Phase-variation of genes is defined as the rapid and reversible switching of expression — either ON-OFF switching or the expression of multiple allelic variants. Switching of expression can be achieved by a number of different mechanisms. Phase-variable genes typically encode bacterial surface structures, such as adhesins, pili, and lipooligosaccharide, and provide an extra contingency strategy in small-genome pathogens that may lack the plethora of ‘sense-and-respond’ gene regulation systems found in other organisms. Many bacterial pathogens also encode phase-variable DNA methyltransferases that control the expression of multiple genes in systems called phasevarions (phase-variable regulons). The presence of phase-variable genes allows a population of bacteria to generate a number of phenotypic variants, some of which may be better suited to either colonising certain host niches, surviving a particular environmental condition and/or evading an immune response. The presence of phase-variable genes complicates the determination of an organism’s stably expressed antigenic repertoire; many phase-variable genes are highly immunogenic, and so would be ideal vaccine candidates, but unstable expression due to phase-variation may allow vaccine escape. This review will summarise our current understanding of phase-variable genes that switch expression by a variety of mechanisms, and describe their role in disease and pathobiology.

Introduction
Phase-variable bacterial loci rapidly and reversibly switch their expression. In many small-genome, host-adapted bacterial pathogens, phase-variation serves as an extra contingency strategy to allow adaptation to changing conditions [1], and form part of the ‘tinkerer’s evolving toolbox’ [2]. The mechanisms and processes behind the evolution of these loci have been discussed in excellent detail previously [2,3], and as such this review will focus on the role, and the advantages, of phase-variable gene expression during pathobiology in a number of important human pathogens. Phase-variable genes often encode bacterial cell-surface features, such as adhesins, iron acquisition proteins, pilis and lipooligosaccharide (LOS) biosynthetic enzymes [4–9]. The suite of phase-variable genes in a particular species is referred to as the ‘phasome’ [10]. Variable expression of the phasome within a bacterial population results in a variety of phenotypically distinct individuals which may be better equipped to colonise certain host niches or better able to evade a pre-primed host immune response. For example, expression of an adhesin may be required for initial colonisation of the host, and variants where the adhesin is expressed may be selected for during this stage of infection. However, this adhesin may be highly immunogenic, and its expression may be selected against if the host has a pre-primed immune response, as those variants expressing high levels are killed by the immune system. Modifications to LOS may impart serum resistance and resistance to neutrophil-mediated killing [11,12], but may be selected against in different host niches [8]. The polysaccharide capsule of Neisseria meningitidis is absolutely required for resistance to serum and always present in invasive isolates [13], but variants expressing high levels of capsule show decreased adherence to host cells [14], meaning they may be
less equipped to initially colonise the host. As such, a ‘back-and-forth’ selection and counter-selection for the different phenotypes resulting from phase-variable gene expression occur as bacteria colonise different niches or encounter different pressures.

The rapid and reversible switching of gene expression means that many antigens encoded by phase-variable genes are often discounted as vaccine candidates. However, under certain circumstances, phase-variable genes can be used as vaccine antigens if they are highly immunogenic or their expression is high during colonisation and/or disease. For example, the NadA protein forms part of the 4CMenB vaccine against *N. meningitidis* serogroup B (licensed as Bexsero) [15], and although it undergoes phase-variation due to DNA repeats in the region upstream of its promoter, it is highly expressed during infection [16], and thus required for a key stage in disease. Therefore, targeting this protein with a vaccine will result in protection against a key stage of disease. Similarly, the adhesin Hia switches between high and low expression states by variation in length of a DNA repeat tract in its promoter region, but is being investigated as a vaccine candidate for non-typeable *Haemophilus influenzae* (NTHi) as it is able to induce high levels of serum anti-Hia antibodies in a Chinchilla model of NTHi disease [17]. Targeting Hia with a vaccine would in theory prevent the initial colonisation of the host as *hia* expression is selected for during a Chinchilla model of host colonisation [4], meaning protection could be achieved by preventing colonisation.

