Large scale genomic analysis shows no evidence for repeated pathogen adaptation during the invasive phase of bacterial meningitis in humans

John A. Lees[1], Philip H.C. Kremer[2], Ana S. Manso[3], Nicholas J. Croucher[4], Bart Ferwerda[2], Mercedes Valls Serón[2], Marco R. Oggioni[3], Julian Parkhill[1], Matthijs C. Brouwer[2], Arie van der Ende[5, 6], Diederik van de Beek[2], Stephen D. Bentley[1]*

Affiliations: [1] Wellcome Trust Sanger Institute, Hinxton, UK; [2]Department of Neurology, Center for Infection and Immunity Amsterdam (CINIMA), Academic Medical Center, Amsterdam, The Netherlands; [3]Department of Genetics, University of Leicester, Leicester, UK; [4]Department of Infectious Disease Epidemiology, St. Mary's Campus, Imperial College London, London, UK; [5]Department of Medical Microbiology, Center for Infection and Immunity Amsterdam (CINIMA), Academic Medical Center, Amsterdam, The Netherlands; [6]Netherlands Reference Laboratory for Bacterial Meningitis, Academic Medical Center, Amsterdam, The Netherlands

*Corresponding author: sdb@sanger.ac.uk

Abstract

Recent studies have provided evidence for rapid pathogen genome variation, some of which could potentially affect the course of disease. We have previously detected such variation by comparing isolates infecting the blood and cerebrospinal fluid (CSF) of a single patient during a case of bacterial meningitis. To determine whether the observed variation repeatedly occurs in cases of disease, we performed whole genome sequencing of paired isolates from blood and CSF of 938 meningitis patients. We also applied the same techniques to 54 paired isolates from the nasopharynx and CSF.

Using a combination of reference-free variant calling approaches we show that no genetic adaptation occurs in the invasive phase of bacterial meningitis for four major pathogen species: *Streptococcus pneumoniae*, *Neisseria meningitidis*, *Listeria monocytogenes* and *Haemophilus influenzae*. From nasopharynx to CSF , no adaptation was seen in *S. pneumoniae*, but in *N. meningitidis* mutations potentially mediating adaptation to the invasive niche were occasionally observed in the *dca* gene.

This study therefore shows that the bacteria capable of causing meningitis are already able to do this upon entering the blood, and no further sequence change is necessary to cross the blood-brain barrier. The variation discovered from nasopharyngeal isolates suggest that larger studies comparing carriage and invasion may help determine the likely mechanisms of invasiveness.

Author Summary

We have analysed the entire DNA sequence from bacterial pathogen isolates from cases of meningitis in 938 Dutch adults, focusing on comparing pairs of isolates from the patient's blood and their cerebrospinal fluid. Previous research has been on only a single patient, but showed possible signs of adaptation to treatment within the host over the course of a single case of disease.

By sequencing many more such paired samples, and including four different bacterial species, we were able to determine that adaptation of the pathogen does not occur after bloodstream invasion during bacterial meningitis.

We also analysed 54 pairs of isolates from pre- and post-invasive niches from the same patient. In *N. meningitidis* we found variation in the sequence of one gene which appears to provide bacteria with an advantage after invasion of the bloodstream.

Overall, our findings indicate that evolution after invasion in bacterial meningitis is not a major contribution to disease pathogenesis. Future studies should involve more extensive sampling between the carriage and disease niches, or on variation of the host.

Introduction

Bacterial meningitis is a severe inflammation of the meninges surrounding the brain as a response to the presence of bacteria [1]. This inflammation can

65 compromise brain function, requiring immediate admission to hospital. In
66 European countries, the four bacteria which most frequently cause meningitis
67 are *Streptococcus pneumoniae*, *Neisseria meningitidis*, *Haemophilus influenzae*
68 and *Listeria monocytogenes* [2].
69
70 The route of infection varies depending on the species of bacteria, though in the
71 majority of invasive cases the final stage is from blood to cerebrospinal fluid
72 (CSF) [1]. Respiratory pathogens (*S. pneumoniae*, *N. meningitidis* and *H.*
73 *influenzae*) are carried asymptomatically in the nasopharynx by a proportion of
74 the population at a given time [3, 4]. *L. monocytogenes* is food-borne infection
75 and can result from consumption of contaminated products [5, 6]. In a small
76 number of cases commensal nasopharyngeal or ingested food-borne bacteria
77 may invade the blood through a single cell bottleneck (bacteraemia) [7], then
78 cross the blood-brain barrier into the CSF where they cause meningitis [8]. In
79 some meningitis patients the CSF may be invaded directly due to CSF leakage or
80 otitis media [9], in which case the progression of bacteria after carriage is
81 reversed: CSF to blood.
82
83 Until recently it was thought that mutation rates in bacterial genomes were low,
84 and as such would not change within a single host [10]. However, many studies
85 sequencing bacterial populations of various different species gave estimates
86 three orders of magnitude higher than previously expected [11-13]. These new
87 estimates of mutation rate also gave evidence that DNA sequence variation can
88 occur over the course of a single infection [14].
89

Such within-host variation has been shown to occur through a variety of mechanisms such as recombination [15], gene loss [16, 17] and variation in regulatory regions [18-20]. The rapid variation that occurs in these regions of the genome can increase the population's fitness as the bacteria adapt to the host environment [21, 22], and potentially affect the course of disease [23]. Previous studies have shown variation between strains even during the rapid clinical progression of bacterial meningitis [24, 25].

It is possible that the existing genetics of the bacterium invading from the carriage population may determine, prior to blood stream invasion, whether CSF invasion is possible. Causing invasive disease is an evolutionary dead end for these pathogens, so studies of carriage will not observe selection for variations that advantage bacteria in the blood or CSF. Current knowledge is mostly focused at the serotype and MLST level, and lacks the resolution and sample size to be able to address this question [26-28]. Though the only whole genome based study suggests this is not case (at the gene level) in *S. pneumoniae* [29], we believe higher powered study designs are needed to better answer this question.

We also hypothesise that bacterial variation may also occur during the invasive phase of meningitis. We have previously reported in a single patient that the bacteria appeared to adapt to the distinct conditions of blood and CSF [24]. These are very different niches from that of nasopharyngeal carriage where this variation is well documented [30], not least because the bacteria are under more intense exposure to immune pressures and have less time over which to accumulate mutations.

To look for adaptation to these three niches, we used samples from the MeninGene study [31, 32], based at the Academic Medical Centre Amsterdam. 938 patients recruited to the study had culture positive bacterial meningitis with samples collected from both their blood and CSF (breakdown by species in Table S1), and 54 with pneumococcal or meningococcal meningitis have a matched sample taken from their nasopharynx. By whole genome sequence analysis of large numbers of paired bacterial isolates cultured from these samples, we have been able to test for repeated variation that occurs during the course of the disease.

Results

We made assumptions about the evolution of bacteria within the host, under which we discuss the power of pairwise comparisons between single colonies taken from each niche to capture repeated evolution occurring post-invasion:

1. There is a bottleneck of a single bacterium upon invasion into the first sterile niche (usually blood), which then founds the post-invasion population [7, 33].
2. A large invasive population is quickly established, as the population size approaches the carrying capacity of the blood/CSF. The population size is large enough for selection to operate efficiently.
3. As infection occurs in a mass transport system, populations are well mixed without any substructure. Therefore, the effective population size equals the census population size.
4. The bacterial growth curve within blood and CSF is similar.

Initially the population size is small, so selection is inefficient and the population-wide mutation rate is low. However, the eventual carrying capacity of the blood and CSF are large enough (>1.5x10^5) [34, 35] for beneficial mutations to fix rapidly. Due to the short generation time of around an hour [36], this carrying capacity is reached early in the course of the disease (after 1-2 days) [37].

Crucially, population sizes where selection acts efficiently [38] are reached even earlier than this – a few hours after invasion. Therefore, mutations with a selective advantage occurring after the first stages of infection will eventually become fixed in the niche's population. So, sequence comparison between colony picks from each niche is likely to find adaptation that has occurred post invasion.

Similarity of the bacterial growth curve within blood and CSF is an important assumption because in 45% of the pneumococcal cases there was evidence that CSF invasion happened before blood invasion (patients had a documented prior CSF leak, otitis media or sinusitis [39, 40]). This allows us to search for post-adaptation invasion that happens in either direction in this species. We investigated the validity of this assumption using analysis of data on the *ivr* locus (see Methods).

In carriage samples, although the population size is small [41] carriage episodes can persist over many months [42], therefore allowing the potential for mutations conferring an advantage in an invasive niche to arise. Additionally,

164 during carriage there is known to be population wide diversity [30] and in some
165 cases competition between strains [43]. We only sample a single strain from this
166 diverse pool, which means we have less power to detect mutations either side of
167 the bottleneck. Combined with the small sample size, this means only adaptive
168 mutations with large selective advantages will be discovered in this part of the
169 study. We therefore discuss our results from blood and CSF comparisons first, as
170 we have a higher power to detect variation between these niches.
171

172 **No repeated post-invasion adaptation in coding regions across species**

173 We performed comparisons between the pan-genome of each pair of blood and
174 CSF isolates, using reference-free variant calling techniques (see Methods). The
175 method was evaluated using simulated data, giving us confidence that it could
176 detect the small amounts of variation expected between each isolate pair (figures
177 S1, S2).
178

179 For each species we then counted the number of variants of any type between
180 each blood/CSF isolate pair taken from a patient. For all species except
181 *N. meningitidis* the majority of paired samples have no variation between them
182 (Fig. 1).
183

184 In *S. pneumoniae* 452 of 674 paired samples (67%) were identical. The
185 distribution is roughly Poisson (mean = 0.547), excluding outliers. In *H.*
186 *influenzae* and *L. monocytogenes* the observed number of mutations between
187 each paired set of strains is low, and similar in distribution to *S. pneumoniae*.

