Aus der Klinik für Innere Medizin
der Medizinischen Fakultät Charité – Universitätsmedizin Berlin und der Unité de recherche sur les Maladies infectieuses et tropicales émergentes Marseille

DISSERTATION

«Genotyping and Genomotyping of Tropheryma whipplei – The Causative Agent of Whipple's Disease»

zur Erlangung des akademischen Grades

Doctor medicinae (Dr. med.)

vorgelegt der Medizinischen Fakultät

Charité – Universitätsmedizin Berlin

von

Nils Wetzstein

aus Hanau

Datum der Promotion: 08.12.2017

Table of Contents

1. Abstracts....7

1.1. Abstract....7

1.2. Zusammenfassung....7

2. Introduction....9

2.1. Historical introduction....9

2.2. Epidemiology and transmission....9

2.3. Clinical manifestations....10

2.4. Antibiotic treatment....12

2.5. Diagnostics....12

2.6. Bacteriology and phylogeny....13

2.7. Bacterial typing systems....14

2.8. Geographical distribution and disease outcome....14

2.9. Bacterial genomics and application in TW....15

2.10. Bacterial core and pan-genomes....15

2.11. Problems addressed in this study....16

3. Materials & Methods....17

3.1. Utilized reagents and devices....17

3.2. Strains of TW....19

3.3. Axenic culture of TW....19

3.4. Harvesting....19

3.5. Verification of purity on cellular level....20

3.5.1. Classical staining methods....21

3.5.2. Immunofluorescence....21

3.6. DNA-extraction....21

3.7. Verification of purity on the DNA-level....22

3.8. Genotyping based on HVGS-regions....23

3.9. Epidemiological, phylogenetic and geographical analysis of typing data....25
3.10. Basic genomic data and sequencing....26
3.11. Genomic analysis....26
3.12. Geographic distribution of sequenced genomes....27
3.13. Calculation of core and pan-genome....28
4. Results....29
4.1. Patients' characteristics and clinics....29
4.2. Genotyping of remaining strains....29
4.3. Analysis of genotyping data in patient group including prior genotyping results....30
4.4. Sequenced genomes, genome length and GC-content....32
4.5. Geographic distribution of sequenced genomes and ANI distance matrix....32
4.6. Phylogenetic analysis of sequenced strains using genotyping data....35
4.7. Phylogenetic analysis of sequenced strains using whole-genomes....35
4.8. ANI-comparison of most important genotypes and different clinical entities....36
4.10. Comparison of TW with other intracellular pathogenic bacteria and other species in the group of Actinobacteria....37
4.11. Comparison of RAST annotation in TW strains....39
4.12. Prediction of antibiotic resistance genes and phage DNA....39
4.13. Core-genome and pan-genome....40
4.14. WiSPs in the core and pan-genome....41
5. Discussion....43
5.1. Patients' characteristics and clinics....43
5.2. Analysis of genotyping data....43
5.3. Sequenced genomes, genome length and GC-content....44
5.4. Geographic distribution of sequenced genomes....45
5.5. Phylogenetic analysis of sequenced strains....46
5.6. Core and pan-genome....46
5.7. Host and pathogen factors in WD....47

5.8. Prediction of antibiotic resistance genes: fluoroquinolone resistance gene.....48

5.9. Limitations of this study.....48

5.10. Conclusion.....48

6. Bibliography.....50

7. Eidesstattliche Erklärung.....55

8. Anteilserklärung an erfolgten Publikationen.....56

9. Curriculum Vitae.....57

10. Acknowledgments.....58

Abbreviations

AC	asymptomatic carriers / asymptomatic carriage
ANI	average nucleotide identity
APHM	Assistance Publique – Hôpitaux de Marseille
BCYE-Agar	buffered charcoal yeast extract agar
BDBH	bidirectional best hit algorithm
BLAST	basic local alignment search tool
CDC	Centers for Disease Control and Prevention
COG	cluster of orthologous groups
COS-Agar	Columbia agar with 5% sheep blood
CSF	cerebrospinal fluid
CWD	classical Whipple’s disease
DMEM F12	Dulbecco's modified Eagle medium F12
DNA	deoxyribonucleic acid
NW	neurological manifestations due to Whipple's disease
EW	endocarditis due to Whipple's disease
FBS	fetal bovine serum
Fig.	figure
Gt	genotype
HACEK	Haemophilus/Actinobacillus/Cardiobacterium hominis/Eikenella corrodens/Kingella
HIV	human immunodeficiency virus
HEL-Cells	Henrietta-Lacks-Cells
HGDI	Hunter-Gaston Discriminatory Index
HLA	human leucocyte antigen
HVGS	hyper variable genomic sequence
IP	isolated pulmonary affection of Whipple's disease
ITS	internal transcribed spacer region
MAFFT	Multiple Alignment using fast Fourier Transform
Mbp	megabasepairs
mM	millimole
MEM	minimum essential medium
MDM	monocyte-derived macrophage
MIC	minimal inhibitory concentration
MSA	multiple sequence alignment
NCBI	National Center for Biotechnology Information
NGS	next generation sequencing
NJM	neighbour-joining method
OMCL	OrthoMCL
N.A.	not applicable
OD	optical density
PAS-stain	periodic-acid Schiff-stain
PHAST	PHAge Search Tool
PCR	polymerase chain reaction
PBS	phosphate buffered saline
RAST	rapid annotation using subsystem technology
RNA	ribonucleic acid
SNP	single nucleotide polymorphism
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
TW	*Tropheryma whipplei*
UPGMA	unweighted pair group method with arithmetic mean

URMITE	Unité de recherche sur les maladies infectieuses et tropicales émergentes
WD	Whipple’s disease
WiSP	Wnt-inducible surface protein

1. Abstracts

1.1. Abstract

Whipple's disease (WD) is a very rare disease caused by *Tropheryma whipplei* (TW), an intracellular bacterium with a reduced genome. Its main clinical manifestations are classical Whipple disease (WD) with gastrointestinal affection, neurological manifestations (NW), endocarditis (EW) and isolated pulmonary affection (IP). An asymptomatic carriage (AC) occurs frequently.

In the present study, 17 strains of TW have been successfully cultured, their DNA extracted and their genomes sequenced by SOLiD-technology. They were set into context with the two reference strains *Twist* and *TW08/27* by whole-genome comparison. Genotyping data according to the HVGS genotyping system for our study group but also for all European samples was examined.

Two predominant genotypes (Gt1 and Gt3) could be described in France and Germany in our group as in all TW genotyping data. Typing resolution could be increased by implementing genomotypes. Hereby, no significant clustering of identical clinical manifestations could be observed. The ANI did not decrease with increasing geographical distance. But same genotypes showed higher ANIs in all of their genomes.

The pan-genome of TW could be constructed. As other intracellular bacteria, TW shows a closed pan-genome, with a reduced genome mirroring a sympatric life style and a putative close relationship to its human host. This underlines the theory that TW is a commensal of the gut and causes disease only in distinct situations with a certain immunological disorder.

1.2. Zusammenfassung

Morbus Whipple ist eine sehr seltene Ekrankung, die durch *Tropheryma whipplei* (TW) verursacht wird. Hierbei handelt es sich um ein intrazelluläres Bakterium mit einem reduzierten Genom. Die hauptsächlichen klinischen Manifestationen der Erkrankung sind klassischer Morbus Whipple (CWD) mit gastrointestinaler Symptomatik, neurologische Manifestationen durch TW (NW), eine Endokarditis durch TW (EW) und eine isoliert pulmonale Beteiligung (IP). Häufig kommt aber auch eine asymptomatische Besiedlung mit dem Erreger vor (AC) .

Im Rahmen dieser Arbeit wurden 17 TW-Stämme erfolgreich kultiviert und ihre Genome mittels SOLiD-Technik sequenziert. Diese wurden mit den Referenzstämmen, *Twist* und *TW08/27*, verglichen und mit vorherigen Genotypisierungsdaten kontextualisiert.

Es zeigten sich in Deutschland und Frankreich zwei prädominante Genotypen (Gt1 und Gt3). Die Typisierungsauflösung konnte mittels Implementierung von Genomotypen erhöht werden. Hierbei waren jedoch keine signifikanten Gruppierungen anhand der klinischen Manifestationen zu verzeichnet. Die durchschnittliche Nukleotididentität (ANI) verhielt sich nicht antipropotional zur geographischen Distanz, jedoch zeigten sich für gleiche Genotypen signifikant höhere ANIs.

Ebenso wurde das Pangenom von TW konstruiert. Wie andere intrazelluläre Erreger hat TW ein geschlossenes Pangenom. Dies kann als Hinweis auf ein enge Beziehung zum menschlichen Wirtsorganismus in der Evolutionsgeschichte gewertet werden und unterstreicht die Theorie, dass TW als Kommensale lebt und nur bei Menschen mit einer bestimmten Immunschwäche Erkrankungen hervorruft.

2. Introduction

2.1. Historical introduction

Whipple's disease (WD) is a rare infectious disease that can affect virtually every organ of the body. Its causal agent is *Tropheryma whipplei* (TW), a bacterium phylogenetically belonging to the G-C-rich gram positive Actinobacteria [1]. WD was first described by George Hoyt Whipple in 1907. He named the disease "intestinal lipodystrophy" because of masses of fat and fatty acids typically found in duodenal biopsies [2].

Long assumed to be bacterial, its origin was not proven until 1961 by electron microscopy that could show bacterial inclusions in macrophages [3], which along with monocytes are the main cells infected in this disease (Fig. 1). Cultivation of this fastidious bacterium remained difficult and was not achieved until the year 2000, when Raoult et al. established the first culture of TW in a human fibroblast cell line (HEL-cells) [4].

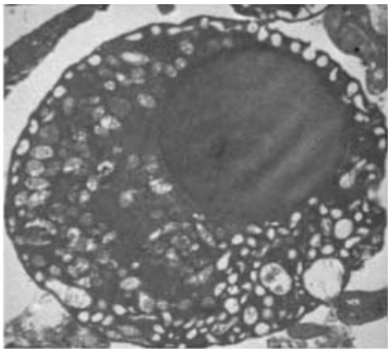

Fig. 1: Transmission electron microscopy picture of a TW-infected monocyte-derived macrophage (MDM). TW rods are visible in bacterial inclusions inside the cell. Photograph courtesy of Gorvel et al.

The cultivation of TW made possible sequencing of its genome in 2003 by two groups, Bentley et al. and Raoult et al. (reference strains *TW08/27* and *Twist;* accession numbers: BX072543 and AE014184) [5,6] which led to the genome-based design of an axenic culture medium [7], making cultivation much easier than in the past, thus simplifying further investigations.

2.2. Epidemiology and transmission

WD affects mainly middle-aged Caucasian men and cases have therefore been described primarily in Europe and the United States [8]. Being a very rare disease, its prevalence is supposed to lie around 1/1,000,000 in these countries and until now, more than 1000 cases have been

described in the literature [8,9]. Nonetheless, exact epidemiological estimations are difficult, given the low incidence of the disease.

4 % of stool samples collected from the general French population has been tested positive for TW DNA, whereas 48 % of the general population has antibodies against TW surface proteins [10]. Thus TW appears to be a highly ubiquitous bacterium and some authors consider it even to be a commensal [11]. Its ecological niche remains unknown, though it has been found in sewage waters, both in Germany and in France, as well as in sewage treatment plants in both countries [12]. Water samples of two highly endemic regions in Senegal (villages of Dielmo and Ndiop) tested by polymerase chain reaction (PCR) showed no traces of TW DNA [13], while 44 % of tested children had TW DNA in their stools and overall seroprevalence was estimated at 72.8% [14]. So probably humans are the main reservoir of the bacterium.

Mode of transmission and natural habitat of this pathogen still remain unknown. The supposed feco-oral transmission route could not be confirmed until now, but appears highly probable, as higher incidence of the same circulating genotype in the same family stresses a possible interhuman transmission of the bacterium, but not of the disease [10].

2.3. Clinical manifestations

WD can be divided into four clinical entities: 1) classical Whipple's disease (CWD), 2) isolated infections, mainly endocarditis (EW) and neurological manifestations due to TW (NW), 3) acute infections, such as gastroenteritis in children, pneumonia (IP) or bacteremia and 4) asymptomatic carriage in healthy subjects (AC) [15].

1) Classical Whipple's disease (CWD)

CWD affects mainly the human gastrointestinal tract. Clinically, a malabsorption syndrome and diarrhea are observed. These symptoms are often secondary to diffuse joint pains that typically precede bowel symptoms by up to six years [16]. Arthralgia is one of the most frequent symptoms in CWD-patients. Weight loss is also very common and found in 79 to 92 % of all cases [16,17]**.**

2) Isolated infections

Endocarditis (EW): TW is able to cause endocarditis and has to be taken into account if a patient is suffering from culture-negative endocarditis. This is the second most frequent clinical manifestation caused by TW [16]. A study conducted in four German hospitals found TW to be the most common pathogen responsible for infections of cardiac valves in which bacterial cultures

remained negative, thus even outnumbering the HACEK group [18]**.** Additionally, TW can be the cause of constrictive pericarditis [19].

Neurological manifestations (NW): TW may infect the central nervous system by entering it via macrophages and causing variable neurological symptoms [16]. Neurological involvement appears secondary to WD or it is isolated as a primary infection. Case reports have shown that intracranial infection by TW can result in various neurological and neuropsychological symptoms. Even hyperphagia and weight gain have been thought to be caused by TW [20]. As in other focal cerebral infections, neurological symptoms vary depending on the localization of a lesion [21].

3) Acute infections

Gastroenteritis: In a French hospital, an outbreak of gastroenteritis in children between 2 and 4 years has been observed. This epidemic outbreak was probably caused by a clonal strain of TW (Gt 3) that has been found in 10 out of 34 samples. TW might have acted as a co-pathogen during this outbreak [22]**.** Thus, primo-infection in humans might take place during childhood and cause gastrointestinal symptoms. Following this primo-infection, WD might follow much later in adulthood in certain predisposed individuals.

Pulmonary (IP): TW has been shown to be responsible for pneumonia and other pulmonary affections. Route of infection might be the aspiration of TW issued from the gastro-intestinal tract [23]. Interestingly, TW was found to be the predominant pathogen in the pulmonary microbiome of HIV-positive patients, but no correlation between the immune deficiency caused by HIV and WD has been described so far [24,25].

4) Asymptomatic carriage (AC)

Different epidemiological studies have shown the occurrence of TW in stools of asymptomatic carriers (AC) [12,14]. These comprise children and adults in Senegal as well as different populations in Europe. Sewage workers in Marseilles and around Vienna have been shown to have a higher prevalence of TW carriage in their stools than a control population (12% in comparison to 2-4%) [12,26].

The existence of AC, acute infections not preceding CWD, and the fact that the disease is mainly found in Caucasian white men led to the assumption that there might be immunological host factors involved in the pathogenesis of WD [10,27,27–32]. Accordingly, a number of polymorphisms

of genes involved in the presentation of antigens or in the establishment of inflammatory immune reactions are associated with WD [33–35]. For example, the HLA genotype seems to influence the course of the infection [33].

2.4. Antibiotic treatment

Antibiotic treatment of WD remains controversial. While some propose even lifelong antibiotic treatment with tetracyclines - namely doxycycline - to prevent relapses [36,37], there is evidence that a 14-day course of treatment with a third-generation cephalosporin (e.g. ceftriaxone) followed by three or twelve months of treatment with trimethoprim-sulfamethoxazole might be sufficient in the treatment of WD [38,39]. Although *in vitro* studies have shown that all TW strains are resistant to the trimethoprim component, and some strains to sulfonamides as well [40].

2.5. Diagnostics

Histology: Since its first description in 1907, WD is diagnosed by duodenal biopsies that show foamy macrophages with PAS-positive inclusions intruding the *lamina propria*. This can be considered the hallmark of CWD [41]. Positive PAS-staining is caused by a glycosylated protein on the bacterium's surface [41]. However, negative PAS-staining of duodenal specimens does not rule out isolated WD with other manifestations, such as for example NW.

Molecular biology: A universal 16 PCR, sequencing and subsequent search in the BLAST-Database, can identify TW-DNA. This technique can be used for various specimens, such as CSF, stool-samples, cardiac valves and others [42]**.** Due to AC, positive results from gastrointestinal specimens have to be interpreted cautiously. As in every other bacterial infection, the presence of bacterial DNA does not prove that there is an active infection. A recent study has shown a novel rpoB-assay to be very sensitive, facilitating the screening of Whipple's disease in clinical specimens [43].

TW-specific primers are available as well and permit the specific detection of the causative agent[44]**.** For epidemiological purposes, specimens can be typed by the TW-specific genotyping system using HVGS-genotypes (see below) [45].

Microbiological culture: Culture can be achieved in specialized laboratories (for example at the URMITE, Marseilles, France). Therefore, TW-strains are cultured on cell cultures and in axenic medium. At the present, this technique is not applicable for daily routine diagnostics, it

postpones clinical decisions and is used primarily for research purposes, as the bacterium has very slow growth and still remains difficult to culture in some cases [4,7].

