BMC Genomics

RESEARCH ARTICLE Open Access

# *In-silico* prediction and deep-DNA sequencing validation indicate phase variation in 115 *Neisseria meningitidis* genes

Emilio Siena[1*], Romina D'Aurizio[1,4], David Riley[2,5], Hervé Tettelin[2], Silvia Guidotti[1], Giulia Torricelli[1], E. Richard Moxon[3] and Duccio Medini[1*]

## Abstract

**Background:** The *Neisseria meningitidis* (*Nm*) chromosome shows a high abundance of simple sequence DNA repeats (SSRs) that undergo stochastic, reversible mutations at high frequency. This mechanism is reflected in an extensive phenotypic diversity that facilitates *Nm* adaptation to dynamic environmental changes. To date, phase-variable phenotypes mediated by SSRs variation have been experimentally confirmed for 26 *Nm* genes.

**Results:** Here we present a population-scale comparative genomic analysis that identified 277 genes and classified them into 52 strong, 60 moderate and 165 weak candidates for phase variation. Deep-coverage DNA sequencing of single colonies grown overnight under non-selective conditions confirmed the presence of high-frequency, stochastic variation in 115 of them, providing circumstantial evidence for their phase variability.
We confirmed previous observations of a predominance of variable SSRs within genes for components located on the cell surface or DNA metabolism. However, in addition we identified an unexpectedly broad spectrum of other metabolic functions, and most of the variable SSRs were predicted to induce phenotypic changes by modulating gene expression at a transcriptional level or by producing different protein isoforms rather than mediating on/off translational switching through frameshifts.
Investigation of the evolutionary history of SSR contingency loci revealed that these loci were inherited from a *Nm* ancestor, evolved independently within *Nm*, or were acquired by *Nm* through lateral DNA exchange.

**Conclusions:** Overall, our results have identified a broader and qualitatively different phenotypic diversification of SSRs-mediated stochastic variation than previously documented, including its impact on central *Nm* metabolism.

**Keywords:** *Neisseria meningitidis*, Phase variation, Contingency loci, Comparative genomics, Host-pathogen interaction

## Background

*Neisseria meningitidis (Nm)* is a Gram-negative, encapsulated bacterium that is present, as a commensal organism, in the nasopharyngeal cavity of five to ten percent of the adult population [1, 2]. Despite its prevalence as a harmless organism some strains, for reasons not yet completely understood, can cross the epithelial barrier and enter the bloodstream, causing septicemia and life-threatening disease [3–6].

In order to maximize its fitness in the diverse and changing environments offered by the host-pathogen interplay, *Nm* has evolved multiple and complementary adaptive strategies [7, 8]. One such mechanism is represented by Simple Sequence Repeats (SSRs), contiguous iterations of short DNA motifs, generally assumed to range from 1 to 10 nucleotides in length [9].

SSRs, often called contingency loci, are hyper-mutable DNA sequences mediating high frequency, stochastic, reversible and heritable genotypic switching [10], whose number of tandemly repeated motifs can vary with relatively high frequencies [11, 12]. A variable number SSR (VNSSR) located within a coding sequence can change

* Correspondence: emilio.x.siena@gsk.com; duccio.x.medini@gsk.com
[1]GSK Vaccines, 53100 Siena, Italy
Full list of author information is available at the end of the article

translation by introducing frameshifts in the reading frame [13–15] or, when located in the proximity of a promoter, can modulate transcription either by switching between alternative translational start sites [11, 16] or by altering the promoter sequence [17–19].

The first *Nm* genome-wide SSRs analyses were based on the single genome sequences determined for the MC58 [9] and Z2491 [20] strains. Putative phase-variable genes were predicted on the basis of *cis* factors such as SSR sequence, number of repetitions and sequence context. For a subset of genes phase variation was confirmed experimentally.

Comparative investigations, in which the genomic sequences of strains *N. gonorrhoeae* FA1090 and *Nm* FAM18 were added to the analysis, suggested that the presence of length polymorphisms among orthologous SSRs present in different genomes is a reliable predictor of phase variation [21, 22]. Overall, 78 putative phase variable genes were reported in the *Neisseria* genus, of which 67 were specific to the *Nm* species and included genes involved in cell adhesion (adhesins), capsule formation (evasins), biosynthesis of the lipopolysaccharide layer (e.g. *LgtA* and *lgtE*), cell-surface receptors (e.g. the *HmbR* iron-acquisition receptor) and restriction/modification systems [22]. Among these, phase variation was confirmed experimentally for only 26 genes, primarily combining PCR with immunoblotting techniques [17, 23–32] or with northern blotting and quantitative PCR [13, 17, 33–35].

Here, we report that the limited number of sequences employed in previous VNSSRs studies, together with the lack of high-throughput technologies (next-generation sequencing, NGS) suitable for a large-scale validation of putative phase variable genes, have potentially led to an underestimation of the overall impact of SSR-mediated phase variation on *Nm* phenotype and fitness. Also, the evolutionary dynamics underlying the generation of VNSSRs in *Nm* and other species has not been fully elucidated.

Recently, new insights into the *Nm* population structure and dynamics have been generated through a comparative genomic analysis based on the complete genome sequences of 20 *Nm* strains, including multiple isolates from each of the most virulent clonal complexes [36]. Also, the advent of NGS technologies have allowed for an unprecedented sequencing depth across the whole genome, such that the presence of mixed populations at specific loci of a single genome can be detected with high throughput, including those variants occurring at low frequencies during bacterial replication [37–40].

In the present study we took advantage of *Nm* pangenome data to map, characterize and infer functional impact and evolutionary properties of SSR contingency loci, and NGS to experimentally validate *in silico* predictions on stochastic genotypic switching.

## Results

### Two hundred seventy-seven meningococcal genes are associated with simple sequence repeats that show inter-strain length polymorphisms

An average of 4243 SSRs were identified in each genome, the majority of which (95 %) were represented by homopolymeric tracts (Additional file 1). The number of SSRs was similar across genomes (coefficient of variation = 3 %) and their abundance was found to be significantly higher than random expectation ($p = 6.8e\text{-}8$; Additional file 2: Figure S1). Significant deviation from neutrality was also observed for each individual SSR type.

We identified 6295 clusters of orthologous SSRs, 35 % (2183) of which were represented in every genome analyzed. Among these, 324 clusters had polymorphisms in the repeat tract length across different genomes, located either in intragenic regions (166, 51 %) or in the associated gene's upstream region (158, 49 %). These were selected for further analyses based on their potential to change gene expression.

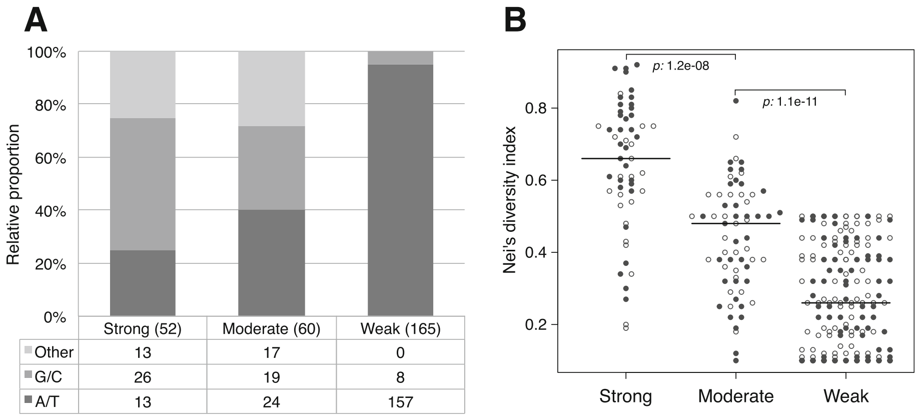
Consistent with previous findings [9], homopolymeric tracts were the most abundant simple sequence repeat type, with 288 VNSSRs clusters divided into 229 A/T and 59 G/C tracts. The remaining 36 clusters were represented by tandem repeats of two- to nine-nucleotides motifs (Additional file 3).

Positional analysis identified 166 (51 %) intragenic and 158 (49 %) intergenic VNSSRs, respectively. Interestingly, all VNSSRs with a unit motif of three-, six- and nine-nucleotides, for which a length polymorphism would not result in the disruption of the reading frame, were found to be located within coding sequences (Additional file 3). Finally, thirty-two genes were associated with more than a single VNSSR, resulting in a total of 277 phase variable gene candidates.

### SSR contingency loci show different genotype switching frequencies

Based on repeat variability and intra-strain phylogenetic relationships we stratified the 277 VNSSR-associated genes into 52 strong, 60 moderate and 165 weak candidates for phase variation. Most weak candidate phase variable genes (157, 95 %) were associated with A/T homopolymeric tracts while G/C repeats were the most abundant VNSSR type among the strong candidate genes [26, 50 % (Fig. 1a)]. VNSSRs other than homopolymeric repeats were found in 17 moderate (28 %) and 13 strong (25 %) candidate genes, respectively (Fig. 1a).

