Available online at www.sciencedirect.com
ScienceDirect
www.sciencedirect.com/science/journal/16720229

GENOMICS
PROTEOMICS &
BIOINFORMATICS

Article

A Novel Interpretation of Structural Dot Plots of Genomes Derived from the Analysis of Two Strains of *Neisseria meningitidis*

Wilfred R. Cuff[1*], Venkata R.S.K. Duvvuri[2,3], Binhua Liang[4], Bhargavi Duvvuri[5], Gillian E. Wu[3,5], Jianhong Wu[2,6], and Raymond S.W. Tsang[1]

[1]*Public Health Agency of Canada, Canadian Science Centre for Human and Animal Health, Winnipeg, Canada;*
[2]*Center for Disease Modelling, York Institute for Health Research, York University, Toronto, Canada;*
[3]*Department of Biology, York University, Toronto, Canada;*
[4]*Department of Medical Microbiology, University of Manitoba, Winnipeg, Canada;*
[5]*School of Kinesiology and Health Sciences, York University, Toronto, Canada;*
[6]*Department of Mathematics and Statistics, York University, Toronto, Canada.*

Genomics Proteomics Bioinformatics 2010 Sep; 8(3): 159-169 DOI: 10.1016/S1672-0229(10)60018-6

Abstract

Neisseria meningitidis is the agent of invasive meningococcal disease, including cerebral meningitis and septicemia. Because the diseases caused by different clonal groups (sequence types) have their own epidemiological characteristics, it is important to understand the differences among the genomes of the *N. meningitidis* clonal groups. To this end, a novel interpretation of a structural dot plot of genomes was devised and applied; exact nucleotide matches between the genomes of *N. meningitidis* serogroup A strain Z2491 and serogroup B strain MC58 were identified, leading to the specification of various structural regions. Known and putative virulence genes for each *N. meningitidis* strain were then classified into these regions. We found that virulence genes of MC58 tend more to the translocated regions (chromosomal segments in new sequence contexts) than do those of Z2491, notably tending towards the interface between one of the translocated regions and the collinear region. Within the collinear region, virulence genes tend to occur within 16 kb of gaps in the exact matches. Verification of these tendencies using genes clustered in the cps locus was sufficiently supportive to suggest that these tendencies can be used to focus the search for and understanding of virulence genes and mechanisms of pathogenicity in these two organisms.

Key words: structural dot plots, virulence genes, translocation, match regions, MUMmer, *Neisseria meningitidis*

Introduction

Neisseria meningitidis is the agent of invasive meningococcal disease including cerebral meningitis and septicemia, or focal disease involving the joints or other organs. The overall mortality rate from meningococcal meningitis is 5%-10% in industrial countries (http://www.who.int/immunization/topics/meningitis/en/index.html) but it can reach up to 20%-30% (*1*).

N. meningitidis is divided into at least 12 serogroups, depending upon the characteristics of the polysaccharide capsule, and more than 20 serotypes

*Corresponding author.
E-mail: Wilfred_Cuff@phac-aspc.gc.ca

and serosubtypes, depending upon antigens expressed on two sets of outer membrane proteins (*2, 3*). Yet most of the epidemic disease worldwide is caused by a handful of clonal groups designated as hypervirulent clones (*4*) and defined as sets of sequence types (*5*). Diseases caused by the different hypervirulent clones have their own epidemiological characteristics and hence it is important to understand the differences among the genomes of *N. meningitidis* clonal groups. The first two complete nucleotide sequences of meningococci were published in year 2000 [serogroup A strain Z2491 by Parkhill *et al* (*6*) and serogroup B strain MC58 by Tettelin *et al* (*7*)] and these well-known strains are the two used in this study. While an effective polysaccharide vaccine is available for the control of all serogroup A disease, there is no universal vaccine for the control of serogroup B meningococcal disease (*8*). This suggests the value of comparative genomic studies.

Experiment-based comparisons of *Neisseria* strains have a long history (*9*) and investigators were quick to exploit the availability of chromosome sequences with *in silico* comparisons of pathogenic *Neisseria* genomes (*10*). Based on the annotation of the Z2491 genome, these authors compared the predicted proteins of this genome to the translated DNA sequences of two other *Neisseria* strains and calculated the frequency distribution of the homology (percent similar amino acids) of Z2491 open reading frames to each of *N. meningitidis* MC58 and *N. gonorrhoeae* FA1090. They also plotted the percentage amino acid similarity of predicted proteins, along the Z2491 chromosome. As an exercise in exploiting existing DNA sequence information, Perrin *et al* (*10*) follow a practice commonly used in comparative genomics, that is, a focus on variables associated with genes, proteins, and metabolic pathways (*11*): the "first step of genome analysis commonly aims to identify the gene repertoire emphasizing similarities, differences, and uniqueness among genes". This orientation towards high-level "functional variables" has occurred even though knowledge of fundamental concepts at this level continues to advance at a rapid rate. Data quality is thus an important consideration in the selection of the types of data to use in comparative studies, and it is obvious that nucleotide data undergo the least amount of change over time.

