinations. Another apparent inconsistency was observed with the modern sequence s15 (type 3L). Depending on the trees, the 3L strain clusters either with 3K strains or with 3I strains. This phenomenon has been observed by the authors of the publication which mentioned s15 (Schuenemann et al. 2013). It is believed to be a result of to the high number of variants present in this strain compared to the others. Moreover, the 3L branch only has one representative while the others possess several representatives. Therefore, the uncertainty in the location of this strain might be a consequence of its under-representation.

4.5.2 Placement of the low-coverage genomes

The approach which combined the branch assignment of the low-coverage genomes to a previously-built tree and partial tree reconstruction for each branch allowed adding five strains to the set of newly-described strains. The low-coverage strains typed fell into Branch 3I and Branch 2F. Moreover, the proportion of 3I and 2F found in the low-coverage genomes seems to support the previously-observed dominance of 3I. Because of the principle of the branch assignment, no new strain can be discovered via this method. A straightforward follow-up would be to apply the same approach using all previously-published strains as anchors, thus enabling the assignment of low-coverage samples to branches other than 3I or 2F. However, the coverage of the previously-published strains has been observed to be greatly variable, with several strains being themselves at very low coverage. In consequence, this follow-up was not implemented, as it would make little sense to use low-coverage genomes as anchors. It is likely that future research will provide more high-coverage genomes to increase the resolution of *M. leprae's* phylogeny, allowing the use of more sequences as anchors for the kind of phylogenetic placement strategies which are currently being developed for *M. tuberculosis* (Kay et al. 2015).

4.6 *M. leprae* in Northern Europe

4.6.1 *M. leprae* in the medieval St. Jørgen leprosarium

Table 29: SNP types of the recovered genomes and geographical context

Sample	SNP type	Radiocarbon age (cal AD)	Modern geographical association (Monot et al. 2009)	Ancient geographical association (Mendum et al. 2014)
SJG 189	3I		America & Europe	Sweden, Denmark, UK
SJG 611	3I		America & Europe	Sweden, Denmark, UK
SJG 658	2F		Middle East	Sweden, Denmark, UK
SJG 507	3K	1170-1214	Middle East & Asia	Egypt, Hungary, Turkey
SJG 533	3I		America & Europe	Sweden, Denmark, UK
SJG 427	3I	1177-1250	America & Europe	Sweden, Denmark, UK
SJG 404	3I	1226-1268	America & Europe	Sweden, Denmark, UK
SJG 749	2F	1230-1277	Middle East	Sweden, Denmark, UK
SJG 472	3I	1252-1280	America & Europe	Sweden, Denmark, UK
SJG 722	3I	1269-1285	America & Europe	Sweden, Denmark, UK
SJG 1137	3I	1282-1380	America & Europe	Sweden, Denmark, UK
SJG 978	2F	1283-1381	Middle East	Sweden, Denmark, UK

The radiocarbon dating results suggest the possible co-existence of several *M. leprae* strains within the leprosarium (Table 29). Strains 3K and 3I in particular seem to have co-existed from the 12th-14th centuries. Strain 2F seems to have also been present during and after the 13th century. In the absence of registers indicating the admission and death dates of leprosy sufferers, it is difficult to assess if the studied individuals were in contact with each other during their stay in the leprosarium. It is interesting to notice, however, that none of the individuals showed evidence of infection by several strains. Indeed, after arrival in the leprosarium, the leprosy sufferer would have been exposed to various strains of the pathogen for a prolonged period of time. During this exposure time, little is known regarding the spread of the strains. It has been suggested that the most contagious strain would spread to all patients, at the expense of less contagious ones. In a case in which several strains would present similar contagion capacity and virulence, a few co-infection cases would still be expected. The results presented above do not seem to support this hypothesis. This might be explained by the long incubation time of leprosy. Indeed, stays in leprosariums have been reported to have lasted for several years before the death of the patients. As leprosy can take up to 20 years to develop, a leprosy patient would have been likely to die before the spread of a secondary infection. Nevertheless, the presence of various strains in the hospital in such a close time frame suggests two possible alternate explanations: the large geographical influence of the St. Jørgen leprosarium, and/or very rapid changes in the distribution of *M. leprae* strains between the 12th and the 14th centuries.

The observation regarding the influence of St. Jørgen originates from the hypothesis that if all *M. leprae* strains were present homogeneously in southern Denmark without one strain overpowering the others, then infections with multiple strains should be found. The results of this study instead suggest that *M. leprae* strains might have had more localised distribution areas in which a single village may have been mainly affected by a specific strain. The major strain determination might have been related to the various living standards, levels of animal contact and/or integration into local village networks. The region of origin of the leprosy sufferers is difficult to determine using only the results of ancient DNA. In the case of the St. Jørgen cemetery, no leprosarium register was available as a historical source. Moreover, no isotopic studies have been performed to date. Mitochondrial DNA typing of the human genomes performed by colleagues showed no direct maternal relationship between the individuals, supporting the suggestion made using the *M. leprae* data. Therefore, the leprosy sufferers likely came not only from the nearby city of Odense, but also from remote villages.

The second hypothesis comes from the data suggesting that strain 2F was present only in the second part of the leprosarium's active period as well as the impossibility of evaluating the contemporaneousness of the individuals studied here. If none of the recovered strains were present simultaneously, strains 3K, 3I and then, finally, 2F would have been present in the region between the 12th and 14th centuries. The succession of three strains in a 300 year time frame with the accepted *M. leprae* incubation period of 2-20 years would suggest that *M. leprae* strains were completely replaced in the region in less than 80 years, which equates to about two to three human generations. So far, there is not enough data available on the dynamics of *M. leprae* strains in medieval times to evaluate the plausibility of this calculation. It is worth noting, however, that the 80- year replacement time is not biologically impossible, especially given the efforts which were made to isolate leprosy sufferers within dedicated hospices.

4.6.2 Medieval *M. leprae* genetic diversity in Northern Europe

All the 16 ancient *M. leprae* genomes recovered varied from each other by at least one SNP (Supplementary Table 14) and clustered in various branches in the pathogen phylogeny. The St. Jørgen cemetery presents, therefore, the most genetically diverse collection of medieval *M. leprae* genomes from one locale studied to date. It also brings to light one genome typed as strain 3K which had never previously been described in medieval Northern Europe. This suggests that the genetic diversity of *M. leprae* in southern 12th -14th century Denmark has likely been underestimated in the literature. In addition, not all individuals infected with the disease would have been sent to the leprosarium. Commonly, only persons who showed striking symptoms (such as disfiguration or the loss of fingers) were ostracized. Persons with mild symptoms were able to avoid marginalization by hiding those symptoms. Moreover, individuals who developed the more extreme forms of the disease might have died before being sent to specialized establishments. Therefore, the observed genetic diversity of *M. leprae* in the St. Jørgen individuals is itself likely an under-representation the overall genetic diversity of the pathogen in Northern Europe. While it is still impossible at present to accurately quantify the genetic diversity of medieval European *M. leprae*, it is likely to have been similar to the one observed today in countries where the disease is still endemic.

4.6.4 Spread of *M. leprae* strains into and out of Europe

Strains 3I and 2F have been previously documented in medieval Northern Europe (Mendum et al. 2014), although strain 3K has not yet been documented within that time frame. A few previous identifications of this strain in ancient remains exist, albeit in older samples (4^{th}-5^{th} century Egypt, 7^{th} century Hungary and 8^{th}-9^{th} century Turkey (Monot et al. 2009)). Its presence in 12^{th} century Denmark highlights the wide distribution of the 3K strain in medieval times. Moreover, the ancient 3K clusters with modern-day 3K strains from China and New Caledonia. The branch containing 3K strains has been suggested as a separate branch (Branch 0) corresponding to a fifth *M. leprae* type (Schuenemann et al. 2013). These results seem to corroborate this hypothesis and support the very early branching point previously observed. Indeed, in this data, Branch 0 appears to be separate from the ancestral type 2 before its diversification into subtypes E, F, G and H.

According to the evolutionary history of the pathogen as it is currently understood (see part 1.2), *M. leprae* type 3 likely derived from a type 2-related ancestor in Europe (Monot et al. 2009). Afterwards, type 3 strains probably evolved into 4 subtypes (I, J, K, and L) during the time in which they were spreading into Europe. Each of the four subtypes has been shown to be associated with human migrations out of Europe, with strain 3K being associated with eastward human migrations to Asia (Monot et al. 2009; Monot et al. 2005). The current geographical distribution of type 3K is consistent with this association, with the majority of 3K isolates originating from China and southwest Asia. In some respects, this view contradicts the phylogeny results from Schuenemann and colleagues as well as those from this study. Indeed, Branch 0 does not separate from the ancestral Branch 2, as would be expected when following the idea described above. On the contrary, Branch 2 seems to be as ancient as the other branches. The modern geographical repartition of *M. leprae* types and subtypes shows that the spread of leprosy followed human dispersals (Monot et al. 2009). Therefore, if strain 3K did not separate from an ancestral type 2 in Europe, it was likely brought into Europe along with a type 2 ancestor during successive human migration events. According to this data, the other type 3 strains seem to have evolved as has been suggested by Monot and colleagues.

After spreading into Europe, strain 3K is believed to have migrated out of Europe towards Asia. This has been supported by the discovery of *M. leprae* 3K isolates from 7^{th}-10^{th} century Hungary and Turkey. The presence at a later time of 3K in Denmark suggests that the out-of-Europe migration of this strain was likely very progressive, following several centuries of human warfare and trade in this direction. This might also be the case for other type 3 strains (including 3L, which has not yet been found in ancient European remains).

The 3K and 3I types can still found to this day in the rare cases of European leprosy, in contrast to the 2F type. This observation is interesting because in the *M. leprae* genomes recovered, the 2F strain seemed to appear later than the type 3 strains. Not enough genomes were recovered to reliably establish the timespan of 2F's presence. If confirmed, this trend would suggest that the 2F strain might have played a major role in the disappearance of leprosy from Europe. However, the absence of type 2F in the modern day could be due to the 20th century's tremendous improvements in medical care and life quality.

4.6.5 Outlook

The results presented in this dissertation characterize medieval *M. leprae* as a genetically highly diverse bacteria with complex distribution patterns in southern Denmark. It is currently impossible to generalise the observations drawn from the St. Jørgen data, since it represents a very specific case (leprosarium hospital) and as the two other cemeteries did not yield positive results. Nevertheless, the high genetic diversity observed in southern Denmark should be used as an indication that isolated studies of a couple of cases from various periods and locations are likely to largely underestimate the genetic diversity of the pathogen in a specific area and time. It would be interesting to study other medieval leprosaria from Europe via the same method in order to evaluate the hypotheses drawn from this study and allow for some degree of generalisation. In addition, while this study only focussed on the causative agent of leprosy, the same approach might be applied to other ancient bacterial pathogens, such as *M. tuberculosis*. Indeed, other pathogens studied through ancient DNA methods are equally as likely to exhibit be underestimated in terms of their genetic diversity.

The data presented in this study represent a first step towards understanding the etiology of leprosy in the late Middle-Ages. Combined with future data on the subject, this study might provide key information regarding the disappearance of the disease from medieval Europe, especially by allowing a direct comparison between ancient and modern European strains of the same type. Moreover, modern studies on leprosy would likely benefit from efforts made in this direction. Indeed, leprosy is still highly prevalent in some developing countries; about 200,000 cases are still reported each year. Unravelling the reasons behind leprosy's disappearance from Europe would certainly improve the chances of the eradication attempts made in countries where the disease is still endemic.

5 Summary

Leprosy is an infectious disease mainly caused by the obligate intracellular pathogen *Mycobacterium leprae (M. leprae)*. In Northern Europe, the disease reached its highest prevalence in the Middle-Ages, between the 11th and 15th centuries. It disappeared around the 16th century before the introduction of modern medicine. Besides archaeological and medical studies, the direct recovery and analysis of medieval *M. leprae* genomes can greatly improve our understanding of the history and epidemiology of the disease. This thesis aims to analyse numerous medieval *M. leprae* genomes from Northern Europe at high genomic coverage using next-generation sequencing. Teeth from 140 human skeletal remains were collected from three cemeteries in Denmark and Germany (1000-1560 AD) in order to 1) screen the remains for *M. leprae* DNA with PCR, 2) use high-throughput sequencing to recover high-quality *M. leprae* genomes, 3) place the recovered genomes within the *M. leprae* phylogeny, and 4) investigate *in-silico* the possible effects of the genomic variations. The St. Jørgen leprosarium in Denmark yielded the highest number of genetically confirmed leprosy cases ever reported from one cemetery (16 out of 34 specimens collected). None of the leprosy-positive DNA samples showed evidence of co-infection with tuberculosis. Of the 16 ancient *M. leprae* genomes obtained, 6 had complete high-coverage, while for the remaining 10 genomes, partial drafts were recovered. Strain SNP typing of the genomes showed the presence of strains 3I and 2F, already found in previous European medieval leprosy studies, and strain 3K, so far never described in medieval Northern Europe. The presence of 3K in late medieval Denmark changes the current hypothesis about the spread of this strain out of Europe towards China. The high quality of the genomic data allowed the construction of a new phylogenetic tree for *M. leprae,* which for the first time also included low-coverage genomes. Although the samples were obtained from one locale (St. Jørgen), the genomes all differed from each other and clustered in various branches. This finding indicates that medieval *M. leprae* genetic diversity has so far been greatly underestimated and suggests that it is likely to have been similar to the one observed today. The ancient type 3K grouped with the modern 3K strains from China and New Caledonia, confirming the existence of the recently described Branch 0. According to modern data, none of the 696 detected genetic variants seemed to be associated with a change in the virulence of the pathogen. However, the *in-silico* methods used to evaluate the variant effects are hampered by our current limited knowledge of protein expression and activity in *M. leprae*. Taken together, the results presented in this thesis draw a more detailed picture of the genetic diversity of medieval *M. leprae* and might contribute to the debate regarding both the disappearance of the disease from medieval Northern Europe and the current re-emergence of the disease in some developing countries.

6 Zusammenfassung

Lepra ist eine chronische Infektionskrankheit, welche durch das obligat intrazelluläre Bakterium *Mycobacterium leprae (M. leprae)* ausgelöst wird. In Nordeuropa erreichte die Prävalenz der Krankheit im Mittelalter zwischen dem 11. und 15. Jahrhundert ihren Höhepunkt. Im 16. Jahrhundert verschwand die Krankheit weit vor der Einführung moderner Medizin. Neben archäologischen sowie medizinischen Studien kann die Isolierung und Analyse des mittelalterlichen *M. leprae* Genoms zum Verständnis der Geschichte sowie der Epidemiologie der Krankheit beitragen. Die vorliegende Arbeit hat die Analyse zahlreicher, gut abgedeckter mittelalterlicher *M. leprae* Genome aus Nordeuropa unter Verwendung von Hochdurchsatzsequenzierung zum Ziel. Im Laufe des Projektes wurden Zähne von insgesamt 140 Individuen von drei unterschiedlichen mittelalterlichen Friedhöfen in Dänemark sowie Deutschland gesammelt (1000 – 1560 AD). Das Projekt umfasste folgende Arbeitsschritte: 1) Test auf *M. leprae* positive Proben mittels PCR, 2) Rekonstruktion möglichst gut abgedeckter *M. leprae* Genome unter Verwendung der Hochdurchsatzsequenzierung, 3) Einbindung der alten *M. leprae* Genome in einen phylogenetischen Baum, und 4) In silico-Modellierung der möglichen Effekte genomischer Variationen. Das St. Jørgen Leprosarium in Dänemark wies die höchste Anzahl genetisch bestätigter Lepra-Fälle auf, die bisher jemals mittels molekularbiologischer Methoden innerhalb eines Friedhofs nachgewiesen wurden (16 von 32 gesammelten Proben). Eine Ko-Infektion mit Tuberkulose war in keiner der Lepra-positiven Proben genetisch zu bestätigen. Insgesamt konnten aus den 16 *M. leprae* positiven Proben sechs komplette Genome mit einer hohen Abdeckung sowie 10 Genome partiell rekonstruiert werden. Die Erregerstämme 3I und 2F, welche schon in anderen Publikationen für das mittelalterliche Nordeuropa beschrieben worden sind, wurden auch in den vorliegenden Proben mittels SNP-Typisierung identifiziert. Des Weiteren konnte ein bisher in Nordeuropa unbekannter Stamm (3K) beobachtet werden. Die Anwesenheit dieses Stammes im spätmittelalterlichen Dänemark verändert die bisher aufgestellte Hypothese über die Verbreitung des Stammes von Europa nach China. Die hohe Qualität der generierten Daten erlaubte die Erstellung eines phylogenetischen Baums für *M. leprae*, welcher erstmals nun auch Genome mit einer geringen Abdeckung beinhaltete. Obwohl die Proben aus einem engen geografischen Herkunftsraum stammten, unterschieden sich die Genome voneinander und konnten unterschiedlichen monophyletischen Gruppen zugeordnet werden. Der Befund weist darauf hin, dass die genetische Diversität von *M. leprae* im Mittelalter bisher stark unterschätzt worden ist und sie der heutigen Diversität moderner Stämme ähnelt. Der gefundene mittelalterliche Erregerstamm 3K fiel in dieselbe Gruppe wie die modernen 3K Stämmen aus China und Neukaledonien, was die Existenz der monophyletischen Gruppe 0 bestätigt. Eine im

Vergleich zu modernen Daten gesteigerte Virulenz des Pathogens konnte nicht nachgewiesen werden, da keine der 696 identifizierten Varianten auf eine solche Veränderung hinwiest. Nichtsdestotrotz sollte bemerkt werden, dass die verwendeten *in-silico* Methoden durch das fehlende Wissen über Proteinexpression sowie -aktivität in *M. leprae* limitiert sind. Zusammengefasst erweitert diese Arbeit unser Verständnis der genetischen Diversität der *M. leprae* Stämme im Mittelalter. Überdies tragen die gewonnenen Ergebnisse zu Debatten um das Verschwinden der Krankheit im 16. Jahrhundert aus Nordeuropa sowie um das Wiederauftreten der Krankheit in einigen Entwicklungsländern bei.

7 Résumé

La lèpre est une maladie infectieuse dont l'agent principal est *Mycobacterium leprae* (*M. leprae*), une bactérie parasite intracellulaire. En Europe, la prévalence de atteint son paroxysme à la fin du Moyen-Age entre le 11ème et le 15ème siècle. Elle disparait ensuite entre le 15ème et le 16ème siècle, avant l'introduction de la médecine moderne. En complément des approches archéologiques et médicales, l'étude directe de génomes d'origine médiévale de *M. leprae* peut grandement améliorer notre compréhension de l'histoire et de l'épidémiologie de la maladie. Cette thèse a pour but d'analyser de nombreux génomes de *M. leprae* médiévaux avec une grande profondeur de séquençage en utilisant les nouvelles techniques de séquençage haut débit. 140 dents humaines provenant de trois cimetières d'Allemagne et du Danemark (1000-1560 ap. JC) ont été collectées et analysées afin de : 1) sélectionner via PCR les échantillons où l'ADN de *M. leprae* est préservé, 2) utiliser le séquençage haut-débit pour récupérer des génomes de *M. leprae* de haute qualité, 3) placer les génomes obtenus dans la phylogénie de *M. leprae,* et 4) étudier *in-silico* les effets possibles des variations génétiques. L'hôpital médiéval danois St. Jørgen, réservé aux lépreux, présente le plus grand nombre de cas de lèpre confirmés génétiquement pour un même endroit (16 génomes obtenus pour 34 individus testés). Aucun des échantillons positifs pour la lèpre ne présente de preuves de co-infection avec la tuberculose. Parmi les 16 génomes obtenus, 6 ont pu être séquencés complètement et en profondeur. Pour les 10 génomes restants, seuls des assemblages incomplets ont pu être obtenus. L'analyse des variations génetiques montre la présence des souches 3I et 2F, déjà documentées dans d'autres populations médiévales en Europe, ainsi que de la souche 3K, dont c'est le premier cas identifié dans le nord de l'Europe avant le 16ème siècle. L'existence de la souche 3K au Danemark à la fin du Moyen-Age affine l'hypothèse existante concernant la migration de cette souche hors d'Europe vers la Chine. La haute qualité des donnés a permis de replacer tous les génomes dans la phylogénie de *M. leprae*, y compris les génomes incomplets. Malgré leur origine géographique commune (St. Jørgen), les génomes obtenus sont tous différents et se placent dans différentes branches. Ce résultat indique que la diversité génétique de *M. leprae* au Moyen-Age est certainement grandement sous-estimée. L'ancien type 3K se place avec les souches 3K modernes de Chine et de Nouvelle-Calédonie dans la branche 0 décrite récemment. D'après des données modernes, aucune des 696 variations génétiques observées ne semble liée à un changement de virulence du pathogène. Cependant, le peu d'informations disponibles au sujet de l'expression et de l'activité des protéines chez *M. leprae* limite les méthodes de prédiction. Dans l'ensemble, les résultats présentés dans ce document permettent de mieux appréhender la diversité génétique de *M.*

leprae au Moyen Age, et pourront dans le futur contribuer à comprendre la disparition spontanée de la lèpre en Europe et sa réémergence actuelle dans certains pays en développement.

8 References

Alford, P.L., Lee, D.R., Binhazim, A.A., Hubbard, G.B. & Matherne, C.M., 1996. Naturally acquired leprosy in two wild-born chimpanzees. *Lab. Anim. Sci*, 46(3), pp.341–346.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J., 1990. Basic local alignment search tool. *J. Mol. Biol.*, 215(3), pp.403–410.

Anastasiou, E. & Mitchell, P.D., 2013. Palaeopathology and genes: Investigating the genetics of infectious diseases in excavated human skeletal remains and mummies from past populations. *Gene*, 528(1), pp.33–40.

Andersen, J.G. & Manchester, K., 1992. The rhinomaxillary syndrome in leprosy : a clinical , radiological and palaeopathological study. *Int. J. Osteoarchaeol.*, 2, pp.121–129.

Andersen, J.G., Manchester, K. & Roberts, C., 1994. Septic bone changes in leprosy : a clinical , radiological and palaeopathological review. *Int. J. Osteoarchaeol.*, 4, pp.21–30.

Anderson, R. & May, R., 1982. Coevolution of host and parasites. *Parasitology*, 85, pp.411–426.

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., De Castro, E., Duvaud, S., Flegel, V., Fortier, A., Gasteiger, E., Grosdidier, A., Hernandez, C., Ioannidis, V., Kuznetsov, D., Liechti, R., Moretti, S., Mostaguir, K., Redaschi, N., Rossier, G., Xenarios, I. & Stockinger, H., 2012. ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res.*, 40(W1), pp.597–603.

Aufderheide, C. & Rodriguez-Martin, C., 1998. *The Cambridge encyclopedia of human paleopathology*, Cambridge: Cambridge University Press.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S. & DePristo, M.A., 2013. From FastQ data to high-confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics*, pp.11.10.1–11.10.33.

Barnes, I. & Thomas, M.G., 2006. Evaluating bacterial pathogen DNA preservation in museum osteological collections. *P. Roy. Soc. B-Biol. Sci.*, 273(1587), pp.645–653.

Bateman, A., Martin, M.J., O'Donovan, C., Magrane, M., Apweiler, R., Zhang, J. et al., 2015. UniProt: A hub for protein information. *Nucleic Acids Res.*, 43(D1), pp.D204–D212.

Bates, J.H. & Stead, W.W., 1993. The history of tuberculosis as a global epidemic. *Med. Clin. N. Am.*, 77(6), pp.1205–1217.

Belda, E., Moya, A., Bentley, S. & Silva, F.J., 2010. Mobile genetic element proliferation and gene inactivation impact over the genome structure and metabolic capabilities of Sodalis glossinidius, the secondary endosymbiont of tsetse flies. *BMC Genomics*, 11, p.449.

Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C. & Apweiler, R., 2009. QuickGO: A web-based tool for Gene Ontology searching. *Bioinformatics*, 25(22), pp.3045–3046.

Blondiaux, J., 2002. Microscopic study and X-ray analysis of two 5th century cases of leprosy: paleoepidemiological inferences. In C. A. Roberts, M. Lewis, & K. Manchester, eds. *The past and present of leprosy. Archeological, historical, paleopathological and clinical approaches*. Oxford: Archaeopress, pp. 105–110.

Bloom, B.R. & Godal, T., 1983. Selective primary health care: strategies for control of disease in the developing world. V. Leprosy. *Rev. Infect. Dis.*, 5(4), pp.765–780.

Bloom, B.R. & Mehra, V., 1984. Immunological unresponsiveness in leprosy. *Immunol. Rev.*, 80, pp.5–28.

Boldsen, J.L., 2001. Epidemiological approach to the paleopathological diagnosis of leprosy. *Am. J. Phys. Anthropol.*, 115(4), pp.380–387.

Boldsen, J.L., 2005a. Leprosy and mortality in the medieval Danish village of Tirup. *Am. J. Phys. Anthropol.*, 126(2), pp.159–168.

Boldsen, J.L., 2007. *Leprosy in medieval Denmark – A comprehensive analysis*, Odense: University of Southern Denmark.

Boldsen, J.L., 2009a. Leprosy in medieval Denmark - Osteological and epidemiological analyses. *Anthropol. Anz.*, 67(4), pp.407–425.

Boldsen, J.L., 2005b. Testing conditional independence in diagnostic palaeoepidemiology. *Am. J. Phys. Anthropol.*, 128(3), pp.586–592.

Boldsen, J.L. & Mollerup, L., 2006. Outside St. Jørgen: Leprosy in the medieval Danish city of Odense. *Am. J. Phys. Anthropol.*, 130(3), pp.344–351.

Boldsen, J.L., Rasmussen, K.L., Riis, T. & Weise, S., 2013. Schleswig : Medieval leprosy on the boundary between Germany and Denmark. *Anthropol. Anz.*, 3, pp.273–287.

Bouwman, A.S., Kennedy, S.L., Müller, R., Stephens, R.H., Holst, M., Caffell, A.C., Roberts, C.A. & Brown, T.A., 2012. Genotype of a historic strain of Mycobacterium tuberculosis. *Proc. Natl. Acad. Sci. U.S.A.*, 109(45), pp.18511–18516.

Briggs, A.W., Stenzel, U., Johnson, P.L.F., Green, R.E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M.T., Lachmann, M. & Pääbo, S., 2007. Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci. U.S.A.*, 104(37), pp.14616–14621.

Britton, W.J. & Lockwood, D.N., 2004. Leprosy. *Lancet*, 363(9416), pp.1209–1219.

Brotherton, P., Endicott, P., Sanchez, J.J., 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(17), pp.5717–5728.

Bryceson, A. & Pfaltzgraff, R.E., 1990. *Leprosy (Medicine in the tropics)* 3rd ed., Churchill Livingstone.

Chaussinand, R., 1953. Tuberculosis and leprosy; mutually antagonistic diseases. *Leprosy Rev.*, 24(2), pp.90–94.

Chou, P. & Fasman, G., 1974a. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry*, 13(2), pp.211–222.

Chou, P. & Fasman, G., 1974b. Prediction of protein conformation. *Biochemistry*, 13(2), pp.222–245.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. & Ruden, D.M., 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, 6(2), pp.80–92.

Clark, G.A., Kelley, M.A., Grange, J.M. & Hill, M.C., 1987. The evolution of mycobacterial disease in human populations: a reevaluation. *Curr. Anthropol.*, 28(1), pp.45–62.

Clark, K.A., Kim, S.H., Boening, L.F., Taylor, M.J., Betz, T.G. & McCasland, F. V, 1987. Leprosy in armadillos (Dasypus novemcinctus) from Texas. *J. Wildlife. Dis.*, 23(0090-3558), pp.220–224.

Clark-Curtiss, J.E., Jacobs, W.R., Docherty, M.A., Ritchie, L.R. & Curtiss, R., 1985. Molecular analysis of DNA and construction of genomic libraries of Mycobacterium leprae. *J. Bacteriol.*, 161(3), pp.1093–102.

Cole, S.T., Eiglmeier, K., Parkhill, J., James, K.D., Thomson, N.R., Wheeler, P.R., Honoré, N., Garnier, T., Churcher, C., Harris, D., Mungall, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R.M., Devlin, K., Duthoy, S., Feltwell, T., Fraser, A., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Lacroix, C., Maclean, J., Moule, S., Murphy, L., Oliver, K., Quail, M.A., Rajandream, M.A., Rutherford, K.M., Rutter, S., Seeger, K., Simon, S., Simmonds, M., Skelton, J., Squares, R., Squares, S., Stevens, K., Taylor, K., Whitehead, S., Woodward, J.R. & Barrell, B.G., 2001. Massive gene decay in the leprosy bacillus. *Nature*, 409(6823), pp.1007–11.

Combet, C., Blanchet, C., Geourjon, C. & Deléage, G., 2000. NPS@: Network protein sequence analysis. *Trends Biochem. Sci.*, 25(3), pp.147–150.

Cooper, A. & Poinar, H.N., 2000. Ancient DNA: Do it right or not at all. *Science*, 289(5482), p.1139.

Cooper, A. & Waynet, R., 1998. New uses for old DNA. *Curr. Opin. Biotech.*, 9(1), pp.49–53.

Cowdry, E. V, 1978. Cytological studies on globi in leprosy b Edmund V. Cowdry, reprinted from American Journal of Pathology, Vol. 16, No. 2, March 1940. *Int. J. Leprosy*, 46(2), pp.175–201.

Daffe, M. & Draper, P., 1998. The envelope layers of Mycobacteriawith reference to their pathogenicity. *Adv. Microb, Physiol.*, 39, pp.131–203.

Darling, A.E., Mau, B. & Perna, N.T., 2010. progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE*, 5(6), pp.1–17.

