The comparative population genetics of *Neisseria meningitidis* and *Neisseria gonorrhoeae*

Lucile Vigue [1] , **Adam Eyre-Walker** [Corresp. 2]

1 Ecole Polytechnique, Paris, France

2 School of Life Sciences, University of Sussex, Brighton, United Kingdom

Corresponding Author: Adam Eyre-Walker
Email address: a.c.eyre-walker@sussex.ac.uk

Neisseria meningitidis (Nm) *and N. gonorrhoeae* (Ng) are closely related pathogenic bacteria. To compare their population genetics we compiled a dataset of 1145 genes found across 20 Nm and 15 Ng genomes. We find that Nm is seven-times more diverse than Ng in their combined core genome. Both species have acquired the majority of their diversity by recombination with divergent strains, however we find that Nm has acquired more of its diversity by recombination than Ng. We find that linkage disequilibrium declines rapidly across the genomes of both species. Several observations suggest that Nm has a higher effective population size than Ng; it is more diverse, the ratio of non-synonymous to synonymous polymorphism is lower, and linkage disequilibrium declines more rapidly to a lower asymptote in Nm. The two species share a modest amount of variation, half of which seems to have been acquired by lateral gene transfer and half from their common ancestor. We investigate whether diversity varies across the genome of each species and find that it does. Much of this variation is due to different levels of lateral gene transfer. However, we also find some evidence that the effective population size varies across the genome. We test for adaptive evolution in the core genome using a McDonald-Kreitman test and by considering the diversity around non-synonymous sites that are fixed for different alleles in the two species. We find some evidence for adaptive evolution using both approaches.

1 **The comparative population genetics of *Neisseria meningitidis* and *Neisseria gonorrhoeae***

3 Lucile Vigue[1]

4 Adam Eyre-Walker[2]

6 1. Ecole Polytechnique

7 Paris

8 France

10 2. School of Life Sciences

11 University of Sussex

12 Brighton

13 United Kingdom

15 Correspondence : a.c.eyre-walker@sussex.ac.uk

17 **Abstract**

18 *Neisseria meningitidis* (Nm) *and N. gonorrhoeae* (Ng) are closely related pathogenic bacteria. To

19 compare their population genetics we compiled a dataset of 1145 genes found across 20 Nm

20 and 15 Ng genomes. We find that Nm is seven-times more diverse than Ng in their combined

21 core genome. Both species have acquired the majority of their diversity by recombination with

22 divergent strains, however we find that Nm has acquired more of its diversity by recombination

23 than Ng. We find that linkage disequilibrium declines rapidly across the genomes of both

24 species. Several observations suggest that Nm has a higher effective population size than Ng; it

25 is more diverse, the ratio of non-synonymous to synonymous polymorphism is lower, and

26 linkage disequilibrium declines more rapidly to a lower asymptote in Nm. The two species share

27 a modest amount of variation, half of which seems to have been acquired by lateral gene

28 transfer and half from their common ancestor. We investigate whether diversity varies across

29 the genome of each species and find that it does. Much of this variation is due to different

30 levels of lateral gene transfer. However, we also find some evidence that the effective

31 population size varies across the genome. We test for adaptive evolution in the core genome

32 using a McDonald-Kreitman test and by considering the diversity around non-synonymous sites

33 that are fixed for different alleles in the two species. We find some evidence for adaptive

34 evolution using both approaches.

37 **Introduction**

38 The two closely related bacteria *Neisseria meningitidis* (Nm) and *Neisseria gonorhoeae* (Ng) are

39 major human pathogens. Ng is the causative agent of the sexually transmitted disease

40 gonorrhoeae which currently infects 106 million people each year worldwide (WHO 2012).

