

**Epigenetic variability in the facultative human pathogen
*Neisseria meningitidis***

INAUGURALDISSERTATION

zur
Erlangung der Würde eines Doktors der Philosophie

vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Mohamad Rustom Abdul Sater

aus Baalbeck, Libanon

Basel 2015

Originaldokument gespeichert auf dem Dokumentenserver der Universität Basel
edoc.unibas.ch



This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von

Prof. Dr. Marcel Tanner

Dr. Christoph Schmid

Prof. Dr. Urs Jenal

Basel, den 21.4.2015

Prof. Dr. Jörg Schibler

Dekan

Table of Contents

1	General Introduction.....	1
1.1	Epigenetics.....	1
1.2	DNA methylation and Prokaryotic Epigenetics	2
1.3	Detection of DNA modifications using SMRT sequencing.....	4
1.4	<i>Neisseria meningitidis</i>	6
1.4.1	Meningitis belt.....	6
1.4.2	Classification	7
1.4.3	<i>Neisseria</i> genomic plasticity	8
	Recombination	9
	Phase-variation.....	10
2	Objectives.....	13
3	DNA methylation assessed by SMRT sequencing is linked to mutations in <i>Neisseria meningitidis</i> isolates.....	15
3.1	Abstract:	16
3.2	Introduction.....	17
3.2.1	<i>Neisseria</i> : pathogenicity and genomic plasticity	17
3.2.2	Prokaryotic epigenetics and detection of DNA modifications using SMRT.....	17
3.3	Materials and methods	19
3.3.1	Cultivation of strains of <i>N. meningitidis</i> , isolation of genomic DNA	19
3.3.2	Methylation sensitive restriction digest.....	19
3.3.3	SMRT sequencing	19
3.3.4	Local deviations in positional distributions of methylation motifs	20
3.3.5	Identification of DNA methyltransferase genes	20
3.3.6	SNP calling	20
3.3.7	Co-occurrence of SNPs at methylation motifs	21
3.4	Results	22
3.4.1	SMRT sequencing determines divergent DNA modification profiles	22
3.4.2	Methylation target motifs with biased distributions in regulatory genomic regions ...	24
3.4.3	Variable set of active DNA methyltransferase genes in serogroup A <i>N. meningitidis</i> isolates	25
3.4.4	Mutations overrepresented at DNA methylation target motifs	26
3.5	Discussion	27
	<i>Detection limits for 5mC modifications</i>	28
	<i>Comparable sequencing accuracy of SMRT sequencing</i>	29

<i>Sequence variability in clonal populations</i>	29
<i>Functional consequences of highly variable DNA modifications</i>	29
<i>Conclusions</i>	31
<i>Acknowledgments</i>	31
3.6 Tables.....	33
3.7 Figures	34
3.8 Supplementary material.....	41
3.9 References.....	43
4 Exploring the phase-variable genome of <i>Neisseria meningitidis</i> from massively parallel sequencing data	47
4.1 Abstract	48
4.2 Introduction.....	49
4.3 Methods	52
4.3.1 Identify short tandem repeats using Phobos	52
4.3.2 Regular expression fast approach for an exact sequence matching.....	52
4.3.3 RepHMM an exhaustive approach for approximate sequence matching.....	53
4.3.4 Generation of simulated data and comparison of performance	54
4.3.5 Bacterial cultivation, PCR and Sanger Sequencing	55
4.3.6 SNP calling	55
4.4 Results	57
4.4.1 Development of a flexible microsatellite repeat typing tool (RepHMM)	57
4.4.2 RepHMM outperforms alternative approaches	58
4.4.3 Evaluating RepHMM at multi-copy gene duplicates	63
4.4.4 Integrated Pipeline for identification of Phase-variable Genome	64
Pipeline development	64
<i>Analysis of pipeline output</i>	67
4.4.5 A Predominant OFF state of type III methyltransferases.	70
4.4.6 Repeat region length evolution at <i>modA12</i> locus.....	71
4.5 Discussion	73
<i>Summary</i>	73
<i>Consistent mutation rate of phase-variable short tandem repeats</i>	73
<i>Pipeline adaptable to other sequencing technologies</i>	74
<i>Closely related bacterial population ideal for identifying phase-variable genome</i>	74
<i>Phase-variable surface components and regulatory genes</i>	75
<i>Phase-variation mediates a reduced expression status of genes</i>	75

<i>Acknowledgments</i>	76
4.6 Supplementary materials	77
5 General Discussion	81
5.1 Summary.....	81
5.2 DNA methylation interplay with other adaptation mechanisms	82
5.3 Relevance to public health	83
5.4 Considerations for future sequencing approaches	84
6 References	87
7 Abbreviations	97
8 Acknowledgements	99
9 CURRICULUM VITAE	101

1 General Introduction

Heritable genetic information is encoded by DNA (Watson and Crick, 1953) which is transcribed into RNA and ultimately translated to proteins, determining thus the cells regulatory mechanisms, metabolic function as well as structure and shape. Cells sharing same DNA (genotype) differentiate to form multicellular organisms with a wide range of functional and structural characteristics (phenotype). Apart from a set of essential housekeeping genes, cells exhibit a highly regulated tissue specific gene expression patterns. Bone, muscle, skin are tissues made up of fundamentally divergent cell types despite having the same genotype.

Phenotypic heterogeneity is not restricted to differentiated cells in multicellular organisms. In microbiology, advances in single cell analysis revealed cell-to-cell variability to be common within populations of isogenic bacteria. A few examples of phenotypic variability evident in bacteria include; persistence where dormant bacterial cells are formed within the population leading to antibiotic resistance (Balaban et al., 2004), lactose utilization and chemotaxis in *Escherichia coli* (Davidson and Surette, 2008), bistability in genes expression of extracellular matrix and spore formation in *Bacillus subtilis* (Chai et al., 2007; Veening et al., 2008). Hence a genome cannot be considered as a deterministic phenotypic blueprint. Instead a phenotype is a product of combinatorial expression and repression of specific sets of genes.

1.1 Epigenetics

“If you want to start an argument, ask the person who just said “epigenetic” what it really means” (Language and dispute 2008). Although it is undisputed that epigenetics is of high biological importance there is still a big debate on what it actually means. The Greek term “Epi” translates to “above” or “in addition”. The term Epigenetics is used typically to describe heritable non-genetic based changes in gene expression.

Before epigenetics became a major term in current scientific vocabulary as well as a vital field of research the word had historically been used to describe two different biological processes. In 1957, Conrad Waddington first used the term epigenetic to link two separate fields at the time, developmental biology and genetics, describing Epigenesis that is the development of a phenotype from genotype (Waddington, 1957). In 1958, epigenetics was used by David Nanney to describe inherited events deviating from conventional genetics.

About 30 years later in the 1990s, an explosion in the use of the word epigenetics started notably after reports of DNA methylation status being transmitted through the germ line and altering gene expression (Doerfler, 1981; Harrison et al., 1983; Jones, 1985; Holliday, 2006). In his book *Epigenetic Mechanisms of Gene Regulation*, Arthur Riggs defined epigenetics as “the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence” (Russo et al., 1996). This definition seems nevertheless very lossless, assigning the unexplainable to epigenetics.

To present several biological processes including cellular differentiation, gene regulation, aging, response to environmental effectors, embryology, tumors, and other diseases have been directly linked to epigenetics. Such events are associated to direct DNA and chromatin modification marks, RNA interference, as well as higher order structures of chromosomes and nucleus (Holliday, 2006).

Epigenetics redefines great works of Mendel and Darwin and is suggested to revive Lamarckian theories mainly through the controversial heritable response to environmental stress (Heard and Martienssen, 2014).

1.2 DNA methylation and Prokaryotic Epigenetics

Methylation is a chemical post-synthetic modification of DNA. In eukaryotes epigenetics has emerged as a significant phenotypic determinant, in addition to the sequence of nucleotides in a genome. Eukaryotic DNA methylation is predominantly mediated by the DNA methyltransferase (dnmt) gene family modifying cytosine bases into 5-methylcytosine (5mC) in the context of CpG dinucleotide or CpHpG (H = A, T, C). 5mC has been correlated with gene silencing and given its significance, 5mC has been termed the 5th base (Bird, 2007). Recently, additional DNA modifications such as 5-hydroxymethylcytosine or DNA glycosylation have been described in eukaryotes (Yu et al., 2012).

Unlike methylation in eukaryotes, prokaryotes exhibit chemically more diverse modifications including 6-methyladenine (6mA), 4-methylcytosine (4mC) and 5mC (Bestor, 1990; Pomraning et al., 2009). DNA methylation in bacteria is not confined to CpG dinucleotide, instead a diverse set of methyltransferases bind the DNA at specific sequence motifs adding a methyl-group at a particular base (Bestor, 1990). The DNA modification process starts by the binding of a methyltransferase enzyme to a sequence motif, followed by flipping of the target base and transferring a methyl group using S-adenosyl-L-methionine (AdoMet) as the group donor (Casadesús and Low, 2006).

Historically, prokaryotic DNA methylation was mainly characterized as part of restriction-modification systems (RMS) and its antiviral defense mechanisms (Arber and Linn, 1969). RMS typically consists of a restriction enzyme and a cognate methyltransferase sharing the same DNA target motif. Host genome is methylated and protected, whereas un-methylated foreign DNA, for instance of viral origin, gets cleaved by the restriction enzyme. The sheer diversity of RMS is collected in a dedicated database (REBASE) containing information from a large number of microorganisms (Roberts et al., 2010).

With time, the mechanisms and functions of DNA methylation revealed to be more complex. Evidence of hemi-methylated DNA and methylation-dependent restriction enzymes cleaving DNA only in the presence of specific DNA methylation patterns challenged the assigned function of RMS as a defense mechanism (Camacho and Casadesús, 2005). Similarly, the identification of “orphan” methyltransferases, such as the DNA Adenine Methylase (Dam), in several organisms lacking a cognate restriction enzyme suggested methylation to have additional biological aspects (Palmer and Marinus, 1994). Adding to the puzzle was the characterization of many different DNA methyltransferases potentially active within a bacterial cell (Ishikawa et al., 2010).

A detailed exploration of numerous bacterial transcriptomes suggests that transcription in bacteria resembles that of eukaryotes in terms of complexity more closely than was previously thought (Güell et al., 2011). Although considered simpler model organisms to study, prokaryotic epigenetics gained less attraction compared to eukaryotes due to the lack of tools to detect the diverse types and specificities of methylation in bacteria. Until recently, DNA methylation detection methods mostly targeted the simpler 5mC CpG patterns of eukaryotes, however novel sequencing technologies (section 1.3) have closed this gap enabling a more broad detection of DNA modifications and paving the way for a boom in bacterial epigenetic research (Flusberg et al., 2010).

With the advances in DNA methylation detection, recent studies link prokaryotic DNA methylation with several biological functions. In *Mycobacterium tuberculosis*, the *mamA* methyltransferase activity was correlated with the regulation of gene expression ensuring hypoxic survival (Shell et al., 2013). *Caulobacter crescentus ccrM* methyltransferase regulation was in return found to be crucial for cell cycle regulation (Kozdon et al., 2013). A similar mechanism was also proposed for *Mycoplasma pneumoniae* (Lluch-Senar et al., 2013). Epigenetic effects are further postulated to enabling a small fraction of cells in isogenic *Mycobacteria* populations to resist antibiotics (Wakamoto et al., 2013). In addition, genetic re-arrangements was reported to generate genetic and epigenetic diversity in cell populations of *Streptococcus pneumoniae* and *Helicobacter pylori* (Manso et al., 2014; Furuta et al., 2014). DNA methylation was also linked to *E. coli mtuH* DNA mismatch

recognition and repair mechanisms (Casadesús and Low, 2013a).

1.3 Detection of DNA modifications using SMRT sequencing

Pacific Biosciences' Single Molecule Real-Time (SMRT) sequencing is based on the direct monitoring of DNA polymerase during its processing of single DNA molecules. The monitoring of individual DNA polymerase complexes is enabled by the scattering of light through a small aperture in the bottom of Zero Mode Waveguides (ZMW) chambers with ~70 nm diameter (Figure 1-1A). The laser light from below penetrates only the lower 20-30 nm of the ZMW where a DNA molecule and polymerase complex is immobilized at the bottom, reducing interference and background noise (Flusberg et al., 2010). If associated with an extension reaction, phosphor-linked fluorescent labels on nucleotides are cleaved and light pulse is emitted. Not only the color and thereby the nucleotide sequence is recorded, but also the kinetics of base incorporations by engineered DNA polymerase complexes. Due to direct effects of steric hindrance, different modifications may be discriminated by divergent specific kinetic signatures (Figure 1-1B). Modifications displaying weak SMRT signals such as 5mC can be enzymatically oxidized and converted with the Tet1 enzymes into the bulkier 5-carboxylcytosine (5caC) (Tahiliani et al., 2009) and thereby the SMRT signal on 5mC is enhanced, while maintaining detectability of other modifications (Clark et al., 2013). Reproducible alterations in the kinetics allow the detection of modifications on the specific DNA strand serving as template for DNA synthesis. The chemistry and analysis methods of this novel third generation DNA sequencing are continuously developing with modeling approaches further improving the localization of kinetic variation events (Fang et al., 2012).

Sequencing small genomes with SMRT represents a highly accurate sequencing method and the extended lengths of sequence reads can enable the closure of bacterial genomes. The determination of DNA modification profiles represents a unique advantage of SMRT technology, which may have a profound effect on our understanding of this biology (Roberts et al., 2013).

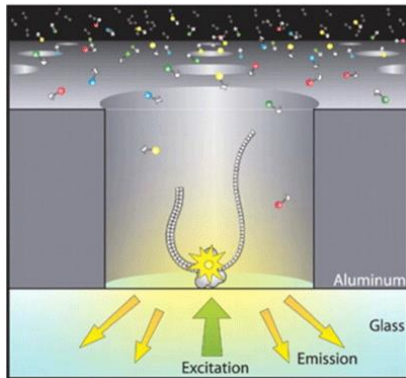
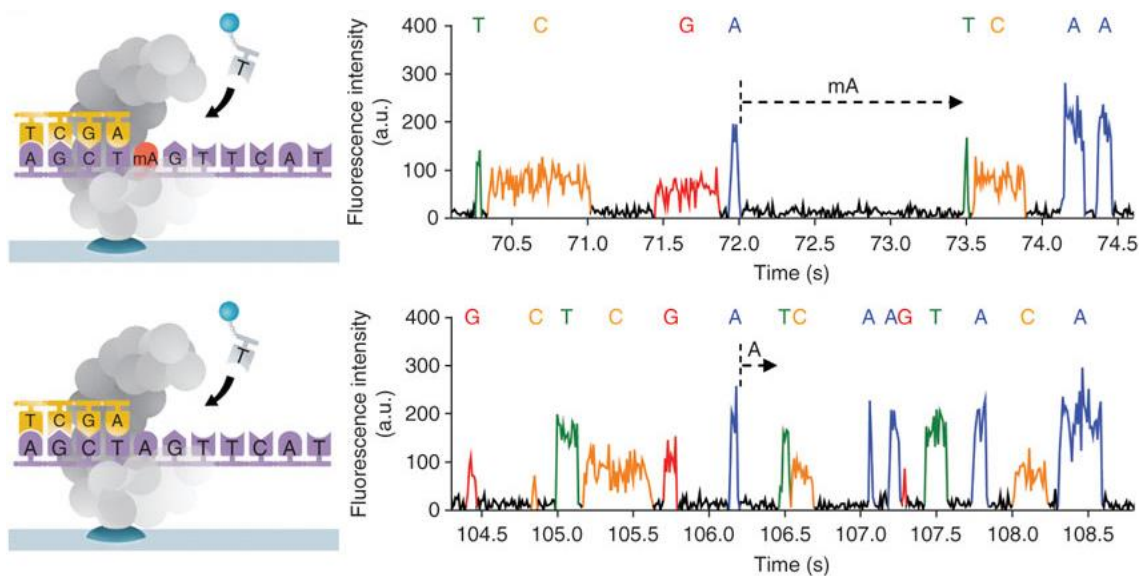
A**B**

Figure 1-1: (a) a cartoon of the ~70 nm Zero Mode Waveguides (ZMW) chambers with the polymerase-DNA molecule complex fixed at the bottom. (b) Schema of DNA Synthesis with methylated (top) and unmethylated (bottom) bases and the corresponding typical SMRT sequencing fluorescence traces of nucleotide incorporation rate. Letters on top refer to the nucleotides on the synthesized strand. Dashed arrows before thymine base indicate the Inter-pulse duration (IPD) of nucleotide incorporation. At the example of 6^mA in this sequence context, the IPD is 5 times larger than un-methylated Adenosine. Adapted from (Flusberg et al., 2010).

1.4 *Neisseria meningitidis*

Neisseria meningitidis is a gram-negative diplococcal proteobacterium and one of the major causes of bacterial meningitis and sepsis worldwide. Infectious meningococcal meningitis is likely a newly emerged health care problem (Greenwood, 2006). The disease was first described by Vieusseux after an outbreak in Geneva Switzerland causing 33 deaths in 1805 (Vieusseux, 1805). Shortly after in 1806 two American physicians Elias Mann and Lothario Danielson reported another outbreak in Massachusetts, USA (Danielson and Mann, 1806). In 1887, the Viennese doctor Anton Weichselbaum first isolated the bacterium from patients, it was initially named *Diplococcus intracellularis* (Weichselbaum, 1887).

Neisseria meningitidis is an obligate human pathogen and a regular commensal of the nasopharyngeal mucosa. The rate of asymptomatic carriers within a population is in general around 10% (Caugant and Maiden, 2009), but also varies considerably depending on age and conditions peaking among adolescents, as well as in military and university dormitories (Caugant and Maiden, 2009; Christensen et al., 2010). Early diagnosis is key to treat the disease, after onset of symptoms it could rapidly lead to disability or death in as short as 24 hours (Stephens et al., 2007). Consequently, meningococcal meningitis is a deadly disease worldwide especially in regions with little to no access to medical care.

In comparison to carriage, invasive meningococcal disease is rare with incidents rate between 0.5 to 1000 cases/100,000 individuals depending on the epidemiological region (Pizza and Rappuoli, 2015). In high income countries the disease is in continuous decline (0.15 per 100,000 in the USA in 2012) (Andrews and Pollard, 2014). However, Sub-Saharan Africa still suffers the highest burden of meningococcal meningitis (Leimkugel et al., 2007).

1.4.1 Meningitis belt

After an extensive survey, Lapeyssonnie published his “La méningite cérébro-spinale en Afrique” in 1963, where he first introduced the term Meningitis Belt. Lapeyssonnie’s comprehensive epidemiological survey allowed him to recognize unusual meningococcal infection patterns unique to the region bounded between the Sahara and the tropical forests. His pioneering work included a description of the disease causative agent, asymptomatic carriage, periodicity of epidemics, climate influence and clinical aspects (Lapeyssonnie, 1963) which are still effectively valid to present.

Within the meningitis belt (Figure 1-2) meningococcal disease occur in recurring epidemic

cycles every eight to 12 years leading up to 10,000 deaths annually (Leimkugel et al., 2007; Teyssou and Muros-Le Rouzic, 2007). Another peculiar behavior is the epidemics seasonality, peaking towards the end of the dry season, stopping abruptly at the arrival of the rainy season only to start over again with the start of the dry season. Typically during the dry season the temperatures can drop below 10 degrees leading to population congregation, and strong blowing winds (Harmattan) carrying fine desert dust disrupting nasopharyngeal mucosa likely to facilitate invasion (Molesworth et al., 2003).

1.4.2 Classification

Clinical infections have almost entirely been attributed to encapsulated strains. Among 13 *N. meningitidis* serogroups, defined based on the capsular polysaccharide structure, six serogroups (A, B, C, W-135, X, and Y) have mainly been responsible for the large majority of infections (Virji, 2009). The biochemical composition of the polysaccharide capsule determines strain's serogroup. Invasive serogroups A & X are formed by N-acetyl-d-mannosamine-6-phosphate and N-acetylglucosamine 1-phosphate whereas serogroups B,C, W-135 and Y capsule is composed of sialic acid (Tzeng et al., 2003). Nevertheless, recombination events allow *N. meningitidis* to alter its capsular polysaccharide phenotype (Swartley et al., 1997).

Until World War II, most epidemics in North America and Europe were attributed to serogroup A (Greenwood, 1999). Since then serogroup A practically diminished, responsible for <1% in the late 1990s of the cases, serogroup B is however causing the majority of cases as well as local outbreaks caused by serogroup C (Harrison et al., 2009). On the other hand, serogroup A remained a major contributor to infections in Asia and Africa, however the introduction of monovalent A conjugate vaccine in 2011 is reducing the infection rate caused by this serogroup (Daugla et al., 2014). Figure 1-2 denotes the distribution of the major serogroups worldwide.

Neisseria meningitidis serogroups can be further sub-classified into serotypes, subtypes and immunotypes based on outer membrane proteins and lipopolysaccharides antigens (Poolman et al., 1995). In addition, in 1998 Maiden et al. developed Multilocus Subtyping (MLST) whereby unique alleles defined by 500 basepair fragments of seven housekeeping genes are identified. The combination of alleles result in a sequence type (ST) and clusters of closely related strains are grouped into clonal complexes (CC) (Maiden et al., 1998). Over the past 20 years most serogroup A epidemics within the African meningitis belt were caused by ST 5, 7 and 2859 (Lamelas et al., 2014; Caugant et al., 2012; Teyssou and Muros-Le

Rouzic, 2007).



Figure 1-2: The African meningitis belt in grey shades with the global spread of *N. meningitidis* serogroups. Figure adapted from (Harrison et al., 2009)

1.4.3 *Neisseria* genomic plasticity

The asexual reproduction of bacteria through “binary fission”, whereby a mother cell divides into two genetically identical daughter cells (clones) theoretically limits genetic variation. Consequently, genotypic variants would only be the outcome of *de novo* mutagenesis and selection coupled with the accumulation and propagation across generations. Based on MLST profile, some organisms such as *Salmonella enterica*, *Mycobacterium tuberculosis* and *Bacillus anthracis* are indeed genetically clonal with largely uniform MLST profiles (Achtman, 2004; Boyd et al., 1997). On the other hand, although meningococcal clonal clusters remain detectable through their MLST profile, the introduction of pulsed field gel electrophoresis in 1990 quickly revealed a strikingly dynamic chromosomal structure of *N. meningitidis*. Some strains, even belonging to the same clonal complex, had evidently divergent restriction patterns (Smith et al., 1993; Bautsch, 1998; Gagneux et al., 2000; Schoen et al., 2009). Notably, large genomic rearrangements (40 kb deletion) were also observed during the course of infection of a single strain (Vogel and Frosch, 2002).

The high throughput sequencing era revealed more peculiar plasticity of the meningococcal genome. In the year 2000, whole genome sequences of two strains of serogroup B (Tettelin

et al., 2000) and serogroup A (Parkhill et al., 2000) were published. Since then a steadily increasing number of sequenced genomes is becoming available (18 closed genomes so far) which have shed the light on mechanism mediating remarkable genome plasticity allowing genetic and antigenic variation (Schoen et al., 2009). Comparative genomics of sequenced strains, including carriage and invasive strains as well as closely related commensal species like *Neisseria lactamica* helped to identify several virulence factors (Bentley et al., 2007). Nevertheless, no pathogenic genotypes or classic pathogenicity islands could be identified so far differentiating invasive from commensal isolates (Perrin et al., 2002; Virji, 2009).

Recombination

Meningococcal competence for natural transformation and recombination is a determinant factor driving its dynamic genome structure at the level of gene content and sequence diversity (Jolley et al., 2005). High frequency of recombination has been reported in several studies (Hao et al., 2011; Holmes et al., 1999; Kong et al., 2013; Lamelas et al., 2014). The genomic flexibility of *Neisseria meningitidis* allows for horizontal transfer of entire genes as well as intragenic fragments. Hence, mutations occurring in different genomes could be shared within the population thus having a profound impact on biological processes, phenotype and adaptation (Kong et al., 2013). The meningococcus has also been reported to acquire DNA fragments from other commensal species such as *N. cinerea* and *N. flavescens* sharing the nasopharyngeal environment (Bowler et al., 1994).

A diversity of genes associated with virulence and contributing to surface and antigenic variation in *N. meningitidis* have been reported as recombination hot spots including, penicillin binding proteins (*penA*), pili (*pil* locus) and adhesion genes (*maf* locus), surface antigens glycosylation genes (*pgl* locus) as well as capsule and vaccine target genes, (Bowler et al., 1994; Hao et al., 2011; Joseph et al., 2011; Kong et al., 2013; Lamelas et al., 2014).

Restriction modification systems (RMS) have been suggested as a tool utilized by a bacteria to control genetic exchange. RMS could theoretically block homologous recombination between strains having non-matching methylation patterns by cleaving such DNA fragments at corresponding recognition sequences (Budroni et al., 2011; Jeltsch, 2003). Such a mechanism remains controversial and species-specific. *Helicobacter pylori* for example, displays a significant correlation of methylation target sequences occurrence at end points of identified recombination fragments (Lin et al., 2009). On the other hand, several reports on recombination between different meningococcal clonal complexes or even different *Neisseria*

species having detectable divergent DNA methylation pattern have been published (Hao et al., 2011; Holmes et al., 1999). Recently, few studies have suggested a transient effect of restriction modification systems on recombination efficiency, attributed partially to the plasticity of RMS and differential distribution across clonal complexes as well as the meningococcus competence for single stranded DNA recombination (Budroni et al., 2011; Kong et al., 2013).

Phase-variation

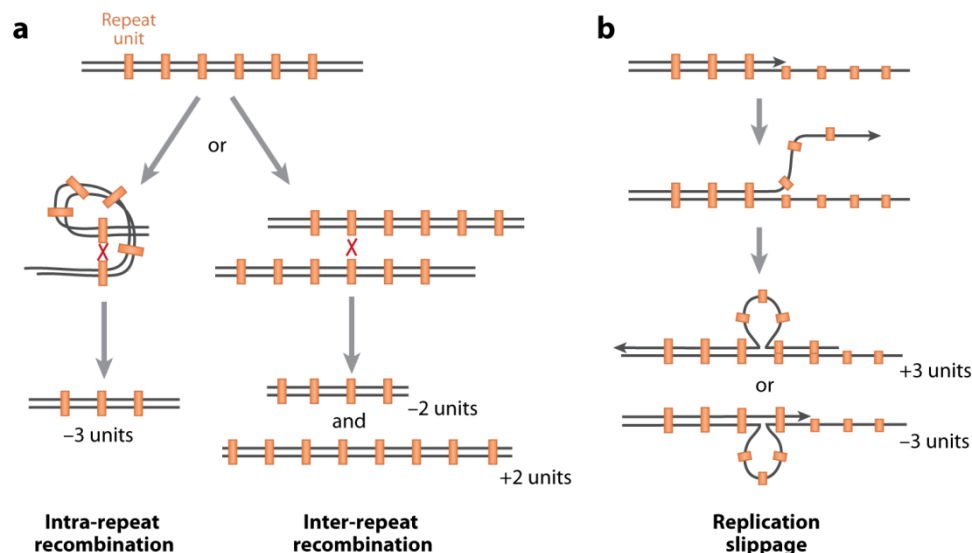
Adaptability is a vital strategy that allows pathogens to endure stress conditions such as rapidly changing environment, variable nutrient sources, host and tissue specific immune stress (Balaban et al., 2004; Zhou et al., 2014). A rapid response by a pathogen requires a prompt ability to modulate gene expression (Avery, 2006). Stochastic *de novo* mutations and selection does indeed produce genotypic variants, however random mutations are often deleterious, less likely to be reversible and the selection process occurs over several generations (Rando and Verstrepen, 2007). Phenotypic variability in clonal (genetically uniform) population of bacteria is observed in several pathogenic species. Transient phenotypes have been traditionally associated with non-genetic mechanism; nevertheless, reversible genotypic variations have been also identified to mediate phenotypic heterogeneity (Goldberg et al., 2014).