Degang, Y., Nakamura, K., Akama, T., Ishido, Y., Luo, Y. & Ishii, N., 2014. Leprosy as a Model of Immunity. *Future Microbiol.*, 9(1), pp.43–54.

Delmotte, F., Rispe, C., Schaber, J., Silva, F.J. & Moya, A., 2006. Tempo and mode of early gene loss in endosymbiotic bacteria from insects. *BMC Evol. Biol.*, 6(1), p.56.

Desikan, K. V & Job, C.K., 1968. A review of postmortem findings in 37 cases of leprosy. *Int. J. Leprosy*, 36(1), pp.32–44.

Dols, M.W., 1979. Leprosy in Medieval Arabic Medicine. *J. Hist. Med. All. Sci.*, XXXIV(3), pp.314–333.

Donoghue, H.D., 2009. Human tuberculosis - an ancient disease, as elucidated by ancient microbial biomolecules. *Microbes Infect.*, 11(14-15), pp.1156–1162.

Donoghue, H.D., 2011. Insights gained from palaeomicrobiology into ancient and modern tuberculosis. *Clin. Microbiol. Infect*, 17(6), pp.821–829.

Donoghue, H.D., 2013. Insights into ancient leprosy and tuberculosis using metagenomics. *Trends Microbiol.*, 21(9), pp.448–450.

Donoghue, H.D., Holton, J. & Spigelman, M., 2001. PCR primers that can detect low levels of Mycobacterium leprae DNA. *J. Med. Microbiol.*, 50(2), pp.177–182.

Donoghue, H.D., Marcsik, A., Matheson, C., Vernon, K., Nuorala, E., Molto, J.E., Greenblatt, C.L. & Spigelman, M., 2005. Co-infection of Mycobacterium tuberculosis and Mycobacterium leprae in human archaeological samples: a possible explanation for the historical decline of leprosy. *P. Roy. Soc. B-Biol. Sci.*, 272(1561), pp.389–394.

Donoghue, H.D., Spigelman, M., Greenblatt, C.L., Lev-Maor, G., Kahila Bar-Gal, G., Matheson, C., Vernon, K., Nerlich, A.G. & Zink, A.R., 2004. Tuberculosis: From prehistory to Robert Koch, as revealed by ancient DNA. *Lancet Infect. Dis.*, 4(9), pp.584–592.

Donoghue, H.D., Spigelman, M., O'Grady, J., Szikossy, I., Pap, I., Lee, O.Y.-C., Wu, H.H.T., Besra, G.S. & Minnikin, D.E., 2015. Ancient DNA analysis – An established technique in charting the evolution of tuberculosis and leprosy. *Tuberculosis*, 95, pp.S140–S144.

Draper, P., 1983. The bacteriology of Mycobacterium leprae. *Tubercle*, 64(1), pp.43–56.

Draper, P., Kandler, O. & Darbre, A., 1987. Peptidoglycan and Arabinogalactan of Mycobacterium Leprae. *Microbiology*, 133(5), pp.1187–1194.

Drozdetskiy, A., Cole, C., Procter, J. & Barton, G., 2015. JPred4 : a protein secondary structure prediction server. *Nucleic Acids Res.*, 43(1), pp.389–394.

Drummond, A., Ashton, B. & Cheung, M., 2010. Geneious v 6.1. 5. , pp.1–206.

Dubos, R.J., 1980. *Man Adapting*, Yale University Press.

Faerman, M. & Jankauskas, R., 2000. Paleopathological and molecular evidence of human bone tuberculosis in Iron Age Lithuania. *Anthropol. Anz.*, 58(1), pp.57–62.

Felsenstein, J., 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39(4), pp.783–791.

Fine, P., 1984. Leprosy and tuberculosis - an epidemiological comparison. *Tubercle*, 65(2), pp.137–153.

Fine, P.E., 1981. Immunogenetics of susceptibility to leprosy, tuberculosis, and leishmaniasis. An epidemiological perspective. *Int. J. Leprosy*, 49(4), pp.437–454.

Fine, P.E., 1982. Leprosy: the epidemiology of a slow bacterium. *Epidemiol. Rev.*, 4, pp.161–88.

Fletcher, H.A., Donoghue, H.D., Taylor, M.G., van der Zanden, A.G.M. & Spigelman, M., 2003. Molecular analysis of Mycobacterium tuberculosis DNA from a family of 18th century Hungarians. *Microbiology*, 149(1), pp.143–151.

Gagneux, S., 2012. Host-pathogen coevolution in human tuberculosis. *P. Roy. Soc. B-Biol. Sci.*, 367(1590), pp.850–9.

Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., B.A., 2005. Protein identification and analysis tools on the ExPASy server. In J. M. Walker, ed. *The Proteomics*

Protocols Handbook. Humana Press, pp. 571–607.
Gil, R., Belda, E., Gosalbes, M.J., Delaye, L., Vallier, A., Vincent-Monegat, C., Heddi, A., Silva, F.J., Moya, A. & Latorre, A., 2008. Massive presence of insertion sequences in the genome of SOPE, the primary endosymbiont of the rice weevil Sitophilus oryzae. *Int. Microbiol.*, 11(1), pp.41–48.
Ginolhac, A., Rasmussen, M., Gilbert, M.T.P., Willerslev, E. & Orlando, L., 2011. mapDamage: Testing for damage patterns in ancient DNA sequences. *Bioinformatics*, 27(15), pp.2153–2155.
Gomez-Valero, L., Rocha, E.P.C., Latorre, A. & Silva, F.J., 2007. Reconstructing the ancestor of Mycobacterium leprae: The dynamics of gene loss and genome reduction. *Genome Res.*, 17(8), pp.1178–1185.
Grantham, R., 1974. Amino acid difference formula to help explain protein evolution. *Science*, 185(4154), pp.862–864.
Green, R.E., Briggs, A.W., Krause, J., Prüfer, K., Burbano, H. A, Siebauer, M., Lachmann, M. & Pääbo, S., 2009. The Neandertal genome and ancient DNA authenticity. *EMBO J.*, 28(17), pp.2494–2502.
Gross, W.M. & Wayne, L.G., 1970. Nucleic acid homology in ghe genus Mycobacterium. *J. Bacteriol.*, 104(2), pp.630–634.
Grupe, G., 1997. Die anthropologische Bearbeitung der Skelettserie von Schleswig, Ausgrabung Rathausmarkt : Rekonstruktion einer mittelalterlichen Bevölkerung und ihrer Umweltbeziehungen. In *Kirche und Gräberfeld des 11 : -13. Jahrhunderts unter dem Rathausmarkt von Schleswig.* Neumünster: Wachholtz, pp. 147–209.
Haas, C.J., Zink, A., Pálfi, G., Szeimies, U. & Nerlich, A.G., 2000. Detection of leprosy in ancient human skeletal remains by molecular identification of Mycobacterium leprae. *Am. J. Clin. Pathol.*, 114(3), pp.428–36.
Hagelberg, E. & Clegg, J.B., 1991. Isolation and characterization of DNA from archaeological bone. *P. Roy. Soc. B-Biol. Sci.*, 244(1309), pp.45–50.
Han, X.Y., Mistry, N.A., Thompson, E.J., Tang, H. & Khanna, K., 2015. Draft genome sequence of new leprosy agent Mycobacterium lepromatosis. *Genome Announcement*, 3(3), pp.9–10.
Han, X.Y., Seo, Y.H., Sizer, K.C., Schoberle, T., May, G.S., Spencer, J.S., Li, W. & Nair, R.G., 2008. A new Mycobacterium species causing diffuse lepromatous leprosy. *Am. J. Clin. Pathol.*, 130(6), pp.856–864.
Han, X.Y. & Silva, F.J., 2014. On the Age of Leprosy. *PLoS Neglect. Trop. Dis.*, 8(2), p.e2544.
Han, X.Y., Sizer, K.C., Thompson, E.J., Kabanja, J., Li, J., Hu, P., Gomez-Valero, L. & Silva, F.J., 2009. Comparative sequence analysis of Mycobacterium leprae and the new leprosy-causing Mycobacterium lepromatosis. *J. Bacteriol.*, 191(19), pp.6067–6074.
Harkins, K.M., Buikstra, J.E., Campbell, T., Bos, K.I., Johnson, E.D., Krause, J., Stone, A.C., Campbell, T., Ki, B., Ed, J., Krause, J. & Stone, A.C., 2015. Screening ancient tuberculosis with qPCR : challenges and opportunities. *P. Roy. Soc. B-Biol. Sci.*, 370, pp.1–10.
Heesterbeek, H., Anderson, R.M., Andreasen, V., Bansal, S., De Angelis, D., Dye, C., Eames, K.T.D., Edmunds, W.J., Frost, S.D.W., Funk, S., Hollingsworth, T.D., House, T., Isham, V., Klepac, P., Lessler, J., Lloyd-Smith, J.O., Metcalf, C.J.E., Mollison, D., Pellis, L., Pulliam, J.R.C., Roberts, M.G. & Viboud, C., 2015. Modeling infectious disease dynamics in the complex landscape of global health. *Science*, 347(6227), pp.aaa4339–aaa4339.
Hershkovitz, I., Donoghue, H.D., Minnikin, D.E., Besra, G.S., Lee, O.Y.-C., Gernaey, A.M., Galili, E., Eshed, V., Greenblatt, C.L., Lemma, E., Bar-Gal, G.K. & Spigelman, M., 2008. Detection and Molecular Characterization of 9000-Year-Old Mycobacterium tuberculosis from a Neolithic Settlement in the Eastern Mediterranean N. Ahmed, ed. *PLoS ONE*, 3(10), p.e3426.
Higuchi, R., Bowman, B., Freiberger, M., Ryder, O.A. & Wilson, A.C., 1984. DNA sequences from the quagga, an extinct member of the horse family. *Nat.*, 312(5991), pp.282–284.
Hirokawa, T., Boon-Chieng, S. & Mitaku, S., 1998. SOSUI: classification and secondary structure

prediction system for membrane proteins. *Bioinformatics*, 14(4), pp.378–379.
Hofreiter, M., 2008. Palaeogenomics. *C. R. Palevol*, 7(2-3), pp.113–124.
Hofreiter, M., Jaenicke, V., Serre, D., Haeseler Av, A. & Pääbo, S., 2001. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res.*, 29(23), pp.4793–4799.
Hofreiter, M., Serre, D., Poinar, H.D., Kuch, M. & Pääbo, S., 2001. Ancient DNA. *Nat. Rev. Genet.*, 2, pp.353–359.
Holloway, K.L., Henneberg, R.J., de Barros Lopes, M. & Henneberg, M., 2011. Evolution of human tuberculosis: A systematic review and meta-analysis of paleopathological evidence. *HOMO*, 62(6), pp.402–458.
Hoss, M., Jaruga, P., Zastawny, T.H., Dizdaroglu, M. & Paabo, S., 1996. DNA damage and DNA sequence retrieval from ancient tissues. *Nucleic Acids Res.*, 24(7), pp.1304–1307.
Hulse, E. V, 1972. Leprosy and ancient Egypt. *Lancet*, 2(7788), pp.1203–1204.
Hunter, J.M. & Thomas, M.O., 1984. Hypothesis of leprosy, tuberculosis and urbanization in Africa. *Soc. Sci. & Med.*, 19(1), pp.27–57.
Imaeda, T., Kirchheimer, W.F. & Barksdale, L., 1982. DNA isolated from Mycobacterium leprae: Genome size, base ratio, and homology with other related bacteria as determined by optical DNA-DNA reassociation. *J. Bacteriol.*, 150(1), pp.414–417.
Janin, J., 1979. Surface and inside volumes in globular proteins. *Nat.*, 277(5696), pp.491–492.
Jessamine, P.G., Desjardins, M., Gillis, T., Scollard, D., Jamieson, F., Broukhanski, G., Chedore, P. & McCarthy, A., 2012. Leprosy-like illness in a patient with Mycobacterium lepromatosis from Ontario, Canada. *J. Drugs Dermatol.*, 11(2), pp.229–233.
Jonsson, H., Ginolhac, A., Schubert, M., Johnson, P.L.F. & Orlando, L., 2013. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, 29(13), pp.1682–1684.
Jopling, W. & McDougall, A., 1988. *Handbook of leprosy* 4th ed., Heinemann Professional.
Kaplan, G. & Cohn, Z.A., 1986. The immunobiology of leprosy. *Int. Rev. Exp. Pathol.*, 28, pp.45–78.
Kapopoulou, A., Lew, J.M. & Cole, S.T., 2011. The MycoBrowser portal: A comprehensive and manually annotated resource for mycobacterial genomes. *Tuberculosis*, 91(1), pp.8–13.
Katoh, K. & Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.*, 30(4), pp.772–780.
Kay, G.L., Sergeant, M.J., Zhou, Z., Chan, J.Z.-M., Millard, A., Quick, J., Szikossy, I., Pap, I., Spigelman, M., Loman, N.J., Achtman, M., Donoghue, H.D. & Pallen, M.J., 2015. Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. *Nat. Commun.*, 6, p.6717.
Kircher, M., Sawyer, S. & Meyer, M., 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.*, 40(1), pp.1–8.
Kirchheimer, W.F. & Storrs, E.E., 1971. Attempts to establish the armadillo (Dasypus novemcinctus Linn.) as a model for the study of leprosy. I. Report of lepromatoid leprosy in an experimentally infected armadillo. *Int. J. Leprosy*, 39(3), pp.693–702.
Kjellström, A., 2012. Possible cases of leprosy and tuberculosis in medieval Sigtuna, Sweden. *Int. J. Osteoarchaeol.*, 22(3), pp.261–283.
Knapp, M., Clarke, A.C., Horsburgh, K.A. & Matisoo-Smith, E.A., 2012. Setting the stage - Building and working in an ancient DNA laboratory. *Ann. Anat.*, 194(1), pp.3–6.
Krause, J., 2010. From Genes to Genomes : What is new in ancient DNA ? *Mitteilungen der Gesellschaft für Urgeschichte*, 19, pp.11–33.
Kriventseva, E. V., Tegenfeldt, F., Petty, T.J., Waterhouse, R.M., Simão, F.A., Pozdnyakov, I.A., Ioannidis, P. & Zdobnov, E.M., 2015. OrthoDB v8: Update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Res.*, 43(D1), pp.D250–D256.
Kyte, J. & Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, 157(1), pp.105–132.

Lamers, R., Hayter, S. & Matheson, C.D., 2009. Postmortem miscoding lesions in sequence analysis of human ancient mitochondrial DNA. *J. Mol. Evol.*, 68(1), pp.40–55.

Lee, E.J., Makarewicz, C., Renneberg, R., Harder, M., Krause-Kyora, B., Müller, S., Ostritz, S., Fehren-Schmitz, L., Schreiber, S., Müller, J., Von Wurmb-Schwark, N. & Nebel, A., 2012. Emerging genetic patterns of the european neolithic: Perspectives from a late neolithic bell beaker burial site in Germany. *Am. J. Phys. Anthropol.*, 148(4), pp.571–579.

Lenski, R.E. & May, R.M., 1994. The evolution of virulence in parasites and pathogens: reconciliation between two competing hypotheses. *J. Theor. Biol.*, 169(3), pp.253–265.

Lietman, T., Porco, T. & Blower, S., 1997. Leprosy and tuberculosis: The epidemiological consequences of cross-immunity. *Am. J. Public Health*, 87(12), pp.1923–1927.

Lindahl, T., 1997. Facts and artifacts of ancient DNA. *Cell*, 90(1), pp.1–3.

Lindahl, T., 1993. Instability and decay of the primary structure of DNA. *Nat.*, 362, pp.709–715.

Lindahl, T., 2013. My Journey to DNA Repair. *Genomics, Proteomics and Bioinformatics*, 11(1), pp.2–7.

Loughry, W.J., Truman, R.W., McDonough, C.M., Tilak, M.-K., Garnier, S. & Delsuc, F., 2009. Is leprosy spreading among nine-banded armadillos in the southeastern United States? *J. Wildlife. Dis.*, 45(1), pp.144–52.

Lüdtke, H., 1997. Die archäologischen Untersuchungen unter dem Schleswiger Rathausmarkt. In *Kirche und Gräberfeld des 11 : -13. Jahrhunderts unter dem Rathausmarkt von Schleswig.* Neumünster: Wachholtz, pp. 9–84.

Lurie, M.B., 1955. *Ciba Foundation Symposium - Experimental Tuberculosis: Bacillus and Host (with an Addendum on Leprosy)* G. E. W. Wolstenholme & M. P. Cameron, eds., Chichester, UK: John Wiley & Sons, Ltd.

Mackaness, G.B., 1968. The immunology of antituberculous immunity. *Am. Rev. Respir. Dis.*, 97(3), pp.337–344.

Manchester, K., 1984. Tuberculosis and leprosy in antiquity: an interpretation. *Med. Hist.*, 28(2), pp.162–173.

Maricic, T., Whitten, M. & Pääbo, S., 2010. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS ONE*, 5(11), pp.9–13.

Matheson, C.D., Vernon, K.K., Lahti, A., Fratpietro, R., Spigelman, M., Gibson, S., Greenblatt, C.L. & Donoghue, H.D., 2009. Molecular exploration of the first-century tomb of the shroud in Akeldama, Jerusalem. *PLoS ONE*, 4(12).

May, R.M. & Anderson, R.M., 1983. Epidemiology and Genetics in the Coevolution of Parasites and Hosts. *P. Roy. Soc. B-Biol. Sci.*, 219(1216), pp.281–313.

Mays, S. & Taylor, G.M., 2003. A first prehistoric case of tuberculosis from Britain. *Int. J. Osteoarchaeol.*, 13(4), pp.189–196.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M.A., 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, 20(9), pp.1297–1303.

McMurray, D.N., 1996. (Book) Mycobacteria and Nocardia. In *Medical Microbiology*. Galveston: University of Texas Medical Branch at Galveston.

Mendum, T. a, Schünemann, V.J., Roffey, S., Taylor, G.M., Wu, H., Singh, P., Tucker, K., Hinds, J., Cole, S.T., Kierzek, A.M., Nieselt, K., Krause, J. & Stewart, G.R., 2014. Mycobacterium leprae genomes from a British medieval leprosy hospital: towards understanding an ancient epidemic. *BMC Genomics*, 15(1), p.270.

Meyer, M. & Kircher, M., 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, 5(6).

Meyers, R.A., 1995. *Molecular biology and biotechnology: A comprehensive desk reference*, Wiley.

Michel, A.L., Müller, B. & van Helden, P.D., 2010. Mycobacterium bovis at the animal-human interface: A problem, or not? *Vet. Microbiol.*, 140(3-4), pp.371–381.

Mira, A., Ochman, H. & Moran, N.A., 2001. Deletional bias and the evolution of bacterial genomes. *Trends Genet.*, 17(10), pp.589–596.

Misch, E. a, Berrington, W.R., Vary, J.C. & Hawn, T.R., 2010. Leprosy and the human genome. *Microbiol. Mol. Biol. R.*, 74(4), pp.589–620.

Mitchell, D., Willerslev, E. & Hansen, A., 2005. Damage and repair of ancient DNA. *Mutat. Res.*, 571(1-2), pp.265–76.

Molak, M. & Ho, S.Y.W., 2011. Evaluating the impact of post-mortem damage in ancient DNA: A theoretical approach. *J. Mol. Evol.*, 73(3-4), pp.244–255.

Möller-Christensen, V., 1978. *Leprosy changes of the skull*, Odense: Odense University Press.

Monot, M., Honoré, N., Garnier, T., Araoz, R., Coppée, J.-Y., Lacroix, C., Sow, S., Spencer, J.S., Truman, R.W., Williams, D.L., Gelber, R., Virmond, M., Flageul, B., Cho, S.-N., Ji, B., Paniz-Mondolfi, A., Convit, J., Young, S., Fine, P.E., Rasolofo, V., Brennan, P.J. & Cole, S.T., 2005. On the origin of leprosy. *Science*, 308(5724), pp.1040–1042.

Monot, M., Honore, N., Garnier, T., Zidane, N., Sherafi, D., Paniz-Mondolfi, A., Matsuoka, M., Taylor, G.M., Donoghue, H.D., Bouwman, A., Mays, S., Watson, C., Lockwood, D., Khamispour, A., Dowlati, Y., Jianping, S., Rea, T.H., Vera-Cabrera, L., Stefani, M.M., Banu, S., Macdonald, M., Sapkota, B.R., Spencer, J.S., Thomas, J., Harshman, K., Singh, P., Busso, P., Gattiker, A., Rougemont, J., Brennan, P.J. & Cole, S.T., 2009. Comparative genomic and phylogeographic analysis of Mycobacterium leprae. *Nat. Genet.*, 41(12), pp.1282–1289.

Müller, R., Roberts, C.A. & Brown, T.A., 2014. Biomolecular identification of ancient Mycobacterium tuberculosis complex DNA in human remains from Britain and continental Europe. *Am. J. Phys. Anthropol.*, 153(2), pp.178–189.

Nei, M. & Kumar, S., 2000. *Molecular Evolution and Phylogenetics*, Oxford University Press.

Ng, V., Zanazzi, G., Timpl, R., Talts, J.F., Salzer, J.L., Brennan, P.J. & Rambukkana, A., 2000. Role of the cell wall phenolic glycolipid-1 in the peripheral nerve predilection of Mycobacterium leprae. *Cell*, 103(3), pp.511–524.

Noonan, J.P., 2005. Genomic sequencing of pleistocene cave bears. *Science*, 309(5734), pp.597–599.

Nuorala, E., 2004. *Molecular palaeopathology: Analyses of the bacterial diseases tuberculosis and leprosy*, Stockholm: The Archaeological Research Laboratory.

O'Reilly, L.M. & Daborn, C.J., 1995. The epidemiology of Mycobacterium bovis infections in animals and man: a review. *Tubercle Lung Dis.*, 76 Suppl 1, pp.1–46.

Overballe-Petersen, S., Orlando, L. & Willerslev, E., 2012. Next-generation sequencing offers new insights into DNA degradation. *Trends Biotechnol.*, 30(7), pp.364–368.

Pääbo, S., 1989. Ancient DNA: extraction, characterization, molecular cloning and enzymatic amplification. *Proc. Natl. Acad. Sci. U.S.A.*, 86, pp.1939–1943.

Pääbo, S., 1985. Molecular cloning of ancient egyptian mummy DNA. *Nat.*, 314(6012), pp.644–645.

Pääbo, S., Poinar, H., Serre, D., Jaenicke-Després, V., Hebler, J., Rohland, N., Kuch, M., Krause, J., Vigilant, L. & Hofreiter, M., 2004. Genetic analyses from ancient DNA. *Annu. Rev. Genet.*, 38(1), pp.645–679.

Palfi, G., Dutour, O., Deak, J. & Hutas, I., 1999. *Tuberculosis. Past and Present* T. Foundation, ed., Budapest: Golden Book Publisher.

Peltzer, A., Jäger, G., Herbig, A., Seitz, A., Kniep, C., Krause, J. & Nieselt, K., 2016. EAGER: efficient ancient genome reconstruction. *Genome Biol.*, 17(1), p.60.

Pilli, E., Modi, A., Serpico, C., Achilli, A., Lancioni, H., Lippi, B., Bertoldi, F., Gelichi, S., Lari, M. & Caramelli, D., 2013. Monitoring DNA contamination in handled vs. directly excavated ancient human skeletal remains. *PLoS ONE*, 8(1), pp.1–6.

Poinar, H.N., Höss, M., Bada, J.L. & Pääbo, S., 2008. Amino acid racemization and the preservation of ancient DNA. *Science*, 272(5263), pp.864–866.

Pozzoli, U., Menozzi, G., Fumagalli, M., Cereda, M., Comi, G.P., Cagliani, R., Bresolin, N. & Sironi,

M., 2008. Both selective and neutral processes drive GC content evolution in the human genome. *BMC Evol. Biol.*, 8(1), p.99.

Rafi, A., Spigelman, M., Stanford, J., Lemma, E., Donoghue, H. & Zias, J., 1994. Mycobacterium leprae DNA from ancient bone detected by PCR. *Lancet*, 343(8909), pp.1360–1361.

Reader, R., 1974. New evidence for the antiquity of leprosy in early Britain. *J. Archaeol. Sci.*, 1, pp.205–207.

Ridley, D.S. & Jopling, W.H., 1966. Classification of leprosy according to immunity. A five-group system. *Int. J. Leprosy*, 34(3), pp.255–273.

Rizzi, E., Lari, M., Gigli, E., De Bellis, G. & Caramelli, D., 2012. Ancient DNA studies: new perspectives on old samples. *Genet. Sel. Evol.*, 44(1), p.21.

Roberts, C.A. & Manchester, K., 1996. *The archaeology of disease* 3rd ed., The History Press.

Rodrigues, L.C. & Lockwood, D.N.J., 2011. Leprosy now: Epidemiology, progress, challenges, and research gaps. *Lancet Infect. Dis.*, 11(6), pp.464–470.

Saitou, N. & Nei, M., 1987. The Neighbor-joining method : A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4(4), pp.406–425.

Sawyer, S., Krause, J., Guschanski, K., Savolainen, V. & Pääbo, S., 2012. Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS ONE*, 7(3).

Scheuer, J.L., 1992. Human Paleopathology - Current Syntheses and Future Options. *J. Anat.*, 180(Pt 1), pp.213–214.

Schuenemann, V.J., Singh, P., Mendum, T. a, Krause-Kyora, B., Jäger, G., Bos, K.I., Herbig, A., Economou, C., Benjak, A., Busso, P., Nebel, A., Boldsen, J.L., Kjellström, A., Wu, H.H.H.T., Stewart, G.R., Taylor, G.M., Bauer, P., Lee, O.Y.-C., Wu, H.H.H.T., Minnikin, D.E., Besra, G.S., Tucker, K., Roffey, S., Sow, S.O., Cole, S.T., Nieselt, K. & Krause, J., 2013. Genome-wide comparison of medieval and modern Mycobacterium leprae. *Science*, 341(6142), pp.179–83.

Scollard, D.M., Adams, L.B., Gillis, T.P., Krahenbuhl, J.L., Truman, R.W. & Williams, D.L., 2006. The continuing challenges of leprosy. *Clinical Microbiol. Rev.*, 19(2), pp.338–381.

Seguin-Orlando, A., Hoover, C.A., Vasiliev, S.K., Ovodov, N.D., Shapiro, B., Cooper, A., Rubin, E.M., Willerslev, E. & Orlando, L., 2015. Amplification of TruSeq ancient DNA libraries with AccuPrime Pfx: consequences on nucleotide misincorporation and methylation patterns. *Sci. & Technol. Archaeol. Res.*, 1(1), pp.1–9.

Shapiro, B. & Hofreiter, M., 2012. *Ancient DNA: Methods and protocols* B. Shapiro & M. Hofreiter, eds., Totowa, NJ: Humana Press.

Sharma, R., Singh, P., Loughry, W.J., Lockhart, J.M., Inman, W.B., Duthie, M.S., Pena, M.T., Marcos, L.A., Scollard, D.M., Cole, S.T. & Truman, R.W., 2015. Zoonotic leprosy in the Southeastern United States. *Emerg. Infect. Dis.*, 21(12), pp.2127–2134.

Shimoji, Y., Ng, V., Matsumura, K., Fischetti, V.A. & Rambukkana, A., 1999. A 21-kDa surface protein of Mycobacterium leprae binds peripheral nerve laminin-2 and mediates Schwann cell invasion. *Proc. Natl. Acad. Sci. U.S.A.*, 96(17), pp.9857–62.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D. & Higgins, D.G., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, 7(1), p.539.

Silva, F.J., Latorre, A. & Moya, A., 2001. Genome size reduction through multiple events of gene disintegration in Buchnera APS. *Trends Genet.*, 17(11), pp.615–618.

Singh, P., Benjak, A., Schuenemann, V.J., Herbig, A., Avanzi, C., Busso, P., Nieselt, K., Krause, J., Vera-Cabrera, L. & Cole, S.T., 2015. Insight into the evolution and origin of leprosy bacilli from the genome sequence of Mycobacterium lepromatosis. *Proc. Natl. Acad. Sci. U.S.A.*, 112(14), pp.4459–64.

Singh, P. & Cole, S.T., 2011. Mycobacterium leprae : genes, pseudogenes and genetic diversity. *Future Microbiol.*, 6(1), pp.57–71.

Skoglund, P., Northoff, B.H., Shunkov, M. V, Derevianko, A.P., Pääbo, S., Krause, J. & Jakobsson,

M., 2014. Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proc. Natl. Acad. Sci. U.S.A.*, 111(6), pp.2229–34.

Spierings, E., de Boer, T., Wieles, B., Adams, L.B., Marani, E. & Ottenhoff, T.H., 2001. Mycobacterium leprae-specific, HLA class II-restricted killing of human Schwann cells by CD4+ Th1 cells: a novel immunopathogenic mechanism of nerve damage in leprosy. *J. Immunol.*, 166(10), pp.5883–5888.

Spigelman, M. & Lemma, E., 1993. The use of the polymerase chain reaction (PCR) to detect Mycobacterium tuberculosis in ancient skeletons. *Int. J. Osteoarchaeol.*, 3(March), pp.137–143.

Stieglmeier, M., Alves, R. & Schleper, C., 2014. *The Prokaryotes* 3rd ed. M. Dworkin et al., eds., Springer Science.

Stone, A.C., Wilbur, A.K., Buikstra, J.E. & Roberts, C.A., 2009. Tuberculosis and leprosy in perspective. *Am. J. Phys. Anthropol.*, 140(SUPPL. 49), pp.66–94.

Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B. & Wu, C.H., 2015. UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6), pp.926–932.