2.6. Bacteriology and phylogeny

TW is an intracellular bacterium which could be classified into the GC-rich Actinobacteria clade[1]. Unlike other bacteria in this clade, GC-content of its genome is very low (strains *Twist* and *TW08/27*: 46.3%) [6]. It is surrounded by a trilamellar cell membrane (Fig. 2) whose outermost layer is composed of various glycoproteins [1]. Genomic sequencing revealed a reduced genome (0.93 mega base pairs, Mbp) - like in other intracellular living bacteria - deficient of certain pathways in energy metabolism, notably in amino acid metabolism. TW can thus be regarded as very dependent on its host cell [5,6]. This is depicted by the formation of the glycopeptide layer as well, which depends mainly on carbohydrate metabolism enzymes of the host cell and is lost in axenic culture after repeated passages, whereas in cell cultures it remains present [46].

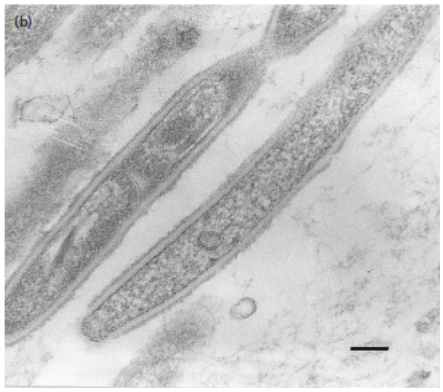

Fig. 2: Transmission electron microscopy image showing the trilamellar cell-membrane of TW and its rod-like structure. One cell is dividing longitudinally. This is believed to be the reason for the typical rope-like structures that are observed in microscopy of TW. Bar length is 100 nm. This picture was taken with friendly permission from LaScola et al. [1].

Staining behavior of TW is variable: Gram stainings appear gram-negative and poorly stained. Gimenez staining colors rods in pale-pink, Ziehl-Neelsen staining does not color TW. In axenic culture, two morphologies are observed: firstly, specific rope-like structures (Fig. 3), that might be caused by the longitudinal replication of TW, and secondly, cellular aggregates, that can be observed after several passages, and are possibly linked to the loss of the glycopeptide layer of TW [1].

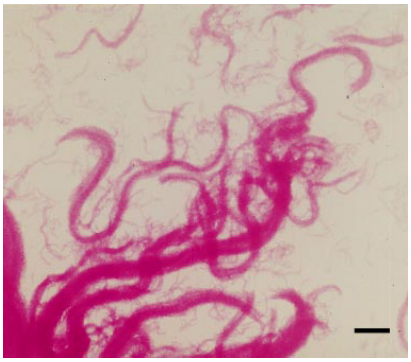

Fig. 3: PAS staining of TW showing rope-like structures being specific for this bacterium, probably caused by longitudinal cell division. Bar length is 15 µm. This photograph was taken with friendly permission from LaScola et al.

2.7. Bacterial typing systems

Genotyping in TW has undergone various changes over time. First typing methods consisted of 16s-typing which permitted its phylogenetic classification into the Actinobacteria clade as noted above [47,48]. This was confirmed by rpoB-sequence analysis [49]. Sequencing of two genomes in 2003 (*TW08/27* and *Twist*) allowed the implementation of another genotyping system based on highly variable genetic regions (HVGS, namely TW133, ProS, SecA, Pro184). This system showed higher discrimination power and found more different genotypes than previous systems, thus increasing "typing resolution" inside the species [45]. It has since been applied to various samples from central Europe (France, Germany, Austria, Switzerland, Italy and Belgium) and sub-Saharan Western Africa (Senegal) and showed a high genetic diversity of TW-specimens.

2.8. Geographical distribution and disease outcome

Until 09/2011, about 300 specimens of TW's have been genotyped in URMITE, Marseilles. Most of them in Europe and some of them in Western Africa (Senegal). Until now, no real pattern in the geographical distribution of genotypes could be observed. But genotypes found in Senegal are manifestly distinct from the ones found in Europe, and European strains constitute a separate cluster when compared to African strains [13].

The correlation between virulence factors, genetic information of the bacterium and disease outcome still remains unclear. It has been shown that there is none between clinics and

genotypes. But the existence of other genetic regions that mirror possible virulence factors – not depicted by the actual typing system – remains another possibility.

2.9. Bacterial genomics and application in TW

Genome sequencing of the two reference strains, *Twist* and *TW08/27*, has demonstrated that TW - as other fastidious, intracellular bacteria - has a reduced genome (0.92 Mbp) comprising just basic metabolic functions [5,6]. The organism is deficient in certain metabolic functions, as for example in amino acid metabolism [7]. Prior studies already addressed intra-species-variation between different strains of TW using microarray-based comparative genomic hybridization. They found a maximum 2.24 % difference in hybridization between strains. The main variability could be demonstrated in genome regions coding for certain surface proteins, the WiSP-family (Wnt-inducible Surface Proteins) [50]. Thus, WiSPs have been assumed to be important putative virulence factors. They might also play an important role in the pathogen-host interaction in this intracellular bacterium.

2.10. Bacterial core and pan-genomes

The past twenty years have seen an immense increase in DNA sequencing velocity, thus increasing output of sequencing data and facilitating high resolution genomic comparisons in eukaryotes and prokaryotes. In the past, bacterial species were defined by morphological features, then by DNA-hybridization tests and finally by 16s RNA sequences. Because more and more complete bacterial genomes are available, the new sequence data leads to a discussion about species borders and allows a new definition of bacterial species.

In order to give answers to this discussion, Tettelin et al. introduced the concept of the bacterial core and pan-genome [51]. Thus, a species can be defined by the genes that are shared or not shared by all strains. The « core-genome » defines the totality of all genes that all the genomes of a species have in common. On the other hand, there are so-called accessory genes that only some of the strains bear in their genomes. The sum of all genes hitherto described in a species can be defined as the « pan-genome » [51]. Some bacteria seem to have a so called “closed” pan-genome which means that there are very few accessory genes and there is only a small difference between the core and the pan-genome. Others have an “open” pan-genome, with a large amount of accessory genes.

Whether a bacterial species has an open or closed pan-genome, can give supportive information about its evolutionary origin and mode of living. Recent studies have shown that so called sympatric living organisms tend to have a closed pan-genome, whereas species that live allopatrically, for example, environmental bacteria, seem to have open pan-genomes [52].

2.11. Problems addressed in this study

Prior genotyping studies have shown that there is no correlation between geographical origin, clinics and genotype of TW [45]. Possible virulence factors could not be found. Genomic and proteomic studies found a striking similarity between the different strains of TW, while the main differences were described in so called WiSPs. Those have then been assumed to play a role in pathogenesis of WD [50,53]. Thus, the main objectives of this study were :

- to evaluate the hitherto used typing system

- to perform another geographic evaluation of new and existing typing data

- to enable a more thorough comparison of the available cultured strains of TW by whole-genome sequencing, thus reevaluating the actual typing system and increasing typing resolution

- and finally to elucidate once again the high genetic diversity in WiSPs.

Next generation sequencing (NGS) makes it possible to analyze genomes very exactly and to describe single nucleotide polymorphisms (SNP) on a large scale. Thus, the genomes that have already been compared by hybridization assays should be compared at an even higher resolution [50]. This study tried to answer once again the question if there is a relationship between genetic information of TW, clinical manifestation and geographic distribution. Therefore, phylogenetic questions should be addressed by genotyping and whole-genome phylogeny, multiple alignments, genome to genome comparisons, and phylogenetic tree building. In addition the geographic data should be correlated to the genomic data.

Furthermore, the genomic data was used to predict possible antibiotic resistances and putative virulence factors. Finally, the core and pan-genome of TW should be examined in order to trace the evolutionary path of TW.

3. Materials & Methods

3.1. Utilized reagents and devices

Table 1: Table depicting reagents, devices and their origins utilized during this study.

Material	Origin
Bacterial culture	
BCYE-Agar	Oxoid, Wesel, Germany
COS-Agar	Biomérieux SA, Craponne, France
DMEM F12	Gibco life technologies, Paisley, UK
Fetal Bovine Serum (FBS), Qualified, Heat inactivated	Gibco life technologies, Paisley, UK
Minimum Essential Medium (MEM) 100x non-essential amino-acids	Invitrogen, Lonza Verviers, Belgium
L-Glutamine (200 mM)	Gibco life technologies, Paisley, UK
Coloring and harvesting	
PBS	Biomérieux SA, Craponne, France
PBS Tween	Gibco life technologies, Paisley, UK
Methanol	Sigma Aldrich, St- Louis, USA
Gimenez staining	URMITE, Marseilles, France
Gram staining	Biomérieux SA, Craponne, France
Acridine orange	URMITE, Marseilles, France
Rabbit antibodies	URMITE, Marseilles, France
Fluoroprep	URMITE, Marseilles, France
Falcon tubes	Becton Dickinson Labware, Franklin Lakes, USA
DNA-extraction	
Proteinase K (> 600mAU/ml)	Qiagen, Hilden, Germany
Buffer AL	Qiagen, Hilden, Germany
QIAamp-DNA Minikit	Qiagen, Hilden, Germany
UltraPure Agarose	Invitrogen, Madrid, Spain
Genotyping and sequencing	
Sequencing mix	Applied Biosystems, Foster City, USA
HotStar Taq Mastermix-Kit	Qiagen SA, Hilden, Germany
Big dye terminator, standard	Gibco life technologies, Paisley, USA
DNA-ase free water	Gibco life technologies, Paisley, USA
MultiScreen plates	Merck Millipore, Billerica, USA
Sephadex G50	Sigma Aldrich, St. Louis, USA
Bdv1 Buffer	Applied Biosystems, Foster City, USA
Nucleo Fast 96 Plate	Macherey Nagel, Düren, Germany
Primers for PCR	
536 F	5’ CAG CAG CCG CGG TAA TAC 3’
Rp2	5’ ACG GCT ACC TTG TTA CGA CTT 3’
800 F	5' TAG ATA TAC CCG GTT AG 3'
1050 R	5' CAC GAG CTG ACG ACA 3'
ITS1	5' TCC GTA GGT GAA CCT GCG G 3'
ITS4	5' TCC TCC GCT TAT TGA TAT GC 3'
HVGS1: TW 133	5’ GCT GCG CGA AGT AAT TTG 3’
HVGS2: ProS	5’ GCC TTG ACT ATG ACA TAA TCA A 3’
HVGS3: SecA	5’ TTT GTC ATA GGC ATT TCT GTA G 3’
HVGS4: 184	5’ CGG ATC TTC ACG AAA TGT CC 3’

Table 1: Table depicting reagents, devices and their origins utilized during this study.

Material	Origin
Software	
Chromas-Pro Version 1.5	Technelysium, South Brisbane, Australia
Libre Office Calc	The Document Foundation
Libre Office Draw	The Document Foundation
Libre Office Writer	The Document Foundation
Epi Info 7	Centers for Disease Control and Prevention, Atlanta, USA
Fig Tree Software v 1.4.	Andrew Rambaut, Edinburgh, UK
Galaxy Project	http://usegalaxy.org
Geographic Distance Matrix Generator v.1.2.3.	Ersts,P.J., American Museum of Natural History, Center for Biodiversity and Conservation.
JspeciesWS	Ribocon GmbH, Bremen, Germany
MEGA 6	Kumar S, Stecher G, and Tamura K
MAFFT	Kazutaka Katoh, Kyoto, Japan
RAST-Server	http://rast.nmpdr.org/
MAUVE	Genome Center and Department of Computer Science, University of Wisconsin, Madison, Wisconsin, United States of America
Get_homologues	Contreras-Moreira B, Vinuesa P
PHAST	Zhou Y, Liang Y, Lynch K, Dennis JJ, Wishart DS
ResFinder	https://cge.cbs.dtu.dk/services/ResFinder/
Technical devices	
2720 Thermal Cycler	Applied Biosystems, Foster City, USA
ABI 3130 Genetic analyzer	Applied Biosystems, Foster City, USA
GENios platform	Tecan Group, Männedorf, Switzerland
Nanodrop 1000	Thermo scientific, Wilmington, USA
Electrophoresis apparatus	Rapid one advance
Eppendorf Biophotometer	Eppendorf AG, Hamburg, Deutschland
MP FastPrep 24	MP Biomedicals, Solon, USA
Laser confocal Fluorescence Microscope	Leica, Lyon, France
Nikon Eclipse E400 light microscope	Nikon Corporation, Tokyo, Japan
Cytospin 4	Thermo scientific, Waltham, USA
GR 422 Centrifuge	Jouan Group, Saint Herblain, France

3.2. Strains of TW

In total, 29 strains of *TW* were available at URMITE, Marseilles, at the time of this study (09/2011). All of them were kept in continuous subculture, the reference strain *Twist* being the one with the longest time of cultivation (since 2000) [1]. Strains were taken from different patients and from different specimens (CSF, small intestine biopsies, bronchoalveolar lavage, heparinized blood, heart valves, lymph nodes, muscles, synovial fluid and feces). Names for strains were chosen according to their origin and numbers according to the date of their sampling. Table 2 gives a summary of origin and nature of the strains used in this study. Informed consent has been given by every patient for further studies with their material. For German patients, the further usage of strains was permitted by the « Ethikkomission der Charité » (Ethikantrag EA4/122/10). For French patients, further usage was approved by the local ethics committee. No tests on animals were involved in this study.

3.3. Axenic culture of TW

TW cultures of less than 15 passages were inoculated into 50 ml of D10F12 medium (88% DMEMF 12, 1% 100x diluted MEM non-essential amino acids, 1% L-glutamine) in big ventilated flasks. For verification of bacterial contamination, COS-agar-plates, as well as BCYE-plates for fungal contamination were inoculated and checked visually at day 3 and day 7, respectively. Because TW does not grow on classical bacterial media, colonies were evaluated as contamination and the corresponding bacterial culture therefore discarded and restarted.

The flasks were incubated for 8 days (37°C, 5% CO2). On day 8, a microscopy check was conducted using Gimenez-Coloration (Basic fuchsin, phenol and ethanol, malachite green, own production URMITE, Marseilles, France), and optical density was determined using a biophotometer. If agar-plates showed no colonies and bacterial staining showed significant bacterial growth, cultivation was continued for another 10 days after adding 300ml of fresh D10F12 medium to each flask.

3.4. Harvesting

If optical density lay in the range between 0.3 and 0.8, controls were negative and microscopy showed a rich culture, the cultures were harvested. In order to achieve sufficient DNA yields for subsequent sequencing, aliquots of culture liquid were put into 50 ml Falcon tubes (Becton

Dickinson labware, Franklin lakes, USA) and centrifuged for 10 minutes at 7500 rpm, in total six times, in a GR 422 Centrifuge. The resulting pellet was washed two times in PBS.

Table 2: Summary of all strains used during this study as well as the specimen they were taken from, the patients' initials, their sex and the date on which the sample was taken. All these strains are still in continuous subculture at URMITE, Marseilles, France.

Strain	Initials	Patient's sex	Geographic origin	Date of sampling	Kind of specimen
TWIST	TW	M	Canada	22/05/98	Aortic valve
SLOW2	AJ	F	France	07/05/97	Small intestine biopsy
ENDO5	PD	M	France	22/03/02	Heparinized blood
NEURO1	KT	M	Germany	10/03/02	CSF
NEURO2	PL	F	France	19/05/04	CSF
ENDO7B	GJ	M	Portugal	10/12/04	Heparinized blood
DIG7	JD	M	France	10/12/04	Heparinized blood
DIG9	EL	M	France	10/02/02	Heparinized blood
DIG10	LB	M	Germany	20/04/05	CSF
DIGADP11	JR	n.a.	France	22/06/05	Mesenterial lymph node
ART1	MG	M	France	07/07/05	Synovial fluid
DIGNEURO14	FW	M	Germany	03/08/05	CSF
DIG15	AR	M	Germany	13/09/05	CSF
DIGMUSC17	TD	n.a.	France	16/11/05	Muscle
SLOW1B	MD	F	France	02/02/05	Feces
DIGNEURO18	SE	M	France	21/12/05	CSF
ENDO19	JG	M	Germany	11/07/06	Cardiac valve
NEURO20	BM	M	Germany	16/05/06	CSF
NEURO21	KB	M	Germany	11/07/06	CSF
DIGNEURO23	CA	F	France	28/09/06	CSF
ENDO24	GR	M	France	24/04/07	Mitral valve
DIGADP25	RR	M	France	29/01/09	Mesenterial lymph node
TWBCU26	GF	F	France	12/02/09	Cutaneous biopsy
ENDO27	PM	M	France	09/06/10	Aortic valve
SALI28	MS	M	France	01/07/10	Saliva
ART29	GT	M	France	09/09/10	Synovial fluid
PNEUMO30	ND	F	France	28/09/10	Bronchoalveolar lavage
ART31	DH	M	France	05/01/11	Synovial fluid
ENDO32	KM	M	France	12/05/11	Aortic valve

3.5. Verification of purity on cellular level

Before DNA-extraction was conducted, purity of cultures was tested by different means: classical staining methods (Gram- and Gimenez-staining), Immunostaining and staining with acridine orange (own production URMITE, Marseilles, France), in order to evaluate if bacteria were vital and therefore would lead to sufficient DNA yields (Fig. 4).