188 Variation between *N. meningitidis* pairs also followed a roughly Poisson
189 distribution (mean = 2.34), which when compared to other species showed a
190 higher number of variants between blood and CSF isolates (Wilcoxon rank-sum
191 test, W = 25790, p-value < 10^{-16}) such that most pairs have at least one variant
192 between the blood and CSF samples.
193
194 In *N. meningitidis* these mutations may be a signal of repeated adaptation
195 between the two niches if they cause the same functional change. Similarly, rare
196 mutations in the other three species, if they cause the same functional change,
197 could represent a signal of adaptation. To determine whether this is the case, we
198 counted the number of times each gene annotation contained variation between
199 the blood and CSF isolate over all the pairs collected, and used a Poisson test to
200 determine whether this was more than expected for each gene (see Methods).
201 For *L. monocytogenes* and *H. influenzae* there was not enough variation measured
202 in the samples to show any such signal. For *S. pneumoniae* the results are shown
203 in table 1.
204
205 **Table 1: Genes containing significantly repeated mutations between blood**
206 **and CSF isolate pairs in *S. pneumoniae*.** Ordered by increasing p-value; locus
207 tags refer to the D39 genome, if present.

Gene name	Gene length (bp)	Mutations between blood and CSF	p-value
pde1 (SPD_2032)	1973	19	$<10^{-10}$
dltD (SPD_2002)	1269	13	$<10^{-10}$
dltB (SPD_2004)	1245	12	$<10^{-10}$
dltA (SPD_2005)	1551	11	$<10^{-10}$
clpX (SPD_1399)	1233	7	1.3×10^{-8}
wcaJ (SPD_1620)	693	6	3.4×10^{-8}
cysB (SPD_0513)	909	5	1.6×10^{-5}

cbpJ	1122	5	4.7×10^{-5}
amiC (SPD_1670)	1332	4	6.0×10^{-3}
marR	435	3	9.6×10^{-3}
fhuC	519	3	1.6×10^{-2}

208

209 The *dlt* operon, responsible for D-alanylation in teichoic acids in the cell wall [44-
210 46], was the most frequently mutated region: 36 mutations in 31 sample pairs
211 (Poisson test $p<10^{-10}$).

212

213 Mapping the variation between sample pairs to the R6 *S. pneumoniae* strain,
214 which has a functional *dlt* operon, the variations were annotated with their
215 predicted function. There was no directionality to the mutations: 19 occurred in
216 the blood, and 11 in the CSF. Only seven of the patients infected by these strains
217 showed signs of blood invasion before CSF invasion (sinusitis or otitis); this also
218 did not show directionality. The nature of the mutations is shown in Fig. 2 and
219 Table S2. Most of these mutations would be expected to cause a loss of function
220 (LoF) in the operon. Though this suggests this locus has a deleterious effect in
221 invasive disease generally, the lack of directionality to the mutations means it
222 does not show evidence of adaptation to either the blood or CSF specifically.

223

224 In all the other genes in table 1 the variants are non-synonymous SNPs
225 distributed evenly between blood and CSF, therefore also showing no adaptation
226 specific to either niche.

227

228 The most frequently mutated genes between pairs in *N. meningitidis* are shown
229 in table 2.

230

Table 2: Genes containing significantly repeated mutations between blood and CSF isolate pairs in *N. meningitidis*. Ordered by increasing p-value; locus tags refer to the MC58 genome, if present.

Gene name	Gene length (bp)	Mutations between blood and CSF	p-value
pilE (NMB0018)	384	18	$<10^{-10}$
lgtC	189	16	$<10^{-10}$
hyaD	327	14	$<10^{-10}$
oatA	1869	19	$<10^{-10}$
hpuB (NMB1668)	2382	17	$<10^{-10}$
porA (NMB1429)	1178	12	$<10^{-10}$
lgtA (NMB1929)	1050	10	$<10^{-10}$
kfoC	360	7	$<10^{-10}$
cotSA	1134	7	9.2×10^{-9}
ssa1	3252	6	3.9×10^{-4}

Top ranked are those relating to the pilus: *pilE* (19), *pilC* (6) and *pilQ* (4). *pilE* genes are associated with immune interaction [47], and are therefore expected to be under diversifying selection; an excess of non-synonymous mutations (46/49 observed; 33/49 expected for neutral selection) is consistent with this. The other notable gene with more mutations than expected in *N. meningitidis* was *porA*, a variable region which is a major determinant of immune reaction [48], in which 12 samples had frameshift mutations in one of two positions. Phase variation in the gene's promoter region, affecting its expression, is discussed in more detail below.

The mutations in table 2 show no association with blood or CSF specifically, so do not represent adaptation to either niche. Genetic variation in *pilE*, *hpuA*, *wbpC*, *porA* and *lgtB* within host has been observed previously in a single patient with a hypermutating *N. meningitidis* infection [25]. The coding sequences with excess variants that are found in the samples analysed here include these genes. This

also suggests an elevated background rate in these sequences, rather than strong selection between the blood and CSF niches.

No evidence for repeated adaptation in intergenic regions in *S. pneumoniae* and *N. meningitidis*

Our previous result suggesting adaptation from blood to CSF was an intergenic change affecting the transcription the *patAB* genes, encoding an efflux pump. In general it is known that in pathogenic bacteria a common form of adaptation is mutation in intergenic regions, which may affect global transcription levels, causing a virulent phenotype [49, 50], antimicrobial resistance [51] and changing interaction with the host immune system [52]. Changes in these regions have previously been shown to display signs of adaptation during single cases of bacterial disease [18].

We therefore separately investigated the mutations in non-coding regions, which were only observed in *S. pneumoniae* and *N. meningitidis*. As genome annotations from the calling above are not consistent between samples outside of CDSs, we mapped the variation in intergenic regions to the coordinates of a reference genome. In a subset of samples we had carriage isolates corresponding to blood and CSF isolate pairs (discussed further below); we used these carriage isolates as the reference genome to determine whether these mutations occur in the blood or CSF isolate.

Figure S3a shows all mutations plotted genome-wide in *S. pneumoniae*. The peaks correspond to mutations in genes described in table 1. In the remaining 121 mutations in non-coding regions we observed no clustering by position. Over all pairs of samples, intergenic mutations were spread between blood and CSF isolates when compared to a carriage reference. This suggests none of the intergenic mutations are providing a selective advantage in either invasive niche.

The mutations in *N. meningitidis* are plotted in figure S4a, 110 of which were in non-coding regions. We observed enrichment, but no niche specificity, in the upstream region of six genes. These mutations are listed in table 3. Some of the mutations upstream of *porA* and *opc* are in phase variable homopolymeric tracts, which are discussed more fully in the section below. The other mutations are upstream of the adhesins *hsf*/NMB0992 and NMB1994, which are involved in colonisation [53] and immune interaction during invasion [54], and *frpB*/NMB1988 which is a surface antigen involved in iron uptake [55]. Differential expression of these genes may be an important factor affecting invasion, but the mutations we observed that may affect this do not appear to be specific to blood or CSF.

Table 3: Intergenic regions containing significantly repeated mutations between CSF and blood isolate pairs in *N. meningitidis*. Ordered by increasing number of mutations; coordinates refer to the MC58 genome.

Coordinates	Downstream gene(s)	Mutations between blood and CSF
1468329-1468331	*porA* (NMB1429)	7
1072215-1072328	*opc* (NMB1429)	7
1008872-1008985	*hsf* (NMB0992)	6

1315621-1315672	NMB1299	6
2092257-2092552	*frpB* (NMB1988)	5
2100124-2100258	NMB1994	4

295

296 **No evidence for repeated adaptation in phase variable regions in *S. pneumoniae***

297 **and *N. meningitidis***

298 Phase variable regions, which may also be intergenic, can mutate rapidly and are

299 known to be a significant source of variation in pathogenic bacteria [56]. This

300 mutation is an important mechanism of adaptation [57], and meningococcal

301 genomes in particular contain many of these elements [58].

302

303 In *N. meningitidis* we observed six samples with single base changes in length of

304 the phase-variable homopolymeric tract in the *porA* gene's promoter sequence,

305 and five samples with the single base length changes in the analogous promoter

306 sequence of *opc*. While changes in the length of these tracts will affect expression

307 of the corresponding genes, both of which are major determinants of immune

308 response [59, 60], the tract length does not correlate with blood or CSF

309 specifically. *porA* expression has previously been found to be independent of

310 whether isolates were taken from CSF, blood or throat [59].

311

312 In *S. pneumoniae*, recent publications highlight a potential role in virulence for

313 the *ivr* locus, a type I restriction-modification system with a phase-variable

314 specificity gene allele of *hsdS* in the host specificity domain (Fig. 3) [19, 20, 61].

315 There are six possible different alleles A-F (Figure S5) for *hsdS*, each

316 corresponding to a different level of capsule expression. Some of these alleles are

317 more successful in a murine model of invasion, whereas others are more

successful in carriage. We used a mapping based approach (see methods) to determine whether any of these alleles were associated with either the blood or CSF niche specifically, which could be a sign of adaptation.

As the locus inversion is rapid and occurs within host, we first ensured that cultured samples are representative of the original clinical samples using PCR quantification of each allele. However, as even a single colony contains heterogeneity at this locus, simply taking the allele with the most reads mapping to it in each sample gives a poor estimate of the overall presence of each allele in the blood and CSF niches. To take into account the mix of alleles present in each sample, and to calculate confidence intervals, we developed a hierarchical Bayesian model for the *ivr* allele (see Fig S6 and Methods). This simultaneously estimates the proportion of each colony pick with alleles A-F for both each individual isolate (π), and summed over all the samples in each niche (μ). We apply this over *i* samples and *c* niches (in this case *c* can be blood or CSF).

For each pair of blood and CSF samples listed in Table S1, the difference in allele prevalence $\pi_{CSF} - \pi_{blood}$ was calculated (table S3). All *S. pneumoniae* samples had a difference in mean of at least one allele (as the confidence intervals overlap zero), highlighting the speed at which this locus inverts.

While this means that between a single CSF and blood pair the allele at this locus usually changes, it is the mean of μ_c (corresponding to the mean allele frequency in each niche over all sample pairs) which tells us whether selection of an allele occurs in either the blood or CSF more generally. This is plotted in Fig. 4. As the

343 confidence intervals overlap, no particular allele is associated with either blood
344 or CSF *S. pneumoniae* isolates.
345
346 Previous work on a murine invasion model [19] has shown an increase in
347 proportion of alleles A and B over the course of infection. We did not observe the
348 same effect in our clinical samples, though the large confidence intervals from
349 the mathematical model suggest that genomic data with a small insert size
350 relative to the size of repeats in the locus is limited in resolving changes in this
351 allele. A small selective effect of *ivr* allele between these niches would therefore
352 not be detected, but we can rule out strong selection for a particular allele (odds
353 ratio > 2). Application to read data from a large carriage dataset may help
354 resolve whether the same effect does occur in humans, as it would provide a
355 greater temporal range over the course of pathogenesis.