Our interest is in using genome-scale analyses to help focus laboratory investigations into pathogenesis in *Neisseria*, as a step towards the goal of identifying epitopes. To date, the objective of focusing pathogenesis studies has happened as a consequence of functional studies leading to the compilation of lists of virulence genes. Early examples of gene lists include those described in Perrin *et al* (*10*), Tettelin *et al* (*7*), Liò and Vannucci (*12*), and Sun *et al* (*13*). Table S2 of Hotopp *et al* (*14*) provides a recent list of comparative information, but it does not suggest that a convergence to a common set of virulence genes is likely. There also exists a discussion in the literature (*15-17*) about the reasons for the lack of overlap between their respective lists of virulence genes. Snyder and Saunders (*15*) speculated that "the virulence of the pathogenic *Neisseria* spp. may not lie within the genes they possess *per se,* but rather in a 'genetic personality' which is a result of the combinations of these genes".

These observations argue in favor of identifying regions where the probability of finding virulence genes is higher than elsewhere in the genome. Existing concepts of pathogenicity islands (PAIs) (*18*) and islands of horizontal transfer (IHTs) (*7*) have been used for identifying such regions. These methods for identifying "virulence regions" are based on the analysis of single genomes, not on comparative analyses.

In this study, we propose a novel interpretation of a dot plot of exact nucleotide matches between two genomes towards the identification of virulence regions, including an application to *N. meningitidis* serogroup A strain Z2491 (vaccine available) and serogroup B strain MC58 (no vaccine). To this end, we identified those regions of the genome that summarize structural (*i.e.*, nucleotide) differences between the two strains and allocated known virulence genes (including putative ones) relative to these regions. Our analysis suggests a predictable distribution of virulence genes along the genome. We conclude that there may be merit in large-scale structural comparisons of the DNA sequences of closely-related microorganisms, followed by a search for functional correlates.

Results

The genomes of *N. meningitidis* serogroup A strain Z2491 and serogroup B strain MC58 were compared using software MUMmer3, which identified 11,542 forward matches, with a median match length of 48 bp (range: 20-3,203). The sum of the forward matches is 1,070,698 bp, 49% of Z2491 (2,184,406 bp) and 47% of MC58 (2,272,360 bp). MUMmer3 also identified 8,279 reverse-complement matches, with a median match length of 48 bp (range: 20-2,751). The sum of the reverse complement matches is 785,626 bp, 36% of Z2491 and 35% of MC58.

Structural features

Dot plots of the maximal unique matches (MUMs) are shown in **Figure 1**. The first base in the GenBank sequence for MC58 was assigned by Tettelin *et al* (*7*) at the origin of replication (*oriC*), but for Z2491 it was assigned at about 247,600 bp counterclockwise of *oriC* (*6*). Figure 1A presents data as they exist in GenBank, with *oriC* shown, and Figure 1B presents *oriC*-standardized data. We base our analyses mainly on the non-standardized dot plot to enable ready verification using the GenBank files.

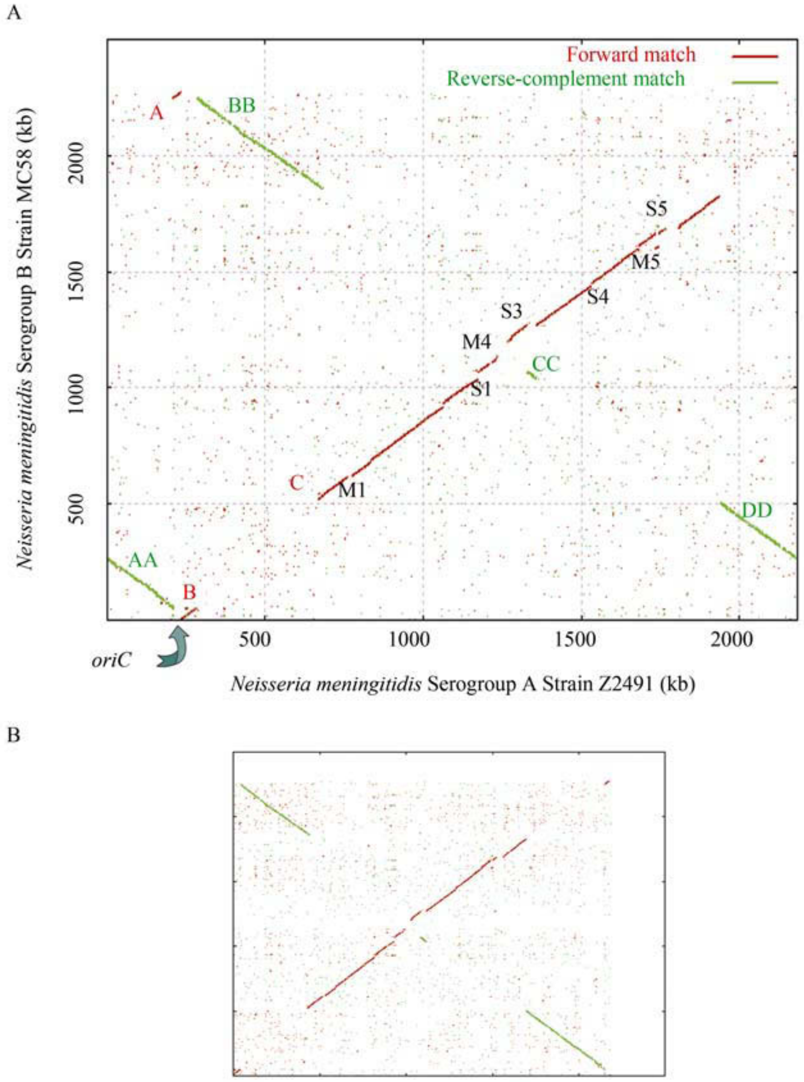

Figure 1 Dot plots of the maximal unique matches. The MUMs were identified by the application of MUMmer3 to the *N. meningitidis* strains Z2491 and MC58. Forward strands are illustrated, numbering from 5' to 3' ends. Forward matches are colored red and reverse-complement matches are colored green. **A**. Non-standardized plot. **B**. *oriC* standardized plot (presented without annotation, for comparative purposes only).