41 When untreated, gonoccocal infections can result in long-term problems such as persistent

42 urethritis, cervicitis, proctitis, pelvic inflammatory disease, infertility, first-trimester abortion,

43 ectopic pregnancy and maternal death (WHO 2012). They also increase the risk of acquiring and

44 transmitting HIV. In cases of pregnancy, Ng infections can cause severe damages to neonatal

45 health (WHO 2012). In contrast, Nm is a human commensal infecting approximately 10% of the

46 healthy human population (Claus et al. 2005; Yazdankhah et al. 2004), which only occasionally

47 causes disease. However, it can cause meningococcal meningitidis and septicaemia with

48 mortality rates that can reach 50% when untreated, and the global disease burden is estimated

49 to be ~500,000 cases a year (Roberts 2008). Among the different micro-organisms that can

50 cause meningitidis, it is regarded as one of the most important because of its ability to cause

51 large epidemics.

52

53 Here we consider several aspects of the population genetics of these bacterial species. The two

54 species are sister taxa (Bennett et al. 2012), and Nm is known to be considerably more diverse

55 than Ng within the genes that they share in common (Bennett et al. 2012; Bennett et al. 2007).

56 The first problem we address is why the two taxa differ in their diversities. There are several

57 potential explanations. First, Ng might have a lower effective population size, either because it

58 evolved from Nm and went through a bottleneck when the species was formed (Vazquez et al.

59   1993), or because it generally has a lower effective population size, possibly because it has a

60   lower census population size. Second, Ng might have a lower mutation rate than Nm. Third, Ng

61   might acquire less diversity through recombination than Nm. Both Ng and Nm are known to be

62   naturally transformable, and it has been known for many years that both species acquire

63   diversity, within their core genome, by homologous recombination with genetically divergent

64   strains (Spratt 1988; Spratt et al. 1989). We refer to this process as homologous lateral gene

65   transfer (hLGT), to differentiate it from the acquisition of accessory genes by non-homologous

66   lateral gene transfer (nhLGT) (however, note that the acquisition of new genes generally

67   involves homologous recombination with flanking genes, so nhLGT will typically involve some

68   hLGT (Kong et al. 2013)). hLGT leads to mosaic genes, in which parts of the gene have been

69   acquired from a highly divergent strain or a different bacterial species. In fact, Nm and Ng were

70   some of the first bacteria in which this form of recombination was demonstrated (Spratt 1988;

71   Spratt et al. 1989). It has been estimated that Nm acquires single nucleotide polymorphisms

72   (SNPs) through hLGT at a rate between 4 and 100x higher than via mutation (Feil et al. 2001;

73   Hao et al. 2011; Kong et al. 2013; Vos & Didelot 2009). In contrast this ratio has recently been

74   estimated to be only about two-fold in Ng (Ezewudo et al. 2015). It is unclear whether these

75   ratios are significantly different. We investigate this here.

76

77   The second question, we address is whether diversity varies across the core genome of the two

78   species. Genetic diversity is known to vary across the genome of many species. This was

79   originally demonstrated in *Drosophila melanogaster* by Begun and Aquadro (Begun & Aquadro

80   1992) who showed that diversity was positively correlated to the rate of recombination. This

81   was thought to be due to the effects of linked selection, in the form of genetic hitch-hiking

82   (Maynard Smith & Haigh 1974) and background selection (Charlesworth et al. 1993), depressing

83   diversity in regions of the genome with low rates of recombination. Variation in diversity across

84   the genome has been demonstrated in many other species including the bacterium *Escherichia*

85   *coli* (Maddamsetti et al. 2015; Martincorena et al. 2012). The reasons for this variation remain

86   unclear (Chen & Zhang 2013; Maddamsetti et al. 2015; Martincorena & Luscombe 2013).

87

88   The final question we address is whether Nm and Ng have undergone adaptive evolution. Nm

89   and Ng inhabit different niches and one presumes they have undergone adaptive evolution to

90   allow them to do this. Some of this adaptation may have been through the acquisition of new

91   genes via nhLGT, but there might also be adaptation in the core genome. Two recent analyses

92   using the d_N/d_S test on the core genome have found limited evidence for adaptive evolution in

93   Nm (Yu et al. 2014) and Ng (Ezewudo et al. 2015), but this test is known to be very

94   conservative. Here we apply two additional tests.

