Slipped strand mispairing in the gene encoding sialidase NanH3 in *Gardnerella* spp.

Shakya P. Kurukulasuriya, Mo H. Patterson, Janet E. Hill[#]

Department of Veterinary Microbiology, University of Saskatchewan, Saskatoon, Canada

Running head: Slipped-strand mispairing in *Gardnerella* sialidase

#Address correspondence to Janet E. Hill, Janet.Hill@usask.ca

Abstract

Cell wall proteins with sialidase activity are involved in carbohydrate assimilation, adhesion to mucosal surfaces, and biofilm formation. *Gardnerella* spp. inhabit the human vaginal microbiome and encode up to three sialidase enzymes, two of which are suspected to be cell wall associated. Here we demonstrate that the gene encoding extracellular sialidase NanH3 is found almost exclusively in *G. piotii* and closely related *Gardnerella* genome sp. 3, and its presence correlates with sialidase positive phenotype in a collection of 112 *Gardnerella* isolates. The *nanH3* gene sequence includes a homopolymeric repeat of cytosines that varies in length within cell populations, indicating that this gene is subject to slipped-strand mispairing, a mechanisms of phase variation in bacteria. Variation in the length of the homopolymer sequence results in encoding of either the full length sialidase protein or truncated peptides lacking the sialidase domain due to introduction of reading-frame shifts and premature stop codons. Phase variation in NanH3 may be involved in immune evasion or modulation of adhesion to host epithelial cells, and formation of biofilms characteristic of the vaginal dysbiosis known as bacterial vaginosis.

28 **Introduction**

29 Bacterial vaginosis (BV) is a condition that is characterized by altered composition of the
30 vaginal microbiota and occurs when the healthy microbiota is replaced by an overgrowth of mixed
31 aerobic and anaerobic species, including *Gardnerella* spp. (1, 2). An abundance of *Gardnerella*
32 spp. is often found in cases of symptomatic BV, although they are also found in healthy women
33 with no clinical signs or symptoms of BV (3). *Gardnerella* spp. can be resolved into four subgroups
34 based on cpn60 barcodes sequencing (4) or whole genome sequencing (5). Recently, the
35 description of *Gardnerella vaginalis* was amended and three new species were defined within the
36 genus *Gardnerella*: *G. leopoldii*, *G. swidsinskii*, and *G. piotii* (6). *G. piotii* and *G. vaginalis*
37 correspond to cpn60 subgroup B and C, respectively. *G. leopoldii* and G. *swidsinskii* were
38 previously grouped together as subgroup A based on cpn60 sequences of the isolates available at
39 that time. Isolates belonging to cpn60 subgroup D corresponded to three distinct genome species:
40 genome species 8, 9 and 10 (7).

41 In addition to the characteristic change in microbiota, and elevated pH, sialidase activity in
42 vaginal fluid is a diagnostic marker of BV (8, 9). Sialidase enzymes cleave the glycosidic linkages
43 of sialic acids from terminal glycans including vaginal mucins, immunoglobulin A molecules, and
44 epithelial cell surface glycoproteins (10, 11). Activity on the latter may be related to the adhesion
45 of bacteria to epithelial cells; the initial step in biofilm formation (12). Previously, we have
46 demonstrated that sialidase activity is almost exclusively confined to *G. piotii* and *Gardnerella*
47 genome species 3 (cpn60 subgroup B) (13).

48 A putative sialidase gene, *nanH1* (sialidase A) was identified in *Gardnerella* spp. and was
49 initially thought to be the gene responsible for sialidase activity (14). Although *nanH1* appears to
50 be present in all sialidase positive strains, it is also found in sialidase activity negative strains (13).

51 This observation combined with the lack of signal peptide on the NanH1 proteins suggests that
52 this protein is an intracellular enzyme, likely involved in manipulation of nutritional substrates as
53 is the case in many bacteria (15). The discrepancy between *nanH1* presence and enzyme activity
54 led to the recognition of two additional sialidase genes (*nanH2* and *nanH3*) in some *Gardnerella*
55 isolates (16), which are unevenly distributed among *Gardnerella* spp. (17). Based on biochemical
56 studies, Robinson et al. (16) concluded that NanH2 and NanH3 are the proteins responsible for
57 sialidase activity detected in vaginal fluid. The presence of signal peptides and/or C-terminal
58 transmembrane alpha-helices in some of the identified sequences further suggested that they are
59 secreted and tethered to the cell, although some proportion may subsequently be released into the
60 extracellular environment due to proteolytic activity (16). Cell wall anchored proteins with
61 sialidase activity in other species have been found to be involved in adhesion to mucosal surfaces
62 and biofilm initiation as well as carbohydrate assimilation (18–21).

63 Interestingly, in our examination of genome sequences of our collection of *Gardnerella*
64 isolates, we noted that *nanH3* contains a homopolymeric tract of cytosine residues. Genomic
65 regions that contain short, homogenous or heterogenous repeats are susceptible to slipped-strand
66 mispairing in which the length of the repeat region can change with each replication (22). The
67 result of this modulation is phase variation: a reversible process in which the expression of the
68 encoded protein can be rapidly switched on and off (23). Phase variation in cell surface proteins
69 can result in immune evasion and alteration of biofilm phenotypes (24, 25).

70 Here we determined the distribution of genes encoding sialidases NanH2 and NanH3 in
71 the context of newly reclassified *Gardnerella* spp. and demonstrated that *nanH3* is subject to
72 slipped-strand mispairing.

73

74 **Methods**

75 **Protein domain identification and sequence alignment**

76 DNA and protein sequence alignments were performed with Clustal Omega (EMBL_EBI)
77 and NCBI BLAST (basic local alignment search tool). InterproScan
78 (https://www.ebi.ac.uk/interpro/) and SignalP were used to predict the location of functional
79 domains and signal peptides (26).

80 **Bacterial strains and culture conditions**

81 *Gardnerella* strains (n = 112) from a previously described culture collection were used in
82 the study (13). Sialidase activity for all isolates had been determined previously (13) using a
83 quantitative assay and fluorogenic substrate 2'-(4-methylumbelliferyl)-α-D-N-acetylneuraminic
84 acid sodium salt hydrate (11), and whole genome sequences for 36 of these isolates were
85 previously determined (27). Complete strain information and sequence accessions are provided in
86 Table S1.