Intrinsic variability of VNSSRs clusters, defined as the Nei's diversity index [41], was then computed for the



| | Strong (52) | Moderate (60) | Weak (165) |
|---|---|---|---|
| ▨ Other | 13 | 17 | 0 |
| ▨ G/C | 26 | 19 | 8 |
| ▨ A/T | 13 | 24 | 157 |

**Fig. 1** Properties of identified VNSSRs. **a** Relative contribution of different VNSSRs types to the Strong, Moderate and Weak candidate phase variable genes. Other: all non-homopolymeric repeats. G/C: guanine or cytosine variable homopolymeric tracts. A/T: adenine or thymine variable homopolymeric tracts. **b** Nei's diversity index distribution of VNSSRs associated with the 52 strong, 60 moderate and 165 weak candidate phase variable genes. Genes associated with intragenic and intergenic VNSSRs are represented full and empty circles, respectively. The thick horizontal lines represent distributions median. *p*: Wilcoxon-Mann-Whitney *p* value

three gene sets. VNSSRs associated to strong candidate genes were significantly more variable than those associated to moderate genes, which in turn showed higher variability than those linked to weak phase variable gene candidates (Fig. 1b), providing supporting evidence that genes associated with a different likelihood of undergoing phase variation exist and that these can be identified through the analysis of a suitable genome dataset.

A positive correlation between the SSR tract length and repeat instability has been proposed [9, 42], possibly as a result of a decreased proof-reading efficiency of the DNA polymerase over the longer tracts [43]. We could confirm such association for the strong [Pearson correlation coefficient: $R = 0.6$ (homopolymeric tracts) and $R = 0.73$ (non-homopolymeric repeats)] but not for the moderate or weak candidate gene pools (Additional file 2: Figure S2).

## VNSSRs variation is detectable in single colonies grown overnight under non-selective conditions

The genomes of five of the analyzed strains were re-sequenced at high depth using a next generation sequencing platform with the goal of detecting repeats length polymorphisms occurring in single colonies during an overnight growth. Analysis of the sequencing reads confirmed repeat length polymorphisms in 115 of the 277 predicted phase variable genes, distributed in 42 strong (81 %), 31 moderate (52 %) and 42 weak (25 %) putative phase variable genes (Fig. 2 and Additional file 4). The same procedure applied to 100 randomly chosen, non-variable SSRs detected length polymorphisms in 11 loci (11 %, Fig. 2), a significantly lower proportion compared to VNSSRs (chi-squared test for proportions $p \leq 0.01$). Additionally, the average frequencies of polymorphic reads (those containing an SSR showing length polymorphism) observed for VNSSRs was 4.5 %, a higher proportion compared to the 11 false positive SSRs (0.7 %) or to the allowed read mapping error rate (≤0.1 %; Phred-scaled MAPping quality ≥30). Frequency distributions of polymorphic reads identified in the 5 genomes are reported in Additional file 2: Figure S3.

Seventy of the 115 validated VNSSRs were present in all five re-sequenced genomes. Among these, 26 (37 %) were shown to be variable in all the 5 genomes analyzed, while 44 (63 %) were confirmed in four or fewer

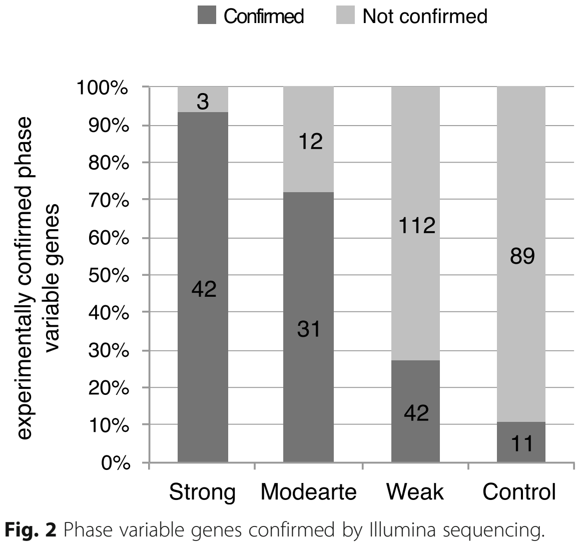


**Fig. 2** Phase variable genes confirmed by Illumina sequencing. Numbers inside bars indicate the frequency of phase variable genes present in each category. Confirmed: predicted phase variable genes confirmed by deep sequencing data. Not confirmed: predicted phase variable genes not showing evidence of variation. Control: analysis of 100 randomly selected SSRs that were predicted as non-variable by the genomic comparative analysis

genomes (Additional file 2: Figure S4A). Interestingly, for any one genome there was a strong association between SSR size (number of unit motif repetitions) and the likelihood of finding it variable. Specifically, homopolymeric tracts of 8 or more nucleotides were predominantly found to be variable in four or five genomes (chi-square $p = 1e\text{-}11$; Additional file 2: Figure S4B). Because there were only three non-homopolymeric tracts, it was not possible to generalize this conclusion to longer repeats. Additionally, for the 70 experimentally confirmed core VNSSRs, the proportion of polymorphic reads was found to correlate (adjusted $R^2 = 0.28$; $p = 1.6e\text{-}6$; Additional file 2: Figure S5A) with the Nei's diversity index derived from the comparison of the 20 available genome sequences. Specifically, VNSSRs associated to a Nei's index >0.5 were mainly characterized by above-average polymorphic reads proportions (chi-square $p = 0.004$; Additional file 2: Figure S5B), meaning that the extent of VNSSRs variability can be predicted with good confidence by the comparison of multiple genomic sequences.

Finally, the number of phase variable genes that would be confirmed by deep sequencing of a wider genome collection was estimated by applying an approach originally proposed for estimating the bacterial pan-genome size [44]. The regression analysis based on the number of validated phase variable genes predicted that if all the 20 genomes were re-sequenced, the number of validated phase-variable genes would have been 146 ± 10 (95 % CI; Additional file 2: Figure S6), while the analysis of 100 genomes could raise the number to 190 ± 23 (95 % CI).

## The pool of genes whose expression is influenced by VNSSRs is wider than previously hypothesized

The pool of putative phase variable genes predicted in this study was compared with the 69 genes previously described or predicted to be phase-variable in *Nm* [22]. Our approach confirmed 24 of the 26 genes whose phase variation had been previously demonstrated in *Nm*, as well as 15 of the 19 putative phase variable genes previously proposed as strong candidates and 8 of the 24 genes previously reported as either moderate or weak candidates (Additional file 2: Figure S7 and Additional file 3). The 22 genes not confirmed by our method, either didn't satisfy the criteria for SSR search or didn't show inter-strain length polymorphisms. In the particular case of a poly-C tract located within the NMB1760 coding region, the repeat showed between-strain variation, alternating between the C5 and C6 states. However, since the C5 was below the minimum length cut off (see materials and methods) our approach failed to identify this variable SSR. Remarkably, the wider genome collection allowed for the identification of 230 new putative phase variable genes, predicted as 19 strong, 50 moderate and 161 weak candidates respectively (Additional file 2: Figure S7 and Additional file 3).

## Candidate phase variable genes related to cell surface structure or involved in DNA metabolism are overrepresented

The 112 strong and moderate candidate genes, described in results section 2, distributed across 15 different functional roles [defined as TIGR main roles [45], Fig. 3].

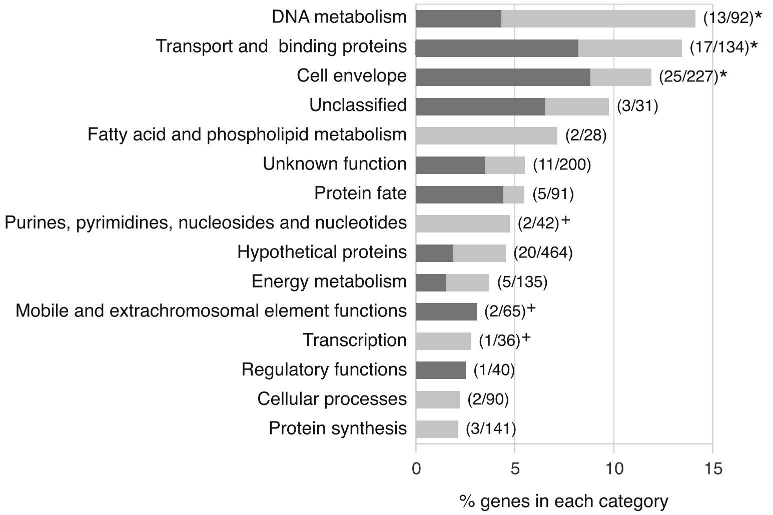


**Fig. 3** TIGR functional roles represented by the 112 strong and moderate candidate phase variable genes. X-axis represents the proportion of putative phase variable genes present in each annotation. Numbers in brackets represent the number of putative phase variable genes over the total number of genes that are associated to a specific function. Genes associated with intragenic and intergenic VNSSRs are represented in dark and light grey, respectively. '*': Enriched functional roles (Bonferroni adjusted $p \leq 0.01$). '+': Functional roles not previously described to be associated with SSR-mediated regulation in *Nm*

Consistent with previous studies [9, 21] VNSSR-associated genes were found to be primarily involved in host-microbial interplay, with "Cell envelope", "Transport and binding" and "DNA metabolism" functional roles being significantly overrepresented. Intragenic and intergenic VNSSRs were both present in the enriched categories. Other biological processes included functions related to the peptide synthesis machinery ("Protein synthesis" and "Protein fate"). Additionally, 5 putative phase variable genes were found in functional categories that had not been previously associated to SSR-mediated regulation, such as genes encoding for transposons and prophage related functions, transcription factors and proteins participating in nucleic acids biosynthesis (Fig. 3).