Special genomic sequences referred to as short tandem repeats (STRs) have been reported as unstable loci subject to reversible extension/contraction via insertion/deletion mutations of repeat units leading to divergent yet interchangeable phases. The average rate of typical mutations of a gene in bacteria is 10^{-9} mutations per division (Bayliss, 2009); certain microsatellite sequences (1-10 unit size) can however reach up to 10^{-3} mutations per division (Bayliss, 2009). This special reversible, localized, stochastic and rapid mechanism is termed phase-variation (van Ham et al., 1993).

Two hypothetical models have been proposed for the observed high frequency length variability of STRs associated with phase-variable genes: replication induced strand slippage (slipped-strand mispairing) and intra/inter repeat recombination (Figure 1-3). Although the first model is less characterized, it's presumed that self-pairing of the repeat region during the replication process causes DNA polymerase slippage. Looping of the nascent strand causes elongation of the repeat region, conversely looping of the template strand results in a shortening of the repeat region (Figure 1-3B) (Gemayel et al., 2010; Tachida and Iizuka, 1992). Recombination is however more potent in altering longer tandem repeats by unequal

crossing over (Figure 1-3A) (Zhou et al., 2014). In addition, a double strand breakage and repair model have been proposed leading to expansion and contraction of tandem repeat loci (Pâques et al., 1998).

Phase-variation provides pathogens with an additional layer of genome plasticity allowing some genes to be transiently expressed. Early reports of phase-variation described ON/OFF switching of Opa gene family in *Neisseria gonorrhoeae* caused by frameshifts introduced by a CTCTT microsatellite variable repeat region located inside the open reading frame (ORF) (Stern et al., 1986). Since then several phase-variable loci have been reported in pathogenic and commensal bacteria such as *Haemophilus influenzae*, *Neisseria meningitidis* and *Campylobacter jejuni* (Bayliss, 2009; Parkhill et al., 2000; Saunders et al., 2000). In *N. meningitidis* several phase-variable loci have been reported (Table 1-1), some of which are located within ORF sequence leading to ON/OFF switching by causing frameshifts (Saunders et al., 2000), others are located in promoter regions and could influence gene expression by complex mechanisms such as interaction with transcription factors binding sites or altering mRNA stability (Loh et al., 2013).




 Gemayel R, et al. 2010.
Annu. Rev. Genet. 44:445–77

Figure 1-3: Simplified models illustrating mechanisms of tandem repeat length variation. Source (Gemayel et al., 2010).

Growing evidence suggest phase-variation as one of the vital mechanisms triggering immune evasion through rapid antigenic variation and apt response to stress within a bacterial population. In addition, identification of an increasing number of phase-variable genes within the genome therefore allows for combinatorial expression profiles. A cell phasotype denotes the combinatorial expression state (ON/OFF) of a set of phase-variable genes. The phasevariome however, signify the cumulative percentage of an expression state of individual genes within the whole population (Bidmos and Bayliss, 2014).

Besides driving antigenic variation, phase-variation also contributes to epigenetic variability. In several bacterial species including meningococci, a number of methyltransferase genes are also reported to undergo phase-variation. These are mainly type I and type III restriction modification systems (Zhou et al., 2014). Meningococci have two phase-variable type III RMS genes (even three in some strains) which were reported to have an effect on gene expression (Table 1-1) (Seib et al., 2011; Srikhanta et al., 2009).

Table 1-1: Confirmed phase-variable loci in *N. meningitidis*.

Moiety	Locus	Microsatellite	Reference
Adhesins	<i>Opa</i>	CTCTT	(Stern et al., 1986)
	<i>NadA</i>	TAAA	(Martin et al., 2005)
Capsule	<i>siaD</i>	C	(Loh et al., 2013)
	<i>CssA</i>	TATACTTA	
Iron binding	<i>hpuA</i>	G	(Lewis et al., 1999)
	<i>hmbR</i>	G	(Tauseef et al., 2011)
Lipopolysaccharides	<i>lgtA</i>	C	(Saunders et al., 2000)
Outer membrane protein	<i>porA</i>	G	(Jennings et al., 1999)
Glycosylation	<i>pglA</i>	G	(Snyder et al., 2001)
Restriction-modification systems	<i>modA</i>	AGCC	(Srikhanta et al., 2009)
	<i>modB</i>	CCCAA	(Srikhanta et al., 2010)
	<i>modD</i>	ACCGA	(Seib et al., 2011)

2 Objectives

The aim of this PhD thesis was to develop bioinformatic tools to investigate genetic and epigenetic variation in *Neisseria meningitidis* population.

Specific aims:

- Assay and compare the DNA methylome of two serogroup A *Neisseria meningitidis* isolates using the recent single molecule real-time sequencing technology.
- Analyze the consequences of DNA methylation in the sequenced isolates.
- Develop bioinformatic tools to analyze microsatellite repeat length variation using sequencing data.
- Identify phase-variable genes using the developed tools by comparing a closely related set of meningococcal genomes.

3 DNA methylation assessed by SMRT sequencing is linked to mutations in *Neisseria meningitidis* isolates

Mohamad R. Abdul Sater^{1,2,5}, Araceli Lamelas^{1,2}, Guilin Wang³, Tyson A. Clark⁴, Katharina Roeltgen^{1,2}, Shrikant Mane³, Jonas Korlach⁴, Gerd Pluschke^{1,2}, and Christoph D. Schmid^{1,2,5,*}

¹ Swiss Tropical and Public Health Institute, Socinstrasse 57, P.O. Box, CH-4002 Basel, Switzerland;

² Universität Basel, Petersplatz 1, CH-4003 Basel, Switzerland;

³ Yale Center for Genomic Analysis, Yale University, CT 06516-0972, USA;

⁴ Pacific Biosciences, Menlo Park, CA 94025, USA;

⁵ SIB Swiss Institute of Bioinformatics, Basel 4056, Switzerland

This manuscript has been submitted to the Journal Genome Biology and Evolution on 18 March 2015.

3.1 Abstract:

The gram-negative prokaryote *Neisseria meningitidis* features extensive genetic variability. To present, proposed virulence genotypes are also detected in isolates from asymptomatic carriers, indicating more complex mechanisms underlying variable colonization modes of *N. meningitidis*.

We applied the SMRT sequencing method from Pacific Biosciences to assess the genome-wide DNA modification profiles of two closely related *N. meningitidis* strains of serogroup A. The resulting DNA methylomes revealed high divergence, represented by the detection of shared target motifs and of one novel strain-specific DNA methylation target motif. The positional distribution of these methylated target sites within the genomic sequences displayed clear biases, which suggests a functional role of DNA methylation related to the regulation of genes.

DNA methylation in *N. meningitidis* has a likely underestimated potential for variability, as evidenced by a careful analysis of the ORF status of a panel of confirmed and predicted DNA methyltransferase genes in an extended collection of *N. meningitidis* strains of serogroup A. Based on high coverage short sequence reads, we find phase-variability as a major contributor to the variability in DNA methylation. Taking into account the phase-variable loci, the inferred functional status of DNA methyltransferase genes matched the observed methylation profiles.

Towards an elucidation of presently incompletely characterized functional consequences of DNA methylation in *N. meningitidis*, we reveal a prominent co-localization of methylated bases with Single Nucleotide Polymorphisms (SNPs) detected within our genomic sequence collection.

These findings suggest a more diverse role of DNA methylation and Restriction-Modification systems in the evolution of prokaryotic genomes.

3.2 Introduction

3.2.1 *Neisseria*: pathogenicity and genomic plasticity

Neisseria meningitidis is a commensal Gram-negative bacterium exclusively found in the human nasopharyngeal mucosa and is readily transmitted via respiratory secretions or saliva (Trivedi et al. 2011). A small proportion of individuals colonized by a virulent strain may develop invasive disease including sepsis or meningitis (Caugant & Maiden 2009), especially devastating as epidemics in the African 'meningitis belt' (Leimkugel et al. 2007). Regular transmission events in meningitis outbreaks indicate that the disease causing invasive colonization mode is at least in part 'inheritable', in other words a bacterial population can maintain its 'invasive' phenotype. However not all transmissions necessarily lead to disease, a complex interplay of host-pathogen interactions influences the outcome of invasive infections (Stephens et al. 2007). Vaccination projects have dramatically lowered the incidence of meningococcal disease, yet the asymptomatic carriage and the high genetic variability of meningococci (Dunning Hotopp et al. 2006) might be responsible for occasional reemergence of epidemics (Maiden 2013). Genome sequencing of steadily increasing numbers of *N. meningitidis* strains suggested a number of genotypes associated with virulence including genes involved in the synthesis of the polysaccharide capsule. Yet to present no strict pathogenic genotype is defined which would allow to distinguish disease-causing strains from inoffensive carrier strains (Maiden 2008).

3.2.2 Prokaryotic epigenetics and detection of DNA modifications using SMRT

In eukaryotes, epigenetics has emerged as a significant phenotypic determinant representing an additional layer to the sequence of nucleotides in a genome, as showcased by the epigenetic roadmap project (Bernstein et al. 2010). DNA methylation in prokaryotes differs by more diverse modification types including 6-methyladenine (6mA), 4-methylcytosine (4mC) and 5-methylcytosine (5mC), deposited by a diverse set of methyltransferases at specific target sequences (motifs). Prokaryotic DNA methylation is therefore not concentrated to the CpG dinucleotide context and was in the past mainly characterized as part of restriction-modification (R-M) systems and its antiviral defense mechanisms cleaving any unmodified 'non-self' DNA (Arber 2000). Contemporary sequencing methods enable the determination of genome-wide epigenetic DNA modification maps. Pacific Biosciences' Single Molecule, Real-Time (SMRT) sequencing method is based on the direct monitoring of the processing of single DNA molecules by DNA polymerase (Eid et al. 2009). The kinetics of DNA synthesis enables the genome-wide determination of diverse DNA modifications

(Cao et al. 2014), which represents a unique advantage for studying prokaryotic epigenetics (Roberts et al. 2013). The approach has previously been successfully applied to the genome-wide mapping of methylated adenine and cytosine residues in multiple organisms including pathogenic *Escherichia coli* (Powers et al. 2013), *Helicobacter pylori* (Krebes et al. 2014), *Caulobacter crescentus* (Kozdon et al. 2013), and *Mycoplasma* (Lluch-Senar et al. 2013). SMRT sequencing enabled to determine previously unknown target sequences and the exact site of methylation of specific methyltransferases (Clark et al. 2012). Yet these experiments revealed also considerable divergence in the target sequences and/or methylation efficiency, if comparing homologous alleles of methylation enzymes in related strains differing by only a few amino acids (Furuta et al. 2014).

A number of studies in diverse prokaryotic systems have linked deficiencies in DNA methylation with altered gene expression patterns (Srikhanta et al. 2009), (Fang et al. 2012), (Furuta et al. 2014), (Manso et al. 2014). However the molecular mechanisms for direct effects of DNA methylation on prokaryotic gene expression are presently not elucidated, and only in single cases for instance a positional overlap of differentially methylated target sites with binding sites of transcription factors could be shown (Shell et al. 2013) (Kozdon et al. 2013). In many cases the detected methylation sites cannot be directly linked to a larger number of differentially expressed genes. Accordingly alternative molecular effects of DNA methylation are proposed, including interactions at the origin of replication and an involvement in genome replication (Bendall et al. 2013).

Variable DNA methylation in *Neisseria* species has been reported previously (Ritchot & Roy 1990), yet no direct association of the activity of a specific DNA adenine methyltransferase (Dam) with virulence was found (Jolley et al. 2004). More recently different alleles of the *mod* DNA methyltransferase gene family undergoing phase-variability were associated to divergent cellular phenotypes (Furuta et al. 2014).

Given the described variability in genomic sequences and phenotypes, we set out to investigate the epigenetic DNA modification profiles in *N. meningitidis* isolates. We determine DNA methylation target motifs (one or several DNA sequences), and our analysis reveals biased distributions of these target sequences in the genomes. We observe high variability in the methylation profiles among a population of closely related bacterial isolates. Strikingly, we also discover enrichments of SNPs at the precise positions of methylated bases in the genomes, pointing to a role of DNA methylation in the evolution of favorable genome configurations.

3.3 Materials and methods

3.3.1 Cultivation of strains of *N. meningitidis*, isolation of genomic DNA

Neisseria meningitidis reference strain Z2491 (DSM No. 15465) was obtained from DSMZ (Braunschweig, Germany). *N. meningitidis* isolates were previously collected over a time period of ~10 years during meningococcal meningitis epidemics in Sub-Saharan Africa (two sequence types ST2859 and ST7). Isolates underwent typically 2 rounds of single colony sub-culturing and overnight expansion *in vitro*. For genomic DNA preparation, strains were grown on supplemented GC agar base (Oxoid) plates for 20-24 hours in 5% CO₂ at 37°C. Single colonies were transferred into liquid Brain Heart Infusion (Bacto™) medium and again incubated overnight in 5% CO₂ at 37°C. Genomic DNA was extracted as described previously (Marri et al. 2010). SMRT sequencing of strain NM1264 was performed on aliquots of a genomic DNA sample previously subjected to the Illumina sequencing method (Lamelas et al. 2014).

3.3.2 Methylation sensitive restriction digest

NlaIV Restriction enzymes (methylation sensitive target sequence GGNNCC) were obtained from New England Biolabs (catalog #R0126) and used according to manufacturer specifications to digest 1 ug of genomic DNA of each strain.

3.3.3 SMRT sequencing

Genomic DNA preparations were sheared by sonication to ~500bp fragments, aiming at shorter reads with an increased coverage for DNA modification detection. To enhance detection of 5mC modifications, enzymatic conversion of 5- methylcytosine (5mC) to 5-carboxylcytosine (5caC) was carried out using the 5mC Tet1 oxidation kit (WiseGene) with an input of ~500ng of genomic DNA (Clark et al. 2013). Generation of SMRTbell libraries and SMRT sequencing were performed following manufacturer instructions (Flusberg et al. 2010) to obtain a strand-specific sequencing coverage of about 50X on a standard PacBio RS instrument at the Yale Center for Genomic Analysis. Sequencing reads were aligned to Z2491 reference genome (AL157959) or to the genome assembly of strain NM1264 (344 contigs in supp. dataset 6). To identify modified positions, we used Pacific Biosciences' SMRTPortal analysis platform, v. 1.3.1. In brief, at each genomic position, modification scores (modQV) were computed as the -10 log of a p-value for

representing a modified base position, based on the distributions of the kinetics of base incorporation (IPD ratios) from all reads covering this position and from *in silico* kinetic reference values (details are available at

http://www.pacb.com/pdf/TN_Detecting_DNA_Base_Modifications.pdf, (Feng et al. 2013)).

Methylated sequence motifs were identified as previously described (Furuta et al. 2014).

3.3.4 Local deviations in positional distributions of methylation motifs

Occurrences of methylation target sequences in genome sequences were determined using the fetchGWI tool (Iseli et al. 2007). The start positions and orientations of 1997 annotated ORFs (Parkhill et al. 2000) were used as 'reference feature' to sum up the occurrence counts for each methylation target motif ('target feature') using the ChIP-Cor tool (http://ccg.vital-it.ch/chipseq/chip_cor.php). Thereby motif counts were aggregated within 50bp windows positioned relative to the start (position zero) of each ORF. Statistical significance for the observed depletions/enrichments in the plotted counts was derived from a comparison to 1000 sets of simulated reference features with 2000 random genomic loci each. P-values represent the fraction of random reference feature sets exhibiting aggregate motif counts across their corresponding 50bp windows more extreme than the count observed across the 50bp windows of the ORF set.

3.3.5 Identification of DNA methyltransferase genes

Protein sequences of methyltransferases as obtained from REBASE (rebase.neb.com) were used to identify genes with >80% identity via BLAST searches. Potential methyltransferase ORFs were attributed the REBASE annotation, as available for the reference strain Z2491. For each of our isolate strains each methyltransferase ORF was verified for indels and SNPs (see SNP calling below) altering the frame or introducing premature stop mutations and thereby deactivating the enzyme.

3.3.6 SNP calling

Single nucleotide polymorphisms (SNP) detection was performed as described in (Lamelas et al. 2014), (sequence data available at

http://www.sanger.ac.uk/resources/downloads/bacteria/neisseria.html#t_2).

In brief, sequence variations relative to the *N. meningitidis* serogroup A, ST4 strain Z2491 (Parkhill

et al. 2000) were determined, excluding SNPs in phage sequences, recombinant fragments (Croucher et al. 2011), and repetitive regions (>50bp) of the reference genome, as identified using repeat-match (Holt et al. 2008), (Kurtz et al. 2004).

3.3.7 Co-occurrence of SNPs at methylation motifs

Based on coordinates in BED format of SNPs and of individual bases within target motifs (or non-target control motifs), we determined the number of overlapping positions using the intersect and count commands of BEDTools (Quinlan & Hall 2010). For plotting, the overlap counts between mutated bases and methylation sites were normalized by the number of genome wide motif occurrences and multiplied by a scaling factor x1000. The specificity of the overlaps to methylated positions was ascertained by the comparison to unmethylated positions within methylation target sites, as well as within 2 similar control sequences not known as DNA-methylation targets. To test the statistical significance of the observed increased overlaps, we assumed a random distribution of SNPs over the genome. The null hypothesis of independence between mutations and methylations was tested using the Chi-square approximation to the hyper-geometric distribution

3.4 Results

3.4.1 SMRT sequencing determines divergent DNA modification profiles

We assayed the DNA methylation profiles of 2 *N. meningitidis* strains (Z2491 and NM1264) using SMRT sequencing at a coverage for each strand approximating 50x on Tet1 converted genomic DNA samples.

The kinetics of polymerase extension steps were compared with previously recorded control values for highly similar, unmodified reference sequences (Schadt et al. 2012). We observed diverse kinetic variation signals, some of which could be attributed to known modification events such as DNA methylation. DNA methylation on each genomic position was represented by a probabilistic modification score ("modQV") comprising base incorporation rates differing from that of the unmodified reference sequences. A genomic position is covered by several sequenced DNA fragments, and the modification scores include the consistency by which a specific modification was observed (supp. datasets S4, S5). SMRT sequencing assessed both DNA strands independently, accordingly we determined for strain Z2491 comparable average modification scores of 78.97 over 5237 sites with a modification score > 50 on the forward strand versus an average of 80.27 over 5246 sites on the reverse strand. In a plot of modification scores against sequencing coverage (Figure 3-1), both strains displayed a signal for modified cytosines (green dots). Spurious signals on non-cytosine bases in strain Z2491 are due to secondary peaks from nearby modified cytosines (see Figure 3-2B). Modification scores on adenosine bases (red dots) were clearly dominant in strain NM1264. If comparing to SMRT sequencing of unmodified aliquots of identical DNA samples (Figure S3-1), we find a satisfactory specificity of the Tet1-conversion for 5mC, with a minor reduction of the modification scores for 6mA.

In order to identify DNA recognition sequences of prokaryotic methyltransferases, we applied the SMRT[®] Analysis software suite from Pacific Biosciences to interpret the kinetic variation data on a genome-wide scale. We identified sequence motifs associated with a consistent kinetic variation pattern. Table 3-1 summarizes sequence motifs with a stringent modification score threshold >50.

To relate the discovered sequence motifs with information from REBASE (Roberts et al. 2010) and the ORF status of the corresponding gene in the genome sequences, we assessed the presence of functional ORFs of DNA methyltransferase genes in the assembled genome sequences. We compiled a set of 13 DNA methyltransferase genes (RM genes) occurring in our genomes (Z2491 and NM1264), based on sequence similarity with established DNA methyltransferase genes in all bacterial species in REBASE.

This comparison allowed attributing the identified motifs to established DNA methylation target

motifs (Table 3-1). Two DNA methylation motifs were identified to be common in both *N. meningitidis* strains. A closely similar sequence motif predicted in both strains perfectly fit the C^{5m}CGG target motif of the methyltransferase gene M.NmeAI active in both strains. Multiple partially overlapping motifs could be attributed to either the T^{5m}CTGG target motif of M.NmeAORF1035P or to the related CC[AT]GG target motif of the methyltransferase gene M.NmeAORF1500P. Given the considerable similarity of these two target sequences including ambiguous positions, we cannot completely exclude technical artifacts in the motif discovery defining the target sequence motifs and improvements of the sequence specificity description in future REBASE releases.

Two adenosine methylation motifs were detected exclusively in strain NM1264, consistent with the global DNA modification scores in Figure 3-1. The motif ATGC^{6m}AT matches the (predicted) target sequence for M.Nme2594ORF759P in REBASE. As a novel finding the motif AC^{6m}ACC can be attributed to modA12 (M.NmeAORF1589P), which is the only remaining DNA methyltransferase with functional ORF solely in strain NM1264 (Table 3-1). Notably this target specificity differs from the 5'-AGAAA-3' recognition site of a related modA13 allele in *N. gonorrhoeae* (Srikhanta et al. 2009). Our SMRT sequencing results resolved furthermore the position of the modified base within target sequences with a yet undetermined position as reported by REBASE, exemplified by ATGC^{6m}AT for M.Nme2594ORF759P (Table 3-1). Given the still limited positional resolution of 5mC even after Tet1 conversion (see also Figure 3-2B), the position calls were considered particularly reliable for 6mA modifications.

The SMRT sequencing results moreover revealed a modification of the sequence motif GGNN^{5m}CC, which strain-specific detection associated with an ORF for the gene M.NmeAORF1453P complete solely in the strain Z2491. The existence of a methylation-sensitive restriction enzyme NlaIV targeting an identical sequence motif (GGNNCC) allowed validating the differential methylation as detected by SMRT sequencing. Accordingly NlaIV fragmented the genome of strain NM1264, whereas the Z2491 genome methylated at GGNNCC sites resisted NlaIV digestion (Figure 3-2A).

The results of these restriction digests indicated a complete protection and therefore a genome-wide methylation of 'GGNNCC' sequences in the strain Z2491. However only 48% of the 1817 instances of 'GGNNCC' sequences were called as modified in SMRT sequencing, despite the genome-wide methylation (Figure 3-2B). This limited sensitivity was presumably due to a very stringent threshold >50 for the SMRT modification score, to an incomplete enzymatic Tet1 conversion, and/or to limited positional precision of the kinetic signature of 5caC (Tet1-modified 5mC). In clear contrast, the fractions of modified bases were below 1% for the NlaIV restriction sensitive strain NM1264.

Most of the discovered sequence motifs were palindromic, and accordingly a modification signal was also detected on the 'mirror' base on the opposite strand. The motif AC^{6m}ACC is exemplifying the strand-specificity and sensitivity of the SMRT sequencing on adenosine methylation, for this non-palindromic motif consequently no signal was observed on the opposite strand (Figure 3-2C). Given the limited sensitivity and positional precision for 5mC modifications, instead of using the actual SMRT modification scores, in subsequent analysis we considered all sequences matching the methylation target motifs identified by SMRT. In conclusion, SMRT sequencing of 2 closely related *N. meningitidis* strains of serogroup A revealed highly divergent DNA methylation profiles associated with the functional status of DNA methyltransferase genes. In addition our approach enabled the confirmation and identification of novel target motifs for predicted DNA methyltransferase genes.

3.4.2 Methylation target motifs with biased distributions in regulatory genomic regions

Functional consequences of DNA methylation are incompletely characterized. Moreover the genomic locations of DNA parts with regulatory functions are not precisely established in *N. meningitidis*. We therefore focused on sequences immediately upstream from genes, which were suggested to harbor a considerable proportion of loci under purifying selection based on the analysis of phylogenetically conserved sequences in prokaryotes (Molina & van Nimwegen 2007). We applied a cumulative analysis of the occurrence of methylation motifs relative to a set of 1997 start positions of annotated ORFs. The aggregation over a large set of loci renders this ChIP-cor analysis (see methods for details) very sensitive for recurring local deviations in linear distributions. At distances up to 1kb to ORF start positions, methylation motifs were detected at frequencies in general closely approaching the average genome wide frequencies (Figure 3-3). Only the motif occurrences immediately upstream from ORFs displayed a significant deviation (p value < 0.05), if compared to motif counts in equally sized sets of random loci. The observed deviations displayed a larger magnitude than the average GC content, which is only slightly decreased at the ORFs (Figure 3-3). To further control for base composition effects, we assessed the positional distributions of a set of non-methylated sequence motifs without overlaps with target motifs described in this study, with similar base composition as the two non-palindromic target motifs, and not specifying exclusively G and C bases. Unlike methylation target motifs, these control sequence motifs displayed no significant deviation, if compared to motif counts at random loci as described above.

We have extracted 120 ORFs displaying at least one AC^{6m}ACC motif within the interval from -75bp to their start position, but the current annotations of the large majority of those genes (hypothetical

protein, unknown function) did not allow to identify particular functional groups sharing methylation target sequences in their regulatory sequences. An analogous analysis for each of the 5-methylcytosine motifs neither led to the identification of over-represented gene categories, functions or localization. Nevertheless the observed clear biases in the positional distribution of methylated target sites strongly suggests a functional role of DNA methylation likely related to the regulation of genes.

3.4.3 Variable set of active DNA methyltransferase genes in serogroup A *N. meningitidis* isolates

In order to establish the potential of DNA methylation in the genomes of a collection of *N. meningitidis* strains, we extended the assessment of the presence of functional ORFs of DNA methyltransferase genes to assembled genome sequences of 101 strains of *N. meningitidis* previously collected over a time period of ~10 years during meningococcal meningitis epidemics in Sub-Saharan Africa, clustering into two sequence types (ST2859 and ST7) (Lamelas et al. 2014). We included two reference strains of serogroup A, namely WUE2594 (Schoen et al. 2011) and Z2491 (Parkhill et al. 2000).

Our analysis of the matrix of predicted DNA methylation activities revealed the genomic diversity within the 101 serogroup A strains assessed here. While the majority of DNA methyltransferase genes display constant presence/absence (ORF ON/OFF) patterns (Figure 3-4), selected genes featured a larger diversity than to be expected from the global genome sequence similarity. Contributing to the ON/OFF diversity, we detected point mutations leading to premature stop codons (M.NmeAORF1453P in all strains except Z2491), or deletion of complete genes (M.Nme2594ORF759P =NMAA_0759) likely related to genome rearrangement events and horizontal gene transfers. The largest part of divergence between the strains is however due to phase-variability in two type III methyltransferase genes modB2 (M.NmeAORF1467P) and modA12 (M.NmeAORF1589P).

We used for SMRT sequencing an aliquot of the genomic DNA preparation of strain NM1264 previously subjected to the Illumina sequencing method. Thereby we detected only 198 sequence variants (supp. dataset S3) if mapping circular consensus reads from SMRT sequencing at an average coverage of approximately 100x (twice 50x from each strand) to contigs assembled from Illumina reads (~300x coverage (Lamelas et al. 2014), supp. dataset S6). Hence the augmented number of indels in individual sub-reads of the SMRT method are effectively averaged out if DNA fragments are read multiple times and unified into circular consensus sequences.