Tamura, K., 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Mol. Biol. Evol.*, 9(4), pp.678–687.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S., 2011. MEGA5 : Molecular evolutionary genetics analysis using Maximum Likelihood , Evolutionary Distance , and Maximum Parsimony methods research resource. *Mol. Biol. Evol.*, 28(10), pp.2731–2739.

Taylor, G.M., Tucker, K., Butler, R., Pike, A.W.G., Lewis, J., Roffey, S., Marter, P., Lee, O.Y.C., Wu, H.H.T., Minnikin, D.E., Besra, G.S., Singh, P., Cole, S.T. & Stewart, G.R., 2013. Detection and strain typing of ancient Mycobacterium leprae from a medieval leprosy hospital. *PLoS ONE*, 8(4).

Team, R. development core, 2008. *R: A language and environment for statistical computing.*, Vienna: R Foundation for Statistical Computing.

Toh, H., Weiss, B.L., Perkin, S.A.H., Yamashita, A., Oshima, K., Hattori, M. & Aksoy, S., 2006. Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of Sodalis glossinidius in the tsetse host. *Genome Res.*, 16(2), pp.149–156.

Trautman, J.R., 1984. A brief history of Hansen's disease. *B. New York Acad. Med.*, 60(7), pp.689–95.

Truman, R.W., Singh, P., Sharma, R., Busso, P., Rougemont, J., Paniz-Mondolfi, A., Kapopoulou, A., Brisse, S., Scollard, D.M., Gillis, T.P. & Cole, S.T., 2011. Probable zoonotic leprosy in the southern United States. *New. Engl. J. Med.*, 364(17), pp.1626–1633.

Valverde, C.R., Canfield, D., Tarara, R., Esteves, M.I. & Gormus, B.J., 1998. Spontaneous leprosy in a wild-caught cynomolgus macaque. *Int. J. Leprosy*, 66(2), pp.140–148.

Vera-Cabrera, L., Escalante-Fuentes, W.G., Gomez-Flores, M., Ocampo-Candiani, J., Busso, P., Singh, P. & Cole, S.T., 2011. Case of diffuse lepromatous leprosy associated with "Mycobacterium lepromatosis." *J. Clin. Microbiol.*, 49(12), pp.4366–4368.

Walther, B.A. & Ewald, P.W., 2004. Pathogen survival in the external environment and the evolution of virulence. *Biol. Rev. Camb. Philos.*, 79(4), pp.849–869.

Wang, H., Avican, K., Fahlgren, A., Erttmann, S.F., Nuss, A.M., Dersch, P., Fallman, M., Edgren, T. & Wolf-Watz, H., 2016. Increased plasmid copy number is essential for Yersinia T3SS function and virulence. *Science*, 353(6298), pp.492–495.

Wayne, L.G. & Gross, W.M., 1968. Isolation of deoxyribonucleic acid from mycobacteria. *J. Bacteriol.*, 95(4), pp.1481–1482.

WHO, 2016. Leprosy fact sheet.

Wilbur, A.K., Bouwman, A.S., Stone, A.C., Roberts, C.A., Pfister, L.A., Buikstra, J.E. & Brown, T.A., 2009. Deficiencies and challenges in the study of ancient tuberculosis DNA. *J. Archaeol. Sci.*, 36(9), pp.1990–1997.

Willerslev, E. & Cooper, A., 2005. Ancient DNA. *P. Roy. Soc. B-Biol. Sci.*, 272(1558), pp.3–16.
Wirth, T., Hildebrand, F., Allix-Béguec, C., Wölbeling, F., Kubica, T., Kremer, K., van Soolingen, D., Rüsch-Gerdes, S., Locht, C., Brisse, S., Meyer, A., Supply, P. & Niemann, S., 2008. Origin, spread and demography of the Mycobacterium tuberculosis complex M. Achtman, ed. *PLoS Pathog.*, 4(9), p.e1000160.
Yang, D.Y. & Watt, K., 2005. Contamination controls when preparing archaeological remains for ancient DNA analysis. *J. Archaeol. Sci.*, 32(3), pp.331–336.
Zias, J., 2002. New evidence for the history of leprosy in the ancient Near East: An overview. In C. A. Roberts, M. Lewis, & K. Manchester, eds. *The past and present of leprosy. Archeological, historical, paleopathological and clinical approaches*. Owxford: Archaeopress, pp. 259–268.

9 Declaration

Herewith, I confirm that the submitted thesis content and study design is the result of my own work. Apart from the advice of my supervisors, all sources and cooperation partners are listed within the thesis. This thesis has not been wholly or partially submitted elsewhere as part of a doctoral degree, nor has it been published or submitted for publication. This thesis has been carried out in strict accordance with the Rules of Good Scientific Practice of the Deutsche Forschungsgesellschaft.

Signature Date

10 Curriculum Vitae

Name: Marion Bonazzi

Date of birth: 20th of June 1989

Place of Birth: Paris

Nationality: French

Education and degrees:

2004 Brevet des Colleges

2007 Scientific Baccalaureat with distinctions

2012 Biotechnology Engineer (eq. Master of Science), ESBS Biotechnology School, University of Strasbourg, with the project "*Quantitative study of DNA loss caused by the Computed Tomography exposure of ancient remains".*

Publications:

A. Immel, D. Drucker, M. Bonazzi, T. Jahnke, S. Münzel, V. Schuenemann, A. Herbig, C.-J. Kind, J. Krause, 2016, Mitochondrial Genomes of Giant Deers Suggest their Late Survival in Central Europe, *Sci. Rep.*

M. Bonazzi*, A. Immel*, A. Le Cabec*, A. Herbig, H. Temming, V.J. Schuenemann, K. Bos, K. Harvati, A. Bridault, G. Pion, N. Conard, S. Münzel, D Drucker, B. Viola, J.-J. Hublin, P. Tafforeau, J. Krause, 2016, Effect of x-ray irradiation on ancient DNA in sub-fossil bones - Guidelines for safe x-ray imaging, *Sci. Rep*.

11 Acknowledgements

Numerous institutions and persons provided me with the guidance and support which allowed this PhD study to reach completion. Therefore, I wish to express my gratitude for to all who have taken on themselves to entrust me with their help.

The Graduate School Human Development in Landscapes and the Institute of Clinical Molecular Biology for funding my research and giving me the opportunity to do my PhD in an excellent work environment, and for providing me with technical and logistical support for the duration of the project.

Prof. Dr. Almut Nebel and Prof. Dr. Manuela Dittmar for their scientific guidance and support during the study and the writing of the dissertation.

Jun.-Prof. Dr. Ben Krause Kyora for welcoming me into the Ancient DNA Lab Kiel and granting me access to the laboratories. His patience and will to provide scientific as well as moral support were most appreciated in times of doubt.

Prof. Dr. Jesper Boldsen and Dorthe Pedersen for sharing their archaeological and palaeopathological expertise, and for their help in accessing and sampling the remains.

Prof. Dr. Wiebke Kirleis for being a supportive Graduate School mentor and understanding the difficulties of interdisciplinary studies.

The technicians from the IKMB sequencing platforms for their excellent work and their will to help when modern protocols had to be adapted to fit this aDNA study.

All my colleagues, especially Marcel Nutsua, Lisa Böhme, Lena Möbus, Julian Suzat, Sabin Kornell, Yara Burmeister and Guillermo Torres for their constant support, advice and helpful comments during the study. In addition, Lisa Böhme and Marcel Nutsua greatly helped with the writing of the dissertation.

Finally, I would like to send my best regards and gratitude to my dear friends Jessica Krause, Artur Ribeiro, Marco Zanon, Shénila Katchera, Delphine Cheron, Camille Butruille, Florian Bauer, Nicole Taylor and Steven Rogers, as well as to Stéphane Brand and my family, for their constant moral support. They were always there to pull me back on my feet when I needed it.

12 Appendices

12.1 Supplementary material

12.1.1 Lists of samples

The archaeological and paleopathological data was provided by Dorthe Petersen and Jesper Boldsen. Disease status was based on observed disease-specific bone lesions. Undetermined status (und.) was used when the lesions observed were suggestive of a disease but not specific. When available, archaeological dating of the samples is provided.

Supplementary Table 1: St Jørgen cemetery samples (Odense, Denmark)
Thirty-four individuals from a leprosarium cemetery in use between 1270 and 1560

Excavation ID	Osteological sex	Age	Time period (arm position)	Leprosy status*	Rhinomaxillary lesions	TB status
SJG_G022	F	14-17	1000-1250	Und.	yes	0
SJG_G035	F	28-35	1000-1375	+	yes	0
SJG_G1035	F	60-75	1000-1250	Und.	yes	Und.
SJG_G1087	? (child)	12-14	1000-1375	+	yes	Und.
SJG_G1137	F	37-47	1000-1250	+	yes	+
SJG_G1159	F	15-17	1000-1250	Und.	yes	Und.
SJG_G131	F	16-20	1000-1375	+	yes	Und.
SJG_G149	F	21-24	1000-1375	+	yes	Und.
SJG_G189	F	22-25	1000-1375	+	yes	0
SJG_G271	F	55-65	1000-1375	+	yes	Und.
SJG_G289	F	19-22	1000-1375	+	yes	+
SJG_G295	F	32-40	1000-1375	+	yes	Und.
SJG_G379	M	60-70	1000-1375	+	yes	Und.
SJG_G404	M	23-26	1000-1375	+	yes	+
SJG_G427	F	22-24	1000-1375	+	yes	Und.
SJG_G472	F	19-22	1000-1375	+	yes	+
SJG_G507	F	16-19	1000-1375	+	yes	0
SJG_G533	F	23-26	1000-1375	+	yes	Und.
SJG_G604	M	38-48	1000-1375	+	yes	+
SJG_G611	M	35-45	1000-1250	+	yes	Und.
SJG_G658	M	23-27	1000-1375	Und.	no	Und.
SJG_G711	M	24-26	1000-1250	+	yes	Und.
SJG_G718	F	16-18	1000-1375	+	yes	Und.
SJG_G722	M	32-42	1000-1250	Und.	no	+
SJG_G749	F	18-22	1000-1375	Und.	yes	Und.
SJG_G792	M	35-45	1000-1375	+	yes	+
SJG_G846	M	35-45	1000-1375	Und.	no	Und.
SJG_G859	M	26-32	1000-1375	+	yes	+
SJG_G872	M	32-42	1000-1375	Und.	no	+
SJG_G877	F	30-40	1000-1375	Und.	yes	+
SJG_G928	F	28-34	1000-1375	+	yes	+
SJG_G947	nd	nd	nd	nd	nd	nd
SJG_G978	M	29-36	1000-1375	+	yes	+
SJG_G988	M	35-45	1000-1375	Und.	yes	Und.

Supplementary Table 2: Samples from the Rathausmarkt cemetery (Schleswig, Germany)

Seventy-nine samples collected from the cemetery in use between 1000 and 1250

Excavation ID	Osteological sex	Age	Time period	Leprosy status*	Rhinomaxillary lesions	TB status
2	M	38-52	1070-1210	0	no	+
7	F	18-30	1070-1210	0	no	Und.
8	M	23-25	1070-1210	Und.	yes	Und.
9	F	30-45	1070-1210	0	no	0
20	F	16-18	1070-1210	Und.	yes	0
21	F	35-50	1070-1210	+	yes	+
23	F	30-35	1070-1210	0	no	Und.
28	F	50-70	1070-1210	Und.	yes	Und.
42	M	37-45	1070-1210	Und.	yes	+
49	F	35-50	1070-1210	0	no	+
52	F	50-70	1070-1210	Und.	yes	Und.
57	M	60-75	1070-1210	Und.	yes	+
58	M	35-45	1070-1210	0	no	nd
60	M	35-50	1070-1210	+	yes	Und.
70	M	35-45	1070-1210	Und.	yes	Und.
73	F	38-53	1070-1210	0	no	Und.
78	F	40-50	1070-1210	Und.	yes	Und.
80	M	28-35	1070-1210	Und.	yes	0
81	M	40-65	1070-1210	0	no	+
86	M	45-60	1070-1210	Und.	yes	Und.
88	M	33-40	1070-1210	Und.	no	Und.
90	F	55-65	1070-1210	0	no	+
91	M	16-18	1070-1210	0	no	Und.
92	M	30-45	1070-1210	Und.	yes	Und.
94	M	28-33	1070-1210	Und.	yes	+
95	M	47-60	1070-1210	0	no	Und.
100	F	18-20	1070-1210	0	no	0
101	M	25-30	1070-1210	0	no	Und.
103	M	55-65	1070-1210	Und.	yes	+
105	M	25-28	1070-1210	Und.	yes	Und.
109	M	17-20	1070-1210	0	no	Und.
113	M	26-36	1070-1210	0	no	Und.
114	M	28-32	1070-1210	+	yes	+
120	M	55-70	1070-1210	+	yes	+
121	M	38-48	1070-1210	Und.	yes	nd
125	F	40-50	1070-1210	Und.	yes	Und.
131	F	50-70	1070-1210	0	no	Und.
133	M	45-60	1070-1210	Und.	yes	+
134	3	55-70	1070-1210	Und.	yes	nd
135	F	20-24	1070-1210	Und.	yes	0
137	M	26-32	1070-1210	0	no	Und.
139	F	45-55	1070-1210	Und.	yes	+
140	M	19-21	1070-1210	+	yes	+
142.1	M	21-25	1070-1210	Und.	yes	Und.
146	M	40-60	1070-1210	Und.	yes	Und.
155	M	20-30	1070-1210	Und.	yes	nd
156	M	22-25	1070-1210	0	no	0
158	F	18-20	1070-1210	0	no	0
159	M	24-34	1070-1210	0	no	Und.
160	M	23-26	1070-1210	Und.	yes	0
163.9	F	30-70	1070-1210	0	no	nd
165	F	30-45	1070-1210	Und.	yes	Und.
167	M	45-60	1070-1210	Und.	yes	0
170	M	22-26	1070-1210	0	no	Und.
171	M	30-45	1070-1210	Und.	yes	Und.
172	F	27-35	1070-1210	Und.	yes	Und.
173	F	23-25	1070-1210	Und.	yes	Und.
178	M	40-70	1070-1210	0	no	nd
179	M	22-25	1070-1210	+	yes	Und.
183	M	50-65	1070-1210	0	no	Und.
185	M	47-63	1070-1210	Und.	yes	+
186	M	38-50	1070-1210	0	no	Und.
191	M	25-30	1070-1210	Und.	no	Und.

193	M	22-25	1070-1210	0	no	+
195	M	25-40	1070-1210	Und.	yes	nd
197	M	23-26	1070-1210	0	no	Und.
200	M	25-45	1070-1210	Und.	yes	Und.
202	F	47-63	1070-1210	0	no	+
203	M	50-65	1070-1210	+	yes	Und.
205	M	45-55	1070-1210	0	no	+
207	M	36-46	1070-1210	+	yes	Und.
208	F	45-60	1070-1210	0	no	+
214	M	22-33	1070-1210	+	yes	Und.
215	F	45-65	1070-1210	Und.	yes	+
219	F	35-45	1070-1210	Und.	yes	Und.
221	M	22-24	1070-1210	Und.	no	0
230	M	17-19	1070-1210	0	no	Und.
232	F	30-65	1070-1210	Und.	yes	Und.
233	M	22-27	1070-1210	Und.	yes	0

Supplementary Table 3: Samples from the Dagmargården cemetery (Ribe, Denmark)

Excavation ID	Osteological sex	Age	Time period	Leprosy status	Rhinomaxillary lesions	TB status
G17	M	25-30	1275-1536	0	no	Und.
G23	M	32-40	1275-1536	Und.	no	Und.
G25A	M	40-50	1275-1536	Und.	no	+
G26	M	28-34	1275-1536	0	no	+
G35	M	50-60	1275-1536	Und.	yes	+
G40	M	20-22	1275-1536	Und.	yes	+
G43	M	30-40	1275-1536	0	no	Und.
G44	M	42-52	1275-1536	Und.	yes	+
G48	M	32-40	1275-1536	Und.	yes	+
G50	M	22-24	1275-1536	0	no	Und.
G56	M	38-48	1275-1536	Und.	yes	+
G58	M	28-33	1275-1536	0	no	0
G63	M	30-40	1275-1536	0	no	Und.
G69	M	28-34	1275-1536	+	yes	Und.
G89	M	55-65	1275-1536	Und.	yes	+
G99	M	24-28	1275-1536	Und.	yes	0
G100	M	20-23	1275-1536	Und.	yes	und.
G108	M	24-26	1275-1536	0	no	+
G113	M	27-33	1275-1536	0	no	+
G117	M	23-26	1275-1536	Und.	yes	Und.
G120	? (child)	11-13	1275-1536	0	no	+
G131	M	23-25	1275-1536	0	no	Und.
G138	M	35-45	1275-1536	0	no	+
G159	M	30-38	1275-1536	0	no	0
G160	M	35-43	1275-1536	0	no	0
G176	M	37-47	1275-1536	0	no	+
G178	M	20-23	1275-1536	+	yes	Und.

12.1.2 List of enzymes, reagents and kits

Product	Specificities	Catalogue number	Supplier
AccuPrime pfx DNA amplification kit	200rxns	12344-024	Life Technologies
Adapter Mix	prepared according to (Meyer & Kircher 2010)		
Agilent DNA 1000 kit		5067-1504	Agilent technologies
ATP	10mM	P0756S	NEB
Bleach	100%		
BSA			NEB
BSM DNA Polymerase LF	1600U	EP0691	Fermentas
Buffer 2		B7002S6	NEB
DMSO	100%	D2650	Sigma-Aldrich
dNTPs mix	2.5mM each	BIO-39053	Bioline
dNTPs mix	2.5mM each	R1121	Fermentas
EDTA	pH8, 0.5M	A4895.1000	Applichem
Ethanol	96%	T171.	Roth
EZ1 DNA Investigator Extraction Kit		952034	Qiagen
Immolase DNA amplification kit	200U	BIO-21046	Bioline
Minelute PCR product purification kit	250 rxns	28006	Qiagen
Quick ligation kit		M2200S	NEB
T4 DNA Polymerase	750U	M0203L	NEB
T4 Polynucleotide Kinase	2,500U	M0201L	NEB
Thermopol Buffer		B9004S	NEB
USER Enzyme	50U	M550S	NEB
Water for molecular biology	DNA-free	W4502	Sigma-Aldrich

12.1.3 List of instruments

Product	Catalogue number / type	Supplier
Mixer Mill MM 200	20.746.0001	Retsch
EZ1 DNA Investigator	9016387	Qiagen
PCR-hood with UV-lights	UVC/T-M-AR	Grant
Table centrifuge	5424	Eppendorf
Dremel	3000	Dremel
pipets 2.5µL, 200µL, 500µL	-	Eppendorf
Thermomixer	type TS1	Biometra

12.1.4 List of consumables necessary for working with ancient DNA

Consumables	Catalogue number	Supplier
Lab coveralls	Through central IKMB stock	
Latex gloves	Through central IKMB stock	
Hair nets	Through central IKMB stock	
Surgical masks	Through central IKMB stock	
Rubber shoes	Through central IKMB stock	
Aluminium foils		
DNA free pipet tips with filters (20µL, 200µL, 1000µL)	70.760.592, 70.760.213	Sarstedt
Parafilm	RO/H95.1.1	Th. Geyer
DNA LoBing tubes (1.0 and 2.0mL)	EPP/0030108051, EPP/0030108.078	Th. Geyer

Common molecular biology consumables are not listed

12.1.5 List of online databases, tools and software

Sofware/databank	Version/ Distribution/ Last access date	Usage	Reference
Geneious	R7	PCR product alignment	(Drummond et al. 2010)
BLAST	/	PCR product control	(Altschul et al. 1990)
NCBI	June 2016	/	/
Mycobrowser	June 2016	/	(Kapopoulou et al. 2011)
MAUVE		Multiple alignments	(Darling et al. 2010)
R		Statistical tests	
SNPeff		Variant annotation	(Cingolani et al. 2012)
Uniprot	June 2016	/	(Bateman et al. 2015)
OrthoDB	June 2016	/	(Kriventseva et al. 2015)
EMBL QuickGo Browser	June 2016	GO annotations	(Binns et al. 2009)
CFSSP	June 2016	Secondary structure prediction	(Chou & Fasman 1974b; Chou & Fasman 1974a)
Jpred	version 4	Secondary structure prediction	(Drozdetskiy et al. 2015)
SOPMA	June 2016	Secondary structure prediction	(Combet et al. 2000)
ExPASy	June 2016	Resource portal	(Artimo et al. 2012)
ProtScale	June 2016	Protein profiling	(Janin 1979; Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D. 2005; Grantham 1974; Kyte & Doolittle 1982)
ProtParam	/	Protein characterisation	(Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D. 2005)
SOSUI	/	Hydrophobic protein characterisation	(Hirokawa et al. 1998)
EAGER	/	Analysis of NGS data	(Peltzer et al. 2016)

12.1.6 Oligonucleotide sequences and references

Supplementary Table 4: Oligonucleotides primers for the PCR screening

Target reference	Primer name	Reference
Human mtDNA	mt1	(Lee et al. 2012)
	mt2	(Lee et al. 2012)
M. leprae	LP1 - LP2	(Matheson et al. 2009)
	LP11 - LP12	(Schuenemann et al. 2013)
M. tuberculosis	IS3 - IS4	(Matheson et al. 2009)
	lns1 - Ins2	(Matheson et al. 2009)
	Tb-A - Tb-B	(Matheson et al. 2009)
	Tb-C - Tb-D	(Matheson et al. 2009)
	F-F2 - R-R3	(Matheson et al. 2009)
	IS1081-F2 - IS1081-F3	(Matheson et al. 2009)

12.2 Supplementary methods

12.2.1 PCR screening target sequences

Supplementary Table 5: List of PCR screening target sequences

Target reference	Primer pair	Target sequence (5' -> 3')
Human mtDNA	mt1	CTCCACCATTAGCACCCAAAGCTAAGATTCTAATTTAAACTATTCTCTGT TCTTTCATGGGGAAGCAGATTTGGGTACCACCCAAGTATTGACTCACCCA TCAACAACCGCTATGTATTTCGTACATTACTGCCAGCCACCATGAATAT GTACGGTACCATAAATACTTGACCACCTGTAGTA
	mt2	GCCAGCCACCATGAATATGTACGGTACCATAAATACTTGACCACCTGTAG TACATAAAAACCCAATCCACATCAAAACCCCCTCCCCATGCTTACAAGCA AGTACAGCAATCAACCCTCAACTATCACACATCAACTGCAACTCCAAAGC
M. leprae	LP1 - LP2	TTGCATGTCATGGCCTTGAGGTGTCGGCGTGGTCAATGTGGCCGCACCTG AACAGGCACGTCCCCGTGCACGGTATAACTATTCGCACCTGATGTTATCC CTTGCACCATTTCTGCCGCTGGTATCGGTG
	LP11 - LP12	GAGCTGCTCACCACAACAAATAGAACAAATAGGGTGGTTCTGCTTCTATT GCACCGACCAACAGTAGGAATGGTCTGAACAGGTGCAACGGATACACA
M. tuberculosis	IS3 - IS4	
	Ins1 - Ins2	
	Tb-A - Tb-B	
	Tb-C - Tb-D	
	F-F2 - R-R3	
	IS1081-F2 - IS1081-F3	

M. tuberculosis *target sequences were not used because no clear PCR results could be obtained. They are therefore not shown*

12.2.2 MinElute purification of DNA fragments

During the library preparation, indexing and amplification, several purifications are performed with MinElute columns from Qiagen. The manufacturer's protocol was used with the following modifications:

Reagent volumes:

PB Buffer	5X volume of DNA to purify (250µL for library preparation, 500µL for PCR purification)
PE Buffer	700µL per column
EB Buffer	To fit next step requirements (18µL or 20µL for library preparation, 50µL for PCR purification)

Number of columns per sample:

During library preparation, each sample was purified in one Minelute column. During the purification of the indexing and amplification PCRs products, all 4 PCR products of each sample were purified together in a column to optimize DNA recovery. To that end, 2 PCR products were mixed and bound to the column together and the process was repeated for the 2 remaining PCR products before washing the membrane with PE buffer.

Centrifugation steps

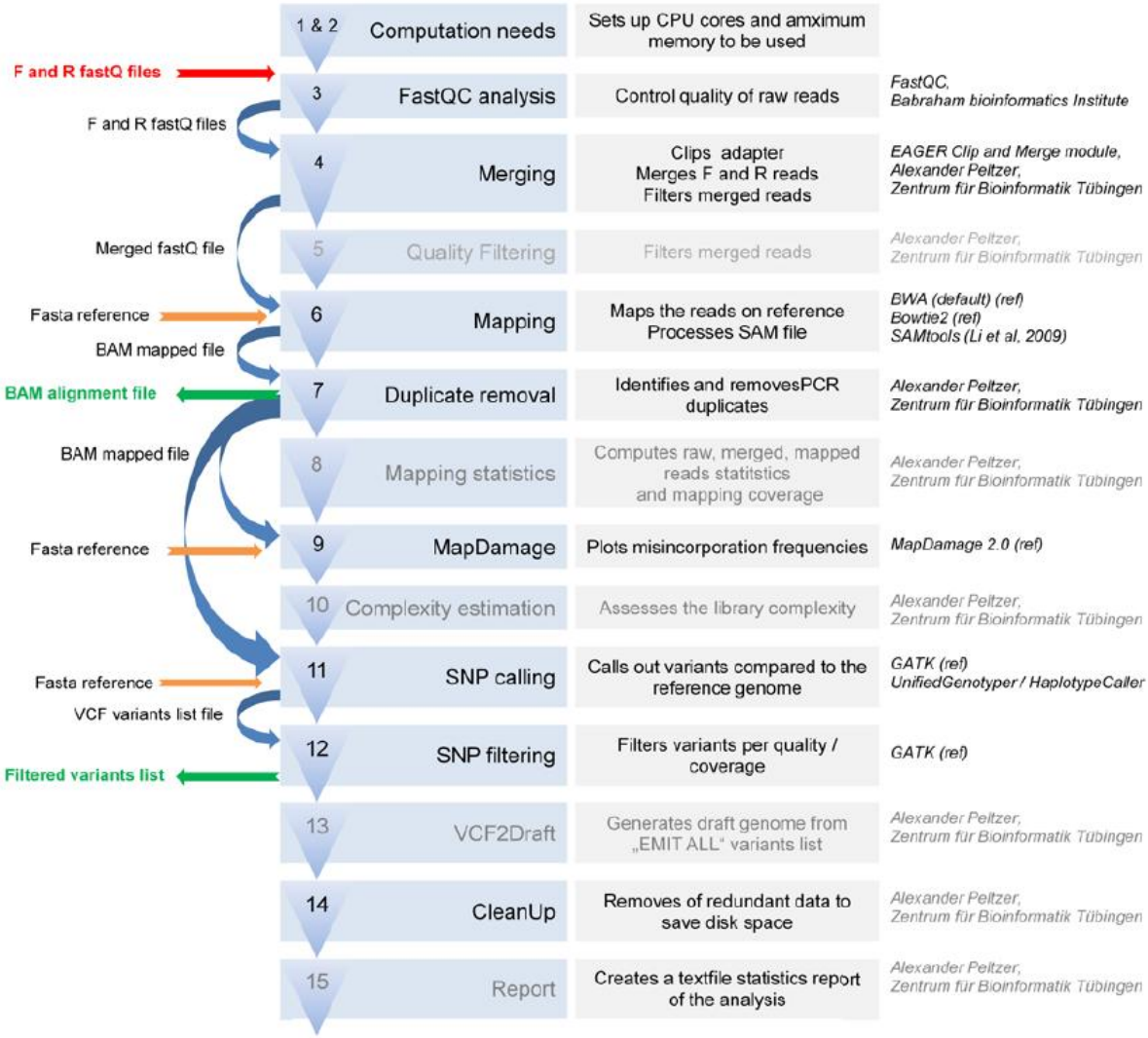
DNA binding	30s at max. speed
Washing with PE buffer	30s at max. speed
Dry-clean to remove PE traces	60s at max. speed
DNA elution in EB buffer	60s at max. speed

12.2.3 Sex determination from NGS data

Ideally, a female individual displays two copies of the X chromosome and a male individual one copy of X and one copy of Y. Therefore, comparing the coverage statistics on the X and Y chromosomes should provide a good proxy for sex determination. The statistics tables provided by the Module δ include the coverage of the reference as well as the read depth on that reference. The EAGER pipeline and Modules α to δ were used to obtain statistics about the coverage of the human X and Y chromosomes. The read depths on the two references were compared and a Y/X ratio was calculated for each sample. The number of unique reads mapping to each chromosome was used as a reference for the interpretation of the ratios, as lower number of reads would suggest less reliable mapping results. The Y/X ratio is expected to vary between 0 (Y coverage null, female individual) and 1 (Y = X situation, male individual)

12.2.4 Bioinformatic pipeline descriptions

The raw reads were processed using the EAGER (Efficient Algorithms for Ancient Human Genome Reconstruction, A. Peltzer, (Peltzer et al. 2016)) pipeline, which incorporates published bioinformatic softwares into a framework optimized for aDNA analysis. The softwares are implemented in the form of modules that can be called and set-up independently through a graphical user interface (GUI). The pipeline allows each module to be performed one after the other by automatically converting or preparing the output files from one step to fit input files requirements for the next step (see Supplementary Figure 1)



Supplementary Figure 1: Schematic overview of the EAGER analysis pipeline
The modules are listed in the blue text boxes, their function and references in the grey and transparent boxes, respectively. Modules with grey text were not used in this study, either because they were not relevant with our datasets or because they were replaced by manual analysis or self-made scripts. On the left part of the figure, an overview of the types of files used and created by each module is provided. Temporary files are not listed.