95 **Materials and methods**

96 *Dataset*

97 All 15 genomes of *Neisseria gonorrhoeae* that were present in Genbank in April 2018

98 (NCCP11945 (Chung et al. 2008), FA19 (Abrams et al. 2015), FA6140 (Abrams et al. 2015), 35/02

99 (Abrams et al. 2015), FA 1090, MS11, FA19, FA6140, 35/02, 32867, 34530, 34769, FDAARGOS

100 204, FDAARGOS 205, FDAARGOS 207, NCTC13799, NCTC13798, NCTC13800) and 20 randomly

101 selected genomes of *Neisseria meningitidis* (MC58 (Tettelin et al. 2000), Z2491 (Parkhill et al.

102 2000), FAM18 (Bentley et al. 2007), 053442 (Peng et al. 2008), alpha14 (Schoen et al. 2008),

103 8013 (Rusniok et al. 2009), alpha710 (Joseph et al. 2010), WUE 2594 (Schoen et al. 2011),

104 G2136 (Budroni et al. 2011), M01-240149 (Budroni et al. 2011), M04-240196 (Budroni et al.

105 2011), H44/76 (Budroni et al. 2011), M01-240355 (Budroni et al. 2011), NZ-05/33 (Budroni et al.

106 2011), 510612 (Zhang et al. 2014), NM3686, M7124, NM3682, NM3683, L91543) were

107 downloaded from Genbank. From these all protein coding sequences were extracted. We

108 retained those coding sequences that started NTG, terminated with TAA, TAG or TGA and had a

109 length that is a multiple of three. We identified orthologs using reciprocal BLAST, with an e-

110 value threshold of 0.00001; i.e. each protein coding gene in each genome was BLASTed against

111 the genes of FA1090, and then the best hit was BLASTed back onto the original genome,

112 retaining only those hits in which the original query sequence was the best hit. Similar

113 selections of genes were obtained using alternative starting genomes. The protein sequences of

114 the orthologs were aligned using MUSCLE (Edgar 2004). We selected genes where the

115 alignments meet these criteria: the number of gaps is lower than 1% of the length of the

116 sequence and the total number of nucleotides in gaps is lower than 10% of the total number of

117 nucleotides in the sequence. Sequences with internal stop codons were removed. This resulted

118 in a dataset of 1145 genes belonging to the core genome of both *Neisseria gonorrhoeae* and

119 *Neisseria meningitidis*. We used the BioPython Phylo library (Cock et al. 2009) to estimate a

120 phylogeny of the strains based on the core genome alignment.

121

122 *Analyses*

123   In most analyses we treated genes independently. However, to detect hLGT we ran

124   ClonalFrameML (Didelot & Wilson 2015) on a concatenation of the protein coding sequences

125   from the core genome of both species. Genes were concatenated randomly without respect for

126   synteny. For some analyses we masked those regions inferred to be due to hLGT in the strains

127   affected.

128

129   We investigated whether linkage disequilibrium (LD) declines with the distance between sites

130   by measuring the LD between all pairs of polymorphisms within each gene; we did not

131   concatenate the genes or align whole genomes, because with the gain and loss of genes the

132   distance between sites differs depending on the strains being analysed. We measured LD using

133   the r^2 statistic (Hill & Robertson 1968). LD values were then assigned to bins based on the

134   distance between the two sites – 10bp bins between 1-100bp, a bin from 101-200bp and then

135   200bp bins between 201-800bp. We took the average LD and distance between sites for each

136   bin in a manner which weighted each gene equally – we estimated the average LD and distance

137   for pairs of sites in each bin for each gene and then averaged those values across genes. To

138   estimate the approximate half-life of LD, we found the distance between sites that gave

139   approximately half the LD between the LD for the 1-10bp bin and the asymptotic value of the