As standard genome assembly and read mapping algorithms consistently failed especially at

longer microsatellite repeat regions (Treangen & Salzberg 2012), we determined the repeat unit numbers directly from Illumina reads covering the corresponding locus (Figure 3-5). The determined repeat numbers enabled to call the ORF status at the ModA12 locus (ON: 18 strains; OFF: 59) and at the ModB2 (ON: 4; OFF: 62). The read length of 75bp represented a limit to determine the number of microsatellite repeat units ('AGCC' for modA12 and 'TTGGG' for modB2) flanked by at least 5bp of non-repeat sequence. We could therefore not determine the ORF status at modA12 for 22 strains or at modB2 loci for 33 strains, respectively. These genomes contain in all likelihood repeats of a lengths exceeding the read length, for instance more than 15 x (AGCC) repeat units at the modA12 locus (Figure 3-4). Strikingly a few genomic DNA preparations yielded in a limited number of sequence reads containing repeat units divergent from the majority of reads covering the corresponding locus. Assuming no cross-contaminations from other samples, these reads might be products of intra-clonal variability, consistent with increased mutations rates at phase-variable loci (Gemayel et al. 2010). In conclusion, our careful analysis of the ORF status of a panel of DNA methyltransferase genes revealed phase-variability as a major contributor to variability in the DNA methylomes of isolates assessed here.

3.4.4 Mutations overrepresented at DNA methylation target motifs

We set out to investigate correlations of DNA modifications as determined in this study to the mutations as observed in the genomes of our serogroup A strain collection (Lamelas et al. 2014). The single nucleotide polymorphisms were determined based on the genome sequence of strain Z2491 as reference and presumably reflect the *in vivo* mutation and selection processes within the bacterial population associated with the meningitis epidemics.

From the total number of 6031 SNPs filtered for repeats and for recombinant fragments in the genomes of these strains and from the 20537 methylated nucleotides based on the consistent DNA methylation target sequences (AC^{6m}ACC, C^{5m}CGG, Y^{5m}CTGG, GGN^{5m}C) we would expect from a random distribution a total of 6031 SNPs / 1.6Mb * 20537bp = ~77 SNPs occurring per chance on a methylation target site in a 1.6Mb Mb repeat-excluded genome length. We actually observed a total number of 201 SNPs overlapping a methylation site, representing a 2.6 fold over-representation. This global approach indicated that methylated nucleotides indeed have an increased likelihood of mutation in settings with *in vivo* mutation and selection processes. The corresponding 201 methylation sites detected in the Z2491 genome did lose their function as target sites by the occurrence of the SNPs in the sequences of our serogroup A strain collection.

To highlight the specificity of this effect to the methylated base position, we assessed the average number of SNPs at each motif position, normalized by the number of genome-wide occurrences of

the motif. Given that our SNP calling could not determine the strand affected by a mutation, we considered both complementary bases. Figure 3-6 represents five of the methylation target motifs detected in this study. We compared the SNP counts (C/G→N or A/T → N) at each position of methylated motifs, or of scrambled non-methylated sequence motifs. The cytosine positions (T^{5m}CTGG, C^{5m}CWGG) consistently methylated in both strains as well as the methylated adenosines in the phase-variable AC^{6m}ACC target motif displayed a ~2-3 fold significantly higher co-occurrence rate of SNPs, if compared to corresponding positions within scrambled motifs with unmethylated bases (p-value < 10⁻⁵). Non-methylated nucleotides in neighboring positions within the same motif, or within motifs not identified as methylation targets featured SNP occurrence rates close to the expected overlap if assuming randomly distributed SNPs. SNP classes (synonymous, non-synonymous, intergenic) might reveal divergent selective pressures, we did however not observe significant differences for SNPs overlapping methylated bases (Figure S3-2). The target motif ATGC^{6m}AT detected in this study displayed a tendency to increased co-occurrence rates with SNPs at methylated positions, however the motif was excluded due to a low number of only 128 occurrences in the Z2491 genome. For palindromic sequence motifs only the occurrence on the forward strand was considered. Consistent with a full methylation on both strands the palindromic motif C^{5m}CGG showed a mirroring peak at the guanosine in the third motif position, which correspond to the methyl-cytosine on the reverse strand. The methylated positions in the palindromic motif GGNN^{5m}CC displayed a barely increased overlap with SNPs. The corresponding methyltransferase (M.NmeAORF1453P) is only active in strain Z2491 (Figure 3-4, Table 3-1). From the uniform inactivation of the methyltransferase in all our isolates by an identical premature stop mutation we can assume an early time point of this mutation event in the evolutionary history separating our genomes from a common ancestor genome. Therefore the limited overlap of SNPs is consistent with a loss of methylation at GGNNCC, further supporting mutation rates depending on the duration of DNA methylation during evolution of the genomes.

3.5 Discussion

We applied SMRT sequencing to genomes of the facultative human pathogen *Neisseria meningitidis*. The thereby determined DNA modification profiles of closely related isolates revealed similarities and differences in DNA methylation motifs, which could be associated with the presence of intact ORFs of a set of methyltransferase genes. Part of the differential DNA methylation could be attributed to the phase-variable state of corresponding DNA methyltransferase genes. We furthermore assessed the positional distribution of the detected methylation target motifs within the genome assemblies. Clear occurrence biases of methylation

motifs within presumable regulatory sequences are suggesting a role of DNA methylation in gene regulation, possibly related to proposed differences in antimicrobial susceptibility (Jen et al. 2014). Moreover in our cross-sectional analysis of *Neisseria* isolates, we detected a prominent co-localization of SNPs with methylated bases, demonstrating an association of DNA methylation with mutagenesis and the evolution of genomes. This may have general implications for other prokaryotic populations.

Detection limits for 5mC modifications

Our genomic DNA samples were subjected to an enzymatic conversion of 5mC to 5caC via the Tet1 enzyme. We detected kinetic signals for DNA methylation at only a fraction of the 5mC target sites in the genome. In contrast we observed a genome-wide complete protection of GGNNC^{5m}C target sequences in a methylation-sensitive restriction digest. As a likely explanation the conversion of 5mC by Tet1 may have been incomplete. In addition, we detected a modification signal only on 22% of the CCGG sequences, which may hint for motif discovery artifacts due to partial overlaps with GGNNC^{5m}C target sequences. We therefore excluded the 358 instances of DDGGNNCCGG or CCGGNNCCHH sequences (D= not-C; H= not-G) in the Z2491 genome and still detected a SMRT modification signal at 19% of C^{5m}CGG sequences.

Another potential reason for partial modification signals at a discovered sequence motif is the arbitrarily chosen threshold for the modification score. A threshold of 50 applied in this study represents a compromise to minimize presumable false positive calls while still enabling to discover target motifs including also cytosine methylation with reduced kinetic signatures. Our data analysis procedures required some consistency in DNA modifications within the pool of individual cells subjected to SMRT sequencing. Consequently the absence of SMRT modification calls cannot be interpreted as complete absence of any DNA modification at this locus. Conceptually intra-sample variations in DNA modifications could occur at different loci in the genome, for instance if some of the potential target sequences are 'masked' for the DNA modification enzymes by other DNA binding complexes. Such divergent modification patterns at specific loci reminding of eukaryotic cellular differentiation mechanisms have been previously observed (Kozdon et al. 2013), however only for modifications on adenosines bases, presumably due to the experimental limitations mentioned above.

At present potential biological variability can thus not be distinguished from the technical variability, specifically for 5mC. Further development of analysis methods might enable the detection of additional divergence in individual cells or at specific loci.

Comparable sequencing accuracy of SMRT sequencing

The design of our SMRT sequencing assays was aiming at a reliable detection of DNA modifications with high sequencing coverage via relatively short sequence inserts (average ~260bp), therefore current hybrid assembly algorithms did not lead to an improvement of the genome assembly of strain NM1264. Circular consensus sequences covering the phase-variable loci confirmed the divergent genotypes for the two SMRT sequenced strains, as originally derived from Illumina sequences, or from the reference genome assembly, respectively. Overall the SMRT sequencing in our approaches displayed a sequencing accuracy very comparable to that of other standard sequencing methods.

Sequence variability in clonal populations

In addition to the differences between strains we also observe divergent repeat unit counts in genomic sequences extracted from clonal populations. Extending the considerations above, we do not observe such variability for all samples and for all loci, suggesting site-specific mutagenesis mechanisms. Albeit the sequencing coverages applied in these studies did yield in only a few reads with divergent repeat unit counts, we might nevertheless be detecting the products of phase-variable mutations occurring during expansion of a clonal cell population. Verification of this hypothesis might require extreme precautions in preparations of genomic DNA samples and high-coverage sequencing in distant facilities to exclude the possibility of cross-contaminations of samples.

Functional consequences of highly variable DNA modifications

Previous studies described the direct consequences of variable numbers of microsatellite repeat units within the ORF of two methyltransferase genes (modB2 and modA12) on the ORF status and consequently the DNA methylation profiles (Srikhanta et al. 2009). The rate at which these phase-variability mutations occur and underlying mechanisms are ill characterized (Gemayel et al. 2010). Reports describing mutations rates at other phase-variable loci in *N. meningitidis* described drastic differences between serotypes, possibly linked to mismatch repair systems (Richardson & Stojiljkovic 2001).

We do reveal in our study a non-random positioning of the methylation target sites which might suggest an involvement in gene regulation. Altered RNA abundance levels of a set of genes are

however difficult to interpret as direct consequence of divergent global DNA methylation activities, as typically there is a very limited correlation of deregulated transcripts and DNA methylation target sequences observed (Bendall et al. 2013). Mutations in target sequences associated with divergent expression levels of specific genes could reveal more mechanistic insights. Our observation of increased numbers of SNPs precisely at methylation sites may however indicate that the regulatory mechanisms feature large degrees of plasticity and redundancy.

An additional consequence of the presence of restriction sites was reported on the DNA uptake sequence-dependent transformation (Ambur et al. 2012). Specific DNA methylation profiles in *N. meningitidis* strains might thus define 'compatibility groups' for horizontal DNA transfers within the microbial community in the human nasopharynx (Claus et al. 2000) (Bart et al. 2001). We could not identify significant biases in the presence or absence of DNA methylation target sites within putative recombinant fragments. These fragments putatively originating from *N. lactamica* or *N. gonorrhoeae* constituted about 20% of the genome length, and are containing a matching proportion of methylated motifs.

In the present study we detected a significant enrichment of SNPs between genomes of *N. meningitidis* serogroup A strains at positions of methylated bases within DNA methylation target motifs. The causes of this correlation and the consequences on genome evolution are at present not clear. Strikingly, of the 4096 instances of the non-palindromic AC^{6m}ACC target motifs within the genome of the Z2491 strain, only 33% occur on the leading strand, which might relate to the observation of a clear bias (60.2%) of ORFs on the leading strand in one replicore (Parkhill et al. 2000). Essential genes are favored to occur on the leading strand, hypothetically due to lower mutation rates resulting from reduced replication–transcription conflicts (Paul et al. 2013). Accordingly the observed biased distributions of specific DNA methylation target motifs might be either the consequence of increased mutations at these sites or represent selection pressure to exclude or maintain DNA methylations sites at intergenic regulatory regions. Strikingly we could not discern major biases, SNPs overlapping methylated nucleotides showed a similar distribution between intergenic and coding regions. Similarly no bias was observed between synonymous and non-synonymous SNPs within the coding region deviating from the overall SNPs segregation (Figure S3-2). Therefore we conclude that selective pressures are similar on mutations associated with DNA methylation.

Recent comparative genome analysis has considerably expanded our knowledge of prokaryotic defense systems (Makarova et al. 2013). Specifically the presence of apparent conflicts between restriction systems (Ishikawa et al. 2010), or orphan methyltransferases lacking cognate restriction enzymes (Marinus & Casadesus 2009) hint for more complex biological roles of prokaryotic DNA methylation. The precise effect of DNA methylation on mutation rates in prokaryotes is presently

unclear due to multiple levels of mutational, mismatch repair, and selection mechanisms (Casadesús & Low 2006). The damage of an uracil base resulting from deamination of an unmodified cytosines can typically be corrected (Walsh & Xu 2006). Original studies describe an increased rate of spontaneous deamination of 5-methylcytosine compared to cytosine residues (Ehrlich et al. 1986). Deamination of 5-methylcytosine results in a 'genuine' thymine base. In the context of double stranded DNA molecules, mismatch repair mechanisms have therefore limited means to detect and repair unequivocally the newly mutated nucleotide in a G/T mismatched pair. Repair systems counteracting the mutagenic effects of hydrolytic deamination of m5C (Vsr endonucleases) have been described in *Neisseria gonorrhoeae* (Kwiatek et al. 2010), yet we have no evidence for activities of orthologous genes (V.NmeIP) in *Neisseria meningitidis*. Methylated bases have been reported to be mutational hotspots for instance in mutation-accumulation studies in *E. coli* in laboratory settings (Lee et al. 2012). Our present study addresses for the first time the association of experimentally confirmed DNA methylation and the genome evolution in an *in vivo* setting. Here a number of additional processes are involved in the selection of favorable configurations of genome structures at different scales (Rocha 2008). Our results suggest that DNA methylation and evolutionary processes are two processes intimately correlated. Despite the highly variable activities of DNA methyltransferase genes in evolutionary timescales, genomic and epigenomic factors contribute in a complex interplay to the evolution of the optimally adapted prokaryotic populations.

Conclusions

SMRT sequencing determines DNA methylation profiles of prokaryotes at a genome-wide level.

This study contributes to the recognition of a previously underestimated potential for variability in DNA methylation. The discovery of biased presence of methylation target motifs in genomic sequences may indicate a role in gene regulation. The increased occurrences of mutations precisely at methylation target positions suggests additional yet unidentified functional consequences of DNA methylation and Restriction-Modification systems in the evolution of prokaryotic genomes.

Acknowledgments

This work was supported by the Forschungsfonds of the University of Basel (grant “DZX2056” to CS). The authors wish to acknowledge Till Voss (Swiss TPH) for advice on molecular biology

approaches, Julia Hauser (Swiss TPH) for help with the bacterial cultures, and Christian Schindler (Swiss TPH) for statistical advice.

3.6 Tables

Table 3-1: summary of detected modifications at DNA target motifs

SMRT identified motifs		Closest motif in REBASE	GeneID / REBASE entry	ORF status	
Z2491	NM1264			Z2491	NM1264
C ^{5m} CGG	C ^{5m} CGG	C ^{5m} CGG	M.NmeAI	ON	ON
T ^{5m} CTGG	T ^{5m} CTGG	T ^{5m} CTGG	M.NmeAORF1035P	ON	ON
C ^{5m} CWGG	C ^{5m} CWGG	CCWGG	M.NmeAORF1500P	ON	ON
GGNN ^{5m} C C	—	GGNNCC	M.NmeAORF1453P	ON	early stop
—	AC ^{6m} ACC	—	<i>ModA12</i> M.NmeAORF1589P	early stop*	ON *
—	—	—	<i>ModB2</i> M.NmeAORF1467P	early stop*	early stop*
—	ATGC ^{6m} AT	ATGCAT	M.Nme2594ORF759P	not present	ON
—	—	CATG	M.NmeAORF59P	early stop	early stop
—	—	GGTGA	M.NmeAORF191P	early stop	early stop
—	—	GCCG ^{6m} AG	NMAIII RM	early stop	early stop
—	—	[C/T/G]A	M.NmeAIV	ON	ON
—	—	—	M.NmeAORF1038P	early stop **	early stop**
—	—	—	M.NmeAORF1385P	early stop	early stop

DNA methylation target motifs as inferred from sites featuring a SMRT modification score >50. *N. meningitidis* strains Z2491 and NM1264 feature shared and strain-specific motifs, mostly with a correspondence in REBASE. Suffix 'P' in the GeneID specifies a methyl-transferase gene predicted by REBASE based on sequence homology. * Premature stop due to phase-variable mutation ** premature stop codon in specificity subunit of type I RM system comprising M.NmeAORF1038P.

3.7 Figures

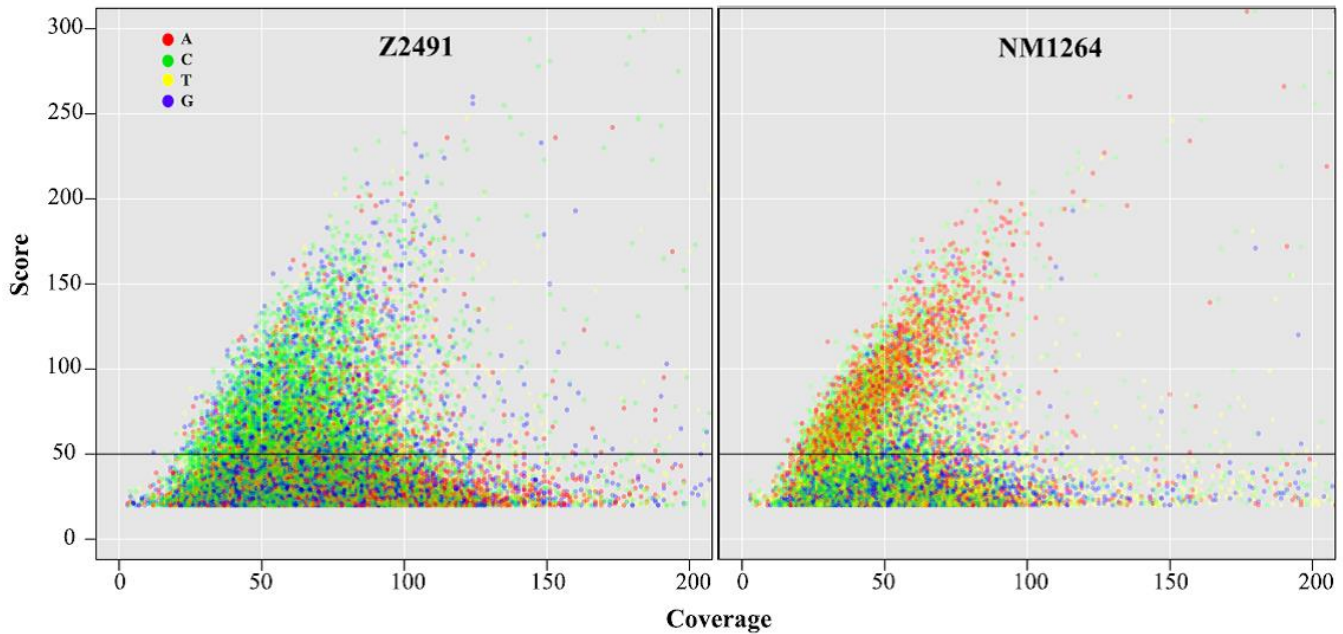


Figure 3-1: Two *N. meningitidis* serogroup A strains Z2491 and NM1264 display divergent DNA modifications.

DNA modification scores are plotted against the coverage in SMRT sequencing of Tet1 converted samples. Each dot represents a position on either strand with a modification score larger than 20, the color specifying the nucleotide base, on which the modification was detected. Modified adenosines (red dots) are predominantly detected in strain NM1264. The horizontal line indicates the threshold score 50 applied for subsequent motif finding.

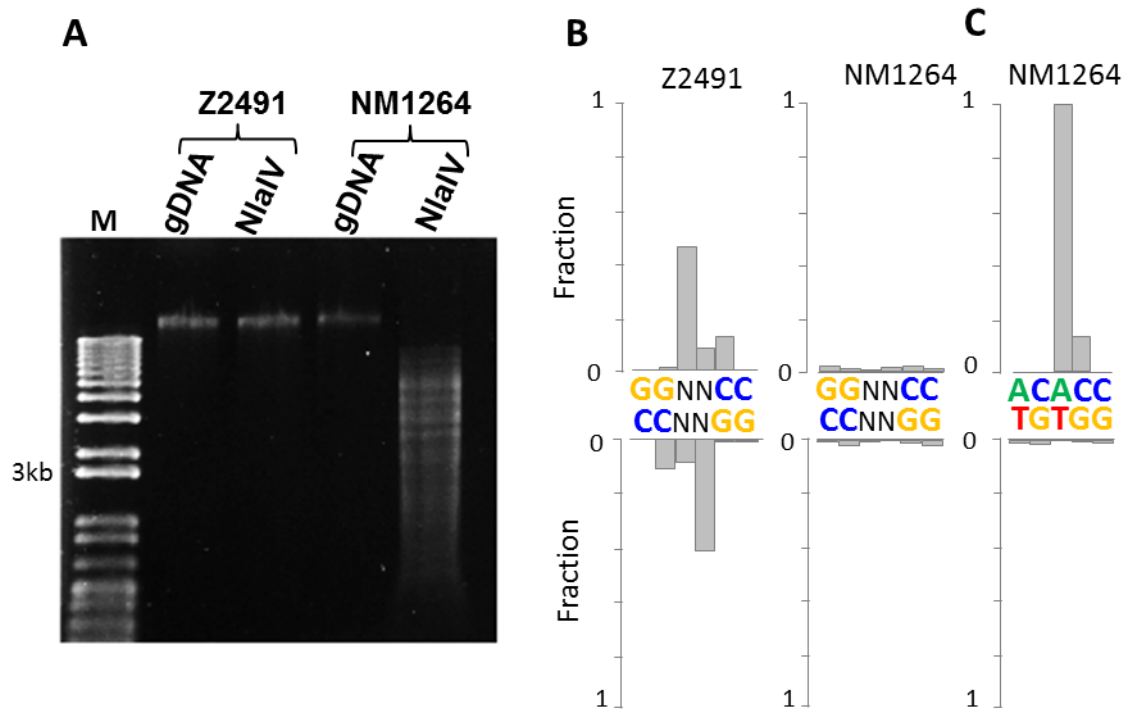


Figure 3-2: Methylation-sensitive restriction assays to validate DNA methylation derived from SMRT sequencing

(A) DNA samples separated via gel electrophoresis, lane labels indicating (M) size marker, (gDNA) whole genomic DNA preparations, and (NlaIV) restriction digest products targeting sites 'GGNNCC' in a methylation-sensitive manner. Samples from 2 different strains (Z2491 and NM1264, genome size: 2.18 Mb) display differential resistance consistent with DNA methylation profiles. (B) single position resolution of modification signal averaged over ~1800 'GGNNCC5mC' sites in the respective genome assemblies. The fractions of sites exhibiting a modification score above 50 are displayed for each position and strand. (C) For comparison, adenosine methylation featuring enhanced sensitivity and positional resolution averaged over ~4000 'AC6mACC' sites.

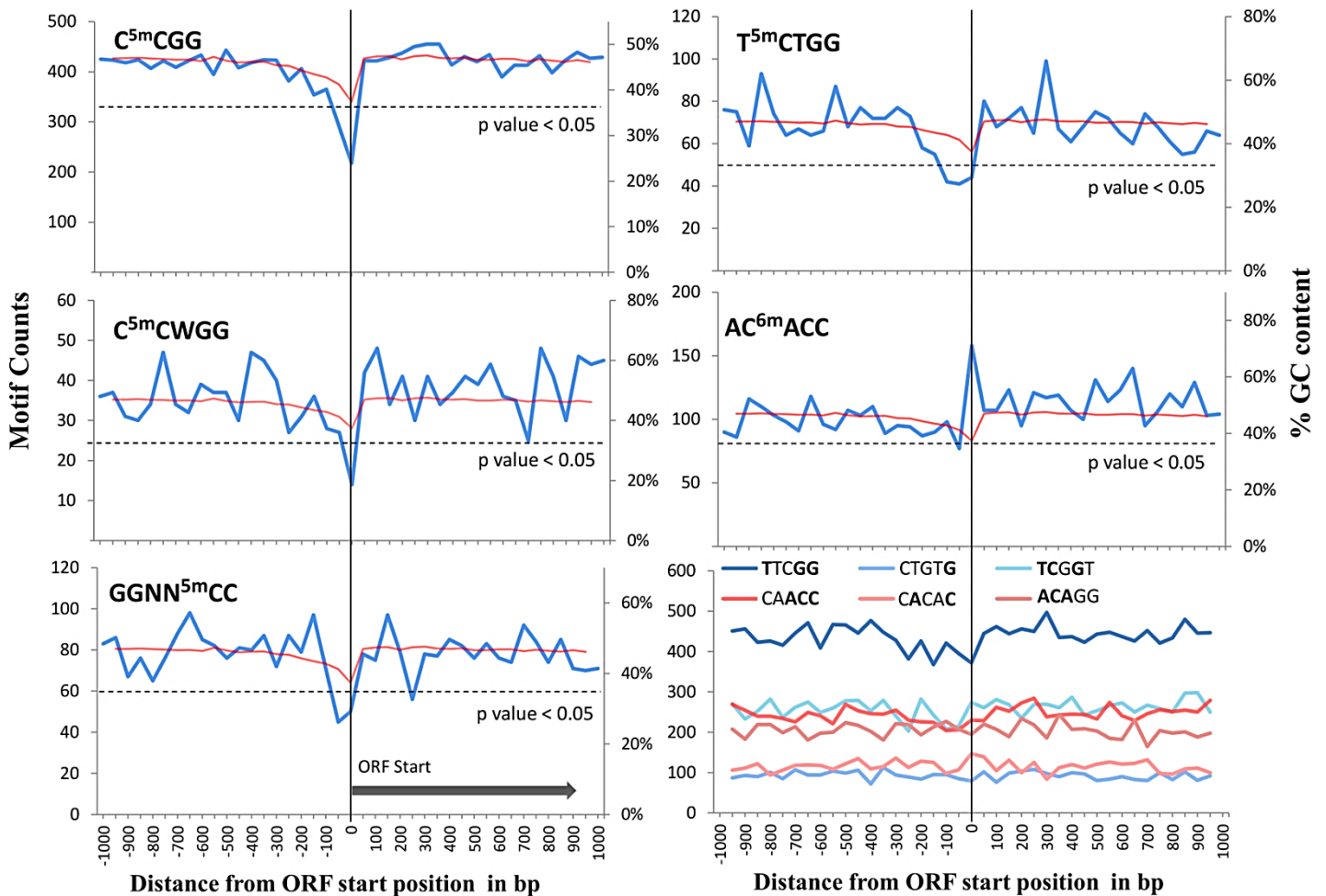


Figure 3-3: Depletion of cytosine methylated motifs in the immediate upstream region of ORFs.

Occurrence counts of five (most frequent) methylation target motifs are plotted against their position (in bp) relative to 1997 oriented ORFs (genome annotation *N. meningitidis* strain Z2491). Motif counts are presented as sum over all ORFs within 50 bp windows, centered at position zero. Red lines in each panel compare to the GC content percentage (y-axis label to the right), averaged over all ORF regions. Dashed horizontal lines represent the averaged motif occurrences corresponding to statistically significant (p value 0.05) depletions of the corresponding motif, derived from a comparison to equally sized sets of random loci. The lower right panel represents occurrence counts of a set of six non-methylated control motifs with similar base composition (identical positions in bold), and similar occurrence frequencies as the non-palindromic target

motifs T^mCTGG and AC^mACC. None of the control motifs display significant depletions at ORFs comparable to that of methylated motifs.