356 **Carriage and invasive disease sample pairs show some evidence of repeated**
357 **adaptation**

358 Using the same methods, we also sequenced and analysed pairs of genomes from
359 54 patients that were collected from the nasopharynx and CSF. Six of these were
360 *S. pneumoniae*. In these *S. pneumoniae* samples, we detected only one sample
361 with any variation, which was a two base insertion upstream of the *gph* gene
362 (Figure 5). This is similar to the amount of mutation observed between the blood
363 and CSF isolates, which is expected given the similar sampling timeframes. While
364 we found that a functional *dlt* operon appears to have a deleterious effect in
365 invasive disease, we did not observe mutation between our carriage and disease

366 samples. However, this was expected given the small number of carriage samples
367 relative to the effect size detected for this operon.
368
369 Between the 48 *N. meningitidis* carriage and CSF isolate pairs small numbers of
370 mutations were common. We went on to search for regions enriched for
371 mutation, however in 8 samples we observed large numbers of mutations
372 clustered close together. These represent single recombination events, so when
373 analysing genes enriched for mutation we counted each recombination as a
374 single event [62, 63].
375
376 Table 4 shows the results of this analysis. Similar genes are mutated as in the
377 blood/CSF pairs, again with no specificity to either niche. In phase variable
378 intergenic regions, we observed four sample pairs with an insertion or deletion
379 in the *porA* promoter tract with no niche specificity. Otherwise, none of the
380 regions above showed enrichment for mutation in either niche. These
381 observations support the theory that these genes mutate at a higher rate but do
382 not confer a selective advantage in any of the three niches studied.
383
384 **Table 4: Genes containing significantly repeated mutations between**
385 **nasopharyngeal and CSF isolate pairs in *N. meningitidis*.** Ordered by
386 increasing p-value; locus tags refer to the MC58 genome, if present.

Gene name	Gene length (bp)	Mutations between nasopharynx and CSF	p-value
lgtA (NMB1929)	1050	6	5.0×10^{-7}
oatA	1869	6	1.5×10^{-5}
hyaD	327	4	2.6×10^{-5}
pilE (NMB0018)	384	4	3.8×10^{-3}
pilT (NMB0052)	1131	4	3.5×10^{-3}

dca (NMB0415)	444	3	1.1x10-2

387

388 A notable exception to this is the *dca* gene, a phase variable gene involved in
389 competence in *Neisseria gonorrhoea* but of unknown function in *N. meningitidis*
390 [64, 65], in which all mutations are protein truncating variants in the invasive
391 isolate. Similarly, though not reaching significance (due to the long length of the
392 genes) were the *ggt* (NMB1057) and *czcD* (NMB1732) genes in which three
393 recombinations occurred, all of which were in the invasive isolate of the pair.

394

395 The mutations in these three genes therefore may confer a selective advantage in
396 the invasive niche; the sequence at these loci in the invasive strains are the same
397 as the MC58 reference, an invasive isolate itself. *ggt* has previously shown to be
398 essential for *N. meningitidis* growth in CSF in rats [66], and metal exporters such
399 as *czcD* have been shown to increase virulence in a mouse sepsis model [67].
400 More such paired carriage and invasion samples would be needed to confirm this
401 is the case in human invasive disease.

402

403 ## Discussion

404 Previous studies have shown that substantial levels of genomic DNA sequence
405 variation occur in bacteria colonising or infecting human hosts [12, 14, 15] and
406 suggest that some of this variation may be due to selective adaptation [18, 22-24,
407 68]. Such adaptations during invasive bacterial disease could lead to new
408 insights into the processes of pathogenesis with the potential to inform
409 therapies[69, 70]. Therefore it is important to capture evidence of such
410 variation; we sought to do this through large-scale genomic analysis. We have

411 searched for variation in four different bacterial species comparing genomes
412 from bacteria isolated from both blood and CSF from the same individuals in 938
413 bacterial meningitis cases.
414
415 We found overall that the amount of variation within bacteria that occurs after
416 infection is low. The mutation observed is not randomly distributed throughout
417 the genome, but is randomly distributed between blood and CSF isolates. These
418 mutations are therefore an observation of a higher mutation rate in these
419 regions during invasion (for example the pilus in *N. meningitidis*, which is known
420 to be under diversifying selection) but not repeated adaptation to either niche.
421 This study indicates that our previous observation of variation between blood
422 and CSF isolates from a single case of meningitis was a rare event most likely
423 driven by antibiotic selection pressure during treatment.
424
425 We went on to analyse 54 samples comparing carriage and invasive isolates from
426 the same patient. Though the sample size was lower, and we did not fully sample
427 diversity within the nasopharynx, we were able to get an insight into potential
428 genetic differences between bacteria in these niches. We see some of the same
429 genes mutate rapidly between blood and CSF isolates also doing this between
430 carriage and invasion. This supports the conclusion that these genes have a
431 higher mutation rate, rather than giving a selective advantage to a niche. Finally,
432 we saw possible evidence for selection on the *dca* gene in *N. meningitidis*.
433
434 This study eliminates the need to search for bacterial diversity between invaded
435 host niches (blood and CSF) when trying to explain pathogenesis of meningitis.

436 However, our comparison between the genomes of carriage and invasive isolates
437 did show some weak signals of adaptation. Our power in these comparisons was
438 limited by small sample size and single colony sequencing. Future studies are
439 needed to comprehensively identify whether adaptation occurs through
440 bacterial genetics between carriage and invasion. This should be addressed
441 through either comparing more pairs of carriage and invasive isolates with
442 deeper sequencing, or large scale genome wide association studies.

443 ## Methods

444 ### Ethics statement

445 Written informed consent was obtained from all patients or their legally
446 authorized representatives. The studies were approved by the Medical Ethics
447 Committee of the Academic Medical Center, Amsterdam, The Netherlands
448 (approval number: NL43784.018.13).

449 ### Reference free variant calling

450 We extracted DNA from positive blood and CSF cultures from adults with
451 bacterial meningitis in the Netherlands from 2006 to 2012, and sequenced this
452 with 100bp paired end reads on the Illumina HiSeq platform.
453
454 To avoid reference bias, and missing variants in regions not present in an
455 arbitrarily chosen reference genome, we then performed reference free variant
456 calling between all sequence pairs of isolates using two methods: the 'hybrid'
457 method [71] and Cortex [72]. The former uses *de novo* assembly of the CSF
458 sequence reads, mapping of reads from both the blood and CSF samples back to
459 this sequence, then calling variants based on this mapping. Cortex uses an

assembly method that keeps track of variation between samples as it traverses the de Bruijn graph.

In the hybrid method we created draft assemblies of the CSF samples using SPAdes v3.5 [73], corrected these with SSpace and GapFiller [74], then mapped reads from both blood and CSF samples to this reference using SNAP [75] followed by variant calling with bcftools v1.1 [76].

For Cortex we first error corrected sample reads using quake [77], preventing false positive calls supported by very low coverage of reads. The joint workflow of cortex was then used with each set of corrected reads in its own path in the de Bruijn graph, and bubble calling was used to produce a second set of variants between samples. SNPs in the error corrected reads were also called using the graph-diff mode of SGA [78].

Simulations of closely related genomes

As the rate of variation is very low, we needed to ensure we had sufficient power to call variants and didn't suffer from an elevated false negative rate. We did this by simulating evolution of *S. pneumoniae* genomes along the branch of the tree between *S. pneumoniae* R6 [79] and the common ancestor with *Streptococcus mitis* B6 [80] (figure S7). The rates in the GTR matrix and insertion/deletion frequency distributions were estimated by aligning the R6 and B6 reference sequences with Progressive Cactus [81]. An average of 200 mutations with these rates were created in 100 sequences, and Illumina paired end read data at 200x coverage simulated using pIRS [82]. Variants between these sequences and a

484 draft R6 assembly from simulated read data were then called using both of the
485 above methods; comparison with the mutations known to be introduced allowed
486 power and false positive rate to be calculated – separately for SNPs (single base
487 substitution) and INDELs (one or more bases inserted or deleted).
488
489 In addition to *in silico* simulation, we cultured blood/CSF paired strains 4038
490 and 4039 [24] and resequenced them using the same 100bp Illumina paired end
491 sequencing as the rest of the isolates in the study. The genomes of strains 4038
492 and 4039 have been exhaustively analysed using multiple sequencing
493 technologies, so represent high quality positive control data to assess the calling
494 methods. Both methods were tested on these data.
495
496 The highest power was achieved using hybrid mapping for SNPs and Cortex for
497 INDELs: median power for calling SNPs was 90% using hybrid mapping, and
498 74% for INDELs using cortex (figures S1 and S2). This combination of methods,
499 mapping for SNPs and cortex for INDELs, was therefore used across all samples.
500 When applied to the paired strains 4038/4039 the same mutations as originally
501 reported are recovered, plus a 37bp insertion in *cysB* which was found to be
502 introduced during culturing.
503
504 We used simulations to compare against a simple method of mapping against an
505 arbitrary reference, in this case TIGR4 [83]. We found our reference free method
506 has greater power, especially for INDELs (figure S8), and a markedly reduced
507 false positive rate. We also used an assembly method alone to compare gene

508 presence and absence, but this too suffered from a vastly elevated false positive
509 rate (figure S9).

510 ## Tests for genes, intergenic regions and genotypes enriched for mutation

511 To scan for repeated variation, the number of mutations in each CDS annotation
512 (adjusted for CDS length) was counted. We then performed a single-tailed
513 Poisson test using the genome wide mutation rate per base pair multiplied by
514 the gene length as the expected number of mutations. The resulting p-values we
515 corrected for multiple testing using a Bonferroni correction with the total
516 number of genes tested as the *m* tests. Tables 1 and 2 report those CDS with an
517 adjusted p-value less than the significance level of 0.05. For intergenic regions,
518 anywhere with more than one variant is reported.
519
520 To test whether certain genotypic backgrounds are associated with a higher
521 number of mutations that occurs post-invasion, a linear fit of each MLST against
522 number of mutations between blood and CSF isolates was performed. The p-
523 values of the slope for each MLST were Bonferroni corrected; at a significance
524 level of 0.05 no MLST was associated with an increased number of mutations.
525 For genes the same test was performed, except samples were coded as one and
526 zero based on whether they had a mutation in the gene being tested or not. We
527 performed a logistic regression for each gene with over ten mutations in tables 1
528 and 2: no genes being mutated post invasion were associated with an MLST.