Module 1 and 2:

Module 1 and 2 only consists in setting the memory requirements in the EAGER GUI before starting the analyses. All pooled sequencing runs were processed through the EAGER pipeline with 8 cores and 42GB of memory requirements. The single-lane re-sequencing runs were processed with 8 cores and 64BG of memory.

Module 3: FastQC analysis

FastQC is a software application that provides a quality control report from a high-throughput sequencing file, allowing the sequencing issues or problems in the library preparation to be identified. It was developed by the Babraham Bioinformatics institute. For each quality control performed (see list below), FastQC provides warning or error notifications if the data present any problematic feature:

Basic statistics	Includes the basic information about the input file, the number of raw sequences as well as average GC content and sequence length
Per base sequence quality	WhiskerBox plot of the sequencing quality (as provided by the sequencer) for each position along the read.
Per tile sequence quality	Heatmap representing the sequencing quality on the flowcell during the sequencing run cycles.
Per base GC content	Plot representing the GC content distribution in the sequences and builds a model normal distribution.
Per sequence quality scores	Plot representing the average read quality distribution in the sequences.
Per base sequence content	Plots of the A, T, C and G proportions for each position along the reads
Per base N content	Sequencers usually write an N in a sequence when unable to confidently call a conventional base. This module plots the percentage of Ns for each position along the reads.
Sequence length distribution	Plots the read length (bp) distribution in the sequencing run
Sequence duplication levels	A subset of the sequencing file is used to estimate the number of copies for each read.
Overrepresented sequences	Uses the duplication level results to display the sequences that are present in unexpectedly high number
Adapter content	Cumulative percentage plot of the percentage of sequencing library adapters along the sequencing reads.
Kmer content	Relative enrichment of 7-mers along the sequencing reads

For each raw reads file received from the sequencing platform, the FastQC controls were performed and the warnings manually evaluated to identify possible problems and inconsistencies with aDNA characteristics.

Module 4: Merging

This module is a java script written by A. Peltzer that combines adapter clipping, merging of overlapping forward and reverse reads and quality filtering of the merged reads.

Since aDNA molecules are usually short, it is common to find belonging to the library or the indexing adapters at the ends of the reads sequences. If not removed, those adapter sequences might impair the rest of the analysis. Consequently, the first step of most aDNA analysis pipelines is to trim the end of the raw reads where the adapter is starting. All the adapter sequences used during library preparation and indexing are listed in a configuration file, allowing the Clip and Merge module to find them in the sequencing reads. The Clip and Merge module also integrates an option to merge forward and reverse reads from paired-end datasets. First, overlapping regions are detected in the forward and reverse FastQ files and when the overlap length is greater than the chosen threshold (10bp by default) the two reads are merged. Afterwards, new quality scores are calculated for the merged reads, taking into account the forward and reverse reads qualities. Clip and Merge writes an output file with merged reads marked as M while non-merged read keep their F or R identifiers. Finally, the module trims the reads when the base call is lower than 20 (which correspond to a 99% accuracy on Illumina sequencers) and filters out reads shorter than 25bp. A single compressed FastQ file is written in the output directory. The temporary outputs from the individual steps are removed.

Module 6: Mapping

Most DNA sequencing projects fall into the “re-sequencing” category, in which a reference genome is used as model to assemble the reads obtained by NGS sequencing. In aDNA studies, the use of a reference is even more important since reads are commonly short and genome coverage usually incomplete. The 3rd module of the EAGER pipeline performs the mapping of the sequencing reads after Module 1 (or 2 when relevant) to a chosen reference in Fasta format. Both linear and circular reference genomes are supported. Several published alignment methods are implemented into the module and can be selected by the user. By default, the mapping is done by BWA (Burrows-Wheeler Alignment tool), an algorithm that combines local sequence alignments from Smith and Waterman (1981) and the Burrows-Wheeler data compression transform

(Burrows and Wheeler, 1994). The EAGER pipeline GUI allows setting up some parameters for the alignment (See list below). With BWA, the beginning nucleotides of each read (called seed) are used as an anchor point to align the read. Once an alignment is found between the seed and the reference genome, the algorithm progressively extends the alignment until the overall alignment quality drops (Li and Durbin, 2010; BWA sourceforge website, 2015). BWA alignment output format is SAM (Sequence Alignment/Map), a file format that gathers the reads alignments to the reference as well as information about the number of mismatches, gaps, mapping quality and specificity (existence of multiple hits).

EAGER option for BWA	description	default value
-n	maximal number of mismatches allowed in the alignment	5
-l	seed length (bp) BWA includes an option to set the number of allowed mismatches in the seed alignment itself (2 per default). In EAGER this parameter cannot be modified	32

For the analyses performed in this project, BWA mapping with default settings was used.

The mapping results saved in SAM format require large data storage capacity. In addition, non-mapping reads usually form the bulk of the dataset for shotgun sequencing files, since only abokljut 1% of the reads are expected to hit the target species. To avoid rapid congestion of storing spaces, EAGER can perform several tools from the SAMtools package (Han et al. 2009; Peltzer et al. 2016). By default, EAGER sorts the mapping reads by position along the reference and compresses the file into a binary BAM file. In addition, EAGER was set up to remove non-mapping reads from the SAM file before conversion to BAM.

Module 8: Duplicate removal

Reads present in high copy numbers artificially increase the read depth while decreasing the library complexity. Moreover, variant calling might be biased in highly-duplicated regions as the confidence score for variants takes into account the number of reads covering the position of interest (Isakov et al, 2013; Walker 2009). This module was developed by A. Petlzer to remove sequences originating from PCR amplification during the NGS library preparation. In this regard, two approaches are usually available: 1) keep only one read from all the reads starting and ending at the same position (Kircher ref) or 2) keep only one read from all reads with a similar sequence. The duplicates removal module uses a combination of both approaches. The algorithm isolates

reads that start at the same position and display strong similarity (find numbers), and keeps only the read with the highest overall quality.

Module 10: MapDamage

Ancient DNA typically displays short fragments length (Green et al. 2009) and increased occurrence of G and T towards the 3'-end and 5'-end of the DNA strand, respectively (Briggs et al. 2007; Overballe-Petersen et al. 2012; Brotherton et al. 2007; Lamers et al. 2009). Therefore, the study of the DNA fragments molecular characteristics (damage patterns) can differentiate between a young DNA sample (below 20yo) and an ancient DNA sample (Molak & Ho 2011; Krause 2010; Sawyer et al. 2012), although the molecular age of the sample cannot be used as a proxy to estimate the actual sample age (Sawyer et al. 2012). For a specific reference sequence, length distribution of the mapping reads as well as frequencies of misincorporated bases can be calculated (Briggs et al. 2007; Ginolhac et al. 2011; Jonsson et al. 2013).

The MapDamage module, which applies the mapDamage2.0 alogrithms, uses the BAM file generated by the Duplicate removal module as well as the corresponding FASTA reference genome (Ginolhac et al. 2011; Jonsson et al. 2013). First of all, the BAM file is analysed to record the lenght of DNA fragments (in base pairs) and the number of occurrences of A, C, T or G bases for each position along each. Then, the reference sequence is used in comparison to compute the frequency at which a specific position on the end of the DNA fragment displays either a C->T or G->A misincorporation. Finally, as DNA fragment length is not constant and since the misincorporation frequencies are expected to increase towards the 3'- and 5'- ends, only the frequencies over the first and last 25bp are given as output.

Module 12 and 13: SNP calling and filtering with GATK

Modules 13 and 13 implement the Genome Analysis Toolkit (Van der Auwera et al. 2013; McKenna et al. 2010) software to call and filter the genetic variants seen on the sequenced genome compared to the reference genome. Two genotyping algorithms are available: UnifiedGenotyper (by default) and HaplotypeCaller. In Module 12 several advanced options are available to set up a SNP reference, the organism ploidy, the confidence of the variant calling or the type of output. In module 13 minimum coverage and base quality can be set to filter the variants called. All variants that do not fulfil those criteria will be marked as LowQuality.

12.2.5 Internally designed scripts

Several scripts were designed internally by M. Nutsua (aDNA research group, IKMB, Kiel). Those modules are identifies with greek letters.

Module α:

Module α is a python script developed to create user-specific hard links to allow each user to work on the raw file from their own directories. It requires a text file containing the names of the sample-specific raw data directories and the location of the output hard links. The EAGER pipeline can then be started directly from the hard links.

Module β:

EAGER does not separate output files depending on the reference genome. However, it skips modules when the outputs for the modules are present in the output directory. In case multiple references were used for a single sample, there was a risk of false completion of the analysis since all output files could already present from a run with a different reference genome. To allow the analysis of several species per sample with the EAGER pipeline, Module β was created to move to reference-specific outputs to subfolders within the sample output directory.

Module δ:

Module δ was designed to combine and replace Modules 8, 10 and 15 of the EAGER pipeline. It consists in a perl script that gathers statistics from each step of the EAGER pipeline into a general CSV textfile with all the samples. This file can then be easily imported into Microsoft Excel or R for sample comparison. This module was designed to access the EAGER output files also from reference-specific subfolders, when necessary. The statistics include the number of reads after each step (raw reads, clip and merged reads, mapping reads, unique mapping reads) as well as coverage and read depth information for the reference genome used. It also offers the possibility to compute the coverage and read depth information over a target subset of the reference genome.

12.3 Supplementary results

12.3.1 PCR results per sample for each primer pair

A sample is considered positive for a primer pair when both PCR replicates yielded products of the expected length and sequence. In green are highlighted samples with a positive results for all the primers tested for a species .In red are shown the samples were results were negative or inconclusive. For every PCR batch, the extraction blanks and PCR blanks showed no signs of contamination.

Supplementary Table 6: PCR screening results detailed for each sample

	Human mtDNA		*Mycobacterium leprae*		*Mycobacterium tuberculosis*						
Sample	mt1	mt2	LP 1-2	LP 11-12	TB1	TB2	TB3	TB4	TB5	TB6	TB7
Ribe G100	+	+	-	-	+	-	-	-	-	-	-
Ribe G108	-	+	-	-	-	-	-	-	-	-	+
Ribe G113	+	+	-	-	+	-	-	-	-	-	-
Ribe G117	+	+	-	-	-	-	-	-	-	-	-
Ribe G120	+	+	+	-	+	-	-	-	-	-	-
Ribe G131	+	+	+	-	+	-	-	-	-	-	-
Ribe G138	-	+	-	-	+	-	-	-	-	-	+
Ribe G159	+	+	+	-	+	-	-	+	-	-	-
Ribe G160	+	+	-	-	-	-	-	-	-	-	-
Ribe G17	-	+	-	-	+	-	-	-	-	-	-
Ribe G176	-	+	-	-	-	-	-	+	-	-	-
Ribe G178	+	+	-	-	-	-	-	-	-	-	-
Ribe G23	+	+	+	-	-	-	-	+	-	-	-
Ribe G25A	-	+	-	-	-	-	-	-	-	-	-
Ribe G26	+	+	+	-	-	-	-	-	-	-	+
Ribe G35	-	+	-	-	-	-	-	+	-	-	-
Ribe G40	+	+	+	-	-	-	-	-	-	-	-
Ribe G43	+	+	-	-	-	-	-	-	-	-	-
Ribe G44	+	+	-	-	-	-	-	-	-	-	-
Ribe G48	+	+	+	-	-	-	-	-	-	-	-
Ribe G50	+	+	-	-	-	-	-	-	-	-	-
Ribe G56	+	+	-	-	-	-	-	+	-	-	-
Ribe G58	+	+	-	-	-	-	-	-	-	-	-
Ribe G63	+	+	-	-	-	-	-	+	-	-	-
Ribe G69	+	+	-	-	-	-	-	-	-	-	-
Ribe G89	-	+	+	-	+	-	-	-	-	-	+
Ribe G99	+	+	-	-	+	-	-	-	-	-	-
Rathausmarkt 100	+	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 101	+	+	-	-	-	-	-	-	-	-	+
Rathausmarkt 103	+	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 105	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 109	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 113	+	+	-	-	-	-	-	-	-	-	-

Rathausmarkt 114	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 120	+	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 121	+	+	-	-	-	-	-	-	-	-	+
Rathausmarkt 125	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 131	+	+	-	-	-	-	-	+	-	-	-
Rathausmarkt 133	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 134	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 135	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 137	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 139	+	+	-	-	-	-	-	+	-	-	-
Rathausmarkt 140	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 142.1	+	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 146	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 155	-	-	-	-	-	-	-	-	-	-	-
Rathausmarkt 156	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 158	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 159	-	-	-	-	-	-	-	-	-	-	-
Rathausmarkt 160	-	+	-	-	-	-	-	+	-	-	-
Rathausmarkt 163.9	-	-	-	-	-	-	-	-	-	-	-
Rathausmarkt 165	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 167	+	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 170	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 171	-	+	-	-	-	-	-	+	-	-	-
Rathausmarkt 172	+	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 173	+	+	-	-	-	-	-	+	-	-	-
Rathausmarkt 178	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 179	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 183	-	-	-	-	-	-	-	-	-	-	-
Rathausmarkt 185	+	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 186	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 191	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 193	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 195	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 197	-	+	-	-	-	-	-	-	-	-	+
Rathausmarkt 2	+	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 20	+	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 200	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 202	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 203	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 205	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 207	-	+	-	-	-	-	-	-	-	-	+
Rathausmarkt 208	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 21	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 214	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 215	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 219	+	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 221	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 23	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 230	+	+	-	-	-	-	-	-	-	-	+
Rathausmarkt 232	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 233	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 28	-	+	-	-	-	-	-	-	-	-	+
Rathausmarkt 42	-	-	-	-	-	-	-	-	-	-	-
Rathausmarkt 49	+	+	-	-	-	-	-	-	-	-	-

Rathausmarkt 52	-	-	-	-	-	-	-	-	-	-	+
Rathausmarkt 57	-	-	-	-	-	-	-	-	-	-	+
Rathausmarkt 58	-	-	-	-	-	-	-	-	-	-	+
Rathausmarkt 60	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 7	+	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 70	-	-	-	-	-	-	-	-	-	-	-
Rathausmarkt 73	-	-	-	-	-	-	-	-	-	-	-
Rathausmarkt 78	+	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 8	+	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 80	+	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 81	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 86	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 88	-	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 9	+	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 90	+	+	-	-	-	-	-	-	-	-	+
Rathausmarkt 91	+	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 92	-	-	-	-	-	-	-	-	-	-	-
Rathausmarkt 94	+	+	-	-	-	-	-	-	-	-	-
Rathausmarkt 95	-	+	-	-	-	-	-	-	-	-	-
SJG 022	+	+	+	+	-	-	-	-	-	-	-
SJG 035	+	+	-	-	-	-	-	-	-	-	-
SJG 1035	+	+	-	-	-	-	-	-	-	-	-
SJG 1087	+	+	-	-	-	-	-	-	-	-	-
SJG 1137	+	+	+	+	-	-	-	-	-	-	-
SJG 1159	+	+	-	-	-	-	-	-	-	-	-
SJG 131	+	+	+	+	-	-	-	-	-	-	-
SJG 149	-	+	+	+	-	-	-	-	-	-	-
SJG 189	+	+	+	+	-	-	-	-	-	-	-
SJG 271	+	+	-	-	-	-	-	-	-	-	-
SJG 289	+	+	+	+	-	-	-	+	-	-	-
SJG 295	+	+	-	-	-	-	-	-	-	-	-
SJG 379	+	+	-	-	-	-	-	-	-	-	-
SJG 404	+	+	+	+	-	-	-	-	-	-	-
SJG 427	+	+	+	+	-	-	-	+	+	-	-
SJG 472	+	+	+	+	-	-	-	-	-	-	-
SJG 507	+	+	+	+	-	-	-	-	-	-	-
SJG 533	+	+	+	+	-	-	-	-	-	-	-
SJG 604	+	+	-	-	-	-	-	-	-	-	-
SJG 611	+	-	+	+	-	-	-	-	-	-	-
SJG 658	+	+	+	+	-	-	-	-	-	-	-
SJG 711	+	+	-	-	-	-	-	-	-	-	-
SJG 718	+	+	-	-	-	-	-	-	-	-	-
SJG 722	+	+	+	+	-	-	-	-	-	-	-
SJG 749	+	+	+	+	-	-	-	-	-	-	-
SJG 792	+	+	-	-	-	-	-	-	-	-	-
SJG 846	+	+	-	-	-	-	-	-	-	-	-
SJG 859	+	+	-	-	-	-	-	+	-	-	-
SJG 872	+	+	-	-	-	-	-	-	-	-	-
SJG 887	+	+	-	-	-	-	-	-	-	-	-
SJG 928	+	+	-	-	-	-	-	-	-	-	-
SJG 947	+	+	-	-	-	-	-	-	-	-	-
SJG 978	+	+	+	+	-	-	-	+	-	-	-
SJG 988	+	+	-	-	-	-	-	-	-	-	-

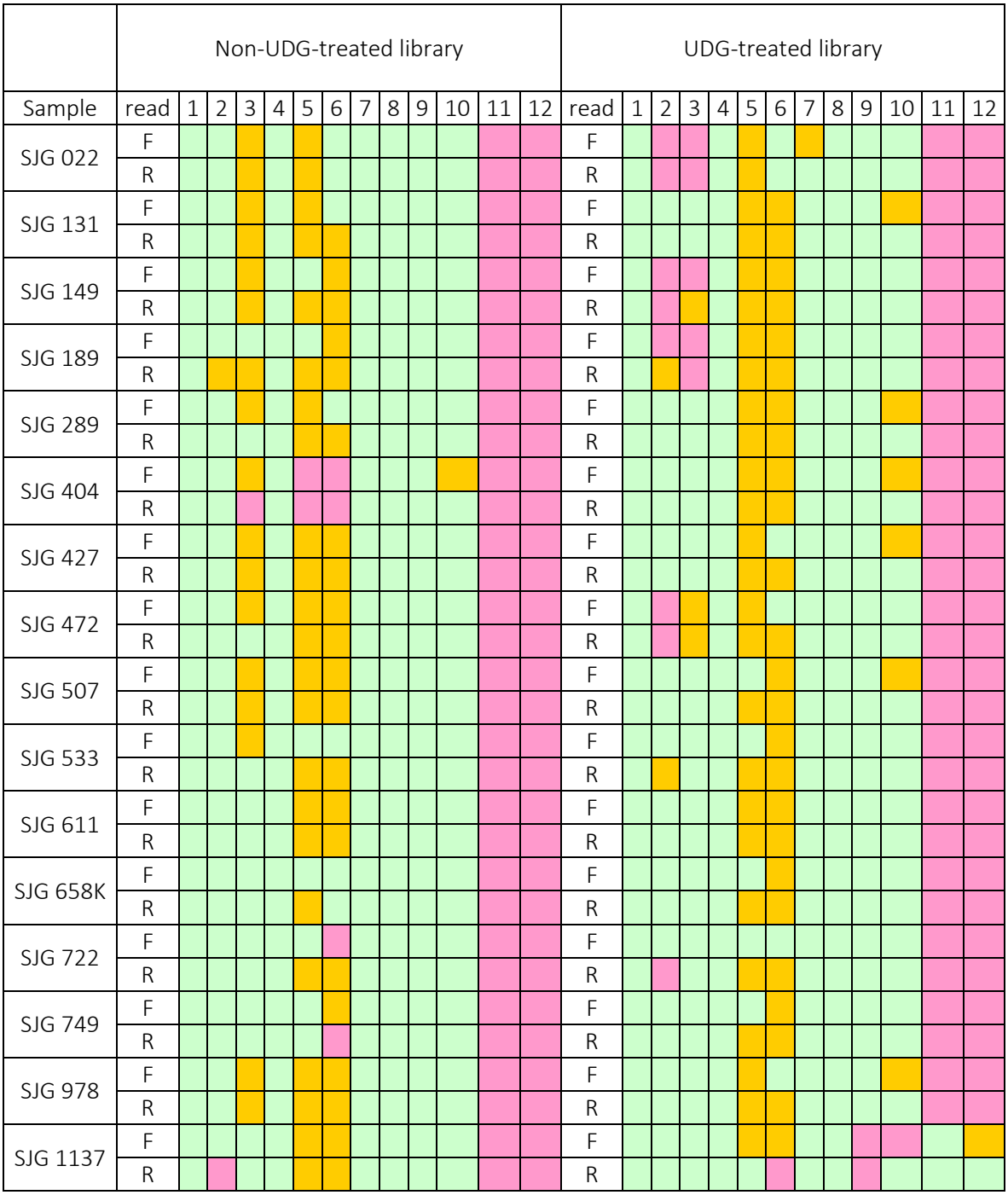
12.3.2 Library quality control results

Supplementary Table 7: Library quantification and quality control results

Sample	Library type	Concentration (ng/µL)	Fragment length (bp)	Peak size (bp)	Adapters chimeras
SJG 022	non UDG	1.91	170-250	231	low
	UDG	18.67	170-250	203	no
SJG 131	non UDG	1.63	150-250	220	no
	UDG	50.70	150-250	205	no
SJG 149	non UDG	39.65	170-250	207	low
	UDG	15.08	170-250	196	no
SJG 189	non UDG	21.0	150-250	205	low
	UDG	11.50	170-250	197	low
SJG 289	non UDG	28.5	170-250	175	low
	UDG	34,96	150-250	198	no
SJG 404	non UDG	7.27	150-250	206	no
	UDG	35.42	170-250	202	low
SJG 427	non UDG	41.5	150-300	203	low
	UDG	28.96	150-250	217	no
SJG 472	non UDG	41.5	170-250	203	low
	UDG	14.66	170-250	206	no
SJG 507	non UDG	26.9	150-300	221	low
	UDG	7.53	170-250	208	low
SJG 533	non UDG	12.0	150-250	224	no
	UDG	20.5	150-250	215	no
SJG 611	non UDG	9.65	170-250	197	no
	UDG	6.06	150-250	198	no
SJG 658	non UDG	2.47	150-250	212	no
	UDG	7.02	170-250	208	no
SJG 722	non UDG	12.8	150-250	192	no
	UDG	11.09	170-250	197	no
SJG 749	non UDG	63.6	150-250	213	low
	UDG	64.0	150-250	210	low
SJG 978	non UDG	29.7	150-250	230	no
	UDG	18.2	150-250	238	no
SJG 1137	non UDG	8.53	150-250	207	no
	UDG	25.27	170-250	217	no

12.3.2 FastQC results

Supplementary Table 8: Summary tables of the FastQC results



Legend:

1- Basic statistics
2- Per base sequence quality
3- Per tile sequence quality
4- Per sequence quality score
5- Per base sequence content
6- Per base GC content
7- Per base N content
8- Sequence length distribution
9- Sequence duplication levels
10- Overrepresented sequences
11- Adapter content
12- K-mer content

Green: Passed
Orange: Warning
Red: Failed

12.3.3 Mapping results for the non-UDG-treated sequencing libraries

Supplementary Table 9: Mapping of the pooled-sequenced samples against the human mtDNA

Sample	Number of raw reads (millions)	Reads mapping (%)	Reference 1X covered (%)	Reference 4X covered (%)
SJG 022	7.1	0.02	98.42	73.09
SJG 131	17.6	0.01	99.56	82.27
SJG 149	8.0	0.01	97.46	71.52
SJG 189	6.4	0.03	69.99	71.52
SJG 289	17.1	0.01	99.79	92.32
SJG 404	21.6	0.01	99.23	80.99
SJG 427	28.9	0.01	99.92	99.19
SJG 472	9.7	0.02	99.44	92.14
SJG 507	19.6	0.01	99.96	99.95
SJG 533	28.4	0.03	99.97	99.94
SJG 611	27.2	0.01	99.78	95.18
SJG 658	38.8	0.01	99.79	92.32
SJG 722	25.0	0.02	99.91	99.79
SJG 749	22.8	0.04	99.99	99.91
SJG 978	22.3	0.01	99.57	88.30
SJG 1137	9.6	0.01	98.32	66.32

Supplementary Table 10: Mapping of the pooled-sequenced samples against the pathogens reference

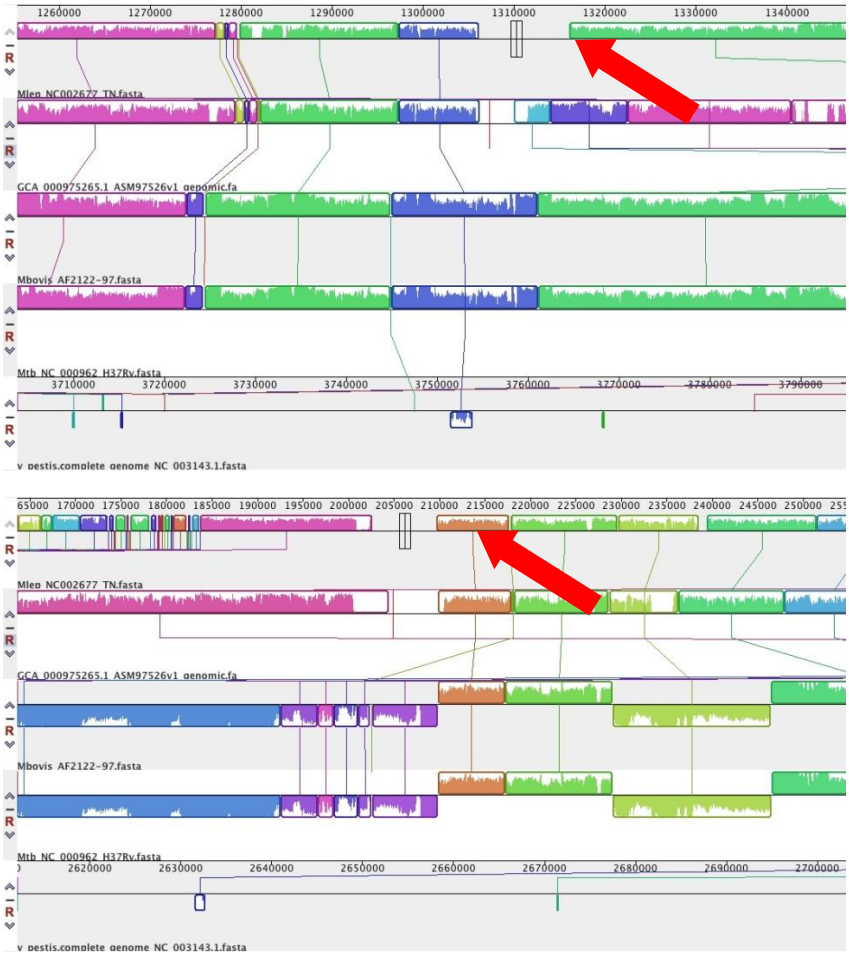
Samples	*M. leprae*			*M. lepromatosis*			*M. bovis*			*M. tuberculosis*			*Y. pestis*		
	Reads mapping (%)	Reference covered (%)		Reads mapping (%)	Reference covered (%)		Reads mapping (%)	Reference covered (%)		Reads mapping (%)	Reference covered (%)		Reads mapping (%)	Reference covered (%)	
		1X	4X		1X	4X		1X	4X		1X	4X		1X	4X
SJG 022	0.10	16.07	0.05	0.02	2.77	0.04	0.01	0.46	0.04	0.01	0.45	0.04	<0.01	0.19	0.04
SJG 131	0.05	15.70	0.09	0.02	3.27	0.10	0.02	1.05	0.11	0.02	1.03	0.11	0.08	0.41	0.13
SJG 149	0.02	2.03	0.05	0.01	0.59	0.05	0.01	0.49	0.05	0.01	0.48	0.05	<0.01	0.23	0.05
SJG 189	0.22	28.50	0.08	0.05	5.07	0.03	0.01	0.44	0.03	0.01	0.44	0.03	<0.01	0.28	0.03
SJG 289	0.03	6.08	0.13	0.02	1.70	0.13	0.03	1.31	0.18	0.03	1.30	0.17	0.01	0.34	0.11
SJG 404	0.26	73.81	5.01	0.07	16.27	0.47	0.03	1.92	0.25	0.03	1.89	0.25	0.01	0.34	0.13
SJG 427	0.64	98.63	65.40	0.15	32.48	7.07	0.04	4.17	0.40	0.04	4.13	0.39	<0.01	0.34	0.13
SJG 472	0.14	25.73	0.14	0.04	4.97	0.09	0.03	1.01	0.11	0.03	0.99	0.11	<0.01	0.21	0.05
SJG 507	12.63	99.99	98.99	2.41	54.72	41.60	0.11	5.80	1.69	0.11	5.74	1.67	<0.01	0.27	0.06
SJG 533	1.44	99.97	98.48	0.27	39.62	17.76	0.02	2.99	0.33	0.02	2.95	0.32	<0.01	0.34	0.16
SJG 611	0.03	11.36	0.13	0.01	2.71	0.13	0.02	1.30	0.17	0.02	1.29	0.16	<0.01	0.51	0.20
SJG 658	0.04	20.97	0.19	0.02	5.07	0.16	0.02	2.03	0.22	0.02	1.98	0.21	<0.01	0.44	0.18
SJG 722	0.33	84.24	13.36	0.08	20.04	1.01	0.04	2.34	0.32	0.04	2.30	0.32	0.01	0.41	0.15
SJG 749	7.16	99.99	99.99	1.34	52.26	36.66	0.07	5.25	1.16	0.07	5.19	1.15	0.01	0.43	0.17
SJG 978	0.11	35.45	0.41	0.04	3.76	0.11	0.05	3.60	0.47	0.05	3.55	0.46	0.01	0.57	0.20
SJG 1137	0.10	15.99	0.12	0.04	3.76	0.11	0.04	1.22	0.15	0.04	1.20	0.15	0.01	0.30	0.10

Supplementary Table 11: Mapping of the Non-UDG-treated libraries pooled datasets, for the human mtDNA and the M. leprae *DNA*