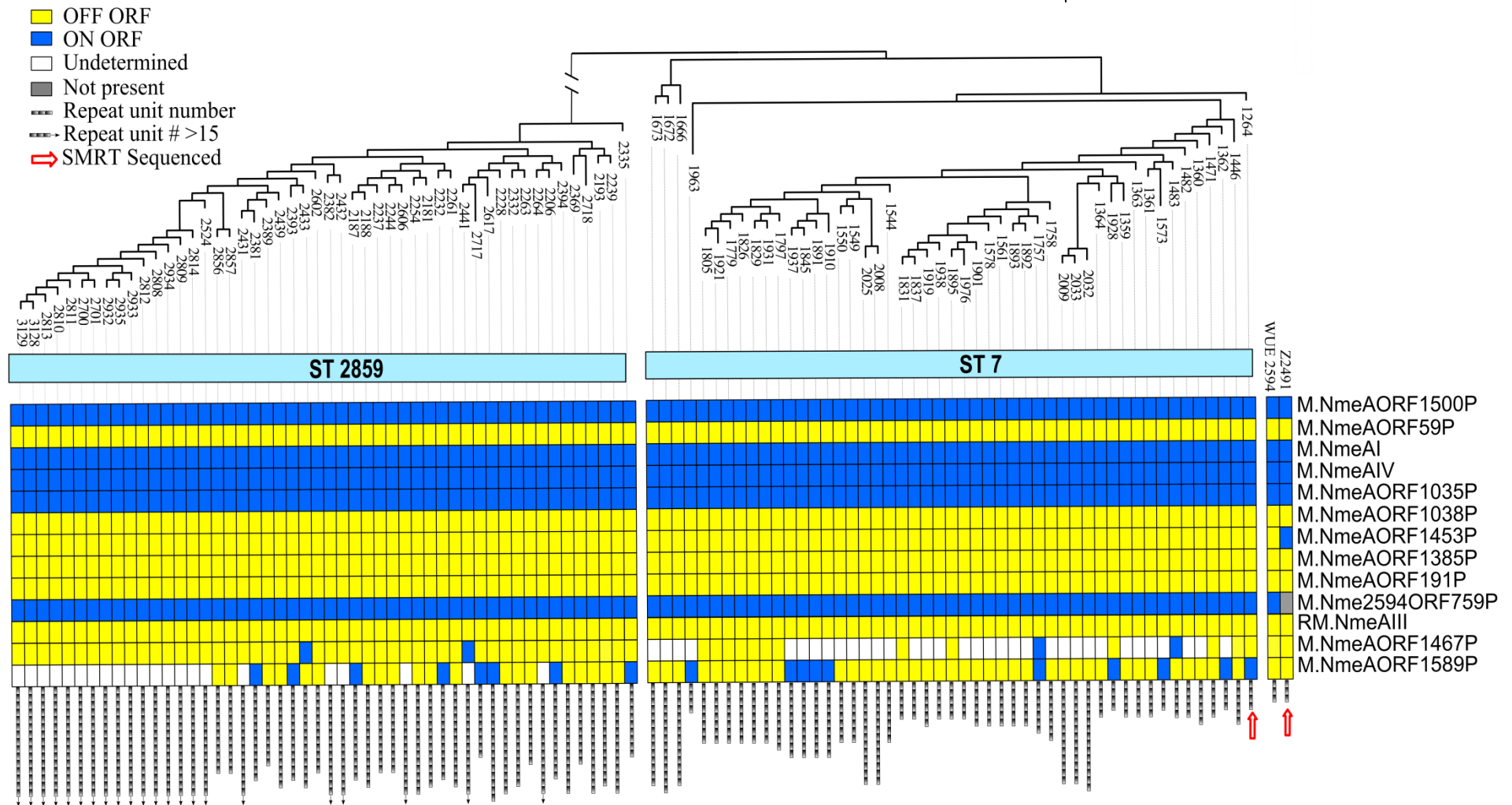


Figure 3-4: Variability at DNA methyltransferase loci. 101 *N. meningitidis* isolates clustered according to SNP distance, yielding in two sequence type (ST) groups. Each column represents an isolate and rows specify the ORF status of 13 DNA methyltransferases (Rebase gene IDs of Z2491 reference strain). Bars in grey at the bottom represent the number of repeat units determining ON/OFF status of the phase-variable modA12 (M.NmeAORF1589P)

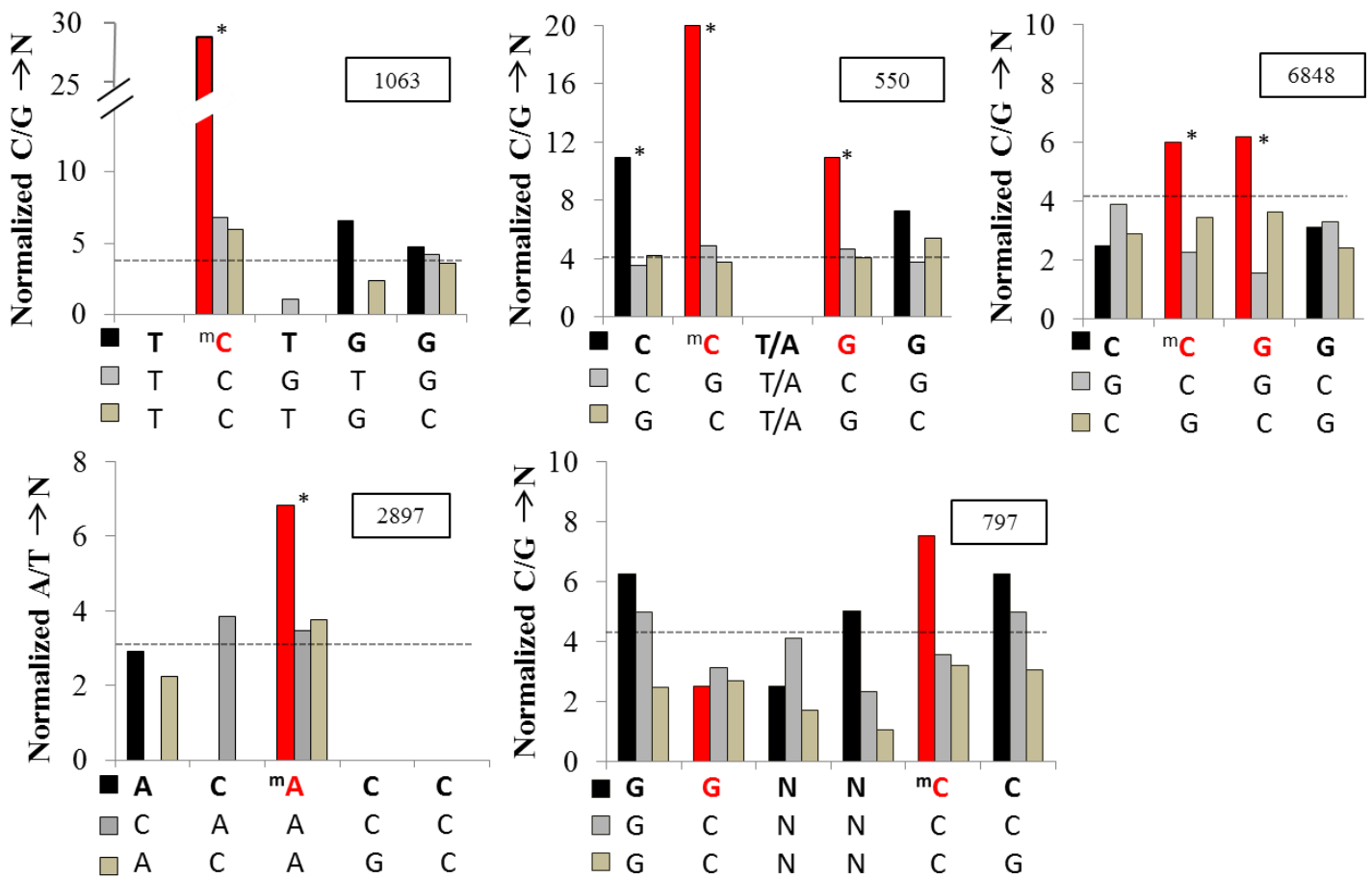


Figure 3-6: Methylated nucleotides display higher mutation rates than non-methylated positions.

Positional co-occurrences of (in total 6031) SNPs at positions within (methylation) target motifs. Methylated positions highlighted in red within five target motifs (**bold**), as detected in the present study, with, each compared to two similar control sequences. Black bars in histogram represent nucleotides in methylated motifs, gray shades represent sequences not known as DNA-methylation targets. For each motif, counts of overlapping SNPs (for 5m-cytosine motifs: C/G in reference →N; or for 6m-adenosine: A/T→N) at each position are normalized by the genome-wide motif occurrences (numbers for methylated motifs in inset box). The dashed lines indicate the corresponding number of SNPs expected from random occurrence (G/C or A/T) across the genome and over-representation was tested with the χ^2 statistics (*p-value < 10⁻⁵).

3.8 Supplementary material

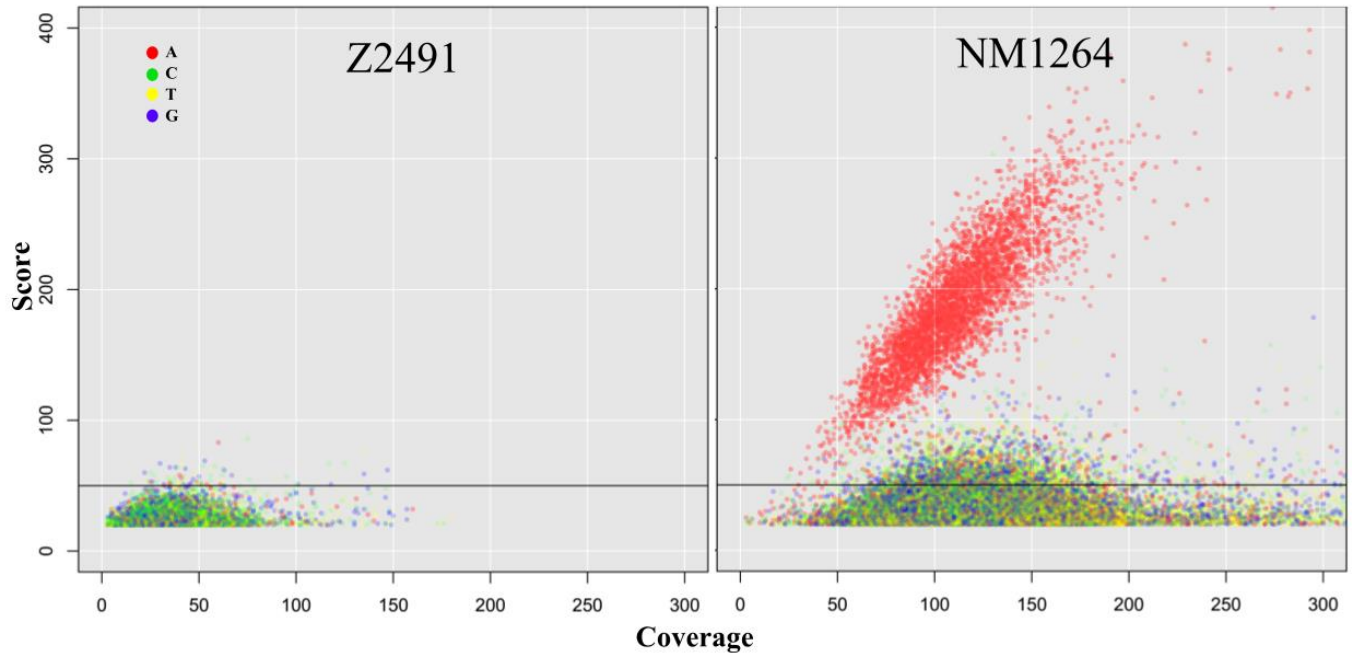


Figure S3-1: SMRT sequencing on samples without Tet1 conversion detects modified adenosines.

DNA modification scores are plotted against the coverage in SMRT sequencing of samples. Each dot represents a position on either strand with a modification score larger than 20, the color specifying the nucleotide base, on which the modification was detected. Modified adenosines (red dots) are predominantly detected in strain NM1264. The horizontal line indicates the threshold score 50 applied for subsequent motif finding. (See figure 3-1 for Tet1 converted samples)

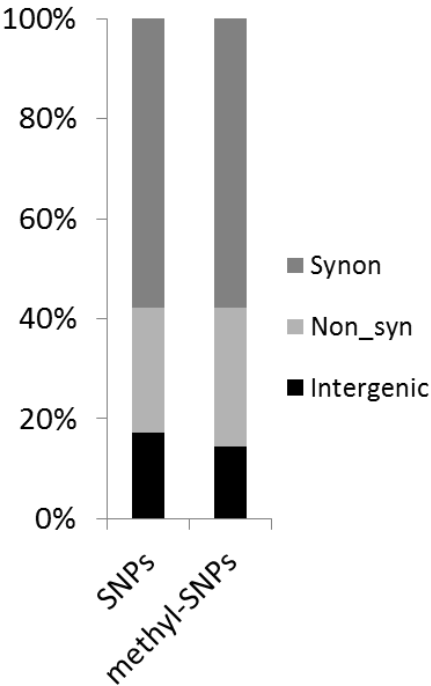


Figure S3-2: Comparable distribution of SNPs overlapping methylated bases.

6031 SNPs as observed within the repeat-filtered genome assemblies of a strain collection of *N. meningitidis* are classified into synonymous and non-synonymous mutations in coding sequences, or attributed to intergenic. SNPs overlapping methylated bases display a very similar distribution, indicating that selective pressures are similar on mutations associated with DNA methylation.

3.9 References

- Ambur OH, Frye SA, Nilsen M, Hovland E, Tønjum T. 2012. Restriction and Sequence Alterations Affect DNA Uptake Sequence-Dependent Transformation in *Neisseria meningitidis*. *PloS One*. 7:e39742. doi: 10.1371/journal.pone.0039742.
- Arber W. 2000. Genetic variation: molecular mechanisms and impact on microbial evolution. *FEMS Microbiol. Rev.* 24:1–7.
- Bart A, Pannekoek Y, Dankert J, van der Ende A. 2001. NmeSI restriction-modification system identified by representational difference analysis of a hypervirulent *Neisseria meningitidis* strain. *Infect. Immun.* 69:1816–1820. doi: 10.1128/IAI.69.3.1816-1820.2001.
- Bendall ML et al. 2013. Exploring the roles of DNA methylation in the metal-reducing bacterium *Shewanella oneidensis* MR-1. *J. Bacteriol.* 195:4966–4974. doi: 10.1128/JB.00935-13.
- Bernstein BE et al. 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* 28:1045–1048. doi: 10.1038/nbt1010-1045.
- Cao B et al. 2014. Genomic mapping of phosphorothioates reveals partial modification of short consensus sequences. *Nat. Commun.* 5:3951. doi: 10.1038/ncomms4951.
- Casadesús J, Low D. 2006. Epigenetic gene regulation in the bacterial world. *Microbiol. Mol. Biol. Rev. MMBR.* 70:830–856. doi: 10.1128/MMBR.00016-06.
- Caugant DA, Maiden MCJ. 2009. Meningococcal carriage and disease--population biology and evolution. *Vaccine.* 27 Suppl 2:B64–70. doi: 10.1016/j.vaccine.2009.04.061.
- Clark TA et al. 2012. Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic Acids Res.* 40:e29. doi: 10.1093/nar/gkr1146.
- Clark TA et al. 2013. Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation. *BMC Biol.* 11:4. doi: 10.1186/1741-7007-11-4.
- Claus H, Friedrich A, Frosch M, Vogel U. 2000. Differential distribution of novel restriction-modification systems in clonal lineages of *Neisseria meningitidis*. *J. Bacteriol.* 182:1296–1303.
- Croucher NJ et al. 2011. Rapid Pneumococcal Evolution in Response to Clinical Interventions. *Science.* 331:430–434. doi: 10.1126/science.1198545.
- Dunning Hotopp JC et al. 2006. Comparative genomics of *Neisseria meningitidis*: core genome, islands of horizontal transfer and pathogen-specific genes. *Microbiol. Read. Engl.* 152:3733–3749. doi: 10.1099/mic.0.29261-0.
- Ehrlich M, Norris KF, Wang RY, Kuo KC, Gehrke CW. 1986. DNA cytosine methylation and heat-induced deamination. *Biosci. Rep.* 6:387–393.
- Eid J et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science.* 323:133–138. doi: 10.1126/science.1162986.
- Fang G et al. 2012. Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. *Nat. Biotechnol.* 30:1232–9. doi: 10.1038/nbt.2432.
- Feng Z et al. 2013. Detecting DNA modifications from SMRT sequencing data by modeling sequence context dependence of polymerase kinetic. *PLoS Comput. Biol.* 9:e1002935. doi: 10.1371/journal.pcbi.1002935.
- Flusberg BA et al. 2010. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods.* 7:461–465. doi: 10.1038/nmeth.1459.
- Furuta Y et al. 2014. Methylome diversification through changes in DNA methyltransferase sequence specificity. *PLoS Genet.* 10:e1004272. doi: 10.1371/journal.pgen.1004272.
- Gemayel R, Vences MD, Legendre M, Verstrepen KJ. 2010. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.* 44:445–477. doi: 10.1146/annurev-genet-072610-155046.
- Holt KE et al. 2008. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. *Nat. Genet.* 40:987–993. doi: 10.1038/ng.195.

- Iseli C, Ambrosini G, Bucher P, Jongeneel CV. 2007. Indexing strategies for rapid searches of short words in genome sequences. *PLoS ONE*. 2:e579. doi: 10.1371/journal.pone.0000579.
- Ishikawa K, Fukuda E, Kobayashi I. 2010. Conflicts targeting epigenetic systems and their resolution by cell death: novel concepts for methyl-specific and other restriction systems. *DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes*. 17:325–342. doi: 10.1093/dnares/dsq027.
- Jen FE-C, Seib KL, Jennings MP. 2014. Phasevarions mediate epigenetic regulation of antimicrobial susceptibility in *Neisseria meningitidis*. *Antimicrob. Agents Chemother.* doi: 10.1128/AAC.00004-14.
- Jolley KA, Sun L, Moxon ER, Maiden MC. 2004. Dam inactivation in *Neisseria meningitidis*: prevalence among diverse hyperinvasive lineages. *BMC Microbiol.* 4:34.
- Kozdon JB et al. 2013. Global methylation state at base-pair resolution of the *Caulobacter* genome throughout the cell cycle. *Proc. Natl. Acad. Sci. U. S. A.* 110:E4658–67. doi: 10.1073/pnas.1319315110.
- Krebes J et al. 2014. The complex methylome of the human gastric pathogen *Helicobacter pylori*. *Nucleic Acids Res.* 42:2415–2432. doi: 10.1093/nar/gkt1201.
- Kurtz S et al. 2004. Versatile and open software for comparing large genomes. *Genome Biol.* 5:R12. doi: 10.1186/gb-2004-5-2-r12.
- Kwiatek A, Luczkiewicz M, Bandyra K, Stein DC, Piekarowicz A. 2010. *Neisseria gonorrhoeae* FA1090 carries genes encoding two classes of Vsr endonucleases. *J. Bacteriol.* 192:3951–3960. doi: 10.1128/JB.00098-10.
- Lamelas A et al. 2014. Emergence of a New Epidemic *Neisseria meningitidis* Serogroup A Clone in the African Meningitis Belt: High-Resolution Picture of Genomic Changes That Mediate Immune Evasion. *mBio*. 5:e01974–14. doi: 10.1128/mBio.01974-14.
- Lee M et al. 2012. Linezolid for treatment of chronic extensively drug-resistant tuberculosis. *N. Engl. J. Med.* 367:1508–1518. doi: 10.1056/NEJMoa1201964.
- Leimkugel J et al. 2007. Clonal waves of *Neisseria* colonisation and disease in the African meningitis belt: eight-year longitudinal study in northern Ghana. *PLoS Med.* 4:e101. doi: 10.1371/journal.pmed.0040101.
- Lluch-Senar M et al. 2013. Comprehensive methylome characterization of *Mycoplasma genitalium* and *Mycoplasma pneumoniae* at single-base resolution. *PLoS Genet.* 9:e1003191. doi: 10.1371/journal.pgen.1003191.
- Maiden MC. 2008. Population genomics: diversity and virulence in the *Neisseria*. *Curr. Opin. Microbiol.* 11:467–471. doi: 10.1016/j.mib.2008.09.002.
- Maiden MCJ. 2013. The endgame for serogroup a meningococcal disease in Africa? *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.* 56:364–366. doi: 10.1093/cid/cis896.
- Makarova KS, Wolf YI, Koonin EV. 2013. Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Res.* 41:4360–4377. doi: 10.1093/nar/gkt157.
- Manso AS et al. 2014. A random six-phase switch regulates pneumococcal virulence via global epigenetic changes. *Nat. Commun.* 5:5055. doi: 10.1038/ncomms6055.
- Marinus MG, Casadesus J. 2009. Roles of DNA adenine methylation in host-pathogen interactions: mismatch repair, transcriptional regulation, and more. *FEMS Microbiol. Rev.* 33:488–503. doi: 10.1111/j.1574-6976.2008.00159.x.
- Marri PR et al. 2010. Genome sequencing reveals widespread virulence gene exchange among human *Neisseria* species. *PLoS One*. 5:e11835. doi: 10.1371/journal.pone.0011835.
- Molina N, van Nimwegen E. 2007. Universal patterns of purifying selection at noncoding positions in bacteria. *Genome Res.* 18:148–60. doi: 10.1101/gr.6759507.
- Parkhill J et al. 2000. Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature*. 404:502–6. doi: 10.1038/35006655.
- Paul S, Million-Weaver S, Chattopadhyay S, Sokurenko E, Merrikh H. 2013. Accelerated gene evolution through replication-transcription conflicts. *Nature*. 495:512–515. doi: 10.1038/nature11989.
- Powers JG et al. 2013. Efficient and accurate whole genome assembly and methylome profiling of *E. coli*. *BMC Genomics*. 14:675. doi: 10.1186/1471-2164-14-675.
- Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinforma. Oxf. Engl.* 26:841–842. doi: 10.1093/bioinformatics/btq033.

- Richardson AR, Stojiljkovic I. 2001. Mismatch repair and the regulation of phase variation in *Neisseria meningitidis*. *Mol. Microbiol.* 40:645–55.
- Ritchot N, Roy PH. 1990. DNA methylation in *Neisseria gonorrhoeae* and other *Neisseriae*. *Gene.* 86:103–106.
- Roberts RJ, Carneiro MO, Schatz MC. 2013. The advantages of SMRT sequencing. *Genome Biol.* 14:405. doi: 10.1186/gb-2013-14-6-405.
- Roberts RJ, Vincze T, Posfai J, Macelis D. 2010. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.* 38:D234–236. doi: 10.1093/nar/gkp874.
- Rocha EPC. 2008. The organization of the bacterial genome. *Annu. Rev. Genet.* 42:211–233. doi: 10.1146/annurev.genet.42.110807.091653.
- Schadt E et al. 2012. Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases. *Genome Res.* 23:129–41. doi: 10.1101/gr.136739.111.
- Schoen C et al. 2011. Whole-genome sequence of the transformable *Neisseria meningitidis* serogroup A strain WUE2594. *J. Bacteriol.* 193:2064–2065. doi: 10.1128/JB.00084-11.
- Shell SS et al. 2013. DNA Methylation Impacts Gene Expression and Ensures Hypoxic Survival of *Mycobacterium tuberculosis*. *PLoS Pathog.* 9:e1003419. doi: 10.1371/journal.ppat.1003419.
- Srikhanta YN et al. 2009. Phasevarions mediate random switching of gene expression in pathogenic *Neisseria*. *PLoS Pathog.* 5:e1000400.
- Stephens DS, Greenwood B, Brandtzaeg P. 2007. Epidemic meningitis, meningococcaemia, and *Neisseria meningitidis*. *Lancet.* 369:2196–2210. doi: 10.1016/S0140-6736(07)61016-2.
- Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* 13:36–46. doi: 10.1038/nrg3117.
- Trivedi K, Tang CM, Exley RM. 2011. Mechanisms of meningococcal colonisation. *Trends Microbiol.* 19:456–63. doi: 10.1016/j.tim.2011.06.006.
- Walsh CP, Xu GL. 2006. Cytosine methylation and DNA repair. *Curr. Top. Microbiol. Immunol.* 301:283–315.

4 Exploring the phase-variable genome of *Neisseria meningitidis* from massively parallel sequencing data

Mohamad R. Abdul Sater^{1,2,3}, Araceli Lamelas^{1,2}, Nagwa Peter Thelma Niba^{1,2}, Gerd Pluschke and Christoph D. Schmid^{1,2,3*}

¹Swiss Tropical and Public Health Institute, Socinstrasse 57, CH-4002 Basel, Switzerland

²Universität Basel, Petersplatz 1, CH-4003 Basel, Switzerland

³ SIB Swiss Institute of Bioinformatics, Basel 4056, Switzerland

* Corresponding author

Manuscript in preparation

4.1 Abstract

Massively parallel sequencing methods are becoming routine, yet repetitive sequences present in most genomes still poses numerous difficulties in data analysis. The commonly applied exclusion of these sequence parts appears not satisfactory, as they are informative for genomic diversity or the prediction of gene functionality. Hyper-mutable short tandem repeats cause ON/OFF switching of phase-variable genes in bacterial genomes. This rapid and reversible mechanism offers bacteria a prompt response to environmental stress.

We present here a novel method to infer the precise number of repeat units at specific tandem repeat loci based on the raw read sequences resulting from large scale sequencing assays. We demonstrate that the probabilistic approach, based on Hidden Markov Model (HMM), detects divergent repeat length configurations and therefore functional states of ORFs directly from raw sequencing data, offering an enhanced detection power and accuracy over conventional tools. We integrated our tools into a fast and efficient computational pipeline to detect genome wide phase-variation events by comparing a large number of sequenced meningococcal genomes. Our comprehensive approach identified a high number of hyper-variable repeat regions. Our ongoing analysis, have so far revealed the top ranking phase-variable loci to be mostly associated with outer membrane components and other virulence determinants.

4.2 Introduction

N. meningitidis has a very high potential for genetic change and surface structural variability. The major outer membrane components (capsular polysaccharide, outer membrane proteins, and lipopolysaccharide (endotoxin); are linked to meningococcal virulence (Stephens et al., 2007). Variable expression of surface antigens in *Neisseria* is suggested to be key in facilitating colonization of the host, adaptation to host environments and evasion of immune-responses (Stephens, 2009). Genome analysis and phylogenetics revealed high recombination rate as a result of horizontal gene transfer between closely related *Neisseria* species (Schoen et al., 2009). Frequent recombination within meningococcal populations was proposed to contribute to immune evasion by hyper virulent clones via the introduction of intricate changes in the antigenic makeup of the bacterial cell surface components (Lamelas et al., 2014). However, growing evidence also suggests phase-variation as a mechanism triggering immune evasion by host-adapted bacterial pathogens through rapid and reversible ON /OFF switching of expression of specific genes mediated by unstable short tandem repeat sequences (Goldberg et al., 2014).

Novel sequencing approaches reveal considerable variability in genomic sequences, even so in cells derived from an individual organism or from closely related populations of the same bacterial strain. In particular, prokaryotic genomes underlay reduced evolutionary pressure to conserve global genome structure due to the lack of chromosome pairing in meiosis. Commonly, single nucleotide polymorphisms (SNPs) and copy number variation (CNV) are used for comparing genomes. A third class of genotypic variation is tandem repeats including micro- and mini-satellites, also known as simple sequence repeats (SSRs) or short tandem repeats (STRs). These sequences have historically been ignored as “junk DNA”, in part due to the difficulties associated with the study of these unstable repeat tracts. Variations in those repetitive sequences can have significant phenotypic consequences, if located within regulatory sequence elements or within coding regions (Gemayel et al., 2010).