529 ## Copy number variation

530 We called copy number variations (CNVs) between samples by first mapping
531 each species to a single reference genome, then fitting the coverage of mapped

532 reads with a mixture of Poisson distributions [84]. Regions were ranked by the
533 number of sample pairs containing a discordant CNV call. As with regions of
534 enhanced SNP/INDEL variation, we performed a Poisson test for enrichment on
535 these regions.
536
537 In *S. pneumoniae* the most frequently varying region was due to poor quality
538 mapping of a prophage region. The only other region with $p < 0.05$ was a change
539 in copy number of 23S rRNA seen in a small number of sample pairs. In *N.*
540 *meningitidis* mismapping in the *pilE*/*pilS* region accounts for the only CNV
541 change. No CNV changes were observed in multiple sample pairs of either *L.*
542 *monocytogenes* or *H. influenzae*.

543 ## Variant annotation

544 To then be able to compare between samples using a consistent annotation, we
545 mapped the called variants to the *S. pneumoniae* ATCC 700669 reference [85] for
546 *S. pneumoniae*, and MC58 [86] for *N. meningitidis*. This was done by taking a 300
547 base window around each variant and using blastn on these with the reference
548 sequence. We used VEP [87] to annotate consequences of each variant. The
549 variants were visualised by plotting variant start positions between all pairs
550 against a reference genome, which allowed identification of clustering of
551 variation in unannotated regions.

552 ## *ivr* locus allele determination directly measured from clinical samples

553 As the rate of variation at this locus is fast compared to other types of variation
554 measured, we checked that culturing the bacteria does not cause the allele to
555 change from what is observed in the original clinical sample. We therefore

556 extracted DNA from a subset of 53 of 674 paired clinical CSF samples and the
557 respective bacterial isolates.
558
559 Allele prevalence was quantified using a combined nested PCR protocol based on
560 PCR amplification of the *ivr* locus, digesting with DraI and PleI followed by
561 quantitative analysis of banding patterns on a capillary electrophoresis [19].
562 Allele prevalence was identical between the original clinical sample and cultured
563 bacteria in 50 out of the 53 samples. The predictive power of the *in vitro*
564 detected *ivr* allele prevalence in a pneumococcal culture for the original allele
565 prevalence within the clinical sample is therefore sufficient to draw conclusions
566 from.

567 ### *ivr* locus allele determination from genomic data

568 There are six possible alleles A-F at this locus, though due to the high variation
569 rate and structural rearrangement mediating the change the allele cannot
570 reliably be determined using assembly and/or standard mapping of short read
571 data.
572
573 Instead, mates of reads mapping to the reverse strand of the conserved 5' region
574 were extracted for each sample, and mapped with BLAT [88] to the possible
575 alleles in position 1. This forms a vector r_i of length two for each sample i, with
576 the number of reads mapped to 1.1 and 1.2. Similarly, to determine the 3' allele
577 (position 2), pairs of reads mapping to each of the reverse strand of allele 1.1 and
578 the forward strand of allele 1.2 were extracted and mapped to the three possible

579 alleles in position 2 (figure S5). This forms a vector q_i of length six for each
580 sample *i*, with the number of reads mapped to each allele A-F.
581
582 From mapping, we found 621 sample pairs had at least one read mapping to an
583 allele of the *ivr* locus *hsdS* gene (table S4). Those without any reads mapping had
584 either a deletion of one component of the locus, or a large insertion mediated by
585 the *ivr* recombinase.

586 **Bayesian model for *ivr* allele**

587 We first modelled the state of the 5' allele (TRD1.*j*) only. For the two possible
588 alleles 1.1 and 1.2, the number of reads mapping to each allele (a 2-vector r_i) was
589 used as the number of successes in multinomial distribution z_c (*c* – index for
590 niche). From these we infered the proportion of each allele in each individual
591 sample π_i, and in each niche overall μ_c. This was done by defining Dirichlet priors
592 expressing the expected proportion of an allele in a given sample π_i to be drawn
593 from a Dirichlet hyperprior representing the proportion of the allele that is
594 found in each niche as a whole μ_c. The κ parameter sets the variance of all the
595 individual sample allele distributions π_{ic} about the tissue average μ_c, with a
596 higher κ corresponding to a smaller variance. This model is represented in figure
597 S6.
598
599 The hyperparameter $A\mu$, which encodes the total proportion of each allele we
600 expected to see over all samples, was set to the average amount of the allele
601 observed from the long range PCR in a subset of 53 paired samples, as described
602 above.

603

604 The observed number of reads mapping to each allele, prior distributions
605 defined above, and structure of the model in figure S6 defines a likelihood which
606 can be used to infer the most likely values of the parameters of interest π and μ.
607 We used Rjags to perform MCMC sampling to simulate the posterior distribution
608 of these parameters. We used 3 different starting points, and took and discarded
609 30000 burn in steps, followed by 45000 sampling steps. Noticeable auto-
610 correlation was seen between consecutive samples, so only every third step in
611 the chain was kept in sampling from the posterior. We manually inspected plots
612 of each hyperparameter value and mean at each point in the chain, as well as the
613 Gelman and Rubin convergence diagnostic, which showed that the chain had
614 converged over the sampling interval (figure S10).

615

616 To model both the 5' end (TRD 1.1 and 1.2) and the 3' end (TRD 2.1, 2.2 and 2.3)
617 together, so each isolate *i* is represented by an allele A-F, for each isolate the total
618 number of reads mapping n_i was drawn from the distribution in equation (1)

$$n_i \sim \pi_i \cdot r_i \tag{1}$$

619 where *j* is the index of the TRD region, r_{ij} is the number of reads in sample *i* that
620 had a mate pair downstream from TRD1.j mapping to any TRD2 region, and π_i is
621 the posterior for allele frequency in the sample.

622

623 The distribution for the number of reads mapping to each allele *j*, z_{ij} was defined
624 as in equation (2)

$$z_{i,j} \sim \begin{cases} n_i \cdot \dfrac{q_{i,j}}{\vec{q_i}} \cdot \pi_{i,1.1}, & \text{if } j \in \text{A, B, E} \\ n_i \cdot \dfrac{q_{i,j}}{\vec{q_i}} \cdot \pi_{i,1.2}, & \text{if } j \in \text{C, D, F} \end{cases} \tag{2}$$

where q_i is a vector of length six which contains the number of reads mapped to each allele A-F as described above, and π, i and n are as previously. A single sample for z was taken for each isolate i. This 6-vector z_{ij} is then used as the observed data in the same model as above to infer π_i, and μ_c for the whole locus allele (A-F) rather than just the 5' end.

For the 5' allele (TRD1.j) a model using a single κ parameter rather than a κ indexed by tissue c was preferred (change in deviance information criterion [89] ΔDIC = −0.523). For the 3' allele (TRD2.j), a model with a single κ parameter did not converge. A model with κ indexed by allele was used instead.

Diversity of *ivr* allele within samples

As the speed of inversion is rapid, the subsequent polymorphism of this locus was also used to evaluate our assumptions about diversity of the bacterial population within each niche. We calculated the Shannon index of each sample's vectors π_{CSF} and π_{blood} to measure diversity of the sample in each niche. The mean Shannon index across CSF samples was 1.01 (95% highest posterior density (HPD) 0.39-1.51) and 0.98 (95% HPD 0.35-1.55) in the blood. Looking at each sample pair individually, the difference between diversity in each niche appears normally distributed with a mean of zero (figure S11). Together, these observations suggest a similar rate of diversity generation in each niche. This is in line with our assumption that the two populations have similar mutation rates, and a similar number of generations between being founded and being sampled.

Acknowledgements

Thanks to Win Kit Man for DNA isolation of bacteria. Simon Harris and James Hadfield developed the visualization tool used plot variants across the genome (http://jameshadfield.github.io/phandango/). The study benefitted from collaboration in ESGIB.

Funding

Work at the Wellcome Trust Sanger Institute was supported by Wellcome Trust core funding (098051 https://wellcome.ac.uk/). J.A.L. is supported by a Medical Research Council studentship grant (1365620 http://www.mrc.ac.uk/). This work was also supported by grants from the European Research Council (ERC Starting Grant [proposal/contract 281156] https://erc.europa.eu/) and Netherlands Organization for Health Research and Development (ZonMw; NWO-Vidi grant 2010 [proposal/contract 016.116.358] http://www.zonmw.nl/), both to D.v.d.B. The Netherlands Reference Laboratory for Bacterial Meningitis is supported by the National Institute for Health and Environmental Protection, Bilthoven (http://www.rivm.nl). NJC is funded by a Sir Henry Dale Fellowship, jointly funded by the Wellcome Trust and Royal Society (Grant Number 104169/Z/14/Z https://wellcome.ac.uk/, https://royalsociety.org/). MRO was supported by Medical Research Council grant (MR/M003078/1 http://www.mrc.ac.uk/).

Figure captions

Fig 1: Histograms binned by number of variants between a blood/CSF sample pair, for each bacterial species. SNPs are from mapping, INDELs are from cortex. Three *S. pneumoniae* and one *N. meningitidis* sample with over 10 variants not shown.

Fig 2: Mutations observed between all paired samples in the *dlt* operon. The operon consists of four genes in the three reading frames of the reverse strand. Mutations, displayed by type, in the blood strains are shown above the operon, and in the CSF strains below the operon.

Fig 3: The structure of the *ivr* type I restriction-modification locus in *S. pneumoniae*. The restriction (*hsdR*) and methylation (*hsdM*) subunits, and the 5' end of the specificity subunit (*hsdS*) are generally conserved. Inverted repeats IR1 (85bp) and IR2 (333bp) facilitate switching of downstream incomplete *hsdS* elements into the transcribed region. Top: The green read pair has the expected insert size, and suggests allele A (1.1, 2.1) is present. The red read pair is in the wrong orientation and has an anomalously large insert size. Bottom: The red read pair is consistent with the displayed inversion, suggesting allele D (1.2, 2.1) is present.

Fig 4: Mean and 95% highest posterior density (HPD) for μ_c. This shows the proportion of each allele present in each of blood (red) and CSF (turquoise) tissues pooling across all samples.