The 165 weak candidate genes were associated with 18 different functional roles, fourteen of which were common to the strong and moderate groups. This subset showed a different profile, whereby genes were significantly overrepresented in the "hypothetical proteins" and "mobile and extrachromosomal element functions" functional annotations (Additional file 2: Figure S8).

### Most VNSSRs are predicted to induce phenotypic changes by modulating the level of gene expression rather than mediating on/off translational switching

We sought to gather new insights into the VNSSR-mediated regulatory mechanisms by analyzing the repeat sequence context in the 52 strong candidate phase variable genes. Nineteen of the 55 VNSSRs associated with strong candidate genes were surrounded by multiple in-frame 5′-ATG-3′ translational initiation codons (Additional file 5). These VNSSRs can shift the origin of translation between alternative start sites, resulting in the modulation of gene expression [11, 46]. Such modulation can arise either from the switching between start sites associated with different expression levels or from the production of different protein isoforms, as may be the case for the NMB0312 (Fig. 4a) and NMB0415 open reading frames (ORFs) (Fig. 4b), respectively. Different peptide isoforms may also result from VNSSRs located in the 3′ portion of the gene [47]. In this location, the VNSSR can bring the C-terminal peptide portion out of frame while still allowing the translation of a functional protein. Two ORFs, NMB1998 (Fig. 4c) and NMB0039, were consistent with this mechanism. In contrast, VNSSRs present within the NMB1969, NMB1818 and NMB0281 reading frames were either a 3- or a 9-nucleotides repeat. These VNSSRs do not interfere with the reading frame but can still influence the cell phenotype by producing proteins with different structures and/or post-translational profiles (Fig. 4d). Also, 15 of the 55 analyzed VNSSRs were predicted to interfere with the associated gene promoters, as represented by the T-homopolymeric tract located upstream of the NMB0056 start (Fig. 4e). The remaining 16 repeats were predicted to introduce frameshifts likely resulting in on-off type of transcriptional regulation, as proposed for NMB0218 (Fig. 4f).

### Most strong candidate phase variable genes are not specific to *N. meningitidis*

Among the 52 strong candidate phase variable genes, 6 were identified only within *Nm* through sequence database comparison. The GC content of these genes largely deviates from the typical meningococcal value (51-52 %), suggesting acquisition through lateral gene transfer from outside the genus. Conversely, orthologous copies of the remaining 46 genes were also identified in other *Neisseria* species (Additional file 2: Figure S9). Among these, 42 were associated with the same SSRs observed in *Nm*, indicating that these SSR-contingency loci arose before meningococcus speciation. In the remaining four cases (NMB0281, NMB0415, NMB1668 and GNMG2136_1693) the SSR associated with the gene was detected in *Nm* only, suggesting that these contingency loci have evolved following *Nm* speciation (Additional file 3).

Orthologous copies of 8 strong candidate phase variable genes were also identified in species outside of the *Neisseria* genus (Additional file 2: Figure S9), including *Cardiobacterium hominis*, *Haemophilus aegyptius*, *Haemophilus haemolyticus*, *Haemophilus influenzae*, *Kigella Dentrificans* and *Streptococcus pneumoniae*. Among these, the NMB1525 encoded gene identified in the three *Haemophilus* species and the 4 opacity proteins identified in *S. pneumoniae,* were associated with the same SSR as in *Nm* (NMB1636 example shown in Additional file 2: Figure S10), raising the hypothesis that SSR contingency loci are transferred across species through lateral DNA exchange.

### VNSSRs reorganization following *N. meningitidis* speciation primarily involved cell surface determinants

Among the 49 VNSSRs associated to the 46 strong candidate genes identified in multiple taxa, 22 (45 %) were found to be significantly longer (greater median number of motif repetitions; Wilcoxon test $p \leq 0.01$) in *Nm* than in other species, while 3 were found to be significantly shorter (Additional file 3). Twelve and 3 of the repeats with increased length in *Nm* were G/C and A/T homopolymeric tracts, respectively, along with 7 other non-homopolymeric SSRs. These repeats were primarily associated with cell surface determinants, including surface receptors, surface transporters, pili and lipopolysaccharide biosynthetic enzymes. The 3 VNSSRs with a reduced size in *Nm* included a poly-T and a poly-C homopolymeric tracts, respectively associated with a hypothetical protein (NMB1209) and to a ferric enterobactin receptor (NMB1988), and a 5′-CTTCT-3′ repeat associated to an opacity protein (NMB0926). Also in this

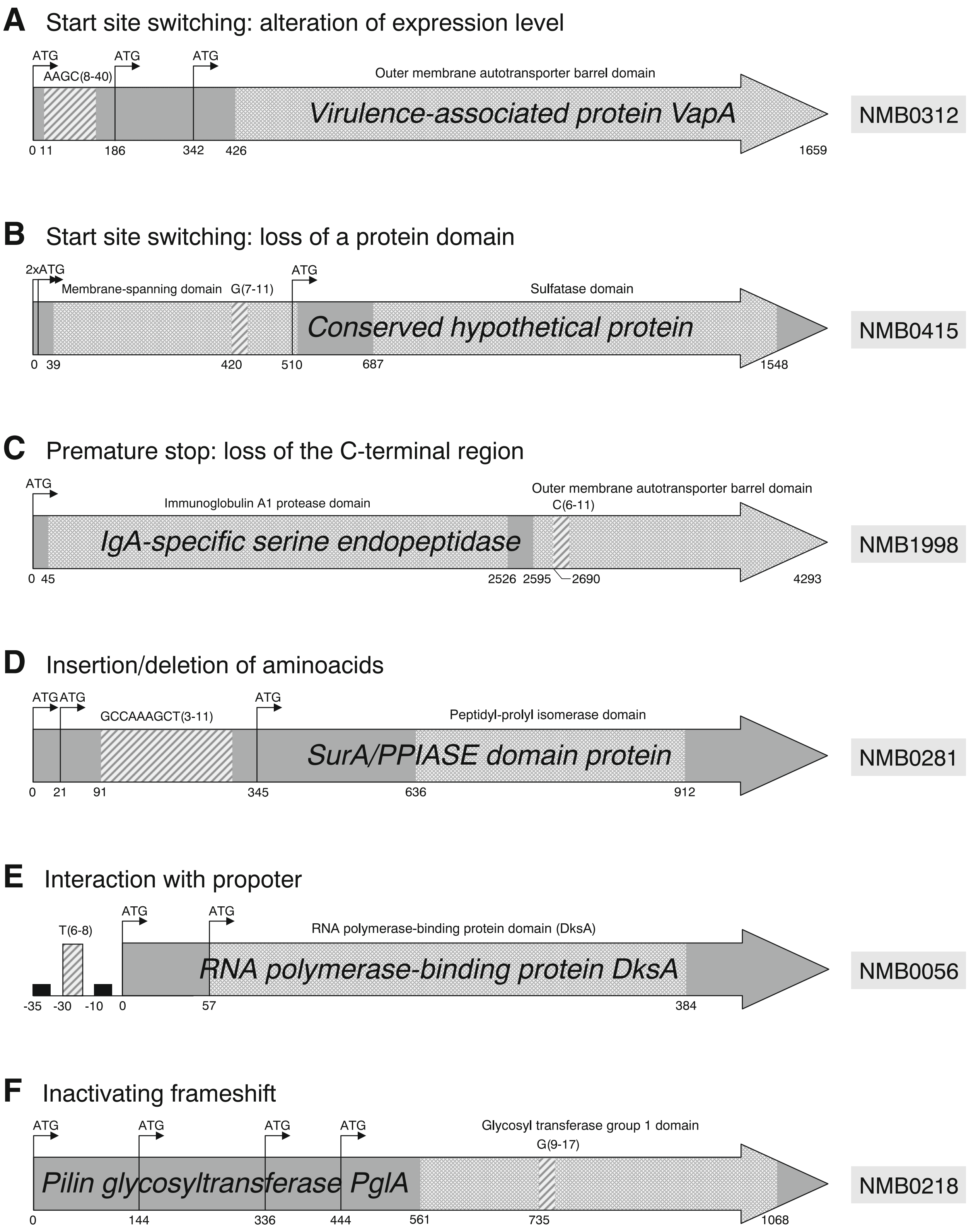

A Start site switching: alteration of expression level
ATG
ATG
ATG
AAGC(8-40)
Outer membrane autotransporter barrel domain
Virulence-associated protein VapA
NMB0312
0 11
186
342
426
1659
B Start site switching: loss of a protein domain
2xATG
ATG
Membrane-spanning domain
G(7-11)
Sulfatase domain
Conserved hypothetical protein
NMB0415
0 39
420
510
687
1548
C Premature stop: loss of the C-terminal region
ATG
Immunoglobulin A1 protease domain
Outer membrane autotransporter barrel domain
C(6-11)
IgA-specific serine endopeptidase
NMB1998
0 45
2526
2595
2690
4293
D Insertion/deletion of aminoacids
ATG ATG
ATG
GCCAAAGCT(3-11)
Peptidyl-prolyl isomerase domain
SurA/PPIASE domain protein
NMB0281
0
21
91
345
636
912
E Interaction with propoter
T(6-8)
ATG
ATG
RNA polymerase-binding protein domain (DksA)
RNA polymerase-binding protein DksA
NMB0056
-35
-30
-10
0
57
384
F Inactivating frameshift
ATG
ATG
ATG
ATG
Glycosyl transferase group 1 domain
G(9-17)
Pilin glycosyltransferase PglA
NMB0218
0
144
336
444
561
735
1068