**Mechanisms of phase-variation**

Genes can phase-variation through a number of genetic mechanisms [1] (Figure 1), including variation in the length of hypermutable simple DNA sequence repeat (SSR) tracts, recombination-mediated shuffling between expressed and silent loci, promoter inversions, and by epigenetic mechanisms [1,18]. SSR tracts are unstable and vary in length through polymerase slippage during replication. Longer SSR tracts exhibit higher rates of phase-variation [19–21]. If an SSR tract is located in the open reading frame (ORF) of a gene, variation in tract length can result in the expression of the gene (ON), or due to a frame-shift mutation downstream of the SSR resulting in a premature stop codon, the gene is not expressed (OFF), or in some cases a truncated protein is expressed. SSR tracts can also be located in the promoter of a gene, where they result in a gradient of protein expression. Alternatively, recombination, or shuffling, between expressed and silent variants of particular loci results in the switching of expression between multiple allelic variants of a single protein. This often occurs via recombination between inverted repeats (IRs) that are present within these loci. This type of phase-variable gene expression is also referred to as antigenic-variation. Furthermore, many bacterial promoters are invertible, which results in ON-OFF switching of their respective genes [22], with promoter inversions often catalysed by an associated recombinase. In addition, differential methylation of DNA at particular target sequences in promoters can lead to up- and down-gene regulation by epigenetic mechanisms, and is dependent on the interaction of the methylated site or the methyltransferase with a regulatory protein at the same site [23]. Phase-variable regulons — phasevarions — control differential regulation of multiple genes through phase-variation of a single gene encoding a methyltransferase [24], adding a further level of complexity to understand gene expression in a number of pathogenic. This review will highlight a variety of interesting and well-studied examples of phase-variable genes in bacteria, and the role that phase-variable gene expression contributes to the biology of the organisms containing them. It will also highlight the implications of phase-variation on vaccine development.

**Phase-variable expression of genes via genetic mechanisms**

**Adhesins**

Many adhesins found in bacteria are phase-variable [1]. Whilst these are required for colonisation, as their expression is required to adhere to host cells, they are also often highly immunogenic; therefore individuals that have lower levels or no expression of the adhesins are better equipped to survive an immune response. For example, the HMW1 and HMW2 adhesins are found in ~75% of NTHi isolates [25]. Genes encoding both adhesins contain heptanucleotide TCTTTCA(n) repeats in their promoter regions. As the number of TCTTTCA(n) repeats in this SSR increases, the level of expression of HMW1/2 decreases [26]. HMW1/2 are required for binding to related host cell receptors in the human airway [27,28], but they are also highly immunogenic, and are currently being investigated as a candidate for a vaccine against NTHi [29,30].
The adhesin SabA in the human gastric pathogen *Helicobacter pylori* is capable of phase-variation through changes in the length of two SSR tracts associated with the *sabA* gene. The *sabA* gene contains a CT(n) tract in its ORF, and a T(n) tract in its promoter [31]. ON-OFF switching occurs due to variation in the length of the SSR tracts.
CT_{(n)} tract [31], with fine tuning of expression occurring via T_{(n)} tract variation [32]. This fine tuning of sabA expression is a result of changes in the DNA structure of the promoter region through T tract length variation, leading to differences in RNA polymerase binding and therefore levels of transcription.

The opacity proteins of Neisseria gonorrhoeae phase-vary through both ON-OFF switching and by expression of multiple allelic variants [33]. Expression of allelic variants is the result of the number of CTCTT_{(n)} repeats in the ORF of multiple variable copies of the *opa* genes [33,34]. The selective pressure of the host immune system drives variation of Opa protein expression [35], allowing evasion of pre-primed immune responses against this organism.

**Autotransporter proteins**

Autotransporters, also called Type V secretion systems, are characterised by an N-terminal passenger domain and a C-terminal transmembrane domain, through which the functional passenger domain is translocated to the bacterial surface [36,37]. Many autotransporters function as adhesins, and a selection for and against their expression during pathobiology would similarly provide advantages to the organisms containing them. The adhesin Hia in NTHi contains a T_{(n)} SSR tract in its promoter, and switches between low and high expression levels due to variation in the length of this tract [4]. Selection for Hia expression occurs during colonisation of the host nasopharynx, and was commensurate with 34T residues present in the T_{(n)} tract. However, Hia is immunogenic, and it was demonstrated that T_{(n)} tract lengths that result in low Hia protein expression levels (30T residues) are selected for during *in vitro* opsonophagocytic killing assays [4]. Thus, selection and counter-selection for Hia protein expression levels, mediated by phase-variation, occur during NTHi colonisation and pathogenesis.