Variable tandem repeat sequences can for instance lead to a reversible, random, high frequency gain or loss of a phenotype, a phenomenon known as phase-variation. An insertion or deletion of a repeat unit with a length divergent from a multiple of 3bp within a coding sequence will lead to frame shift mutations. Consequently the translation into proteins is switched OFF or conversely reverted back to an ON state with a full-length open reading frame (Figure 4-1). At the example of *Neisseria* subspecies, a number of phase-variable genes have been proposed (summarized in section 1.4.5).

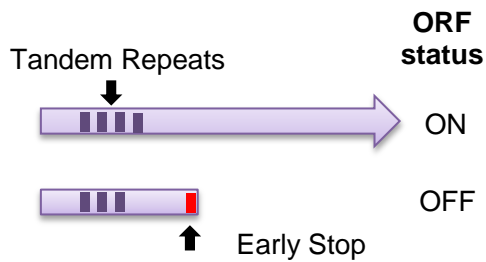


Figure 4-1: ON/OFF switching mechanism of Phase-variable gene determined by repeat unit number (dark bars). Adding or deleting a repeat unit leads to a frame shift and an early stop codon or reversibly restoring original frame.

Recent massively parallel sequencing methods enable the efficient determination of nucleotide sequences of a large number of short DNA fragments. Those fragment sequences are then typically assembled into larger contigs and eventually to complete chromosome assemblies. Repetitive regions are however very problematic for both mapping based approaches as well as *de novo* assembly of genomic sequences. In both approaches, repetitive regions interfere with reads scaffolding as it is impossible to attribute such reads to their original locus. In order to determine SNPs, the read sequences are frequently aligned to the best matching position in a reference genome sequence assembly. A high coverage of the genome by short reads makes this a very powerful approach to call SNPs and short indels (Koboldt et al., 2009). However, larger divergences due to variable unit numbers of tandem repeats pose considerable problems to assembly and mapping approaches (Treangen and Salzberg, 2012). Therefore, the results of high-throughput sequencing methods have to date mainly been exploited to predict suitable PCR primer sequences for further analysis of potentially divergent tandem repeat regions (Guichoux et al., 2011) and only a few studies attempted to assess the microsatellite directly from sequencing data (Kim et al., 2013). The precise resolution of longer microsatellite repeat stretches obviously benefits from sequencing methods providing longer read lengths, yet for instance the 454 pyrosequencing method suffers from reported short comings to achieve sufficient coverage and problems with homopolymer stretches.

A rapidly growing number of genome sequences with the lower sequencing costs enable large scale sequencing experiments. Analyzing the enormous amount of generated data represents the bottleneck in current research. A common task is to compare specific regions in a large number of sequenced genomes with fast and precise methods providing a simple to interpret output. At the example of bacterial repeat sequences and phase-variable genes whole genome mapping and/or *de novo* assembly of a large number of genomes could be a time consuming and a computationally heavy approach which would ultimately require additional analysis steps to compare such regions. Moreover, longer repetitive sequences

have always been problematic for mapping and assembly programs. Consequently, sequence variants resulting from events such as phase-variation are ignored or vastly under estimated.

In this chapter we introduce a fast and accurate analysis pipeline for a direct quantification and comparison of repeat regions variability from sequencing experiments involving a large number of genomes circumventing mapping and genome assembly. Our bioinformatic pipeline is the first comprehensive attempt to identify *Neisseria meningitidis* phase-variable genome from a sequenced population. The pipeline contains a newly developed tool that provides an extended detection power for longer repeat region and allows detection of subtle and rare sequence variants tolerating sequence mismatches as compared to other conventional approaches. The present pipeline relies on both efficient string matching and probabilistic methods (Durbin et al., 1999), a combination of both approaches is required by the large number of read sequences typically required to obtain sufficient sequencing coverage and the occurrence of mutations and sequencing errors disabling 'simple' pattern matching approaches. The resultant is a simple and easy to analyze tabular output (figure 4-6).

4.3 Methods

4.3.1 Identify short tandem repeats using Phobos

STRs, as potentially phase variable loci, were detected from *N. meningitidis* strain WUE 2594 reference genome (Schoen et al., 2011) (accession number: FR774048.1) using Phobos tandem repeat search tool (Mayer, Christoph, Phobos 3.3.11) (Mayer et al., 2010). Program parameters were set to match exact repeats ranging from 1 and up to 15 nucleotides per unit and score set to maximum with zero gaps allowed. An early filtering step depending on the repeat unit length was applied, whereby repeat units consisting of a multiple of 3 nucleotides and occurring inside ORFs were discarded as these would not lead to frame shift mutations. In addition, homopolymer tracts length threshold was set to 5 nucleotides, following reports on homopolymer relative stability (Jennings et al., 1999; Snyder et al., 2001).

Using custom Perl scripts a 5 nucleotide flanking sequence was extracted acting as upstream and downstream anchor of the repeat region. In cases where the flanking sequence was not sufficiently divergent from the repeat unit sequence (<2 different nucleotide) the flanking sequence was extended to eight nucleotides.

4.3.2 Regular expression fast approach for an exact sequence matching

A set of 100 strains derived from *Neisseria meningitidis* isolates were sequenced to an average coverage of ~300x using a standard paired-end 75bp read-length Illumina protocol (Lamelas et al., 2014). Quality scores and paired-end information contained in the original raw sequencing fastq files were not considered.

Using the ~9000 short tandem repeats identified by the genome wide scan for repetitive sequences, we designed custom Perl scripts to scan raw fastq files for reads containing matching repeat sequences. Our approach utilizes string matching algorithm (RegEx) to match anchor sequences flanking any repeat unit conformation in the form of [5' flanking — (Repeat Unit)*n* — 3' flanking]. For each matched read, the number of repeat units (*n*) is calculated. Restricted by the read length the upper limit of *n* is dependent on the repeat unit length allowing a 5bp flanking sequences. The algorithm reports for each locus the number of reads matching a discrete repeat unit conformation. The final output is a tabular format comparing for each locus the repeat length from each of the sequenced isolates.

A Perl script is provided whereby the flanking and repeat unit sequences are fed in tabular format as well as the input fastq files.

4.3.3 RepHMM an exhaustive approach for approximate sequence matching

RepHMM in a nutshell:

- (1) Identification of candidate loci with short tandem repeats in reference genome assembly.
- (2) Pre-selection of read sequences from raw sequencing data (fastq format) containing at least two perfect repeat units.
- (3) Generate a probabilistic model based on Hidden Markov Model (HMM) encoding for a range of distinct conformations represented by repeat unit(s) and flanking sequence.
- (4) Score all pre-selected reads with probabilistic model of repeat configurations (HMM), select reads with high-probability in Viterbi decoding.
- (5) Infer predominant repeat conformation based on number of reads.

Generating HMM model and scoring of preselected reads

For each read sequence, a precise count of repeat units in the context of specific flanking sequences was determined by applying a Hidden Markov Model (HMM) approach. The HMM encodes a set of paths with flanking sequences and divergent numbers of repeat units (schema of HMM Figure 4-2). Based on an estimated Illumina sequencing error frequency, the emission probabilities for mismatches within the repeats state was set to 0.01. Acting as boundaries, flanking sequences of the repeat region are very crucial for an accurate and specific counting of repeat unit number, thus a higher stringency was set by applying reduced emission probabilities for mismatches within the flanking sequences to increase specificity at this critical region. Within the random (R-) state equal emission probabilities for each base (A, T, G or C) were set to 0.25. A Perl script (Mamot_model.pl) is provided to generate the HMM model as input for MAMOT (see below). Simulated read sets with specified numbers of repeat units and randomly inserted mismatches were used to evaluate our model. Emission and transition probabilities were optimized for best distinction of each repeat path based on score threshold.

The tool automatically calculates the optimal score threshold adapting to read length and the user input for the preferred number of allowed mismatches resulting from sequencing error.

Decoding

Each of the preselected read sequences was decoded by the Viterbi algorithm using the Markov Modeling Tool MAMOT (revision 77M) (Schutz and Delorenzi, 2008). From the MAMOT output the most probable path with a score above threshold was extracted using Perl scripts (Supplementary material). The scoring and decoding of the 4 million 75 bp preselected reads for *modA12* locus required 2 hours run time using a standard single core system. For the visualization of read diversity, sequences were aligned using Jalview (Clamp et al., 2004). An example of a RepHMM model and Perl scripts along with a training dataset is available as supplementary materials at (<http://www.swisstph.ch/Rep-HMM>).

Pre-selection of candidate reads

Using the Linux command-line pattern matching utility *grep*, we extracted from the .fastq files for each *Neisseria* strain all raw read sequences containing at least 2 perfectly matching repeat units. Quality scores and paired-end information contained in the original fastq files were not considered.

4.3.4 Generation of simulated data and comparison of performance

In phase-variation, sequence variants deviating from the ubiquitous repeat unit genotype are relatively infrequent. In order to quantify the potentials of RepHMM to detect divergent repeat unit genotypes within the same pool of reads we generated, using custom Perl scripts, a simulated data set of sequences using the phase-variable *modA12* locus (flanking) (AGCC x N) (flanking) as template where “AGCC” is the repeat unit sequence. Limited by read length of 75bp matching our actual Illumina read length the maximum number of AGCC repeats was set to 16x units which could still include anchor flanking sequence. For each repeat genotype (2-16x repeats) a simulated data set of 75bp reads with 5% mismatch rate and 300x coverage was generated. In addition, 3-5 random genotype variants of ± 3 repeat units constituting up to 10% of total read count were added to the read pool.

Each data set was analyzed using *de novo* assembly, simple pattern matching RegEx and RepHMM. Performance of each method was assessed by calculating the ratio of called genotypes relative to actual number of simulated genotypes. *De novo* assembly was attempted using Velvet ($3 \leq \text{kmer} \leq 31$), RegEx recognition pattern was set to match

(flanking)(AGCC{1,16})(flanking) counting the number of AGCC repeat units. 100x Simulation runs were performed.

4.3.5 Bacterial cultivation, PCR and Sanger Sequencing

Single colonies of *Neisseria meningitidis* strains were grown on supplemented GC agar base (Oxoid) plates for 20-24 hours in 5% CO₂ at 37°C. Colonies were transferred into liquid Brain Heart Infusion (Bacto™) medium and again incubated overnight in 5% CO₂ at 37°C. Chromosomal DNA was extracted as described previously in (Marri et al., 2010). DNA extraction was performed using Promega DNA extraction kit according to manufacturer specifications. PCR Primers were designed, avoiding identified, SNPs using *N. meningitidis* Z2491 reference genome. Forward primer: 'aatagccaaccgcatgg', Reverse primer: 'ttgttgccgtctgcgg' generating a 300bp amplicon spanning flanking and repeat regions of the modA12 locus. Sequencing of the PCR amplicons by the Sanger method was performed at Macrogen (Korea).

4.3.6 SNP calling

SNP calling was performed as described in the previous chapter, with 100 sequenced genomes mapped to the reference genome *N. meningitidis* serogroup A, strain WUE 2594 (Schoen et al., 2011) using SMALT (<http://www.sanger.ac.uk/resources/software/smalt/>). Sequence variations were determined using SAMtools (Li et al., 2009), excluding phage sequences, recombinant fragments and repetitive regions (>50bp) (Croucher et al., 2011). Anchor sequences of the reference genome overlapping SNPs were identified using Bed intersect tool (Quinlan and Hall, 2010).

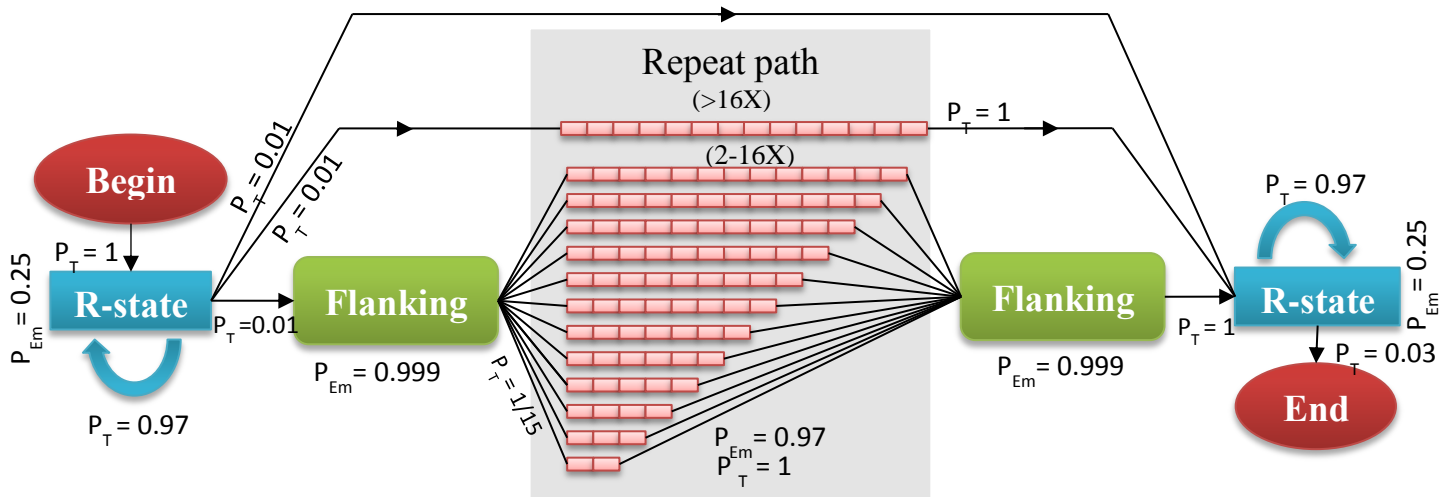


Figure 4-2: Schematic view of RepHMM used for assessing phase-variable repeat sequences. A probabilistic approach based on a Hidden Markov Model (HMM) for scoring high-throughput sequencing reads. The model encodes DNA sequences in a range of distinct conformations denoted by repeat units represented by pink blocks ('AGCC' for the *modA12* locus) and sequences flanking the repeats (4-6 bases each). Additional flanking sequences are encoded by the R-state. Transition and emission probabilities at each sequence position are specified in the model, the schema indicates the major probabilities. P_{Em} indicates emission probabilities for bases matching the sequences of repeat units, or flanking sequences, respectively. Given the max. read-length of 75bp in this study, longer repeat stretches (>16x) are matched by a path lacking flanking sequences.

4.4 Results

Phase-variable loci covered by large scale sequencing experiments of bacterial genomes led us to develop a bioinformatic pipeline to detect phase-variation events within *N. meningitidis* population by counting and comparing short tandem repeat unit numbers.

Comparing a large number of genomes using approaches such as mapping and *de novo* assembly was inefficient. Moreover an assembly approach, if successful, yielded in only the predominant length of the repeat loci, and did not allow the detection of repeat variants within the “clonal” population. We thus developed computational tools, adaptable to sequence data from all organisms for an enhanced detection and analysis of repetitive sequences.

4.4.1 Development of a flexible microsatellite repeat typing tool (RepHMM)

The high sequencing coverage guarantees a locus to be spanned by multiple reads and thus overcoming random sequencing errors. The number of reads is however inversely proportional to a locus sequence length, thus longer repeat stretches are completely covered by fewer numbers of sequencing reads. In addition, minor repeat variants signifying clonal variability also have a reduced coverage. Simple pattern matching via RegEx failed at longer repeat regions due to its intolerance of sequencing errors. Likewise SNPs with respect to reference genome also crippled RegEx matching. We therefore developed a novel tool that allows for an extended detection power for longer repeat region, detection of subtle and rare sequence variants as well as tolerating SNPs and sequencing errors.

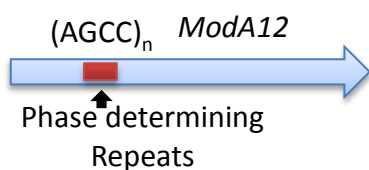


Figure 4-3: *modA12* gene encoding type III methyltransferase with the phase determining repeat region (AGCC) inside the ORF.

For this part we will focus on established phase-variable loci of the *mod* and *opa* gene families. *modA12* is a member of the *mod* gene family associated with type III restriction–modification systems and phase-variable tandem repeats in the N-terminal region of the ORF (Srikhanta et al., 2010) can lead to 'pseudogene' annotations at these loci in several genome assemblies (Figure 4-3). The repeat sequence 'AGCC' in the context of 5' and 3' flanking sequences 'CAGTAAT' and 'AATTAT' respectively, is specific for the *modA12* locus in the assemblies of the *N. meningitidis* genomes.

The workflow starts with a pre-selection of sequence reads containing at least two repeat units as identified by Phobos (see methods). This pre-selection step reduced the total amount of reads to be scored by the probabilistic model, thereby reducing the computational time required. At the example of the repeat sequences of the *modA12* loci (Figure 4-3), the pre-selection of reads containing 'AGCCAGCC' reduced the number of reads to assess from over 6 billion to 4 million sequencing reads. The assessed data set contained reads with a maximal length of 75bp.

Scoring the pre-selected reads with an HMM tolerated defined rate of (sequencing) errors while specifically identifying read sequences covering a phase-variable locus. For each set of read sequences, the number of reads best matching a specific conformation with repeat units and locus-specific flanking sequences was reported (Table 4-1). Specificity was ensured by stringent matching of the flanking sequences (Figure 4-2).

#Repeats Strain	4	5	6	7	8	9	10	11	12	13	14	15	16	>16
NM1264*	310													
NM1325		1			2	9	52	2	1					
NM1446*		1	199	2										
NM1471*		288	1											
NM1673	2	1	3				1						2	
NM1901*			285	3	1		1							
NM2335												3		
NM2193	1	1								1	10			
NM2369			1					26				1		
NM2933														2

Table 4-1: RepHMM Output, a direct read count comparison of repeat unit conformation between and within sequenced strains.

A representative set of strains showing the number of sequencing reads detected for each repeat length conformation identifying variability at the *modA12* locus between different strains as well as within strains. Green background indicates repeat unit numbers resulting in a complete ORF (gene ON). Grey background marks repeat unit conformations confirmed by Sanger sequencing. (*) Representative reads from these strains are extracted and aligned in Figure 4-5. Table S4-1 contains the complete dataset of all sequenced strains.

4.4.2 RepHMM outperforms alternative approaches

RepHMM featured substantial repeat length quantification advantages mainly allowing for longer repeat conformation quantification in both simulated data as well as sequencing data. Notably, RepHMM also provided a higher detection resolution for read variants deviating from the predominant genotype (repeat number) denoting clonal phase-variation.

Comparing RepHMM versus RegEx matching using simulated reads of ModA12 locus

Phase-variation events leading to few read variants conferring a divergent repeat genotype usually constitute a minor fraction of the predominant repeat genotype. Thus, in our simulated reads dataset we generated in addition to the predominant read conformation a few number of reads of ± 3 repeat units resulting in a total of 3-5 divergent repeat genotypes. We also introduced sporadic mismatches, mimicking random sequencing errors or SNPs, at a 5% rate for each generated read matching the *modA12* locus (methods). The ratio of accurately called genotypes to the actual simulated genotypes assesses the performance of each approach.

Using the simulated reads dataset, we compared our RepHMM to RegEx pattern matching (Figure 4-4) in the form of (flanking)(AGCC){1,16}(flanking). Sequencing errors are overcome by higher sequencing coverage, thus the probability to obtain error free sequences increases with number of sequenced reads covering a particular locus. Consequently, the likelihood to find the predominant repeat genotype using RegEx is high; conversely the likelihood of error free sequences drops with longer repeat stretches or for rare genotype variants (Figure 4-4). The number of reads spanning whole repeat length is inversely-proportional to the repeat unit number hence the observed drop in genotype calling using RegEx towards longer repeats (Figure 4-4). Therefore, longer repeat stretches and sequence mismatches rendered RegEx inefficient compared to a much improved detection power using RepHMM.

Velvet de novo assembly performance was severely impaired after introducing repeat genotype variability (scanning all $3 \leq kmer \leq 31$ options), whereby the contig assembly failed if the repeat unit number is >4 (Figure 4-4) whereas a uniform genotype allowed the assembly of longer repeat stretches (Table 4-2).

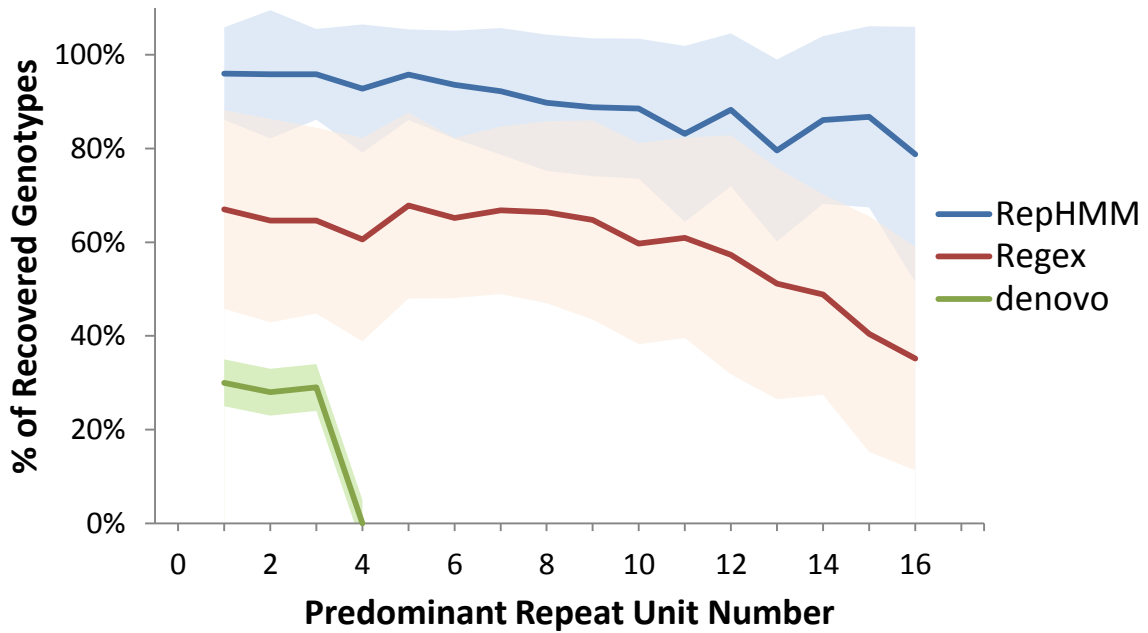


Figure 4-4: RepHMM efficiently quantifies genotypic variability from simulated pool of reads. Comparing performance of RepHMM, RegEx and *de novo* assembly in detecting genotype variants. For each repeat unit genotype (number of repeats), constituting the predominant read count, there exists 3-5 random genotype variants of ± 3 repeat units constituting up to 10% of the simulated read count (methods). Plotted is the mean ratio of genotypes called by each approach shaded by the corresponding standard deviation of 100 simulation rounds.

RepHMM efficiently quantifies population and clonal variability of modA12 locus from Illumina sequenced genomes

In order to quantify the actual phase-variation at *modA12* locus between the sequenced *Neisseria meningitidis* isolates, we applied our RepHMM approach to the Illumina sequenced genomes. Similar to the simulated data, we also compared RepHMM performance to RegEx and *de novo* assembly. Both approaches detected the predominant repeat conformations at *modA12* locus, provided the length of the repeats remained below 14 units (Table 4-2, Figure 4-5A).

Longer repeat stretches (14-16 repeat units) can still include flanking sequences given the 75bp read length; however restricting the precise coordinates of the genomic origins of the reads reduces also the expected coverage. In these situations RepHMM was clearly more sensitive to correctly quantifying repeat unit numbers than both *de novo* assembly and RegEx. Correctness of RepHMM was verified for 45 selected genomes via PCR sequencing where we found a 100% co-incidence of the predominant repeat unit number determined by RepHMM (Table 4-2).

Repeat units #	Count of strains				Number of reads	
	<i>De novo Assembly</i>	RegEx	RepHMM	PCR sequencing	RegEx	RepHMM
2-13	56	56	56	3*	5390	7620
14	2	9	14	14	27	83
15	0 †	1	7	7	3	12
16	0 †	0	1	1	0	4
>16	0 †	0	22	20	6	106
ND	42	34	0	7	-	-

Table 4-2: RepHMM efficiently detects longer repeat conformations.

Comparing aggregate counts of reads from the sequencing of 100 *N. meningitidis* strains targeting sequences derived from the *modA12* loci. Comparing *de novo* assembly and simple string matching (RegEx) to RepHMM specificity using PCR Sanger sequencing as the gold mark standard. Count of strains represents the number of similar cases out of 100 sequenced strains. (†) Contig break. (ND): Number of strains with a non-determined repeat conformation by each approach. (*) 4 strains tested with PCR sequencing as control.

The read length inherently limits the number of detectable repeat units by RepHMM, nevertheless the model is still capable of correctly reporting repeat unit number above a certain length compared to a contig break (Table 4-2) or even wrongly reported repeat length

in *de novo* assemblies where 40% of the cases with >14 units were incorrectly reported as shorter repeat length.

RepHMM with its tolerance of sequencing errors did not only recover reads with longer repeats but also allowed detecting rare genotype variants. Figure 4-5B shows extracted and aligned reads with a divergent repeat length from the same strain most likely resulting from clonal phase-variation events. As described in the introduction, such phase-variation occurring within a clonal population is rare (~1 in 300); likewise the corresponding number of reads would be rare. Thus, analogous to what we observe with the simulation data set described earlier, sequencing errors as observed in figure 4-5B hampered the detection of rare read variants using RegEx. *De novo* assembly and PCR sequencing also could not detect such 'mixed' repeat genotype conformations only reporting the predominant repeat genotype (Table 4-2).

In cases where we observe a comparable number of reads at two or more distinct repeat genotypes such as NM1673 in Table 4-1, we derived a mathematical formula in order to infer the predominant repeat length. The formula calculates the expected number of reads using the average coverage at adjacent genomic regions and average read length. The predominant clonal repeat genotype is inferred by comparing the ratio of the number of reads observed/expected for each genotype variant, a higher ratio designates the predominant genotype. In the case of NM1673 (Table 4-1), the highest ratio corresponds to repeat unit conformation (16) compared to the other conformations, indicating the predominant clonal genotype is 16 repeat units and an ON ORF status.

$$N_{\text{reads}} = X_{\text{cov}} - \frac{X_{\text{cov}}}{L_{\text{rd}}} [(U_{\text{rpt}} \times n_{\text{rpt}}) + L_{\text{flk}} - 1]$$

Xcov standing for coverage, L_{rd} read length and L_{flk} for pre-specified flanking length and U_{rpt} for repeat unit length and n_{rpt} for total number of tandem repeats.

With further optimization, the formula could be utilized to estimate the repeat unit number in cases where the repeat region is longer than the read length (>16xAGCC). At this stage we can only presume, the higher the ratio of the number of reads observed/expected >1 the longer the repeat region.