Match regions are immediately apparent from the figures. Forward match regions show a positive slope and are colored red; reverse-complement match regions show a negative slope and are colored green.

The dominant feature of the dot plot is the long forward match (collinear) region between approximately 660-1,950 kb (length=1,290 kb) on the Z2491 genome and approximately 500-1,850 kb (length= 1,350 kb) on the MC58 genome (label C in **Table 1**). There also exist a number of smaller regions of forward matches, the larger two being listed in Table 1 (labels A and B). Four small regions between approximately 171.0 and 190.3 kb on the Z2491 genome (not visible in Figure 1A) can be considered to be part of the collinear region because each region shows about the same slope and intercept as the collinear region. One forward match region (labeled as A in Figure 1A and Table 1) has the same slope but not the same intercept as that of the collinear region and hence appears to be a translocated region; however, this is easily shown to be a consequence of the choice of origin and hence can also be considered to be part of the collinear region (Figure 1B).

There exist three relatively large translocations (and one relatively small translocation), referred to as reverse-complement match regions AA, BB, DD (and CC) (Table 1 and Figure 1A). Regions AA and DD are shown to be one region when both genomes are standardized to *oriC* (Figure 1B).

Discontinuities are visible in the regions of high-density MUMs (Figure 1), notably in the collinear region. **Table 2** identifies a total of 137 kb (Z2491) and 162 kb (MC58) of forward match simple insertions (S*x*), polymorphic regions (Pf*x*), and “intermediate regions” between simple insertion and polymorphic match regions (M*x*). For Z2491, about 42% of matches were identified, *i.e.*, about half of the simple insertion and polymorphic match regions are less than 5 kb.

Allocation of virulence genes to match regions

The virulence genes considered in this study are not uniformly distributed along either genome. A frequency distribution plot for Z2491 (not shown) identified the largest frequency to be at 1,500-1,600 kb (n=11), but there were only a few large groups: 100-200 kb (3), 500-600 kb (5), 600-700 kb (10), and 1,900-2,200 kb (5). The largest frequencies for the virulence genes of MC58 are at 0-100 kb (n=22) and at 1,400-1,500 kb (12), with most genes within small frequency groups. Eleven of 41 known and putative virulence genes for Z2491 lie outside of the collinear region: *i.e.*, ~27% of the virulence genes lie in the 41% translocated regions. The probability of this result occurring under a binomial model is ~0.02, with a mode of about 18. Also, 67 of 121 virulence genes for MC58 lie outside of the collinear region: *i.e.*, ~55% of the virulence genes lie in the 41% translocated regions (p<0.001, with mode of 30).

The virulence genes of strain Z2491 exhibit some strong clusters (*i.e.*, ten virulence genes in each of the NMA0687-0696 and NMA1617-1626 clusters); if each cluster is assumed to be co-regulated and hence treated as a single gene, then 48% of the virulence genes lie in the 41% translocated regions of Z2491 (p=0.13). The virulence genes of strain MC58 presented in Tettelin *et al* (*7*) are not clustered but those identified by Liò and Vannucci (*12*) exist (with one exception) only in clusters: ten virulence genes in NMB1399-1419 and five in NMB2105-2126. With clustering taken in account, 63% of the virulence genes lie in the 41% of the translocated regions; the binomial probability is <0.0001 with mode of 45. In other words, when clustering is considered, the virulence genes of MC58 are even more over-represented in the translocated regions.

Table 2 of Perrin *et al* (*10*) categorizes Z2491 virulence genes into various functional groups: *N. meningitidis* specific, possible virulence associated (modal frequency at 600-700 kb, with 10 out of 13 virulence genes in this class); *N. meningitidis* specific, virulence associated (1,500-1,600 kb, 10 out of 15); pathogen specific, possible virulence associated (500-600 kb, 4 out of 6); and pathogen specific, virulence associated (two genes, one at 500-600 kb and one at 800-900 kb). Hence in Z2491 the modes of these categorizations are located in region BB or around the middle of the collinear region C.

Table 1 of Tettelin *et al* (*7*) groups the MC58 virulence genes into various functional groups: acquisition (modal frequency at 1,900-2,000 kb, with 5 out of 20 virulence genes in this class); colonization (2,000-2,100 kb, 6 out of 37); evasion (70-80 kb, 8 out of 18); and toxins (1,800-1,900 kb, 5 out of 17).

Table 1 Regions of high exact match concentrations

	Z2491 (kb)	MC58 (kb)	Length (kb)	Label
Forward matches	210-234	2,248-2,270	24 (22)	A
	233-283	0-50	50 (50)	B
	660-1,950	500-1,850	1,290 (1,350)	C, collinear
Reverse-complement matches	0-215	272-42	215 (230)	AA
	280-690	2,252-1,852	410 (400)	BB
	1,335-1,360	1,062-1,038	25 (24)	CC
	1,940-2,190	502-252	250 (250)	DD

Note: The column "Length" presents data for Z2491 followed by MC58 in brackets. Labeled regions are shown in Figure 1A. Ranges apply to non-standardized data.