87 *Gardnerella* isolates were grown on Columbia Sheep Blood Agar plates (BBL, Becton,
88 Dickinson and Company, Sparks, MD, USA) at 37°C for 48 hours with anaerobic BD GasPak EZ
89 (Becton, Dickinson and Company, Sparks, MD, USA). For broth cultures, a few colonies from the
90 plate were collected with a 10 µl inoculation loop and used to inoculate NYC III (ATCC 1685
91 medium; per litre: 2.4g HEPES, 15 g Proteose peptone, 3.8 g Yeast extract, 5 g NaCl, 5 g Glucose).
92 Broth cultures were incubated at 37°C for 48 hours in anaerobic conditions.

93 **PCR screen for *nanH3***

94 Genomic DNA was purified from broth cultures using a modified salting-out procedure
95 (28). All DNA extracts were initially tested by PCR for the universal cpn60 barcode to confirm
96 the quality of the DNA.

97 To screen isolates for the presence of *nanH3*, PCR primers were designed based on
98 multiple sequence alignments of 15 *nanH3* sequences obtained from the Integrated Microbial
99 Genomes database (https://img.jgi.doe.gov/). Degenerate primers were designed to account for
100 sequence variability within the gene sequence and to amplify a product of 375 bp in length
101 (JH0684: 5'-GTT GTA GAR CTT TCT GAT GG-3', JH0685: 5'-YRY TAT TAT CGC CCT CAT
102 ATA-3'). PCR reactions contained 1 × PCR Buffer (0.2 M Tris-HCl at pH 8.4, 0.5 M KCl), 2.5
103 μM $MgCl_2$, 0.40 μM dNTP, 0.20 μM forward primer, 0.20 μM reverse primer, 2 U Taq DNA
104 Polymerase, ultrapure water and 2 μl of template DNA in a final volume of 50 μl. PCR reactions
105 were conducted using the following thermocycling parameters in a Mastercycler Pro 6321
106 (Eppendorf AG, Hamburg, Germany): 94 °C for 3 minutes, 40 cycles of (94 °C for 30 seconds, 55
107 °C for 30 seconds, 72 °C for 30 seconds), 72 °C for 1 minute, hold at 20 °C. PCR products were
108 visualized under UV light on a 1.0 % (w/v) agarose gel containing ethidium bromide.

109 **Homopolymer PCR, cloning and sequencing**

110 To determine the length of the homopolymer region of *nanH3,* four strains (W11, VN014,
111 VN015, NR032) were cultured on Columbia Agar plates with 5% (v/v) sheep blood. Primers
112 JH0780 (5'-ATG ATT GGA ACA GCG CAT AAA G-3') and JH0781 (5'-GAT TTC TCC ACC
113 TAC AGT TAC C-3') were designed to PCR amplify a region including basepairs 2-310 of the
114 open reading frame of *nanH3*.

115 The DNA sequence (308 bp) was amplified by PCR in a Mastercycler Pro 6321 (Eppendorf
116 AG, Hamburg, Germany). The components of the PCR reaction mix (50 μl per reaction) were
117 added to achieve final concentrations of 1×High Fidelity PCR buffer (60 mM Tris-SO_4 (pH 8.9),
118 18 mM $(NH_4)_2SO_4$), 0.2 mM dNTP mix, 2 mM $MgSO_4$, 1U Platinum High Fidelity (Hi-Fi) proof
119 reading Taq polymerase (Invitrogen, Carlsbad, CA, USA). Two colonies of each *G. piotii* or

120 genome sp. 3 strain were randomly picked and added to separate PCR reactions using sterile
121 toothpicks. Thermocycling conditions included 35 cycles of denaturation at 94°C for 15 seconds,
122 annealing at 55°C for 30 seconds, extension at 68°C for 1 minute, and final extension at 68°C for
123 5 minutes. PCR products were visualized on a 1% (w/v) agarose gel. PCR products were purified
124 using QiaQuick PCR purification kit (Qiagen, Hilden, Germany) and purified PCR products were
125 sequenced using the amplification primers (JH0780, JH0781).

126 To clarify the exact length of the homopolymer region, amplicons generated from the two
127 colonies of each of the four strains were A-tailed in 10 µl reactions containing 1× Platinum PCR
128 buffer (20 mM Tris HCl (pH 8.4), 50 mM KCl), 2 mM $MgCl_2$, 0.5 mM dATP, 5 U Platinum Taq
129 polymerase (Invitrogen, Carlsbad, CA, USA) and less than 500 ng of the purified PCR product.
130 The reaction mixture was incubated at 72 °C for 20 minutes in Mastercycler Pro 6321 (Eppendorf
131 AG, Hamburg, Germany). End-modified PCR products were ligated into pGEM-T Easy vector
132 (Promega, Madison, WI, USA). The vector-insert construct was used to transform chemically
133 competent DH5α cells or OneShot Top 10 *E. coli* (Invitrogen, Carlsbad, CA, USA) and plated on
134 LB/AMP/X-gal agar media. Ten white colonies were randomly selected from the transformants of
135 each *Gardnerella* strain and transferred into LB + ampicillin broth. Cultures were grown overnight
136 at 37 °C. Plasmid DNA was isolated using QiaPrep Spin Miniprep kit (Qiagen, Hilden, Germany).
137 Plasmids were sequenced using vector primers T7 and SP6.

138

139 **Results**

140 **Distribution of extracellular sialidases in *Gardnerella* spp. isolates**.

141 Robinson et al. (16) had previously reported that sialidase activity was associated with the
142 presence of either *nanH2* or *nanH3* in a collection of 34 *Gardnerella* isolates but did not report

143 the species or subgroup affiliation. In order to confirm this observation and to reconcile the
144 distribution of genes and activities with the new *Gardnerella* taxonomic framework (6), we
145 queried an unrelated set of 36 isolates in our culture collection for which whole genome sequence
146 data and sialidase activity data were available (Table S1). The presence of the genes was assessed
147 by aligning *nanH2* of *G. piotii* JCP8151B (ATJH01000056) and *nanH3* of *Gardnerella* genome
148 sp. 3 strain W11 to the genome sequences using BLASTn. BLASTn results were dichotomous,
149 with either one clear match per genome (E = 0.0, bitscore 1275-5049, percent identity 95-99%) or
150 no significant hits reported. Sialidase activity was reported in 13/36 isolates (Table 1) including
151 12 *G. piotii* or genome sp. 3 (cpn60 subgroup B) isolates and one *G. vaginalis* isolate. The *nanH3*
152 gene was found in 12/13 sialidase activity positive isolates while *nanH2* was present in only 7/13
153 isolates. Six isolates possessed both *nanH2* and *nanH3* genes in their genomes. VN002 was
154 sialidase activity positive but only *nanH2* was identified in the genome sequence. This isolate
155 subsequently screen PCR positive for *nanH3* (see next section), suggesting that the gene was not
156 included in the shotgun assembly of the genome, which consisted of multiple gap-containing
157 scaffolds. None of the *G. swidsinskii, G. leopoldii* or subgroup D isolates were sialidase positive
158 and none contained *nanH2* or *nanH3.*