695 **Fig 5: Histograms binned by number of variants between a carriage/CSF**

696 **sample pair, for each bacterial species.** a) As figure 1. In *N. meningitidis* eleven

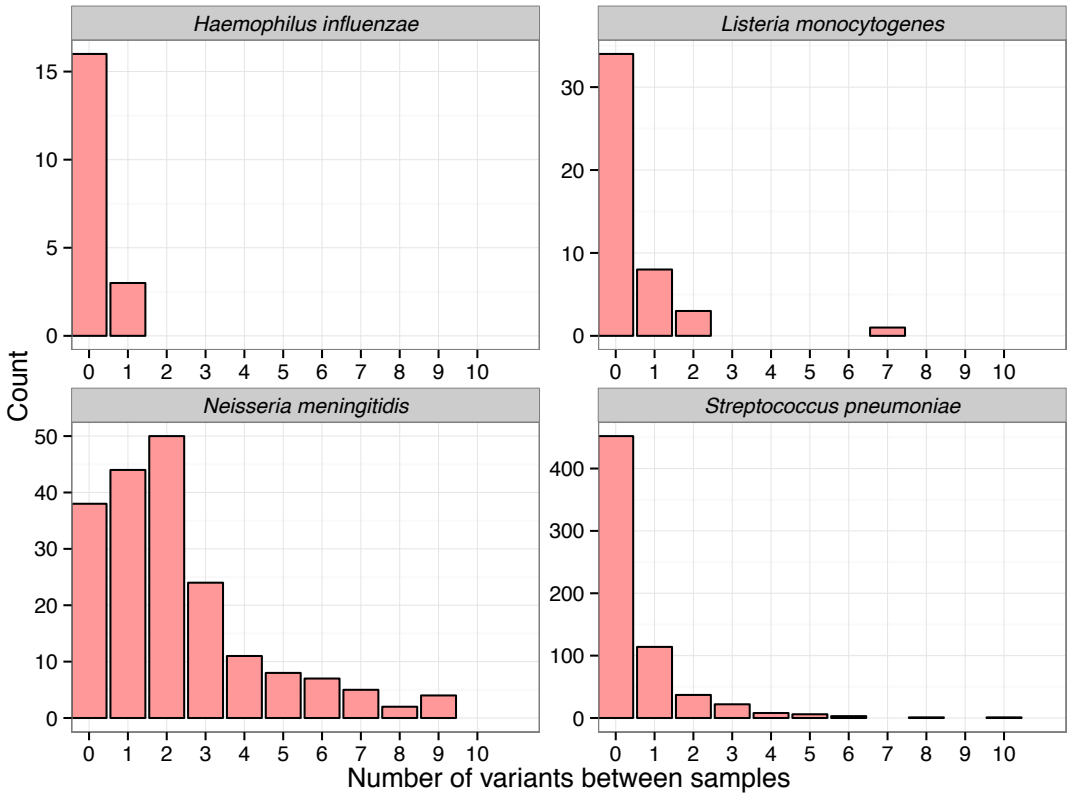
697 samples with over ten variants between them due to recombination events are