Table 2 Simple-insertion regions, polymorphic regions, and intermediate regions

	Z2491			MC58		
	Start (kb)	End (kb)	Length (kb)	Start (kb)	End (kb)	Length (kb)
			Forward matches			
Simple insertions						
S1	1,175	1,175	0	1,037	1,068	31
S2	1,228	1,228	0	1,116	1,125	9
S3	1,328	1,358	30	1,268	1,269	1
S4	1,530	1,533	3	1,435	1,449	14
S5	1,768	1,807	39	1,685	1,685	0
Polymorphisms						
Pf1	1,842	1,847	5	1,725	1,732	7
Intermediates (between simple insertions and polymorphisms)						
M1	762	775	13	613	618	5
M2	1,063	1,067	4	917	933	16
M3	1,212	1,220	8	1,104	1,108	4
M4	1,237	1,270	33	1,135	1,200	65
M5	1,681	1,683	2	1,600	1,610	10
			Reverse-complement matches			
Polymorphisms						
Pr1	182	187	5	81	74	7
Pr2	194	200	6	67	61	6
Pr3	392	397	5	2,138	2,132	6
Pr4	2,130	2,137	7	314	307	7

Note: Ranges apply to non-standardized data.

Hence, the modal frequencies for acquisition, colonization, and toxins are all in translocated region BB, located near the end closest to the collinear region, or in the collinear region near the BB end. Evasion genes have a modal frequency in translocation AA. Hence all modes are near *oriC* (Figure 1B).

To visualize the distribution of virulence genes that fall in the collinear region, we plotted the distance between consecutive pairs of virulence genes ("Gap Length") against the ordinal starting position of the second gene of each pair (**Figure 2**). Large bars represent large gaps between consecutive virulence genes. The textual annotations identify the forward match "breaks" in the collinear region, *i.e.*, the simple insertions, intermediates, and polymorphic regions as identified in Table 2. It appears that the large gaps are associated with breaks, implying that virulence genes occur around breaks. Strain Z2491 shows a tight cluster of 11 virulence genes around S4 (Figure 2A). Strain MC58, with a larger set of known and putative