A number of autotransporters in *N. meningitidis* are phase-variable. For example, NadA is a major adhesin in *N. meningitidis*, and forms part of the 4CMenB (Bexsero) vaccine against *N. meningitidis* serogroup B [15]. The nadA promoter contains a TAAA_{(n)} repeat, with the number of TAAA_{(n)} repeats affecting the spacing of key regulatory elements [38], and consequently differential expression. NadA has been shown to be highly expressed during infection [16], meaning that NadA could be targeted by a vaccine during disease even though it is phase-variable. MspA is an outer-membrane serine protease in *N. meningitidis* that can also be found in the extracellular medium due to auto-catalytic cleavage that releases the passenger domain. ON-OFF switching of expression of MspA occurs due to variation in the length of a C_{(n)} SSR tract located in the ORF of the *mspA* gene [39]. MspA is expressed at high levels during invasive disease, [39,40] and is required for adhesion to human epithelial and endothelial cells [41]. The AutB adhesin switches ON-OFF due to changes in the length of an AAGC_{(n)} SSR tract in its ORF [42]. Expression of AutB results in increased biofilm formation, but when phase-varied OFF, *N. meningitidis* is able to cross epithelial layers at a higher rate [42]. AutB is highly immunogenic, so switching OFF of expression would also allow evasion of an immune response. Thus, like Hia in NTHi, selection and counter-selection for AutB expression likely occur during meningococcal colonisation and disease progression.

**Pili, fimbriae, and flagella**

Type IV pili in the pathogenic *Neisseria* can phase-vary by multiple mechanisms. For example, the expression of *pilC*, encoding the putative tip adhesin of the Type IV pili [43], switches ON-OFF by variation in the length of a G_{(n)} tract [44,45]. The major pili protein subunit, encoded by *pilE*, shuffles between multiple allelic variants by recombining with silent variable *pilS* loci [46,47]. *N. gonorrhoeae* has one expressed *pilE* gene and up to 19 silent variable *pilS* genes, distributed in four or five locations across the genome [48]. *N. meningitidis* typically encodes four to eight variable *pilS* sequences, contained in a single locus on the chromosome [49]. Shuffling between *pilE* and *pilS* loci can also lead to non-functional sequences in the *pilE* locus, meaning cells are non-piliated [50]. RecA, Rep, and RecF proteins, involved in DNA recombination, are all essential for pilin antigenic-variation in *N. meningitidis* [51]. RecA is amongst the seven proteins that are absolutely essential for pilin variation in *N. gonorrhoeae* [48,52], in addition to a guanine quadruplex motif upstream of the *pilE* gene [53]. Interaction between cells expressing different pilin variants within *N. gonorrhoeae* populations results in variable colony formation [54], which could, for example, influence biofilm formation and therefore gonococcal pathogenesis and treatment.

The best-studied example of fimbrial phase-variation occurs in Type I fimbriae of *Escherichia coli* [55]. An invertible DNA element, *fimS*, which contains the *fimA* promoter, is encoded upstream of the major fimbrial
subunit gene, *fimA*. Inversions in *fimS* result in ON-OFF switching of the *fimA* gene [55] and consequently production, or not, of Type I fimbriae.

*Clostridium difficile* is a major nosocomial pathogen and the cause of potentially fatal colitis. It is able to switch expression of its flagellar ON-OFF via inversions in the DNA associated with the *flaB* gene [56], termed the ‘flagellar switch’. Inversions in this promoter also alter production of the toxins TcdA and TcdB by *C. difficile* [57], which could have implications in *C. difficile* pathobiology.