698 grouped. b) The number of recombination and SNP/INDEL events in samples in

699 the group with over ten detected variants

700 ## Figures

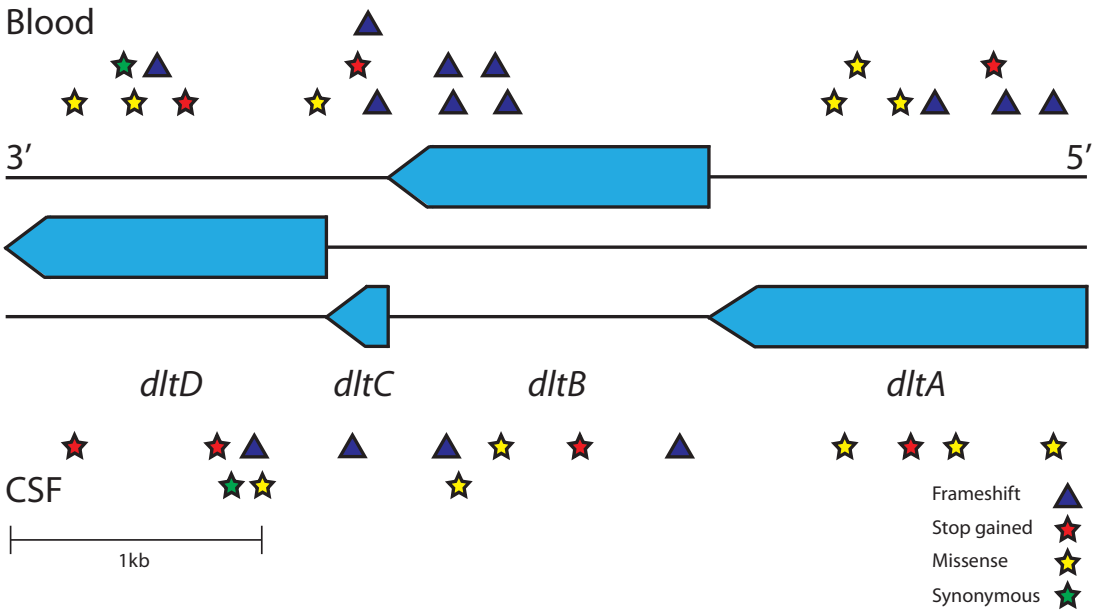
701 Figure 1



702

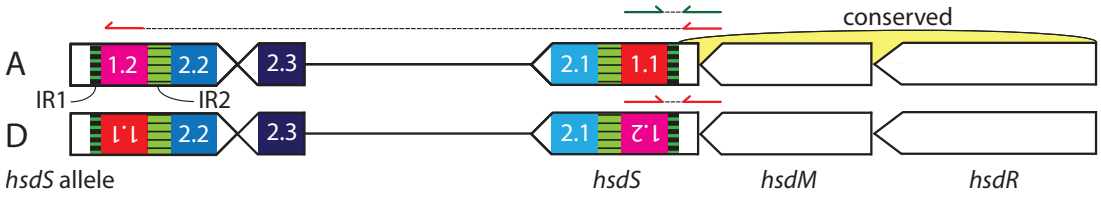
703

704 Figure 2



705

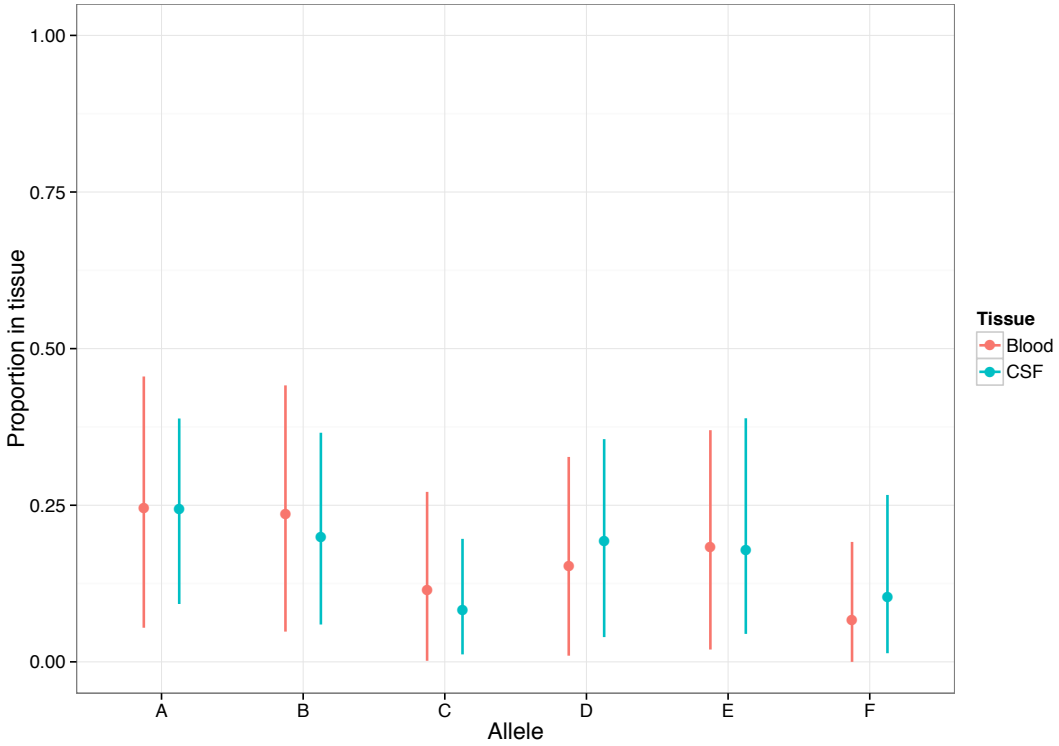
706 Figure 3



707

708

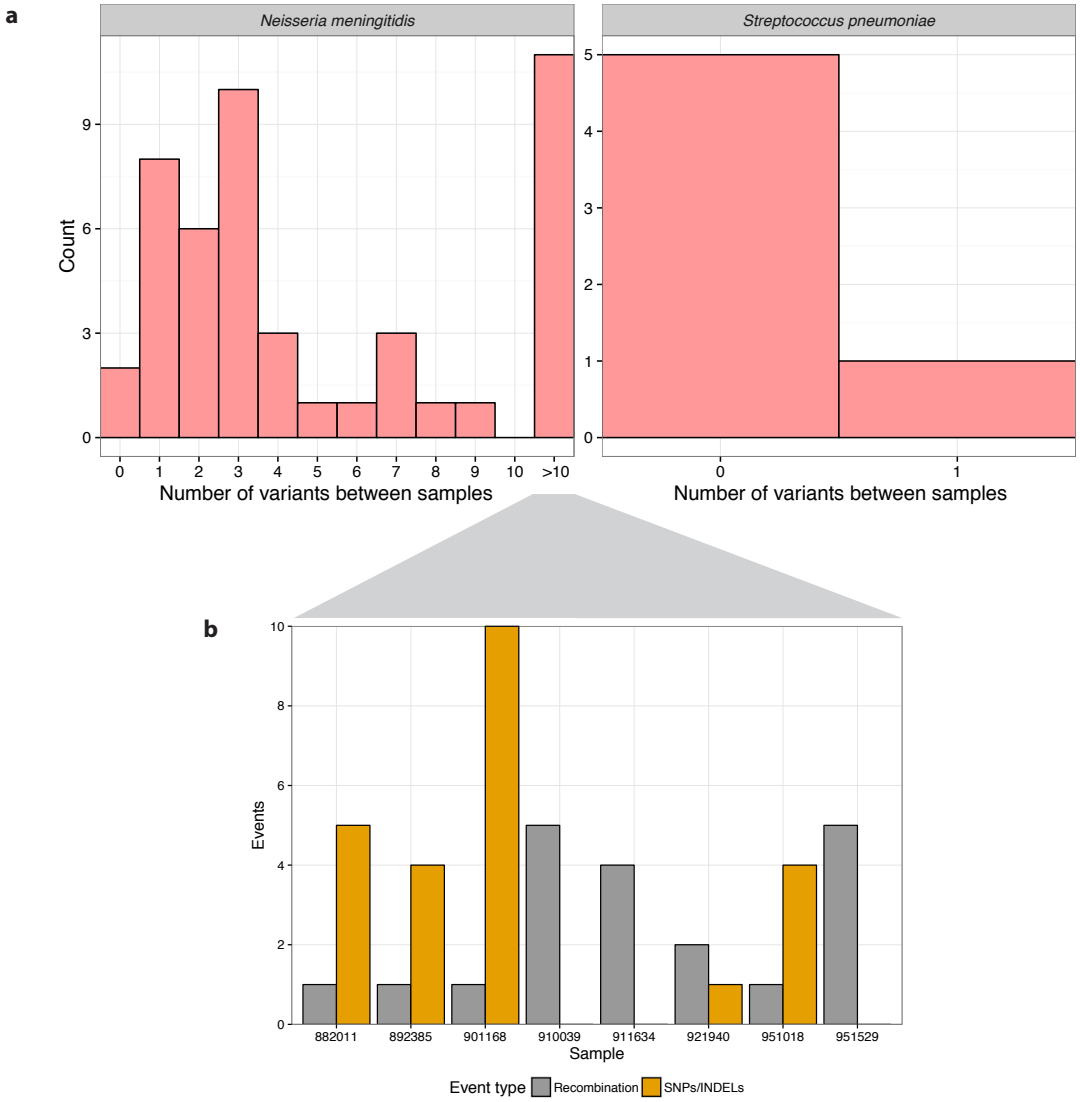
709 Figure 4



710

711

712 Figure 5



713

714

References

1. Mook-Kanamori BB, Geldhoff M, van der Poll T, van de Beek D. Pathogenesis and pathophysiology of pneumococcal meningitis. Clinical microbiology reviews. 2011;24:557-91. doi: 10.1128/CMR.00008-11. PubMed PMID: 21734248.

2. Brouwer MC, Tunkel AR, van de Beek D. Epidemiology, diagnosis, and antimicrobial treatment of acute bacterial meningitis. Clinical microbiology reviews. 2010;23:467-92. doi: 10.1128/CMR.00070-09. PubMed PMID: 20610819.

3. Hammitt LL, Bruden DL, Butler JC, Baggett HC, Hurlburt DA, Reasonover A, et al. Indirect effect of conjugate vaccine on adult carriage of Streptococcus pneumoniae: an explanation of trends in invasive pneumococcal disease. The Journal of infectious diseases. 2006;193:1487-94. doi: 10.1086/503805. PubMed PMID: 16652275.

4. Caugant DA, Hoiby EA, Magnus P, Scheel O, Hoel T, Bjune G, et al. Asymptomatic carriage of Neisseria meningitidis in a randomly sampled population. J Clin Microbiol. 1994;32:323-30.

5. Farber JM, Peterkin PI. Listeria monocytogenes, a food-borne pathogen. Microbiological Reviews. 1991;55:476-511.

6. Doyle MP, Glass KA, Beery JT, Garcia GA, Pollard DJ, Schultz RD. Survival of Listeria monocytogenes in milk during high-temperature, short-time pasteurization. Applied and Environmental Microbiology. 1987;53:1433-8.

7. Gerlini A, Colomba L, Furi L, Braccini T, Manso AS, Pammolli A, et al. The Role of Host and Microbial Factors in the Pathogenesis of Pneumococcal

739 Bacteraemia Arising from a Single Bacterial Cell Bottleneck. PLoS Pathogens.
740 2014;10. doi: 10.1371/journal.ppat.1004026. PubMed PMID: 24651834.
741 8. Weisfelt M, van de Beek D, Spanjaard L, Reitsma JB, de Gans J. Clinical
742 features, complications, and outcome in adults with pneumococcal meningitis: a
743 prospective case series. The Lancet Neurology. 2006;5:123-9. doi:
744 10.1016/S1474-4422(05)70288-X. PubMed PMID: 16426988.
745 9. Adriani KS, Brouwer MC, Beek DVD. Risk factors for community-acquired
746 bacterial meningitis in adults. The Netherlands Journal of Medicine. 2015;73:53-
747 60. PubMed PMID: 25753069.
748 10. Ochman H, Elwyn S, Moran NA. Calibrating bacterial evolution. Proc Natl
749 Acad Sci U S A. 1999;96(22):12638-43. PubMed PMID: 10535975; PubMed
750 Central PMCID: PMCPMC23026.
751 11. Wilson DJ, Gabriel E, Leatherbarrow AJ, Cheesbrough J, Gee S, Bolton E, et
752 al. Rapid evolution and the importance of recombination to the gastroenteric
753 pathogen Campylobacter jejuni. Mol Biol Evol. 2009;26(2):385-97. doi:
754 10.1093/molbev/msn264. PubMed PMID: 19008526; PubMed Central PMCID:
755 PMCPMC2639114.
756 12. Morelli G, Didelot X, Kusecek B, Schwarz S, Bahlawane C, Falush D, et al.
757 Microevolution of Helicobacter pylori during prolonged infection of single hosts
758 and within families. PLoS Genet. 2010;6(7):e1001036. doi:
759 10.1371/journal.pgen.1001036. PubMed PMID: 20661309; PubMed Central
760 PMCID: PMCPMC2908706.
761 13. Bryant JM, Grogono DM, Greaves D, Foweraker J, Roddick I, Inns T, et al.
762 Whole-genome sequencing to identify transmission of Mycobacterium abscessus
763 between patients with cystic fibrosis: a retrospective cohort study. Lancet.

764 2013;381(9877):1551-60. doi: 10.1016/S0140-6736(13)60632-7. PubMed
765 PMID: 23541540; PubMed Central PMCID: PMCPMC3664974.
766 14. Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A, O'Connor L, et al.
767 Diverse sources of C. difficile infection identified on whole-genome sequencing. N
768 Engl J Med. 2013;369(13):1195-205. doi: 10.1056/NEJMoa1216064. PubMed
769 PMID: 24066741; PubMed Central PMCID: PMCPMC3868928.
770 15. Kennemann L, Didelot X, Aebischer T, Kuhn S, Drescher B, Droege M, et al.
771 Helicobacter pylori genome evolution during human infection. Proc Natl Acad Sci
772 U S A. 2011;108(12):5033-8. doi: 10.1073/pnas.1018444108. PubMed PMID:
773 21383187; PubMed Central PMCID: PMCPMC3064335.
774 16. Rau MH, Marvig RL, Ehrlich GD, Molin S, Jelsbak L. Deletion and
775 acquisition of genomic content during early stage adaptation of Pseudomonas
776 aeruginosa to a human host environment. Environ Microbiol. 2012;14(8):2200-
777 11. doi: 10.1111/j.1462-2920.2012.02795.x. PubMed PMID: 22672046.
778 17. Ehrlich GD, Ahmed A, Earl J, Hiller NL, Costerton JW, Stoodley P, et al. The
779 distributed genome hypothesis as a rubric for understanding evolution in situ
780 during chronic bacterial biofilm infectious processes. FEMS Immunology and
781 Medical Microbiology. 2010;59:269-79. doi: 10.1111/j.1574-695X.2010.00704.x.
782 PubMed PMID: 20618850.
783 18. Marvig RL, Sommer LM, Molin S, Johansen HK. Convergent evolution and
784 adaptation of Pseudomonas aeruginosa within patients with cystic fibrosis. Nat
785 Genet. 2015;47(1):57-64. doi: 10.1038/ng.3148. PubMed PMID: 25401299.
786 19. Manso AS, Chai MH, Atack JM, Furi L, De Ste Croix M, Haigh R, et al. A
787 random six-phase switch regulates pneumococcal virulence via global epigenetic
788 changes. Nature Communications. 2014;5:5055. doi: 10.1038/ncomms6055.

789 20. Li J, Li JW, Feng Z, Wang J, An H, Liu Y, et al. Epigenetic Switch Driven by
790 DNA Inversions Dictates Phase Variation in Streptococcus pneumoniae. PLoS
791 Pathog. 2016;12(7):e1005762. doi: 10.1371/journal.ppat.1005762. PubMed
792 PMID: 27427949; PubMed Central PMCID: PMCPMC4948785.
793 21. Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, et al. Genome
794 evolution and adaptation in a long-term experiment with Escherichia coli.
795 Nature. 2009;461(7268):1243-7. doi: 10.1038/nature08480. PubMed PMID:
796 19838166.
797 22. Yang L, Jelsbak L, Marvig RL, Damkiaer S, Workman CT, Rau MH, et al.
798 Evolutionary dynamics of bacteria in a human host environment. Proc Natl Acad
799 Sci U S A. 2011;108(18):7481-6. doi: 10.1073/pnas.1018249108. PubMed PMID:
800 21518885; PubMed Central PMCID: PMCPMC3088582.
801 23. Young BC, Golubchik T, Batty EM, Fung R, Larner-Svensson H, Votintseva
802 AA, et al. Evolutionary dynamics of Staphylococcus aureus during progression
803 from carriage to disease. Proc Natl Acad Sci U S A. 2012;109(12):4550-5. doi:
804 10.1073/pnas.1113219109. PubMed PMID: 22393007; PubMed Central PMCID:
805 PMCPMC3311376.
806 24. Croucher NJ, Mitchell AM, Gould Ka, Inverarity D, Barquist L, Feltwell T, et
807 al. Dominant role of nucleotide substitution in the diversification of serotype 3
808 pneumococci over decades and during a single infection. PLoS genetics.
809 2013;9:e1003868. doi: 10.1371/journal.pgen.1003868. PubMed PMID:
810 24130509.
811 25. Omer H, Rose G, Jolley Ka, Frapy E, Zahar JR, Maiden MCJ, et al. Genotypic
812 and phenotypic modifications of Neisseria meningitidis after an accidental