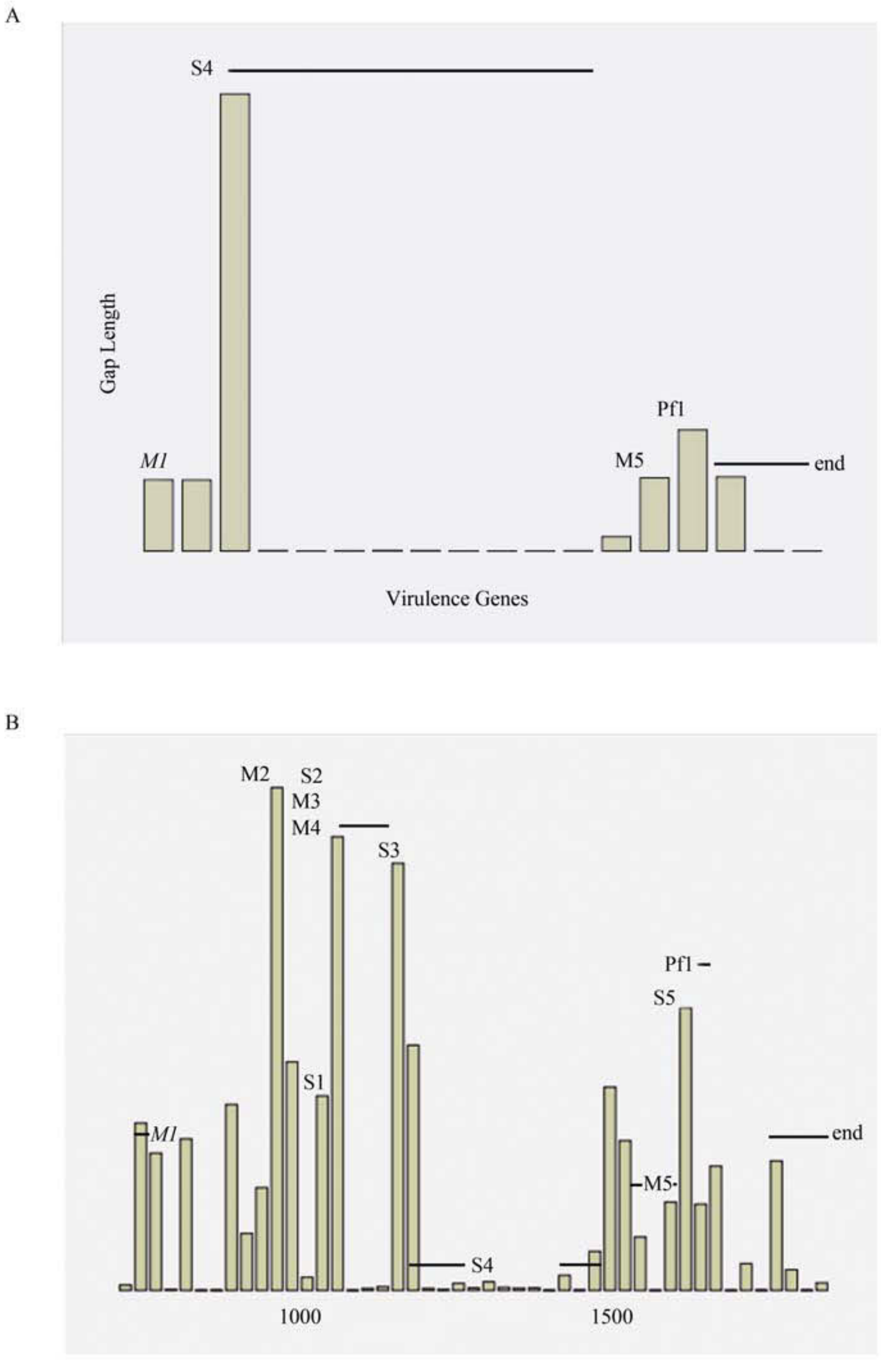

Figure 2 Bar plots of the distance between consecutive pairs of virulence genes. The distance between genes ("Gap Length") is relative to the ordinal starting position of the second gene of each pair ("Virulence Genes"). The locations of the breaks in the collinear region (*i.e.*, forward match simple insertions, polymorphisms, intermediates, and beginning and end of the collinear region) are annotated relative to the nearest gene. The annotations are placed over the breaks they encompass, with italics indicating an exact overlap. The horizontal bars cover the ±16 kb range in the virulence genes that are closest to the annotated break. **A**. Z2491. **B**. MC58 (same axes as Z2491).

virulence genes, shows a similar pattern, but with more variety. The large gaps are again associated with breaks but the clusters of virulence genes around the breaks show a variety of patterns. S4 contains six tightly-clustered virulence genes within itself (space between horizontal lines) and is surrounded on both sides by more genes. M4 is beside a tightly-clustered set of four genes. Virulence genes within and around

M1, M5, Pf1, and end are only partially clustered (Figure 2B).

Both strains show a frequency distribution of "distance from the nearest break" strongly skewed to the right. Although the limited sample size (especially Z2491) precludes the fitting of a density function, perusal of the data suggested to us that virulence genes tend to be situated within 16 kb of the nearest break. Hence, using the proportion of the collinear region occupied by the forward match simple insertions and polymorphisms (±16 kb) as the expected probability, we estimated the probability of observing 16 of 19 virulence genes within 16 kb of a break for Z2491, and of observing 30 out of 48 virulence genes for MC58. The binomial probability for both strains was <0.001, with a mode of 8 for Z2491 and 19 for MC58.

Discussion

The aim of this study was to identify structural (nucleotide) properties from a genomic comparison in order to help focus investigations at a functional (gene, protein, and metabolic pathway) level. A general approach, based on a novel interpretation of structural dot plots of genomes, was devised and applied. Dot plots are one of the simplest bioinformatics methods and are frequently used. We claim a novel interpretation of their use, based on exact matches to enhance data quality. It should be useful to provide focus in searching for phenological characteristics (including virulence genes) in the high-dimensional genomes.

It is surprising, perhaps, that such a simple and potentially powerful interpretation has not been identified in the literature. Only one recent application exists. Using dot plots in the same way that we propose (but using inexact matches), Bentley *et al* (*19*) highlight certain genes at the interfaces of major translocations. However, they do not discuss the logic behind doing so, or suggest that this is a different way of looking at dot plots. In comparison, our manuscript provides the logic for understanding and using structural dot plots.

Resulting match regions (structural) have interpretations and associations at the functional level. Any forward match at the same position in each *oriC*-standardized genome is part of the collinear region, the subsequence assumed to be shared in evolutionary time. Forward match regions at different positions in each genome [the transpositions of Delcher *et al* (*20*)] are consistent with translocations—the MUMs appear to be strings placed in a "new sequence context elsewhere in the genome" (*21*). Reverse-complement match regions at the same position in both genomes are consistent with inversions. A reverse-complement match region at a different position is consistent with the combination of a translocation and an inversion—a "translocated inversion" perhaps. Simple-insertion regions are consistent with "lateral transfer, simple deletions or other evolutionary processes" (*20*).

According to a review by Arber (*22*), there are "three major strategies serving ... for the generation of genetic variation": small changes in nucleotide sequences (*i.e.*, structural polymorphisms), intragenomic reshuffling of DNA segments (*i.e.,* structural translocations and inversions), and intergenomic reshuffling (*i.e.*, structural simple insertions). Hacker and Kaper (*23*) draw similar conclusions about pathogenesis. Thus, in general it seems useful to generate structural dot plots of the genomes of closely-chosen strains to identify structural regions, draw functional associations and implications, and use this information to identify genomic regions (relating to pathogenesis) that deserve further laboratory evaluation and analysis.

We found that the sets of known and putative virulence genes (for each strain) we studied are not distributed at random along the genome. It is, of course, possible that the non-uniform distribution of these sets of virulence genes is an unbiased reflection of the non-uniform distribution of the full sets of genes. However, histograms suggested that the full sets of genes are evenly distributed along each genome. This was confirmed by fitting continuous uniform distributions and testing goodness of fit with the two-tail Kolmogorov-Smirnov statistic (p=0.05 for strain MC58; p=0.07 for strain Z2491).