159 Interestingly, the NanH3 of *G. piotii* strain JCP8151B investigated by Robinson et al. was
160 reported to lack a signal peptide (16). When we examined the sequence upstream of *nanH3* in the
161 JCP8151B sequence (Genbank accession ATJH01000033), we found that there is an alternative
162 start codon, and a cytosine homopolymer. The frame-shift that resulted in the annotation of the
163 gene without the signal peptide-encoding N-terminus is caused by the homopolymer. This was
164 also the case for isolates GH020, GH022, N170, and NR026 where the homopolymer created a
165 frame-shift that resulted in a protein annotation lacking the signal peptide encoded by sequence

166 immediately upstream of the homopolymer. In N144, W11, N101, N153 and N95, the
167 homopolymer length did not create a frame-shift and the annotated NanH3 protein sequence
168 includes a signal peptide. In GH007, GH019, and GH015 we were unable to determine if a signal-
169 peptide encoding sequence was present due to assembly gaps immediately upstream of the
170 homopolymer in these genomes (Table S1).

171 Signal peptides were detected in 4/7 predicted NanH2 sequences (GH019, GH020, N101
172 and N153). In the other three cases where a *nanH2* sequence was identified, we could not conclude
173 if a signal peptide was included in the protein sequence provided in the Genbank annotation since
174 assembly gaps in those regions of the genomes affected the annotation of these proteins.

175 **Correlation of *nanH3* with sialidase activity**

176 In order to examine further the relationship of *nanH3* to sialidase activity, 112 *Gardnerella*
177 spp. isolates for which sialidase activity data was available (13) were screened for the presence of
178 *nanH3* using PCR primers JH0684/JH0685 (Figure 1). Sialidase activity was detected in 32/33 *G.*
179 *piotii*/genome sp. 3 isolates and 3/35 *G. vaginalis* isolates. All sialidase activity positive isolates
180 were positive for *nanH3* by PCR with the exception of one *G. piotii* isolate (WP027) that was PCR
181 negative. Unfortunately, genome sequence data was not available for WP027.

182 **Characterization of a poly-C homopolymer in *nanH3***

183 The protein sequences encoded by all nine *nanH3* sequences without assembly gaps were
184 predicted to encode a signal peptide (amino acids 1-32), a sialidase domain (amino acids 183-547
185 in W11) and a C-terminal transmembrane domain (amino acids 789-807 in W11) (Figure 2A). We
186 observed a poly-cytosine homopolymer (8-14 cytosines) in all *nanH3* genes identified in the whole
187 genome sequences, approximately 100 bases from the start of the open reading frame. The
188 homopolymer occurred immediately following the region of the sequence predicted to encode the

189 signal peptide (amino acids 1-32) (Figure 2B). When Sanger sequencing was performed on PCR
190 products corresponding to nucleotides 2-310 of *nanH3* amplified from genomic DNA extracted
191 from broth cultures of strain W11, the results suggested variable lengths of the homopolymer.
192 Specifically, clean data was obtained upstream of the poly-C tract but sequence 3' to that region
193 was indicative of a mixed template (Figure S1). Since it has been demonstrated that simple repeats
194 such as homopolymers can be subject to slipped strand mispairing, we set out to determine if the
195 homopolymer region of *nanH3* varied in length within and between strains of *G. piotii*/genome
196 species 3.

197 PCR product libraries were made from two colonies each of isolates W11, VN014, VN015
198 and NR032 (8 clone libraries total). Plasmids were purified from 10 colonies from each of the 8
199 clone libraries and sequenced. High quality sequence data was obtained from 15, 14, 18 and 16
200 clones from the W11, VN014, VN015 and NR032 libraries, respectively. Example results are
201 shown in Figure S2. The length of the homopolymer varied from 8 to 14 among all strains (Table
202 2), and within each strain the sequence flanking the homopolymer was identical. *In silico*
203 translation of the encoded polypeptides showed that most homopolymer lengths resulted in a
204 truncated peptide (37-44 amino acids), while full-length protein (812 aa) would result when there
205 were nine or twelve cytosines in the homopolymer region. Truncated peptides would include the
206 signal peptide (amino acids 1-32) but no part of the predicted sialidase domain (amino acids 183-
207 547) (Figure 2). Interestingly, the most common length of the poly-C region varied among strains.
208 C_{10} was most common in VN014 and VN015, while C_9 and C_{12} were most frequently observed in
209 NR032 and W11, respectively. VN015 had the highest number of homopolymer variants (Figure
210 3). A second open reading frame within the *nanH3* open reading frame was also identified,
211 corresponding to amino acids Met_{397}-Tyr_{812} of the NanH3 protein. These 415 aa polypeptides

212 encompass part of the predicted sialidase domain, and the C-terminal transmembrane domain but
213 lacks a signal peptide.
214
215 **Discussion**
216 The genus *Gardnerella* is known as a hallmark of BV and their abundance is used as a
217 criterion for the laboratory diagnosis of the condition (29). In addition to participating in biofilms
218 that coat epithelial cells in BV, *Gardnerella* spp. produce enzymes and a cholesterol dependent
219 cytolysin (vaginolysin) that contribute to degrading the protective barriers of the vaginal mucosa
220 (30). Sialidase activity can provide nutrients by releasing sialic acid moieties from vaginal
221 sialylated mucins (11), altering the physical properties of vaginal mucus. Removal of sialic acid
222 residues from epithelial cell surface glycans can facilitate bacterial adhesion and initiation of
223 biofilm formation (12, 18, 20, 21).
224 The two putative sialidase genes (*nanH2* and *nanH3*) in *Gardnerella* spp. have been
225 demonstrated to encode proteins with high enzymatic potency (16). Taken together, the result of
226 this study and our current study show that *nanH3* is more common and that *nanH2* is virtually
227 never found without *nanH3.* It is now clear that extracellular sialidase activity is a property of *G.*
228 *piotii* and the closely related *Gardnerella* genome sp. 3. We found three occurrences of sialidase
229 positive, *nanH3* positive *G. vaginalis* isolates (3/35 isolates screened), and Vaneechoutte et al. (6)
230 reported sialidase activity in 1/4 *G. vaginalis* isolates used in the amendment of the genus. This
231 infrequent prevalence of *nanH3* in *G. vaginalis* could be the result of lateral gene transfer from *G.*
232 *piotii* since these species co-exist in the same microbiome and women are usually colonized by
233 more than one species (7, 31, 32).