**Iron acquisition proteins**

*N. meningitidis* encodes two separate haemoglobin receptors that are phase-variably expressed: HpuAB and HmbR [58]. ON-OFF switching occurs commensurate with the length of a *G(n)* tract located within the ORF of the *hpuA* and *hmbR* genes [6,58]. HmbR also shows allelic variation between different meningococcal strains, implying that the selection for variants occurs *in vivo* [59]. A strain lacking both HpuAB and HmbR was less virulent in a rat model of infection [60], but not impaired in its growth in human blood [61]. Examination of *in vivo* isolates showed that the majority of strains were phase-ON for *hpuA* or *hmbR*, implying that haemoglobin acquisition is key for systemic disease [62,63].

Haemoglobin and haptoglobin binding proteins are also phase-variable in *H. influenzae*. Genes for the related proteins HgpA, HgpB, and HgpC contain CCA*G(n)* SSR tracts in their ORFs, and show ON-OFF switching of expression [7]. HgpA is required for full virulence in an infant rat model of invasive disease [64], implying that the selection for HgpA ON would likely occur in invasive *H. influenzae/NTHi* isolates, although this remains to be investigated.

A novel iron acquisition protein, Irp, expressed by the bovine pathogen *Mannheimia haemolytica* has been reported to undergo phase-variation [65]. The *irp* gene contains an *A(n)* SSR tract, with ON-OFF switching of *irp* expression resulting from changes in the number of adenine residues present [65]. In addition, a stem-loop structure formed by a short IR upstream of this *A(n)* tract contributes to rates of phase-variation [65], demonstrating a complex mode of phase-variable expression of this iron acquisition protein in *M. haemolytica*.

**Lipooligosaccharide biosynthetic enzymes**

Lipooligosaccharide (LOS) is a major virulence factor in a number of bacterial pathogens, such as NTHi and the pathogenic *Neisseria*, and has been proposed as a vaccine candidate for *N. gonorrhoeae* [66]. LOS contributes to NTHi survival *in vivo* [8,67]. Many LOS biosynthetic genes contain SSR tracts and are phase-variably expressed [68,69]. Variation of the expression of enzymes required for LOS biosynthesis results in a heterogeneous LOS. In NTHi, at least seven LOS biosynthetic genes are phase-variable: *lic1A*, encoding a phosphorylcholine transferase [70], *lic2A* encoding a galactosyltransferase [71], *lic3A* and *lic3B* encoding related sialyltransferases [67,72], *lex2A* encoding a glucosyltransferase [73], *lgIC* encoding a galactosyltransferase [74], and *oafla* encoding an O-acetyltransferase [12]. Variable expression of these genes is selected for when NTHi colonise or cause disease in humans. For example, the addition of galactose by Lic2A is required for resistance to human serum [75], and protects cells from neutrophil-mediated killing [11], but *lic2A* expression is switched OFF in the majority of invasive NTHi isolates [76]. This indicates a complex role for LOS modified by Lic2A, and demonstrates the fine tuning of phenotype afforded by phase-variable loci. Expression of the O-acetyltransferase Oafla is turned OFF in NTHi during middle ear infection [8], but turned ON during invasive disease [76]. During experimental infection of human volunteers with NTHi, both *lex2A* and *lic1A* were shown to switch from OFF to ON during nasopharyngeal colonisation [9]. These findings together show the rapid adaptability to be gained by LOS phase-variation during host colonisation and disease.

*Clostridium jejuni*, a major human gastric pathogen, also switches the expression of many LOS biosynthetic enzymes due to the presence of SSRs in these genes [77]. The addition of terminal GM1 or GM2 gangliosides is dependent on variation in the length of a *G(n)* tract in the *wlaN* gene, encoding a beta-1,3 galactosyltransferase [78]. These structures mimic host glycans, allowing *C. jejuni* to evade host immune responses [79], and is also the basis for the auto-immune disease Guillain–Barre syndrome [80]. LOS phase-variation can also result from shuffling between variable biosynthetic loci, leading to ‘mosaic’ LOS structures, caused by different specificities of the encoded enzymes [81]. *C. jejuni* isolates do not necessarily encode all LOS biosynthetic genes or contain point mutations in one or more biosynthetic loci [82,83], so in addition to switching expression of genes ON-OFF, LOS heterogeneity can occur through lack of functional biosynthetic enzymes.
Phase-variation of gene expression through epigenetic regulation