For all staining methods, 200 µl of bacterial culture was put into a Cytospin 4 device and centrifuged for 5 min at 1500 rpm in order to fix bacteria to microscopy slides and increase bacterial concentration for subsequent microscopy.

3.5.1. Classical staining methods

For Gram staining slides were colored according to the manufacturer's recommendations and then dried at room temperature. Gimenez stainings were conducted by incubating with carbol fuchsin solution for 3 minutes and 1% malachite green for 10 seconds, respectively. Slides were examined using a Nikon Eclipse E400 light microscope at a 1000x magnification.

3.5.2. Immunofluorescence

After Cytospin treatment, slides were fixed for ten minutes in methanol. 300 µl of rabbit anti-TW antibodies (own production, URMITE, Marseilles, France) were deposited on each spot and incubated at 37 °C in a humid chamber for 30 minutes, then washed twice: 10 minutes in PBS Tween and 10 minutes in PBS, and dried at room temperature. In a second step, 300 µl of fluorescent anti-rabbit antibodies (own production, URMITE, Marseilles) were put onto each spot and subjected to the same procedure. Slides were prepared for microscopy with Fluoroprep solution. Microscopy device was a laser confocal fluorescence microscope equipped with a ×100 oil immersion lens (Fig. 4).

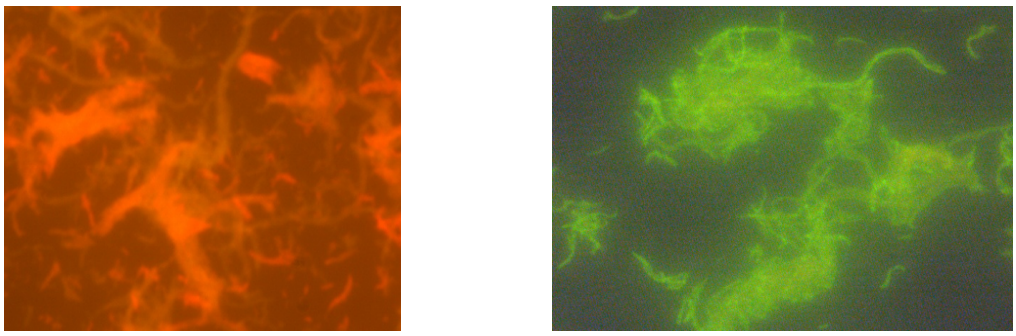

Fig. 4: *Left:* Acridine orange staining of TW in axenic culture. 1000X magnification. Orange stained bacteria are viable. Rope-like structures of TW are visible. *Right*: Fluorescent microscopy of TW from axenic culture. Bacteria show typical rope-like structures. Shot at a 1000x magnification.

3.6. DNA-extraction

250 ml of concentrated bacterial culture was centrifuged at 4000 rpm for twenty minutes. The resulting bacterial pellet was resuspended in 2ml of sterile PBS. Subsequently, the new solution

was treated in five portions (200 µl each) with a mechanical cell wall disruption in the MP-FastPrep 24 device at 5.5 m/s for 45 seconds, before 20 µl 600 mAU/ml Proteinase K and 200 µl of Buffer AL was added. Incubation time was at least 12 hours overnight at 56 °C. On the next morning, another fast-prep treatment was executed under the same conditions as mentioned above. Then DNA-samples were applied to the QIAamp-DNA mini kit according to the manufacturer's recommendations. The single five portions were pooled into one DNA-sample at the end of the procedure in order to maximize DNA output.

Fragment size and grade of shearing of resulting DNA were assessed by gel-electrophoresis with a 1.5 % UltraPure agarose-gel at 135V for 25 min. Purity and protein-content, as well as the approximate concentration of DNA-samples were tested by photometry in Nanodrop measurements on a Windows-based personal computer (Fig. 5)**.** Exact DNA-concentration was assessed by fluorometry using the pico-green method on a GENios- platform.

Concentrations higher than 20 ng/µl were considered sufficient and utilized for genome sequencing.

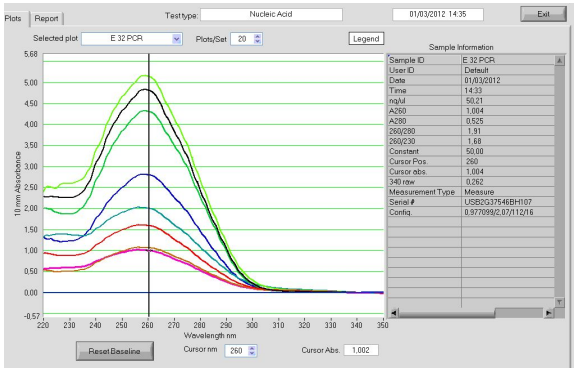

Fig. 5: Screenshot depicting photometric measurements of DNA-concentrations using the Nanodrop platform (Nanodrop 1000, Thermo Scientific, Wilmington, USA). In this case of the ENDO32 strain.

3.7. Verification of purity on the DNA-level

Absence of any other bacterial DNA was tested by PCR with universal 16s primers with the program described below (Table 3). If gel electrophoresis showed positive results, a sequencing reaction with all four 16s primers was performed with the program described below. Subsequent sequencing on an ABI 3130 genetic analyzer and application to nBLAST (National Center for Biotechnological Information, Bethesda, USA) were conducted to evaluate the results. If sequences showed single peaks and were identical to the 16s region in TW by more than 98 %, DNA was considered clean and admitted to further processing.

Absence of any eukaryotic cells, especially fungal contaminations, was tested by PCR with universal 18s RNA primers ITS1 and ITS4 (Table 4). If gel electrophoresis showed bands as in negative controls, the sample was considered to be free of eukaryotic DNA.

Table 3: Universal 16s-Primers used for identification of cultured bacteria. Primer-sequences and PCR-programs are depicted as well.

Primer	Sequence	Program	
536 F	5’ CAG CAG CCG CGG TAA TAC 3’	1. 95°C 15’ 2. 95°C 30’’ 3. 62°C 30’’	4. 72°C 1’ 5. 72°C 5’ 6. 4°C infinite
Rp2	5’ ACG GCT ACC TTG TTA CGA CTT 3’	1. 95°C 15’ 2. 95°C 30’’ 3. 62°C 30’’	4. 72°C 1’ 5. 72°C 5’ 6. 4°C infinite
800 F	5' TAG ATA TAC CCG GTT AG 3'	1. 95°C 15’ 2. 95°C 30’’ 3. 62°C 30’’	4. 72°C 1’ 5. 72°C 5’ 6. 4°C infinite
1050 R	5' CAC GAG CTG ACG ACA 3'	1. 95°C 15’ 2. 95°C 30’’ 3. 62°C 30’’	4. 72°C 1’ 5. 72°C 5’ 6. 4°C infinite

Table 4: Utilized 18s primers to clarify the absence of eukaryotic cells, especially fungal contaminations. Sequence of the internal transcribed spacer regions and used PCR-programs are depicted as well.

Primer	Sequence	Program	
ITS1	5' TCC GTA GGT GAA CCT GCG G 3'	1. 94 °C 2' 2. 94 °C 30" 3. 54 °C 1'	4. 72 °C 1.5' 5. 72 °C 8' 6. 4°C infinite
ITS4	5' TCC TCC GCT TAT TGA TAT GC 3'	1. 94 °C 2' 2. 94 °C 30" 3. 54 °C 1'	4. 72 °C 1.5' 5. 72 °C 8' 6. 4°C infinite

3.8. Genotyping based on HVGS-regions

As ten samples in total were not yet typed with the HVGS-genotyping system (DIG9, DIGADP11, DIGNEURO14, DIGNEURO18, NEURO20, NEURO21, DIGNEURO23, BCU26, ART31) these samples were subjected to genotyping before sequencing.

For this purpose, a tenfold-diluted sample of axenic cultures was taken for genotyping. Four primers specific for TW were used as described elsewhere [45]. PCR programs were conducted as indicated below with TW-specific primers (Table 5). PCR-products were applied to a 0.7 % UltraPure agarose-gel at 135 V for 25 minutes. As negative control, 5 µl of DNA-free water was added.

PCR products were subjected to DNA-purification with a Nucleo Fast 96-plate under a negative pressure of 20mmHG for 10 minutes. Afterwards, DNA was diluted in 50 µl of distilled water and shaken at 500 rpm for another ten minutes. Then a sequencing reaction for all positive

samples was conducted with primers and mixes as indicated in the tables below (Table 6, Table 7). Cycler program was the same for all primers.

Table 5: Primers used for genotyping TW-samples. Sequences and utilized PCR-programs are depicted as well.

Primer	Sequence	Program	
HVGS1: TW 133	5’ GCT GCG CGA AGT AAT TTG 3’	1. 95°C 15’ 2. 95°C 30’’ 3. 55°C 45’’	4. 72°C 90’’ 5. 72°C 5’ 6. 4°C infinite
HVGS2: ProS	5’ GCC TTG ACT ATG ACA TAA TCA A 3’	1. 95°C 15’ 2. 95°C 30’’ 3. 60°C 45’’	4. 72°C 90’’ 5. 72°C 5’ 6. 4°C infinite
HVGS3: SecA	5’ TTT GTC ATA GGC ATT TCT GTA G 3’	1. 95°C 15’ 2. 95°C 30’’ 3. 55°C 45’’	4. 72°C 90’’ 5. 72°C 5’ 6. 4°C infinite
HVGS4: 184	5’ CGG ATC TTC ACG AAA TGT CC 3’	1. 95°C 15’ 2. 95°C 30’’ 3. 55°C 45’’	4. 72°C 90’’ 5. 72°C 5’ 6. 4°C infinite

Table 6: Utilized primers for the genotyping sequencing reaction.

Primer	Sequence	Primer	Sequence
TW 133 F	5’ GCT GCG CGA AGT AAT TTG 3’	**TW 133 R**	5' AGA TAC ATG CGG AGA TAC T 3'
ProS F	5’ GCC TTG ACT ATG ACA TAA TCA A 3’	**ProS R**	5' TCG GAC TAA AAG TGC GAC AC 3'
SecA F	5’ TTT GTC ATA GGC ATT TCT GTA G 3’	**SecA R**	5' AGA CCT CAC TGT TAT ACG GAT 3'
184 F	5’ CGG ATC TTC ACG AAA TGT CC 3’	**184 R**	5' ATA ACA AGA AGC TGG ATA TGC 3'

Table 7: Sequencing reaction program and mastermix applied for the genotyping sequencing reaction.

Mix	1.5 µl BDV1 1 µl Big Dye 0.5 µl Primer solution 3 µl H20 4 µl PCR-Products	**Sequencing reaction program**	96 °C 1 min 96 °C 1 min 50 °C 10 sec 60 °C 3min 4°C infinite

After sequencing reaction, samples were put into a Millipore MultiScreen purification plate filled with Sephadex and centrifuged at 2600g for 6 minutes.

Purified sequencing products were put into an ABI 3130 Genetic analyzer. Sequence data was analyzed with the ChromasPro Software package version 1.5 (Fig. 6). Resulting sequences for spacer regions were put into the FASTA file format and blasted versus the nucleotide data base (nBLAST) and our internal data base with all TW-spacers. The concatenated sequences (TW133, ProS, SecA and Pro 184) were assigned the known genotype numbers if combinations already existed. If not, new genotype numbers were given in order of their exploration.

Spacer stability was evaluated by conducting a typing experiment with the reference strain *Twist* after nearly ten years of serial subculture and comparison to its prior typing results. The typing resolution was evaluated using the Hunter-Gaston discriminatory index (HGDI) [54].

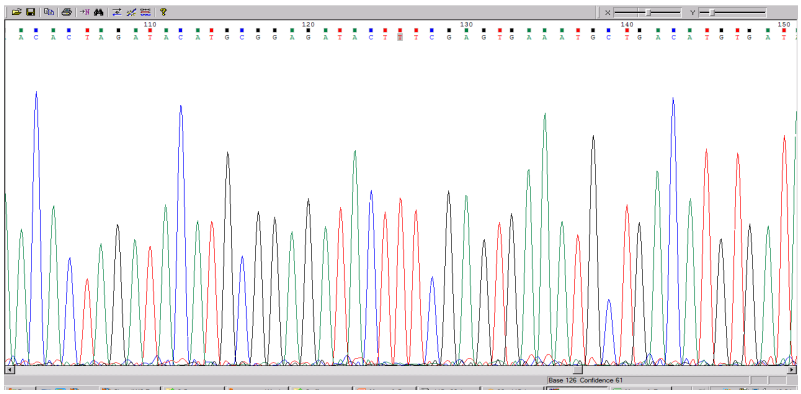

Fig. 6 : Screenshot of a PCR-sequence analyzed with the ChromasPro Software package version 1.5. Sequences were put into the FASTA file format and subsequently blasted versus URMITEs internal database.

3.9. Epidemiological, phylogenetic and geographical analysis of typing data

Coordinates of patient's home cities or – if unknown – of their physicians were determined using http://itouchmap.com/latlong.html. Then case cluster maps of the tested specimen were made by using Epi Info 7 (Centers for Disease Control and Prevention, Atlanta, USA). Multiple-Sequence-alignment (MSA) of genotype-sequences was performed with MEGA6 [55]. For this purpose, all four spacer-regions (TW133, ProS, SecA, Pro184) were concatenated. Afterwards, phylogenetic trees were constructed with the same software package using the neighbor joining method (NJ), bootstrap values above 70% were considered reliable.

Differences that appeared in frequencies were tested for statistical significance by using the exact Fisher-Test. A p-value smaller than 0.05 was considered statistically significant. Anonymous patients' data of our patient group was obtained from the WD database (personal communication Verena Moos, Charité Universitätsmedizin, Berlin, Germany) and research in the local clinical data network of the APHM, Marseilles (SMARLAB). Epidemiological data and frequencies were calculated using LibreOffice Calc. Genotyping data of other TW strains was obtained from the internal URMITE database in anonymous form (196 European patients).

3.10. Basic genomic data and sequencing (not performed by the author)

Sequencing itself was not conducted by the author but by the Genoscope Unit inside URMITE.

The complete genome sequences of *TW08/27* (accession number: NC_004551.1) and *Twist* (accession number: NC_004572.3) are available in the NCBI database. The sequences of the other 17 strains were obtained from SOLiD data. The paired-end library was constructed from 1 µg of purified genomic DNA of each of the 17 TW samples, and sequencing was carried out to 50x35 base pairs (bp) using SOLiDTM V4 chemistry on one full slide that was associated with 96 others projects on an Applied Biosystems SOLiD 4 machine. All of the 96 genomic DNA samples were barcoded with module 1-96 barcodes that were provided by Life Technologies, the libraries were pooled in equimolar ratios, and ePCR was performed according to Life Technologies' specifications: templated bead preparation kits were used with EZ beads to automate emulsification and amplification; and enrichment of E80 was used for full-scale preparation. For each run, 708 million P2-positive beads were loaded onto the flow cell, and the output read length was 99 expected to be 85 bp (50x35 bp).

Subsequently, the genome assembly was performed by Laetitia Rouli, PhD. For further genomic analysis, the genomes publicly available in Genbank were used.

3.11. Genomic analysis

Genomes were annotated using the RAST-server (Rapid Annotations using subsystems technology) for bacterial genome annotation [56]. Following the RAST algorithm, gene groups were subdivided into: Cofactors, Vitamins, Prosthetic Groups, Pigments (A); Cell Wall and Capsule (B); Virulence, Disease and Defense (C); Miscellaneous (D); Membrane Transport (E); RNA Metabolism (F); Nucleosides and Nucleotides (G) ; Protein Metabolism (H); Cell Division and Cell Cycle (I); Regulation and Cell signaling (J); DNA Metabolism (K); Fatty Acids, Lipids, and Isoprenoids (L); Dormancy and Sporulation (M); Cellular Respiration (N); Stress Response (O); Metabolism of Aromatic compounds (P); Amino acids and Derivatives (Q); Sulfur metabolism (R); Phosphorus Metabolism (S) and Carbohydrate Metabolism (T).

MSA of concatenated HVGS genotypes was conducted using the Clustal-W algorithm. In order to evaluate the existing typing system, a phylogenetic tree using the neighbour-joining method was constructed with all sequenced TW strains. Then multiple alignment with whole-genomes according to the MAFFT algorithm was performed using the galaxy online platform [57]. A

phylogenetic tree was calculated with MEGA 6 using the neighbour-joining method [55]. Bootstrap values above 70 % were considered reliable. Multiple alignment of all genomes was achieved using MAUVE (Multiple Alignment of Conserved Genomic Sequence With Rearrangements) by subgrouping them into strains causing CWD, NW, EW, AC and IP, respectively, in order to visualize the genome alignment [58]**.** All sequenced Genomes were tested for prophage DNA using the PHAST-application [59]. Predictions about antibiotic resistance genes were made using the ResFinder application and were searched for in the RAST annotation [60].

For different clinical manifestations (CWD, EW, NW, AC, IP) and different genotypes, the ANIs were calculated using JSpeciesWS and box-plotted against each other and against the average ANI of all clinical manifestations [61]. Differences in ANIs were tested using the two tailed Student's t-test.