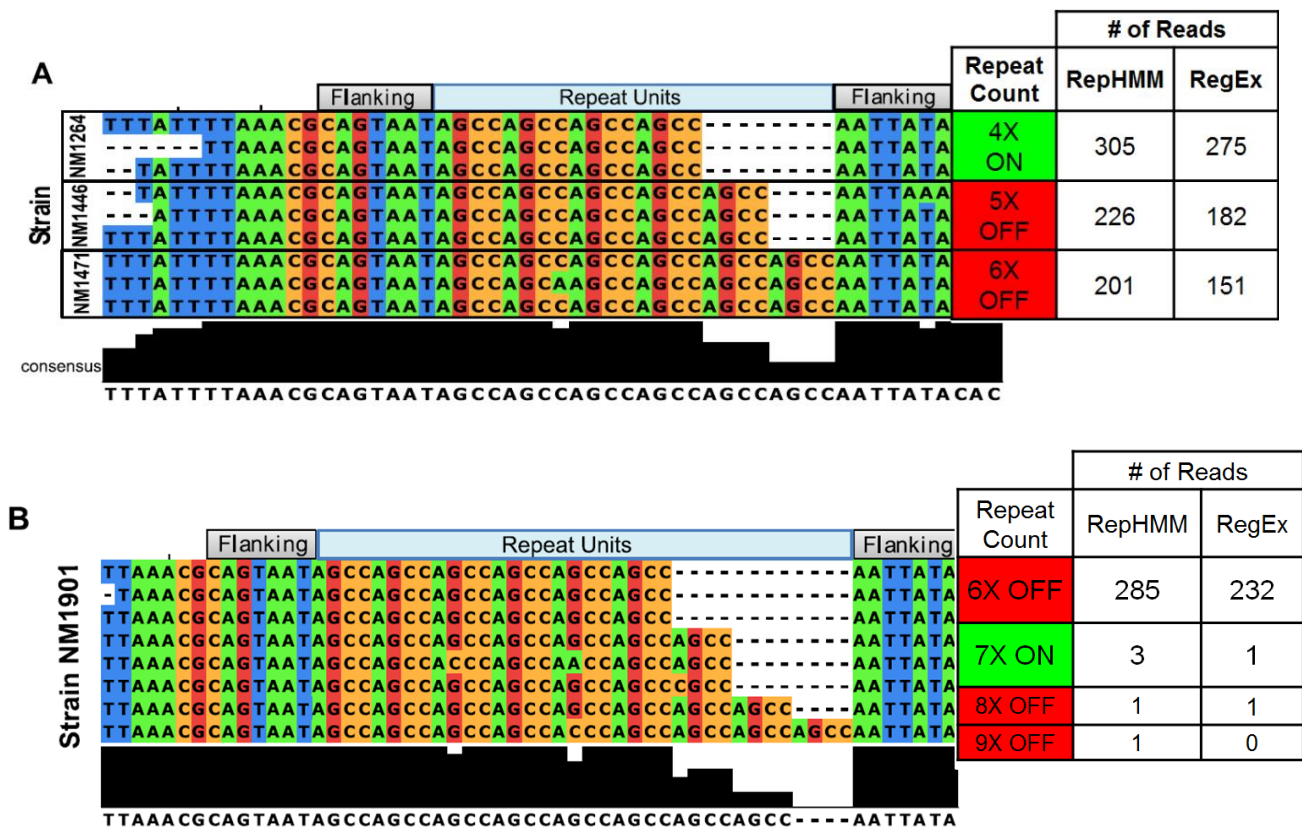


Figure 4-5: Alignment of reads sequences at *modA12* locus. (A) divergent counts of 'AGCC' repeat units in 3 different *N. meningitidis* strains (NM1264, NM1446, and NM1471) separated by boxes as identified by RepHMM. Grey boxes on top of the alignment indicate flanking sequences used to delineate the repeat region. The table to the right indicates the number of reads derived from each strain and repeat unit conformation as identified by RepHMM compared to RegEx. (B) Divergent repeat conformations within the same clonal population are aligned displaying a high read count for predominant repeat length conformation and lower counts for read variants. Please note interspersed sequence variants disabling RegEx pattern matching approaches.

4.4.3 Evaluating RepHMM at multi-copy gene duplicates.

Tandem repeats are a popular target in molecular epidemiology, yet current genotyping assays including multiplex PCR remain laborious and time consuming. We therefore compared the repeat unit counts at the *opa* loci determined by RepHMM with genotyping assays applied to matching DNA samples (Huber et al., 2012). 4 *opa* gene copies are present in the *N. meningitidis* genomes with identical repeat and immediate flanking sequences, yet having variable repeat unit numbers. Specific PCR primer sequences located at larger distances to the repeats allowed assessing individual *opa* loci, while RepHMM determined repeat unit counts combined over all loci sharing identical flanking sequences.

Table 4-3 summarizes our findings showing RepHMM results perfectly correlating with PCR sequencing results combined over the 4 *opa* loci, if considering read counts >12 as detection threshold, due to longer repeat unit length ('CTTCT'). Moreover, similar to *modA12*, RepHMM also detects read variants representing clonal variability at the phase-variable *opa* locus.

Neisseria isolate	Repeat unit number						
	7x	8x	9x	10x	11x	12x	>12x
NM2008	0	0	7	119	6	22	na*
NM2009	0	0	9	68	47	17	na*
NM2700	0	80	3	3	35	7	na*
NM2701	0	127	4	9	53	12	na*

Table 4-3. Validation of *opa* loci by direct PCR sequencing

RepHMM determines for specific strains (in rows) the predominant counts of 'CTTCT' repeat units at *opa* loci. Each cell of the table indicates the number of sequencing reads best matching with the corresponding repeat conformation (and flanking sequences) specifying the 4 *opa* loci. Each individual *opa* locus in the corresponding strain displays a specific repeat conformation as detected by direct PCR-sequencing, summarized in the table by grey background. (*): the maximal read length of 75bp in the present sequencing experiments disables the detection of repeat configurations with more than 12 repeats (12x5bp repeats + 2x6bp flanking sequences).

4.4.4 Integrated Pipeline for identification of Phase-variable Genome

Pipeline development

Having established the utility of RepHMM in detecting repeat variation, we integrated RepHMM and RegEx into a pipeline for a comprehensive identification of phase-variable genes from massively parallel sequencing data of closely related bacterial population (Figure 4-6). The pipeline begins with a thorough scan of reference genome, in this case *N. meningitidis* WUE 2594, for short tandem repeats using Phobos repeat finder (section 4.3.1). The highly sensitive scan identified 9000 loci containing a repeat sequence.

Using two synergic methods, sequencing reads matching a flexible length of each of the identified STRs in the context of fixed flanking sequences (anchors) are detected directly from raw sequencing data (fastq files). For each detected read the corresponding repeat unit number is calculated. The approach integrates the fast and exact Regular Expression

(Regex) string matching algorithm (section 4.3.2) with the slower yet exhaustive RepHMM tool, based on a Hidden Markov Model (HMM) read scoring (sections 4.3.3; 4.4.1).

SNPs, sequencing errors and reduced coverage severely cripple exact matching algorithms especially at longer repeat sequences. Our probabilistic tool RepHMM tolerates mismatches and allows a thorough detection of repeat variant complementing the fast and exact Regex sequence matching. Despite its advantages, the complexity of the HMM algorithms would nevertheless render scanning six billion reads (of 100 genomes) computationally demanding and time wise futile. Thus, loci found to contain SNPs in the flanking sequences (~450 loci) are scanned for by the approximate matching tool RepHMM, whereas conserved loci are scanned for by the custom designed Regex based Perl scripts. In cases where Regex fails for reasons such as longer repetitive region and sequencing errors, or requiring a higher resolution to detect minor genotype variants, a targeted RepHMM is applied per genome or locus. A combination of those two methods vastly reduces the computational and time constraints.

A variable repeat number at each locus across the sequenced isolates indicates a hyper-variable microsatellite region undergoing extension/contraction mechanism. Loci falling inside ORFs lead to direct phase-variation events thereby allow inferring the phase-variable genome within the studied population. Loci occurring in the intergenic region might have an indirect effect and would be unfeasible to computationally predict their outcome.

An isolate phasotype is deduced by analyzing combinatorial ON/OFF conformation of each phase-variable gene, and the aggregate proportion of ON/OFF variants within a population denotes the phasevariome.

The output is sorted according to degree of repeat length variability within the population. In the case of homopolymer repeat regions it is impossible to distinguish an insertion or deletion (indel) from a phase-variation event as both would have the same outcome. Therefore, we set a rather stringent threshold of >10 variants across the 100 genomes representing a high frequency of localized variability consistent with phase-variation rather than indel mutation or sequencing errors which are expected to be more sporadic.

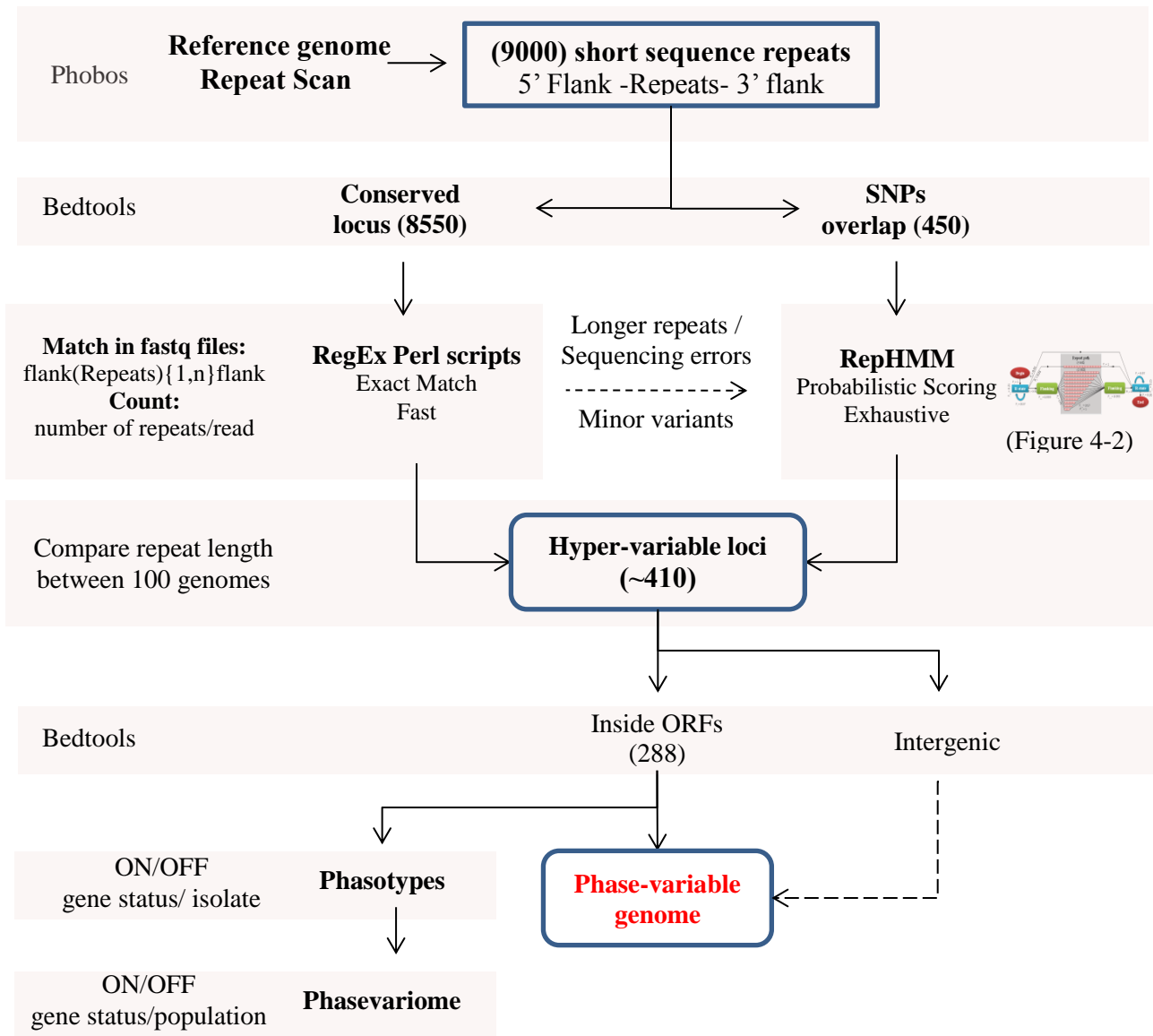


Figure 4-6: Pipeline to identify phase-variable genome from a sequenced population of *N. meningitidis* isolates. The method or tool used is indicated adjacent to each step. In brackets the number of loci determined by the corresponding step.

Analysis of pipeline output

Our ongoing analysis of the large output has so far revealed 288 hyper-variable repeat regions occurring inside gene bodies. Table 4-4 summarizes the top 70 genes based on repeat unit length and degree of variability within our population of isolates. Our pipeline successfully identifies genes known to undergo phase-variation (blue cells, Table 4-4) such as the *mod* and *opa* genes. Analyzing and publishing the first whole genome sequence of *N. meningitidis* strain Z2491 (Parkhill et al., 2000) described 27 repeat regions potentially associated with phase-variation in a serogroup A. The following year, Snyder et al. expanded the predicted list to 57 putative phase-variable genes by conducting a basic comparison of repetitive loci from three distinct whole genome sequences of *Neisseria* genus available at the time. Analyzing our 100 closely related genomes, 21 out of 57 putative phase-variable genes appear in the top ranking phase-variable loci (yellow cells, Table 4-4). 15 loci appear to have no variation or are not present in our reference genome. The rest of the putative loci appear to have a reduced degree of variation in our population.

Examining the main association of the top ranking hits shows the majority of the genes in Table 4-4 are associated with cell surface structures, adhesion and other virulence factors. In particular, 14 genes are associated with adhesion and pili (*Opa* alleles, *nhhA*), pilus-associated proteins (*pilC*) and glycosylation (*pgl*, *igtG*, *rfaG1*), five lipopolysaccharide associated genes (*Lgt*, *lpxC*) and two capsule biosynthesis proteins (*sacB*, *sacD*). Ten additional genes are associated with membrane proteins, six of which are outer membrane proteins (*porA*, *autB*). Genes involved in DNA methylation (*modA*, *modB*), iron binding (*tonB*, *fetA*) and phage associated (*intA*) also appear to phase-vary.

Table 4-4. Top loci having variable repeat unit length with in the meningococcal serogroup A population. The gene and the corresponding repeat unit and the degree of variability based on the number of variants within the population is indicated (H= High >10 variants, M= Medium >5 variants, L=Low >3). Where available the product and main association are listed for each gene. Blue highlights genes reported as phase variable, yellow highlights indicates putative phase-variable genes predicted by (Snyder et al., 2001).

Gene	Repeat Unit	Degree of variability	Main Association	Product
opaA	CTTCT	H	Adhesion	Opacity protein
opaB	CTTCT	H	Adhesion	Opacity protein
opaC	CTTCT	H	Adhesion	Opacity protein
opaD	CTTCT	H	Adhesion	Opacity protein
modA12	AGCC	H	DNA methylation	type III restriction/modification system
modB2	TTGGG	H	DNA methylation	type III restriction-modification system
cotSA	C	H	pili	pilin glycosyl transferase A
UbiE	T	H		Ubiquinone/menaquinone biosynthesis
SacD	T	H	capsule	Capsule biosynthesis protein
pcm	T	H		Protein-L-isoaspartate O-methyltransferase
pilC1	C	H	pili	Type IV pilus-associated protein
NMAA_1262	A	H		Uncharacterized protein
M.NlaIV	A	H	DNA methylation	methyltransferase
SstT	A	H	membrane	inner membrane Serine/threonine transporter
glmS	A	H		Glutamine-fructose6phosphateamino-transfer
icd	T	H		Isocitrate dehydrogenase
pilc2	G	H	pili	Type IV pilus-associated protein
mfpsA (lgtA)	G	H	LPS	lipooligosaccharide glycosyl transferase A
lgtG	G	H	LPS	lipooligosaccharide glycosyl transferase G
pgII	G	H	pili	putative LipO-oligosaccharide acyltransferase
pgIA	G	H	pili	pilin glycosyl transferase A
ssa1	G	H		extracellular serine protease precursor
porA	C	H	membrane	outer membrane porin protein
NMAA_0422	T	H		flavoprotein-ubiquinone oxidoreductase
NMAA_1672	T	H		Putative metallopeptidase
sacB	A	H	capsule	Capsular polysaccharide phosphotransferase
NMAA_0828	T	H		Putative phosphatase
thiC	A	H		Phosphomethylpyrimidine synthase
Alx	A	H	membrane	membrane protein
NMAA_0064	T	H		Uncharacterized protein
NMAA_0123	C	H		Uncharacterized protein
tbp2	T	H	membrane	outer membrane Transferrin-binding protein 2

lpxC	CG	M	LPS	lipid A biosynthesis, LPS anchor to outer membrane
pdxA	GC	M		NAD(P)-dependent oxidation
potG	GGCAA	M		Putrescine transport ATP-binding protein
NMAA_1079	AAGAA	M		Uncharacterized protein
ssb	GCGGC	M		recombination and DNA repair
MtrC	GC	M	membrane	Cell membrane lipoprotein
topA	T	M		DNA topoisomerase 1
NMAA_1879	T	M	membrane	Putative membrane protein
ZipA	T	M		Cell division protein
rfaG1	T	M	pili	Glycosyl transferase group 1
NhhA	T	M	adhesion	outer membrane protein
NMAA_0797	T	M		Uncharacterized protein
FtsK 2	T	M		integral component of membrane DNA translocase
hscA	A	M		Chaperone protein HscA homolog
NMAA_0920	A	M		Putative cytochrome
NMAA_1126	T	M	membrane	Putative membrane protein
NM1626_01938	C	M	pili	glucosyl transferase
NM1626_00384	G	M	LPS	lipopolysaccharide-1,6-D-galactosyltransferase
Opc (promoter)	C	M	Membrane	Outer membrane protein
pntA_serB	G	M		NAD(P) transhydrogenase subunit alpha-and-SerB protein
intA_1	TG	M	phage associated	putative phage integrase
NMAA_0143	A	M		Uncharacterized protein
fetA	T	M	Iron binding	TonB-dependent enterobactin receptor FetA
NMA0643	GGCG	L	LPS	Lipopolysaccharide biosynthesis translocase
PglE	AAACAAC	L	pili	glycosyltransferase
ilvD	CG	L		Dihydroxy-acid dehydratase, cofactor Fe
NMAA_0927	AC	L		Putative HTH-type transcriptional regulator
thiL	AC	L		Thiamine-monophosphate kinase
autB	TGCT	L	membrane	outer membrane transporter
drg	TTCC	L	DNA methylation	Dam-replacing protein
NMAA_0690	AT	L		Uncharacterized protein
NMA_0299	CG	L		Putative ATP-binding protein
RibD	GA	L		Riboflavin biosynthesis protein
NM1626_01372	CAAG	L		hypothetical protein
vapA_2	GCTT	L	adhesion	Aida-related Type V secretory pathway adhesin
tonB	GT	L	Iron binding	tonB-dependent ferric siderophore receptor
hpuB	T	L	Iron binding	Hemoglobin-haptoglobin utilization protein B

4.4.5 A Predominant OFF state of type III methyltransferases.

Within a bacterial population, a straightforward analysis would be to determine the ON versus OFF frequency of a particular phase-variable gene. In the case of multiple phase-variable genes, each cell would have a combinatorial ON/OFF conformation of these genes denoting a cell's phasotype. On the other hand, the cumulative frequency of each phase-variable gene status within the population denotes the phasevariome.

A higher number of phase-variable genes would lead to an exponential increase in possible genotype combinations (2^n ; n denoting number of genes). In the case of genes having regulatory functions or antigenic determinants such a mechanism could generate a remarkable number of possible phasotypes thereby giving the pathogen a huge repertoire of rapidly accessible ON/OFF genotypes maximizing variation potentials. At the example of the phase-variable opacity genes, four allele copies are present in WUE 2594 (Table 4-4). With each allele having 2 expression states (ON or OFF) this would result in 16 possible phasotypes combinations significantly increasing potential genotype variability.

Two phase-variable type III methyltransferase genes are present in our reference strain WUE 2594, *modA12* and *modB2*. Using our pipeline we have confirmed a high phase-variation rate of both genes (Table 4-1; Table 4-4). For some isolates, our computational approach indicated longer repeat stretches but failed to resolve the exact repeat unit number due to read length limitations. For a complete overview of the population at these two *mod* genes, we complemented our data with PCR and Sanger sequencing to determine repeat unit number and subsequently the gene status (Table S4-1).

In order to identify isolate phasotype, we used the predominant repeat conformation as the most likely approximation for the (original) single cell status. The combinatorial ON/OFF status of both *mod* genes allow defining for each isolate a methyltransferase phasotype. The frequency of each possible phasotypes of the two *mod* genes in our 100 meningococcal isolates indicates that the majority of isolates have both methyltransferase genes switched off (58%). A mere 5% have both genes ON whereas 30% of the isolates have an alternate ON/OFF status of the two genes (Table 4-5). The three possible frames would suggest a 33% ON status versus 67% off status to be expected from a random fluctuation of repeat number. Instead, we observe a consistent reduced proportion of ON status for both *modA12* and *modB2* genes at 18% and 22% respectively (Table 4-5). Both phasotype and phasevariome of the *mod* genes indicate a predominant OFF status within the population which hints potential selective pressure to maintain an inactive status of these two genes.

		ModB2		Phase-variome
		ON	OFF	
ModA12	ON	5%	13%	18%
	OFF	17%	58%	75%
phasevariome		22%	71%	7% *

Table 4-5: Type III methyltransferases phasotypes (purple) and phasevariome (orange). (*) Total undetermined where pipeline and PCR failed.

4.4.6 Repeat region length evolution at *modA12* locus

The 100 sequenced *Neisseria meningitidis* strains represent epidemics isolates with a hypothetic common ancestor strain. In order to investigate the evolutionary dynamics of a phase-variable repeat region within the 100 sequenced strains, we used pairwise genetic distances based on genome wide SNP count. From the RepHMM inferred *modA12* repeat length, we calculated a pairwise repeat unit count difference. Based on pairwise genetic distance we observe two main clouds of data points (Figure 4-7) consistent with two main sequence types in the analyzed collection of strains. Within one cluster, strains show a high similarity (<40 SNPs) but are clearly separated between clusters. Pairwise difference of repeat unit count reveals a slight tendency for an increased repeat unit length with increased genetic distance however the two mutation types clearly appear to evolve independently (Figure 4-7). Repeat unit number seem to be highly dynamic with a fast variability rate a process which allows a phase-variable gene to be turned on and off rapidly.

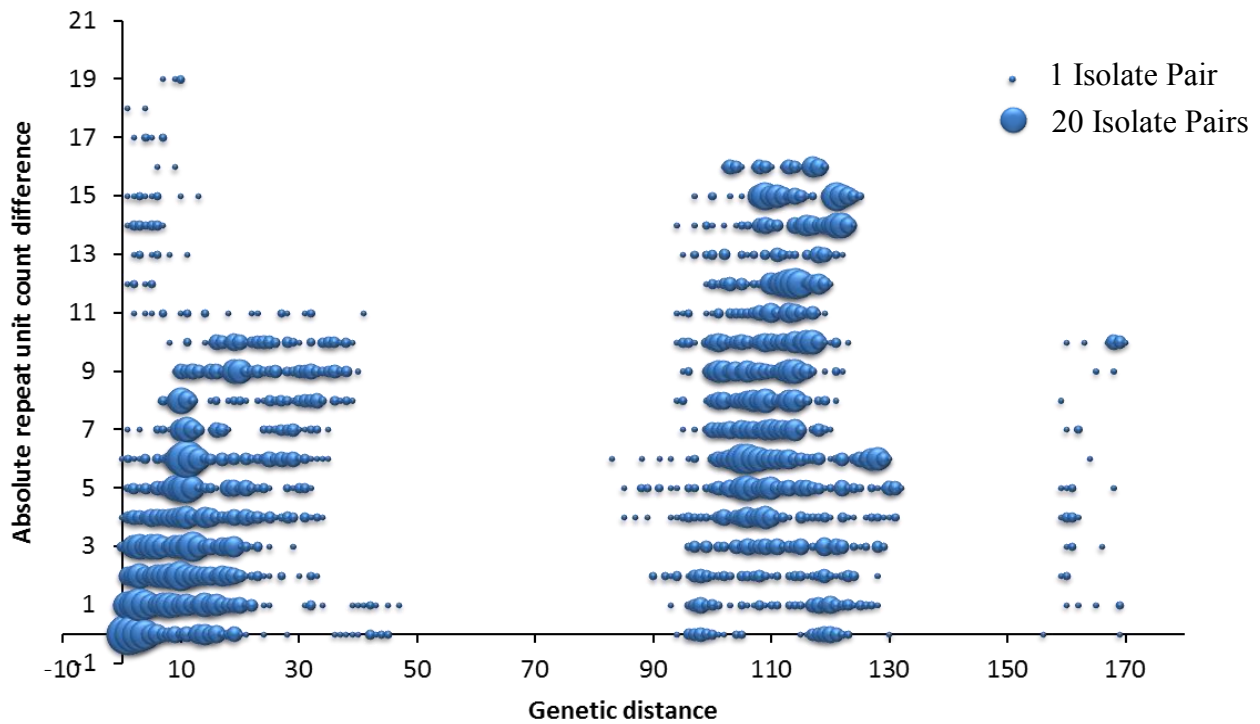


Figure 4-7: Evolution of repeat region length is independent of genetic distance. Pairwise genetic distance of 100 strains plotted against pairwise absolute difference in repeat unit count at the *modA12* repeat locus. Bubble size reflects aggregate number of dots.

4.5 Discussion

Summary

Complementing the fast and exact simple pattern matching algorithm (RegEx), we developed a probabilistic tool 'RepHMM' for an extended repeat length detection by tolerating mismatches resulting from sequencing errors or SNPs. RepHMM also provides an enhanced detection resolution enabling detection of rare repeat genotype variants signifying clonal variation. Our results demonstrate the utility of our methods to discover and assess repeat unit numbers at microsatellite loci directly from high-throughput sequencing data. In the era of big data where large scale whole genome sequencing is becoming routine, we provide a well-designed pipeline for a fast and reliable genome wide assessment of phase-variation events using raw sequencing data. Our direct approach circumvents the need of computationally heavy and time consuming classical approaches to detect phase-variation which are mostly locus specific and/or require genome assembly or PCR amplification. Applied to a large dataset of sequenced *Neisseria meningitidis* genomes our pipeline revealed a rapid ON/OFF switching mechanism that controls an extended set of genes. The identified phase-variable genes are mostly associated with surface antigens, capsule and adhesion elements, which are proposed to contribute to the pathogens adaptability and evasion of the host immune system.

We prove the ability of RepHMM to determine locus specific repeat unit counts from a mix of different sequence variants. Highly similar, recently duplicated genomic sequences such as the *opa* loci can however not be resolved by our method, nevertheless RepHMM is able to provide an aggregate count of the different repeat length conformations present in the genomes of the population of cells subjected to sequencing. The present results suggest that RepHMM can also be applied to monitor highly unstable tandem repeat loci in the (clonal) population of cells pooled for a sequencing assay.

Consistent mutation rate of phase-variable short tandem repeats

Our results are consistent with the reported rapid variability of phase-variable genes in *Neisseria* spp. Mutations at STRs can reach up to 10^{-3} per division (Bayliss, 2009). Divergent repeat conformations from a particular isolate likely represent clonal variants. Calculating the ratio of minor repeat genotype variants to the predominant repeat conformation using sequencing reads count as detected by RepHMM provides an estimate of the rate of phase-variation at a particular locus. At the example of *modA12* locus, our calculated clonal phase-

variation rate matches for most isolates the reported mutation rates. In few extreme cases however we do observe much higher mutation rates (Table 4-1, NM1325) which might be also attributed to contaminations or selection biases in cell culture.

Pipeline adaptable to other sequencing technologies

In addition to analyzing genome wide repeat regions, our pipeline can be adapted to target a specific genomic locus for a direct and detailed analysis using raw sequencing data. Furthermore, the tool can be used with virtually any sequencing technology. We have applied the pipeline also to circular consensus sequences (ccs) reads generated by PacBio single molecule real-time sequencing platform and we were successful in quantifying repeat unit length and detecting single molecule variability at phase-variable regions similar to what we observe in the Illumina data sets (Figure S4-1). Although read length is constantly increasing with advances in sequencing technologies thus covering longer repeat loci, yet rare repeat genotype variants are still ignored using conventional genome mapping and assembly tools.