140   LD.

141

142   Because r^2 is constrained to be positive, the expected value of r^2 is greater than zero even when

143   there is no LD. To calculate the expected value of r^2 when there is no LD, we considered two bi-

144   allelic loci with alleles at frequencies p_1 and p_2. The expected frequencies of the four haplotypes

145   are p_1p_2, $p_1(1-p_2)$…etc. from which we generated four random variates from a multinomial

146   distribution for a sample size of N chromosomes using Mathematica version 11; for each

147   sample of haplotypes we calculated r^2. We repeated this procedure 10,000 times and calculated

148   the mean to estimate the expected value of r^2. We found that the expected value of r^2 is

149   independent of the allele frequencies.

150

151 To investigate the relationship between the non-synonymous, π_N, and synonymous, π_S,

152 nucleotide diversity we used a variation of the method of James et al. (James et al. 2017) to

153 combine data from different genes. If the distribution of fitness effects of new mutations is a

154 gamma distribution (assuming most mutations are deleterious) then $\log(\pi_N)$ is expected to be

155 linearly correlated to $\log(\pi_S)$ if there is variation in N_e (Welch et al. 2008). However, for many

156 genes either π_N or π_S is zero, hence we need to combine genes together. We can do this by

157 splitting the synonymous polymorphisms into two groups according to whether they were in an

158 odd or even numbered codon and then using the two groups to estimate two synonymous

159 nucleotide diversities that have independent sampling errors, π_{S1} and π_{S2}. One of these, π_{S1},

160 was used to rank and group genes, and the other was averaged across genes in the group to

161 give an unbiased estimate of π_S for the group. π_N was also averaged across the genes in the

162 group.

163

164 To investigate the diversity around sites that are fixed between Nm and Ng for different alleles

165 we focused on genes that had at least one synonymous polymorphism and one fixed difference

166 between the two species. For each fixed difference, we identified all the synonymous

167 polymorphisms that were within 1 kb and we grouped them by windows of 100 bp. Since,

168 background selection can potentially lead to a lower dip in diversity around fixed non-

169 synonymous mutations we normalised the diversity around fixed synonymous and non-

170 synonymous substitutions by dividing the number of synonymous polymorphisms in a particular

171 window by the total number of synonymous polymorphisms in the gene, multiplied by the

172 window size over the gene length.

173

174

175 **Results**

176 *Recombination and mutation*

177 We are interested in how genetic variation is generated and distributed in the two *Neisseria*

178 species *N. meningitidis* (Nm) and *N. gonorrhoeae* (Ng). Although, the presence and absence of

179 genes in the strains of the two species is an important aspect of this problem, here we focus on

180   the genetic variation that is present in the core genome that is common to both species. Using

181   reciprocal BLAST, we identified 1145 genes present across the 15 genomes of Ng and 20

182   genomes of Nm that we analysed. The total length of this core genome is 1.1MB long. Defining

183   a polymorphism as a site that contains two or more alleles within either of the two species, we

184   find that Nm is ~7.6 fold more diverse than Ng consistent with previous qualitative reports

185   (Bennett et al. 2012; Bennett et al. 2007). The difference in diversity is more apparent at

186   synonymous (~8.9 fold) than non-synonymous (~5.5 fold) sites (Table 1), a pattern we return to

187   later. The two species share a modest amount of diversity; 35% of all polymorphisms in Ng are

188   shared with Nm, and 4.5% of those in Nm are shared with Ng.