Samples	Total number of reads	*human mtDNA*				*M. leprae*			
		# Reads mapping	Read length (bp)	Reference covered		# Reads mapping	Read length (bp)	Reference covered	
				1X	4X			1X	4X
SJG 022	48964744	2654	88 (67-105)	0.997	0.986	33147	91 (66-103)	0.584	0.018
SJG 131	39151942	1331	87 (64-105)	0.993	0.889	10327	89 (63-104)	0.228	0.001
SJG 149	44446430	6217	77 (58-101)	0.998	0.994	44562	86 (63-103)	0.674	0.043
SJG 189	38183618	5630	76 (58-100)	0.998	0.993	41930	85 (62-101)	0.651.	0.0340.
SJG 289	41799412	1816	101 (77-116)	0.999	0.969	6205	75 (56-101)	0.119	0.001
SJG 404	9890194	1	76 (76-76)	0.005	0.000	128	59 (50-70)	0.003	0.000
SJG 427	39352640	1676	97 (72-115)	0.998	0.955	166792	92 (66-107)	0.987	0.666
SJG 472	63092774	14544	82 (58-101)	1.000	1.000	3964844	101 (73-110)	1.000	1.000
SJG 507	59559570	4964	81 (64-101)	1.000	0.998	56997	84 (61-101)	0.731	0.080
SJG 533	38418000	6831	85 (63-103)	1.000	0.997	345056	84 (61-101)	1.000	0.978
SJG 611	49194026	1529	75 (56-99)	0.9915	0.8874	3916	71 (51-101)	0.0749	0.0005
SJG 658	44552648	2751	85 (64-105)	0.998	0.984	5361	77 (54-101)	0.111	0.001
SJG 722	49835688	3449	88 (65-104)	1.000	0.996	92459	100 (70-107)	0.914	0.277
SJG 749	28980040	4195	87 (61-104)	0.999	0.991	1007860	100 (71-113)	1.000	1.000
SJG 978	30266582	1079	100 (76-114)	0.981	0.824	19311	94 (66-107)	0.390	0.003
SJG 1137	49073730	2448	84 (63-101)	0.998	0.975	14269	90 (63-102)	0.295	0.002

13.3.4 Manual review of specificity of the genomic targets

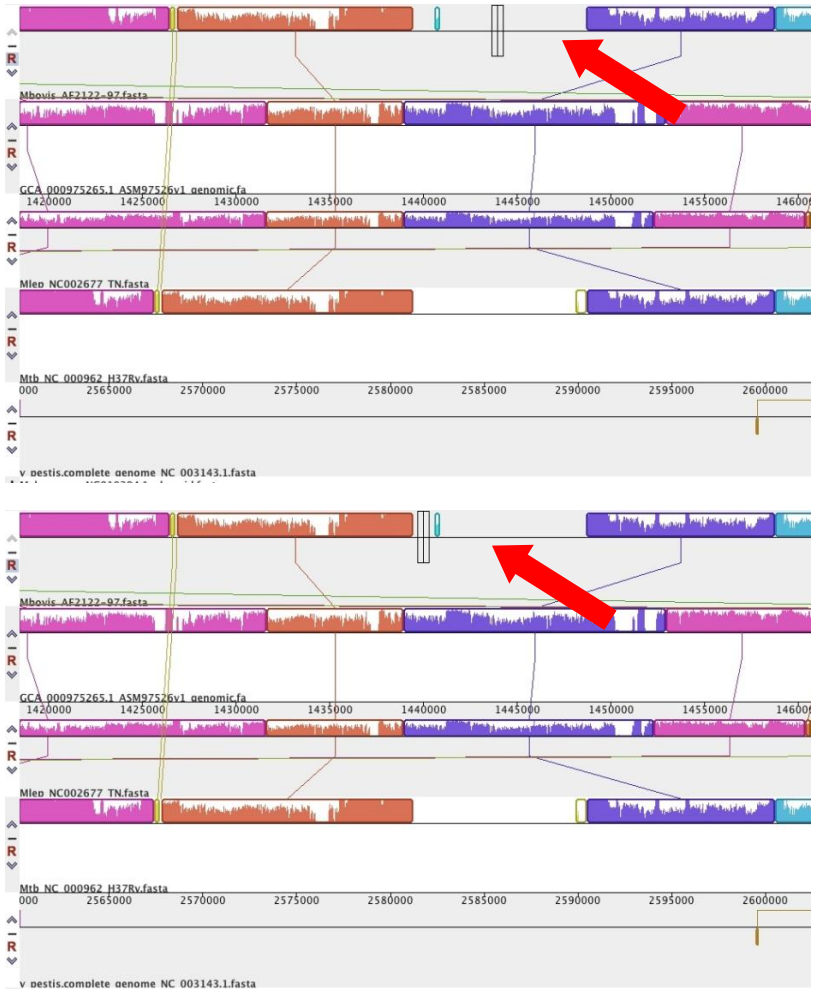
The figures below gather the MAUVE GUI views of each of the regions chosen as species-specific targets within the multiple alignment. For visibility, only *M. leprae*, *M lepromatosis, M. tuberculosis, M. bovis* and *Y. pestis* lines are shown.



Supplementary Figure 2: Genomic synteny in the target regions for M. leprae
Up: target 1, Down: Target 2

Supplementary Figure 3: Genomic synteny in the target regions for M. lepromatosis
Up: target 1, Down: Target 2

Supplementary Figure 4: Genomic synteny in the target regions for M. tuberculosis
Up: target 1, Down: Target 2



Supplementary Figure 5: Genomic synteny in the target regions for M. bovis
Up: target 1, Down: Target 2

12.3.5 Mapping statistics on the species-specific target regions

Supplementary Table 12: Y. pestis targets coverage statistics

Samples	Read Depth			
	Whole reference		*Targets*	
	All	Covered bases	All	Covered bases
SJG 022	0.0020	1.8256	0.0004	1.3542
SJG 131	0.0091	3.7127	0.0073	3.7134
SJG 149	0.0030	2.3890	0.0022	2.4689
SJG 189	0.0026	1.6058	0.0016	1.3843
SJG 289	0.0064	3.4016	0.0058	3.7840
SJG 404	0.0091	4.7609	0.0080	5.3561
SJG 427	0.0097	5.2885	0.0066	5.1432
SJG 472	0.0031	2.6223	0.0029	2.9697
SJG 507	0.0036	2.4256	0.0023	2.3705
SJG 533	0.0070	4.2372	0.0051	4.5644
SJG 611	0.0152	5.9490	0.0119	5.9263
SJG 658	0.0159	6.7292	0.0114	7.0229
SJG 722	0.0106	5.1263	0.0072	4.7481
SJG 749	0.0127	5.8911	0.0101	6.6945
SJG 978	0.0168	5.9223	0.0124	5.6450
SJG 1137	0.0058	3.1993	0.0046	2.9189

Statistics computed from the UDG-treated sequencing reads obtained during the 1st sequencing phase (1/8th of lane per sample)

Supplementary Table 13: M. tuberculosis targets coverage statistics

Samples	*M. leprae*		*M. tuberculosis*			
	Whole reference		Whole reference		Targets	
	1X	4X	1X	4X	1X	4X
SJG 022	0.3190	0.0086	0.0016	0.0003	0	0
SJG 131	0.2104	0.0022	0.0027	0.0006	0	0
SJG 149	0.0616	0.0070	0.0021	0.0007	0	0
SJG 189	0.6970	0.2055	0.0028	0.0007	0	0
SJG 289	0.0765	0.0008	0.0034	0.0007	0.0013	0.0014
SJG 404	0.9999	0.9979	0.0167	0.0045	0.0014	0.0004
SJG 427	1.0000	1.0000	0.0623	0.0478	0.0090	0.0017
SJG 472	0.9966	0.9060	0.0150	0.0048	0.0014	0.0013
SJG 507	1.0000	1.0000	0.0330	0.0169	0.0020	0.0019
SJG 533	0.9999	0.9998	0.0082	0.0022	0.0018	0.0010
SJG 611	0.1102	0.0007	0.0026	0.0005	0	0
SJG 658	0.5462	0.1894	0.0187	0.0053	0.0014	0.0013
SJG 722	1.0000	0.9999	0.0218	0.0080	0.0016	0.0014
SJG 749	1.0000	1.0000	0.0145	0.0056	0.0016	0.0013
SJG 978	0.9867	0.8081	0.0376	0.0246	0.0015	0.0014
SJG 1137	0.8732	0.3175	0.0127	0.0042	0.0016	0.0014

Statistics computed from the pooled datasets of all the available UDG-treated sequencing reads

12.3.6 Variant calling and annotation

Supplementary Table 14: List of all the annoted variants

Position	Referee (TN)	G022	G131	G149	G189	G289	G404	G427	G472	G507	G533	G611	G658	G722	G749	G978	G1137	SNP effect	Impact	Gene ID	Gene Name (if not ID)	Gene product	Futional category (Mycobrowser)	AA change	Prot. lenght (bp)
73	A						G	G	G	G	G			G	A		G	miss.	Moder.	ML0001	dnaA	chromosomal replication initiator	Inf. pathways	Ser25Gly	521
883	G						A	A	A	G	A			A	G	G		miss.	Moder.	ML0001	dnaA	chromosomal replication initiator	Inf. pathways	Gly295Ser	521
4180	G						G	A		G	G			G	G			syn.	Low	ML0002	dnaN	DNA polymerase III subunit DnaN	Inf. pathways	Ser300Ser	385
7614	C						T	T	T	C	T			T	C			syn.	Low	ML0006	gyrA	DNA gyrase subunit A	Inf. pathways	Arg99Arg	1249
7870	G						G	G		G	A			G	G			miss.	Moder.	ML0006	gyrA	DNA gyrase subunit A	Inf. pathways	Glu185Lys	1249
8453	T						C	C	C	C	C			C	C	C		miss.	Moder.	ML0006	gyrA	DNA gyrase subunit A	Inf. pathways	Leu379Pro	1249
12220	A						A	A	A	A	A			A	A	a		syn.	Low	MLP000001	ileT	tRNA-Ile anticodon GAT	Stable RNAs	Ala9Ala	24
13907	C						C	C		C	C			C	T	T		miss.	Moder.	ML0009		hypothetical protein	Unknown	Pro45Ser	63
14222	C									t								interg.	Modif.						
14226	C							t		T					c			interg.	Modif.						
14610	T						t	T		T	T			T	T			interg.	Modif.						
17157	G						G	G		A	G			G	G	G		syn.	Low	ML0013		transmembrane protein	Cell wall and cell processes	Leu87Leu	93
17669	C							C		T	C			C	C			intrag.	Modif.	ML0014		pseudogene	Pseudogene		
17928	G							a		G					G			intrag.	Modif.	ML0014		pseudogene	Pseudogene		
19071	G						a	A	A	G	G			A	G			downst. 48b.	Modif.	ML0015	trpG	anthranilate synthase component II	Metabolism and respiration		233

19071	G						a	A	A	G	G			A	G			miss.	Moder.	ML0016	pknB	serine/threonine protein kinase B	Reg. proteins	Ala600Val	622
28938	C							c		C	C			C	C	t		downst. 16b.	Modif.	ML0025		hypothetical protein	Unknown		97
28938	C							c		C	C			C	C	t		miss.	Moder.	ML0024		hypothetical protein	Unknown	Leu17Phe	84
44832	G						a	A	A	G	G			A	G			intrag.	Modif.	ML0037	eccD2	pseudogene	Pseudogene		
45407	C							C		T	C			C	C			intrag.	Modif.	ML0038	eccE2	pseudogene	Pseudogene		
48196	G						a	A	A	G	G			A	G		A	intrag.	Modif.	ML0039	eccA2	pseudogene	Pseudogene		
52084	C						C	C		C	C			C	T			miss.	Moder.	ML0042	eccE1	hypothetical protein	Unknown	Thr31Ile	467
52084	C						C	C		C	C			C	T			downst. 88b.	Modif.	ML0041	mycP1	secreted protease	Metabolism and respiration		447
52851	T						C	C	C	C	C			C	C		C	miss.	Moder.	ML0042	eccE1	hypothetical protein	Unknown	Trp287Arg	467
61425	A						G	G	G	G	G		G	G	G	G	G	miss.	Moder.	ML0049	esxA	6 kDa early secretory antigenic target	Cell wall and cell processes	Met90Thr	95
61530	C						T	T	T	T	T			T	T	T		miss.	Moder.	ML0049	esxA	6 kDa early secretory antigenic target	Cell wall and cell processes	Arg55Gln	95
71687	G							G		A					G			intrag.	Modif.	ML0057	espG1	pseudogene	Pseudogene		
71721	G									g					t			intrag.	Modif.	ML0057	espG1	pseudogene	Pseudogene		
71722	T									t								intrag.	Modif.	ML0057	espG1	pseudogene	Pseudogene		
71739	G									A					G			intrag.	Modif.	ML0057	espG1	pseudogene	Pseudogene		
73090	A									a					A			intrag.	Modif.	ML0058	espE	pseudogene	Pseudogene		
73094	T							t		t	t				t			intrag.	Modif.	ML0058	espE	pseudogene	Pseudogene		
77864	T						C	C	C	C	C		C	C	C	C		syn.	Low	ML0061	gltB	ferredoxin-dependant glutamate synthase	Metabolism and respiration	Gly52Gly	1527
85630	T						T	T		T	T			T	T	C		intrag.	Modif.	ML0063		pseudogene	Pseudogene		
86851	C						C	C		C	C			T	C			syn.	Low	ML0065	ethA	monooxygenase	Metabolism and respiration	Gly57Gly	494
94314	G						G	G		G	G			G	A			miss.	Moder.	ML0073		hypothetical protein	Unknown	Gly75Ser	344

100574	A						G	G	G	G	G			G	A			downst. 74b.	Modif.	ML0081		hypothetical protein	Unknown		
100574	A						G	G	G	G	G			G	A			intrag.	Modif.	ML0080		pseudogene	Pseudogene		
116225	C						T	T	T	C	T		T	T	C	C		syn.	Low	ML0093	ML0093	hypothetical protein	Unknown	Gly285Gly	643
122485	C						C	C		C	G			C	C			miss.	Moder.	ML0098	fbpD	secreted MPB51/MPT51 antigen	Lipid metabolism	Ala59Gly	301
156556	G						G	G	G	A	G			G	G			syn.	Low	ML0116		hypothetical protein	Unknown	Ile458Ile	654
156948	C				T		T	T	T	C	T	T		T	C			miss.	Moder.	ML0118		oxidoreductase	Metabolism and respiration	Val328Ile	654
165152	C							C	c	C	C			C	C	C		syn.	Low	MLP00006	argU	tRNA-Arg	Stable RNAs	Pro2Pro	23
165152	C							C	c	C	C			C	C	C		downst.25 b.	Modif.	MLP00005	serT	tRNA-Ser	Stable RNAs		23
165195	A						g	a	a	A	A		g	a	A	a	g	miss.	Moder.	MLP00006	argU	tRNA-Arg	Stable RNAs	Arg17Gly	23
165195	A						g	a	a	A	A		g	a	A	a	g	downst. 68b.	Modif.	MLP00005	serT	tRNA-Ser	Stable RNAs		23
168079	A						A	A		G	A			A	A			miss.	Moder.	ML0125		glycosyl transferase	Metabolism and respiration	Tyr161Cys	438
193848	C				T		T	T	T	C	T		T	T	C			upst. 96b.	Modif.	ML0138	fadD28	acyl-CoA synthetase	Lipid metabolism		580
193848	C				T		T	T	T	C	T		T	T	C			interg.	Modif.						
202440	A						A	A		a	A			A	A			miss.	Moder.	ML0141		hypothetical protein	Unknown	Leu41Pro	84
202441	G						G	G		g	G			G	G			miss.	Moder.	ML0141		hypothetical protein	Unknown	Leu41Phe	84
202443	G						G	G		g	G			G	G			miss.	Moder.	ML0141		hypothetical protein	Unknown	Ser40Phe	84
227077	A						g	G	G	A	A			G	A			syn.	Low	ML0162		hypothetical protein	Unknown	His70His	92
233893	G						G	G		G	A			G	G			intrag.	Modif.	ML0167		pseudogene	Pseudogene		
243860	G						G	G	T	G	G			G	G			syn.	Low	ML0176	pepD	serine protease	Metabolism and respiration	Gly269Gly	382
250837	C						C	C		T	C			C	C			miss.	Moder.	ML0185		hypothetical protein	Unknown	Thr206Ile	342
253623	C						C	C		C	C			C	T	T		syn.	Low	ML0187		hypothetical protein	Unknown	Val322Val	353

261744	C						C	C		C	C			C	C	T		intrag.	Modif.	ML0195	ML0195	pseudogene	Pseudogene		
264501	A						C	C	C	C	C			C	C	C		intrag.	Modif.	ML0197		pseudogene	Pseudogene		
298572	A				G		G	G	G	G	G			G	G	G		syn.	Low	ML0227	ML0227	hypothetical protein	Unknown	Val19Val	158
298572	A				G		G	G	G	G	G			G	G	G		downst. 56b.	Modif.	ML0226	folK	pyrophosphokinase	Metabolism and respiration		192
305585	G							G	G	G	A			G	G	G		miss.	Moder.	ML0234	lsr2	lsr2 protein precursor	Cell wall and cell processes	Gly73Glu	112
305654	C						C	C	C	C	C			C	T	C		miss.	Moder.	ML0234	lsr2	lsr2 protein precursor	Cell wall and cell processes	Thr96Ile	112
313361	A				G		G	G	G	G	G			G	G	G		miss.	Moder.	ML0238	metS	methionyl-tRNA synthetase	Inf. pathways	Leu154Pro	537
326140	T						T	T		T	T			T	T	t		miss.	Moder.	MLP000009	glnT	tRNA-Gln anticodon TTG	Stable RNAs	Thr22Ala	23
328604	G						a	A	A	G	G			A	G	G		miss.	Moder.	ML0252	mfd	transcription-repair coupling factor	Inf. pathways	Asp180Asn	1224
330125	G						A	A	A	A	A			A	G	G	A	miss.	Moder.	ML0252	mfd	transcription-repair coupling factor	Inf. pathways	Ala687Thr	1224
330478	G						G	G		G	G			G	G	g		syn.	Low	ML0252	mfd	transcription-repair coupling factor	Inf. pathways	Leu804Leu	1224
330912	A						A	A		G	A			A	A			miss.	Moder.	ML0252	mfd	transcription-repair coupling factor	Inf. pathways	His949Arg	1224
336266	T						T	T		C	T			T	T			syn.	Low	ML0257		hypothetical protein	Unknown	Ala154Ala	167
336266	T						T	T		C	T			T	T			upst. 39b.	Modif.	ML0258		hypothetical protein	Unknown		
347385	T						C	C	C	C	C			C	C	C		intrag.	Modif.	ML0266		pseudogene	Pseudogene		
347385	T						C	C	C	C	C			C	C	C		intrag.	Modif.	ML0266		pseudogene	Pseudogene		
349747	G							G		A	G			G	G			interg.	Modif.						
365411	A						G	G	G	A	G			G	A			miss.	Moder.	ML0283		cation-efflux transporter component	Cell wall and cell processes	Lys166Glu	268
365540	G						G	G		T	G			G	G			miss.	Moder.	ML0283		cation-efflux transporter component	Cell wall and cell processes	Val209Leu	268
373059	C						C	C		C	C			C	T			interg.	Modif.						
383599	C						G	G	G	G	G		G	G	G	G		intrag.	Modif.	ML0301	mutT3	pseudogene	Pseudogene		

383599	C						G	G	G	G	G		G	G	G	G	G		downst.: 57b.	Modif.	ML0300	thiE	thiamine-phosphate pyrophosphorylase	Metabolism and respiration		708
384392	C									c									intrag.	Modif.	ML0302		pseudogene	Pseudogene		
384393	G									c									intrag.	Modif.	ML0302		pseudogene	Pseudogene		
393043	G						G	G		A	G			G	G				intrag.	Modif.	ML0306		pseudogene	Pseudogene		
396059	G							G	G	T	G			G	G				intrag.	Modif.	ML0309		pseudogene	Pseudogene		
405673	C						C	C		C	C			C	C	c			syn.	Low	ML0317	groEL2	65 kDa chaperonin 2	Virulee, detoxification, adaptation	Gly52Gly	542
408063	A						A	A		a	A			A	A				interg.	Modif.						
408068	A						A	A		c	A			A	A				interg.	Modif.						
408072	C						C	C		A	C			C	C				interg.	Modif.						
416972	C						T	T	T	C	C			T	C		T		intrag.	Modif.	ML0327	esxS	pseudogene	Pseudogene		
417340	C						C	C	T	C	C			C	C				intrag.	Modif.	ML0328	PPE	pseudogene	Pseudogene		
436217	T						T	T		C	T			T	T				intrag.	Modif.	ML0354	ilvX	acetohydroxyacid synthase	Metabolism and respiration		516
451498	C						C	C		T	C			C	C				intrag.	Modif.	ML0357	fadE26	pseudogene	Pseudogene		
459545	G						G	G	G	G	G			G	T				intrag.	Modif.	ML0367		pseudogene	Pseudogene		
474071	G			C			G	G		G	G			G	G	G			miss.	Moder.	ML0381	groEL1	60 kDa chaperonin 1	Virulee, detoxification, adaptation	Val172Ile	537
481476	A	G					G	G	G	G	G	G		G	G	G	G		syn.	Low	ML0387	guaB2	inosine 5'-monophosphate dehydrogenase	Metabolism and respiration	Leu272Leu	529
485138	T						C	C	C	C	C		C	C	C	C	C		miss.	Moder.	ML0389	choD	cholesterol oxidase precursor	Lipid metabolism	Leu563Pro	569
494674	T	G			G		G	G	G	G	G	G		G	G	G	G		miss.	Moder.	ML0397		hypothetical protein	Unknown	Glu358Asp	602
508481	T						C	C	C	C	C		C	C	C	C	C		miss.	Moder.	ML0410	PE3	PE-family protein	PE/PPE	Val52Ala	408
509065	G						G	G		A	G			G	G				miss.	Moder.	ML0411		serine-rich antigen protein	PE/PPE	Gly104Asp	408
509178	G						G	G		A	G			G	G				miss.	Moder.	ML0411		serine-rich antigen protein	PE/PPE	Asp142Asn	408

509325	C						G	G	G	G	G			G	G	G		miss.	Moder.	ML0411		serine-rich antigen protein	PE/PPE	.His191Asp	408
509662	G						G	G		G	G			G	G	A		miss.	Moder.	ML0411		serine-rich antigen protein	PE/PPE	Gly303Asp	408
509673	G						G	G		G	G			G	A			miss.	Moder.	ML0411		serine-rich antigen protein	PE/PPE	Val307Met	408
509674	T						T	T		C	T			T	T			miss.	Moder.	ML0411		serine-rich antigen protein	PE/PPE	Val307Ala	408
509851	G						G	G	G	T	G			G	G			miss.	Moder.	ML0411		serine-rich antigen protein	PE/PPE	Gly366Val	408
523366	G									a								interg.	Modif.						
523810	C						C	C		C	C			C	T	T		miss.	Moder.	ML0424	bcp	bacterioferritin comigratory antioxidant	Virulee, detoxification, adaptation	Cys160Tyr	161
533403	A				G		G	G	G	G	G			G	G	G		intrag.	Modif.	ML0433	fadD35	pseudogene	Pseudogene		
537503	G						G	G		G	G			G	T			intrag.	Modif.	ML0437	accA1	pseudogene	Pseudogene		
558254	C							C		C	A			C	C			intrag.	Modif.	ML0468		hypothetical protein	Cell wall and cell processes		261
563796	G						G	G		C	G			G	G			miss.	Moder.	ML0466		hypothetical protein	Unknown	Glu55Asp	301
564029	A						A	a		A	A			A	A			miss.	Moder.	ML0466		hypothetical protein	Unknown	Asp133Gly	301
578472	C									C					A			intrag.	Modif.	ML0476	speE	pseudogene	Pseudogene		
578476	A							A		A					C			intrag.	Modif.	ML0476	speE	pseudogene	Pseudogene		
601404	C						C	C		C	C			C	T			intrag.	Modif.	ML0476	speE	pseudogene	Pseudogene		
603386	A							A		G	A			A	A			intrag.	Modif.	ML0476	speE	pseudogene	Pseudogene		
604325	T				C		C	C	C	T	C			C	T		C	intrag.	Modif.	ML0498		pseudogene	Pseudogene		
614606	A						A	A	A	a	A			A	A			interg.	Modif.						
614607	A						A	A	A	a	A			A	A			interg.	Modif.						
614610	G						G	G	G	g	G			G	G			interg.	Modif.						
637185	C						C	C		T	C			C	C	C		miss.	Moder.	ML0254	adi	amino acid decarboxylase	Metabolism and respiration	Thr848Ile	950

637535	G						G	G		T	G			G	G			miss.	Moder.	ML0525		hypothetical protein	Unknown	Arg15Leu	58
637535	G						G	G		T	G			G	G			downst. 40b.	Modif.	ML0254	adi	amino acid decarboxylase	Metabolism and respiration		950
648620	C						C	C		C	C			C	C	g		miss.	Moder.	ML0535	carA	carbamoyl-phosphate synthase small chain	Metabolism and respiration	Gln67Glu	376
656930	G						G	G		G	G			G	G	g		syn.	Low	ML0540	mihF	integration host factor MihF	Inf. pathways	Ala12Ala	106
657736	G						C	C	C	C	C			C	C	C		syn.	Low	ML0541	gmk	guanylate kinase	Metabolism and respiration	Arg157Arg	210
687105	T							T		t				T	T			miss.	Moder.	ML0568		hypothetical protein	Unknown	Leu98Pro	197
687132	C							C		t				C	C			miss.	Moder.	ML0568		hypothetical protein	Unknown	Pro107Leu	197
692487	C						C	C	c	C	C			C	C			interg.	Modif.						
694090	T						C	C	C	C	C		C	C	C	C		miss.	Moder.	ML0569		hypothetical protein	Unknown	Thr113Ala	271
711197	T						T	T		C	T			T	T			intrag.	Modif.	ML0585	qor	pseudogene	Pseudogene		
736584	G							g		G	G				G			miss.	Moder.	ML0605		hypothetical protein	Unknown	Glu12Asp	561
736585	A							a		A	A				A			miss.	Moder.	ML0605		hypothetical protein	Unknown	Thr13Pro	561
749178	C							C		T				C	C			miss.	Moder.	ML0615	subI	sulphate-binding lipoprotein	Cell wall and cell processes	Thr266Ile	348
750910	G						a	G		G	G			G	G			intrag.	Modif.	ML0617	cysW	pseudogene	Pseudogene		
751982	T						T	T	T	G	T			T	T			intrag.	Modif.	ML0619		pseudogene	Pseudogene		
763572	G						G	G		A	G			G	G			miss.	Moder.	ML0631	era	GTP-binding protein	Metabolism and respiration	Val33Ile	302
780280	T						T	T		T	T			t	T	t		upst. 4 b.	Modif.	MLP000013	metU	tRNA-Met	Stable RNAs		24
780280	T						T	T		T	T			t	T	t		interg.	Modif.						
780299	C						C	C	C	C	C		c	C	C	C		miss.	Moder.	MLP000013	metU	tRNA-Met	Stable RNAs	Leu6Phe	24
780341	G		A				g	g	g	G	G		a	G	G	a	G	miss.	Moder.	MLP000013	metU	tRNA-Met	Stable RNAs	Glu20Lys	24
805290	C						C	C		T	C			C	C	C		intrag.	Modif.	ML0668		pseudogene	Pseudogene		

816835	C						C	C		C	C			C	T			intrag.	Modif.	ML0681		pseudogene	Pseudogene		
832152	T						C	C	C	C	C			C	C	C		intrag.	Modif.	ML0694		pseudogene	Pseudogene		
840584	A						G	G	G	G	G			G	G		G	intrag.	Modif.	ML0702		pseudogene	Pseudogene		
840974	C						g	C		C	C			C				miss.	Moder.	ML0703		hypothetical protein	Unknown	Ala394Pro	423
870381	G						G	G	G	A	G			G	G	G		miss.	Moder.	ML0728	sseA	thiosulfate sulfurtransferase	Metabolism and respiration	Leu180Phe	296
889694	G			T			G	G	G	G	G			G	G	G		miss.	Moder.	ML0747	ctpC	metal cation-transporting P-type ATPase C	Cell wall and cell processes	Pro426His	725
890453	A						G	G	G	G	G			G	G	G	G	miss.	Moder.	ML0747	ctpC	metal cation-transporting P-type ATPase C	Cell wall and cell processes	Val173Ala	725
904824	G						C	C	C	C	C			C	C	C	C	syn.	Low	ML0763	pmmA	phosphomannomutase	Metabolism and respiration	Val50Val	468
906512	T						T	T		T	T			C	T			syn.	Low	ML0764		hypothetical protein	Unknown	Arg145Arg	365
919482	G						G	G	G	C	G			G	G			miss.	Moder.	ML0775	lpqB	lipoprotein	Cell wall and cell processes	Gly372Ala	589
921523	C							C		C	C			C	C	T		miss.	Moder.	ML0778		hypothetical protein	Inf. pathways	Leu191Phe	230
924882	G						G	G		T	G			G	G			intrag.	Modif.	ML0780		pseudogene	Pseudogene		
924883	C						C	C		c	C			C	C			intrag.	Modif.	ML0780		pseudogene	Pseudogene		
926861	G						G	G		G	A			G	G			miss.	Moder.	ML0782		transmembrane transport protein	Cell wall and cell processes	Val232Met	385
938787	T						g	T		T	T			T	T			intrag.	Modif.	ML0794		pseudogene	Pseudogene		
944190	A									a								intrag.	Modif.	ML0797	PPE	pseudogene	Pseudogene		
944191	C									a								intrag.	Modif.	ML0797	PPE	pseudogene	Pseudogene		
944193	G									g								intrag.	Modif.	ML0797	PPE	pseudogene	Pseudogene		
954037	C							C		G	C			C	C			interg.	Modif.						
958228	A							a		a					A			intrag.	Modif.	ML0809		pseudogene	Pseudogene		
958231	T									t					T			intrag.	Modif.	ML0809		pseudogene	Pseudogene		