234 The *nanH3* sequences we examined were predicted to encode a signal peptide, a sialidase
235 domain and a C-terminal membrane domain, suggesting a cell-wall tethered protein, similar to
236 SiaBb2 of *Bifidobacterium bifidum*, which enhances adhesion to intestinal mucosal surfaces and
237 contributes to carbohydrate assimilation (21). The presence of the homopolymer immediately
238 downstream of the signal peptide encoding region presents a challenge to automated annotation of
239 this sequence. Sequencing and assembly methods commonly used in bacterial whole genome
240 sequencing struggle with homopolymeric sequences, which can result in frame-shifts such as we
241 observed in the *nanH3* sequence of JCP8151B.

242 Genomic regions that contain homogenous or heterogenous repeats are prone to changes
243 in length of the repeat at each replication due to slipped strand mispairing (22). This can lead to
244 consequent changes in the transcription or translation product of a gene depending on whether the
245 slipped strand mispairing occurs within the open reading frame or in extragenic regions such as
246 promoter sequences. Phase variation results when slipped strand mispairing creates "on" and "off"
247 states of expression of a protein. Phase variation of cell surface proteins has been documented in
248 many bacterial species including *Neisseria* spp. (33–35), *Salmonella* spp. (36–38), *Treponema*
249 *pallidum* (39, 40) and *Helicobacter pylori* (25). It was first described in the *opa* genes that encode
250 opacity surface proteins of *Neisseria* spp. (41). The *opa* genes contain CTCTT pentamer repeats
251 within the signal peptide encoding region and the expression of the Opa protein is regulated by
252 slipped-strand mispairing. With six, nine or twelve CTCTT repeats the initiation codon is in frame
253 with the remaining *opa* gene translating the Opa protein. Four or eight coding repeats makes the
254 initiation codon out of frame with rest of the *opa* codons. The phase "on" state allows them to
255 adhere to specific surface receptors on host cells (42).

256 Our results clearly show variation of the homopolymer length among *G. piotii* and
257 *Gardnerella* genome sp. 3 within colonies growing on agar, with anywhere from 8 to 14 C's
258 observed. The effect of this change would be a mixture of cells with translation of NanH3 on or
259 off. Why would *G. piotii* have a sialidase subject to phase variation? Two possible reasons are
260 immune evasion and cell adhesion (biofilm initiation and dispersal).

261 Although BV is often referred to as a non-inflammatory condition because of the lack of
262 typical clinical signs of inflammation (leukocyte infiltration, pain, redness, swelling), there is
263 evidence of a host immune response to bacteria associated with BV, including *Gardnerella* spp..
264 IgA specific for *Gardnerella* produced vaginolysin has been detected in women with clinical signs
265 of BV (43), and IgA levels correlate with IL-8 expression in vaginal secretions (44). *Gardnerella*
266 has also been shown to stimulate pro-inflammatory cytokines *in vitro* (45). Balancing this
267 inflammatory response are the actions of bacterial enzymes like prolidases and sialidases that can
268 degrade the effectors of inflammation. Whether there is a specific host response to NanH3 remains
269 to be determined.

270 The initial step in biofilm formation is adhesion, which in the case of BV associated
271 biofilms is to epithelial cells (2). *Streptococcus pneumoniae* surface proteins with sialidase activity
272 have been shown to reveal carbohydrate ligands for bacterial adhesion to host cells (18). Similarly,
273 Soong et al. (19) showed that a cell wall associated sialidase of *Pseudomonas aeruginosa* plays a
274 critical role in facilitating respiratory mucosa infection through biofilm formation. Specific
275 antibodies against *B. bifidum* SiaBb2, a cell wall tethered sialidase, inhibit adhesion to cells (21).
276 The involvement of *Gardnerella* sialidases in adhesion has not been demonstrated directly,
277 although it has been reported that inhibition of sialidase activity in *Gardnerella* strain JCP8066 by
278 an antiviral drug (Zanamavir) resulted in reduction of adherence to vaginal epithelial cells *in vitro*

279 (12). *Gardnerella* spp. can form multi-species biofilms *in vitro* (46) and although "*Gardnerella*
280 *vaginalis"* has been shown to participate in multispecies biofilms with other BV associated
281 bacteria *in vivo* (47, 48) and *in vitro* (49), it is not yet known how each of the *Gardnerella* spp.
282 participate in this process. Interestingly, *G. piotii* and *Gardnerella* genome sp. 3 (cpn60 subgroup
283 B) have been previously associated with "intermediate" grades of vaginal dysbiosis (7, 31, 50, 51)
284 and may contribute to transition from eubyosis to dysbiosis and enhance colonization by other BV
285 associated anaerobes. Phase variation in a cell surface associated sialidase enzyme might be critical
286 for turning on or off the initial adhesion of *Gardnerella* to epithelial cells and thus the cascade of
287 events following that result in the biofilm covered "clue cells" typical of BV (52).

288 It seems likely that some of the NanH3 protein produced by *G. piotii* does not remain
289 tethered to the cell surface, as is the case with *S. pneumoniae* (53), since the sequence upstream of
290 the C-terminal transmembrane domain may be susceptible to proteolytic cleavage, and
291 heterologously expressed NanH3 lacking the transmembrane domain is active (16). Taken
292 together, the activity of NanH3 while anchored to the cell surface or released into the extracellular
293 environment, and the likelihood that it is subject to phase variation through slipped-strand
294 mispairing, make this protein an intriguing puzzle for future study. Development of genetic tools
295 to create mutants where NanH3 production is locked "on" or "off" would be a significant step
296 toward understanding the role of this protein in the initiation and maintenance of vaginal dysbiosis.
297

298 **Acknowledgements**

299 This research was funded by a Natural Sciences and Engineering Council of Canada Discovery
300 Grant to JEH. The authors are grateful to Champika Fernando for excellent technical assistance,
301 and to all our colleagues in the Hill Lab for helpful discussions and feedback.

302

303 **References**

304 1. Nugent RP, Krohn MA, Hillier SL. 1991. Reliability of diagnosing bacterial vaginosis is improved
305 by a standardized method of gram stain interpretation. J Clin Microbiol 29:297–301.

306 2. Rosca AS, Castro J, Sousa LGV, Cerca N. 2020. *Gardnerella* and vaginal health: the truth is out
307 there. FEMS Microbiol Rev 44:73–105.