Epigenetic regulation by DNA methyltransferases

Epigenetics is the study of heritable gene expression changes that occur without a change in the DNA sequence [84]. DNA methylation at adenine residues is the most common form of epigenetic regulation in bacteria [85]. The DNA methyltransferase Dam (DNA adenine methyltransferase) is a well-studied example of epigenetic regulation in bacteria. Dam regulates genes by methylating DNA or binding and competing with regulatory proteins at specific GATC target sites [23,86]. For example, variable expression of the Pap pilus and antigen 43 (Ag43) in E. coli is mediated by Dam methylation at their respective promoters. Methylation of the pap promoter by Dam alters the affinity of the LRP regulatory protein for DNA, and results in ON-OFF phase-variable switching of Pap pilus expression [5,87]. Mutants lacking OxyR are locked ON for Ag43 expression, whereas strains lacking Dam are locked OFF for Ag43 expression, implying OxyR competes with Dam for unmethylated GATC sites in the promoter region of the ag43 gene [88].

Phase-variable DNA methyltransferases

In addition to phase-variable genes encoding surface features, many bacterial pathogens encode cytoplasmic methyltransferases, associated with restriction-modification (R-M) systems, that are subject to phase-variation. Phase-variation of methyltransferase expression results in differential methylation throughout the genome, leading to variable expression of multiple genes through epigenetic mechanisms. These systems are called phasevarions (phase-variable regulons), and have been described in a number of human-adapted pathogens [89–95]. All phasevarions described to date regulate expression of multiple genes, and many include proteins involved in host colonisation, survival, and pathogenesis, and many regulate putative vaccine candidates. The genes regulated by phasevarions do not contain any identifiable features, and so complicate the identification of an organism’s stably expressed protein repertoire. Phasevarions and their role in pathobiology and vaccine development have been described in detail in a number of recent reviews [18,24,96], and as such this review will only briefly describe them.

Many of the phasevarions described to date are controlled by phase-variation of Type III mod genes [97]. In these systems, the methyltransferase (Mod) phase-varies between two states (ON or OFF) by variation in the number of SSRs in the encoding mod gene [98]. A recent survey of all Type III methyltransferases in REBASE showed that nearly 20% of Type III mod genes contain SSRs, are therefore able to phase-vary, and potentially able to control a phasevarion [99]. mod genes are highly conserved (>95% DNA sequence identity) in their 5' and 3' regions, but contain a highly divergent central domain, the TRD (for Target Recognition Domain). The TRD determines the DNA sequence that is methylated by the Mod methyltransferase. Thus, mod genes can exist as multiple allelic variants, due to TRD variation, and which therefore encode enzymes that methylate a different DNA target sequence. Methylation of a different DNA target sequence means different Mod proteins regulate the expression of a different suite of genes; i.e. they control different phasevarions. For example, H. influenzae contains 21 different modA alleles [95]; Neisseria species contain 7 modB alleles [100]; and H. pylori contains 17 modH alleles [93].

Many Type I R-M systems contain multiple variable hsdS loci that contain IRs, and shuffle between multiple allelic variants. The specificity of Type I R-M systems is dictated by the encoded specificity gene, hsdS; if the expressed HsdS protein changes, so does the sequence methylated. Therefore, rather than Type III mod genes that reversibly switch ON-OFF, these Type I methyltransferases are always expressed, but as multiple allelic variants, dependent on the sequence of the expressed hsdS gene. These phase-variable Type I methyltransferases have been termed ‘inverting’ systems, as their specificity varies by DNA inversions [101]. Inverting Type I methyltransferases have been described and studied in Streptococcus pneumoniae [91] and Streptococcus suis [102], and observed in Listeria monocytogenes and Enterococcus faecalis [101], suggesting these systems are widespread in the bacterial domain. The inverting Type I methyltransferase in S. pneumoniae shuffles between six methyltransferase specificities (SpnD39III HsdS alleles A-F) [91], and is in part catalysed by a recombinase encoded at the Type I locus [103]. Phase-variation of the SpnD39III system results in the variable expression of the pneumococcal capsule [91], one of the main virulence factors of this organism. In S. suis, the inverting Type I system shuffles between four unique HsdS proteins [102].