963339	C							C		T	C			C	C	C		upst. 45b.	Modif.	ML0813		hypothetical protein	Unknown		
963339	C							C		T	C			C	C	C		interg.	Modif.						
966690	G							G		C	G			G	G	G		miss.	Moder.	ML0817	moeB1	molybdopterin cofactor biosynthesis protein	Metabolism and respiration	Arg254Thr	395
972005	T				G		G	G	G	G	G			G	G		G	intrag.	Modif.	ML0821		pseudogene	Pseudogene		
981132	C						C	C		T	C			C	C			interg.	Modif.						
997947	C						C	C	C	C	C			C	C	T		miss.	Moder.	ML0840		hypothetical protein	Unknown	Val215Ile	434
1017407	G						A	A	A	G	A		A	A	G			miss.	Moder.	ML0855	pepR	zi protease	Metabolism and respiration	Glu387Lys	424
1044447	C						C	C		C	C			C	C	A		miss.	Moder.	ML0879	qcrB	cytochrome C reductase cytochrome B subunit	Metabolism and respiration	Ala122Ser	551
1044678	A						A	A	a	A	A			A	A	A		syn.	Low	ML0879	qcrB	-cytochrome C reductase cytochrome B subunit	Metabolism and respiration	Leu45Leu	551
1056594	A						A	A		G	A			A	A			syn.	Low	ML0890		hypothetical protein	Unknown	Ser118Ser	415
1067418	C						C	C		T	C			C	C			interg.	Modif.						
1073641	C						C	C	C	T	C			C	C			syn.	Low	ML0907		membrane protein	Cell wall and cell processes	Leu314Leu	377
1081267	A						g	A		A	A			A	A			syn.	Low	ML0912	murD	UDP-N-actylmuramoyl-L-alanyl-D-glutamate ligase	Cell wall and cell processes	Val349Val	490
1087397	T						C	C		C	C			C	C	C	C	syn.	Low	ML0917	ftsZ	cell division protein FtsZ	Cell wall and cell processes	Gly34Gly	379
1089560	T						T	T	T	C	T			T	T			syn.	Low	ML0919		hypothetical protein	Unknown	Thr119Thr	260
1092198	A							A		A	A			A	G			downstr. 50b.	Modif.	ML0922	wag31	hypothetical protein	Cell wall and cell processes		267
1092198	A							A		A	A			A	G			upst. 23b.	Modif.	ML0923		transmembrane protein	Cell wall and cell processes		131
1092198	A							A		A	A			A	G			interg.	Modif.						
1093114	C							C		G					C			interg.	Modif.						
1093117	G							G		T					G			interg.	Modif.						
1104232	C						G	G	G	G	G			G	C			intrag.	Modif.	ML0934		pseudogene	Pseudogene		

1126158	G						G	G		G	G			G	g			miss.	Moder.	ML0957		hypothetical protein	Unknown	Pro24Leu	124
1126158	G						G	G		G	G			G	g			upst. 81b.	Modif.	ML0958		pseudogene	Pseudogene		
1127192	A							A		A	A			A	A	C		miss.	Moder.	ML0959		hypothetical protein	Unknown	Asp94Ala	117
1133492	T						T	T		G	T			T	T			interg.	Modif.						
1141251	C						C	c		C	C			C	C			intrag.	Modif.	ML0973		pseudogene	Pseudogene		
1143423	T						C	C	C	C	C		C	C	C	C	C	intrag.	Modif.	ML0975		pseudogene	Pseudogene		
1143812	G						a	A	A	G	G			A	G		A	intrag.	Modif.	ML0975		pseudogene	Pseudogene		
1147526	G							G		A	G			G	G	G		syn.	Low	ML0978		acetyltransferase	Metabolism and respiration	His61His	180
1149607	T						T	T		C	T			T	T			intrag.	Modif.	ML0981	35kd_ag	pseudogene	Pseudogene		
1155582	T						G	G	G	G	G			G	G			syn.	Low	ML0988	recX	recombination regulator	Reg. proteins	Ala142Ala	171
1160426	C						g	G	G	C	C			G	C		G	interg.	Modif.						
1161063	A						A	A		G	A			A	A			intrag.	Modif.	ML0993		pseudogene	Pseudogene		
1162797	G						g	T		G	G			T	G		T	miss.	Moder.	ML0994		hypothetical protein	Unknown	Gly60Cys	232
1179260	G						G	A		G	G			G	G		A	intrag.	Modif.	ML1008		pseudogene	Pseudogene		
1191821	G						a	A		G	G			A	G			miss.	Moder.	ML1023	ppgK	polyphosphate glucokinase	Metabolism and respiration	Ala25Val	324
1205320	C						c	C		C	C			C	C			miss.	Moder.	ML1040		hypothetical protein	Unknown	Gly406Val	429
1205773	A							A		A	G			A	A			miss.	Moder.	ML1040		hypothetical protein	Unknown	Leu255Ser	429
1205818	A							A		g	A			A	A			miss.	Moder.	ML1040		hypothetical protein	Unknown	Ile240Thr	429
1205821	C						c	C		C	C			C	C			miss.	Moder.	ML1040		hypothetical protein	Unknown	Gly239Asp	429
1205841	G							G		g	G			G	G			miss.	Moder.	ML1040		hypothetical protein	Unknown	Asp232Glu	429
1205842	T							T		t	T			T	T			miss.	Moder.	ML1040		hypothetical protein	Unknown	Asp232Val	429

1206130	C						C	c		C	C			C	C			miss.	Moder.	ML1040		hypothetical protein	Unknown	Arg136His	429
1218705	C						c	C		C	C			C	C			miss.	Moder.	ML1052		hypothetical protein	Unknown	Glu375Gln	421
1227631	C				A		A	A	A	A	A			A	A			intrag.	Modif.	ML1062	fadD6	pseudogene	Pseudogene		
1232637	G						G	G		G	G			G	A			intrag.	Modif.	ML1068	glgA	pseudogene	Pseudogene		
1256127	G						G	G		G	G			A	G			intrag.	Modif.	ML1092	PE-PGRS3	pseudogene	Pseudogene		
1257009	A						A	A		G	A			A	A			interg.	Modif.						
1257185	T						C	C	C	C	C		C	C	C	C		interg.	Modif.						
1262010	T						T	T	T	C	T			T	T	T		miss.	Moder.	ML1095	sucA	2-oxoglutarate dehydrogenase	Metabolism and respiration	Thr579Ala	1238
1265267	T				G		G	G	G	G	G			G	G	G		intrag.	Modif.	ML1097		pseudogene	Pseudogene		
1276011	C						C	C	C	T	C			C	C			intrag.	Modif.	ML1102	cyp130	pseudogene	Pseudogene		
1276847	C						C	C	C	T	C			C	C			miss.	Moder.	ML1103		oxidoreductase	Metabolism and respiration	Ala325Thr	453
1280194	G						g	C	C	G	G			C	G		C	interg.	Modif.						
1291018	G						A	A	A	G	A			A	G		A	syn.	Low	ML1114		transmembrane ATP-binding protein ABC transporter.	Cell wall and cell processes	Asn120Asn	584
1295192	A						A	A		G	A			A	A			miss.	Moder.	ML1119		hypothetical protein	Unknown		384
1324009	C						G	G		G	G		G	G	G	G		miss.	Moder.	ML1132	rho	transcription termination factor rho	Inf. pathways	Gln227Glu	610
1329459	C						T	T	T	C	T			T	C	C	T	syn.	Low	ML1137	rfe	acetylglucosaminyltransferase	Cell wall and cell processes	Arg320Arg	399
1339813	T						C	C	C	C	C		C	C	C	C		syn.	Low	ML1150	murA	carboxyvinyltransferase	Cell wall and cell processes	Asp78Asp	418
1341139	G							G		G	G			G	G	g		upst. 5 b.	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341139	G							G		G	G			G	G	g		interg.	Modif.						
1341142	T						g	t	g	T	T		G	t	T	t	g	upst. 2 b.	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341142	T						g	t	g	T	T		G	t	T	t	g	interg.	Modif.						

1341179	C		t	C		C	C	C	C	C	C	t	C	C	C	C	C	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341215	A						a	A		A	A		a	a	A	A	g	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341417	C		T				c	c		C	C		T	t	C	T	t	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341460	A							c		A	A		c	A	A	C		_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341481	T		C	c	c	c	t	t	c	T	t		t	c	T	t	c	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341523	C	g	C	c			C	C	C	C	c	g	C	C	C	C	C	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341534	C		g				c	C	c	C	C		c	c	C	c	c	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341549	G					G	G	G	G	G			G	G	G	G	G	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341560	G					A		G	a	G	G		a	G	G			_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341572	T						T	t		T	T		c	T	T	c		_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341636	G						G	G		G	G			G	G	G		_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341637	G						G	g		G	G			G	G	c		_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341653	C						c	C	c	C	C			c	C	c	t	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341688	T					t	g		t	T	T		T	t	T	t	t	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341709	C						C	C	c	C	C		C	c	C	C		_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341747	T						T	T	T	T	T		T	T	T	c		_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341755	T						T	T		T	T			T	T	t		_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341773	A						A	a		A	A			A	A	a		_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341798	G						G	G		G	G		G	G	G	g		_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341807	C						C	C		C	C		C	C	C	c	C	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341816	C						C	C	C	C	C		C	C	C	C	c	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		

1341822	T						t			T	T		t	t	T	T	t	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341866	A					G	g	g	g	A			G	g	A	G	G	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341907	A	T	t		t	a	a	a	a	A	a	T	a	a	A		a	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341925	A	A	a	a	A	a	g	a	a	A	A		a	g	A	g	a	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341964	G					G		G	a	G	G		g	G	G		G	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341991	T						t	T		T	T		T	T	T	T		_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1341992	T						t	T		T	T		T	t	T	g		_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342001	T						T	T	T	T	T		T	T	T	T	t	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342022	T		T			T	T	T	t	T	T		T	T	T	T	T	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342023	A		t			g	g	a	g	A	A		g	g	A	g	g	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342064	G	G	G	G	G	G	G	G	G	G	G	G		G	G	G	G	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342099	T	T	t	c	t	C	c	t	c	T	t	T	c	c		c	T	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342153	T						T	T		T	T			T	T	c		_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342170	A						A	A		A	A			A	A	g		_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342171	C						C	C		C	C			C	C	t		_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342189	G							G	g	G	G		G		G	G	G	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342196	G	c	g	g	c	g	g	g	g	g	g	c	g	g		g	g	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342204	C		C	C		c	c	c	c	C	C		c	c	c	c	c	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342260	T						t	T	t	T	T		T		T	T	t	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342261	T						T	T		T	T		T		T	t	T	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342264	C						C	c	C	C	C		C	C	C	C	C	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		

1342266	C						C	C		C	C		C	C	C	C	c	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342267	A							A		A	A			a	A	A		_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342282	A						A	a		A	A		A	A	A	A		_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342310	C	G			c		c	g	G	C	C	G	G	g	C	G	g	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342344	G				G		G	G	G	G	G		G	G	G	G		_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342353	C				C		c	C	C	C	C		C	C	C	C	C	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342375	A						a	A	a	A			a	a	A	a	g	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342397	A						A	A	a	A	A		A	a	A	A	A	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342409	G						G			G	G			G	G	a		_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342455	C							C	c	C			t	c	C	t	c	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342476	G		G		G		g	g	g	G	g		g	G	G	a	G	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342483	C		C	C	C		c	C	C	C	C		C	c	C	C	C	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342531	C		C	C	C	t	c	c	c	C	C		c	c	C	c	c	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342557	C		C		C	C	C	C	C	C	C		C	C	C	c	C	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342562	T		t	c	c		T	T	c	T	T		t	t	T	t	T	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342562	T		t	c	c		T	T	c	T	T		t	t	T	t	T	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342572	T		t	c	T		t	T	t	T	t		t	t	T	T	T	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342572	T		t	c	T		t	T	t	T	t		t	t	T	T	T	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342586	A						a	a	a	A	A		a	a	A	a	A	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342593	T						T	T		T	T		T		T	T	C	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342611	T						T	T		T	T			T	T	t		_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		

1342615	C						C	c		C	C		g	C	C	g		_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342626	C						c	c		C	C			c	C	C	c	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342627	G						G	G	G	G	G			G	G	g	G	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342629	C						c	c	c	C	C		c	c	C	c	t	_exon	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342694	A						A	A	a	A	A		a	A	A	A		downst. 2 b.	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342694	A						A	A	a	A	A		a	A	A	A		interg.	Modif.						
1342697	A						a	A		A	A		A		A	A	a	downst. 5 b.	Modif.	MLP000016	rrs	16S ribosomal RNA	Stable RNAs		
1342697	A						a	A		A	A		A		A	A	a	interg.	Modif.						
1343066	T							T	T	T	T		T	T	T	c		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1343072	G						G	G		G	G		G	G	G	g		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1343076	G						G	G		G	G		G	G	G	g		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1343090	T						T	T		T	T		G	T	T	g		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1343434	T							T		T	T			T	T	t		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1343496	A						A	a		A	A			A	A	g		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1343523	T						T	T	T	T	T		T	T	T	t		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1343524	C						c	C	c	C	C		c	c	C	c	G	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1343571	T					c	t	T	t	T	T		c	t	T	t	t	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1343585	G						G	G	G	G	G		G	G	G	g		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1343631	T						T	t		T	T			T	T	T		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1343681	T						T	t		T	T			T	T	T		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1343683	T						T	T		T	T			T	T	T		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		

1343751	C						C	c		C	C			C	C	g		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1343829	A						A	a		A	A			A	A	g		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1343838	T							t		T	T			T	T	c	c	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1343850	G							G	G	G	G			G	G	G	G	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1343851	T							t	c	T	T		t	t	T	T	T	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1343857	G						G	G	G	G	G		g	g	G	g	G	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1343881	T						t	T	t	T	T		t	t	T	T	t	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1343925	T							T	t	T	T			t	T	t	T	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1343927	C						c	C	C	C	C		C	C	C	C	c	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1343928	G						G	G	G	G	G		G	G	G	g		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1343931	G						G	G		G	G				G	g		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1343975	T						T	T	T	T	T			T	T		A	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1344053	A						A	a		A	A		g	A	A	a		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1344060	A							a	a	A	A		g	a	A	g		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1344094	T						T	t	T	T	T		T	T	T	a	T	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1344103	G							G	G	G	G		G	G	G	G	G	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1344123	C							C	C	C	C		C	c	C	C		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1344130	A							g		A	A		G	a	A	g		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1344164	C						c	c	c	C	C		c	c	C	c	c	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1344169	G						g	G	G	G	G		G	G	G	G		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1344204	G							G	G	G	G		G	g	G	G		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		

1344253	C						C	C		C	C			C	C	c		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1344254	A						A	A		A	A			A	A	a		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1344364	A						A	A		A	A		A	A	A	a		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1344390	T							t		T	T		A	T	T	a		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1344422	G						G	g	G	G	G		G	G	G	t		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1344426	A						a	a		A	A		A	a	A	A		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1344431	G		A				G	g	g	G	G		g	G	G	g	G	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1344504	T						T	T		T	T			T	T	t		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1344665	A						A	A		A	A			A	A	a		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1344880	T				T			T	t	T	T		c	t	T	c		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1344953	G						G	G	G	G	G			G	G	g	G	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1344959	T						T	T	t	T	T				T	T	T	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1344961	A						A	A	A	A	A			A	A	a	g	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345003	T			c	C			t	c	T	T		t	t	T	c	c	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345007	A							A	A	A	A		a	a	A	A	A	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345010	A						A	A	A	A	A		a	A	A	A	A	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345044	G						G	G	g	G	G		G	G	G	G	G	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345047	T						t	t		T	T				T			_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345057	T						T	T		T	T			T	T	g	T	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345079	G				t		G	g	g	G	G		g	g	G	t		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345081	A							a	a	A	A		a		A	a	a	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		

1345085	T						t	T	T	T	T			t	T	t	t	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345087	T						t	t	T	T	T		T		T	T	T	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345092	C		G			C		C	c	C	C		c	c	C	c	C	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345098	A		A	A		a	a	a	a	A	A		a	a	A	a	a	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345183	T			T	T		t	T	T	T	T		T		T	T	t	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345185	T			C				T	t	T					T	T	c	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345187	A				g		A	A	g	A	A			a	A	A	a	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345188	A							A		A	A		g	a	A	A		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345220	C						C	c	C	C	C		C	C	C	T		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345244	T						T	t	t	T	T		c		T	c		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345435	C							c	t	C	C			c	C	t	t	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345469	T						T	T		T	T			T	T		T	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345504	T						T	T		T	T			T	T	a		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345564	A							A		A	A			A	A	a		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345585	T						T	t		T	T			T	T	c		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345601	T							t		T	T		C	T	T	t		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345603	T							t		T	T		C	T	T	t		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345611	A						A	A	A	A	A		a	A	A	A		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345613	G						g	G		G	G		G	g	G	G		_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345616	T							T	t	T	T			t	T	t	T	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		
1345629	A		A	A			A	A	A	A			a	A	A	A	A	_exon	Modif.	MLP000017	rrl	23S ribosomal RNA	Stable RNAs		

1345 663	C		C	C		C	c	C	c	C	C		C	C	C	C	c	_exon	Modi f.	MLP000 017	rrl	23S ribosomal RNA	Stable RNAs		
1345 685	G		G	G		g	g	g	g	G	G		g	g	G	g	c	_exon	Modi f.	MLP000 017	rrl	23S ribosomal RNA	Stable RNAs		
1345 712	G		G	G	G	G	G	G	G	G	G		G	G	G	G		_exon	Modi f.	MLP000 017	rrl	23S ribosomal RNA	Stable RNAs		
1345 729	A		g	G		g	g	g	g	A	a		g	g		g	g	_exon	Modi f.	MLP000 017	rrl	23S ribosomal RNA	Stable RNAs		
1345 732	C	T	t	T	t	c	c	C	t	C	c	T	c	c	c	C	c	_exon	Modi f.	MLP000 017	rrl	23S ribosomal RNA	Stable RNAs		
1345 770	C		c			t	c	c	c	C	C		c	c	C	c	t	_exon	Modi f.	MLP000 017	rrl	23S ribosomal RNA	Stable RNAs		
1345 785	T			T	t	t		T	t	T	t		t	T	t	T	t	_exon	Modi f.	MLP000 017	rrl	23S ribosomal RNA	Stable RNAs		
1345 795	T		T	c	T	T	T	T	T	T	T		t	T	T	T	T	_exon	Modi f.	MLP000 017	rrl	23S ribosomal RNA	Stable RNAs		
1345 810	T						t	T	c	T			c	t	T	T	c	_exon	Modi f.	MLP000 017	rrl	23S ribosomal RNA	Stable RNAs		
1345 833	A						A	a	A	A	A		G	A	A	t		_exon	Modi f.	MLP000 017	rrl	23S ribosomal RNA	Stable RNAs		
1345 843	A						a	A		A	A			a	A	A		_exon	Modi f.	MLP000 017	rrl	23S ribosomal RNA	Stable RNAs		
1345 844	T						t	T	t	T	T		C	t	T	c		_exon	Modi f.	MLP000 017	rrl	23S ribosomal RNA	Stable RNAs		
1346 226	A						A	A		A	A			A	A	a		_exon	Modi f.	MLP000 017	rrl	23S ribosomal RNA	Stable RNAs		
1346 270	A						A	A		A	A			A	A	a		_exon	Modi f.	MLP000 017	rrl	23S ribosomal RNA	Stable RNAs		
1348 426	T						C	C	C	C	C			C	C	C	C	interg.	Modi f.						
1351 149	C						G	G	G	G	G			G	G	G	G	interg.	Modi f.						
1368 031	C						C	C		T	C			C	C			intrag.	Modi f.	ML1170	cysM	pseudogene	Pseudogene		
1372 166	C						C	C	C	C	C			C	T			miss.	Mod er.	ML1175		hypothetical protein	Unknown	Pro129S er	220
1409 826	G							G		A	G			G	G			intrag.	Modi f.	ML1194		pseudogene	Pseudogene		
1414 666	A									a								interg.	Modi f.						
1414 674	A									a								interg.	Modi f.						

1415370	C						C	C		T	C			C	C			intrag.	Modif.	ML1197	dinX	pseudogene	Pseudogene		
1456091	C						C	C		C	A			C	C			miss.	Moder.	ML1229	pks4 pks3	polyketide beta-ketoacyl synthase	Lipid metabolism	Thr598Lys	2118
1459878	C						C	C		T	C			C	C			syn.	Low	ML1229	pks4	polyketide beta-ketoacyl synthase	Lipid metabolism	Val1860Val	2118
1477964	T						T	T		t	T			T	T			interg.	Modif.						
1477965	T						T	T		T	T			T	T			interg.	Modif.						
1500834	C														T			interg.	Modif.						
1529088	A						G	G	G	G	G			G	G	G		intrag.	Modif.	ML1284	cydA	pseudogene	Pseudogene		
1530476	T							T	t	T	T			T	T			intrag.	Modif.	ML1285	cya	pseudogene	Pseudogene		
1530718	A						A	A	A	G	A			A	A			intrag.	Modif.	ML1285	cya	pseudogene	Pseudogene		
1531185	A						A	A		a	A				A			intrag.	Modif.	ML1285	cya	pseudogene	Pseudogene		
1533315	C									c								intrag.	Modif.	ML1287		pseudogene	Pseudogene		
1533316	A									a								intrag.	Modif.	ML1287		pseudogene	Pseudogene		
1533317	T									t								intrag.	Modif.	ML1287		pseudogene	Pseudogene		
1542248	C						c	C	C	C	C			C	C			interg.	Modif.						
1548486	T						t	T	T	T	T			T	T	T		miss.	Moder.	ML1300		hypothetical protein	Unknown	Val32Ala	262
1579036	T						T	T		C	T			T	T			intrag.	Modif.	ML1327		pseudogene	Pseudogene		
1579109	T						G	G	G	G	G			G	G	G		intrag.	Modif.	ML1327		pseudogene	Pseudogene		
1585800	A						g	G	G	A	A			G	A			miss.	Moder.	ML1333	helY	DNA helicase	Inf. pathways	Thr436Ala	920
1587661	C							C		t	C			C	C			miss.	Moder.	ML1334		hypothetical protein	Unknown	Pro114Ser	272
1587697	C									C	C			t	C			miss.	Moder.	ML1334		hypothetical protein	Unknown	Pro126Ser	272
1601902	C						C	C		A	C			C	C			interg.	Modif.						

1602395	T							T	T	T	T				T	t		downst. 70 b.	Modif.	MLP000024	cysU	tRNA-Cys anticodon GCA	Stable RNAs		24
1602395	T							T	T	T	T				T	t		miss.	Moder.	MLP000023	valU	tRNA-Val anticodon GAC	Stable RNAs	Gln19Arg	24
1602426	T							T	t	T	T				T	T		downst. 39 b.	Modif.	MLP000024	cysU	tRNA-Cys anticodon GCA	Stable RNAs		24
1602426	T							T	t	T	T				T	T		miss.	Moder.	MLP000023	valU	tRNA-Val anticodon GAC	Stable RNAs	Thr9Ala	24
1602539	A						A	A	A	A	A			A	A	a		UPSTREAM: 1 base	Modif.	MLP000024	cysU	tRNA-Cys anticodon GCA	Stable RNAs		24
1602539	A						A	A	A	A	A			A	A	a		downst. 25 b.	Modif.	MLP000025	glyT	tRNA-Gly anticodon GCC	Stable RNAs		24
1602539	A						A	A	A	A	A			A	A	a		interg.	Modif.						
1602573	T						T	T	t	T	T			T	T		T	upst. 35 b.	Modif.	MLP000024	cysU	tRNA-Cys anticodon GCA	Stable RNAs		24
1602573	T						T	T	t	T	T			T	T		T	miss.	Moder.	MLP000025	glyT	tRNA-Gly anticodon GCC	Stable RNAs	Ile23Val	24
1602581	T					C	c	T	t	T	T			t	T	c	t	UPSTR 43 b.	Modif.	MLP000024	cysU	tRNA-Cys anticodon GCA	Stable RNAs		24
1602581	T					C	c	T	t	T	T			t	T	c	t	miss.	Moder.	MLP000025	glyT	tRNA-Gly anticodon GCC	Stable RNAs	Asn20Ser	24
1602617	G							G		G	G		t		G	g		upst. 79 b.	Modif.	MLP000024	cysU	tRNA-Cys anticodon GCA	Stable RNAs		24
1602617	G							G		G	G		t		G	g		miss.	Moder.	MLP000025	glyT	tRNA-Gly anticodon GCC	Stable RNAs	Ala8Glu	24
1602624	A						A	A	a	A	A				A	a		upst. 86 b.	Modif.	MLP000024	cysU	tRNA-Cys anticodon GCA	Stable RNAs		24
1602624	A						A	A	a	A	A				A	a		syn.	Low	MLP000025	glyT	tRNA-Gly anticodon GCC	Stable RNAs	Leu6Leu	24
1604320	A						A	A		A	A			A	G			intrag.	Modif.	ML1345		pseudogene	Pseudogene		
1605956	G						A	A	A	A	A			A	A		A	interg.	Modif.						
1614069	T						C	C		C	C			C	C	C	C	intrag.	Modif.	ML1353		pseudogene	Pseudogene		
1625045	T	G					G	G	G	G	G	G		G	G	G	G	syn.	Low	ML1363	pyrG	CTP synthase	Metabolism and respiration	Pro298Pro	590
1635967	G						g	G	G	G	G			G	G			interg.	Modif.						
1635969	G						g	G	G	G	G			G	G			interg.	Modif.						