Closely related bacterial population ideal for identifying phase-variable genome

We utilize a sequenced population of 100 meningococcal serogroup A isolates which were collected within the framework of longitudinal micro-epidemiological studies in Ghana and Burkina Faso between 1998 and 2011 (Leimkugel et al., 2007b; Sié et al., 2008, Lamelas et al. 2014). Although the genomes belong to two sequence types, ST7 strains are only single locus variants of ST2859 strains. While the MLST profile underestimates the overall genome-wide differences, the population is highly closely related with a total of only 361 SNPs occurring within the 100 isolates population (Lamelas et al. 2014). Such population is therefore an ideal set up to infer genome wide phase-variation events and ultimately a serogroup (A) meningococcal phase-variable genome. The assembly and mapping complications relating to repetitive sequences as described in the introduction prevents classical comparative genomics analysis as performed by (Lamelas et al.) to easily detect and compare phase-variation events. Therefore, a potential role of phase-variation in adaptation and pathogenesis would be severely underestimated or even neglected due to technical limitations.

Genes known to phase-vary and are present in our WUE 2594 reference genome are successfully detected by our approach. Complementing the pioneering work of (Parkhill et al.) and (Snyder et al.), we have validated some of the putative phase-variable genes which

also occur in our reference genome. In comparison, our data provide a more reliable and solid assessment of genome wide phase-variation events as we compare highly closely related population of isolates whereas (Snyder et al.) analysis compared only three distant strains of *Neisseria* spp., namely two *N. meningitidis* strains Z2491 and MC58 and *N. gonorrhoeae* strain FA1090. Such a limited number of compared genomes and their evolutionary distance could very likely confound phase-variation with simple mutagenesis and selection.

Phase-variable surface components and regulatory genes

Adding to the sheer number of genes (288) containing hyper-variable repeat sequences within our isolates population, another striking feature is the diversity of functions of the top ranking genes. The majority of our identified phase-variable genes can be associated with pathogenicity determinants and surface components in *Neisseria*. The 13 genes already known to phase-vary (blue cells Table 4-4) are also known as virulence factors in *Neisseria* (Loh et al., 2013). Phase-variation of enzymes involved in LPS biosynthesis would lead to antigenic variation typical of meningococci. In addition, adhesion associated genes including pili and opacity proteins are characterized virulence traits important for establishment of *N. meningitidis* on mucosal surfaces. The constant switching of surface components help maintain colonization as well as facilitate immune evasion by the bacterium (Virji, 2009).

Remarkably genes not associated with *Neisseria* surface structures but are nevertheless crucial pathogenicity determinants such as iron acquisition; regulatory mechanisms and DNA methylation are also phase-variable. Minor yet rapid changes in repeat regions of regulatory genes would have genome wide ramifications augmenting the consequences of phase-variation. The phasevarion is one such system whereby phase-variation of type III DNA methyltransferases (*mod* genes Table 4-4) has been associated with a coordinated switching of multiple genes in several species including *Neisseria* spp. and *Helicobacter pylori* (Gawthorne et al., 2012a; Srikhanta et al., 2010).

Phase-variation mediates a reduced expression status of genes

Phase-variation provides an additional layer of gene expression control. To our knowledge no biases in the extension/contraction mechanism of repeat unit number has been reported so far. Accordingly, the three reading frames would theoretically render phase-variable genes to have a reduced expression status (33% ON versus 67% OFF) compared to other non-

phase-variable genes. Analyzing phasotypes and phasevariome of type III methyltransferases, both *modA12* and *modB2* display an even lower ON percentage within our population than expected from arbitrary repeat unit number variation (Table 4-5), suggesting a pressure to keep these methyltransferase enzymes with potential regulatory functions (Srikhanta et al., 2010) at a low frequency expression state. It is worth noting that despite the observed predominant OFF status, mod gene family appear to be highly conserved in *Neisseria* spp. (Seib et al., 2011). A similar observation of reduced expression was reported for phase-variable outer membrane proteins. *NadA* and *porA*, which are components of the recently licensed Bexsero vaccine (Vogel et al., 2013) and *fetA* another vaccine candidate (Urwin et al., 2004) were shown to be down regulated as a consequence of phase-variation (Alamro et al., 2014). The reduced expression of cell surface components in *N. meningitidis* is suggested to facilitate host persistence by evasion of adaptive immune responses (Alamro et al., 2014). In contrast, the autotransporters *mspA* phase-variable gene from serogroup B strains displayed a predominant ON state (86%) suggesting an important role in certain epidemiological settings (Oldfield et al., 2013). Nevertheless, Phasotypes and phasevariome are indicative of the bacterial response but cannot be used to predict *in vivo* phenotype since external factors and post-transcriptional modifications could also control gene expression (Bidmos and Bayliss, 2014).

Acknowledgments

We would like to thank the Molecular Immunology and Gene Regulation groups at the Swiss TPH for the training and assistance with the bacterial cultures and PCRs.

4.6 Supplementary materials

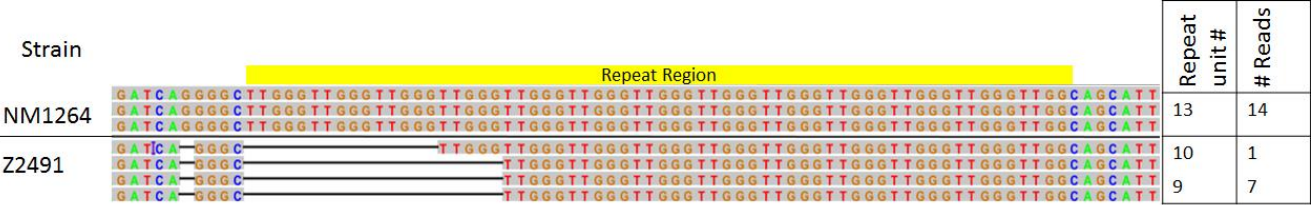


Figure S4-1: Pipeline applied to CCS reads from Pacbio SMRT sequencing detecting repeat unit variation at *modB2* locus.

Table S4-1: Complete set of 100 isolates *modA12* and *modB2* genes status including the predominant repeat length as determined by RepHMM (repeat number *modA12*<16; *modB2*<12) or by PCR/Sanger sequencing (repeat number >16 and >12). (*) Accession number at the Sanger data resource

http://www.sanger.ac.uk/resources/downloads/bacteria/neisseria.html#project_1893

Isolate	Accession Number*	ModA12 Repeat units #	ORF	ModB2 Repeat units #	ORF
1264	ERS040961	4	On	12	Off
1325	ERS040962	10	On	7	Off
1359	ERS040963	5	Off	14	On
1360	ERS040964	6	Off	12	Off
1361	ERS040965	5	Off	14	On
1362	ERS040966	4	On	14	On
1363	ERS040967	5	Off	14	On
1364	ERS040968	5	Off	14	On
1446	ERS040969	6	Off	12	Off
1471	ERS040970	5	Off	12	Off
1482	ERS040971	5	Off	14	On
1483	ERS040972	5	Off	13	Off
1544	ERS040973	8	Off	15	Off
1549	ERS040974	8	Off	15	Off
1550	ERS040975	8	Off	16	Off
1561	ERS040976	6	Off	15	Off
1573	ERS040977	4	On	9	Off
1578	ERS040978	6	Off	15	Off
1666	ERS040979	14	Off	7	OFF
1672	ERS040980	15	Off	14	On
1673	ERS040981	14	Off	14	On
1757	ERS040982	7	On	13	Off
1758	ERS040983	8	Off	14	On
1779	ERS040984	8	Off	10	Off
1797	ERS040985	9	Off	10	Off
1805	ERS040986	8	Off	10	Off
1826	ERS040987	8	Off	10	Off
1829	ERS040988	8	Off	10	Off
1831	ERS040989	5	Off	8	On
1837	ERS040990	5	Off	18	Off
1845	ERS040991	10	On	14	On
1891	ERS040992	10	On	14	On
1892	ERS040993	6	Off	15	Off
1893	ERS040994	6	Off	15	Off

Isolate	Accession Number*	ModA 12 Repeat units #	ORF	ModB2 Repeat units #	ORF
2181	ERS041011	15	Off	9	Off
2187	ERS041012	13	ON	9	Off
2188	ERS041013	14	Off	9	Off
2193	ERS041014	14	Off	9	Off
2206	ERS041015	13	On	10	Off
2228	ERS041016	15	Off	9	Off
2232	ERS041017	13	On	10	Off
2237	ERS041018	12	Off	9	Off
2239	ERS041019	14	Off	9	Off
2244	ERS041020	12	Off	10	Off
2254	ERS041021	15	Off	9	Off
2261	ERS041022	14	Off	10	Off
2263	ERS041023	11	Off	9	Off
2264	ERS041024	18	Off	9	Off
2332	ERS041025	14	Off	9	Off
2335	ERS041026	10	ON	10	Off
2369	ERS041027	11	Off	11	On
2381	ERS041028	13	On	11	On
2382	ERS041029	17	Off	10	Off
2389	ERS041030	11	Off	9	Off
2393	ERS041031	14	Off	9	Off
2394	ERS041032	15	Off	10	Off
2431	ERS041033	19	On	11	On
2432	ERS041034	17	Off	10	Off
2433	ERS041035	14	Off	8	On
2439	ERS041036	14	Off	9	Off
2441	ERS041037	17	Off	8	On
2524	ERS041038			10	Off
2602	ERS041039	12	Off	9	Off
2606	ERS041040			10	Off
2617	ERS041041	16	On	9	Off
2700	ERS041042	20	Off	9	Off
2701	ERS041043	20	Off	9	Off
2717	ERS041044	10	ON	10	Off

1895	ERS040995	5	Off	12	Off
1901	ERS040996	6	Off	21	Off
1910	ERS040997	10	ON	-	-
1919	ERS040998	6	Off	-	-
1921	ERS040999	8	Off	10	Off
1928	ERS041000	4	On	12	Off
1931	ERS041001	7	Off	10	Off
1937	ERS041002	10	ON	-	-
1938	ERS041003	5	Off	9	Off
1963	ERS041004	4	ON	12	Off
1976	ERS041005	5	Off	19	Off
2008	ERS041006	14	Off	17	On
2009	ERS041007	14	Off	15	Off
2025	ERS041008	14	Off	15	Off
2032	ERS041009	15	Off	-	-
2033	ERS041010	14	Off	-	-

2718	ERS041045	15	Off	9	Off
2808	ERS041046	20	Off	9	Off
2809	ERS041047	19	On	9	Off
2810	ERS041048	20	Off	9	Off
2811	ERS041049	20	Off	9	Off
2812	ERS041050	20	Off	9	Off
2813	ERS041051	20	Off	9	Off
2814	ERS041052	20	Off	9	Off
2856	ERS041053	12	Off	10	Off
2857	ERS041054	12	Off	10	Off
2932	ERS041055	20	Off	9	Off
2933	ERS041056	20	Off	8	On
2934	ERS041057	20	Off	9	Off
2935	ERS041058	20	Off	8	On
3128	ERS041059	20	Off	8	On
3129	ERS041060	20	Off	8	On

5 General Discussion

5.1 Summary

Meningococcal disease occurs worldwide as endemic infections. The gram-negative bacteria *Neisseria meningitidis* is a predominant cause of septicemia and meningitis, at the same time it is also a frequent commensal of the human nasopharynx. Meningococci displays a high genetic diversity, yet to present the results from multilocus sequence typing (MLST) and whole genome sequencing have not defined a strict core patho-genome enabling the prediction of virulence on a genetic basis.

Towards a better understanding of meningococcal complexity and genomic plasticity, we address in this thesis the epigenetic diversity and the consequences of DNA methylation as well as developing tools to assess phase-variation events within a population of *Neisseria meningitidis* isolates.

We applied single-molecule real-time (SMRT) sequencing technology to establish genome wide DNA modification profiles of two closely related *N. meningitidis* strains. DNA modifications, and in particular DNA methylation as the most common DNA modification, are thereby detected based on delay in the kinetics of the DNA synthesis *in vitro*. Our approach revealed a high diversity in DNA methylation between closely related *Neisseria* strains. The methylated sequences as defined by the SMRT sequencing results largely correspond to putative target sequence of DNA methyltransferases identified in *N. meningitidis* and to the observed protection from digestion by methylation-sensitive cognate restriction enzymes.

Association of epigenetic modifications with phenotypes is still poorly characterized in bacteria. Our analysis of the methylation patterns along the genome showed a biased distribution evident by a clear depletion of 5-methylcytosine motifs relative to gene start positions. These results suggest a more complex role for DNA methylation in terms of the hypothesized regulatory functions. In addition, we identified a striking co-localization of mutations at methylated bases indicating additional un-known consequences of DNA methylation in relation to bacterial evolution and adaptation.

We also developed a novel bioinformatic tool to infer the precise number of repeat units at specific tandem repeat loci exploiting increasing read length resulting from recent versions of large scale sequencing assays. Our probabilistic approach detects divergent repeat length configurations and therefore functional states of ORFs directly from raw sequencing data, offering an enhanced detection power and accuracy over conventional tools. We integrated our tool into a fast and efficient computational pipeline to detect genome wide phase-

variation events by comparing a large number of sequenced meningococcal genomes. Our comprehensive approach identified a high number of hyper-variable repeat regions mostly associated with genes encoding outer membrane components and other virulence determinants.

5.2 DNA methylation interplay with other adaptation mechanisms

Several mechanisms contributing to *Neisseria meningitidis* genomic plasticity and adaptation have been identified over the years. These mechanisms include point mutations, homologous recombination and horizontal gene transfer as well as phase-variation (Jeltsch, 2003; Jolley et al., 2005; Casadesús and Low, 2013b; Lamelas et al., 2014). Taken individually each mechanism provides the pathogen with distinctive variability and adaptation tools. An interplay between these mechanisms augments the pathogen complexity and adaptability (Virji, 2009).

Our results integrate DNA methylation into an established system of evolution and adaptation (Figure 5-1). We provide evidence of methylated nucleotides being more prone to mutation within a selective environment. Conversely, mutations lead to a localized loss of methylation by altering the methylation target sequence and rendering it unrecognizable by the methyltransferase. Mutations occurring in the DNA recognition domain of a methyltransferase gene could in return modify the methylation target motif and thereby altering the methylome of the bacterium (Gawthorne et al., 2012b).

In the same line, our results are consistent with previous reports linking methyltransferase enzymes with phase-variation (Seib et al., 2011; Srikhanta et al., 2010). We demonstrate phase-variation to be the major contributor to the observed variability in DNA methylation between meningococcal isolates. The ON/OFF switching of a gene primarily affects its own function and possibly in some cases other downstream functions. In the case of genes encoding a protein with a regulatory function, as proposed for some DNA methyltransferases (phasevarion), the switching mechanism has genome wide consequences involving a number of biological processes (Srikhanta et al., 2010).

DNA methylation has been proposed to have a selective role in horizontal gene transfer (Budroni et al., 2011). Our data however did not support this hypothesis perhaps for reasons related to size and properties of our studied bacterial population. While this hypothesis remains debatable (Bart et al., 2001; Kong et al., 2013), a potential role of restriction

modification systems in defining compatible fragments for horizontal transfer requires additional investigation.

On the other hand, horizontal gene transfer has been proven to play a role in methyltransferase plasticity. In *Helicobacter pylori*, methyltransferases DNA recognition sequences have been described as recombination hotspots (Lin et al., 2009). In addition, the high number and diversity of methyltransferase genes in *Neisseria* spp. has been attributed to horizontal gene transfer (Budroni et al., 2011). Inter repeat recombination is also proposed as one of the mechanisms causing extension and contraction of tandem repeats regions and consequently leading to phase-variation events (Gemayel et al., 2010).

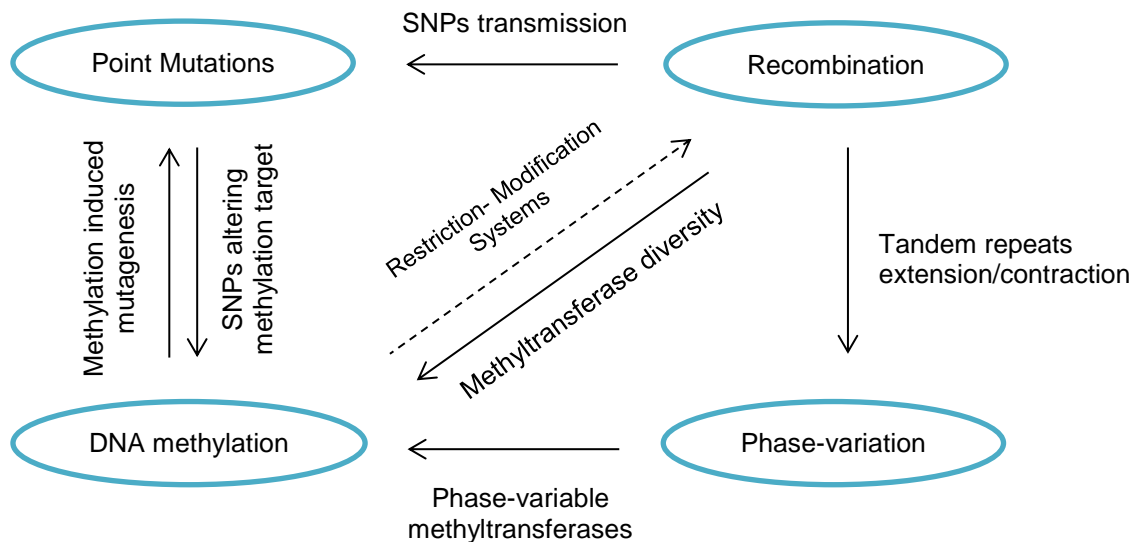


Figure 5-1: Integrating DNA methylation into an established concept of mutagenesis and selection to obtain genomes encoding for better adapted phenotypes.

5.3 Relevance to public health

Advances in bioinformatics, sequencing technologies and microbiology are expanding the scope of epidemiology to include not only the disease but also the pathogens biology. Genomics tools can recognise single nucleotide variations across sequenced genomes which allow exploring diversity among microbial populations with high resolution, replacing methods that depended on phenotype or specific genetic markers. In addition to basic scientific understanding of the bacterium, these advances have provided insights into public health issues such as mechanisms of drug resistance in bacteria, epidemiological linking of infected individuals and transmission dynamics in several pathogens including

Mycobacterium tuberculosis (Coscolla et al., 2015; Stucki et al., 2014), *Staphylococcus aureus* (Harris et al., 2013) and *Neisseria spp.* (Grad et al., 2014).

The field of prokaryotic epigenetics is still in its early stages. Our effort at this stage was primarily focused on the basic implications of DNA modifications in bacteria. Nevertheless, our successful characterisation of dynamic and variable DNA modification patterns between *meningococcal* isolates, suggest an “epi-genotype” could potentially complement genetic basis for phenotype diversity, which in return could be investigated in an epidemiological context of the host population.

Short tandem repeats are popular markers used in epidemiological studies. In the era of large scale sequencing experiments an increasing number of whole genome sequences of bacterial isolates are becoming routine. However, there exists a large gap between the sheer amount of generated data and the lack of analysis capacity. We provide an efficient bioinformatic pipeline to analyze and compare tandem repeat regions, adaptable to any organism and sequencing experiment.

High resolution computational tools similar to what we present in this thesis contribute to the advancement of pathogen genome analysis methods which can uncover otherwise unknown transmission dynamics and give insight into organism evolution and the genetic basis for phenotypes.

5.4 Considerations for future sequencing approaches

Despite its outstanding advantages, Pacific Biosciences' SMRT sequencing so far has a complicated and not fully transparent analysis pipeline, especially for detecting DNA modifications. For instance, the modification score calculation remains a black box. The technology suffers from short comings due to a low detection power for 5-methylcytosines, which is the most common DNA modification. The Tet1 enzymatic conversion used to enhance the 5mC signal does not have a 100% yield, therefore only a methylation target motif can be reliably inferred but individual genomic positions cannot be consistently detected. Thus, advanced exploration targeting specific genomic regions and/or differential methylation analysis could only be conducted for 6-methyladenosines and 4-methylcytosines. Both SMRT hardware and analysis pipeline is under continuous development and hopefully future improvements could also reduce the cost of this third generation sequencing technology, which would open new horizons in epigenetic research (Roberts et al., 2013).

Having said that, it would be advantageous to use SMRT sequencing especially for small bacterial genomes. The long read length would allow to have closed full genomes in addition to the bacterial methylome. The usefulness of having a full genome with its corresponding epigenome could prove to be highly valuable for our understanding of bacterial biology.

We compared the epigenetic potentials of a population of strains based on *in silico* prediction of gene status compared to only two actually SMRT sequenced genomes. Such predictions does not ensure *in vivo* activity of the corresponding methyltransferase. Lower sequencing costs would enable a more thorough assessment of epigenetic phenotype of bacterial population.

Sequencing technologies maintaining accuracy but requiring a less amount of genomic DNA would enable studying isolate samples directly from patients. The current need for secondary laboratory culturing steps is likely to confound original isolate status by propagating generations in the new culture environment. In addition, Current culture based methods assume a single colony selection accurately represents the whole plated population. However, our data shows phase-variation to be caused by dynamic and rapid genetic changes even at the clonal level. Single colony based culture and sequencing could therefore underestimate actual within host diversity of bacterial isolates.

6 References

- Achtman, M. (2004). Population structure of pathogenic bacteria revisited. *Int. J. Med. Microbiol. IJMM* 294, 67–73.
- Alamro, M., Bidmos, F.A., Chan, H., Oldfield, N.J., Newton, E., Bai, X., Aidley, J., Care, R., Mattick, C., Turner, D.P.J., et al. (2014). Phase Variation Mediates Reductions in Expression of Surface Proteins during Persistent Meningococcal Carriage. *Infect. Immun.* 82, 2472–2484.
- Andrews, S.M., and Pollard, A.J. (2014). A vaccine against serogroup B *Neisseria meningitidis*: dealing with uncertainty. *Lancet Infect. Dis.* 14, 426–434.
- Arber, W., and Linn, S. (1969). DNA modification and restriction. *Annu. Rev. Biochem.* 38, 467–500.
- Avery, S.V. (2006). Microbial cell individuality and the underlying sources of heterogeneity. *Nat. Rev. Microbiol.* 4, 577–587.
- Balaban, N.Q., Merrin, J., Chait, R., Kowalik, L., and Leibler, S. (2004). Bacterial persistence as a phenotypic switch. *Science* 305, 1622–1625.
- Bart, A., Pannekoek, Y., Dankert, J., and van der Ende, A. (2001). NmeSI restriction-modification system identified by representational difference analysis of a hypervirulent *Neisseria meningitidis* strain. *Infect. Immun.* 69, 1816–1820.
- Bautsch, W. (1998). Comparison of the genome organization of pathogenic neisseriae. *Electrophoresis* 19, 577–581.
- Bayliss, C.D. (2009). Determinants of phase variation rate and the fitness implications of differing rates for bacterial pathogens and commensals. *FEMS Microbiol. Rev.* 33, 504–520.
- Bentley, S.D., Vernikos, G.S., Snyder, L.A., Churcher, C., Arrowsmith, C., Chillingworth, T., Cronin, A., Davis, P.H., Holroyd, N.E., Jags, K., et al. (2007). Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18. *PLoS Genet.* 3, e23.
- Bestor, T.H. (1990). DNA methylation: evolution of a bacterial immune function into a regulator of gene expression and genome structure in higher eukaryotes. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 326, 179–187.
- Bidmos, F., and Bayliss, C. (2014). Genomic and Global Approaches to Unravelling How Hypermutable Sequences Influence Bacterial Pathogenesis. *Pathogens* 3, 164–184.
- Bird, A. (2007). Perceptions of epigenetics. *Nature* 447, 396–398.
- Bowler, L.D., Zhang, Q.Y., Riou, J.Y., and Spratt, B.G. (1994). Interspecies recombination between the *penA* genes of *Neisseria meningitidis* and commensal *Neisseria* species during the emergence of penicillin resistance in *N. meningitidis*: natural events and laboratory simulation. *J. Bacteriol.* 176, 333–337.
- Boyd, E.F., Li, J., Ochman, H., and Selander, R.K. (1997). Comparative genetics of the *inv-spa* invasion gene complex of *Salmonella enterica*. *J. Bacteriol.* 179, 1985–1991.
- Budroni, S., Siena, E., Hotopp, J.C.D., Seib, K.L., Serruto, D., Nofroni, C., Comanducci, M., Riley, D.R., Daugherty, S.C., Angiuoli, S.V., et al. (2011). *Neisseria meningitidis* is structured in clades associated

with restriction modification systems that modulate homologous recombination. *Proc. Natl. Acad. Sci. U. S. A.* 108, 4494–4499.

Camacho, E.M., and Casadesús, J. (2005). Regulation of *traJ* transcription in the Salmonella virulence plasmid by strand-specific DNA adenine hemimethylation: Regulation of *traJ* transcription by Dam methylation. *Mol. Microbiol.* 57, 1700–1718.

Casadesús, J., and Low, D. (2006). Epigenetic gene regulation in the bacterial world. *Microbiol. Mol. Biol. Rev. MMBR* 70, 830–856.

Casadesús, J., and Low, D.A. (2013a). Programmed heterogeneity: epigenetic mechanisms in bacteria. *J. Biol. Chem.* 288, 13929–13935.

Casadesús, J., and Low, D.A. (2013b). Programmed heterogeneity: epigenetic mechanisms in bacteria. *J. Biol. Chem.* 288, 13929–13935.

Caugant, D.A., and Maiden, M.C.J. (2009). Meningococcal carriage and disease--population biology and evolution. *Vaccine* 27 Suppl 2, B64–B70.

Caugant, D.A., Kristiansen, P.A., Wang, X., Mayer, L.W., Taha, M.-K., Ouédraogo, R., Kandolo, D., Bougoudogo, F., Sow, S., and Bonte, L. (2012). Molecular characterization of invasive meningococcal isolates from countries in the African meningitis belt before introduction of a serogroup A conjugate vaccine. *PloS One* 7, e46019.

Chai, Y., Chu, F., Kolter, R., and Losick, R. (2007). Bistability and biofilm formation in *Bacillus subtilis*: Bistability and biofilm formation in *Bacillus subtilis*. *Mol. Microbiol.* 67, 254–263.

Christensen, H., May, M., Bowen, L., Hickman, M., and Trotter, C.L. (2010). Meningococcal carriage by age: a systematic review and meta-analysis. *Lancet Infect. Dis.* 10, 853–861.

Clamp, M., Cuff, J., Searle, S.M., and Barton, G.J. (2004). The Jalview Java alignment editor. *Bioinforma. Oxf. Engl.* 20, 426–427.

Clark, T.A., Lu, X., Luong, K., Dai, Q., Boitano, M., Turner, S.W., He, C., and Korlach, J. (2013). Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation. *BMC Biol.* 11, 4.

Coscolla, M., Barry, P.M., Oeltmann, J.E., Koshinsky, H., Shaw, T., Cilnis, M., Posey, J., Rose, J., Weber, T., Fofanov, V.Y., et al. (2015). Genomic Epidemiology of Multidrug-Resistant *Mycobacterium tuberculosis* During Transcontinental Spread. *J. Infect. Dis.*

Croucher, N.J., Harris, S.R., Fraser, C., Quail, M.A., Burton, J., van der Linden, M., McGee, L., von Gottberg, A., Song, J.H., Ko, K.S., et al. (2011). Rapid Pneumococcal Evolution in Response to Clinical Interventions. *Science* 331, 430–434.