**Fig. 4** (See legend on next page.)

(See figure on previous page.)
**Fig. 4** Schematic representation of five VNSSRs and their sequence context. **a** VNSSR causing translational start site switching. **b** VNSSR causing the loss of a membrane-spanning domain. **c** VNSSR leading to the loss of the peptide C-terminal region. **d** VNSSR introducing changes in the peptide sequence. E) VNSSR influencing the gene promoter. **f** VNSSR introducing an inactivating frameshift. Dark grey arrows represent open reading frames. Black arrows marked with ATG represent in-frame ATG translational start sites. Light grey boxes represent the annotated functional domains. Stripped boxes represent VNSSRs and the related tags indicate the repeat unit motif along with the minimum and maximum number of repetitions observed in the 20 analyzed genomes. Numbers below each gene indicate the position relative to the annotated translational tart site

case, an association (2 genes out of 3) with surface exposed proteins was observed.

Overall, these results suggest that evasion of host immune responses, along with changes in surface adhesion properties and tissue tropism, have likely been the main drivers of SSR contingency loci evolution within *Nm*.

### VNSSRs chromosomic loci are fixed in the meningococcal population

Chromosomic loci containing the 324 identified VNSSRs were tested for signatures of selective pressures that could confirm a beneficial role (increased fitness) of those loci and justify their over-abundance in *Nm* species.

Only 14 (4.3 %) VNSSR loci, all associated with A/T homopolymeric tracts, showed a significant deviation from neutrality with all the applied tests (Additional file 6). Signatures of positive selection (negative Tajima and Fu-Li statistics) were identified in the upstream region of the two strong candidate phase variable genes *mtrF* (NMB1719), a surface-exposed efflux-pump component involved in hydrophobic antimicrobial resistance [48], and the transcription factor *dksA* (NMB0056). A signature of negative selection (positive Tajima and Fu-Li statistics) was detected in the upstream region of the moderate candidate phase variable gene *sstT* (NMB2133), which also encodes for a surface transporter. All other signals were identified in VNSSRs associated to weak candidate phase variable genes. The remaining 310 (95.7 %) tested loci did not show signatures of selection, supporting the hypothesis that SSR contingency loci have reached an equilibrium state and are now fixed in the population.

## Discussion

After the first *Nm* complete genome sequence became available [49], the highly abundant SSRs were largely associated with genes involved in host adaptation, commensalism and virulence. These initial studies on *Nm* SSRs contingency loci were based on analysis of individual genomes, strains MC58 [9] and Z2491 [20]. With the technology available at that time, the generation of new genomic sequences and the validation of the newly proposed phase-variable genes were laborious and time-consuming tasks. Despite the challenges, a few years after the first meningococcal genome sequence was revealed, a study reported a list of 69 *Nm* phase variable genes, 26 of which were experimentally confirmed [22].

Benefiting from the most recent sequencing technologies, the present study employed a 20 *Nm* genomes dataset, with a resolved phylogenetic structure [36], in a comparative genomic analysis aimed at the study of SSR-driven phase variation as a population-based, as opposed to single cell, adaptive strategy.

All analyzed strains were found to have more SSRs than expected in their chromosomes. A comparable number of SSRs were also found in two non-pathogenic strains, α14 and OX99-30304, questioning the hypothesis that SSRs-mediated phase variation is an adaptation associated with the evolution of virulence. Most of the identified SSRs were homopolymeric tracts, with an excess of A/T tandem repeats, suggesting slipped-strand mispairing as the primary mechanism driving the occurrence of SSRs. It has been proposed that slipped-strand mispairing is preferentially associated with homopolymeric tracts due to their high sequence redundancy [50] and favouring A/T repeats due to the lower stability of the DNA-DNA hybrid during bacterial genome replication [51].

The likelihood that repeat-associated genes undergo phase variation is variable. Efforts to identify molecular signatures indicative of such likelihood have defined a number of *cis* factors, including repeat sequence, size (number of motif repetitions) and sequence context, along with the associated gene function [9, 42]. Particularly, genes encoding for surface antigens were regarded as strong candidates for phase variation. We applied a more unsupervised approach, in which only the repeat sequence context and the distribution of its variation over the sampled phylogeny were considered. The over-representation of A/T repeats and their alleged higher propensity to cause transcriptional slippage compared to G/C repeats [51] would suggest these to be the main contributor of SSR mediated phase variation. Conversely, in agreement with previous findings [42], homopolymeric G/C repeats were the most frequent (50 % of cases) among the strong candidate genes pool, while weak candidate genes were almost exclusively associated with A/T repeats (95 % of cases). A primary involvement of C/G homopolymeric in the mediation of phase variation has also been described in *C. jejuni* [52]. This particular mutational pattern, together with the fact that we didn't identify traces of selective pressure for most VNSSRs, suggest that there may be molecular drivers governing the mutability of these contingency loci. In support of this hypothesis, increased rates of phase variation have been associated with the loss of the *Dam* DNA methylase [53] and the two mismatch repair proteins *mutS* and *mutL* [54, 55].

Regarding non-homopolymeric repeats, 3-, 4-, 5-, 6- and 9-nucelotides VNSSRs were also found to be abundant. Three-, 6- and 9-nucleotides VNSSRs do not disrupt the reading frame and their abundance may be justified by a reduced selective pressure for their stability. Abundance of 5-nucleotide VNSSRs is due to the fact that 6 out of 8 cases are represented by the same 5′-CTTCT-3′ repeat associated to different paralogs of an opacity protein (NMB0442). Interestingly, a homolog of this opacity protein, coupled with the same SSR, was also identified in *Streptococcus pneumoniae* genome (Additional file 2: Figure S9), providing evidence that SSR-contingency loci can be transferred across different species, or strains of the same species and even different chromosomic loci of the same strain through lateral gene transfer. Conversely, abundance of 4-nucleotides VNSSRs, which has been reported both in *Nm* and in *Haemophilus influenzae* [56], has no obvious interpretation. Given the conservation of those repeat sequences and their proximity to the translational origin, the interaction with yet unidentified regulatory elements may be involved. One such regulatory mechanism has been described for the TAAA repeat associated with the *NadA* (NMB1994) promoter [35, 57].

The availability of high-throughput sequencing data enabled a high throughput experimental validation of *in silico* predicted SSR contingency loci. The rationale behind this approach was that DNA libraries used for Illumina sequencing were generated from a bacterial culture derived from a single colony. After an 18 h growth and approximately 36 bacterial replication cycles (assuming a growth rate of 2 duplications per hour), a library was expected to contain, and represent in the form of variable alleles, most of the VNSSRs variation. This test provided circumstantial evidence for phase variability of 93, 72 and 27 % of the strong, moderate and weak candidate genes respectively, leading to a total of 115 experimentally confirmed phase variable genes. This approach also allowed estimating VNSSRs variation frequencies. The 115 validated VNSSR showed an average proportion of alternative alleles of 0.71 %, indicating an average mutation rate of once every seven cell replication cycles. However, given the contribution of environmental, population and molecular factors to the mutability of SSRs in phase variable loci, it is likely that real frequencies may differ from those we observed in vitro.

SSR contingency loci were over-represented among genes related to cell surface functions and structure or involved in DNA metabolism, primarily DNA restriction and modification. These results, which are in agreement with previous findings [9], reflect a primary involvement of SSR-mediated phase variation in microbe-host interaction. Dis-regulation of those genes, in fact, alters *Nm* ability to adhere to and invade host cells, its ability to scavenge required nutrients and to evade host immune defenses by generating antigenic variants [58]. Additionally, putative phase variable genes were found in several other functional classes, indicating that SSR-mediated gene regulation interferes with a set of metabolic functions wider than previously hypothesized.