308 3. van de Wijgert JH, Borgdorff H, Verhelst R, Crucitti T, Francis S, Verstraelen H, Jespers V. 2014.
309 The vaginal microbiota: what have we learned after a decade of molecular characterization? PLoS
310 One 9:e105998.

311 4. Paramel Jayaprakash T, Schellenberg JJ, Hill JE. 2012. Resolution and characterization of distinct
312 cpn60-based subgroups of *Gardnerella vaginalis* in the vaginal microbiota. PLoS ONE 7:e43009.

313 5. Ahmed A, Earl J, Retchless A, Hillier SL, Rabe LK, Cherpes TL, Powell E, Janto B, Eutsey R,
314 Hiller NL, Boissy R, Dahlgren ME, Hall BG, Costerton JW, Post JC, Hu FZ, Ehrlich GD. 2012.
315 Comparative genomic analyses of 17 clinical isolates of *Gardnerella vaginalis* provide evidence of
316 multiple genetically isolated clades consistent with subspeciation into genovars. J Bacteriol
317 194:3922–37.

318 6. Vaneechoutte M, Guschin A, Van Simaey L, Gansemans Y, Van Nieuwerburgh F, Cools P. 2019.
319 Emended description of *Gardnerella vaginalis* and description of *Gardnerella leopoldii* sp. nov.,
320 *Gardnerella piotii* sp. nov. and *Gardnerella swidsinskii* sp. nov., with delineation of 13 genomic
321 species within the genus *Gardnerella*. Int J Syst Evol Microbiol 69:679–687.

322 7. Hill JE, Albert AYK. 2019. Resolution and co-occurrence patterns of *Gardnerella leopoldii,*
323 *Gardnerella swidsinskii, Gardnerella piotii* and *Gardnerella vaginalis* within the vaginal
324 microbiome. Infect Immun 87:e00532-19.

325 8. McGregor JA, French JI, Jones W, Milligan K, McKinney PJ, Patterson E, Parker R. 1994.
326 Bacterial vaginosis is associated with prematurity and vaginal fluid mucinase and sialidase: results
327 of a controlled trial of topical clindamycin cream. Am J Obstet Gynecol 170:1048–1059.

328 9. Myziuk L, Romanowski B, Johnson SC. 2003. BVBlue test for diagnosis of bacterial vaginosis. J
329 Clin Microbiol 41:1925–8.

330 10. Briselden AM, Moncla BJ, Stevens CE, Hillier SL. 1992. Sialidases (neuraminidases) in bacterial
331 vaginosis and bacterial vaginosis-associated microflora. J Clin Microbiol 30:663–6.

332 11. Lewis WG, Robinson LS, Gilbert NM, Perry JC, Lewis AL. 2013. Degradation, foraging, and
333 depletion of mucus sialoglycans by the vagina-adapted Actinobacterium *Gardnerella vaginalis*. J
334 Biol Chem 288:12067–79.

335 12. Govinden G, Parker JL, Naylor KL, Frey AM, Anumba DOC, Stafford GP. 2018. Inhibition of
336 sialidase activity and cellular invasion by the bacterial vaginosis pathogen *Gardnerella vaginalis*.
337 Arch Microbiol 200:1129–1133.

338 13. Schellenberg JJ, Paramel Jayaprakash T, Withana Gamage N, Patterson MH, Vaneechoutte M, Hill
339 JE. 2016. *Gardnerella vaginalis* subgroups defined by cpn60 sequencing and sialidase activity in
340 isolates from Canada, Belgium and Kenya. PLoS One 11:e0146510.

341 14. Santiago GL, Deschaght P, El Aila N, Kiama TN, Verstraelen H, Jefferson KK, Temmerman M,
342 Vaneechoutte M. 2011. *Gardnerella vaginalis* comprises three distinct genotypes of which only two
343 produce sialidase. Am J Obstet Gynecol 204:450.e1–7.

344 15. Corfield T. 1992. Bacterial sialidases—roles in pathogenicity and nutrition. Glycobiology 2:509–
345 521.

346 16. Robinson LS, Schwebke J, Lewis WG, Lewis AL. 2019. Identification and characterization of
347 NanH2 and NanH3, enzymes responsible for sialidase activity in the vaginal bacterium *Gardnerella*
348 *vaginalis*. J Biol Chem 294:5230–5245.

349 17. Potter RF, Burnham C-AD, Dantas G. 2019. In silico analysis of *Gardnerella* genomospecies
350 detected in the setting of bacterial vaginosis. Clin Chem 65:1375–1387.

351 18. Navarre WW, Schneewind O. 1999. Surface proteins of Gram-positive bacteria and mechanisms of
352 their targeting to the cell wall envelope. Microbiol Mol Biol Rev MMBR 63:174–229.

353 19. Soong G, Muir A, Gomez MI, Waks J, Reddy B, Planet P, Singh PK, Kanetko Y, Wolfgang MC,
354 Hsiao Y-S, Tong L, Prince A. 2006. Bacterial neuraminidase facilitates mucosal infection by
355 participating in biofilm production. J Clin Invest 116:2297–2305.

356 20. Parker D, Soong G, Planet P, Brower J, Ratner AJ, Prince A. 2009. The NanA neuraminidase of
357 *Streptococcus pneumoniae* is involved in biofilm formation. Infect Immun 77:3722–30.

358 21. Nishiyama K, Yamamoto Y, Sugiyama M, Takaki T, Urashima T, Fukiya S, Yokota A, Okada N,
359 Mukai T. 2017. *Bifidobacterium bifidum* extracellular sialidase enhances adhesion to the mucosal
360 surface and supports carbohydrate assimilation. MBio 8.

361 22. Woude MW van der, Bäumler AJ. 2004. Phase and antigenic variation in bacteria. Clin Microbiol
362 Rev 17:581–611.

363 23. Phillips ZN, Tram G, Seib KL, Atack JM. 2019. Phase-variable bacterial loci: how bacteria gamble
364 to maximise fitness in changing environments. Biochem Soc Trans 47:1131–1141.

365 24. Seib KL, Jen FE-C, Scott AL, Tan A, Jennings MP. 2017. Phase variation of DNA
366 methyltransferases and the regulation of virulence and immune evasion in the pathogenic *Neisseria*.
367 Pathog Dis 75.

368 25. Bergman M, Del Prete G, van Kooyk Y, Appelmelk B. 2006. Helicobacter pylori phase variation,
369 immune modulation and gastric autoimmunity. 2. Nat Rev Microbiol 4:151–159.