Virulence genes of strain Z2491 were biased towards the collinear region (or neutral when clustering was taken into account) whereas those of strain MC58 were biased towards the translocations. In the latter case, virulence genes were disproportionately located around the junction between the match region BB and the collinear region C. In the shared collinear region, the virulence genes tended to be located around breaks in contiguous match regions.

It could be that these biases occurred as a consequence of the approaches used to identify the sets of virulence genes that we studied. To investigate this issue, we compared the gene lists used in this study (*7, 10, 12*) to two other lists (*13, 14*). We found that of the 55 virulence genes of Z2491 listed in Sun *et al* (*13*), only one (NMA0185) was in the list compiled by Perrin *et al* (*10*); the 56 virulence genes listed by Hotopp *et al* (*14*) overlapped the list of Perrin *et al* (*10*) by only 3 genes (NMA0578, NMA1617, and NMA1900). And the two new lists shared no genes in common. Yet the histograms of the virulence genes from the three sources showed strong similarities. Major differences are large frequencies occurring only in the list by Perrin *et al* (*10*) in the NMA1600 and only the list of Hotopp *et al* (*14*) in the ~NMA2000. With the exception of the genes around S4 (~NMB1600), large gaps around the breaks in the collinear region appear to exist in both of the new lists. But notably the absence of the S4-related virulence genes in the new lists reduced the proportion of virulence genes around the breaks down to ~0.3, a non-significant result.

We found that of the 57 virulence genes of MC58 given in Sun *et al* (*13*), only 5 genes (NMB0067, NMB0083, NMB1527, NMB1705, and NMB1929) were in the lists compiled by Tettelin *et al* (*7*) and by Liò and Vannucci (*12*). The 60 virulence genes listed by Hotopp *et al* (*14*) overlapped by only 6 genes. And again the two new lists shared no virulence genes in common. The combined lists of Tettelin *et al* (*7*) and Liò and Vannucci (*12*) showed large frequencies in ~NMB0001-0100, ~NMB1400-1430, and ~NMB1900-2150. With the exception of the genes around S4 (~NMB1400-1430), large gaps around the breaks in the collinear region appear to exist, but again not significantly. These results, for each strain, do not suggest a method effect.

Table S2 of Hotopp *et al* (*14*) provides an interesting source of comparative information including lists not considered in detail herein, but this additional information does not suggest to us that there has been a convergence to a common set of virulence genes. In fact, two recent papers from England, each using microarray data to compile their list of virulence genes (*15, 17*), produced different lists. Snyder and Saunders (*15*) noticed the differences and concluded: "Whether these are due to differences in strain gene complement, micoarray design, or data interpretation can not be determined." This elicited a contrary response from Stabler and Hinds (*16*) arguing for the use of their experimental design and approach to data analysis. Snyder and Saunders (*15*) speculated that "the virulence of the pathogenic *Neisseria* spp. may not lie within the genes they possess *per se,* but rather in a 'genetic personality' which is a result of a combinations of these genes ...". We have attempted to do this by identifying areas in the genome(s) where further investigations into pathogenicity might be most profitable.

Our approach complements the existing concepts of PAIs (*18*) and IHTs (*7*) in identifying such regions. The relative strengths of our approach include: using simpler (and readily available) information than that used in the identification of PAIs (*24*); more factually based (exact nucleotide matches) than indices (%G+C and dinucleotide signatures) of regions of horizontal transfer. The match regions identified by our comparative approach clearly and simply identify the major differences between two genomes. This knowledge is of particular interest to the types of strains considered in this study, given that there is no universal vaccine for serogroup B (MC58) (*8*) but there is one for the control of all serogroup A disease (Z2491).

The major determinant of the different pathogenicities of the various *Neisserial* serogroups is the polysaccharide capsule (*25*). The genes involved are clustered in the cps locus (*26*) and notably include *synX*, *siaB-D*, *ctrA-D*, and *lipA-B* genes (*25*).

A set of BLAST searches showed that of these genes, only the polysialic acid capsule biosynthesis genes (*synX* and *siaB-D*) of MC58 have no homologs in Z2491. In the NCBI datasets, one can find sets of genes in both Z2491 and MC58 strains with products listed as "putative capsule biosynthesis protein", but Vogel *et al* (*25*) have noted that the polysaccharide of the serogroup A capsule is not of sialic acid but of α-1,6-linked N-acetyl-D-mannosamine-1-phosphate, a polymer that is immunogenic in humans. This set of genes is unique among those at the cps locus in being so different in the two strains. The *ctrA* and *ctrD* genes have identities of around 95%, *ctrB* of 81%, and *ctrC* of 77%. The *lipA-B* genes have identities greater than 95%.

We have noted that all these genes are listed as evasion genes in Tettelin *et al* (*7*). They are all found in match region AA and, notably, they are all near the extreme of the AA region, where mutations would be expected following a translocation. Snyder and Champness (*27*) note that "Inversion mutations often cause no phenotype. If the inversion involves a longer sequence, including many genes, generally the only affected regions will be those in the inversion junctions, where the recombination occurred".

We found that most of the published virulence genes lie between BB and C or near breaks in the collinear region, and suggest that a search for low-identity genes in these areas might be profitably investigated to identify possible epitopes.