189

190   It is well known that Nm and Ng undergo substantial levels of homologous recombination with

191   divergent strains, possibly from other species of bacteria. This leads both to the acquisition of

192   new genes, but also to the acquisition of parts of genes that are already present in the genome;

193   we refer to these processes as non-homologous (nhLGT) and homologous lateral gene transfer

194   (hLGT) respectively. To quantify the role that hLGT plays in the acquisition of diversity in the

195   core genome we ran ClonalFrameML (Didelot & Wilson 2015) (Didelot & Wilson 2015). The

196   method estimates the ratio of the rate at which recombination tracts initiate (R) and the rate of

197   mutation (θ), both multiplied by twice the effective population size, N_e, along with the average

198   recombination tract length, δ, and the proportion of sites that differ between the imported and

199   resident sequences, ν. Estimates of these parameters are given in Table 2. The overall effect of

200   recombination relative to mutation can be estimated as $R\delta\nu/\theta = r/m$, where r and m are the

201   rates at which variants are introduced into a genome by recombination and mutation

202   respectively.

203

204   In Nm we find that recombination introduces 6.43 (95% CI = 6.16 to 6.71) times more variation

205   than mutation, whereas in Ng it introduces 1.97 (1.76, 2.19) times as much. In Nm the r/m ratio

206   has previously been estimated to be 5.37(Hao et al. 2011), 6.71 (Vos & Didelot 2009), 16.4

207   (Kong et al. 2013) and 100 (Feil et al. 2001). Our estimate is similar to the first two estimates,

208   but substantially lower than the last two estimates. Both of these latter estimates were

209   obtained from very closely related strains and hence may reflect the value of *r/m* before

210   natural selection has had an opportunity to operate. In Ng it has been estimated that 2.2x as

211   much variation is introduced by recombination (Ezewudo et al. 2015), which is very similar to

212   our estimate. The estimates of *r/m* mean that ~87% of all polymorphisms in Nm are a

213   consequence of recombination, whereas in Ng it is 66%. The difference between the two

214   species in the influence of recombination is largely driven by a difference in the ratio of the rate

215   at which recombination is initiated versus the mutation rate (R/θ), since although the tract

216   lengths are estimated to be on average slightly longer in Nm, they introduce slightly less

217   variation than Ng (Table 2).

218

219   ClonalframeML estimates the ratio of R and θ but not their absolute values. However, we can

220   estimate the absolute value as follows. We note that the nucleotide diversity is due to the input

221   of mutation and the input of recombination: i.e. $\pi = \theta + R\,\delta\,\upsilon$. If we note that ClonalframeML

222   gives us an estimate of R/θ we can rewrite this equation as $\pi = \theta + \theta\,\delta\,\upsilon\,R/\theta$, from which we

223   can estimate $\theta = \pi/(1 + \delta\,\upsilon\,R/\theta)$. Estimates of R and θ are given in Table 2. From this it is evident

224   that the nucleotide diversity is higher in Nm both because of a 3-fold greater mutational input

225   and a nine-fold greater rate at which recombination tracts initiate in Nm, at the population

226   level (i.e. when the tract length initiation rate and mutation rate are multiplied by N_e).

227

228   The parameters R and θ are the rates of recombination initiation and mutation, multiplied by

229   the effective population size. Hence a simple reason why both parameters are higher in Nm

230   might simply be that Nm has a higher N_e than Ng. To test this idea, we masked all sequences

231   that were identified as due to hLGT by ClonalframeML and estimated the levels of non-

232   synonymous and synonymous diversity. Under a model in which synonymous mutations are

233   neutral and non-synonymous mutations are deleterious, but drawn from some distribution, we

234   expect π_N/π_S to be lower in species with high N_e; this is because selection is more effective in

235   species with higher N_e and hence the proportion of mutations that are effectively neutral is

236   lower (Ohta 1972; Ohta 1977; Ohta 1992). This is what we find - $\pi_N/\pi_S = 0.095$ (SE = 0.0023) in

237 Nm versus 0.23 (0.014) in Ng. These are significantly different to each other (normal test z = 9.5,

238 p<0.001).

239

240 As we described above, Nm and Ng share a modest amount of genetic variation. It is of some

241 interest whether this is a consequence of hLGT or the inheritance of genetic variation from

242 their common ancestor. If we exclude those sequences inferred to be due to hLGT we find that

243 the two species still share a modest amount of genetic variation −15.5% of all Ng

244 polymorphisms are shared with Nm and 2.4% of Nm polymorphisms are shared with Ng,

245 approximately half of all shared polymorphisms in each case, suggesting that some proportion

246 of the shared variation originated from their common ancestor.