1638410	T						t		t	T	T		T	t	T	t	g	miss.	Moder.	MLP000027	proT	tRNA-Pro anticodon GGG	Stable RNAs	Ser15Arg	24
1642875	G				T		T	T	T	T	T			T	T	T	T	intrag.	Modif.	ML1378		pseudogene	Pseudogene		
1643162	T						C	C	C	C	C		C	C	C	C	C	intrag.	Modif.	ML1378		pseudogene	Pseudogene		
1648026	T									C								interg.	Modif.						
1648105	C							T		C	T							interg.	Modif.						
1657373	G				T		G	G		G	G			G	G			intrag.	Modif.	ML1385		pseudogene	Pseudogene		
1665902	G						G	G		C	G			G	G			miss.	Moder.	ML1389		transmembrane protein	Cell wall and cell processes	Pro286Ala	527
1677493	G						C	C	C	C	C		C	C	C	C		syn.	Low	ML1397	tsnR	23S rRNA methyltransferase	Inf. pathways	Ala129Ala	260
1679768	A							A	A	G	A			A	A	A		syn.	Low	ML1399		hypothetical protein	Unknown	Arg185Arg	324
1701590	G						C	C	C	C	C			C	C	C		interg.	Modif.						
1731083	C						C	C		C	C			C	T	C		intrag.	Modif.	ML1437	pks12	pseudogene	Pseudogene		
1732115	A						A	A		a	A			A	A			intrag.	Modif.	ML1437	pks12	pseudogene	Pseudogene		
1732116	T						T	T		t	T			T	T			intrag.	Modif.	ML1437	pks12	pseudogene	Pseudogene		
1732117	A						A	A		a	A			A	A			intrag.	Modif.	ML1437	pks12	pseudogene	Pseudogene		
1753548	C						C	C		C	C			C	A			intrag.	Modif.	ML1459		pseudogene	Pseudogene		
1758390	T						t	T		T	T			T	T			intrag.	Modif.	ML1461	rbsK	pseudogene	Pseudogene		
1759651	G						A	A		G	A			A	G			miss.	Moder.	ML1463	nadE	glutamine-dependent NAD(+) synthetase	Metabolism and respiration	Ala566Val	680
1773379	G						G	G	G	G	G			G	A			interg.	Modif.						
1785091	G						G	g		G	G			G	G	a		upst. 2 b.	Modif.	MLP000028	proU	tRNA-Pro anticodon TGG	Stable RNAs		
1785091	G						G	g		G	G			G	G	a		interg.	Modif.						
1787399	C						T	T		C	T			T	C			miss.	Moder.	ML1484		isomerase	Metabolism and respiration	Gly159Asp	162

1792 670	A					c	C	C	A	A			C	A			interg.	Modi f.						
1796 513	G	T				G	G		G	G	T		G	G			intrag.	Modi f.	ML1490	fadH	pseudogene	Pseudogene		
1803 291	C					C	C		T	C			C	C			intrag.	Modi f.	ML1496		pseudogene	Pseudogene		
1819 543	G					t	T	T	G	G			T	G			downst. 14 b.	Modi f.	ML1508		hypothetical protein	Unknown		
1822 963	G					G	a		G	G			G	G			miss.	Mod er.	ML1512		hypothetical protein	Unknown	Pro549S er	558
1828 126	A					A	A	A	G	A			A	A			interg.	Modi f.						
1836 464	T					T	T		C	T			T	T			intrag.	Modi f.	ML1521		pseudogene	Pseudogene		
1841 279	C								c								intrag.	Modi f.	ML1527	dapB	pseudogene	Pseudogene		
1841 281	C						g		c								intrag.	Modi f.	ML1527	dapB	pseudogene	Pseudogene		
1841 282	T								t								intrag.	Modi f.	ML1527	dapB	pseudogene	Pseudogene		
1843 283	C					A	A	A	A	A			A	A	A		intrag.	Modi f.	ML1528		pseudogene	Pseudogene		
1846 778	C					C	T		C	C			C	C			intrag.	Modi f.	ML1531		pseudogene	Pseudogene		
1848 178	G						G		T					G			downst. 97 b.	Modi f.	ML1532	ald	secreted L-alanine dehydrogenase	Metabolism and respiration		
1848 178	G						G		T					G			interg.	Modi f.						
1848 181	T						T		G					T			interg.	Modi f.						
1849 024	T								t								intrag.	Modi f.	ML1533	PPE	pseudogene	Pseudogene		
1849 026	G								g								intrag.	Modi f.	ML1533	PPE	pseudogene	Pseudogene		
1849 029	G								g								intrag.	Modi f.	ML1533	PPE	pseudogene	Pseudogene		
1872 785	C					C	C		C	C			T	C	C		syn.	Low	ML1548		hypothetical protein	Unknown	Pro273P ro	321
1881 671	T					T	T		T	T			T	T	T	g	syn.	Low	ML1556	infB	initiation factor IF-2	Inf. pathways	Val427V al	924
1890 243	C					G	G		G	G			G	G	G	G	intrag.	Modi f.	ML1564	cobB	pseudogene	Pseudogene		

1890244	G						C	C	C	C	C			C	C	C	C	intrag.	Modif.	ML1564	cobB	pseudogene	Pseudogene		
1892076	C						C	C	C	C	C			C	C	T		intrag.	Modif.	ML1567		pseudogene	Pseudogene		
1897187	C				A		C	C		C	C			C	C			intrag.	Modif.	ML1571	nicT	pseudogene	Pseudogene		
1900006	G				A		A	A	A	G	A			A	G			intrag.	Modif.	ML1574	glnA4	pseudogene	Pseudogene		
1912456	A									a								intrag.	Modif.	ML1586	mpt53	pseudogene	Pseudogene		
1923727	A							A		A	A			A	A		a	syn.	Low	ML1598	rpsB	30S ribosomal protein S2	Inf. pathways	Gly98Gly	277
1926696	T						C	C	C	C	C		C	C	T			downst. 58 b.	Modif.	ML1601		hypothetical protein	Unknown		
1926696	T						C	C	C	C	C		C	C	T			interg.	Modif.						
1934293	G						G	G		G	G			G	T			interg.	Modif.						
1937233	G						G	G		A	G			G	G			miss.	Moder.	ML1613	rplS	50S ribosomal protein L19	Inf. pathways	Pro74Leu	113
1944319	C						c	C		C	C			C	C	C		miss.	Moder.	ML1622	ffh	signal recognition particle protein	Cell wall and cell processes	Gly350Ser	521
1962036	C						T	T		C	T		T	T	C	C	T	syn.	Low	ML1630	glnE	glutamate-ammonia-ligase adenyltransferase	Metabolism and respiration	Thr288Thr	1004
1963524	G							G		G	G			G	A	A		syn.	Low	ML1631	glnA2	glutamine synthase class II	Metabolism and respiration	Ser251Se	448
1972146	A						A	A		A	A			A	A	a		_exon	Modif.	ML1636		pseudogene	Pseudogene		
1972214	T						T	T		T	T			T	T	t		upst. 4 b.	Modif.	ML1636		pseudogene	Pseudogene		
1972214	T						T	T		T	T			T	T	t		interg.	Modif.						
2007166	C						C	C		C	C			T	C			intrag.	Modif.	ML1665	pca	pseudogene	Pseudogene		
2008021	G						g	G	G	G	G			G	G	G		intrag.	Modif.	ML1665	pca	pseudogene	Pseudogene		
2011747	T						G	G	G	G	G		G	G	G	G	G	intrag.	Modif.	ML1668	lipN	pseudogene	Pseudogene		
2015320	G						C	C		C	C			C	G			miss.	Moder.	ML1671	recG	ATP-dependent DNA helicase	Inf. pathways	Asp545Glu	743
2017943	A						A	A	A	G	A			A	A	A		intrag.	Modif.	ML1672		pseudogene	Pseudogene		

2023404	A						A	A	A	C	A			A	A	A		miss.	Moder.	ML1679	gpdA	glycerol-3-phosphate dehydrogenase	Lipid metabolism	Val246Gly	351
2029403	C						C	T		C	C			C	C	C		upst. 84 b.	Modif.	ML1683	hupB	DNA-binding protein HU homolog	Inf. pathways		
2029403	C						C	T		C	C			C	C	C		interg.	Modif.						
2039330	C						C	C	C	A	C			C	C			miss.	Moder.	ML1692	serA1	D-3-phosphoglycerate dehydrogenase	Metabolism and respiration	Ala158Ser	528
2043287	A						G	G	G	G	G			G	G	G		syn.	Low	ML1694	ilvC	ketol-acid reductoisomerase	Metabolism and respiration	Pro145Pro	333
2055222	T						T	T		C	T			T	T			syn.	Low	ML1704		hypothetical protein	Unknown	Asp60Asp	232
2065194	C							C	C	C	G			C	C	C		syn.	Low	ML1712	fixA	electron transfer flavoprotein beta-subunit	Metabolism and respiration	Ser229Ser	266
2077004	G						G	G	A	G	G			G	G			miss.	Moder.	ML1720		hypothetical protein	Unknown	Val302Ile	364
2085748	C						C	C	c	C	C			C	C	C		syn.	Low	ML1728	ctaD	cytochrome C oxidase, polypeptide I	Metabolism and respiration	Gly200Gly	574
2100523	C						G	G	G	G	G			G	G	G		miss.	Moder.	ML1740		hypothetical protein	Unknown	Ala133Pro	312
2104127	T						C	C	C	C	C		C	C	C	C		interg.	Modif.						
2104840	G						G	G		G	G			A	G	G		interg.	Modif.						
2114696	G						A	A		A	A			A	A			interg.	Modif.						
2115662	G						G	g		G	G			G	G			miss.	Moder.	ML1750		hypothetical protein	Unknown	Ser444Leu	623
2116298	T						t	T	T	T	T			T	T	T		miss.	Moder.	ML1750		hypothetical protein	Unknown	Asp232Gly	623
2116748	G						G	G	G	G	G			G	g	G	G	miss.	Moder.	ML1750		hypothetical protein	Unknown	Ala82Asp	623
2116866	C						C	C		C	C			C	c	C		miss.	Moder.	ML1750		hypothetical protein	Unknown	Val43Met	623
2119468	T							T	T	C	T			T	T			interg.	Modif.						
2126590	C						C	C		A	C			C	C			miss.	Moder.	ML1753		transcriptional regulatory protein	Reg. proteins	Asp167Tyr	1106
2127769	A							A		a	A			A	A			interg.	Modif.						
2127770	C							C		A	C			C	C			interg.	Modif.						

2127771	G							G		A	G			G	G			interg.	Modif.						
2139675	G							G		G	A			G	G			intrag.	Modif.	ML1764		pseudogene	Pseudogene		
2146897	G						G	G	G	G	A			G	G	G		intrag.	Modif.	ML1771		pseudogene	Pseudogene		
2155013	T						G	G	G	G	G			G	G	G	G	intrag.	Modif.	ML1778		pseudogene	Pseudogene		
2155013	T						G	G	G	G	G			G	G	G	G	intrag.	Modif.	ML1778		pseudogene	Pseudogene		
2165098	C						C	C		C	C			T	C			interg.	Modif.						
2174865	G	C					C	C	C	C	C	C		C	C			interg.	Modif.						
2175587	C						C	C		G	C			C	C			miss.	Moder.	ML1796		hypothetical protein	Unknown	Gln96Glu	137
2179164	G						G	G		A	G			G	G	G		miss.	Moder.	ML1799	mutB	methylmalonyl-CoA mutase	Lipid metabolism	Ala339Val	758
2184984	G							G		G	A			G	G	G		miss.	Moder.	ML1804		hypothetical protein	Unknown	Gly47Asp	298
2202139	T						T	T		T	T			T	C	C		intrag.	Modif.	ML1819	ctpD	pseudogene	Pseudogene		
2204005	T						c	T	C	T	T			C	T			intrag.	Modif.	ML1820		pseudogene	Pseudogene		
2211036	C							C		C	C				c			interg.	Modif.						
2214876	G						G	G		G	A			G	G			miss.	Moder.	ML1829		hypothetical protein	Unknown	Val22Ile	125
2235048	C						C	C		T	C			C	C			intrag.	Modif.	ML1851		pseudogene	Pseudogene		
2235810	C							C		T	C			C	C			intrag.	Modif.	ML1852		pseudogene	Pseudogene		
2256012	T						T	T		T	T		c	T	T			syn.	Low	ML1877	tuf	iron-regulated elongation factor	Inf. pathways	Thr323Th	396
2256708	A						A	A		A	A			a	A			syn.	Low	ML1877	tuf	iron-regulated elongation factor	Inf. pathways	Ile91Ile	396
2269826	A						A	A		C	A			A	A	A		miss.	Moder.	ML1890	rpoC	DNA-directed RNA polymerase, beta chain	Inf. pathways	Ile1132Se	1316
2291484	G							G	G	G	G				G	g		syn.	Low	ML1905	rplK	50S ribosomal protein L11	Inf. pathways	Gly31Gly	142
2292915	T						T	t		T	T				T	t		miss.	Moder.	MLP000035	trpT	tRNA-Trp anticodon CCA	Stable RNAs	Asn20Ser	23

2292968	G						G	G	G	G	G			G	G	G		syn.	Low	MLP000035	trpT	tRNA-Trp anticodon CCA	Stable RNAs	Gly2Gly	23
2297562	T						T	T		A	T			T	T			intrag.	Modif.	ML1914	lpqN	pseudogene	Pseudogene		
2312059	C						C	C		G	C			C	C	C		downst. 33 b.	Modif.	ML1927		hypothetical protein	Unknown		103
2312059	C						C	C		G	C			C	C	C		START_LOST	High	ML1926		hypothetical protein	Unknown	Leu1?	167
2338948	C						C	C	C	G	C			C	C	C		downst. 40 b.	Modif.	ML1947	fum	fumarase	Metabolism and respiration		475
2338948	C						C	C	C	G	C			C	C	C		upst. 28 b.	Modif.	ML1951	phoH2	phoh-like protein	Metabolism and respiration		434
2338948	C						C	C	C	G	C			C	C	C		interg.	Modif.						
2369366	C						C	C		A	C			C	C			interg.	Modif.						
2378341	G						a	A	A	G	G			A	G			interg.	Modif.						
2388071	G						G	G		C	G			G	G	G		miss.	Moder.	ML1997		hypothetical protein	Unknown	Pro32Ala	210
2392817	C						c	C		C	C			C	C			miss.	Moder.	ML2000	ctpB	cation-transporter P-type ATPase B	Cell wall and cell processes	Leu251Phe	750
2396326	C						c	C	C	C	T			C	C			interg.	Modif.						
2411498	T						T	T		T	T			T	G			miss.	Moder.	ML2020		hypothetical protein	Unknown	Tyr62Asp	304
2432314	A						A	A		A	A			A	A	G		syn.	Low	ML2041	oxyR	hydrogen peroxide-inducible genes activator	Reg. proteins	Asp202Asp	311
2433329	G						g	G		G	A			G	G	G		miss.	Moder.	ML2042	ahpC	alkyl hydroperoxide reductase	Virulee, detoxification, adaptation	Glu101Lys	195
2459766	A						G	G	G	G	G			G	G	G	G	miss.	Moder.	ML2069	glcB	malate synthase G	Metabolism and respiration	Lys591Glu	731
2483907	C						C	C		c	C			C	C			intrag.	Modif.	ML2089		pseudogene	Pseudogene		
2483908	A						A	A		a	A			A	A			intrag.	Modif.	ML2089		pseudogene	Pseudogene		
2483909	A						A	A		a	A			A	A			intrag.	Modif.	ML2089		pseudogene	Pseudogene		
2483910	C						C	C		c	C			C	C			intrag.	Modif.	ML2089		pseudogene	Pseudogene		
2494215	T							T		C	T			T	T			intrag.	Modif.	ML2097		pseudogene	Pseudogene		

2498281	C							C		C	C			C	T	T		interg.	Modif.						
2500011	G						T	G		A	G			G	G			interg.	Modif.						
2501479	G						g	G	G	G	G			G	G			intrag.	Modif.	ML2103		pseudogene	Pseudogene		
2503289	A	G						A		A	A	G		A	A			interg.	Modif.						
2504925	C						C	C	T	C	C			C	C			interg.	Modif.						
2514965	A						a	A		A	A			A	A			intrag.	Modif.	ML2115	ctpE	pseudogene	Pseudogene		
2542188	A						A	A		A	A			A	C	C		miss.	Moder.	ML2140		transcriptional regulatory protein	Reg. proteins	Phe77Val	143
2542313	G						A	A	A	G	A			A	G	G		miss.	Moder.	ML2140		transcriptional regulatory protein	Reg. proteins	Ser35Leu	143
2548278	A						A	A	A	A	A			A	A	a		intrag.	Modif.	ML2145	cycA	pseudogene	Pseudogene		
2553176	T				G		G	G	G	G	G		G	G	G	G	G	intrag.	Modif.	ML2149	moaA2	pseudogene	Pseudogene		
2556538	G							G		G	A			G	G			intrag.	Modif.	ML2149	moaC2	pseudogene	Pseudogene		
2564872	T						T	T		C	T			T	T	T		interg.	Modif.						
2567264	G							G	G	g	G			G	t			interg.	Modif.						
2581883	G						G	G		G	G			A	G			intrag.	Modif.	ML2171		pseudogene	Pseudogene		
2582353	A						A	A		A	G			A	A			intrag.	Modif.	ML2171		pseudogene	Pseudogene		
2587080	C			C			C	C		C	C			C	C			interg.	Modif.						
2594702	C						c	c	c	C	C		c	c	C	c	c	miss.	Moder.	MLP000040	aspT	tRNA-Asp anticodon GTC	Stable RNAs	Glu20Lys	24
2594702	C						c	c	c	C	C		c	c	C	c	c	downst. 88 b.	Modif.	MLP000041	gluT	tRNA-Glu anticodon TTC	Stable RNAs		24
2594734	G						G	G	g	G	G		G	G	G	G	G	miss.	Moder.	MLP000040	aspT	tRNA-Asp anticodon GTC	Stable RNAs	Ala9Val	24
2594734	G						G	G	g	G	G		G	G	G	G	G	downst. 56 b.	Modif.	MLP000041	gluT	tRNA-Glu anticodon TTC	Stable RNAs		24
2594734	G						G	G	g	G	G		G	G	G	G	G	upst. 80 b.	Modif.	MLP000039	pheU	tRNA-Phe anticodon GAA	Stable RNAs		24

2594847	A						A	A		A	A			A	A	g		miss.	Moder.	MLP000041	gluT	tRNA-Glu anticodon TTC	Stable RNAs	Trp6Arg	24
2594847	A						A	A		A	A			A	A	g		upst. 88 b.	Modif.	MLP000039	pheU	tRNA-Phe anticodon GAA	Stable RNAs		24
2650240	A						A	A		G	A			A	A			intrag.	Modif.	ML2232		pseudogene	Pseudogene		
2654119	C						C	C		C	C			C	T	T		miss.	Moder.	ML2235	purD	phosphoribosylamine-glycine ligase	Metabolism and respiration	Asp163Asn	422
2659841	G						G	G		G	G			A	G			downst. 75 b.	Modif.	MLP000043	thrV	tRNA-Thr anticodon TGT	Stable RNAs		24
2659841	G						G	G		G	G			A	G			intrag.	Modif.	ML2241	PE_PGRS	pseudogene	Pseudogene		
2662629	C						C	C	C	T	C			C	C			interg.	Modif.						
2706236	T						G	G	G	G	G			G	G		G	intrag.	Modif.	ML2281		pseudogene	Pseudogene		
2710194	A						A	A		G	A			A	A			intrag.	Modif.	ML2286		pseudogene	Pseudogene		
2712842	G						g	G		G	G			G	G			interg.	Modif.						
2720817	G						G	G		A	G			G	G			syn.	Low	ML2295		protease	Metabolism and respiration	Gly148Gly	234
2731405	G						G	G	G	A	G			G	G			downst. 47 b.	Modif.	ML2306		anion transporter ATPase	Cell wall and cell processes		382
2731405	G						G	G	G	A	G			G	G			downst. 58 b.	Modif.	ML2307	whiB4	whiB-like regulatory protein	Reg. proteins		117
2736812	C							C	C	C	C		c		C	c	c	miss.	Moder.	MLP000045	proY	tRNA-Pro anticodon CGG	Stable RNAs	Leu6Phe	23
2736844	C						c	c	c	C	C		t	c	C	c	t	syn.	Low	MLP000045	proY	tRNA-Pro anticodon CGG	Stable RNAs	Gly16Gly	23
2751783	A						G	G	G	G	G			G	A			syn.	Low	ML2322	asd	aspartate semialdehyde dehydrogenase	Metabolism and respiration	Ala299Ala	351
2769047	A							A		A	A			A	A	a		upst. 4 b.	Modif.	ML2336		hypothetical protein	Unknown		427
2769047	A							A		A	A			A	A	a		interg.	Modif.						
2769082	A						A	A		A	A			A	A	a		upst. 39 b.	Modif.	ML2336		hypothetical protein	Unknown		
2769082	A						A	A		A	A			A	A	a		interg.	Modif.						
2775050	G						G	G	G	A	G			G	G			interg.	Modif.						

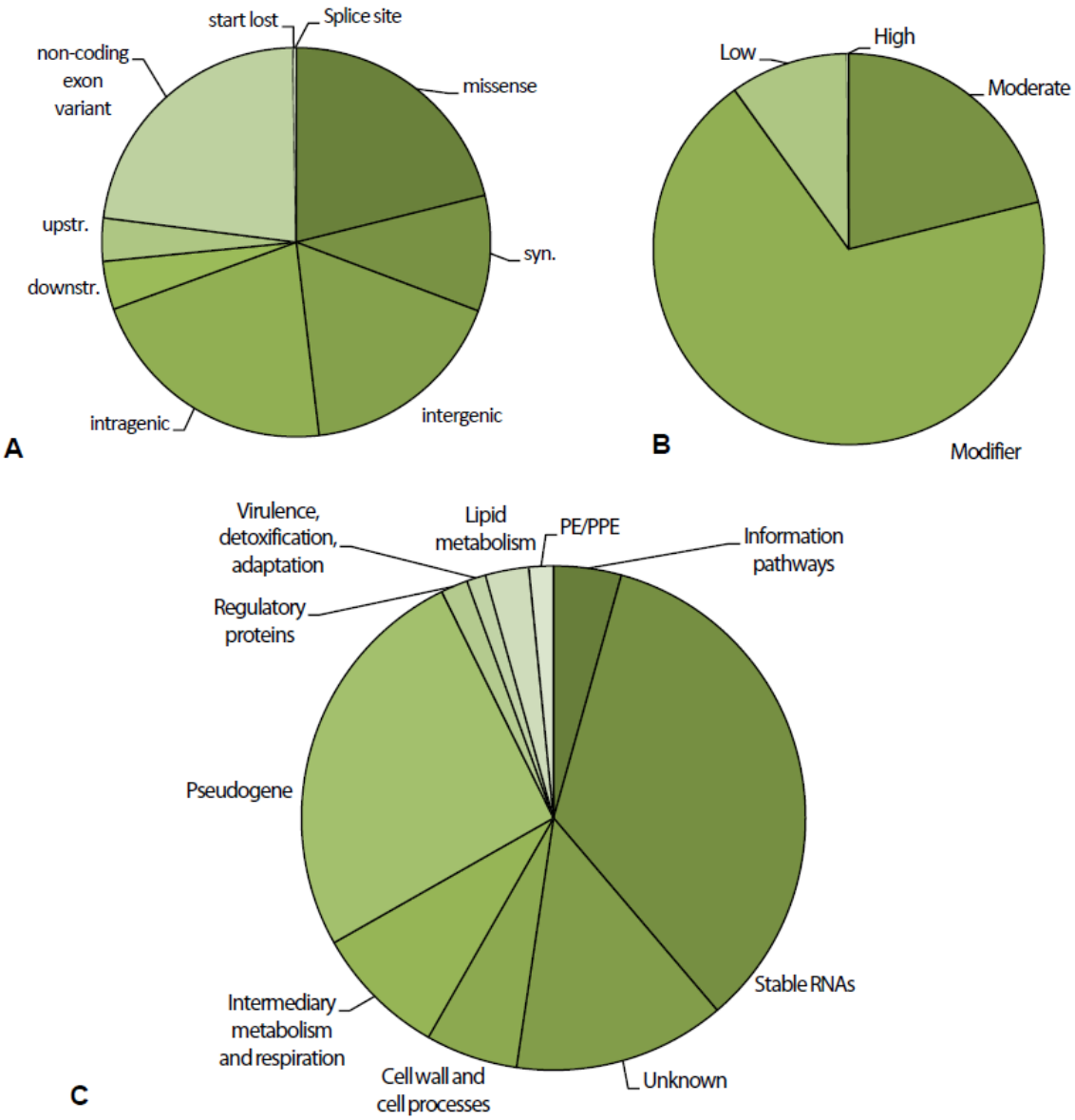
2785462	A							A		a	A				A			interg.	Modif.						
2785466	G									g	G				G			interg.	Modif.						
2789225	G						G	G		G	G		A	G	G			miss.	Moder.	ML2347		membrane protein	Cell wall and cell processes	Arg309His	336
2802569	C						C	C		C	C			C	T			syn.	Low	ML2354	ppsD	type-I polyketide synthase	Lipid metabolism	Ala768Ala	1822
2814758	G							G		G	A			G	G			miss.	Moder.	ML2356	ppsB	type-I polyketide synthase	Lipid metabolism	Leu446Phe	1540
2815232	G						G	G	G	G	G			A	G			syn.	Low	ML2356	ppsB	type-I polyketide synthase	Lipid metabolism	Leu288Leu	1540
2818521	T				C		C	C	C	C	C			C	C	C	C	syn.	Low	ML2357	ppsA	type-I polyketide synthase	Lipid metabolism	Gln1063Gln	1871
2820167	A						A	A	A	G	A			A	A	A		miss.	Moder.	ML2357	ppsA	type-I polyketide synthase	Lipid metabolism	Phe515Leu	1871
2822967	A						A	A		G	A			A	A	A		miss.	Moder.	ML2358	fadD26	fatty-acid-CoA synthetase	Lipid metabolism	Phe164Ser	583
2831819	C						C	C	C	C	C		T	C	C			interg.	Modif.						
2856387	G						G	A	G	G	G			G	G			interg.	Modif.						
2873412	T						T	T	T	C	T			T	T	T		syn.	Low	ML2402	echA8	enoyl-CoA hydratase	Lipid metabolism	Thr194Th	257
2881087	G						G	G	G	A	G			G	G			syn.	Low	ML2410	ML2410	transmembrane protein	Cell wall and cell processes	Ala521Ala	542
2887094	C									C								interg.	Modif.						
2893091	G									t					t			interg.	Modif.						
2893093	G									g								interg.	Modif.						
2900276	C						C	C		G	C			C	C			miss.	Moder.	ML2424	ser1B	phosphoserine phosphatase	Metabolism and respiration	Ala93Gly	300
2902257	C						C	C	C	G	C			C	C			miss.	Moder.	ML2426	cmaA2	cyclopropane mycolic acid synthase	Lipid metabolism	Arg248Gly	308
2905322	G						G	G	G	G	A			G	G	G		miss.	Moder.	ML2429		hypothetical protein	Unknown	Arg10Trp	86
2915098	A						A	A		a	A			A	A	A		upst. 13 b.	Modif.	ML2439	regX3	sensory transduction protein	Reg. proteins		199
2915098	A						A	A		a	A			A	A	A		downst. 30 b.	Modif.	ML2440	senX3	sensor histidine kinase	Reg. proteins		442

2915098	A						A	A		a	A			A	A	A		interg.	Modif.						
2935685	A						C	C	C	C	C			C	A	A	C	intrag.	Modif.	ML2462	icl	pseudogene	Pseudogene		
2936401	G						G	G	G	A	G			G	G			intrag.	Modif.	ML2462	icl	pseudogene	Pseudogene		
2944178	C						C	C		a	C			G	G			interg.	Modif.						
2944181	A						A	A		C	A			A	A			interg.	Modif.						
2945510	A							A		G	A				A			interg.	Modif.						
2945512	A							A		g	A				A			interg.	Modif.						
2945513	A							A		g	A				A			interg.	Modif.						
2945530	A									t	A				A			interg.	Modif.						
2945531	T									T	T				T			interg.	Modif.						
2945532	C									t	C				C			interg.	Modif.						
2948105	G							G		g								miss.	Moder.	ML2472	tyrA	prephenate dehydrogenase	Metabolism and respiration	Arg13Gly	327
2948109	G									g								syn.	Low	ML2472	tyrA	prephenate dehydrogenase	Metabolism and respiration	Val11Val	327
2951822	T									t					T			interg.	Modif.						
2951823	A									a					A			interg.	Modif.						
2961785	G						G	G		G	C			G	G			interg.	Modif.						
2964999	T				C		C	C	C	C	C			C	C	C	C	syn.	Low	ML2490	clpB	endopeptidase ATP binding protein	Virulee, detoxification, adaptation	Glu256Glu	848
2969168	C						C	C		G	C			C	C			interg.	Modif.						
2971055	T			C			T	T	T	T	T			T	T			miss.	Moder.	ML2495	grpE	grpE protein	Virulee, detoxification, adaptation	His172Arg	229
2978023	C						C	C		T	C			C	C			interg.	Modif.						
2978374	C						A	A	A	C	A			A	C		A	interg.	Modif.						

2990150	T						C	C	C	C	C			C	C	C		interg.	Modif.						
2992340	T						T	T		C	T			T	T			interg.	Modif.						
2999976	C						T	T	T	C	T			T	C		T	miss.	Moder.	ML2518		hypothetical protein	Unknown	Gly78Asp	130
3016175	T						C	C	C	C	C			C	C	C		syn.	Low	ML2534	PE10	PE-family protein	PE/PPE	Val79Val	102
3021965	C							C	C	T	C			C	C			syn.	Low	ML2536	eccB3	ESX component protein	Cell wall and cell processes	Val42Val	552
3031636	G						G	G		C	G			G	G			intrag.	Modif.	ML2545	fadE6	pseudogene	Pseudogene		
3041900	T						T	T	T	C	T			T	T			intrag.	Modif.	ML2552	narK3	pseudogene	Pseudogene		
3047644	A						A	A	A	G	A			A	A			intrag.	Modif.	ML2555	nirB	pseudogene	Pseudogene		
3057124	A						A	A		G	A			A	A			downst. 60 b.	Modif.	ML2563	fadE5	acyl-CoA dehydrogenase	Lipid metabolism		612
3057124	A						A	A		G	A			A	A			miss.	Moder.	ML2563	fadE5	acyl-CoA dehydrogenase	Lipid metabolism	Q/R	612
3057124	A						A	A		G	A			A	A			interg.	Modif.						
3076110	G						G	G		G	G			G	A			intrag.	Modif.	ML2575		pseudogene	Pseudogene		
3078037	A						A	A		a	A			A	A			intrag.	Modif.	ML2578	fadE4	pseudogene	Pseudogene		
3078039	C						C	C		a				C	C			intrag.	Modif.	ML2578	fadE4	pseudogene	Pseudogene		
3081021	C						C	C		T	C			C	C			syn.	Low	ML2580		integral membrane acyltransferase	Metabolism and respiration	Val103Val	384
3087916	G						G	G	G	G	G			T	G			intrag.	Modif.	ML2585		pseudogene	Pseudogene		
3100774	G									g								upst. 34 b.	Modif.	ML2595		Mce-associated membrane protein	Cell wall and cell processes		183
3100774	G									g								downst. 99 b.	Modif.	ML2594	mce1F	Mce-family protein	Virulee, detoxification, adaptation		517
3100774	G									g								interg.	Modif.						
3102778	A						C	C	C	C	C			C	C	C	C	miss.	Moder.	ML2597		Mce-associated membrane protein	Cell wall and cell processes	Tyr151Ser	184
3110722	G						G	G	A	G	G			G	G			intrag.	Modif.	ML2606	bglS	pseudogene	Pseudogene		

3145253	G						G	G		A	G			G	G			syn.	Low	ML2630		hypothetical protein	Unknown	Asn115Asn	123
3148092	G						G	G		G	A			G	G			miss.	Moder.	ML2631		oxidoreductase	Metabolism and respiration	Ala318Val	449
3155315	G						G	G		C	G			G	G			intrag.	Modif.	ML2637	fadE2	pseudogene	Pseudogene		
3159408	C						C	C		T	C			C	C			miss.	Moder.	ML2640		hypothetical protein	Unknown	Ala194Thr	310
3159732	C						t	T	T	C	C			T	C		T	miss.	Moder.	ML2640		hypothetical protein	Unknown	Val86Ile	310
3165254	G						G	G		G	A			G	G			intrag.	Modif.	ML2645		pseudogene	Pseudogene		
3175296	A						C	C	C	C	C			C	C	C		interg.	Modif.						
3179300	C						C	C	C	T	C			C	C			interg.	Modif.						
3183835	C						C	C		C	C			C	T			interg.	Modif.						
3192859	C						C	C		T	C			C	C			interg.	Modif.						
3196669	A						g	G	G	A	A			G	A			interg.	Modif.						
3204731	G						G	G		A	G			G	G	G		miss.	Moder.	ML2671	ctpI	cation transporter ATPase	Cell wall and cell processes	Asp8Asn	1609
3212353	G							G	G	G	T			G	G	G		syn.	Low	ML2672	icd2	isocitrate dehydrogenase [NADP]	Metabolism and respiration	Arg179Arg	746
3215535	G						G	G		A	G			G	G			interg.	Modif.						
3224915	G									G	t				G			miss.	Moder.	ML2678		hypothetical protein	Unknown	Trp838Cys	1000
3224919	T							t		T	G				T			miss.	Moder.	ML2678		hypothetical protein	Unknown	Trp840Gly	1000
3234644	G						g	G		G	G			G	G			intrag.	Modif.	ML2686		pseudogene	Pseudogene		
3237159	C						C	C		T	C			C	C			miss.	Moder.	ML2688	ponA	penicillin-binding protein	Cell wall and cell processes	Val543Ile	708
3243731	A						G	G	G	G	G			G	A	A		intrag.	Modif.	ML2694		pseudogene	Pseudogene		
3246728	G							G		G	G			G	A			upst. 54 b.	Modif.	ML2696		transcriptional regulatory protein	Reg. proteins		244
3246728	G							G		G	G			G	A			interg.	Modif.						