Our attempt to understand how VNSSRs have evolved in *Nm* revealed that most strong candidate phase variable genes (46/52) are widespread to the *Neisseria* genus and that about half of them have undergone size change after meningococcal speciation. This is suggestive of the fact that most SSR-associated contingency loci have developed prior to *Nm* speciation and have subsequently evolved, possibly as a consequence of the process of adaptation to the human host environment. In agreement with this is the fact that identified phase variable genes were enriched for cell surface determinants, suggesting that evasion of host immune responses and changes in surface adhesion properties and tissue tropism are the main drivers of SSR contingency loci evolution in *Nm*. Regarding the remaining 6 genes that are specific to *Nm*, they were characterized by a G/C content that deviates from that of meningococcus (data not shown), suggesting a possible acquisition through horizontal gene transfer. Phase variable genes candidates were also identified in species outside the *Neisseria* genus (Additional file 2: Figure S9). Interestingly, all of them are human pathogens that coexist in the respiratory flora of healthy individuals. Moreover, we found that genes shared by *Nm*, *S. pneumoniae* and three *Haemophilus* members were associated with the same repeat element as in *Nm*, supporting the possibility that SSR contingency loci are exchanged across species through lateral DNA exchange. Overall, our findings suggest that SSR contingency loci have been acquired by *Nm* either by spontaneous evolution, inheritance from an ancestral species or acquisition by means of lateral DNA exchange.

In the context of a commensal organism, constantly interacting with its host, the onset of new mutations or allele variants may lead to an increased fitness (mutations are favored in the population in a process called positive selection), a reduced fitness (mutations are selected against in a process called negative selection) or have no effect on fitness (mutations are neither selected for nor against and are said to be neutral) [59]. One of the most accredited hypotheses is that VNSSRs have evolved as an adaptive strategy, in organisms like *Nm*, to increase their fitness in the hostile and mutable host environment [7].

We tested all VNSSR-containing loci for signatures of selection that could confirm a beneficial role of those loci and justify their over-abundance. Surprisingly, the analysis revealed that most SSR-contingency loci are not under selective pressure, suggesting that *Nm* has had enough time to adapt to its niche and has apparently reached an equilibrium state. It must be noted, however, that substantial homologous recombination and repeated population expansions and bottlenecks, along with possible sampling biases, are likely to confound or even obscure neutrality tests results. Nonetheless, we envision that, applying this approach to a wider, unbiased strain collection, coming from a single outbreak, will have the potential of providing valuable insights into the understanding of VNSSRs evolutionary processes.

Despite a proportion of coding DNA in *Nm* higher than 80 % [20, 36, 49], we observed that 51 % of the identified VNSSRs are found within intergenic regions. This bias toward intergenic VNSSRs, which may be explained by a reduced selective pressure along non-coding regions, can also be indicative of a regulation of gene expression occurring at the transcriptional level. This type of VNSSRs can indeed modulate promoter efficiency by changing the relative spacing between key promoter elements [57]. Differently, intragenic VNSSRs, which can induce frameshift mutations, are responsible for an on/of type of regulation occurring at the translational level [13]. Additionally, given the existence of alternative translational start codons, intragenic VNSSRs located in the 5′ portion of a gene can switch between multiple active reading frames, with the potential of affecting the level of gene expression [11] or producing alternative peptide isoforms [46]. In the present analysis we observed that most intragenic VNSSRs are located in the proximity of the translational start site and that among the 52 strong candidate phase variable genes, there are 19 cases in which the VNSSR was found between two or more in-frame ATG translational start sites. Twenty other VNSSRs were predicted to either interact with the gene promoter or to produce alternative peptides isoforms. Overall, collected evidences suggest that SSR-contingency loci have evolved in such a way to allow the maximum degree of flexibility, reflected by their ability to finely modulate gene expression level, or the functionality of the resulting protein, rather than merely switching between the two extreme cases of a gene being expressed or silenced through the introduction of frameshifts.

## Conclusions

In conclusion, this analysis, which capitalized on the most recent DNA sequencing technologies, confirmed the fundamental contribution of SSR contingency loci in promoting phenotypic variation and their deep implications in *Nm* survival and adaptation to the surrounding environment. Our unsupervised approach allowed the generation of a panel of 277 genes whose expression may be controlled or influenced by SSR elements and to provide corroborative evidence for the phase variability for 115 of them. Functional characterization of these genes highlighted an enrichment for cell surface determinants, which included members of the adhesins, evasins, lipopolisaccharide biosynthesis and nutrient scavenging protein families. Such proteins have been attributed multiple roles in attachment of bacterial cells to host membranes [60, 61], in regulating bacterial resistance to both innate and adaptive immune system [62–64] and as major determinants of meningococcal invasive disease [33]. SSR-mediated phase variation has also been reported to control the expression of multiple immunogens currently included in commercial meningococcal vaccines [57, 65, 66] and to modulate resistance to antimicrobial agents [67–69]. We therefore envision that future studies on meningococcal microbe-host interaction and studies aimed at the identification of new vaccine antigens and antimicrobial molecules will benefit from this comprehensive characterization of the meningococcal putative phase variable genes repertoire.

## Methods

### Genomic dataset

This analysis was based on a dataset previously described [36]. The 20 genome sequences available were derived from isolates belonging to phylogenetic clades PC32/269 (MC58 [AE002098.2], H44/76 [CP002420], CU385 [AEQJ 00000000], M04-240196 [CP002423], M01-240013 [AEQ L00000000] and M13399 [AEQG00000000]), PC8/11 (G 2136 [CP002419], 961-5945 [AEQK00000000], FAM18 [A M421808], M6190 [AEQF00000000] and ES14902 [AEQ I00000000]) and PC41-44 (OX99-30304 [AEQE000000 00], M0579 [AEQH00000000], NZ-05/33 [CP002424] and M01-240149 [CP002421]). Five other sequences were derived from isolates belonging to the following MLST groups: CC4281 (053442 [NC_010120.1]), CC53 (α14 [A M889136]), CC213 (M01-240355 [CP002422]), CC4 (Z24 91 [AL157959]) and ST751 (N1568 [AEQD00000000]). Genome sequences are available online from the GenBank database through the reported accession numbers.

### SSRs identification, clustering and comparison

Tandemly repeated motifs, from 1 to 10 nucleotides in length, were identified in each genome using a Perl program developed *ad hoc*. Based on former repeat-associated phase variable genes investigations [9, 17], size cut-offs applied to the repeat search were as follows: 6 repetitions for homopolymeric tracts, 5 for 2-nucleotides, 4 for 3-nucleotides, 3 for 4-nucleotides, 4 for 5-nucleotides and 3 for 6- to 10-nucleotides repeats. The program allowed for the identification of imperfect simple sequence repeats (parameters applied are reported and described in Additional file 7).

In order to establish clusters of orthologous SSRs, all the identified repeats, along with their 50-nucleotides 5′ and 3′ flanking regions, were annotated as features on the respective genome sequences following the GenBank flat file format. All genomes were aligned with the ProgressiveMauve multiple alignment algorithm implemented in the MAUVE v2.3.1 toolkit using the seed--family option [70]. SSRs present in syntenic chromosomic regions, sharing the same repeat sequence and a minimum of 70 % sequence identity over 60 % coverage were identified as orthologous features.

For each cluster of orthologous SSRs, SSR sequences, along with their flanking 50 nucleotides, were aligned using the MUSCLE aligner [71]. Each alignment was then analyzed in order to identify SSR length polymorphisms across different genomes. This resulted in 492 clusters of orthologous SSRs showing evidence of inter-strain length polymorphisms. For each variable number simple sequence repeat (VNSSR) the degree of size heterogeneity was assessed by the Nei's diversity index [DI = 1-$\sum$(allele frequency)$^2$] [41].

### Putative phase variable genes identification

Identity, distance and orientation of genes containing, or located next to VNSSRs were extracted using a Perl program developed *ad hoc*. One hundred and fifty-eight (158) VNSSRs, located within 200 nucleotides preceding the associated genes translational start site, and 166 VNSSRs, located within the genes coding region, were selected for further analyses.

Classification into weak, moderate and strong candidate phase variable genes was based on three parameters: *i*) degree of SSRs overrepresentation compared to a neutral expectation [calculated using a previously described approach [72]], *ii*) the range of VNSSR length variation (difference between the longest and the shortest repeat across multiple genomes) and *iii*) the distribution of the variation over the meningococcal phylogeny, as defined in [36]. Homopolymeric VNSSRs with a length variation range of a single unit motif and not being over-represented (<9 repetitions for A/T tracts and <7 repetitions for G/C tracts) were classified as weak candidates. VNSSRs characterized by a length variation range greater than one repeat unit and showing length variation among members of the same clonal complex in at least two clonal complexes were classified as strong candidates. Remaining VNSSRs were classified as moderate candidates.