TW was compared to other important human pathogenic bacteria and bacteria from the group of Actinobacteria and analyzed for genome size and GC-content. Reference genomes were withdrawn from Genbank for the following organisms: *Mycoplasma pneumoniae* (NC_000912.1), *Chlamydophila pneumoniae* (NC_00922.1), *Rickettsia rickettsii* (NC_010263.3), *Bartonella henselae (NC_005956.1), Coxiella burnetti* (NC_002971.3), *Corynebacterium diphtheriae* (NZ_LN831026.1), *Propionibacterium acnes* (NC_006085.1), *Mycobacterium leprae* (NC_002677.1), *Brucella melitensis* (NC_003317.1), *Legionella pneumophila* (NC_002942.5), *Acinetobacter israelii* (NZ_JON00000000.1), *Cellumonas fimi* (NC_015514.1), *Mycobacterium tuberculosis* (NC_000962.3), *Mycobacterium avium intracellulare* (NC_002944.2), *Mycobacterium abscessus* (NC_010397.1), *Frankia alni* (NC_008278.1), *Nocardia brasiliensis* (NC_018681.1).

3.12. Geographic distribution of sequenced genomes

As for genotyping data, geographical origins of the patients were obtained and a case cluster map was drawn by using Epi Info 7 (CDC, Atlanta, USA). Strains were subdivided according to their clinical manifestations. A geographical distance matrix was calculated using the Geographic Distance Matrix Generator [62]. The average nucelotide identity was calculated based on the blast algorithm (ANIb). A corresponding ANI matrix was then built using the JspeciesWS interface [61,63]. Geographic distance and ANIs were set into relation to show possible correlations. Possible

differences were evaluated by using the two-tailed Student's test to compare ANIs in European strains to non-European strains.

3.13. Calculation of core and pan-genome

For calculation of core and pan-genome, annotation files for all 19 genomes in .gbk format were retrieved from the RAST-annotation server and applied to the GET_HOMOLOGUES application[64]. Clusters were calculated according to the BDHD, OCML and COG algorithm. A consensus core and pan-genome were constructed. Genes present in all 19 genomes were considered the « core-genome », genes present in more than 95 % but not in all strains were defined as the « soft core », genes present only in a few genomes were considered the « shell », and genes present only in one or two genomes were defined as the « cloud genome ».

In order to estimate whether an open or a closed pan-genome is present, the pan-genome size was calculated according to Tettelin et al. by sampling the nineteen genomes [51]. If the graph was asymptotic, the pan-genome was considered closed. Subsequently, WiSPs were analyzed in regard to their appearance inside the core, soft core, shell and cloud genome.

4. Results

4.1. Patients' characteristics and clinics

Of the 29 patients in our study group, 10.3% (n=3) lay in an age range between 30 and 39 years, 17.2% (n=5) between 40 and 49, 24.1% (n=7) were between 50 and 59, 31% (n=9) between 60 and 69 years old, whereas 17.2% (n=5) were older than 69 years at the time of diagnosis. Therefore, the median was 58 years and the average age 56.9 years in our patient group. 75.9% of the patients were male (n=22).

68.9 % of the strains were collected from France (n=20), whereas 24.1 % (n=7) originated in Germany. One strain was from Portugal and one from Canada (3.4% each). Clinical manifestations were 48.2% CWD without neurological involvement (n=14), 24.1 % EW (n=7), 20.7% NW (n=6), 3.4% AC (n=1), 3.4% IP (n=1) (Fig. 7).

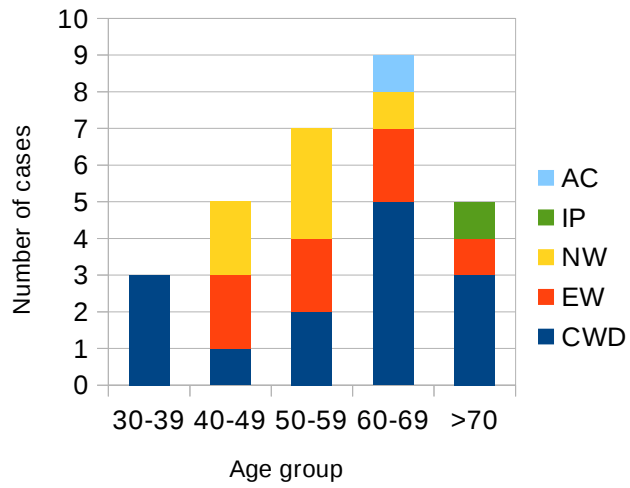

Fig. 7: Clinical manifestations observed in our group of patients. Age has been divided into decennials. Clinical manifestations are classical Whipple's disease (CWD), endocarditis due to Whipple's disease (EW), neurological manifestations due to Whipple's disease (NW), asymptomatic carriage (AC), isolated pulmonary affection (IP).

4.2. Genotyping of remaining strains

Additional genotyping of the strains DIG9, DIGADP11, DIGNEURO14, DIGNEURO18, NEURO20, NEURO21, DIGNEURO23, TWBCU26 and ART31 revealed 4 genotypes that have already been described in the past (Gt 1, 11, 39 and 98) and 4 completely new HVGS genotypes (Gt 102, 114, 115, 116) (Tab. 8). In Gt 102 and 116, two hitherto undescribed spacer sequences could be sequenced (named 30 and 31, Table 8). The two other new genotypes were simply new combinations of already known spacer sequences (Gt 114 and 115). Numbers for genotypes and

spacers were assigned in order of exploration. Sequences were admitted to the internal TW genotyping database at URMITE.

Tab. 8: Genotyped strains, their origins, their spacer regions and therefore their genotype. Numbers of spacer regions and numbers of genotypes were assigned in order of exploration. Completely newly discovered genotypes and spacer regions are marked in red.

Strain	Origin	TW133	ProS	SecA	Pro184	Genotype
DIG9	France	6	8	5	2	114
DIG ADP 11	France	30	6	5	1	116
DIG NEURO 14	Germany	31	1	6	8	102
DIG NEURO 18	France	6	6	5	2	39
NEURO 20	Germany	1	1	1	3	1
NEURO 21	Germany	1	7	1	1	11
DIG-NEURO 23	France	5	6	2	1	115
TW BCU 26	France	1	1	1	3	1
TW ART 31	France	1	6	9	1	98

Tab. 9: Newly discovered spacer regions and their sequences, TW133 No 30 and TW133 No 31. Genetic sequences are written from 5' to 3' end.

Spacer TW133 No 30
5'CGGAGATACTTTCGAGTGAAATGCTGACATGTGATAATTATAGGGTATCTTGGGAGTAAGTGAACCCT ATTGAAGCACCCGTGCAATTGAGGCGATCTGCATTGCAACAGAACTGTTTTGCATACAGTCAACAATA GTTTTTCCGCGCAGTAAAAGAGCATCACAAACCTTTTGATCAAAATCGATCTGCCATACCTGCTGAAA AGACGTGTATTCGCAAA3'
Spacer TW133 No 31
5'CGGAGATACTTTCGAGTGAAATGCTGACATGTGATAATTATATAGGGTATCTTGGGAGTAAGTGAACC CTATTGAAGCACCCGTCCAATTGAGGCGATCTGCATTGCAACAGAACTGTTTTGCATACAGTCAACAA TAGTTTTTCCGCGCAGTAAAAGAGCATCACAAACCTTTTGATCAAAATCGATCTGCCATACCTGCTGA AAAGACGTGTATTCGCAAA3'

4.3. Analysis of genotyping data in patient group including prior genotyping results

HVGS Gt 3 was the most frequent genotype in our patient group (27.6 %, n=8), followed by Gt 1 (13.8%, n=4). Gt 11 was found in 3 strains (10.3 %). All the other genotypes were only found once, respectively (Gt 8, 10, 13, 16, 24, 29, 39, 92, 98, 102, 113, 114, 115, 116) (Fig. 8)**.**

Looking further into detail Gt 3 appeared only in France (100%, n=8, Fisher-Test p=0.0332). Gt 1 was sequenced two times in German specimens and two times in French specimens (Fisher-Test p=0.2381). Whereas Gt 11 was found in 2 German specimens (not significant, Fisher-Test p=0.136). So Gt 3 appeared significantly more frequent in France than in Germany inside our study group.

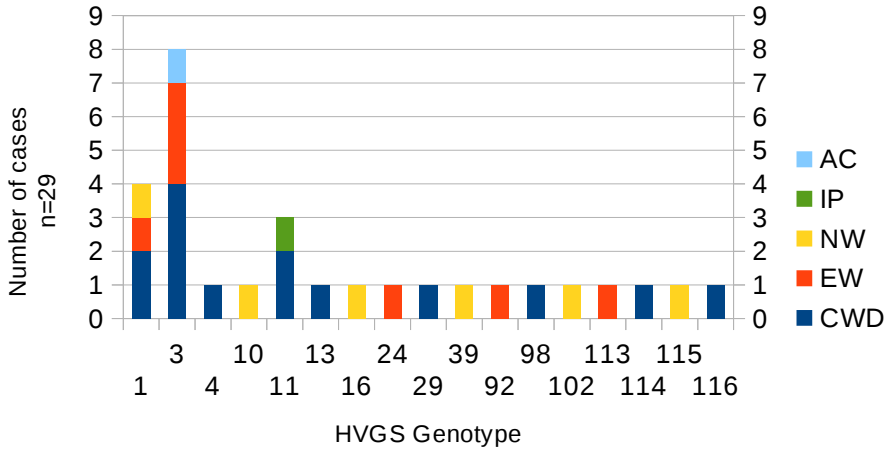

Fig. 8: Stack graph depicting frequency of genotypes. Colors assign different clinical manifestations. Gt 3 and 1 are the most frequent genotypes, followed by Gt 11. All others were just observed once in our group (singletons). Different genotypes cause different clinical manifestations (e.g. CWD, EW, NW in Gt1; CWD, EW, AC in Gt3).

The obtained data was put into the European context and compared to the whole of genotyping data available for TW in URMITE's database containing typing-data of 196 European patients. In total 72 different genotypes could be found in this database, thus demonstrating a very high genetic diversity. The Hunter-Gaston discriminatory index was calculated with a value of 0.9298 for all European samples.

Gt 1 and 3 were the most frequent genotypes both in the group of strains used in our study and in all genotyped strains in Europe until 2011. Gt 3 made up for 27.6 % in our group and 19.9 % in all European samples (n1=8, n2=39). Gt 1 was found in 13,8% and 14,8 % (n1=4, n2=29).

Further geographical examination showed that certain genotypes were found in distinct areas only: Whereas Gt 3 is specific and endemic to France, Italy and Switzerland, it was never described in Germany or Austria [65]. On the other hand Gt 1 is found all over Europe.

34.6 % of all samples in Europe were from Marseilles, France (n=66). In those samples 40 genotypes could be observed showing already a high genetic diversity in a geographic restricted area. Nonetheless, 18.5 % of all Genotypes could be assigned Gt 3 as well. Gt 1 was found only twice in the Marseilles Area (3.1%).

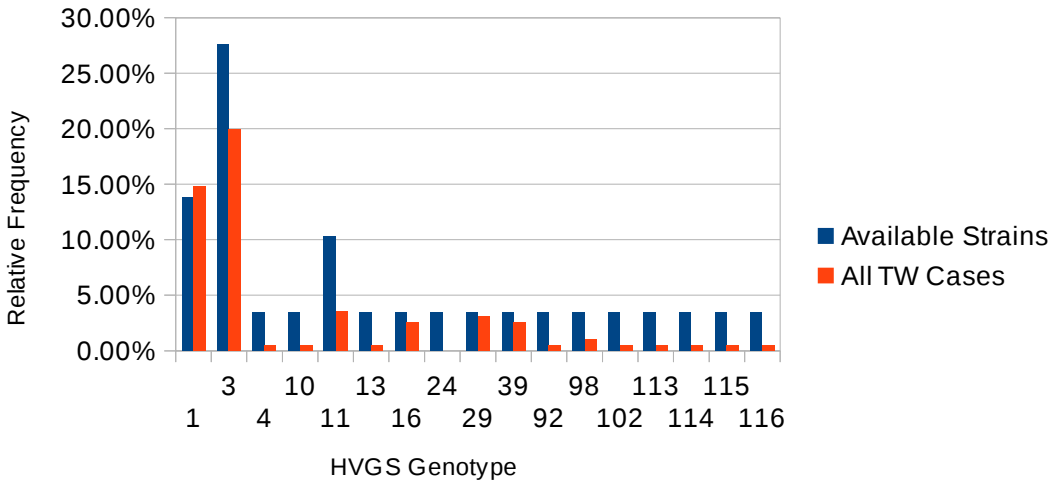

Fig. 9: Comparison of relative genotype frequency in all genotyped strains until 2011 and our study group. Because of the large amount of single strains, those genotypes gain a certain over-representation in our study. Gt 1 and 3 are the most frequent ones in both groups as described elsewhere. Blue for our study group, red for all European TW cases.

4.4. Sequenced genomes, genome length and GC-content

Only 17 out of the 29 planned bacterial strains delivered sufficient DNA-yields for genome sequencing. Namely strains SLOW2, NEURO1, DIG7, DIG9, DIG10, ART1, NEURO14, DIG15, DIGMUSC17, NEURO20, DIGADP25, TWBCU26, ENDO27, SALI28, ART29, PNEUMO30 and ENDO32 could be sequenced using the SOLiD technique. Of these, 11 strains were assembled to the scaffold level, whereas 6 strains remained on contig-level (meaning overlapping sequence data reads). The group was completed by the 2003 sequenced strains *Twist* and *TW08/27*. So 19 strains could be included in the comparative genome analysis (Table 10).

The median genome length was 0.927534 Mb. The average GC-content was 46.34 %. In this group of nineteen 31.6 % of strains had Gt 3 (n=6), 26.3% Gt 1 (n=5) and 10.5 % Gt 11 (n=2), while all the other genotypes appeared only one time (5.3% each, Gt 13, 16, 24, 29, 102, 114).

4.5. Geographic distribution of sequenced genomes and ANI distance matrix

6 out of 19 strains were German strains (31.6%), 12 out of 19 were from French patients (63.2%), and one strain from outside Europe (*Twist*, 5.3%). The case cluster map showed that German strains were responsible for CWD and NW, whereas French strains caused CWD, EW, IP and AC inside our group (Fig. 11).

Tab. 10: Table depicting all strains of TW that delivered sufficient DNA yields for genomic sequencing and were included in further genomic comparison studies. GC-contents, accession codes, genotypes and geographical origin are depicted as well.

Strain	Genome size (Mb)	Total assembly gap length	GC-Content (%)	Level	Accession code	Genotype	Geographical origin
Twist	0.927303	0	46.3	Chromosome	AE014184.1	24	Canada
TW08/27	0.925938	0	46.3	Chromosome	BX072543.1	1	Germany
SLOW2	0.927621	46,268	46.3	Scaffold	HG794425.1	13	France
NEURO1	0.927567	44,540	46.3	Scaffold	NZ_HG421449.1	1	Germany
DIG7	0.927564	45,892	46.3	Scaffold	HG794427.1	3	France
DIG9	0.880115	n.a.	46.4	Contig	CAUY000000000	114	France
DIG10	0.927515	42,964	46.4	Scaffold	HG794428.1	16	Germany
ART1	0.927575	44,340	46.3	Scaffold	HG424698.1	3	France
NEURO14	0.885853	n.a.	46.4	Contig	CAUR000000000	102	Germany
DIG15	0.927582	44,543	46.3	Scaffold	HG794423.1	11	Germany
DIGMUSC17	0.884564	n.a.	46.4	Contig	CAVA000000000	3	France
NEURO20	0.883582	n.a.	46.3	Contig	CAUX000000000	1	Germany
DIGADP25	0.883649	n.a.	46.3	Contig	CAUW000000000	3	France
TWBCU26	0.880271	n.a.	46.4	Contig	CAVB000000000	1	France
ENDO27	0.927598	45,491	46.4	Scaffold	HG794429.1	3	France
SALI28	0.927465	40,981	46.4	Scaffold	HG794430.1	3	France
ART29	0.927595	45,613	46.3	Scaffold	HG794431.1	29	France
PNEUMO30	0.927553	43,244	46.4	Scaffold	HG794432.1	11	France
ENDO32	0.927567	45,143	46.4	Scaffold	HG794424.1	1	France

As already shown above CWD is the most frequent affection caused by TW in our study group (57.9%, n=11). Three strains causing NW were sequenced (15.8%). They were all from Germany. Causing EW three strains could be included as well (15.8%). Their providence is France and Canada. Fortunately one IP and one AC strain could be sequenced and therefore included as well.

In the whole group comparison whole-genome ANIs lay between 98.98% and 99.8% percent. Median ANI between all sequenced strains was at 99.32 %. Geographical distance of the sequenced strains lay between 0 km and 5682.34 km. Median distance was 588.8 km. The geographic distance matrix and the ANI matrix could not demonstrate any significant negative correlation between geographic distance and ANI (Fig. 10). Comparison between European strains and the only Canadian strain, *Twist*, showed no significant differences in ANIs. P-values were all non-significant. Nevertheless all maximum values for ANIs were for intra-European comparisons (99.8 %). The comparison of ANI includes non-coding areas as well.