3256677	A						g	G	G	A	A			G	A		G	SPLICE_SITE	Low	ML2700		transmembrane protein	Cell wall and cell processes	Ter1207Ter	1206
3257070	G						G	G		G	G			G	A	A		intrag.	Modif.	ML2701	sigM	pseudogene	Pseudogene		
3258126	C						c	C	C	C	C			C	C			interg.	Modif.						
3259776	C						C	C		C				T	C	C		syn.	Low	ML2704		hydrolase	Metabolism and respiration	Cys6Cys	406
3261811	G						G	G	G	G	G			G	G	A		upst. 56 b.	Modif.	ML2705		hypothetical protein	Unknown		
3261811	G						G	G	G	G	G			G	G	A		interg.	Modif.						
3261877	G						G	G		A	G			G	G			interg.	Modif.						
3265736	G						G	G	G	A	G			G	G			miss.	Moder.	ML2710		transmembrane protein	Cell wall and cell processes	Pro351Leu	380
3268026	G						G	G		A	G			G	G			interg.	Modif.						



Supplementary Figure 6: M. leprae *variants distribution*
A) SNP distribution per annotated effect, B) SNP distribution per impact, C) SNP distribution per product functional category.

12.3.7 Genotyping results

Supplementary Table 15: SNP typing according to Singh et al, 2011

	Typing SNPs				Subtyping SNPs													
Sample	14676	1642875	2935685	Type	8453	313361	61425	1642875	3102778	1104232	2751783	2935685	1642875	1133945	2312066	413903	20910	Subtype
SJG 022	nc	nc	nc	nd	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SJG 131	nc	g	nc	1	nc	nc	nc	g	-	-	-	-	-	-	-	-	-	D
SJG 149	nc	nc	nc	nd	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SJG 189	C	T	nc	Und.	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SJG 289	nc	nc	nc	nd	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SJG 404	C	T	C	3	-	-	-	-	-	-	-	-	T	T	C	G	G	I
SJG 427	C	T	C	3	-	-	-	-	-	-	-	-	T	T	C	G	G	I
SJG 472	C	T	C	3	-	-	-	-	-	-	-	-	T	T	C	G	G	I
SJG 507	C	T	C	3	-	-	-	-	-	-	-	-	T	G	G	G	G	K
SJG 533	T	C	C	Und.	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SJG 611	nc	nc	nc	nd														-
SJG 658	C	g	nc	1	nc	nc	G	g	-	-	-	-	-	-	-	-	-	D
SJG 722	C	T	C	3	-	-	-	-	-	-	-	-	T	T	C	G	G	I
SJG 749	C	T	A	2	-	-	-	-	C	C	A	A	-	-	-	-	-	F
SJG 978	C	T	A	2	-	-	-	-	C	C	nc	A	-	-	-	-	-	F
SJG 1137	nc	T	C	3	-	-	-	-	-	-	-	-	T	T	C	G	G	I

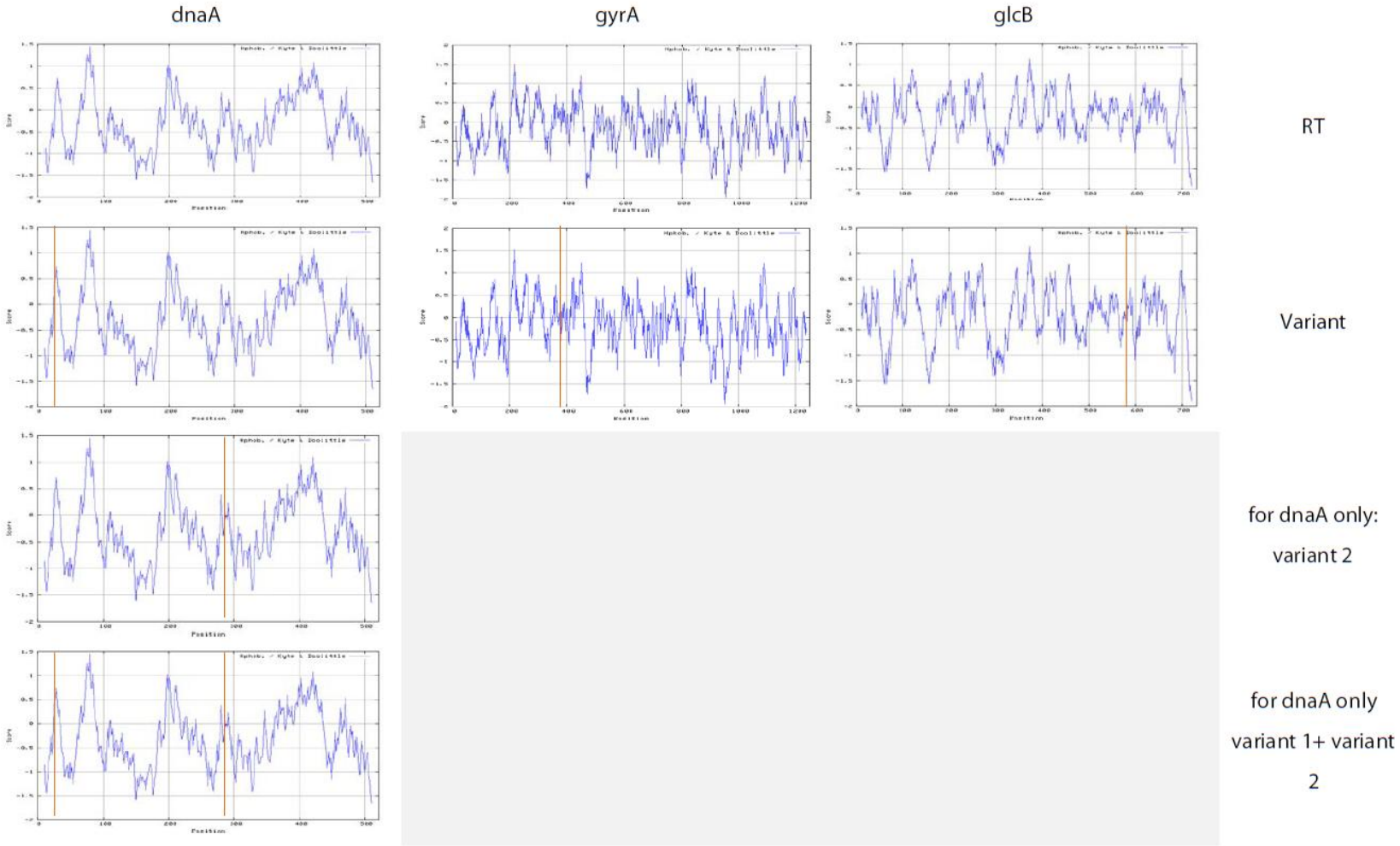
The upper-case bases were obtained from the highest-quality list of SNPS available. The lower-case bases were used when the base was not listed in the high-quality list of SNPs because it had been filtered out beforehand.

Supplementary Table 16: SNP typing according to Monot et Al, 2009

Sample		G022	G1137	G131	G149	G189	G289	G404	G427	G472	G507	G533	G611	G658	G722	G749	G978
Total SNPs		3	24	0	0	6	0	46	46	39	46	46		13	46	46	31
Type	1A	0	0	0	0	0	0	3	3	0	0	3		0	3	10	2
	1B	2	13	0	0	4	0	24	24	18	21	24		7	24	31	22
	1C	3	16	0	0	5	0	30	30	24	27	30		9	30	37	25
	1D	3	17	0	0	5	0	31	31	25	28	31		10	31	38	26
	2E	3	20	0	0	6	0	36	36	30	33	36		12	36	43	30
	2F	3	22	0	0	6	0	38	38	32	35	38		12	38	45	31
	2G	3	22	0	0	6	0	39	39	33	36	39		12	39	44	31
	2H	3	22	0	0	6	0	40	40	34	37	40		12	40	43	31
	3I	3	24	0	0	6	0	45	45	38	42	45		13	45	38	30
	3J	3	24	0	0	6	0	43	43	38	44	43		13	43	36	30
	3K	3	24	0	0	6	0	42	42	38	45	42		13	42	35	29
	3L	3	24	0	0	6	0	42	42	38	45	42		13	42	35	29
	4M	3	24	0	0	6	0	42	42	38	45	42		13	42	35	29
	4N	3	24	0	0	6	0	42	42	38	45	42		13	42	35	29
	4O	3	24	0	0	6	0	42	42	38	45	42		13	42	35	29
	4P	3	24	0	0	6	0	42	42	38	45	42		13	42	35	29
Most likely type		Und.	Und.	Und.	Und.	Und.	Und	3I	3I	Und.	Und.	3I		Und.	3I	2F	Und.

Out of the total 73 genotyping SNPs listed in the reference, the number of SNPs consistent with a type was calculated for each sample. In bracket is the total number of SNPs of the sample.

12.3.8 In-depth effect prediction results

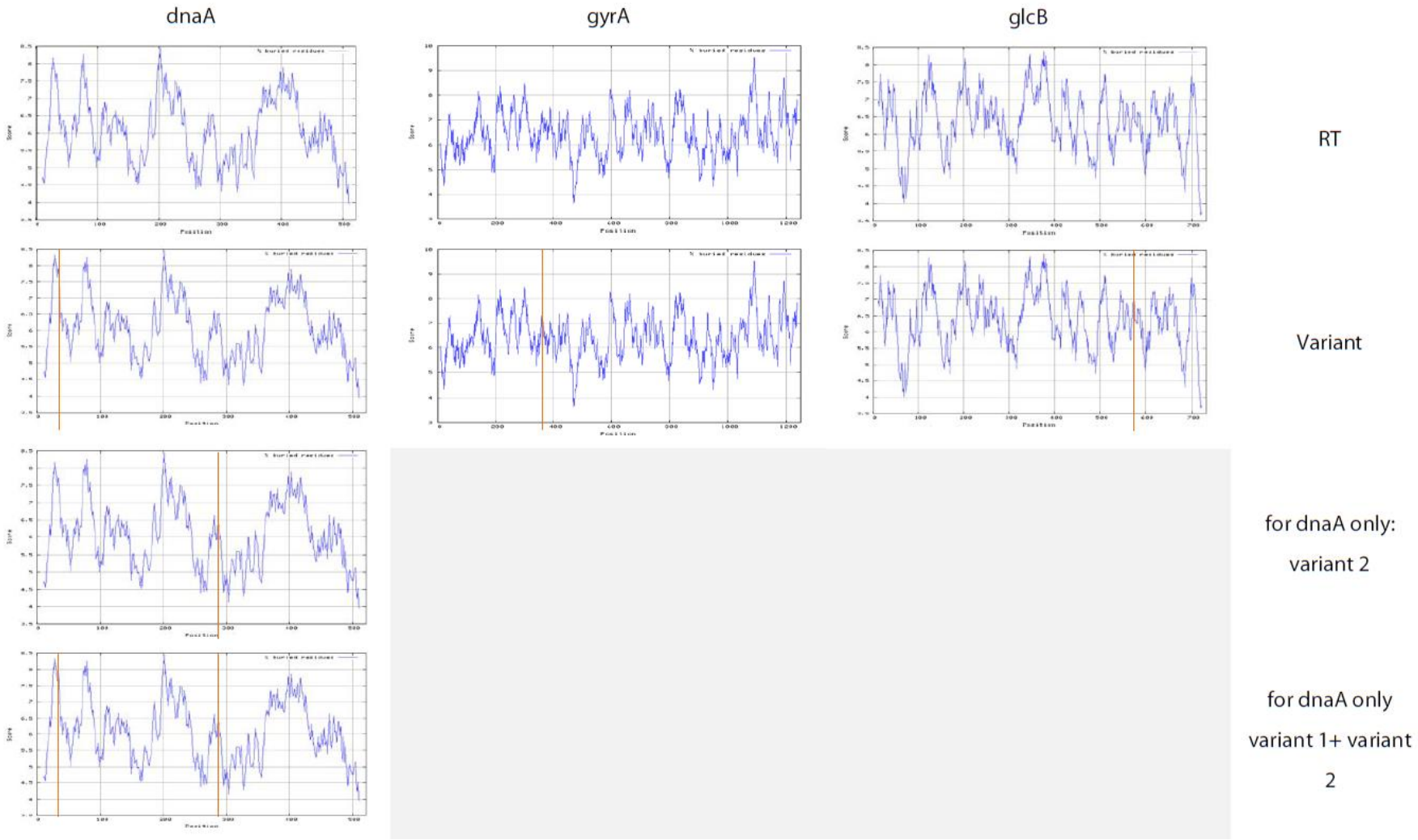


***Supplementary Figure** 7: Hydrophobicity profiles of the proteins*

The profiles were computed with ProtScale from the ExPASy platform using default settings over a 21bp sliding window to highlight changes at the chain scale. The amino acid change locus is indicated by the vertical bars.The amino acids are scored using the Kyte & Doolittle hydrophobicity scale.

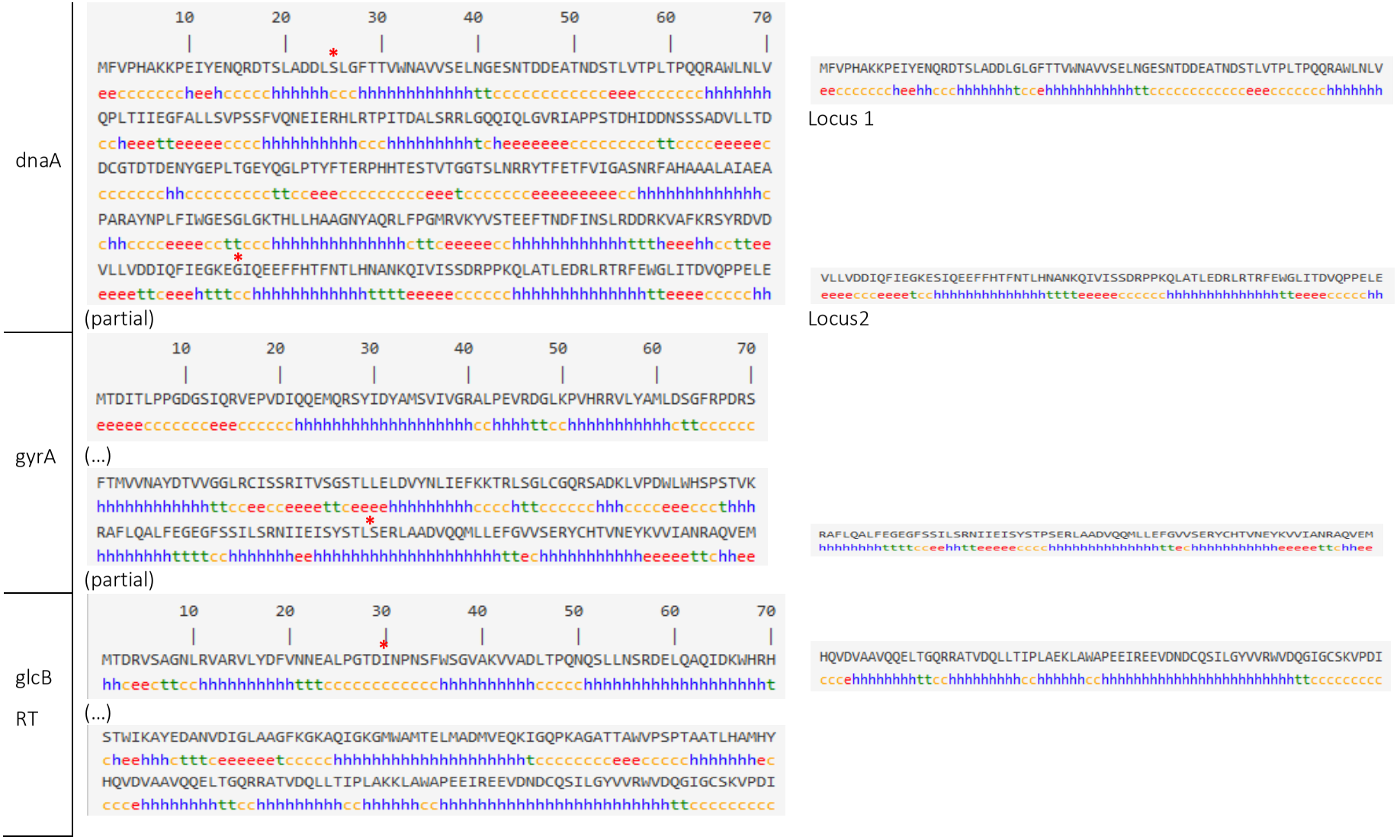
Supplementary Figure 8: Polarity profiles of the proteins

The profiles were computed with ProtScale from the ExPASy platform using default settings over a 21bp sliding window to highlight changes at the chain scale. The amino acid change locus is indicated by the vertical bars. The amino acids are scored using the Grantham polarity scale.



Supplementary Figure 9: Buried residues profiles of the proteins

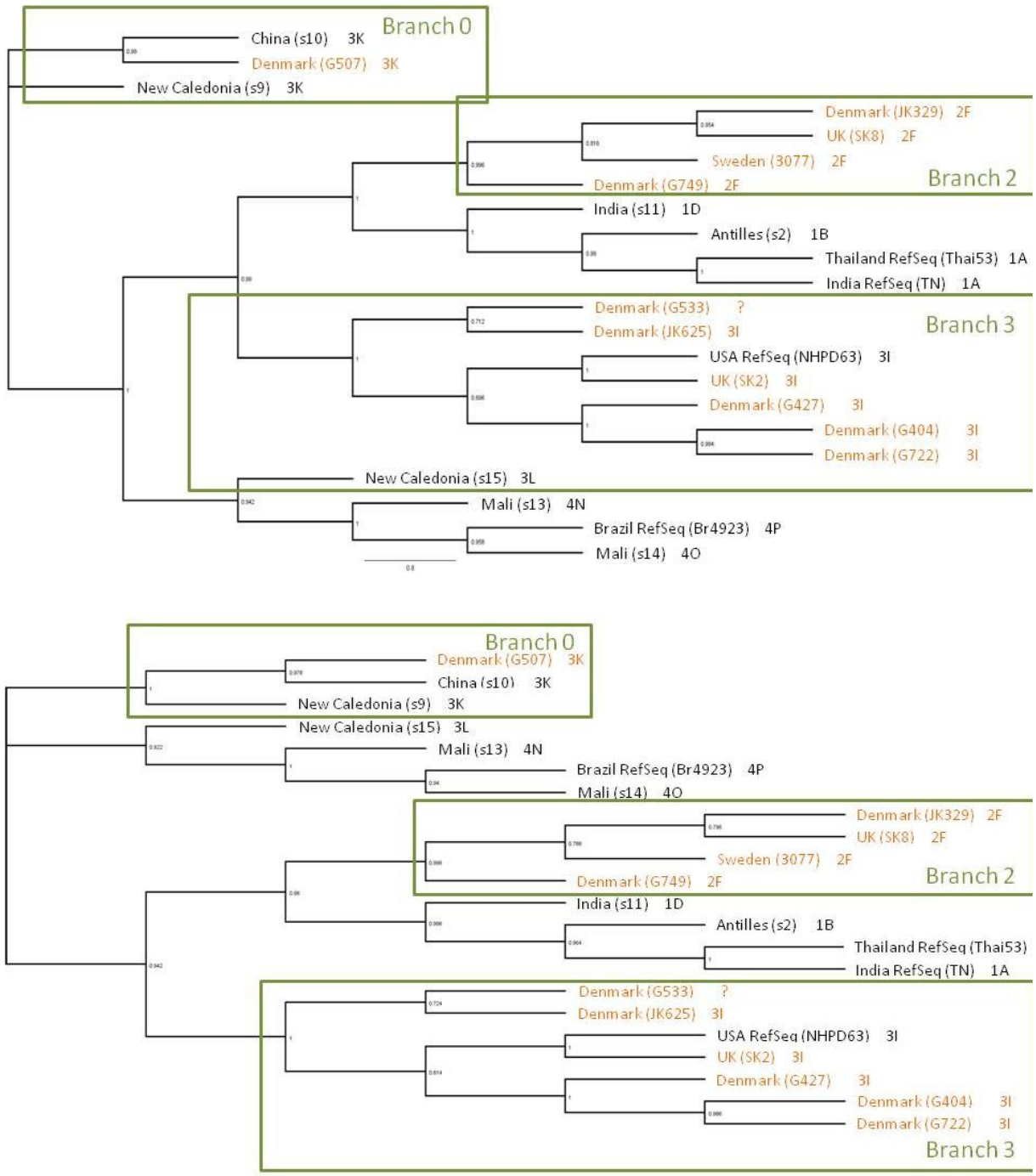
The profiles were computed with ProtScale from the ExPASy platform using default settings over a 21bp sliding window to highlight changes at the chain scale. The amino acid change locus is indicated by the vertical bars. The amino acids are scored using the Janin molar fraction of polar residues.



Supplementary Figure 10: Secondary structure predictions close-up on the variant loci

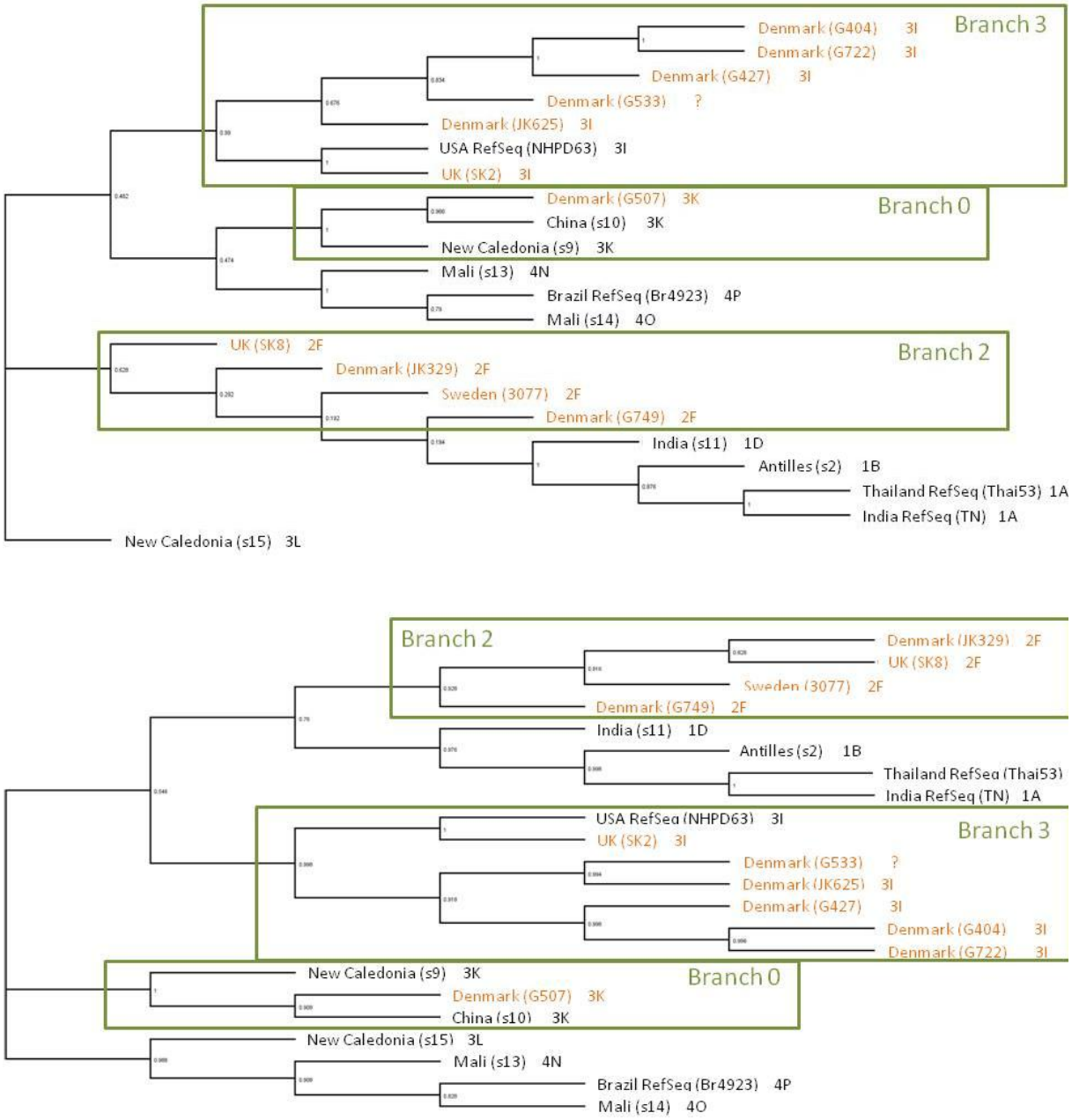
Predictions from the SOPMA software were performed using default settings. A star marks each locus position on the sequence. The variant sequence is shown on the right, with its corresponding secondary structure prediction.

12.3.9 Phylogenetic trees



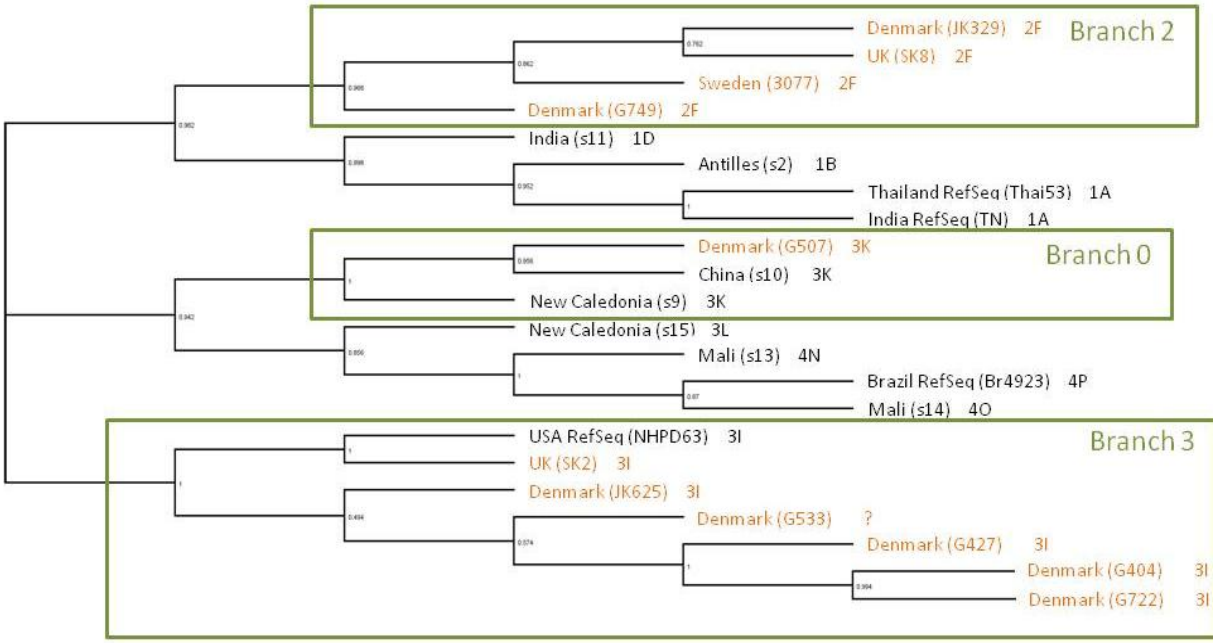
Supplementary Figure 11: Maximum-parsimony tree with the 6 high-coverage genomes

UP: tree constructed from the MAFFT alignment,, DOWN: tree constructed from the MAUVE alignment. Nodes labels indicate the bootstrap values after 500 replications. Ancient genomes are shown in orange, modern ones in black. The mention "RefSeq" is used to identify the four modern reference sequences.The scale reprensents the average character-state changes. Sample s15 shows a different affilitation, with the MAUVE alignment placing it closer to strains 3K, while MAFFT places it closer to strain 3I/4. This has also been noticed when the strain was first published at attributed to an unusual number of genetic variations (Schuenemann et al. 2013).



Supplementary Figure 12: Neighbor-joining tree with the 6 high-coverage genomes

UP: tree constructed from the MAFFT alignment,, DOWN: tree constructed from the MAUVE alignment. Nodes labels indicate the bootstrap values after 500 replications. Ancient genomes are shown in orange, modern ones in black. The mention "RefSeq" is used to identify the four modern reference sequences.



Supplementary Figure 13: Maximum-likelihood tree with the 6 high-coverage genomes

Tree constructed from the MAUVE alignment. Nodes labels indicate the bootstrap values after 500 replications. Ancient genomes are shown in orange, modern ones in black. The mention "RefSeq" is used to identify the four modern reference sequences.