A single Type I R-M system that varies by SSR tract variation has been identified in the human pathogen N. gonorrhoeae [104]. Variation in length of the SSR tract results in a truncated or full-length HsdS protein.
being expressed, resulting in two different methyltransferase specificities [104]. However, apart from the SpnD39III system in *S. pneumoniae*, none of these Type I systems have been shown to result in differential gene regulation commensurate with methyltransferase phase-variation.

**Summary/Conclusion**

Many host-adapted bacterial pathogens use phase-variation as a strategy to generate phenotypic variation within a population, which allows adaptation to changing host micro-environments, and evasion of the immune system. Phase-variable gene expression is rapid and reversible — variants better able to adapt to particular conditions may be at a disadvantage in others, meaning a selection and counter-selection for variants is continually occurring in a bacterial population as it interacts with the host. A thorough and comprehensive understanding of phase-variable gene expression will allow the generation of improved vaccines and treatments; although it appears counter-intuitive to use proteins that show unstable expression in a vaccine, phase-variable proteins are often highly immunogenic, and if we understand their regulation, and the conditions in which they are required, we will have a further tool in our arsenal to combat many important pathogens.

**Perspectives**

- **The importance of the field:** Understanding phase-variable genes in bacteria is key to understanding bacterial adaptation to changing conditions within the host, and in determining the stably expressed protein and antigenic repertoire of organisms encoding phase-variable genes.

- **Summary of current thinking:** Although phase-variable genes, and the genes controlled by phasevarions, are not ideal vaccine candidates as their expression is not stable, they can be included in multi-subunit vaccines if their expression is high and/or essential under certain circumstances, or they are highly immunogenic. In order to design stable and efficacious vaccines and treatments, a thorough understanding of the conditions influencing phase-variable gene expression is required, including pro- and anti-selective pressures, *in vivo* niches where the genes may be required, and the mode of phase-variable expression.

- **Future directions:** Study of *in vivo* selective pressures will be key for understanding phase-variable gene expression, but these are not usually easily replicated *in vitro*. However, a thorough investigation of phase-variable gene expression using multiple variable experimental conditions will provide the best possible information on which phase-variable genes can be included in rationally designed vaccines. This could potentially lead to the generation of vaccines that contain only highly immunogenic phase-variable proteins, that are essential for important aspects of infection, thereby reducing the chances that all genes are switched OFF during infection. This would decrease the chances of vaccine escape.

**Abbreviations**

IR, inverted repeat; LOS, lipoooligosaccharide; NTHi, non-typeable *Haemophilus influenzae*; ORF, open reading frame; SSR, simple DNA sequence repeat.

**Author Contribution**

J.M.A., Z.N.P., and G.T. wrote the article and prepared figures. K.L.S. critically read the manuscript, and provided valuable feedback and input.

**Acknowledgements**

This work was supported by the Australian Research Council (ARC) Discovery Project 180100976 to J.M.A.; Australian National Health and Medical Research Council (NHMRC) Project Grant 1099279 to K.L.S. and J.M.A.; and Garnett Passe & Rodney Williams Memorial Foundation (GPRWMF) Grant-In-Aid (Supplementation) to K.L.S. and J.M.A.

© 2019 The Author(s). This is an open access article published by Portland Press Limited on behalf of the Biochemical Society and distributed under the Creative Commons Attribution License 4.0 (CC BY).
Competing Interests
The Authors declare that there are no competing interests associated with the manuscript.

References
1140

© 2019 The Author(s). This is an open access article published by Portland Press Limited on behalf of the Biochemical Society and distributed under the Creative Commons Attribution License 4.0 (CC BY).
101 De Ste Croix, M., Vacca, I., Kwun, M.J., Ralph, J.D., Bentley, S.D., Haigh, R. et al. (2017) Phase-variable methylation and epigenetic regulation by type
85 De Ste Croix, M., Vaccarino, I., Kwun, M.J., Ralph, J.D., Bentley, S.D., Haigh, R. et al. (2017) Phase-variable methylation and epigenetic regulation by type I restriction-modification systems. Frontiers in Microbiology 8, 1234