The sequence context of the 55 VNSSRs associated with the 52 strong candidate genes was further analysed in order to elucidate possible implications of VNSSRs variation on the associated gene expression. The relative position of the repeats, in-frame ATG translational start-sites and annotated peptide functional domains were extracted (Additional file 5) and manually inspected. The identity and location of peptides functional domains were predicted by HMM motif searches on Pfam [73] and TIGRfam [74] databases.

### Identification of functional roles associated to SSR contingency loci

TIGR main functional roles [45] associated with meningococcal putative phase variable genes were tested for enrichment using the chi-squared test followed by Bonferroni correction for multiple testing. Functional categories associated with a corrected $p$ value $\leq 0.01$ were assumed to be enriched for VNSSR-associated genes.

### Illumina sequencing and data analysis

Genomic sequences of strains G2136, M01-240355, M04-240196, MC58 and NZ-05/33 were re-sequenced using the Illumina HiSeq2000 platform. For each strain, the stocked inoculum was streaked onto agar-chocolate plates and grown overnight (18 h) at 37°. About 30 colonies were harvested into phosphate buffered saline and further processed for phenol-chloroform DNA extraction. Purified DNA was fragmented using the Covaris M220 Focused-ultrasonicator™ and further processed using the Beckman Coulter SPRI-TE™ instrument following the manufacturer's instructions. Clusters of flowcells were generated via the Illumina cBot cluster amplification system and the TruSeq PE Cluster Kit (v.2). Sequencing was carried out on an Illumina HighSeq2000 sequencer following manufacturer guidelines [75].

The 66 nucleotide paired-reads coming from the sequencer were mapped over the corresponding reference genomes using the BWA toolkit v0.5.9 [76] with parameters q = 30 (quality cut-off for read trimming), e = 5 (maximum number of gap extensions), O = 8 (gap-open penalty) and E = 3 (gap-extension penalty). Duplicate reads were filtered out using Picard v1.58. Using this approach it was possible to detect length polymorphisms of SSRs with motifs up to 5-nucleotides long. Reads mapping over the chromosomic regions containing the predicted VNSSRs, spanning the entire repetitive sequence plus at least 4 flanking nucleotides on both 5′ and 3′ extremities, and having a Phred-scaled MAPping quality ≥30 were extracted for each VNSSR and analyzed for SSR length polymorphisms. SSRs showing a proportion of alternative alleles >0.002 were considered to be variable. Due to limitations of the reads mapping procedure, it was not possible to test repeats with a unit motif longer than 5 nucleotides or repeats found inside paralogous chromosomic regions. Finally, it was not possible to test those repeats whose overall length cannot be entirely covered by a single read.

Possible contribution of sequencing errors to the detection of false positive VNSSRs was evaluated by applying the same procedure to 100 randomly selected SSRs that showed no length polymorphisms in the 20-genome comparative analysis.

The regression analysis for new validated phase variable genes with an increasing number of tested genomes was performed as previously described [44]: the Heaps' power law function $n = kN^{\gamma}$ was fitted to the data (for $N > 1$) with a least squares regression, where $n$ is the number of validated phase variable genes, $N$ is the number of genomes and $k$ and $\gamma$ are free parameters.

### Identification of putative phase variable genes across species

In order to assess whether the 52 putative phase variable genes were present in species other than *Nm*, their DNA sequence was searched against the NCBI Reference Sequence [77] and the Whole Genome Shotgun [78] databases using BLAST (BLASTN v.2.2) [79]. Matches with at least 80 % sequence identity over 50 % coverage of the query were extracted and further analyzed.

A similar approach was also used for the analysis of the evolution of VNSSRs associated to strong candidate genes. Orthologous VNSSRs present in different species were extracted and compared through the Wilcoxon test. VNSSRs having a greater or smaller median size (number of motif repetitions) and a difference-associated $p$ value ≤ 0.01 were assumed to have grown or shrunk in meningococcus, respectively.

### Neutrality tests on VNSSR-containing loci

Each cluster of orthologous VNSSRs was extracted, along with their flanking 50 nucleotides, and aligned using the MUSCLE aligner [71]. Each alignment was subsequently tested for signatures of selective pressure using the VariScan software package [80] with the following parameters: StartPos = 1, EndPos = 0, RefPos = 1, Outgroup = none, BlockDataFile = none, RefSeq = 1, RunMode = 12, UseMuts = 1, FixNum = 0, NumNuc = 4 and SlidingWindow = 0. This configuration analyzes the entire alignment region and computes the Tajima D, Fu-Li D* and Fu-Li F* neutrality tests statistics [81, 82]. VNSSRs loci showing deviation from neutral evolution ($p \leq 0.05$) with all 3 tests were considered to be under selective pressure.

## Additional files

**Additional file 1:** Number of SSRs identified in the 20 Neisseria meningitidis genomes collection. (XLSX 37 kb)

**Additional file 2: Figures S1-S10.** with their legends. (PDF 500 kb)

**Additional file 3:** List of putative VNSSRs identified in the 20 *Neisseria meningitidis* genomes collection. (XLSX 48 kb)

**Additional file 4:** VNSSRs validation by Illumina sequencing. (XLSX 108 kb)

**Additional file 5:** Sequence context of the 55 VNSSRs associated to strong candidate genes. (XLSX 58 kb)

**Additional file 6:** Neutrality test results for the 324 VNSSRs-containing loci. (XLSX 63 kb)

**Additional file 7:** Parameters applied for SSR identification. (XLSX 32 kb)

### Abbreviations

A/T: Adenine or thymine nucleotide; BLAST: Basic local alignment search tool; CC: Clonal complex; DI: Nei's diversity index; G/C: Cytosine or guanine nucleotide; HMM: Hidden Markov model; MLST: Multilocus sequence typing; *Nm*: *Neisseria meningitidis*; ORF: Open reading frame; SSR: Simple sequence repeat; TIGR: The Institute for Genomic Research; VNSSR: Variable number simple sequence repeats

### Acknowledgements

We thank Tim Perkins and Nicola Pacchiani for the advice on Illumina sequencing data analysis; Claudio Donati and Stefano Censini for their participation in the generation of Illumina sequencing data; Isabel Delany for the scientific advise.

### Funding

ES, SG, GT and DM are employees of GSK Vaccines. DR and TH were affiliated to the Institute for Genome Sciences at the time the study was carried out. No additional funding was involved.

### Availability of data and material

All public data are available through the references provided in the methods section. Illumina sequencing data are available in the BioProject database using the following link http://www.ncbi.nlm.nih.gov/bioproject/PRJNA341805. Scripts developed for simple sequence repeats identification and analysis are available upon request to the author.

### Authors' contributions

ES carried out the genome comparative analysis, participated in the study design and drafted the manuscript. RD, SG and GT worked at the Illumina sequencing data generation and data preprocessing. TH and DR carried out the gene and protein domains functional annotation. ERM participated in the study design. DM participated in the design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

### Competing interests

ES, SG, GT and DM are employees of GSK Vaccines. TH reports grants from Chiron Vaccines / Novartis Vaccines, during the conduct of the study. TH has a patent 20050191316 issued. ES, DM, TH and DR have a patent 20120070457 issued. The other authors declare that they have no competing interests.

### Consent for publication

Not applicable since the study does not include any individual person's data.

### Ethics approval and consent to participate

The present study did not involve any human or animal related data requiring ethical approval.

### Author details

[1]GSK Vaccines, 53100 Siena, Italy. [2]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201, USA. [3]Medical Sciences Division, Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, Oxford OX3 9DS, UK. [4]Present address: Institute of Informatics and Telematics and Institute of Clinical Physiology, National Research Council, 56124 Pisa, Italy. [5]Present address: Personal Genome Disgnostics inc., Baltimore, MD 21224, USA.

Received: 22 April 2016 Accepted: 22 October 2016
Published online: 28 October 2016

### References

1. Stephens DS, Hoffman LH, McGee ZA. Interaction of Neisseria meningitidis with human nasopharyngeal mucosa: attachment and entry into columnar epithelial cells. J Infect Dis. 1983;148:369–76.
2. Rosenstein NE, Perkins BA, Stephens DS, Popovic T, Hughes JM. Meningococcal disease. N Engl J Med. 2001;344:1378–88.