247

248 *Linkage disequilibrium*

249 Homologous recombination can both increase and decrease linkage disequilibrium (LD);

250 homologous recombination with divergent strains, of the sort detected by ClonalFrameML,

251 generates LD because it simultaneously introduces many polymorphisms that are initially linked

252 to each other. However, homologous recombination amongst a set of closely related strains

253 breaks-up LD. To investigate how these two forces play out, we calculated the LD between all

254 pairs of sites within each gene and plotted these as a function of the distance between sites. As

255 expected we observe a decline in LD with distance (Figure 1A). Both species show similar

256 patterns with LD declining rapidly; in Nm the approximate half-life is 30bp and in Ng it is 100bp.

257 The decline could be due to two processes. If most hLGT fragments tend to be short, with

258 decreasing numbers of long fragments, then LD will be greater between closely linked sites.

259 However, we also expect a decline due to recombination between closely related strains, and in

260 fact we observe a decline even when we focus on those parts of the genome which do not

261 appear to have undergone hLGT (Figure 1B).

262

263 In both species LD asymptotes above zero. The non-zero asymptote could be due to one of

264 three reasons – statistical bias, population substructure and a balance between genetic drift

265 and recombination. The statistical bias arises because our measure of LD, r^2 (Hill & Robertson

266   1968), cannot be negative, so positive values of r^2 are expected even if there is no LD if sample

267   sizes are small; for sample sizes of 15 and 20 strains, the expected value of r^2 is 0.079 and 0.050

268   respectively (see materials and methods), so the asymptote is clearly above this level. Both, Nm

269   and Ng have been shown to have some level of population structure so this is the likely to be

270   part of the explanation (Budroni et al. 2011; Joseph et al. 2011). However, the slower decay in

271   LD, and higher asymptote in Ng, is consistent with Ng having a smaller N_e than Nm – i.e. the

272   non-zero asymptote might in part be caused by a balance between genetic drift creating LD,

273   and recombination breaking it down.

274

275   *Diversity across the genome*

276   Nucleotide diversity is known to vary across the genomes of many organisms. This is largely

277   thought to be driven by variation in the mutation rate or variation in the effects of linked

278   selection. However, in bacteria, and particularly Nm and Ng, it could also be due to variation in

279   the frequency of hLGT. All of these processes are expected to affect synonymous and non-

280   synonymous diversity to greater or lesser extents, and indeed we observe a positive correlation

281   between non-synonymous and synonymous diversity, demonstrating that both vary across the

282   genome in concert. At least part of this pattern is driven by hLGT because genes with hLGT

283   show higher π_N and π_S values than genes without any evidence of hLGT (Figure 2).

284

285   However, to investigate whether there is also variation in the effective population size across

286   the genome we removed sequences inferred to be due to hLGT by ClonalFrameML from our

287   data. This reduces our data substantially and so to reduce statistical sampling issues we used

288   the method of James et al. (2016) to combine data from different genes. We find that π_N and π_S

289   are still significantly correlated suggesting the correlation between them is not just driven by

290   hLGT (Ng slope = 0.23, p<0.001; Nm slope = 0.53, p<0.001)(Figure 3). The remaining correlation

291   could be due to variation in the mutation rate or variation in the effects of linked selection. We

292   can test whether there is variation in the effects of linked selection by considering the slope

293   between $\log(\pi_N)$ and $\log(\pi_S)$. Under a model in which there is no variation in linked selection

294   then the slope of this relationship is expected to be one, and if there is variation in linked

295   selection the slope if expected to be less than one (Galtier 2016; Welch et al. 2008). Linked

296   selection has two consequences. First, it increases the stochasticity in allele frequencies. For

297   example, the spread of an advantageous mutation or the elimination of deleterious genetic