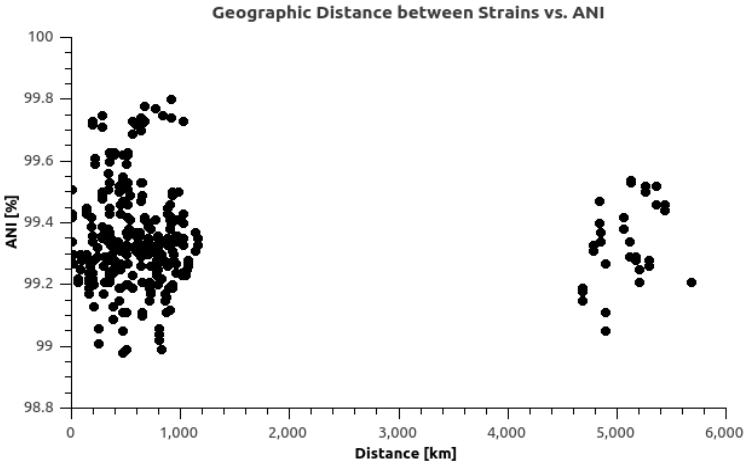

Fig. 10: Scatter plot depicting the geographic distance [km] between strains against the average nucleotide identity [%]. Most comparisons were conducted between European strains (left cluster). All strains were compared to the only Canadian strain Twist (right cluster). No negative correlation could be shown.

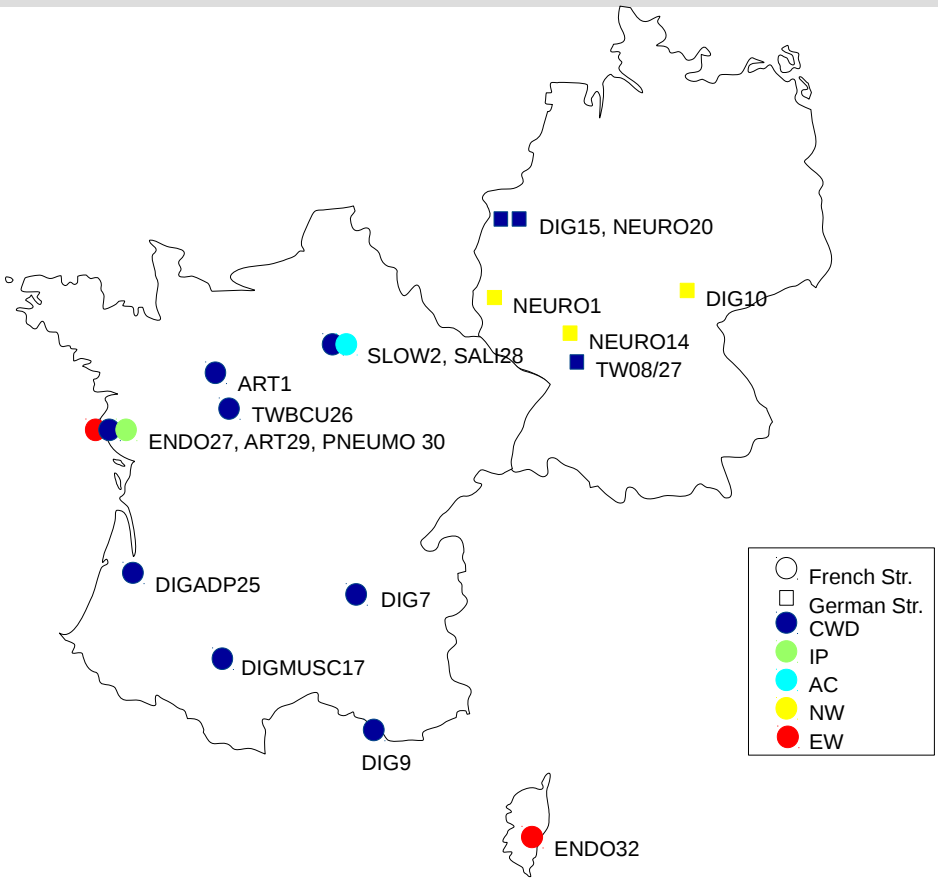

Fig. 11: Map depicting geographic origins and clinical manifestations of the sequenced TW genomes in Europe. The only non-European strain (Twist, from Halifax, Canada) is not shown in this map. Coordinates assign either patients' cities or cities of their physicians. Circles depict French strains, quadrangles mark German strains. Blue for CWD, red for EW, yellow for NW, green for IP and light blue for AC.

4.6. Phylogenetic analysis of sequenced strains using genotyping data

In the phylogenetic tree of sequenced strains using genotyping data identical genotypes grouped together, naturally, as no distinction is possible. Thus, there were three clades formed by Gt 3, 1 and 11. Gt 29 (ART29) showed a close relationship to Gt 3. In general German and French strains mixed together. But all French strains bearing Gt 3 formed the clade mentioned above. And four of the six German strains showed a close relationship: Three strains could be typed with Gt1, and a fourth (NEURO14, Gt102) was closely related to it. Interestingly Gt 24 (Twist), the only strain from outside Europe, grouped alone.

Except for CWD grouping together in Gt 3, clinical manifestation mixed together as well. Bootstrap values for all branches were above 70% and therefore considered reliable.

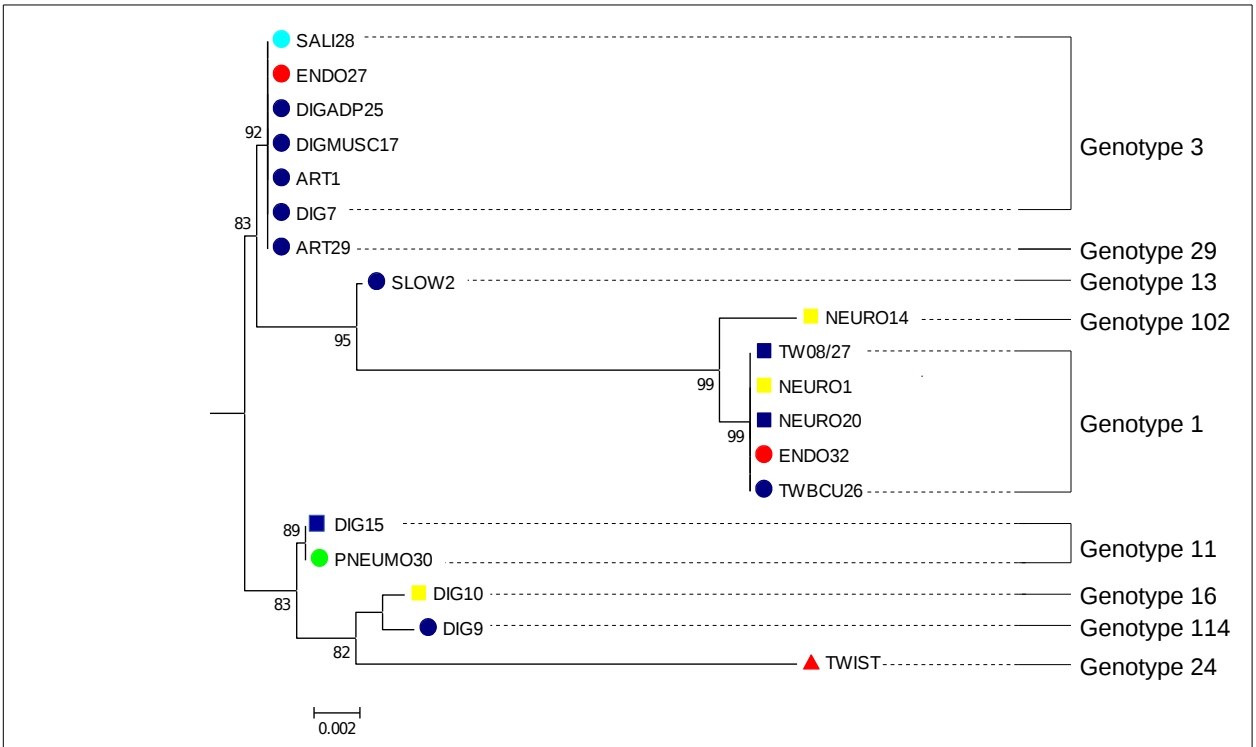

Fig. 12: Phylogenetic tree made with MEGA6 using the neighbour-joining method. This tree was evaluated with the maximum likelihood method. Bootstrap values more than 70% were considered reliable. Circles depict French strains, quadrangles mark German strains, the triangle the Canadian strain. Blue stands for classical WD, red for EW, yellow for NW, green for IP and light blue for AC. Naturally strains having the same genotype group together, as no discrimination is possible. As described above Gt 3 and 1 are the most frequent ones.

4.7. Phylogenetic analysis of sequenced strains using whole-genomes

The phylogenetic tree using a whole-genome alignment showed that most of the French strains group together. German strains DIG10 and NEURO20 were closely related, and DIG15 and NEURO14 grouped together as well. Besides that all geographical origins mixed together.

Except for ART1 all strains bearing Gt 3 grouped together, showing still a high genetic similarity on whole-genome level. ART 29 grouped amidst all Genomes bearing Gt 3.

As in other phylogenetic studies of TW no grouping according to caused clinical manifestations could be observed, the only exception being DIG15 and DIG9 as two strains causing CWD [45]. While genetic diversity was very low, TW08/27 showed the highest genetic difference in comparison to the other strains. In total branch lengths were very short, mirroring the high genetic similarity between strains of TW already described above.

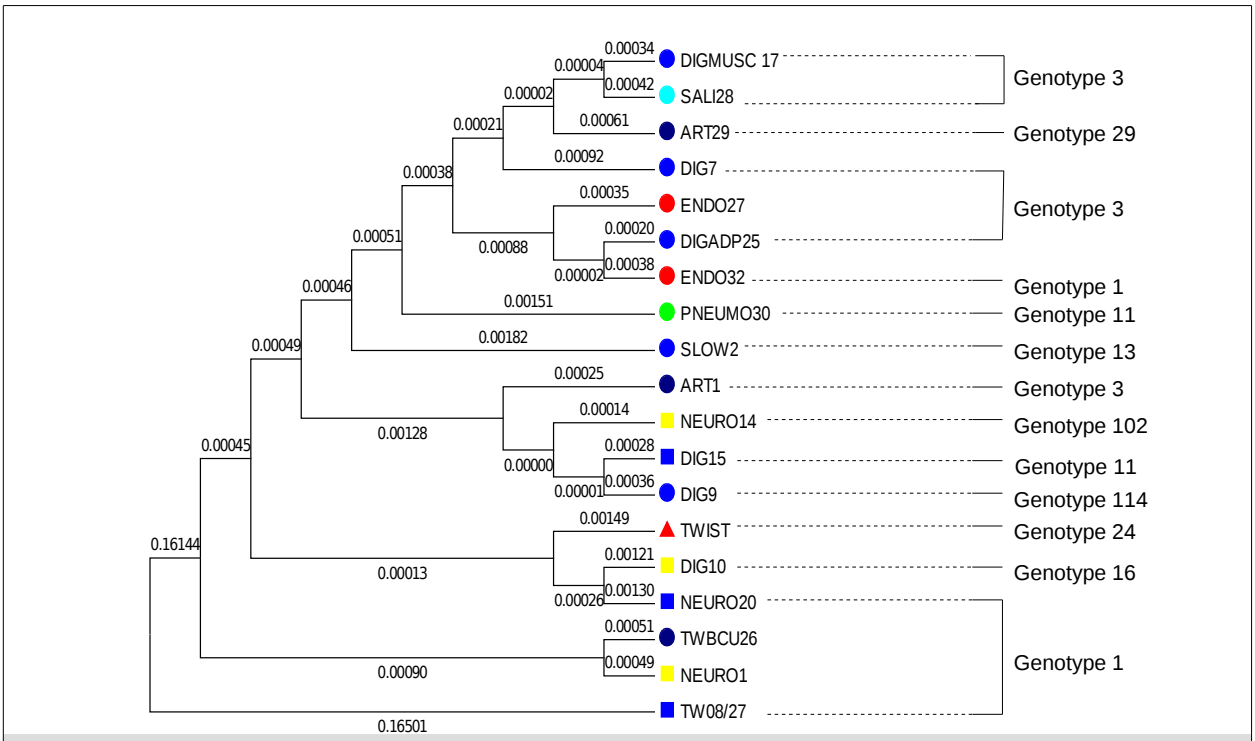

Fig. 13: Phylogenetic tree made with MEGA6 using the neighbour-joining method. This tree was evaluated with the maximum likelihood method. Bootstrap values more than 70% were considered reliable. In this tree almost all nodes are reliable. Circles depict French strains, quadrangles mark German strains, the triangle the Canadian strain. Blue stands for classical WD, red for EW, yellow for EN, green for IP and light blue for AC. In this case a cladogram was used. Branch lengths are written in decimals above the according branch.

4.8. ANI-comparison of most important genotypes and different clinical entities

In all strains bearing Gt 1 the median nucleotide identity was 99.44 %, in those with Gt 3 it was 99.53 % and in all Genotypes (including Gt1 and 3) it was 99.32 % (Fig. 14). A significant difference between the ANI of strains bearing the same Genotype and strains with different genotypes could be found (Gt1 to all genotypes $p=5.4*10^{-10}$, Gt3 to all Gts $p=5.4*10^{-10}$).

There were no significant differences between NW, CWD, EW, and AC/IP (the only AC case and IP case were put into one group in order to make a comparison possible). P-values were all

higher than 0.005. Also comparison of strains causing the same clinical manifestation to the whole of strains showed no significant difference. The median ANI for NW was 99.35 %, for CWD 99.30 %, for EW 99.21 %, for ACIP 99.43 %, and for all strains compared against each other 99.32 %.

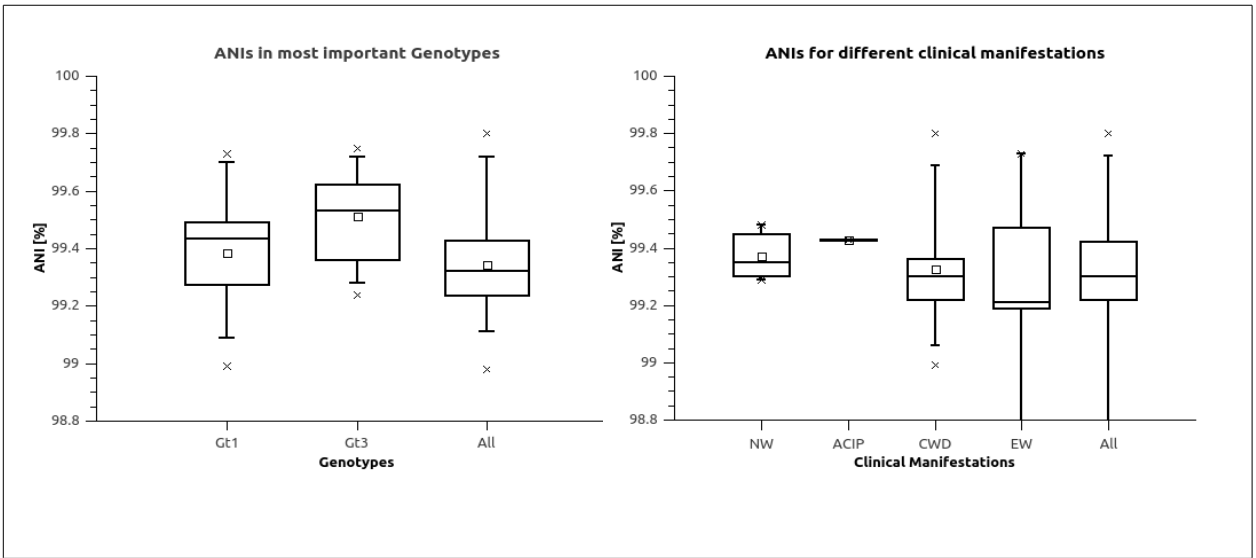

Fig. 14: *Left*: Boxplot showing ANIs for the most important genotypes. Gt 1 and 3 showed a significant higher ANI than all other genotypes combined to each other. *Right*: Boxplot showing ANIs for different clinical manifestations. No significant difference could be shown between the different clinical entities.

4.10. Comparison of TW with other intracellular pathogenic bacteria and other species in the group of Actinobacteria

TW shows a much smaller genome as being comprised of only 0.927534 megabasepairs than other representatives inside the group of Actinobacteria. For example the genome of *Mycobacterium tuberculosis* is 4.41 Mb large. The genome of *Frankia alni*, a symbiotic bacterium living with plants, has a genome size of 7.5 Mb (Table 11). Inside this chosen group of Actinobacteria the average genome size was 4.43 Mbp. GC-content was quite variable as well. TW has a low GC content of 46.34%. The other Actinobacteria range from 53.5 % in *Corynebacterium diphtheriae* to 72.8 % in *Frankia alni*. The average GC-content was at 63.95 %.

When compared to other intracellular pathogenic bacteria there are more similarities (Table 12). *Mycoplasma pneumoniae*, a germ causing pneumonia in humans, shows a smaller genome than TW with 0.81 Mb. *Rickettsia rickettsii*, *Bartonella henselae*, *Chlamydophila pneumoniae* and *Coxiella burnetti* all have genome sizes equal or smaller than 2 Mb. Inside the group of human pathogenic intracellular bacteria the average genome size was at 1.86 Mbp. GC-content ranged

from minimum values of 32.4 % (*Rickettsia rickettsii)* to maximum values of 57.2 % (*Brucella melitensis*) in the group of human pathogenic bacteria. The average GC-content was 41.96 %.