3. Correia JB, Hart CA. Meningococcal disease. N Engl J Med [Internet]. 2001;345:699. Available from: http://www.nejm.org/doi/full/10.1056/NEJM200105033441807.
4. Hart CA, Rogers T. Meningococcal disease. J Exp Med. 1993;39:3–25.
5. Peltola H. Meningococcal disease: still with us. Rev Infect Dis [Internet]. 1983;5:71–91. Available from: http://www.ncbi.nlm.nih.gov/pubmed/6338571.
6. Raman GV. Meningococcal septicaemia and meningitis: a rising tide. Br Med J (Clin Res Ed) [Internet]. 1988;296:1141–2. [cited 2014 Aug 5] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2545613&tool=pmcentrez&rendertype=abstract.
7. Moxon ER, Rainey PB, Nowak MA, Lenski RE. Adaptive evolution of highly mutable loci in pathogenic bacteria. Curr Biol. 1994;4:24–33.
8. Brunham RC, Plummer FA, Stephens RS. Bacterial antigenic variation, host immune response, and pathogen-host coevolution. Infect Immun [Internet]. 1993;61:2273–6. [cited 2014 Aug 6] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=280844&tool=pmcentrez&rendertype=abstract.
9. Saunders NJ, Jeffries AC, Peden JF, Hood DW, Tettelin H, Rappuoli R, et al. Repeat-associated phase variable genes in the complete genome sequence of Neisseria meningitidis strain MC58. Mol Microbiol. 2000;37(1):207–15.
10. Moxon R, Bayliss C, Hood D. Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. Annu Rev Genet [Internet]. 2006;40:307–33. Available from: http://www.annualreviews.org/doi/abs/10.1146/annurev.genet.40.110405.090442.
11. Dixon K, Bayliss CD, Makepeace K, Moxon ER, Hood DW. Identification of the functional initiation codons of a phase-variable gene of Haemophilus influenzae, lic2A, with the potential for differential expression. J Bacteriol. 2007;189:511–21.
12. De Bolle X, Bayliss CD, Field D, Van De Ven T, Saunders NJ, Hood DW, et al. The length of a tetranucleotide repeat tract in Haemophilus influenzae determines the phase variation rate of a gene with homology to type III DNA methyltransferases. Mol Microbiol. 2000;35:211–22.
13. Jonsson AB, Nyberg G, Normark S. Phase variation of gonococcal pili by frameshift mutation in pilC, a novel gene for pilus assembly. EMBO J. 1991;10:477–88.
14. Park SF, Purdy D, Leach S. Localized reversible frameshift mutation in the flhA gene confers phase variability to flagellin gene expression in Campylobacter coli. J Bacteriol. 2000;182:207–10.
15. Stibitz S, Aaronson W, Monack D, Falkow S. Phase variation in Bordetella pertussis by frameshift mutation in a gene for a novel two-component system. Nature. 1989;338:266–9.
16. Srikhanta YN, Maguire TL, Stacey KJ, Grimmond SM, Jennings MP. The phasevarion: a genetic system controlling coordinated, random switching of expression of multiple genes. Proc Natl Acad Sci U S A. 2005;102:5547–51.
17. Martin P, Van De Ven T, Mouchel N, Jeffries AC, Hood DW, Moxon ER. Experimentally revised repertoire of putative contingency loci in Neisseria meningitidis strain MC58: Evidence for a novel mechanism of phase variation. Mol Microbiol. 2003;50:245–57.
18. Van Ham SM, Van Alphen L, Mooi FR, Van Putten JPM. Phase variation of H. influenzae fimbriae: Transcriptional control of two divergent genes through a variable combined promoter region. Cell. 1993;73:1187–96.
19. Winner F, Markovà I, Much P, Lugmair A, Siebert-Gulle K, Vogl G, et al. Phenotypic switching in Mycoplasma gallisepticum hemadsorption is governed by a high-frequency, reversible point mutation. Infect Immun. 2003;71:1265–73.
20. Parkhill J, Achtman M, James KD, Bentley SD, Churcher C, Klee SR, et al. Complete DNA sequence of a serogroup A strain of Neisseria meningitidis Z2491. Nature. 2000;404:502–6.
21. Snyder LA, Butcher SA, Saunders NJ. Comparative whole-genome analyses reveal over 100 putative phase-variable genes in the pathogenic Neisseria spp. Microbiology. 2001;147:2321–32.
22. Bentley SD, Vernikos GS, Snyder LAS, Churcher C, Arrowsmith C, Chillingworth T, et al. Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18. PLoS Genet. 2007;3:0230–40.
23. Stern A, Brown M, Nickel P, Meyer TF. Opacity genes in Neisseria gonorrhoeae: control of phase and antigenic variation. Cell. 1986;47:61–71.
24. van der Ende A, Hopman CT, Zaat S, Essink BB, Berkhout B, Dankert J. Variable expression of class 1 outer membrane protein in Neisseria meningitidis is caused by variation in the spacing between the -10 and -35 regions of the promoter. J Bacteriol. 1995;177:2475–80.
25. Richardson AR, Stojiljkovic I. HmbR, a hemoglobin-binding outer membrane protein of Neisseria meningitidis, undergoes phase variation. J Bacteriol. 1999;181:2067–74.
26. Banerjee A, Wang R, Supernavage SL, Ghosh SK, Parker J, Ganesh NF, et al. Implications of phase variation of a gene (pgtA) encoding a pilin galactosyl transferase in gonococcal pathogenesis. J Exp Med. 2002;196:147–62.
27. Power PM, Roddam LF, Rutter K, Fitzpatrick SZ, Srikhanta YN, Jennings MP. Genetic characterization of pilin glycosylation and phase variation in Neisseria meningitidis. Mol Microbiol. 2003;49:833–47.
28. Warren MJ, Jennings MP. Identification and characterization of pptA: a gene involved in the phase-variable expression of phosphorylcholine on pili of Neisseria meningitidis. Infect Immun. 2003;71:6892–8.
29. Yang QL, Gotschlich EC. Variation of gonococcal lipooligosaccharide structure is due to alterations in poly-G tracts in lgt genes encoding glycosyl transferases. J Exp Med. 1996;183:323–7.
30. Banerjee A, Wang R, Uljon SN, Rice PA, Gotschlich EC, Stein DC. Identification of the gene (lgtG) encoding the lipooligosaccharide beta chain synthesizing glucosyl transferase from Neisseria gonorrhoeae. Proc Natl Acad Sci U S A. 1998;95:10872–7.
31. Chen CJ, Elkins C, Sparling PF. Phase variation of hemoglobin utilization in Neisseria gonorrhoeae. Infect Immun. 1998;66:987–93.
32. Sarkari J, Pandit N, Moxon ER, Achtman M. Variable expression of the Opc outer membrane protein in Neisseria meningitidis is caused by size variation of a promoter containing poly-cytidine. Mol Microbiol. 1994;13:207–17.
33. Hammerschmidt S, Müller A, Sillmann H, Mühlenhoff M, Borrow R, Fox A, et al. Capsule phase variation in Neisseria meningitidis serogroup B by slipped-strand mispairing in the polysialyltransferase gene (siaD): Correlation with bacterial invasion and the outbreak of meningococcal disease. Mol Microbiol. 1996;20:1211–20.
34. Carson SDB, Stone B, Beucher M, Fu J, Sparling PF. Phase variation of the gonococcal siderophore receptor FetA. Mol Microbiol. 2000;36:585–93.
35. Martin P, Makepeace K, Hill SA, Hood DW, Moxon ER. Microsatellite instability regulates transcription factor binding and gene expression. Proc Natl Acad Sci U S A. 2005;102:3800–4.
36. Budroni S, Siena E, Hotopp JCD, Seib KL, Serruto D, Nofroni C, et al. Neisseria meningitidis is structured in clades associated with restriction modification systems that modulate homologous recombination. Proc Natl Acad Sci U S A. 2011;108:4494–9.
37. Jerome JP, Bell JA, Plovanich-Jones AE, Barrick JE, Brown CT, Mansfield LS. Standing genetic variation in contingency loci drives the rapid adaptation of Campylobacter jejuni to a novel host. PLoS One. 2011;6.
38. Herring CD, Raghunathan A, Honisch C, Patel T, Applebee MK, Joyce AR, et al. Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale. Nat Genet. 2006;38:1406–12.
39. Velicer GJ, Raddatz G, Keller H, Deiss S, Lanz C, Dinkelacker I, et al. Comprehensive mutation identification in an evolved bacterial cooperator and its cheating ancestor. Proc Natl Acad Sci U S A. 2006;103:8107–12.
40. Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, et al. Genome evolution and adaptation in a long-term experiment with Escherichia coli. Nature. 2009;461:1243–7.
41. Nei M. Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci U S A. 1973;70:3321–3.
42. Saunders NJ, Peden JF, Hood DW, Moxon ER. Simple sequence repeats in the Helicobacter pylori genome. Mol Microbiol [Internet]. 1998;27:1091–8. Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=9570395.
43. Tran HT, Keen JD, Kricker M, Resnick MA, Gordenin DA. Hypermutability of homonucleotide runs in mismatch repair and DNA polymerase proofreading yeast mutants. Mol Cell Biol. 1997;17:2859–65.
44. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol. 2008;11(5):472–7.
45. Peterson JD. The comprehensive microbial resource. Nucleic Acids Res [Internet]. 2001;29:123–5. [cited 2014 Aug 6] Available from: http://nar.oxfordjournals.org/content/29/1/123.
46. Schweda EKH, Richards JC, Hood DW, Moxon ER. Expression and structural diversity of the lipopolysaccharide of Haemophilus influenzae: Implication in virulence. Int J Med Microbiol. 2007;297(5):297–306.
47. Vakhrusheva AA, Kazanov MD, Mironov AA, Bazykin GA. Evolution of prokaryotic genes by shift of stop codons. J Mol Evol. 2011;72:138–46.