Conclusion

As a step towards the identification of structural (nucleotide) properties from a genomic comparison of two *Neisseria* strains, we developed and applied a new (and general) approach, based on a previously un-utilized interpretation of structural dot plots of genomes. We found that the virulence genes of MC58 tend more to the translocated regions than do those of Z2491, notably tending towards the interface between reverse-complement match region BB and the collinear region C. Within the collinear region, virulence genes tend to occur within 16 kb of gaps in the exact matches. We directed attention to the 16 kb region around simple insertion S4 of MC58. Verification of these tendencies using genes clustered in the cps locus were sufficiently supportive to suggest that these tendencies can be used to focus the search for and understanding of virulence genes and mechanisms of pathogenicity in *Neisseria*.

Materials and Methods

Sequences

Complete sequences of both *N. meningitidis* serogroup A strain Z2491 (*6*) and *N. meningitidis* serogroup B strain MC58 (*7*) were obtained from GenBank (ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/). FASTA file AL157959.fna contained strain Z2491; FASTA file AE002098.fna contained strain MC58.

Exact matches

The genomes were compared using software MUMmer3 (MUMmer version 3.0) (http://mummer.sourceforge.net/). We used only the part concerned with the identification and visualization of MUMs, ignoring the clustering algorithm and the alignment generator. A maximal unique match was defined in Delcher *et al* (*20*) as "a subsequence that occurs exactly once in Genome A and once in Genome B, and is not contained in any longer such sequence".

The file AL157959.fna (Z2491) was used as the reference file and file AE002098.fna (MC58) as the query file. MUMmer3 was run with options: -b (compute both forward and reverse-complement matches); -c (report the query position of a reverse-complement match relative to the forward strand of the query); -mum (compute matches that are unique in both the reference and query files); and default minimum match length (20 bp).

Following the MUMmer manual, we refer to "forward strands" and "reverse strands", and to "forward matches" and "reverse-complement matches". "Match regions" (forward and reverse-complement) were defined as *visible* regions of high-density MUMs. Precise boundaries of match regions are not relevant and were not defined. Match regions were identified from a dot plot of MUMs using a MUMmer utility based on GNUplot (http://www.gnuplot.info/).

When a forward match region is located at the approximately same position in each genome and with unitary slope in a dot plot standardized to the origins of replication (*oriC*), it is referred to as a "collinear region". When a match region appears to have been "deleted from one location (in an assumed common genome) and inserted elsewhere", it would be referred to as a "transposition" in Delcher *et al* (*20*). However, we prefer the words "translocation"/"translocated region", following the definition of Deonier *et al* (*21*) of a translocation as a "placement of a chromosomal segment into a new sequence context elsewhere in the genome".

A "simple-insertion match region"/"polymorphic match region" is defined as an area along the collinear region or a translocated region that has a substantially lower density of MUMs in one/two dimensions. The former is visible in the dot plot as a gap in a region of high MUM density in one genome, which is not matched to a corresponding gap in the other genome. The latter is visible in the dot plot as a gap in a region of high MUM density in one genome, which is matched to a corresponding gap in the other genome (*i.e.*, gaps with approximately the same length in each strain). Simple-insertion and polymorphic match regions were identified first by simple visual inspection and then by calculating the ranked sizes of the gaps between the adjacent sub-regions of high-density MUMs in each genome, and then zooming in on the dot plot to better identify the larger gaps. To keep the number of simple insertions and polymorphisms identified to a tractable set, only those with a gap greater than 5 kb are considered; and this gap size should be interpreted in light of the 2,000 kb genomes being studied.

Virulence genes

A set of 104 virulence genes (including putative ones) was given for strain MC58 in Table 1 of Tettelin *et al* (*7*). Liò and Vannucci (*12*) added an additional 17 putative virulence genes: NMB 0313, 1399, 1400, 1403, 1405, 1409, 1411, 1412, 1414, 1418, 1419 (gene *ruvC*), 2105 (*mafB*), 2107 (between 2,224,569-2,225,333 bp), 2111 (between 2,226,310-2,227,155 bp), 2119, 2122, and 2126.

A set of 41 virulence genes (including putative ones) was given for strain Z2491 in Table 2 of Perrin *et al* (*10*). Those genes that are present in all invasive meningococci and are considered here include NMA 0184-0186 (genes *lipB*, *lipA*), 0788 (between 772,407-774,551 bp), 1617 (gene *sodC*), 1618-1626, 2191, 2193 (virulence associated); 0687-0696, 1994 (gene *natC*), 1995 (*natD'*), and 1996 (also *natD'*) (possible virulence associated). Those genes of strain Z2491 that are shared with the gonococcus while absent from *N. lactamica* (a commensal) and are considered here include NMA 0609 (gene *pilC1*), 0905 (gene *iga*), 1925 (between 1,855,866-1,858,239 bp) (virulence associated); 0478, 0575-0578 (0577 is gene *fetB2*), 1642 (gene *porA*), 1676 (between 1,599,572-1,600,353 bp), 1725, and 1900 (possible virulence associated).

Statistical analyses were conducted with Systat and Matlab software. The analyses included calculation of frequency distributions and their statistics, fitting data to the binomial and uniform distributions, and testing of their goodness of fit.

BLAST analyses were conducted to understand how known virulence genes related to the polysaccharide vaccine for the control of all serogroup A disease are distributed relative to the regions of the genome that summarize structural differences between the two strains.