813 human passage. PloS one. 2011;6. doi: 10.1371/journal.pone.0017145. PubMed
814 PMID: 21386889.
815 26. Brueggemann AB, Griffiths DT, Meats E, Peto T, Crook DW, Spratt BG.
816 Clonal Relationships between Invasive and Carriage Streptococcus pneumoniae
817 and Serotype- and Clone-Specific Differences in Invasive Disease Potential.
818 2003;187.
819 27. Robinson DA, Edwards KM, Waites KB, Briles DE, Crain MJ, Hollingshead
820 SK. Clones of Streptococcus pneumoniae Isolated from Nasopharyngeal Carriage
821 and Invasive Disease in Young Children in Central Tennessee. Journal of
822 Infectious Diseases. 2001;183:1501-7. doi: 10.1086/320194.
823 28. del Amo E, Selva L, de Sevilla MF, Ciruela P, Brotons P, Triviño M, et al.
824 Estimation of the invasive disease potential of Streptococcus pneumoniae in
825 children by the use of direct capsular typing in clinical specimens. European
826 Journal of Clinical Microbiology & Infectious Diseases. 2015;34:705-11. doi:
827 10.1007/s10096-014-2280-y.
828 29. Kulohoma BW, Cornick JE, Chaguza C, Yalcin F, Harris SR, Gray KJ, et al.
829 Comparative genomic analysis of meningitis and bacteremia causing
830 pneumococci identifies a common core genome. Infection and Immunity.
831 2015:IAI.00814-15. doi: 10.1128/IAI.00814-15.
832 30. Cremers AJ, Zomer AL, Gritzfeld JF, Ferwerda G, van Hijum SA, Ferreira
833 DM, et al. The adult nasopharyngeal microbiome as a determinant of
834 pneumococcal acquisition. Microbiome. 2014;2:44. doi: 10.1186/2049-2618-2-
835 44. PubMed PMID: 25671106.
836 31. Van De Beek D. Progress and challenges in bacterial meningitis. The
837 Lancet. 2012;380:1623-4. doi: 10.1016/S0140-6736(12)61808-X.

838 32. Woehrl B, Brouwer MC, Murr C, Heckenberg SGB, Baas F, Pfister HW, et al.
839 Complement component 5 contributes to poor disease outcome in humans and
840 mice with pneumococcal meningitis. Journal of Clinical Investigation.
841 2011;121:3943-53. doi: 10.1172/JCI57522. PubMed PMID: 21926466.
842 33. Moxon ER, Murphy PA. Haemophilus influenzae bacteremia and
843 meningitis resulting from survival of a single organism. Proceedings of the
844 National Academy of Sciences of the United States of America. 1978;75:1534-6.
845 doi: 10.1073/pnas.75.3.1534. PubMed PMID: 306628.
846 34. La Scolea LJ, Dryja D. Quantitation of bacteria in cerebrospinal fluid and
847 blood of children with meningitis and its diagnostic significance. Journal of
848 clinical microbiology. 1984;19:187-90. PubMed PMID: 6365957.
849 35. Brown PD, Davies SL, Speake T, Millar ID. Molecular mechanisms of
850 cerebrospinal fluid production. Neuroscience. 2004;129(4):957-70. doi:
851 10.1016/j.neuroscience.2004.07.003. PubMed PMID: 15561411; PubMed Central
852 PMCID: PMCPMC1890044.
853 36. Allegrucci M, Hu FZ, Shen K, Hayes J, Ehrlich GD, Post JC, et al. Phenotypic
854 characterization of Streptococcus pneumoniae biofilm development. J Bacteriol.
855 2006;188(7):2325-35. doi: 10.1128/JB.188.7.2325-2335.2006. PubMed PMID:
856 16547018; PubMed Central PMCID: PMCPMC1428403.
857 37. Gang TB, Hanley GA, Agrawal A. C-reactive protein protects mice against
858 pneumococcal infection via both phosphocholine-dependent and
859 phosphocholine-independent mechanisms. Infect Immun. 2015;83(5):1845-52.
860 doi: 10.1128/IAI.03058-14. PubMed PMID: 25690104; PubMed Central PMCID:
861 PMCPMC4399050.

862 38. Patwa Z, Wahl LM. The fixation probability of beneficial mutations. J R Soc
863 Interface. 2008;5(28):1279-89. doi: 10.1098/rsif.2008.0248. PubMed PMID:
864 18664425; PubMed Central PMCID: PMCPMC2607448.
865 39. Brouwer MC, Heckenberg SG, de Gans J, Spanjaard L, Reitsma JB, van de
866 Beek D. Nationwide implementation of adjunctive dexamethasone therapy for
867 pneumococcal meningitis. Neurology. 2010;75(17):1533-9. doi:
868 10.1212/WNL.0b013e3181f96297. PubMed PMID: 20881273.
869 40. Heckenberg SG, Brouwer MC, van der Ende A, van de Beek D. Adjunctive
870 dexamethasone in adults with meningococcal meningitis. Neurology.
871 2012;79(15):1563-9. doi: 10.1212/WNL.0b013e31826e2684. PubMed PMID:
872 22972648.
873 41. Li Y, Thompson CM, Trzcinski K, Lipsitch M. Within-host selection is
874 limited by an effective population of Streptococcus pneumoniae during
875 nasopharyngeal colonization. Infect Immun. 2013;81(12):4534-43. doi:
876 10.1128/IAI.00527-13. PubMed PMID: 24082074; PubMed Central PMCID:
877 PMCPMC3837969.
878 42. Turner P, Turner C, Jankhot A, Helen N, Lee SJ, Day NP, et al. A
879 longitudinal study of streptococcus pneumoniae carriage in a cohort of infants
880 and their mothers on the Thailand-Myanmar border. PloS one. 2012;7. doi:
881 10.1371/journal.pone.0038271. PubMed PMID: 22693610.
882 43. Cobey S, Lipsitch M. Niche and Neutral Effects of Acquired Immunity
883 Permit Coexistence of Pneumococcal Serotypes. Science. 2012;335(6074):1376-
884 80. doi: 10.1126/science.1215947.
885 44. Kovács M, Halfmann A, Fedtke I, Heintz M, Peschel A, Vollmer W, et al. A
886 functional dlt operon, encoding proteins required for incorporation of d-alanine

887 in teichoic acids in gram-positive bacteria, confers resistance to cationic
888 antimicrobial peptides in Streptococcus pneumoniae. Journal of bacteriology.
889 2006;188:5797-805. doi: 10.1128/JB.00336-06. PubMed PMID: 16885447.
890 45. Deininger S, Figueroa-Perez I, Sigel S, Stadelmaier A, Schmidt RR, Hartung
891 T, et al. Use of synthetic derivatives to determine the minimal active structure of
892 cytokine-inducing lipoteichoic acid. Clinical and vaccine immunology : CVI.
893 2007;14:1629-33. doi: 10.1128/CVI.00007-07. PubMed PMID: 17928431.
894 46. Habets MGJL, Rozen DE, Brockhurst Ma. Variation in Streptococcus
895 pneumoniae susceptibility to human antimicrobial peptides may mediate
896 intraspecific competition. Proceedings of the Royal Society B: Biological Sciences.
897 2012;279:3803-11. doi: 10.1098/rspb.2012.1118. PubMed PMID: 22764166.
898 47. Wörmann ME, Horien CL, Bennett JS, Jolley Ka, Maiden MCJ, Tang CM, et
899 al. Sequence, distribution and chromosomal context of class I and class II pilin
900 genes of Neisseria meningitidis identified in whole genome sequences. BMC
901 genomics. 2014;15:253. doi: 10.1186/1471-2164-15-253. PubMed PMID:
902 24690385.
903 48. Russell JE, Jolley Ka, Feavers IM, Maiden MCJ, Suker J. PorA Variable
904 Regions of Neisseria meningitidis. Emerging Infectious Diseases. 2004;10:674-8.
905 doi: 10.3201/eid1004.030247. PubMed PMID: 15200858.
906 49. Johansson J, Mandin P, Renzoni A, Chiaruttini C, Springer M, Cossart P. An
907 RNA thermosensor controls expression of virulence genes in Listeria
908 monocytogenes. Cell. 2002;110(5):551-61. PubMed PMID: 12230973.
909 50. Gripenland J, Netterling S, Loh E, Tiensuu T, Toledo-Arana A, Johansson J.
910 RNAs: regulators of bacterial virulence. Nat Rev Microbiol. 2010;8(12):857-66.
911 doi: 10.1038/nrmicro2457. PubMed PMID: 21079634.

912 51. Sreevatsan S, Pan X, Zhang Y, Deretic V, Musser JM. Analysis of the oxyR-
913 ahpC region in isoniazid-resistant and -susceptible Mycobacterium tuberculosis
914 complex organisms recovered from diseased humans and animals in diverse
915 localities. Antimicrob Agents Chemother. 1997;41(3):600-6. PubMed PMID:
916 9056000; PubMed Central PMCID: PMCPMC163758.
917 52. Magnusson M, Tobes R, Sancho J, Pareja E. Cutting edge: natural DNA
918 repetitive extragenic sequences from gram-negative pathogens strongly
919 stimulate TLR9. J Immunol. 2007;179(1):31-5. PubMed PMID: 17579017.
920 53. Hung MC, Christodoulides M. The biology of Neisseria adhesins. Biology
921 (Basel). 2013;2(3):1054-109. doi: 10.3390/biology2031054. PubMed PMID:
922 24833056; PubMed Central PMCID: PMCPMC3960869.
923 54. Griffiths NJ, Hill DJ, Borodina E, Sessions RB, Devos NI, Feron CM, et al.
924 Meningococcal surface fibril (Msf) binds to activated vitronectin and inhibits the
925 terminal complement pathway to increase serum resistance. Mol Microbiol.
926 2011;82(5):1129-49. doi: 10.1111/j.1365-2958.2011.07876.x. PubMed PMID:
927 22050461.
928 55. Delany I, Grifantini R, Bartolini E, Rappuoli R, Scarlato V. Effect of
929 Neisseria meningitidis fur mutations on global control of gene transcription. J
930 Bacteriol. 2006;188(7):2483-92. doi: 10.1128/JB.188.7.2483-2492.2006.
931 PubMed PMID: 16547035; PubMed Central PMCID: PMCPMC1428404.
932 56. Bucci C, Lavitola A, Salvatore P, Del Giudice L, Massardo DR, Bruni CB, et
933 al. Hypermutation in pathogenic bacteria: frequent phase variation in
934 meningococci is a phenotypic trait of a specialized mutator biotype. Mol Cell.
935 1999;3(4):435-45. PubMed PMID: 10230396.