370 26. Nielsen H, Tsirigos KD, Brunak S, von Heijne G. 2019. A brief history of protein sorting prediction.
371 Protein J 38:200–216.

372 27. Khan S, Vancuren SJ, Hill JE. 2020. A generalist lifestyle allows rare *Gardnerella* spp. to persist at
373 low levels in the vaginal microbiome. bioRxiv 2020.08.25.267146.

374 28. Martin-Platero AM, Valdivia E, Maqueda M, Martinez-Bueno M. 2007. Fast, convenient, and
375 economical method for isolating genomic DNA from lactic acid bacteria using a modification of the
376 protein "salting-out" procedure. Anal Biochem 366:102–4.

377 29. Nugent RP, Krohn MA, Hillier SL. 1991. Reliability of diagnosing bacterial vaginosis is improved
378 by a standardized method of gram stain interpretation. J Clin Microbiol 29:297–301.

379 30. Schellenberg JJ, Patterson MH, Hill JE. 2017. *Gardnerella vaginalis* diversity and ecology in
380 relation to vaginal symptoms. Res Microbiol doi:10.1016/j.resmic.2017.02.011.

381 31. Balashov SV, Mordechai E, Adelson ME, Gygax SE. 2014. Identification, quantification and
382 subtyping of *Gardnerella vaginalis* in noncultured clinical vaginal samples by quantitative PCR. J
383 Med Microbiol 63:162–75.

384 32. Albert AY, Chaban B, Wagner EC, Schellenberg JJ, Links MG, van Schalkwyk J, Reid G,
385 Hemmingsen SM, Hill JE, Money D. 2015. A study of the vaginal microbiome in healthy Canadian
386 women utilizing cpn60-based molecular profiling reveals distinct *Gardnerella* subgroup community
387 state types. PLoS One 10:e0135620.

388 33. Bayliss CD, Hoe JC, Makepeace K, Martin P, Hood DW, Moxon ER. 2008. *Neisseria meningitidis*
389 escape from the bactericidal activity of a monoclonal antibody is mediated by phase variation of
390 lgtG and enhanced by a mutator phenotype. Infect Immun 76:5038–5048.

391 34. Apicella MA, Shero M, Jarvis GA, Griffiss JM, Mandrell RE, Schneider H. 1987. Phenotypic
392 variation in epitope expression of the *Neisseria gonorrhoeae* lipooligosaccharide. Infect Immun
393 55:1755–1761.

394 35. van Putten JP. 1993. Phase variation of lipopolysaccharide directs interconversion of invasive and
395 immuno-resistant phenotypes of *Neisseria gonorrhoeae*. EMBO J 12:4043–4051.

396 36. Lederberg J, Iino T. 1956. Phase variation in *Salmonella*. Genetics 41:743–757.

397 37. Silverman M, Zieg J, Hilmen M, Simon M. 1979. Phase variation in *Salmonella*: genetic analysis of
398 a recombinational switch. Proc Natl Acad Sci 76:391–395.

399 38. Bonifield HR, Hughes KT. 2003. Flagellar phase variation in *Salmonella enterica* is mediated by a
400 posttranscriptional control mechanism. J Bacteriol 185:3567–3574.

401 39. Fraser CM, Norris SJ, Weinstock GM, White O, Sutton GG, Dodson R, Gwinn M, Hickey EK,
402 Clayton R, Ketchum KA, Sodergren E, Hardham JM, McLeod MP, Salzberg S, Peterson J, Khalak
403 H, Richardson D, Howell JK, Chidambaram M, Utterback T, McDonald L, Artiach P, Bowman C,
404 Cotton MD, Fujii C, Garland S, Hatch B, Horst K, Roberts K, Sandusky M, Weidman J, Smith HO,
405 Venter JC. 1998. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete.
406 Science 281:375–388.

407 40. Pinto M, Borges V, Antelo M, Pinheiro M, Nunes A, Azevedo J, Borrego MJ, Mendonça J,
408 Carpinteiro D, Vieira L, Gomes JP. 2016. Genome-scale analysis of the non-cultivable *Treponema*
409 *pallidum* reveals extensive within-patient genetic variation. 1. Nat Microbiol 2:1–11.

410 41. Stern A, Brown M, Nickel P, Meyer TF. 1986. Opacity genes in *Neisseria gonorrhoeae*: Control of
411 phase and antigenic variation. Cell 47:61–71.

412 42. Stern A, Nickel P, Meyer TF, So M. 1984. Opacity determinants of *Neisseria gonorrhoeae*: Gene
413 expression and chromosomal linkage to the gonococcal pilus gene. Cell 37:447–456.

414 43. Cauci S, Scrimin F, Driussi S, Ceccone S, Monte R, Fant L, Quadrifoglio F. 1996. Specific immune
415 response against *Gardnerella vaginalis* hemolysin in patients with bacterial vaginosis. Am J Obstet
416 Gynecol 175:1601–5.

417 44. Cauci S, Driussi S, Guaschino S, Isola M, Quadrifoglio F. 2002. Correlation of local Interleukin-
418 1beta levels with specific IgA response against *Gardnerella vaginalis* cytolysin in women with
419 bacterial vaginosis. Am J Reprod Immunol 47:257–264.

420 45. Eade CR, Diaz C, Wood MP, Anastos K, Patterson BK, Gupta P, Cole AL, Cole AM. 2012.
421 Identification and characterization of bacterial vaginosis-associated pathogens using a
422 comprehensive cervical-vaginal epithelial coculture assay. PLoS One 7:e50106.

423 46. Khan S, Voordouw MJ, Hill JE. 2019. Competition among *Gardnerella* subgroups from the human
424 vaginal microbiome. Front Cell Infect Microbiol 9:374.

425 47. Machado A, Cerca N. 2015. Influence of biofilm formation by *Gardnerella vaginalis* and other
426 anaerobes on bacterial vaginosis. J Infect Dis 212:1856–61.

427 48. Machado D, Castro J, Palmeira-de-Oliveira A, Martinez-de-Oliveira J, Cerca N. 2015. Bacterial
428 vaginosis biofilms: Challenges to current therapies and emerging solutions. Front Microbiol 6:1528.

429 49. Castro J, Machado D, Cerca N. 2019. Unveiling the role of Gardnerella vaginalis in polymicrobial
430 Bacterial Vaginosis biofilms: the impact of other vaginal pathogens living as neighbors. ISME J
431 13:1306–1317.