298   variation, removes linked genetic diversity; whether a linked mutation survives either process is

299   a random process depending on whether the advantageous or deleterious mutation occurs in

300   linkage with the target mutation. This can be thought of as reduction in the effective population

301   size. Second, genetic hitch-hiking leads to non-equilibrium dynamics. After a selective sweep,

302   genetic diversity will recover, but this happens faster for deleterious than neutral mutations

303   (Brandvain & Wright 2016; Do et al. 2015; Gordo & Dionisio 2005). In both cases we expect a

304   negative correlation between π_N/π_S and π_S, which manifests itself in a positive correlation

305   between $\log(\pi_N)$ and $\log(\pi_S)$ but with a slope of less than one (James et al. 2017). We find that

306   the slope of the relationship between $\log(\pi_N)$ and $\log(\pi_S)$ is 0.23 (SE = 0.052) and 0.59 (0.070)

307   for Ng and Nm respectively, in both cases significantly less than one (p<0.001); i.e. π_N increases

308   as π_S increases but not as fast. The slopes are significantly different to each other (t-test,

309   p<0.001).

310

311   *Adaptive evolution*

312   Nm and Ng are ecologically quite different and one presumes the two species have undergone

313   adaptation to live in their respective environments. Some of this adaptation will have come

314   about through the acquisition of whole genes through nhLGT. However, some of the adaptation

315   may have occurred within the core genome of the two species either by new mutations,

316   standing genetic variation, or hLGT. To investigate whether there has been adaptation in the

317   core genome we used two approaches. First, we used the McDonald-Kreitman (McDonald &

318   Kreitman 1991) approach to estimate the rate of adaptive evolution (Eyre-Walker 2006; Fay et

319   al. 2001). In this method the numbers of non-synonymous and synonymous substitutions (i.e.

320   differences between the two species, d_N and d_S respectively) are compared to the numbers of

321   non-synonymous and synonymous polymorphisms (p_N and p_S respectively). Under a neutral

322   model in which mutations are either neutral or strongly deleterious we expect $d_N/d_S = p_N/p_S$

323   (McDonald & Kreitman 1991). In contrast if there are slightly deleterious non-synonymous

324   mutations we expect $d_N/d_S < p_N/p_S$, and if there are some advantageous mutations we expect

325     $d_N/d_S > p_N/p_S$ (Eyre-Walker 2006; Fay et al. 2001). Summing d_N, d_S, p_N and p_S we calculate the

326   fixation index $FI = d_N p_S / d_S p_N$ (Gojobori et al. 2007); adaptive evolution is indicated if $FI > 1$.

327

328   We find that our estimate of FI differs if we use the polymorphism data of Nm or Ng; using the

329   SNP data of Nm we estimate that FI is significantly greater than one suggesting adaptive

330   evolution has occurred ($FI = 1.51$ with 95% Cis = 1.41 and 1.61), but if we use the SNP data of

331   Ng, our estimate is significantly less than one ($FI = 0.92$ (0.83, 0.99)). Estimates less than one

332   can occur if there are slightly deleterious mutations (SDMs) segregating, but even if we restrict

333   our analysis to common polymorphisms, which should remove many of the SDMs

334   (Charlesworth & Eyre-Walker 2008; Fay et al. 2001), we find that the $FI<1$ using the SNP data of

335   Ng (using SNPs with allele frequencies above 15%, $FI = 0.78$ (0.78, 0.88)). An explanation for

336   why FI differs between the two species is that either Nm has undergone population expansion,

337   or Ng has undergone contraction. If there are slightly deleterious mutations then population

338   size expansion leads to an overestimate of FI whereas contraction leads to an underestimate

339   (Eyre-Walker 2002; McDonald & Kreitman 1991). As we argue above, a simple explanation for

340   why Nm is more diverse than Ng is that Nm has a higher N_e. We find no evidence of expansion

341   or contraction amongst the current strains – Tajima's D (Tajima 1989), a measure of a skew in

342   the site frequency spectrum away from what we expect for neutral mutations in a stationary

343   population size is close to zero and not significantly different to zero in both species in the

344   regions of the genome that have no evidence of hLGT (Tajima's D = -0.073 and -0.093 in Nm and

345   Ng respectively), consistent with previous analyses in Nm (Joseph et al. 2011). However, the

346   expansion or contraction in either Nm or Ng could have occurred sometime in the past which

347   would not be visible to an analysis using Tajima's D, but which might still affect the FI.