Table 11: Table showing genomes of other important species inside the Actinobacteria, utilized reference genome, life style, pathogenicity, genome size and GC-content.

Group	Species	Reference genome	Life style	Pathogenicity	Size [Mb]	GC-content
Actinomyces	***A. israelii***	*Actinomyces israelii DSM 43320*	Extracellular	Actinomycosis	4.03	71.4 %
Nocardia	***N. brasiliensis***	*Nocardia brasiliensis ATCC 700358*	Extracellular	Systemic Nocardiosis, Lung infections, skin infections	9.44	68.0 %
Mycobacteria	***Mycobacterium tuberculosis***	*Mycobacterium tuberculosis H37Rv*	Extracellular/ intracellular	Tuberculosis	4.41	65.6 %
Mycobacteria	***Mycobacterium leprae***	*Mycobacterium leprae TN*	Intracellular	Leprosic disease	3.27	57.8 %
Mycobacteria	***MAC***	*Mycobacterium avium subsp. paratuberculosis K-10*	Intracellular	Lung diseases / Disseminated infections	4.83	69.3 %
Mycobacteria	***Mycobacterium abscessus***	*Mycobacterium abscessus ATCC 19977*	Intracellular	Lung, skin, wound infections	5.07	64.1 %
Corynebacteria	***Corynebacterium diphtheriae***	*Corynebacterium diphtheriae NCTC11397*	Extracellular	Diphtheria / Endocarditis	2.46	53.5 %
Propionibacteria	***Propionibacterium acnes***	*Propionibacterium acnes KPA171202*	Extracellular	Folliculitis / Acne / Wound infections	2.56	60.0 %
Frankia	***Frankia alni***	*Frankia alni str. ACN14A*	Extracellular	Non pathogenic / Symbiotic relationship with plants	7.5	72.8 %
Cellulomonas	***Cellumonas fimi***	*Cellulomonas fimi ATCC 484*	Extracellular	Non pathogenic / environmental bacterium	4.27	74.7 %
Tropheryma	***Tropheryma whipplei***	*Tropheryma whipplei str. Twist*	Intracellular	CWD, NW, EW, IP, AC	0.93	46.3 %

Tab. 12 : Table showing genomes of other impotrant human pathogenic bacteria, their reference genomes, life styles, genome size and GC contents.

Group	Species	Reference genome	Life style	Pathogenicity	Size [Mb]	GC-content
Rickettsia	***Rickettsia rickettsii***	*Rickettsia rickettsii str. Iowa*	intracellular	Rocky mountain spotted fever	1.26	32.4 %
Bartonella	**Bartonella henselae**	*Bartonella henselae str. Houston-1*	Intracellular / facultative	Cat scratch fever / Bacillary angiomatosis	1.93	38.2 %
Brucella	**Brucella melitensis**	*Brucella melitensis bv. 1 str. 16M*	Intracellular / facultative	Brucellosis	3.30	57.2 %
Coxiella	***Coxiella burnetii***	*Coxiella burnetii RSA 493*	Intracellular	Q-Fever	2.0	42.7 %
Chlamydia	***Chlamydophila pneumoniae***	*Chlamydophila pneumoniae CWL029*	Intracellular	Pneumonia	1.23	40.6 %
Legionella	***Legionella pneumophila***	*Legionella pneumophila subsp. pneumophila str. Philadelphia 1*	Intracellular (Acanthamoeba, Naegleria)	Pneumonia / Pontiac fever	3.4	38.3 %
Mycoplasma	***Mycoplasma pneumoniae***	*Mycoplasma pneumoniae M129*	Intracellular	Pneumonia	0.819394	40.0 %
Tropheryma	***Tropheryma whipplei***	*Tropheryma whipplei str. Twist*	Intracellular	CWD, NW, EW, IP, AC	0.93	46.3 %

4.11. Comparison of RAST annotation in TW strains

An average of 698.63 features was annotated in all TW strains. In most categories the amount of genes was the same. Differences were found in 13 out of 20 gene categories. In only 7 out of 20 gene groups the amount of genes was exactly the same, namely: Miscellaneous (D), Membrane Transport (E), DNA Metabolism (K), Dormancy and Sporulation (M), Stress response (O), Metabolism of aromatic compounds (P) and Sulfur Metabolism (R) (Fig. 16).

The biggest variance could be shown inside genes coding for Protein metabolism (H). The minimum amount of coding genes was found in Strain DIG10 with 126 genes, the maximum amount with a 169 genes in strains *Twist* and *TW08/27*.

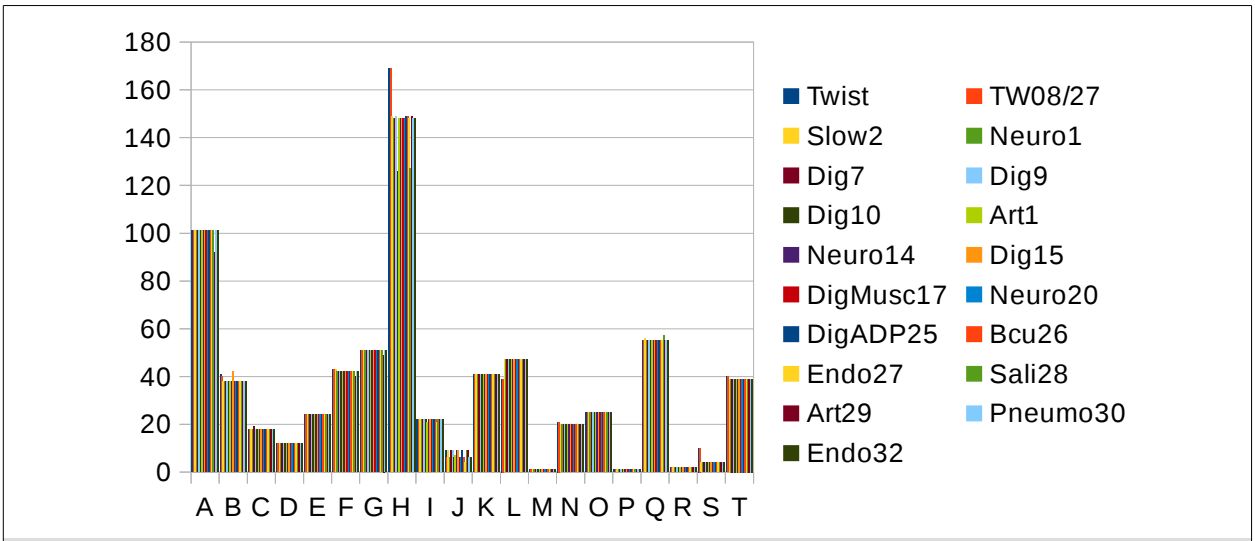

Fig. 16: Bar graph showing the amount of genes found in a category according to RAST annotation. In 13 out of 20 gene groups a difference in gene numbers could be found. Strains of this study are depicted in different colors. Letters assign different gene groups: Cofactors, Vitamins, Prosthetic Groups, Pigments (A); Cell Wall and Capsule (B); Virulence, Disease and Defense (C); Miscellaneous (D); Membrane Transport (E); RNA Metabolism (F); Nucleosides and Nucleotides (G) ; Protein Metabolism (H); Cell Division and Cell Cycle (I); Regulation and Cell signaling (J); DNA Metabolism (K); Fatty Acids, Lipids, and Isoprenoids (L); Dormancy and Sporulation (M); Cellular Respiration (N); Stress Response (O); Metabolism of Aromatic compounds (P); Amino acids and Derivatives (Q); Sulfur metabolism (R); Phosphorus Metabolism (S) and Carbohydrate Metabolism (T).

4.12. Prediction of antibiotic resistance genes and phage DNA

All tested genomes showed no sign for prophage DNA. Neither intact, incomplete or questionable prophage regions were found in any case using the PHAST-application (n=19). Application of all sequenced genomes to ResFinder could not find any antibiotic resistance genes on nucleotide-level. Though in the RAST-Annotation a genotypic antibiotic resistance to fluoroquinolones was described in all strains linked to mutations in the Gyr A Gene (n=19). A

mutation in this gene has been shown to increase minimal inhibitory concentrations (MICs) for fluoroquinolones of 2nd and 3rd generation as Ciprofloxacin and Levofloxacin [66].

4.13. Core-genome and pan-genome

Using the COG algorithm 1.506 gene clusters could be calculated in total. Of those 693 belonged to the core, which means these genes were found in all 19 genomes. 730 gene clusters could be found in 18 genomes and therefore belong to the soft core. 319 gene clusters were found in 3 to 17 genomes thus forming the shell, and 457 gene clusters could be found in less than two or less respectively (Fig. 17). Similar results were calculated using the OMCL and the BDBH algorithm (not shown).

Calculating the core genome using the Tettelin algorithm the core genome size went asymptotic to a 705.6 genes at nineteen genomes. While calculating the pan-genome 1220.2 gene clusters were reached at 19 genomes. For the nineteenth genome still 9 genes were added. But an asymptotic behavior of the graph could be observed (Fig. 18). Regarding this data TW can be considered to have a closed pan-genome.

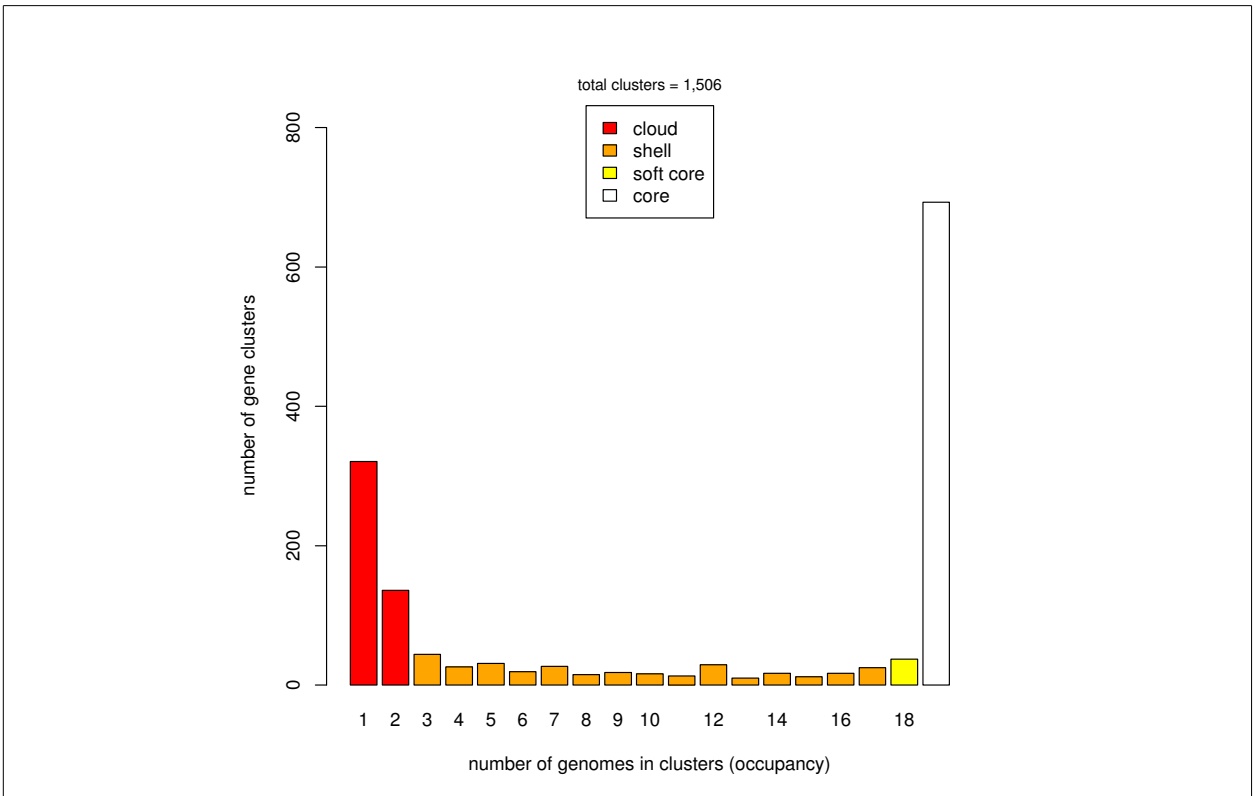

Fig. 17: Bar graph showing the distribution of gene clusters belonging to the core, soft core, shell and cloud genome according to the COG algorithm. In TW 693 gene clusters belong to the core genome, 730 to the soft core, 319 to the shell and 457 to the cloud genome.

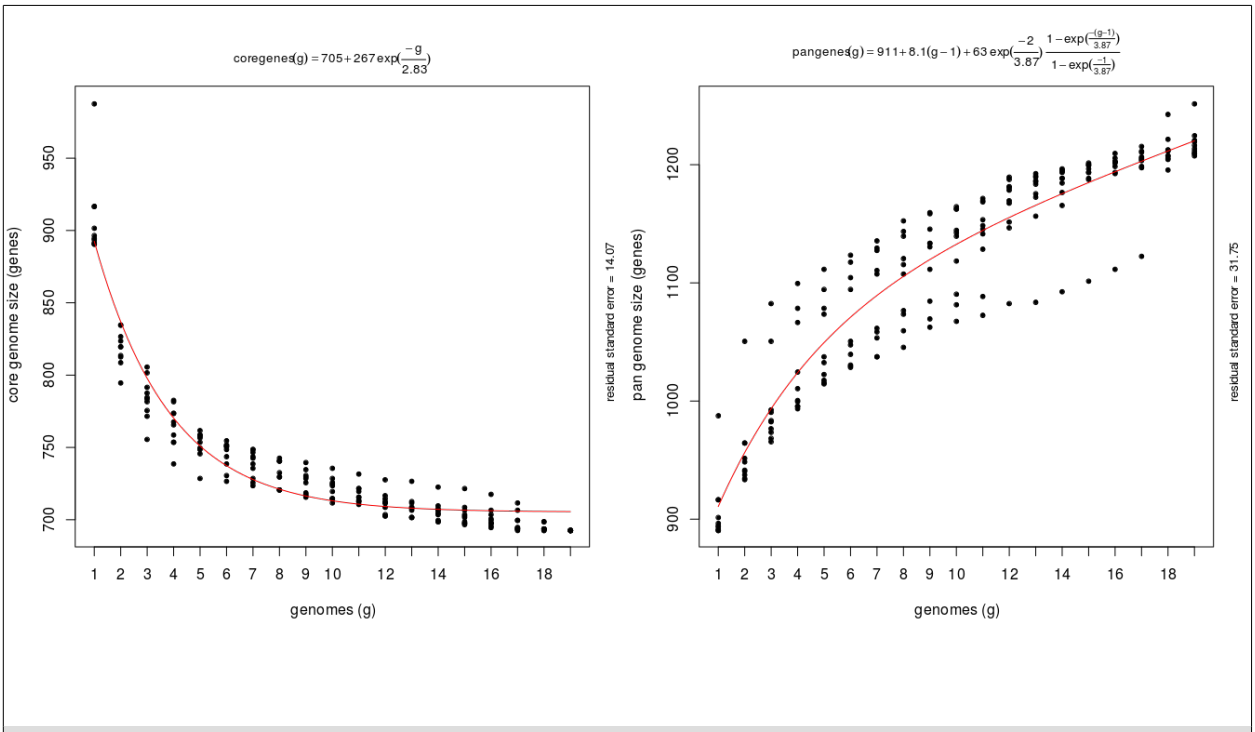

Fig. 18: *Left:* Plot depicting sampling of the core genome size against the number of genomes applied to the study. The graph is asymptotic towards the number 705.6. The core genome graph shows therefore an asymptotic behavior. *Right*: Plot showing the pan-genome reaching its maximum at 1220 genes. For the last genome still 9 genes are added, an asymptotic behavior can be anticipated.

4.14. WiSPs in the core and pan-genome

As WiSP-genes have been shown to be the most variable genetic regions in TWs genome they were examined in more detail and in regard to their appearance inside the core, soft core, shell and cloud genome.

Interestingly, only two WiSP-genes were found in the core genome and made up for only 2,9 % of all core genome genes. In the soft core (95 % of all strains) only 3 WiSP genes could be grouped (4.1 % of soft core genes). The number increased in the shell genome and the cloud genome (17 or 5.3 % of all genes, 77 or 16.8 % of all genes). The number of WiSP genes inside the cloud genome increased significantly thus showing a high genetic variability.

Only two WiSP genes were found in all 19 TW strains inside our group (14499 and 15299, numbers of WiSPs were distributed in order of their appearance during pan-genome annotation process). Alignment and phylogenetic analysis of those two WiSPs could show no grouping according to clinical manifestations either. The heat map showing all WiSPs inside the core, soft core and shell genome showed no clustering but a high genetic variability inside the group of WiSP genes (Fig. 19).

Some WiSPs were even found several times inside the same strain (for example 15364 in Twist, TW08/27, Neuro 1, DigMusc17, Neuro20, DigADP 25). Most WiSPs inside the cloud genome were specific to certain strains.