432 50. Hilbert DW, Schuyler JA, Adelson ME, Mordechai E, Sobel JD, Gygax SE. 2017. *Gardnerella*
433 *vaginalis* population dynamics in bacterial vaginosis. Eur J Clin Microbiol Infect Dis
434 https://doi.org/10.1007/s10096-017-2933-8.

435 51. Shipitsyna E, Krysanova A, Khayrullina G, Shalepo K, Savicheva A, Guschin A, Unemo M. 2019.
436 Quantitation of all four *Gardnerella vaginalis* clades detects abnormal vaginal microbiota
437 characteristic of bacterial vaginosis more accurately than putative *G. vaginalis* sialidase A gene
438 count. Mol Diagn Ther 23:139–147.

439 52. Simoes JA, Discacciati MG, Brolazo EM, Portugal PM, Dini DV, Dantas MC. 2006. Clinical
440 diagnosis of bacterial vaginosis. Int J Gynaecol Obstet 94:28–32.

441 53. Lock RA, Paton JC, Hansman D. 1988. Purification and immunological characterization of
442 neuraminidase produced by *Streptococcus pneumoniae*. Microb Pathog 4:33–43.

443

444

445 **Table 1.** Distribution of *nanH2* and *nanH3* genes in *Gardnerella* spp. whole genome sequences

Isolate	cpn60 Subgroup	Species	Sialidase activity[1]	*nanH2*	*nanH3*
GH007	B	*G. piotii*	+	+	+
GH019		*G. piotii*	+	+	+
GH020		*G. piotii*	+	+	+
GH022		*G. piotii*	+	+	+
N170		Genome species 3	+	-	+
NR026		Genome species 3	+	-	+
VN002		*G. piotii*	+	+	-[2]
N144		Genome species 3	+	-	+
W11		Genome species 3	+	-	+
N101		Genome species 3	+	+	+
N153		Genome species 3	+	+	+
N95		Genome species 3	+	-	+
GH005	A	*G. leopoldii*	-	-	-
NR015		Unknown	-	-	-
NR016		*G. swidsinskii*	-	-	-
NR017		*G. leopoldii*	-	-	-
NR019		*G. leopoldii*	-	-	-
NR020		*G. swidsinskii*	-	-	-
NR021		*G. swidsinskii*	-	-	-
VN003		*G. leopoldii*	-	-	-
WP021		*G. swidsinskii*	-	-	-
WP022		*G. swidsinskii*	-	-	-
NR010		*G. leopoldii*	-	-	-
N072		*G. swidsinskii*	-	-	-
GH021	C	*G. vaginalis*	-	-	-
NR001		*G. vaginalis*	-	-	-
NR037		*G. vaginalis*	-	-	-
NR038		*G. vaginalis*	-	-	-
NR039		*G. vaginalis*	-	-	-
WP023		*G. vaginalis*	-	-	-
N165		*G. vaginalis*	-	-	-
GH015		*G. vaginalis*	+	-	+
NR003	D	Genome sp. 8	-	-	-
NR047		Unknown	-	-	-
WP012		Genome sp. 9	-	-	-
N160		Genome sp. 10	-	-	-

446 [1]Previously determined (13)
447 [2]PCR positive for *nanH3*
448

449

450

451 **Table 2.** Homopolymer length variants in *Gardnerella piotii* and genome sp. 3 isolates.

Isolate	Homopolymer sequence	Predicted peptide length (amino acids)[1]
W11	CAACTA- C_{11}-ATGAACAAA	43
	CAACTA- C_{12}-ATGAACAAA	812
	CAACTA- C_{13}-ATGAACAAA	38
	CAACTA- C_{14}-ATGAACAAA	44
VN014	CAACTA- C_{9}-TCGAACAAA	812
	CAACTA- C_{10}-TCGAACAAA	37
	CAACTA- C_{11}-TCGAACAAA	43
VN015	CAACTA- C_{9}-TCGAACAAA	812
	CAACTA- C_{10}-TCGAACAAA	37
	CAACTA- C_{11}-TCGAACAAA	43
	CAACTA- C_{12}-TCGAACAAA	812
	CAACTA- C_{13}- - -GAACAAA	43
NR032	CAACTA- C_{8}-ATGAACAAA	42
	CAACTA- C_{9}-ATGAACAAA	812

452 [1]Full-length NanH3 is 812 amino acids

453

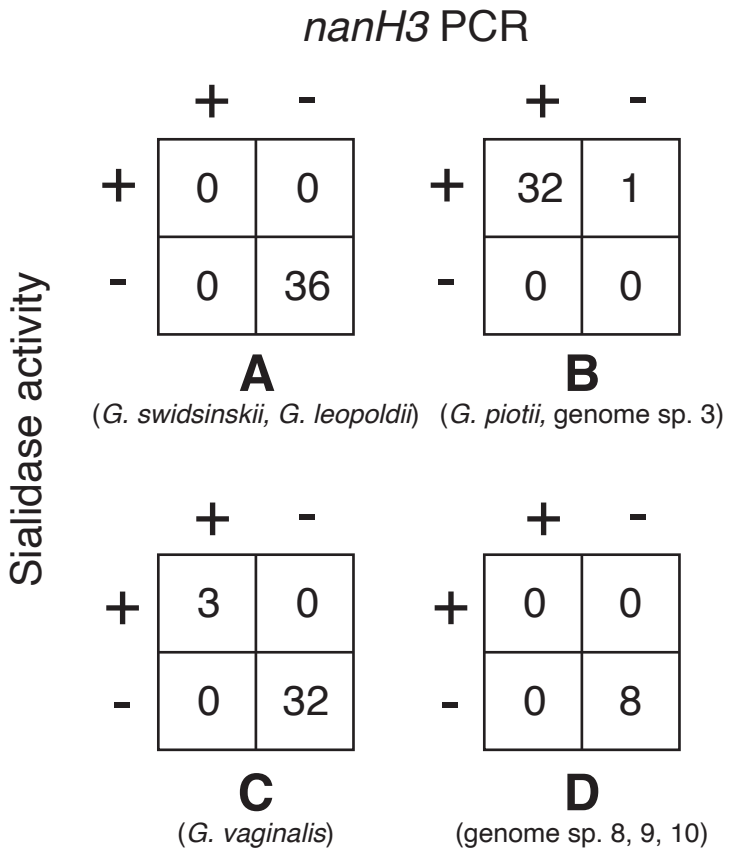
454

455

456

457

458



459

460 **Figure 1.** Correlation of *nanH3* and sialidase activity in 112 *Gardnerella* isolates. Numbers of