936 57. Moxon ER, Rainey PB, Nowak MA, Lenski RE. Adaptive evolution of highly
937 mutable loci in pathogenic bacteria. Curr Biol. 1994;4(1):24-33. PubMed PMID:
938 7922307.
939 58. Snyder LA, Butcher SA, Saunders NJ. Comparative whole-genome analyses
940 reveal over 100 putative phase-variable genes in the pathogenic Neisseria spp.
941 Microbiology. 2001;147(Pt 8):2321-32. doi: 10.1099/00221287-147-8-2321.
942 PubMed PMID: 11496009.
943 59. Van Der Ende AD, Hopman CTP, Dankert J. Multiple mechanisms of phase
944 variation of PorA in Neisseria meningitidis. Infection and Immunity.
945 2000;68:6685-90. doi: 10.1128/IAI.68.12.6685-6690.2000. PubMed PMID:
946 11083782.
947 60. Sarkari J, Pandit N, Moxon ER, Achtman M. Variable expression of the Opc
948 outer membrane protein in Neisseria meningitidis is caused by size variation of a
949 promoter containing poly-cytidine. Mol Microbiol. 1994;13(2):207-17. PubMed
950 PMID: 7984102.
951 61. Croucher NJ, Coupland PG, Stevenson AE, Callendrello A, Bentley SD,
952 Hanage WP. Diversification of bacterial genome content through distinct
953 mechanisms over different timescales. Nature Communications. 2014;5:5471.
954 doi: 10.1038/ncomms6471.
955 62. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, et al.
956 Multilocus sequence typing: a portable approach to the identification of clones
957 within populations of pathogenic microorganisms. Proceedings of the National
958 Academy of Sciences of the United States of America. 1998;95:3140-5. doi:
959 10.1073/pnas.95.6.3140. PubMed PMID: 9501229.

960 63. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al.
961 Rapid phylogenetic analysis of large samples of recombinant bacterial whole
962 genome sequences using Gubbins. Nucleic Acids Research. 2015;43:e15-e. doi:
963 10.1093/nar/gku1196.
964 64. Snyder LA, Saunders NJ, Shafer WM. A putatively phase variable gene
965 (dca) required for natural competence in Neisseria gonorrhoeae but not
966 Neisseria meningitidis is located within the division cell wall (dcw) gene cluster.
967 J Bacteriol. 2001;183(4):1233-41. doi: 10.1128/JB.183.4.1233-1241.2001.
968 PubMed PMID: 11157935; PubMed Central PMCID: PMCPMC94996.
969 65. Snyder LA, Shafer WM, Saunders NJ. Divergence and transcriptional
970 analysis of the division cell wall (dcw) gene cluster in Neisseria spp. Mol
971 Microbiol. 2003;47(2):431-42. PubMed PMID: 12519193.
972 66. Takahashi H, Hirose K, Watanabe H. Necessity of meningococcal gamma-
973 glutamyl aminopeptidase for Neisseria meningitidis growth in rat cerebrospinal
974 fluid (CSF) and CSF-like medium. J Bacteriol. 2004;186(1):244-7. PubMed PMID:
975 14679245; PubMed Central PMCID: PMCPMC303462.
976 67. Veyrier FJ, Boneca IG, Cellier MF, Taha MK. A novel metal transporter
977 mediating manganese export (MntX) regulates the Mn to Fe intracellular ratio
978 and Neisseria meningitidis virulence. PLoS Pathog. 2011;7(9):e1002261. doi:
979 10.1371/journal.ppat.1002261. PubMed PMID: 21980287; PubMed Central
980 PMCID: PMCPMC3182930.
981 68. Jorth P, Staudinger Benjamin J, Wu X, Hisert KB, Hayden H, Garudathri J, et
982 al. Regional Isolation Drives Bacterial Diversification within Cystic Fibrosis
983 Lungs. Cell Host & Microbe. 2015;18:307-19. doi: 10.1016/j.chom.2015.07.006.

984 69. Das S, Lindemann C, Young BC, Muller J, Osterreich B, Ternette N, et al.
985 Natural mutations in a Staphylococcus aureus virulence regulator attenuate
986 cytotoxicity but permit bacteremia and abscess formation. Proc Natl Acad Sci U S
987 A. 2016;113(22):E3101-10. doi: 10.1073/pnas.1520255113. PubMed PMID:
988 27185949; PubMed Central PMCID: PMCPMC4896717.
989 70. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. Within-host
990 evolution of bacterial pathogens. Nature Reviews Microbiology. 2016. doi:
991 10.1038/nrmicro.2015.13.
992 71. Uricaru R, Rizk G, Lacroix V, Quillery E, Plantard O, Chikhi R, et al.
993 Reference-free detection of isolated SNPs. Nucleic Acids Research. 2014;33:1-11.
994 doi: 10.1093/nar/gku1187.
995 72. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. De novo assembly and
996 genotyping of variants using colored de Bruijn graphs. Nature genetics.
997 2012;44:226-32. doi: 10.1038/ng.1028. PubMed PMID: 22231483.
998 73. Bankevich A, Nurk S, Antipov D, Gurevich Aa, Dvorkin M, Kulikov AS, et al.
999 SPAdes: a new genome assembly algorithm and its applications to single-cell
1000 sequencing. Journal of computational biology : a journal of computational
1001 molecular cell biology. 2012;19:455-77. doi: 10.1089/cmb.2012.0021. PubMed
1002 PMID: 22506599.
1003 74. Page AJ, De Silva N, Hunt M, Quail MA, Parkhill J, Harris SR, et al. Robust
1004 high throughput prokaryote de novo assembly and improvement pipeline for
1005 Illumina data. bioRxiv. 2016. doi: 10.1101/052688.
1006 75. Zaharia M, Bolosky WJ, Curtis K, Fox A, Patterson DA, Shenker S, et al.
1007 Faster and More Accurate Sequence Alignment with SNAP. CoRR.
1008 2011;abs/1111.5:1-10.

1009 76. Li H. A statistical framework for SNP calling, mutation discovery,
1010 association mapping and population genetical parameter estimation from
1011 sequencing data. Bioinformatics. 2011;27:2987-93. doi:
1012 10.1093/bioinformatics/btr509.
1013 77. Kelley DR, Schatz MC, Salzberg SL. Quake: quality-aware detection and
1014 correction of sequencing errors. Genome biology. 2010;11:R116. doi:
1015 10.1186/gb-2010-11-11-r116. PubMed PMID: 21114842.
1016 78. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using
1017 compressed data structures. Genome research. 2012;22:549-56. doi:
1018 10.1101/gr.126953.111. PubMed PMID: 22156294.
1019 79. Hoskins J, Alborn WE, Arnold J, Blaszczak LC, Burgett S, DeHoff BS, et al.
1020 Genome of the bacterium Streptococcus pneumoniae strain R6. Journal of
1021 bacteriology. 2001;183:5709-17. doi: 10.1128/JB.183.19.5709. PubMed PMID:
1022 11544234.
1023 80. Denapaite D, Brückner R, Nuhn M, Reichmann P, Henrich B, Maurer P, et
1024 al. The genome of streptococcus mitis B6 - What is a commensal? PloS one.
1025 2010;5. doi: 10.1371/journal.pone.0009426. PubMed PMID: 20195536.
1026 81. Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, Haussler D. Cactus:
1027 Algorithms for genome multiple sequence alignment. Genome research.
1028 2011;21:1512-28. doi: 10.1101/gr.123356.111. PubMed PMID: 21665927.
1029 82. Hu X, Yuan J, Shi Y, Lu J, Liu B, Li Z, et al. pIRS: Profile-based Illumina pair-
1030 end reads simulator. Bioinformatics. 2012;28:1533-5. doi:
1031 10.1093/bioinformatics/bts187.
1032 83. Tettelin H, Nelson KE, Paulsen IT, Eisen Ja, Read TD, Peterson S, et al.
1033 Complete genome sequence of a virulent isolate of Streptococcus pneumoniae.

1034 Science (New York, NY). 2001;293:498-506. doi: 10.1126/science.1061217.
1035 PubMed PMID: 11463916.
1036 84. Klambauer G, Schwarzbauer K, Mayr A, Clevert DA, Mitterecker A,
1037 Bodenhofer U, et al. cn.MOPS: mixture of Poissons for discovering copy number
1038 variations in next-generation sequencing data with a low false discovery rate.
1039 Nucleic Acids Res. 2012;40(9):e69. doi: 10.1093/nar/gks003. PubMed PMID:
1040 22302147; PubMed Central PMCID: PMCPMC3351174.
1041 85. Croucher NJ, Walker D, Romero P, Lennard N, Paterson GK, Bason NC, et
1042 al. Role of conjugative elements in the evolution of the multidrug-resistant
1043 pandemic clone Streptococcus pneumoniaeSpain23F ST81. Journal of
1044 Bacteriology. 2009;191:1480-9. doi: 10.1128/JB.01343-08. PubMed PMID:
1045 19114491.
1046 86. Tettelin H, Saunders NJ, Heidelberg J, Jeffries aC, Nelson KE, Eisen Ja, et al.
1047 Complete genome sequence of Neisseria meningitidis serogroup B strain MC58.
1048 Science. 2000;287:1809-15. doi: 8338 [pii]. PubMed PMID: 10710307.
1049 87. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving
1050 the consequences of genomic variants with the Ensembl API and SNP Effect
1051 Predictor. Bioinformatics. 2010;26:2069-70. doi:
1052 10.1093/bioinformatics/btq330. PubMed PMID: 20562413.
1053 88. Kent WJ. BLAT—The BLAST-Like Alignment Tool. Genome Research.
1054 2002;12:656-64. doi: 10.1101/gr.229202.
1055 89. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures
1056 of model complexity and fit. Journal of the Royal Statistical Society: Series B
1057 (Statistical Methodology). 2002;64:583-639. doi: 10.1111/1467-9868.00353.
1058