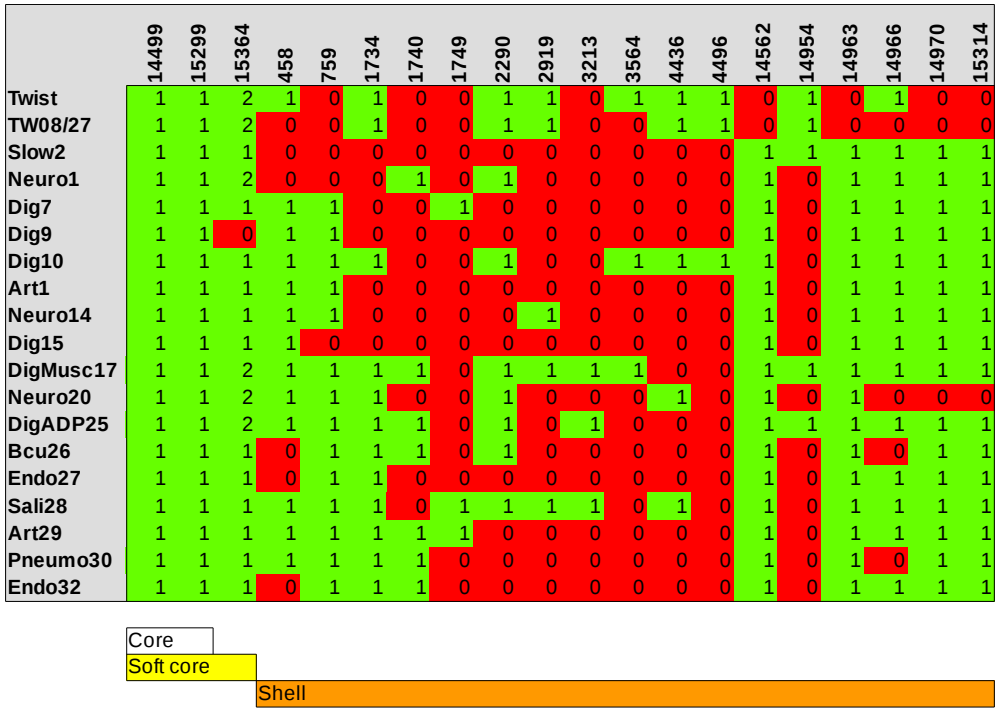

Fig. 19: Heat map showing presence or absence of WiSPs inside the core, soft core and shell genome of our study group. Red for absence, green for presence. Numbers of WiSPs were distributed during the annotation process while calculating the pan and core genome.

5. Discussion

5.1. Patients' characteristics and clinics

Our patient group was comprised of 29 patients. The majority (38,2 %) were older than 60 years. 75.9 % of the patients were male. 48.2 % of the patients showed CWD, 24.1 % showed EW, and 20.7 % showed NW. The majority of patients was from France.

Compared to other epidemiological data in terms of clinical manifestation and patient characteristics, our group seems to be representative, and thus a genomic study performed with this group was considered representative too.

As this is a French-German study, there's a clear over-representation of strains from these two countries. This will limit the applicability of the results to Central Europe only.

Considering the size of our group, the fact that WD is a very rare disorder, and that TW is a bacterium whose culture is still difficult should be taken into account. The permanent culture that is performed in URMITE Marseilles is the world's biggest biobank of TW to the author's knowledge. Thus, the present genomic study was the first full genomic sequence comparison study in TW and the first pangenomic study using NGS-technology so far.

5.2. Analysis of genotyping data

Further typing according to the HVGS-genotyping system showed that Gt 3 was the most frequent genotype in our patient group, followed by Gt 1. Extrapolated to all European samples, these two genotypes are the most frequent ones in all Europe. It is even more interesting to see that Gt 3 appeared in France only and was not described in Germany. Seeing that 196 European samples were examined and genetic diversity is very high, this fact was very astonishing.

Until now, the mode of transmission of TW is not completely clear. But most recent data shows that an oro-oral or feco-oral transmission route by interhuman contact is most probable [22]. Seeing that Gt 3 has been described to be responsible for travelers' diarrhea and small epidemics of diarrhea, could be one explanation for the predominance of Gt 3. This genotype is restricted and endemic to France. Gt 1 showed a strong predominance in Germany and Austria. Any link to possible epidemic factors could not be shown until now [65].

The spacer stability was shown in a typing experiment with the strain Twist. This could demonstrate a stable Gt 24 over a continuous subculture of ten years. Repeated typing experiments between 2007 and 2012 showed spacer stability in time as well. The appearance of

two genotypes in a single patient can thus be interpreted more as a new infection with another genotype than as a mutation in the previous strain.
Discrimination power of the typing system was very high. The Hunter Gaston-Discriminatory Index was calculated at 0.9298. This fact is mirrored in the high amount of different genotypes whereas a single basepair difference makes up for a new genotype. We propose to call these single genotypes “singletons”.
Although the actual typing system allows quite sophisticated answers to epidemiological questions and the propagation of TW inside a population, the real epidemiological character remains unclear. Probably propagation on big a scale is relatively limited, which might be one answer for the high genetic diversity inside the specimen. Other author's propose that TW is a commensal strictly restricted to its host organism [11].

Nevertheless, a typing system that is limited to only small parts of the genome sometimes cannot differentiate between different strains. The strong predominance of Gt 3 could only be an accidental identity in the HVGS genome areas, while the strains do not share the same origin. This underlines the importance of the implementation of genomotyping systems as attempted in the present study for the increase of typing resolution for further epidemiological studies. But with such a high typing resolution, the actual HVGS typing system can be sufficient for epidemiological questions where genomic sequencing is not available or impractical.

5.3. Sequenced genomes, genome length and GC-content

Unfortunately, only 17 out of the planned 28 bacterial strains delivered sufficient DNA yields for sequencing. This was mainly related to the fact that despite the implementation of an axenic culture medium, some strains showed very poor growth.
Nevertheless, 17 strains could be sequenced and set into relation with the two reference genomes *Twist* and *TW08/27*. The median Genome length was 0.927537 Mb. Thus, all TW strains showed a very reduced genome compared to other pathogenic bacteria. Inside the other pathogenic Actinobacteria, no smaller genome could be observed. From other important human pathogenic bacteria, only *M. pneumoniae* showed a smaller genome size. As described in previous studies, this reduced genome could be a sign of a strict intracellular lifestyle, whereas TW is very dependent on its host cell.
The average GC-content was lower than for other bacteria in the Group of Actinobacteria. It lay at only between 46.34%. The most important human pathogenic bacterium of this group, *M.*

tuberculosis, has a genome size of 4.41 Mb and a GC-content of 65.6 %. Like TW, all the other human pathogenic intracellular bacteria (for example *M. pneumoniae)* have a low GC-content. A high GC content and a large genome have been observed in free living organisms in varied environments. This is thought to increase horizontal gene transfer [67]. So it can be hypothesized that TW is a strictly intracellular living organism with a reduced genome and that no significant horizontal gene transfer has taken place in the past. This is underlined by the fact, that the PHAGE application could not find phage DNA in any genome.

5.4. Geographic distribution of sequenced genomes

The geographic distribution shows that in our group, CWD is the most important clinical affection of WD. There are only two European cases of EW in our group, both from France, and one from Canada (strain Twist). Fortunately, one AC and one IP case could be included in the genomic comparison group.

The ANI matrix showed no correlation between genetic and geographic distance. A presupposed negative correlation between average nucleotide identity and geographic distance in kilometers has proved wrong. ANIs of the strains from the same city or from relatively distanced regions such as Canada and central Europe showed no significant difference.

Thus, inter-individual differences seem to be the most important factor. One could also argue that northern America and central Europe both have mainly Caucasian populations (in which WD is most often described). With the colonization of Northern America and emigration waves in the 19th century, populations have been separated quite late, so the genetic distance between microbiota might be only slight. In addition, only one strain from Northern America could be included. So statistical power remains limited.

Unfortunately, in our study group no African or South American strain was represented. Comparison between European strains and Senegalese strains using HVGS genotyping data showed that the Senegalese strains form a distinct clade [13]. If included in a whole-genome study, it would have been perhaps possible to show a correlation between population genetics in humans and TW. Since the pathogen is considered as a commensal by many authors, this hypothesis would be plausible.

5.5. Phylogenetic analysis of sequenced strains

Phylogenetic analysis of genotyping data with the strains that could be sequenced showed that no discrimination between different strains is possible if they bear the same Genotype. Gt 3 grouped together naturally, Gt 1 and 11 did so too. As Gt 3 has only been found in France, all strains bearing that genotype built a cluster. The same effect was observed for Gt 1. Once again, no real clustering according to clinical manifestation could be observed.

Strain *Twist* that was isolated from a Canadian patient, grouped alone and thus showed a high genetic diversity in comparison to the other strains in its HVGS genome area. Nevertheless, this is only a small part of the whole genome.

One of the aims of this study was to increase typing resolution. In the phylogenetic tree using whole genomes, every strain had its own branch as expected. Thus, a discrimination between the different strains was made possible. Interestingly, the French strains bearing Gt 3 still grouped together, except for strain ART 1. Strain ART29 was also still closely related to all strains bearing Gt 3. So a close genetic relationship can still be assumed.

The fact that all Genomes bearing Gt 3 had their own branch and were thus separated strains is an argument against small epidemics caused by TW. In this case, we should have shown the exact same genetic content in all strains bearing that Gt. If used in cases where an epidemic transmission is possible, genomotyping could deliver additional information in the future.

Strains did not group according to clinical manifestations or geographic origin except for strains bearing Gt 3. This time, strain *Twist* from Canada showed a smaller genetic difference in comparison to the other strains. *TW08/27* was the strain with the highest genetic difference. This could be linked to the alignment method used and to the fact that some genomes were only assembled to contig level.

The whole-genome approach was used in order to increase typing resolution, but this has the disadvantage that also non-coding areas are included. Other possibilities would be to choose certain pathogen factors, surface proteins (as WiSPs for example), only coding areas or SNPs. These should be the content of further epidemiological and typing studies with TW.

5.6. Core and pan-genome

Calculation of the core and pan-genome showed that TW can be assumed to have a closed pan-genome. This has also been shown for other intracellular, sympatric bacteria [52]. As TW is considered a commensal by some authors [11], it would reside inside the human gut. There it is

confronted by many other bacterial species of human microbiome and by gene exchange as e.g. horizontal gene transfer. But as an intracellular bacterium, this exchange could be very limited, as shown by the reduced genome and closed pan-genome. This hypothesis is supported by the low GC-content described above.

Interestingly, it was demonstrated that there were significantly more WiSPs inside the cloud and shell genome, and only three WiSP genes in total were found inside the core and soft core genome. So only two identical WiSPs were shared by all strains. As the rest of the genome is very conserved, such a high evolutionary movement inside these gene areas is even more striking. These findings underlined the possible importance of these proteins in the pathogenesis of WD [50].

One possible scenario might be the building of a distinct set of WiSPs according to the host's immune system or to the environment TW is living in. Antigenic variation has also been shown in other pathogenic bacteria, as for example *Neisseria spp.* [68]. This possibility is stressed by the observation in previous bacterial studies that TW loses its set of surface proteins after several passages in axenic cultures [1]. A certain effect of mimicry might be possible as well.

Further studies in vivo and in vitro have to be made in order to understand the role of those proteins better.

5.7. Host and pathogen factors in WD

Our whole-genome study could find no link from genome information to disease outcome. A high variability in WiSPs could be a possible factor, but remains unclear.

As described above, several immunological factors have been found to play a role in the pathogenesis of WD. A number of polymorphisms of genes are associated with the appearance of CWD. The HLA type influences the course of the disease. There are ACs and serological examinations have shown that more than half of the French population have had contact to the pathogen, the share is even higher among children in Senegal. Strains of TW circulate inside healthy individuals in French families, where only one member evolves the disease. TW has been shown to be the most frequent pathogen in bronchial lavages of HIV-positive patients. The exact immunological predisposition to WD remains still unclear but will be the focus of further

experimental studies. The availability of 19 genomes can facilitate these studies in combining bacteriological and immunological research.

5.8. Prediction of antibiotic resistance genes: fluoroquinolone resistance gene

All 19 strains showed a genotypic resistance to fluoroquinolones linked to a mutation in the Gyr-A Gene in the RAST annotation. These resistance genes have already been described in the past for three strains of TW (*Twist*, SLOW2 and ENDO5) [69]. The concept has been proven in vitro with the same three strains [70]. Mutations in Gyr-A and the ParC-Gene have also been shown to promote fluoroquinolone resistance in other bacterial species, as for example Mycobacteria [66]. Therefore, the mutation described in all strains shows that TW is not susceptible to fluoroquinolones in general, *in vitro* or in *silico*.

5.9. Limitations of this study

Firstly, SOLiD sequencing was chosen for this pangenomic study because it allows a very high resolution in sequencing and can show SNPs better than first-generation sequencing methods. Accuracy of SOLiD sequencing is reported to be at 99.94 % [71]. Nonetheless, errors in sequencing could falsify our phylogenetic results, seeing that the genetic differences of TW are so small.

Secondly, a pangenomic study requires as many genomes as possible. Due to slow bacterial growth and difficult DNA-extraction in TW, only 17 out of 29 planned strains could be sequenced. With Twist and TW08/27 19 genomes could be used for calculating the pan-genome.

Thirdly, *Twist* and *TW08/27* were sequenced 2003 in other projects using the Sanger method [5,6]. Annotation showed more similarities between *Twist* and *TW08/27*, although the same annotation method was used for all genomes (RAST). This effect could be increased by the fact that six genomes were only assembled to the contig level. Nevertheless, these six did not group together.

5.10. Conclusion

WD remains a very rare disease of which mainly white Caucasian middle-aged men are affected. The bacterium can be considered a commensal and could be found in many individuals in Central Europe, Africa, North and South America. Genetically different strains of TW show a very close relationship. So far, the classical genotyping method could trace epidemiological relationships. An increase of typing resolution was made possible by the implementation of genomotypes. But even the increased typing resolution could not show any relationship between

genetic content and clinical manifestation. Genetic difference did not increase with increasing geographic distance either. TW seems to have a closed pan-genome.

So firstly, TW can be considered a bacterium that is living in humans for a long period of time and therefore a commensal. Secondly, WD only evolves under certain immunological premises and in predisposed individuals. This predisposition has to be investigated much further in the future in order to understand this potentially lethal disease.

6. Bibliography

1. La Scola B, Fenollar F, Fournier PE, Altwegg M, Mallet MN, Raoult D. Description of Tropheryma whipplei gen. nov., sp. nov., the Whipple's disease bacillus. Int J Syst Evol Microbiol. 2001;51(Pt 4):1471-1479.

2. Whipple GH. A Hitherto Undescribed Disease Characterized Anatomically by Deposits of Fat and Fatty Acids in the Intestinal and Mesenteric Lymphatic Tissues. Baltimore: Johns Hopkins Hospital; 1907.

3. Yardley JH, Hendrix TR. Combined electron and light microscopy in Whipple's disease. Demonstration of "bacillary bodies" in the intestine. Bull Johns Hopkins Hosp. 1961;109:80-98.

4. Raoult D, Birg ML, La Scola B, Fournier PE, Enea M, Lepidi H, Roux V, Piette JC, Vandenesch F, Vital-Durand D, Marrie TJ. Cultivation of the bacillus of Whipple's disease. N Engl J Med. 2000;342(9):620-625.

5. Raoult D, Ogata H, Audic S, Robert C, Suhre K, Drancourt M, Claverie J-M. Tropheryma whipplei Twist: a human pathogenic Actinobacteria with a reduced genome. Genome Res. 2003;13(8):1800-1809.

6. Bentley SD, Maiwald M, Murphy LD, Pallen MJ, Yeats CA, Dover LG, Norbertczak HT, Besra GS, Quail MA, Harris DE, von Herbay A, Goble A, Rutter S, Squares R, Squares S, Barrell BG, Parkhill J, Relman DA. Sequencing and analysis of the genome of the Whipple's disease bacterium Tropheryma whipplei. Lancet Lond Engl. 2003;361(9358):637-644.

7. Renesto P, Crapoulet N, Ogata H, La Scola B, Vestris G, Claverie J-M, Raoult D. Genome-based design of a cell-free culture medium for Tropheryma whipplei. Lancet Lond Engl. 2003;362(9382):447-449.

8. Dobbins WO. Whipple's disease. Mayo Clin Proc. 1988;63(6):623-624.

9. Durand DV, Lecomte C, Cathébras P, Rousset H, Godeau P. Whipple disease. Clinical review of 52 cases. The SNFMI Research Group on Whipple Disease. Société Nationale Française de Médecine Interne. Medicine (Baltimore). 1997;76(3):170-184.

10. Fenollar F, Keita AK, Buffet S, Raoult D. Intrafamilial circulation of Tropheryma whipplei, France. Emerg Infect Dis. 2012;18(6):949-955.

11. Keita AK, Raoult D, Fenollar F. Tropheryma whipplei as a commensal bacterium. Future Microbiol. 2013;8(1):57-71.

12. Fenollar F, Trani M, Davoust B, Salle B, Birg M-L, Rolain J-M, Raoult D. Prevalence of asymptomatic Tropheryma whipplei carriage among humans and nonhuman primates. J Infect Dis. 2008;197(6):880-887.