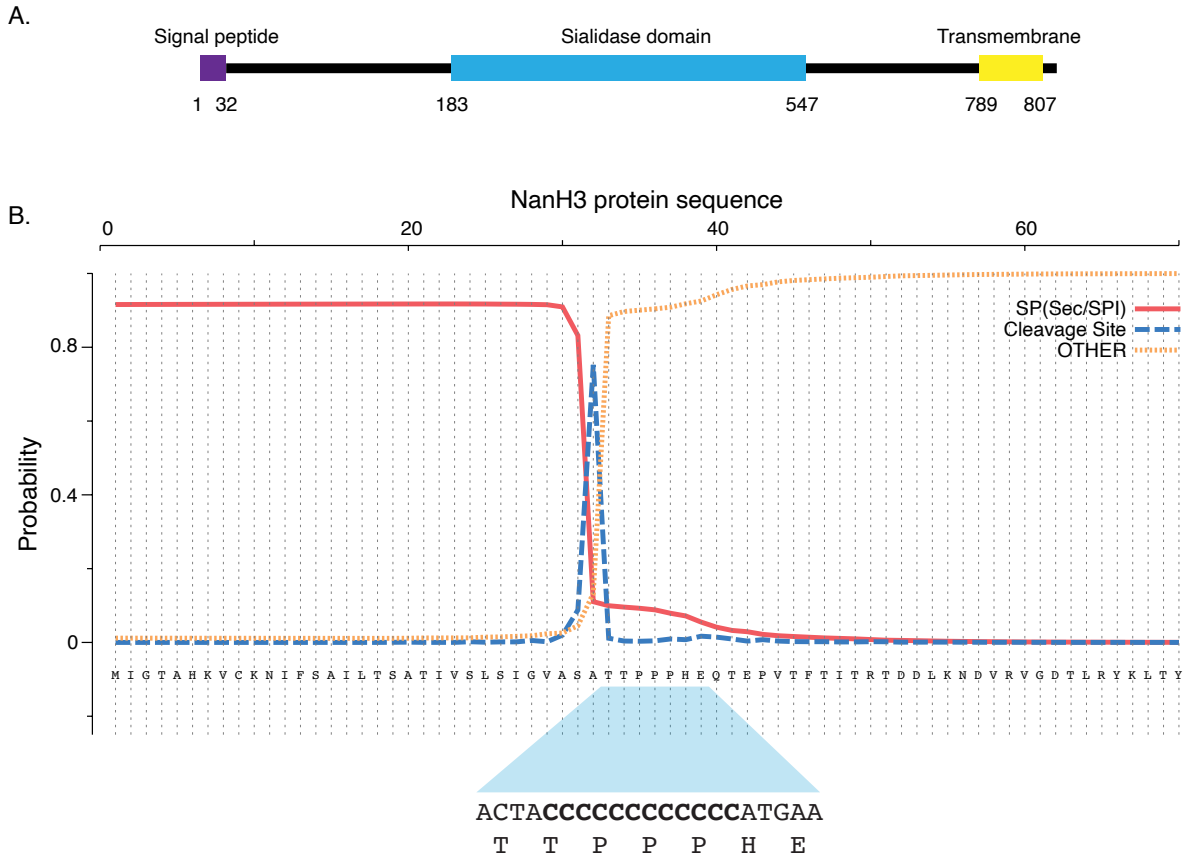
461 isolates positive or negative for sialidases activity and *nanH3* PCR. cpn60 subgroup affiliation and

462 species are indicated below each crosstab. Sialidase results were determined by Schellenberg *et*

463 *al*. (13).

464

465



466

467 **Figure 2.** A. Predicted locations of signal peptide, sialidase and transmembrane domains of

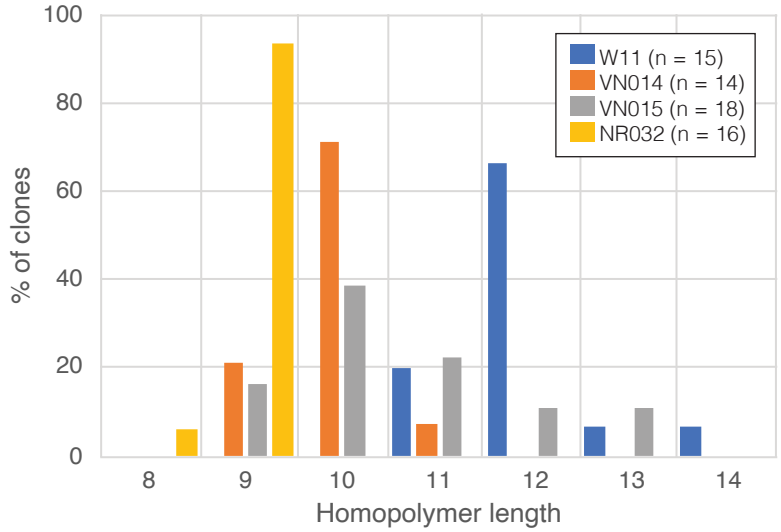
468 NanH3 from strain W11 based on InterProScan analysis. B. Location of predicted signal peptide,

469 cleavage site and adjacent homopolymer in the *nanH3* gene of strain W11. Analysis was

470 conducted with SignalP-5.0.

471

472



473

474 **Figure 3.** Frequencies of the homopolymer length variants in isolates W11, VN014, VN015 and

475 NR032.

476

477

478 **Supplemental Materials**

479

480 **Table S1.** *Gardnerella* strains used in the study, cpn60 subgroup affiliations, whole genome
481 sequence accession numbers, sialidase activity phenotypes, presence/absence of *nanH2, nanH3,*
482 signal peptide encoding sequences and homopolymers.

483

484 **Figure S1.** Sequencing of PCR products amplified directly from colonies of isolate W11.
485 PCR was performed with primers flanking the homopolymer region as described in the methods.
486 In each electropherogram, after eleven peaks of cytosine, the sequencing signal shows multiple
487 overlapping peaks suggesting the presence of multiple template sequences with different
488 homopolymer lengths.

489

490 **Figure S2.** Electropherograms of the homopolymer regions in two W11 colonies. Colony 6 had
491 individuals with 11, 12 and 13 C's in the homopolymeric region. Five out of eight colonies had
492 12C's in the homopolymer region. Colony 8 had individuals with 11, 12 and 14 C's in the
493 homopolymeric region. Five out of eight colonies had 12 C's in the homopolymer region. Having
494 12 C's makes the coding region in-frame. When there is 11,13 or 14 C's, a premature STOP codon
495 is generated possibly making the protein truncated.

496

497