Danielson, L., and Mann, E. (1806). A history of a singular and very noted disease, which lately made its appearance in Medfield. *Med. Agric. Regist.* 1, 65–69.

Daugla, D., Gami, J., Gamougam, K., Naibei, N., Mbainadji, L., Narbé, M., Toralta, J., Kodbesse, B., Ngadoua, C., Coldiron, M., et al. (2014). Effect of a serogroup A meningococcal conjugate vaccine (PsA-TT) on serogroup A meningococcal meningitis and carriage in Chad: a community study. *The Lancet* 383, 40–47.

Davidson, C.J., and Surette, M.G. (2008). Individuality in Bacteria. *Annu. Rev. Genet.* 42, 253–268.

- Doerfler, W. (1981). DNA methylation--a regulatory signal in eukaryotic gene expression. *J. Gen. Virol.* *57*, 1–20.
- Durbin, R., Eddy, S.R., Krogh, A., and Mitchison, G. (1999). *Biological sequence analysis: probabilistic models of proteins and nucleic acids.* (London: Cambridge University Press).
- Fang, G., Munera, D., Friedman, D.I., Mandlik, A., Chao, M.C., Banerjee, O., Feng, Z., Losic, B., Mahajan, M.C., Jabado, O.J., et al. (2012). Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. *Nat. Biotechnol.* *30*, 1232–1239.
- Flusberg, B.A., Webster, D.R., Lee, J.H., Travers, K.J., Olivares, E.C., Clark, T.A., Korlach, J., and Turner, S.W. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* *7*, 461–465.
- Furuta, Y., Namba-Fukuyo, H., Shibata, T.F., Nishiyama, T., Shigenobu, S., Suzuki, Y., Sugano, S., Hasebe, M., and Kobayashi, I. (2014). Methylome Diversification through Changes in DNA Methyltransferase Sequence Specificity. *PLoS Genet.* *10*, e1004272.
- Gagneux, S., Hodgson, A., Ehrhard, I., Morelli, G., Genton, B., Smith, T., Tanner, M., Binka, F., Achtman, M., and Pluschke, G. (2000). Microheterogeneity of serogroup A (subgroup III) *Neisseria meningitidis* during an outbreak in northern Ghana. *Trop. Med. Int. Health* *5*, 280–287.
- Gawthorne, J.A., Beatson, S.A., Srikhanta, Y.N., Fox, K.L., and Jennings, M.P. (2012a). Origin of the diversity in DNA recognition domains in phasevarion associated *modA* genes of pathogenic *Neisseria* and *Haemophilus influenzae*. *PloS One* *7*, e32337.
- Gawthorne, J.A., Beatson, S.A., Srikhanta, Y.N., Fox, K.L., and Jennings, M.P. (2012b). Origin of the diversity in DNA recognition domains in phasevarion associated *modA* genes of pathogenic *Neisseria* and *Haemophilus influenzae*. *PloS One* *7*, e32337.
- Gemayel, R., Vences, M.D., Legendre, M., and Verstrepen, K.J. (2010). Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.* *44*, 445–477.
- Goldberg, A., Fridman, O., Ronin, I., and Balaban, N.Q. (2014). Systematic identification and quantification of phase variation in commensal and pathogenic *Escherichia coli*. *Genome Med.* *6*, 112.
- Grad, Y.H., Kirkcaldy, R.D., Trees, D., Dordel, J., Harris, S.R., Goldstein, E., Weinstock, H., Parkhill, J., Hanage, W.P., Bentley, S., et al. (2014). Genomic epidemiology of *Neisseria gonorrhoeae* with reduced susceptibility to cefixime in the USA: a retrospective observational study. *Lancet Infect. Dis.* *14*, 220–226.
- Greenwood, B. (1999). Manson Lecture. Meningococcal meningitis in Africa. *Trans. R. Soc. Trop. Med. Hyg.* *93*, 341–353.
- Greenwood, B. (2006). Editorial: 100 years of epidemic meningitis in West Africa - has anything changed? *Trop. Med. Int. Health* *11*, 773–780.
- Güell, M., Yus, E., Lluch-Senar, M., and Serrano, L. (2011). Bacterial transcriptomics: what is beyond the RNA hori-z-ome? *Nat. Rev. Microbiol.* *9*, 658–669.

- Guichoux, E., Lagache, L., Wagner, S., Chaumeil, P., Léger, P., Lepais, O., Lepoittevin, C., Malausa, T., Revardel, E., Salin, F., et al. (2011). Current trends in microsatellite genotyping. *Mol. Ecol. Resour.* *11*, 591–611.
- Van Ham, S.M., van Alphen, L., Mooi, F.R., and van Putten, J.P. (1993). Phase variation of *H. influenzae* fimbriae: transcriptional control of two divergent genes through a variable combined promoter region. *Cell* *73*, 1187–1196.
- Hao, W., Ma, J.H., Warren, K., Tsang, R.S.W., Low, D.E., Jamieson, F.B., and Alexander, D.C. (2011). Extensive Genomic Variation within Clonal Complexes of *Neisseria meningitidis*. *Genome Biol. Evol.* *3*, 1406–1418.
- Harris, S.R., Cartwright, E.J.P., Török, M.E., Holden, M.T.G., Brown, N.M., Ogilvy-Stuart, A.L., Ellington, M.J., Quail, M.A., Bentley, S.D., Parkhill, J., et al. (2013). Whole-genome sequencing for analysis of an outbreak of methicillin-resistant *Staphylococcus aureus*: a descriptive study. *Lancet Infect. Dis.* *13*, 130–136.
- Harrison, J.J., Anisowicz, A., Gadi, I.K., Raffeld, M., and Sager, R. (1983). Azacytidine-induced tumorigenesis of CHEF/18 cells: correlated DNA methylation and chromosome changes. *Proc. Natl. Acad. Sci. U. S. A.* *80*, 6606–6610.
- Harrison, L.H., Trotter, C.L., and Ramsay, M.E. (2009). Global epidemiology of meningococcal disease. *Vaccine* *27 Suppl 2*, B51–B63.
- Heard, E., and Martienssen, R.A. (2014). Transgenerational Epigenetic Inheritance: Myths and Mechanisms. *Cell* *157*, 95–109.
- Holliday, R. (2006). Epigenetics: a historical overview. *Epigenetics Off. J. DNA Methylation Soc.* *1*, 76–80.
- Holmes, E.C., Urwin, R., and Maiden, M.C. (1999). The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis*. *Mol. Biol. Evol.* *16*, 741–749.
- Huber, C.A., Pflüger, V., Hamid, A.-W.M., Forgor, A.A., Hodgson, A., Sié, A., Junghanss, T., and Pluschke, G. (2012). Lack of antigenic diversification of major outer membrane proteins during clonal waves of *Neisseria meningitidis* serogroup A colonization and disease. *FEMS Immunol. Med. Microbiol.*
- Ishikawa, K., Fukuda, E., and Kobayashi, I. (2010). Conflicts targeting epigenetic systems and their resolution by cell death: novel concepts for methyl-specific and other restriction systems. *DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes* *17*, 325–342.
- Jeltsch, A. (2003). Maintenance of species identity and controlling speciation of bacteria: a new function for restriction/modification systems? *Gene* *317*, 13–16.
- Jennings, M.P., Srikhanta, Y.N., Moxon, E.R., Kramer, M., Poolman, J.T., Kuipers, B., and van der Ley, P. (1999). The genetic basis of the phase variation repertoire of lipopolysaccharide immunotypes in *Neisseria meningitidis*. *Microbiol. Read. Engl.* *145 (Pt 11)*, 3013–3021.
- Jolley, K.A., Wilson, D.J., Kriz, P., McVean, G., and Maiden, M.C.J. (2005). The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in *Neisseria meningitidis*. *Mol. Biol. Evol.* *22*, 562–569.
- Jones, P.A. (1985). Altering gene expression with 5-azacytidine. *Cell* *40*, 485–486.

- Joseph, B., Schwarz, R.F., Linke, B., Blom, J., Becker, A., Claus, H., Goesmann, A., Frosch, M., Müller, T., Vogel, U., et al. (2011). Virulence Evolution of the Human Pathogen *Neisseria meningitidis* by Recombination in the Core and Accessory Genome. *PLoS ONE* 6, e18441.
- Kim, T.-M., Laird, P.W., and Park, P.J. (2013). The landscape of microsatellite instability in colorectal and endometrial cancer genomes. *Cell* 155, 858–868.
- Koboldt, D.C., Chen, K., Wylie, T., Larson, D.E., McLellan, M.D., Mardis, E.R., Weinstock, G.M., Wilson, R.K., and Ding, L. (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 25, 2283–2285.
- Kong, Y., Ma, J.H., Warren, K., Tsang, R.S.W., Low, D.E., Jamieson, F.B., Alexander, D.C., and Hao, W. (2013). Homologous Recombination Drives both Sequence Diversity and Gene Content Variation in *Neisseria meningitidis*. *Genome Biol. Evol.*
- Lamelas, A., Harris, S.R., Roltgen, K., Dangy, J.-P., Hauser, J., Kingsley, R.A., Connor, T.R., Sie, A., Hodgson, A., Dougan, G., et al. (2014). Emergence of a New Epidemic *Neisseria meningitidis* Serogroup A Clone in the African Meningitis Belt: High-Resolution Picture of Genomic Changes That Mediate Immune Evasion. *mBio* 5, e01974–14 – e01974–14.
- Language: Disputed definitions. (2008). *Nature* 455, 1023–1028.
- Lapeyssonnie, L. (1963). [CEREBROSPINAL MENINGITIS IN AFRICA]. *Bull. World Health Organ.* 28 *Suppl*, 1–114.
- Leimkugel, J., Hodgson, A., Forgor, A.A., Pflüger, V., Dangy, J.-P., Smith, T., Achtman, M., Gagneux, S., and Pluschke, G. (2007). Clonal waves of *Neisseria* colonisation and disease in the African meningitis belt: eight- year longitudinal study in northern Ghana. *PLoS Med.* 4, e101.
- Lewis, L.A., Gipson, M., Hartman, K., Ownbey, T., Vaughn, J., and Dyer, D.W. (1999). Phase variation of HpuAB and HmbR, two distinct haemoglobin receptors of *Neisseria meningitidis* DNM2. *Mol. Microbiol.* 32, 977–989.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* 25, 2078–2079.
- Lin, E.A., Zhang, X.-S., Levine, S.M., Gill, S.R., Falush, D., and Blaser, M.J. (2009). Natural transformation of *helicobacter pylori* involves the integration of short DNA fragments interrupted by gaps of variable size. *PLoS Pathog.* 5, e1000337.
- Lluch-Senar, M., Luong, K., Lloréns-Rico, V., Delgado, J., Fang, G., Spittle, K., Clark, T.A., Schadt, E., Turner, S.W., Korch, J., et al. (2013). Comprehensive methylome characterization of *Mycoplasma genitalium* and *Mycoplasma pneumoniae* at single-base resolution. *PLoS Genet.* 9, e1003191.
- Loh, E., Kugelberg, E., Tracy, A., Zhang, Q., Gollan, B., Ewles, H., Chalmers, R., Pelicic, V., and Tang, C.M. (2013). Temperature triggers immune evasion by *Neisseria meningitidis*. *Nature* 502, 237–240.
- Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., et al. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U. S. A.* 95, 3140–3145.

Manso, A.S., Chai, M.H., Atack, J.M., Furi, L., De Ste Croix, M., Haigh, R., Trappetti, C., Ogunniyi, A.D., Shewell, L.K., Boitano, M., et al. (2014). A random six-phase switch regulates pneumococcal virulence via global epigenetic changes. *Nat. Commun.* 5, 5055.

Marri, P.R., Paniscus, M., Weyand, N.J., Rendón, M.A., Calton, C.M., Hernández, D.R., Higashi, D.L., Sodergren, E., Weinstock, G.M., Rounsley, S.D., et al. (2010). Genome sequencing reveals widespread virulence gene exchange among human *Neisseria* species. *PLoS One* 5, e11835.

Martin, P., Makepeace, K., Hill, S.A., Hood, D.W., and Moxon, E.R. (2005). Microsatellite instability regulates transcription factor binding and gene expression. *Proc. Natl. Acad. Sci. U. S. A.* 102, 3800–3804.

Mayer, C., Leese, F., and Tollrian, R. (2010). Genome-wide analysis of tandem repeats in *Daphnia pulex* - a comparative approach. *BMC Genomics* 11, 277.

Molesworth, A.M., Cuevas, L.E., Connor, S.J., Morse, A.P., and Thomson, M.C. (2003). Environmental Risk and Meningitis Epidemics in Africa. *Emerg. Infect. Dis.* 9, 1287–1293.

Oldfield, N.J., Matar, S., Bidmos, F.A., Alamro, M., Neal, K.R., Turner, D.P.J., Bayliss, C.D., and Ala'aldein, D.A.A. (2013). Prevalence and phase variable expression status of two autotransporters, NalP and MspA, in carriage and disease isolates of *Neisseria meningitidis*. *PLoS One* 8, e69746.

Palmer, B.R., and Marinus, M.G. (1994). The dam and dcm strains of *Escherichia coli*--a review. *Gene* 143, 1–12.

Pâques, F., Leung, W.Y., and Haber, J.E. (1998). Expansions and contractions in a tandem repeat induced by double-strand break repair. *Mol. Cell. Biol.* 18, 2045–2054.

Parkhill, J., Achtman, M., James, K.D., Bentley, S.D., Churcher, C., Klee, S.R., Morelli, G., Basham, D., Brown, D., Chillingworth, T., et al. (2000). Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature* 404, 502–506.

Perrin, A., Bonacorsi, S., Carbonnelle, E., Talibi, D., Dessen, P., Nassif, X., and Tinsley, C. (2002). Comparative genomics identifies the genetic islands that distinguish *Neisseria meningitidis*, the agent of cerebrospinal meningitis, from other *Neisseria* species. *Infect. Immun.* 70, 7063–7072.

Pizza, M., and Rappuoli, R. (2015). *Neisseria meningitidis*: pathogenesis and immunity. *Curr. Opin. Microbiol.* 23, 68–72.

Pomraning, K.R., Smith, K.M., and Freitag, M. (2009). Genome-wide high throughput analysis of DNA methylation in eukaryotes. *Methods San Diego Calif* 47, 142–150.

Poolman, J.T., Kriz-Kuzemenska, P., Ashton, F., Bibb, W., Dankert, J., Demina, A., Frøholm, L.O., Hassan-King, M., Jones, D.M., and Lind, I. (1995). Serotypes and subtypes of *Neisseria meningitidis*: results of an international study comparing sensitivities and specificities of monoclonal antibodies. *Clin. Diagn. Lab. Immunol.* 2, 69–72.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinforma. Oxf. Engl.* 26, 841–842.

Rando, O.J., and Verstrepen, K.J. (2007). Timescales of genetic and epigenetic inheritance. *Cell* 128, 655–668.

- Roberts, R.J., Vincze, T., Posfai, J., and Macelis, D. (2010). REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.* 38, D234–D236.
- Roberts, R.J., Carneiro, M.O., and Schatz, M.C. (2013). The advantages of SMRT sequencing. *Genome Biol.* 14, 405.
- Russo, V.E.A. (Vincenzo E.A., Riggs, A.D., and Martienssen, R.A. (1996). *Epigenetic mechanisms of gene regulation* / edited by Vincenzo E.A. Russo, Robert A. Martienssen, Arthur D. Riggs (Plainview, N.Y. : Cold Spring Harbor Laboratory Press).
- Saunders, N.J., Jeffries, A.C., Peden, J.F., Hood, D.W., Tettelin, H., Rappuoli, R., and Moxon, E.R. (2000). Repeat-associated phase variable genes in the complete genome sequence of *Neisseria meningitidis* strain MC58. *Mol. Microbiol.* 37, 207–215.
- Schoen, C., Tettelin, H., Parkhill, J., and Frosch, M. (2009). Genome flexibility in *Neisseria meningitidis*. *Vaccine* 27 Suppl 2, B103–B111.
- Schoen, C., Weber-Lehmann, J., Blom, J., Joseph, B., Goesmann, A., Strittmatter, A., and Frosch, M. (2011). Whole-genome sequence of the transformable *Neisseria meningitidis* serogroup A strain WUE2594. *J. Bacteriol.* 193, 2064–2065.
- Schutz, F., and Delorenzi, M. (2008). MAMOT: hidden Markov modeling tool. *Bioinforma. Oxf. Engl.* 24, 1399–1400.
- Seib, K.L., Pigozzi, E., Muzzi, A., Gawthorne, J.A., Delany, I., Jennings, M.P., and Rappuoli, R. (2011). A novel epigenetic regulator associated with the hypervirulent *Neisseria meningitidis* clonal complex 41/44. *FASEB J.*
- Smith, J.M., Smith, N.H., O'Rourke, M., and Spratt, B.G. (1993). How clonal are bacteria? *Proc. Natl. Acad. Sci. U. S. A.* 90, 4384–4388.
- Snyder, L.A., Butcher, S.A., and Saunders, N.J. (2001). Comparative whole-genome analyses reveal over 100 putative phase-variable genes in the pathogenic *Neisseria* spp. *Microbiol. Read. Engl.* 147, 2321–2332.
- Srikhanta, Y.N., Dowideit, S.J., Edwards, J.L., Falsetta, M.L., Wu, H.J., Harrison, O.B., Fox, K.L., Seib, K.L., Maguire, T.L., Wang, A.H., et al. (2009). Phasevarions mediate random switching of gene expression in pathogenic *Neisseria*. *PLoS Pathog.* 5, e1000400.
- Srikhanta, Y.N., Fox, K.L., and Jennings, M.P. (2010). The phasevarion: phase variation of type III DNA methyltransferases controls coordinated switching in multiple genes. *Nat. Rev. Microbiol.* 8, 196–206.
- Stephens, D.S. (2009). Biology and pathogenesis of the evolutionarily successful, obligate human bacterium *Neisseria meningitidis*. *Vaccine* 27, B71–B77.
- Stephens, D.S., Greenwood, B., and Brandtzaeg, P. (2007). Epidemic meningitis, meningococcaemia, and *Neisseria meningitidis*. *Lancet* 369, 2196–2210.
- Stern, A., Brown, M., Nickel, P., and Meyer, T.F. (1986). Opacity genes in *Neisseria gonorrhoeae*: control of phase and antigenic variation. *Cell* 47, 61–71.

- Stucki, D., Ballif, M., Bodmer, T., Coscolla, M., Maurer, A.-M., Droz, S., Butz, C., Borrell, S., Längle, C., Feldmann, J., et al. (2014). Tracking a Tuberculosis Outbreak Over 21 Years: Strain-Specific Single-Nucleotide Polymorphism Typing Combined With Targeted Whole-Genome Sequencing. *J. Infect. Dis.*
- Swartley, J.S., Marfin, A.A., Edupuganti, S., Liu, L.J., Cieslak, P., Perkins, B., Wenger, J.D., and Stephens, D.S. (1997). Capsule switching of *Neisseria meningitidis*. *Proc. Natl. Acad. Sci. U. S. A.* *94*, 271–276.
- Tachida, H., and Iizuka, M. (1992). Persistence of repeated sequences that evolve by replication slippage. *Genetics* *131*, 471–478.
- Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L., et al. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* *324*, 930–935.
- Tauseef, I., Harrison, O.B., Wooldridge, K.G., Feavers, I.M., Neal, K.R., Gray, S.J., Kriz, P., Turner, D.P.J., Ala'Aldeen, D.A.A., Maiden, M.C.J., et al. (2011). Influence of the combination and phase variation status of the haemoglobin receptors HmbR and HpuAB on meningococcal virulence. *Microbiology* *157*, 1446–1456.
- Tettelin, H., Saunders, N.J., Heidelberg, J., Jeffries, A.C., Nelson, K.E., Eisen, J.A., Ketchum, K.A., Hood, D.W., Peden, J.F., Dodson, R.J., et al. (2000). Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* *287*, 1809–1815.
- Teyssou, R., and Muros-Le Rouzic, E. (2007). Meningitis epidemics in Africa: a brief overview. *Vaccine* *25 Suppl 1*, A3–A7.
- Treangen, T.J., and Salzberg, S.L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* *13*, 36–46.
- Tzeng, Y.-L., Noble, C., and Stephens, D.S. (2003). Genetic basis for biosynthesis of the (alpha 1-->4)-linked N-acetyl-D-glucosamine 1-phosphate capsule of *Neisseria meningitidis* serogroup X. *Infect. Immun.* *71*, 6712–6720.
- Urwin, R., Russell, J.E., Thompson, E.A.L., Holmes, E.C., Feavers, I.M., and Maiden, M.C.J. (2004). Distribution of surface protein variants among hyperinvasive meningococci: implications for vaccine design. *Infect. Immun.* *72*, 5955–5962.
- Veening, J.-W., Smits, W.K., and Kuipers, O.P. (2008). Bistability, Epigenetics, and Bet-Hedging in Bacteria. *Annu. Rev. Microbiol.* *62*, 193–210.
- Vieusseux, M. (1805). Mmoire sur la maladie qui a regnè a Genkve au printemps de 1805. *J Med Chir Pharmacol* *11*.
- Virji, M. (2009). Pathogenic neisseriae: surface modulation, pathogenesis and infection control. *Nat. Rev. Microbiol.* *7*, 274–286.
- Vogel, U., and Frosch, M. (2002). The genus *Neisseria*: population structure, genome plasticity, and evolution of pathogenicity. *Curr. Top. Microbiol. Immunol.* *264*, 23–45.
- Waddington, C.H. (1957). The strategy of the genes; a discussion of some aspects of theoretical biology. (London: Allen & Unwin).

Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171, 737–738.

Weichselbaum, A. (1887). Ueber die Aetiologie der akuten Meningitis cerebrospinalis. *Fortschritte Med.* 5 573-583.

Yu, M., Hon, G.C., Szulwach, K.E., Song, C.-X., Zhang, L., Kim, A., Li, X., Dai, Q., Shen, Y., Park, B., et al. (2012). Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* 149, 1368–1380.

Zhou, K., Aertsen, A., and Michiels, C.W. (2014). The role of variable DNA tandem repeats in bacterial adaptation. *FEMS Microbiol. Rev.* 38, 119–141.

7 Abbreviations

SNP	Single Nucleotide Polymorphism
dnmt	DNA Methyltransferase
5mC	5-methylcytosine
4mC	4-methylcytosine
6mA	6-methyladenosine
5caC	5-carboxylcytosine
AdoMet	S-adenosyl-L-methionine
RMS	Restriction-Modification systems
Dam	DNA Adenine Methylase
SMRT	Single Molecule Real Time
ZMW	Zero Mode Waveguides
IPD	Interpulse Duration
MLST	Multilocus Sequencetyping
ST	Sequence Type
STR	Short Tandem Repeats
PV	Phase Variation
ORF	Open Reading Frame
modQV	Modification Quality Value
CNV	Copy Number Variation
RegEx	Regular expression
HMM	Hidden Markov Model
Indel	Insertion or Deletion

8 Acknowledgements

First and foremost I offer my sincere appreciation and gratitude to Dr. Christoph Schmid for offering me the chance to be part of this project. With his keen supervision, constant availability and flexibility not only to answer my many questions but also for teaching me where and how to find solutions. Christophs' commitment and dedication have always been inspiring to me. His impeccable attention to details and rich scientific knowledge was equally fascinating me. His complex questions however sometimes surpassed the complexity of *Neisseria* and for that I blame him for the long pathoGenOmics group meetings☺. Christoph always maintained a respectful and exciting working environment and for that I am deeply thankful.

I am grateful to my committee members Prof. Marcel Tanner, Prof. Urs Jenal and Prof Gerd Pluschke for their advice and critical input during our annual meetings.

Special thanks to Araceli Lamelas for all your support and advice. This work would not be the same without you.

Many thanks also go to the molecular immunology group especially Julia Hauser, for their valuable assistance with bacterial cultures. Many thanks also to Prof. Till Voss and all the Gene Regulation team for offering me the laboratory space and resources to carry out my experiments.

I would like to thank the TB group and Prof. Sebastien Gagneux for their advice and stimulating discussion. David, Daniela, Mireia, Sonia, Andrej, Damien, Bijaya and all the third floor friends I am lucky to have met you.

Thanks are also due to the whole Swiss TPH for the fantastic environment. Special thanks to my friend Christine Mensch for her constant support.

My deepest gratitude goes to my family especially my wife for their trust, support and understanding.

Special thanks to the Forschungsfonds of the University of Basel, and the EU FP7 for their generous funding.

9 CURRICULUM VITAE

Surname: Rustom Abdul Sater
 First Name: Mohamad
 Date of birth: 15 July 1985
 Nationality: Lebanon
 E-mail: m.sater@unibas.ch
 Address: Elisabethenstrasse 38, 4051, Basel, Switzerland

EDUCATION

- 04/2015 PhD, Microbiology/Bioinformatics, Swiss TPH & Swiss Institute of Bioinformatics, University of Basel, Switzerland
 - **Supervision:**
 - Dr. Christoph D. Schmid
 - **Projects:**
 - Genetic and epigenetic variability in *N. meningitidis*
 - Electromagnetic field impact on epigenetic stability during stem cell differentiation
 - **Funding:**
 - Forschungsfonds der Universität Basel, Switzerland
 - EU FP7, ARIMMORA
- 02/2011 Masters of Science in infection and Molecular Biology, Molecular Immunology group, , University of Basel, Switzerland
 - **Supervision:** Prof. Gerd Pluschke/Dr. Christoph Schmid
 - **MSc. Dissertation:** DNA methylation and gene regulation in *N. meningitidis*
 - **Grade:** 5.7/6
- 08/2006 Bachelor of Science, Biology Department, American University of Beirut, Lebanon

TRAINING

- 05/2014 Pacific Biosciences Bioinformatics Workshop, ETH Zürich.
- 06/2012 Perl Programming for Biomedical Research, Swiss TPH.
- 11/2011 Next generation sequencing: *De novo* Genome Assembly, Swiss Institute of Bioinformatics.
- 04/2011 UNIX Essentials, Swiss Institute of Bioinformatics.

WORK EXPERIENCE

- 2010 – 2012 Research assistant and insectary Technician, Medical and Diagnostic services, Swiss TPH, Switzerland.
- 2007 –2009 Research Assistant, Marine Biology Laboratory Biology Department, AUB, Lebanon.
- 2006-2007 Museum of Natural History, Biology Department, AUB, Lebanon.

TALKS / POSTERS

- 01/2014 Oral Presentation, Swiss TPH Research Seminars, Basel , “DNA methylation in *Neisseria meningitidis* from an epigenetic perspective”
- 06/2013 Short talk and poster, Systems Biology of Infection Symposium, Ascona
“SMRT sequencing of *N. meningitidis* genomes reveals high variability in DNA modifications”
- 09/2013 Poster, Biovalley Science Day, Basel
- 09/2012 Poster, European Conference on Computational Biology, Basel
- 08/2012 Poster, Global Health meets Infection Biology conference, Lausanne
- 10/2011 Poster, SystemsX conference, Basel

Grants & Awards

- 11/2014 R. Geigy Foundation Stiftung grant
- 06/2013 Systems Biology of Infection award, Ascona, Switzerland

INTERESTS

- Philosophy, Mystic Poetry
- History of Sciences
- Photography
- Scuba diving (NAUI Certified) and bird watching