48. Veal WL, Shafer WM. Identification of a cell envelope protein (MtrF) involved in hydrophobic antimicrobial resistance in Neisseria gonorrhoeae. J Antimicrob Chemother. 2003;51:27–37.
49. Tettelin H, Saunders NJ, Heidelberg J, Jeffries AC, Nelson KE, Eisen JA, et al. Complete genome sequence of Neisseria meningitidis serogroup B strain MC58. Science. 2000;287:1809–15.
50. Levinson G, Gutman GA. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol [Internet]. 1987;4:203–21. [cited 2015 Jun 12] Available from: http://www.ncbi.nlm.nih.gov/pubmed/3328815.
51. Wagner LA, Weiss RB, Driscoll R, Dunn DS, Gesteland RF. Transcriptional slippage occurs during elongation at runs of adenine or thymine in Escherichia coli. Nucleic Acids Res. 1990;18:3529–35.
52. Bayliss CD, Bidmos FA, Anjum A, Manchev VT, Richards RL, Grossier J-P, et al. Phase variable genes of Campylobacter jejuni exhibit high mutation rates and specific mutational patterns but mutability is not the major determinant of population structure during host colonization. Nucleic Acids Res [Internet]. 2012;40:5876–89. [cited 2014 Aug 6] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3401435&tool=pmcentrez&rendertype=abstract.
53. Bucci C, Lavitola A, Salvatore P, Del Giudice L, Massardo DR, Bruni CB, et al. Hypermutation in pathogenic bacteria: frequent phase variation in meningococci is a phenotypic trait of a specialized mutator biotype. Mol Cell [Internet]. 1999;3:435–45. [cited 2014 Aug 6] Available from: http://www.ncbi.nlm.nih.gov/pubmed/10230396.
54. Richardson AR, Stojiljkovic I. Mismatch repair and the regulation of phase variation in Neisseria meningitidis. Mol Microbiol [Internet]. 2001;40:645–55. [cited 2014 Aug 6] Available from: http://www.ncbi.nlm.nih.gov/pubmed/11359570.
55. Richardson AR, Yu Z, Popovic T, Stojiljkovic I. Mutator clones of Neisseria meningitidis in epidemic serogroup A disease. Proc Natl Acad Sci U S A. 2002;99:6103–7.
56. Hood DW, Deadman ME, Jennings MP, Bisercic M, Fleischmann RD, Venter JC, et al. DNA repeats identify novel virulence genes in Haemophilus influenzae. Proc Natl Acad Sci U S A. 1996;93:11121–5.
57. Metruccio MME, Pigozzi E, Roncarati D, Scorza FB, Norais N, Hill SA, et al. A novel phase variation mechanism in the meningococcus driven by a ligand-responsive repressor and differential spacing of distal promoter elements. PLoS Pathog. 2009;5.
58. Bayliss CD, Field D, Moxon ER. The simple sequence contingency loci of Haemophilus influenzae and Neisseria meningitidis. J Clin Invest [Internet]. 2001;107:657–62. [cited 2014 Aug 5] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=208953&tool=pmcentrez&rendertype=abstract.
59. Nielsen R. Molecular signatures of natural selection. Annu Rev Genet. 2005;39:197–218.
60. Dehio C, Gray-Owen SD, Meyer TF. The role of neisserial Opa proteins in interactions with host cells. Trends Microbiol. 1998;6:489–95.
61. Takahashi H, Carlson RW, Muszynski A, Choudhury B, Kim KS, Stephens DS, et al. Modification of lipooligosaccharide with phosphoethanolamine by LptA in Neisseria meningitidis enhances meningococcal adhesion to human endothelial and epithelial cells. Infect Immun [Internet]. 2008;76:5777–89. [cited 2016 Jul 21] Available from: http://www.ncbi.nlm.nih.gov/pubmed/18824535.
62. McNeil LK, Zagursky RJ, Lin SL, Murphy E, Zlotnick GW, Hoiseth SK, et al. Role of factor H binding protein in Neisseria meningitidis virulence and its potential as a vaccine candidate to broadly protect against meningococcal disease. Microbiol Mol Biol Rev. [Internet]. American Society for Microbiology (ASM). 2013;77:234–52. [cited 2016 Jul 21] Available from: http://www.ncbi.nlm.nih.gov/pubmed/23699256
63. Kugelberg E, Gollan B, Tang CM. Mechanisms in Neisseria meningitidis for resistance against complement-mediated killing. Vaccine [Internet]. Elsevier. 2008;26 Suppl 8:I34–9. [cited 2016 Jul 21] Available from: http://www.ncbi.nlm.nih.gov/pubmed/19388162
64. Del Tordello E, Vacca I, Ram S, Rappuoli R, Serruto D. Neisseria meningitidis NalP cleaves human complement C3, facilitating degradation of C3b and survival in human serum. Proc. Natl. Acad. Sci. U. S. A. [Internet]. National Academy of Sciences. 2014;111:427–32. [cited 2016 Jul 21] Available from: http://www.ncbi.nlm.nih.gov/pubmed/24367091
65. Tauseef I, Ali YM, Bayliss CD. Phase variation of PorA, a major outer membrane protein, mediates escape of bactericidal antibodies by Neisseria meningitidis. Infect Immun [Internet]. 2013;81:1374–80. [cited 2016 Jul 21] Available from: http://www.ncbi.nlm.nih.gov/pubmed/23403557.
66. Serruto D, Spadafina T, Ciucchi L, Lewis LA, Ram S, Tontini M, et al. Neisseria meningitidis GNA2132, a heparin-binding protein that induces protective immunity in humans. Proc. Natl. Acad. Sci. [Internet]. National Academy of Sciences. 2010;107:3770–5. [cited 2016 Jul 21] Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0915162107
67. Peak IR, Jennings CD, Jen FE-C, Jennings MP. Role of Neisseria meningitidis PorA and PorB Expression in Antimicrobial Susceptibility. Antimicrob. Agents Chemother. [Internet]. American Society for Microbiology. 2014;58:614–6. [cited 2016 Jul 21] Available from: http://aac.asm.org/cgi/doi/10.1128/AAC.02506-12
68. Kandler JL, Joseph SJ, Balthazar JT, Dhulipala V, Read TD, Jerse AE, et al. Phase-variable expression of lptA modulates the resistance of Neisseria gonorrhoeae to cationic antimicrobial peptides. Antimicrob. Agents Chemother. [Internet]. American Society for Microbiology (ASM). 2014;58:4230–3. [cited 2016 Jul 21] Available from: http://www.ncbi.nlm.nih.gov/pubmed/24820072
69. Jen FE-C, Seib KL, Jennings MP. Phasevarions mediate epigenetic regulation of antimicrobial susceptibility in Neisseria meningitidis. Antimicrob. Agents Chemother. [Internet]. American Society for Microbiology (ASM). 2014;58:4219–21. [cited 2016 Jul 21] Available from: http://www.ncbi.nlm.nih.gov/pubmed/24777094
70. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004;14:1394–403.
71. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res [Internet]. 2004;32:1792–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15034147.
72. Schbath S. An efficient statistic to detect over- and under-represented words in DNA sequences. J Comput Biol. 1997;4:189–92.
73. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res [Internet]. 2014;42:D222–30. [cited 2014 Jul 13] Available from: http://nar.oxfordjournals.org/content/42/D1/D222.long.
74. Haft DH. The TIGRFAMs database of protein families. Nucleic Acids Res [Internet]. 2003;31:371–3. [cited 2015 Sep 23] Available from: http://nar.oxfordjournals.org/content/31/1/371.full.
75. Quail MA, Swerdlow H, Turner DJ. Improved protocols for the illumina genome analyzer sequencing system. Curr. Protoc. Hum. Genet. [Internet]. 2009;Chapter 18:Unit 18.2. [cited 2014 Aug 6] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3849550&tool=pmcentrez&rendertype=abstract
76. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics [Internet]. 2009;25:1754–60. [cited 2014 Jul 9] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2705234&tool=pmcentrez&rendertype=abstract.
77. Pruitt K, Brown G, Tatusova T, Maglott D. The Reference Sequence (RefSeq) Database [Internet]. National Center for Biotechnology Information (US); 2012. [cited 2015 Oct 1] Available from: http://www.ncbi.nlm.nih.gov/books/NBK21091/
78. About WGS [Internet]. [cited 2015 Oct 1] Available from: https://www.ncbi.nlm.nih.gov/genbank/wgs
79. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol [Internet]. 1990;215:403–10. [cited 2014 Jul 10] Available from: http://www.ncbi.nlm.nih.gov/pubmed/2231712.
80. Hutter S, Vilella AJ, Rozas J. Genome-wide DNA polymorphism analyses using VariScan. BMC Bioinformatics. 2006;7:409.
81. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics [Internet]. 1989;123:585–95. [cited 2014 Jul 19] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1203831&tool=pmcentrez&rendertype=abstract.
82. Fu YX, Li WH. Statistical tests of neutrality of mutations. Genetics. 1993;133:693–709.