Acknowledgements

WRC gratefully acknowledges the staff of the Canadian Bioinformatics Workshop series (http://bioinformatics.ca/) for a stimulating introduction to this challenging field. Jason Rancourt offered useful discussion during the early stages of this project. The authors acknowledge the insights of Dr. Clifford Clark for enhancing our understanding of the context of a structural approach in molecular biology. Lastly, we wish to acknowledge the contribution of an unknown reviewer for a number of specific recommendations that greatly helped to clarify our argument. WRC and RSWT are grateful to the Public Health Agency of Canada and Health Canada's Biotechnology Genomic Fund for providing support for this work. VRSKD and JW also would like to acknowledge the support from the Canadian Network of Centers of Excellence, MITACS.

Authors' contributions

WRC conceived the study and performed the analyses. RSWT identified the application to *N. meningitidis*, and collaborated with WRC to apply and improve the approach. WRC and RSWT prepared and revised the first version of the manuscript, with improvements to both the approach and the manuscript contributed by VRSKD, BD, and BL. GEW and JW supervised the work of BD and VRSKD, respectively, and contributed to the robustness of the approach. All authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1 Whalen, C.M., *et al.* 1995. The changing epidemiology of invasive meningococcal disease in Canada, 1985 through 1992. Emergence of a virulent clone of *Neisseria meningitidis*. *JAMA* 273: 390-394.

2 Frasch, C.E., *et al.* 1985. Serotype antigen of *Neisseria meningitidis* and a proposed scheme for designation of serotypes. *Rev. Infect. Dis.* 7: 504-510.

3 Riou, J.Y. and Guibourdenche, M. 1992. *Laboratory Methods: Neisseria and Branhamella.* Institut Pasteur, Paris, France.

4 Caugant, D.A. 1998. Population genetics and molecular epidemiology of *Neisseria meningitidis*. *APMIS* 106: 505-525.

5 Maiden, M.C., *et al*. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. USA* 95: 3140-3145.

6 Parkhill, J., *et al.* 2000. Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature* 404: 502-506.

7 Tettelin, H., *et al.* 2000. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* 287: 1809-1815.

8 Girard, M.P., *et al.* 2006. A review of vaccine research and development: meningococcal disease. *Vaccine* 24: 4692-4700.

9 Kingsbury, D.T. 1967. Deoxyribonucleic acid homologies among species of the genus *Neissera*. *J. Bacteriol.* 94: 870-874.

10 Perrin, A., *et al.* 2002. Comparative genomics identifies the genetic islands that distinguish *Neisseria meningitidis*, the agent of cerebrospinal meningitis, from other *Neisseria* species. *Infect. Immun.* 70: 7063-7072.

11 Karlin, S., *et al.* 1998. Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.* 32: 185-225.

12 Liò, P. and Vannucci, M. 2000. Finding pathogenicity islands and gene transfer events in genome data. *Bioinformatics* 16: 932-940.

13 Sun, Y.H., *et al.* 2000. Functional genomics of *Neisseria meningitidis* pathogenesis. *Nat. Med.* 6: 1269-1273.

14 Hotopp, J.C., *et al.* 2006. Comparative genomics of *Neisseria meningitidis*: core genome, islands of horizontal transfer and pathogen-specific genes. *Microbiology* 152: 3733-3749.

15 Snyder, L.A. and Saunders, N.J. 2006. The majority of genes in the pathogenic *Neisseria* species are present in non-pathogenic *Neisseria lactamica*, including those designated as 'virulence genes'. *BMC Genomics* 7: 128.

16 Stabler, R. and Hinds, J. 2006. The majority of genes in the pathogenic *Neisseria* species are present in non-pathogenic *Neisseria lactamica*, including those designated as virulence genes: response. *BMC Genomics* 7: 129.

17 Stabler, R.A., *et al.* 2005. Identification of pathogen-specific genes through microarray analysis of pathogenic and commensal *Neisseria* species. *Microbiology* 151: 2907-2922.

18 Hacker, J., *et al.* 1997. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol. Microbiol.* 23: 1089-1097.

19 Bentley, S.D., *et al.* 2007. Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18. *PLoS Genet.* 3: e23.

20 Delcher, A.L., *et al.* 1999. Alignment of whole genomes. *Nucleic Acids Res.* 27: 2369-2376.

21 Deonier, R.C., *et al.* 2005. *Computational Genome Analysis: An Introduction.* Springer, New York, USA.

22 Arber, W. 2002. Evolution of prokaryotic gnomes. *Curr. Top. Microbiol. Immunol.* 264: 1-14.

23 Hacker, J. and Kaper, J.B. 2000. Pathogenicity islands and the evolution of microbes. *Annu. Rev. Microbiol.* 54: 641-679.

24 Snyder, L.A., *et al.* 2005. Comparative overview of the genomic and genetic differences between the pathogenic *Neisseria* strains and species. *Plasmid* 54: 191-218.

25 Vogel, U., *et al.* 2001. Capsular operons. In *Meningococcal Disease, Methods and Protocols* (eds. Pollard A.J. and Maiden M.C.), pp.187-201. Humana Press, Totowa, New Jersey, USA.

26 Frosch, M., *et al.* 1989. Molecular characterization and expression in *Escherichia coli* of the gene complex encoding the polysaccharide capsule in *Neisseria meningitidis* group B. *Proc. Natl. Acad. Sci. USA* 86: 1669-1673.

27 Snyder, L. and Champness, W. 2003. *Molecular Genetics of Bacteria*, p.134. ASM Press, Washington, D.C., USA.