13. Keita AK, Bassene H, Tall A, Sokhna C, Ratmanov P, Trape J-F, Raoult D, Fenollar F. Tropheryma whipplei: a common bacterium in rural Senegal. PLoS Negl Trop Dis. 2011;5(12):e1403.

14. Fenollar F, Trape J-F, Bassene H, Sokhna C, Raoult D. Tropheryma whipplei in fecal samples from children, Senegal. Emerg Infect Dis. 2009;15(6):922-924.

15. Desnues B, Al Moussawi K, Fenollar F. New insights into Whipple's disease and Tropheryma whipplei infections. Microbes Infect Inst Pasteur. 2010;12(14-15):1102-1110.

16. Lagier J-C, Lepidi H, Raoult D, Fenollar F. Systemic Tropheryma whipplei: clinical presentation of 142 patients with infections diagnosed or confirmed in a reference center. Medicine (Baltimore). 2010;89(5):337-345.

17. Fenollar F, Puéchal X, Raoult D. Whipple's disease. N Engl J Med. 2007;356(1):55-66.

18. Geissdörfer W, Moos V, Moter A, Loddenkemper C, Jansen A, Tandler R, Morguet AJ, Fenollar F, Raoult D, Bogdan C, Schneider T. High frequency of Tropheryma whipplei in culture-negative endocarditis. J Clin Microbiol. 2012;50(2):216-222.

19. Stojan G, Melia MT, Khandhar SJ, Illei P, Baer AN. Constrictive pleuropericarditis: a dominant clinical manifestation in Whipple's disease. BMC Infect Dis. 2013;13:579.

20. Fenollar F, Nicoli F, Paquet C, Lepidi H, Cozzone P, Antoine J-C, Pouget J, Raoult D. Progressive dementia associated with ataxia or obesity in patients with Tropheryma whipplei encephalitis. BMC Infect Dis. 2011;11:171.

21. Gerard A, Sarrot-Reynauld F, Liozon E, Cathebras P, Besson G, Robin C, Vighetto A, Mosnier J-F, Durieu I, Vital Durand D, Rousset H. Neurologic presentation of Whipple disease: report of 12 cases and review of the literature. Medicine (Baltimore). 2002;81(6):443-457.

22. Raoult D, Fenollar F, Rolain JM, Minodier P, Bosdure E, Li W, Garnier JM, Richet H. Tropheryma whipplei in children with gastroenteritis. Emerg Infect Dis. 2010;16(5):776-782.

23. Fenollar F, Ponge T, La Scola B, Lagier J-C, Lefebvre M, Raoult D. First isolation of Tropheryma whipplei from bronchoalveolar fluid and clinical implications. J Infect. 2012;65(3):275-278.

24. Stein A, Doutchi M, Fenollar F, Raoult D. Tropheryma whipplei pneumonia in a patient with HIV-2 infection. Am J Respir Crit Care Med. 2013;188(8):1036-1037.

25. Lozupone C, Cota-Gomez A, Palmer BE, Linderman DJ, Charlson ES, Sodergren E, Mitreva M, Abubucker S, Martin J, Yao G, Campbell TB, Flores SC, Ackerman G, Stombaugh J, Ursell L, Beck JM, Curtis JL, Young VB, Lynch SV, Huang L, Weinstock GM, Knox KS, Twigg H, Morris A, Ghedin E, Bushman FD, Collman RG, Knight R, Fontenot AP, Lung HIV Microbiome Project. Widespread colonization of the lung by Tropheryma whipplei in HIV infection. Am J Respir Crit Care Med. 2013;187(10):1110-1117.

26. Schöniger-Hekele M, Petermann D, Weber B, Müller C. Tropheryma whipplei in the environment: survey of sewage plant influxes and sewage plant workers. Appl Environ Microbiol. 2007;73(6):2033-2035.

27. Moos V, Kunkel D, Marth T, Feurle GE, LaScola B, Ignatius R, Zeitz M, Schneider T. Reduced peripheral and mucosal Tropheryma whipplei-specific Th1 response in patients with Whipple's disease. J Immunol Baltim Md 1950. 2006;177(3):2015-2022.

28. Geelhaar A, Moos V, Schinnerling K, Allers K, Loddenkemper C, Fenollar F, LaScola B, Raoult D, Schneider T. Specific and nonspecific B-cell function in the small intestines of patients with Whipple's disease. Infect Immun. 2010;78(11):4589-4592.

29. Al Moussawi K, Ghigo E, Kalinke U, Alexopoulou L, Mege J-L, Desnues B. Type I interferon induction is detrimental during infection with the Whipple's disease bacterium, Tropheryma whipplei. PLoS Pathog. 2010;6(1):e1000722.

30. Ghigo E, Barry AO, Pretat L, Al Moussawi K, Desnues B, Capo C, Kornfeld H, Mege J-L. IL-16 promotes T. whipplei replication by inhibiting phagosome conversion and modulating macrophage activation. PloS One. 2010;5(10):e13561.

31. Gorvel L, Al Moussawi K, Ghigo E, Capo C, Mege J-L, Desnues B. Tropheryma whipplei, the Whipple's disease bacillus, induces macrophage apoptosis through the extrinsic pathway. Cell Death Dis. 2010;1:e34.

32. Moos V, Schmidt C, Geelhaar A, Kunkel D, Allers K, Schinnerling K, Loddenkemper C, Fenollar F, Moter A, Raoult D, Ignatius R, Schneider T. Impaired immune functions of monocytes and macrophages in Whipple's disease. Gastroenterology. 2010;138(1):210-220.

33. Martinetti M, Biagi F, Badulli C, Feurle GE, Müller C, Moos V, Schneider T, Marth T, Marchese A, Trotta L, Sachetto S, Pasi A, De Silvestri A, Salvaneschi L, Corazza GR. The HLA alleles DRB1*13 and DQB1*06 are associated to Whipple's disease. Gastroenterology. 2009;136(7):2289-2294.

34. Biagi F, Badulli C, Feurle GE, Müller C, Moos V, Schneider T, Marth T, Mytilineos J, Garlaschelli F, Marchese A, Trotta L, Bianchi PI, Di Stefano M, Cremaschi AL, De Silvestri A, Salvaneschi L, Martinetti M, Corazza GR. Cytokine genetic profile in Whipple's disease. Eur J Clin Microbiol Infect Dis Off Publ Eur Soc Clin Microbiol. 2012;31(11):3145-3150.

35. Biagi F, Schiepatti A, Badulli C, Sbarsi I, Trotta L, Feurle GE, Müller C, Moos V, Schneider T, Marth T, De Amici M, Martinetti M, Corazza GR. -295 T-to-C promoter region IL-16 gene polymorphism is associated with Whipple's disease. Eur J Clin Microbiol Infect Dis Off Publ Eur Soc Clin Microbiol. 2015;34(9):1919-1921.

36. Lagier J-C, Fenollar F, Lepidi H, Raoult D. Evidence of lifetime susceptibility to Tropheryma whipplei in patients with Whipple's disease. J Antimicrob Chemother. 2011;66(5):1188-1189.

37. Lagier J-C, Fenollar F, Lepidi H, Giorgi R, Million M, Raoult D. Treatment of classic Whipple's disease: from in vitro results to clinical outcome. J Antimicrob Chemother. 2014;69(1):219-227.

38. Feurle GE, Moos V, Bläker H, Loddenkemper C, Moter A, Stroux A, Marth T, Schneider T. Intravenous ceftriaxone, followed by 12 or three months of oral treatment with trimethoprim-sulfamethoxazole in Whipple's disease. J Infect. 2013;66(3):263-270.

39. Feurle GE, Junga NS, Marth T. Efficacy of ceftriaxone or meropenem as initial therapies in Whipple's disease. Gastroenterology. 2010;138(2):478-486-12.

40. Fenollar F, Perreal C, Raoult D. Tropheryma whipplei natural resistance to trimethoprim and sulphonamides in vitro. Int J Antimicrob Agents. 2014;43(4):388-390.

41. Dobbins WO. The diagnosis of Whipple's disease. N Engl J Med. 1995;332(6):390-392.

42. Clarridge JE. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clin Microbiol Rev. 2004;17(4):840-862, table of contents.

43. Moter A, Schmiedel D, Petrich A, Wiessner A, Kikhney J, Schneider T, Moos V, Göbel UB, Reischl U. Validation of an rpoB gene PCR assay for detection of Tropheryma whipplei: 10 years' experience in a National Reference Laboratory. J Clin Microbiol. 2013;51(11):3858-3861.

44. Fenollar F, Raoult D. Molecular genetic methods for the diagnosis of fastidious microorganisms. APMIS Acta Pathol Microbiol Immunol Scand. 2004;112(11-12):785-807.

45. Li W, Fenollar F, Rolain J-M, Fournier P-E, Feurle GE, Müller C, Moos V, Marth T, Altwegg M, Calligaris-Maibach RC, Schneider T, Biagi F, La Scola B, Raoult D. Genotyping reveals a wide heterogeneity of Tropheryma whipplei. Microbiol Read Engl. 2008;154(Pt 2):521-527.

46. Bonhomme CJ, Renesto P, Desnues B, Ghigo E, Lepidi H, Fourquet P, Fenollar F, Henrissat B, Mege J-L, Raoult D. Tropheryma whipplei glycosylation in the pathophysiologic profile of Whipple's disease. J Infect Dis. 2009;199(7):1043-1052.

47. Wilson KH, Blitchington R, Frothingham R, Wilson JA. Phylogeny of the Whipple's-disease-associated bacterium. Lancet Lond Engl. 1991;338(8765):474-475.

48. Relman DA, Schmidt TM, MacDermott RP, Falkow S. Identification of the uncultured bacillus of Whipple's disease. N Engl J Med. 1992;327(5):293-301.

49. Drancourt M, Carlioz A, Raoult D. rpoB sequence analysis of cultured Tropheryma whippelii. J Clin Microbiol. 2001;39(7):2425-2430.

50. La M-V, Crapoulet N, Barbry P, Raoult D, Renesto P. Comparative genomic analysis of Tropheryma whipplei strains reveals that diversity among clinical isolates is mainly related to the WiSP proteins. BMC Genomics. 2007;8:349.

51. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome." Proc Natl Acad Sci U S A. 2005;102(39):13950-13955.

52. Rouli L, Merhej V, Fournier P-E, Raoult D. The bacterial pangenome as a new tool for analysing pathogenic bacteria. New Microbes New Infect. 2015;7:72-85.

53. Kowalczewska M, Villard C, Lafitte D, Fenollar F, Raoult D. Global proteomic pattern of Tropheryma whipplei: a Whipple's disease bacterium. Proteomics. 2009;9(6):1593-1616.

54. Hunter PR, Gaston MA. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J Clin Microbiol. 1988;26(11):2465-2466.

55. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol. 2013;30(12):2725-2729.

56. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, Vonstein V, Wattam AR, Xia F, Stevens R. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res. 2014;42(Database issue):D206-214.

57. Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Eberhard C, Grüning B, Guerler A, Hillman-Jackson J, Von Kuster G, Rasche E, Soranzo N, Turaga N, Taylor J, Nekrutenko A, Goecks J. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. 2016;44(W1):W3-W10.

58. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004;14(7):1394-1403.

59. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. PHAST: a fast phage search tool. Nucleic Acids Res. 2011;39(Web Server issue):W347-352.

60. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother. 2012;67(11):2640-2644.

61. Richter M, Rosselló-Móra R, Oliver Glöckner F, Peplies J. JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. Bioinforma Oxf Engl. 2016;32(6):929-931.

62. Ersts PJ. Geographic Distance Matrix Generator (version 1.2.3.). American Museum of Natural History, Center for Biodiversity and Conservation. (Accessed October 9th 2016 at http://biodiversityinformatics.amnh.org/open_source/gdmg/).

63. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 2009;106(45):19126-19131.

64. Contreras-Moreira B, Vinuesa P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol. 2013;79(24):7696-7701.

65. Wetzstein N, Fenollar F, Buffet S, Moos V, Schneider T, Raoult D. Tropheryma whipplei genotypes 1 and 3, Central Europe. Emerg Infect Dis. 2013;19(2):341-342.

66. Sierra JM, Martinez-Martinez L, Vázquez F, Giralt E, Vila J. Relationship between Mutations in the gyrA Gene and Quinolone Resistance in Clinical Isolates of Corynebacterium striatum and Corynebacterium amycolatum. Antimicrob Agents Chemother. 2005;49(5):1714-1719.

67. Mann S, Chen Y-PP. Bacterial genomic G+C composition-eliciting environmental adaptation. Genomics. 2010;95(1):7-15.

68. Finlay BB, McFadden G. Anti-Immunology: Evasion of the Host Immune System by Bacterial and Viral Pathogens. Cell. 2006;124(4):767-782.

69. Masselot F, Boulos A, Maurin M, Rolain JM, Raoult D. Molecular Evaluation of Antibiotic Susceptibility: Tropheryma whipplei Paradigm. Antimicrob Agents Chemother. 2003;47(5):1658-1664.

70. Boulos A, Rolain JM, Mallet MN, Raoult D. Molecular evaluation of antibiotic susceptibility of Tropheryma whipplei in axenic medium. J Antimicrob Chemother. 2005;55(2):178-181.

71. Applied Biosystems. SOLiD™ System Accuracy. 2008. (Accessed December 31st at http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_057511.pdf).

7. Eidesstattliche Erklärung

„Ich, Nils Wetzstein, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Thema: « Genotyping and Genomotyping of Tropheryma whipplei – The Causative Agent of Whipple's Disease » selbstständig und ohne nicht offengelegte Hilfe Dritter verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe.

Alle Stellen, die wörtlich oder dem Sinne nach auf Publikationen oder Vorträgen anderer Autoren beruhen, sind als solche in korrekter Zitierung (siehe „Uniform Requirements for Manuscripts (URM)“ des ICMJE -*www.**icmje**.org*) kenntlich gemacht. Die Abschnitte zu Methodik (insbesondere praktische Arbeiten, Laborbestimmungen, statistische Aufarbeitung) und Resultaten (insbesondere Abbildungen, Graphiken und Tabellen) entsprechen den URM (s.o) und werden von mir verantwortet.

Meine Anteile an etwaigen Publikationen zu dieser Dissertation entsprechen denen, die in der untenstehenden gemeinsamen Erklärung mit dem Betreuer, angegeben sind. Sämtliche Publikationen, die aus dieser Dissertation hervorgegangen sind und bei denen ich Autor bin, entsprechen den URM (s.o) und werden von mir verantwortet.

Die Bedeutung dieser eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unwahren eidesstattlichen Versicherung (§156,161 des Strafgesetzbuches) sind mir bekannt und bewusst.“

Datum **Unterschrift**

8. Anteilserklärung an erfolgten Publikationen

Nils Wetzstein hatte folgenden Anteil an den folgenden Publikationen:

Publikation 1: Nils Wetzstein, Florence Fenollar, Sylvain Buffet, Verena Moos, Thomas Schneider, und Didier Raoult ; *Tropheryma whipplei* Genotypes 1 and 3, Central Europe ; Emerging Infectious diseases ; 19, 341–342, 2013.

Beitrag im Einzelnen: Konzeption des Studiendesigns, Neuauswertung vorliegender Genotypisierungsdaten, Komplettierung fehlender Genotypisierungsdaten, Schreiben des Letters, Führen der Korrespondenz im Peer review.

Unterschrift, Datum und Stempel des betreuenden Hochschullehrers

Unterschrift des Doktoranden

9. Curriculum Vitae

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

10. Acknowledgments

Zunächst möchte ich Prof. Didier Raoult und Prof. Florence Fenollar für die Überlassung des Themas, sowie für die Möglichkeit in ihrer Forschungseinrichtung, dem URMITE Marseille, zu arbeiten, recht herzlich danken. Desweiteren möchte ich mich recht herzlich bei allen Teammitgliedern des URMITE bedanken, ohne deren Hilfe die obenstehenden Experimente niemals zustande gekommen wären. Es sind zu viele hilfreiche Kollegen, um sie alle beim Namen zu nennen.

Insbesondere möchte ich mich aber auch bei den Mitlgiedern des Génoscope bedanken, dass sie die Genomsequenzierung durchgeführt und mir die anschließenden Daten zur Verfügung gestellt haben. Zu erwähnen ist hier vor allem Dr. Laetitia Rouli, die die Assemblierung des Genoms durchgeführt hat.

Schließlich gilt mein Dank meinem Doktorvater, Prof. Thomas Schneider. Für die ausführliche und konstruktive Betreuung möchte ich recht herzlich Dr. Verena Moos danken. Ohne sie hätte ich diese Arbeit nicht zu Ende geführt.

Zu erwähnen ist auch die Studienstiftung des deutschen Volkes, ohne deren Finanzierung und ideelle Unterstürzung meine Laborarbeit in Frankreich nicht möglich gewesen wäre.

Zu guter letzt möchte ich allen guten Freunden und meiner Familie danken, insbesondere meiner Freundin Sophie, für Ihre Geduld und Unterstützung während der Fertigstellung dieser Arbeit.