348

349   A second approach to test for adaptive evolution, is to investigate whether there is a dip in

350   genetic diversity around putatively advantageous mutations (Sattath et al. 2011) – as

351   advantageous mutations spread through a population it reduces diversity in its proximity. We

352   find that synonymous diversity is lower close to sites that are fixed for different nucleotides in

353   the two species, and this dip is significantly greater for non-synonymous than synonymous fixed

354   differences when considering diversity in Nm (p<0.001 for distances 1-100bp, 101-200bp and

355   201-300bp), consistent with a proportion of non-synonymous mutations being fixed by positive

356   adaptive evolution; a similar pattern is not evident in Ng, possibly because it is less diverse.

357

358   There is however an alternative explanation for the greater dip around non-synonymous

359   substitutions; if the strength of background selection (Charlesworth et al. 1995) varies across

360   the genome, then regions with high levels of background selection will have low diversity but

361   will tend to also fix slightly deleterious non-synonymous mutations. To investigate whether

362   there is evidence of this, we considered whether the ratio of the non-synonymous to

363   synonymous substitution rates, $\log(d_N/d_S)$, was correlated to synonymous diversity, $\log(\pi_S)$. We

364   again use the method of James et al. (James et al. 2017) to combine data from different genes

365   and find that a strong negative correlation in Nm (slope = -0.10, p = 0.018) but not in Ng (slope

366   = -0.001, p =0.97). This suggests that background selection might be a factor in Nm. To take into

367   account the potential variation in background selection, we normalised the data from each

368   gene by dividing the number of synonymous SNPs in each window by the average diversity in

369   each gene. This will account for variation in background selection at a gene level, but not at a

370   sub-gene level. The normalised data show a greater dip in diversity for fixed non-synonymous

371   than synonymous substitutions in both Nm (combining t-test results from the three closest

372   points, p<0.001), and Ng (p = 0.0024) (figure 4) although the differences are not large.

373

374   **Discussion**

375   We have investigated several aspects of the comparative population genetics of the two

376   bacteria *Neisseria meningitidis* and *N. gonorrhoeae*. We find, as others have (Bennett et al.

377   2012; Bennett et al. 2007), that Nm is substantially more diverse than Ng, but that the two

378   species share a moderate amount of diversity in the genes that they have in common. This

379   shared diversity could have been a consequence of ancestral polymorphism that has been

380   inherited by both species, or due to hLGT transferring variation between the two. We find a

381   substantial fraction is indeed due to hLGT, since if we remove the fraction of the genome that

382   appears to have undergone hLGT, the fraction of shared polymorphism drops considerably.

383   However, there is some diversity that appears to have been inherited from the ancestor.

384

385   In both species we find that most of their genetic diversity has been acquired by recombination,

386   rather than by mutation. In Nm we estimate that the total input from hLGT is six-fold greater

387   than from mutation; this is in line with the estimates of Hao et al. (Hao et al. 2011) and Vos and

388   Didelot (Vos & Didelot 2009), but lower than two other estimates (Feil et al. 2001; Kong et al.

389   2013). Both of these high estimates were derived by considering very closely related strains. If

390   hLGT events are on average more deleterious than single nucleotide changes then we expect

391   r/m estimates to be greater for more closely related strains, because natural selection has had

392   more opportunity to remove the deleterious mutations in distantly related strains. This has the

393   implication that r/m may be far higher amongst newly arising mutations than often thought. In

394   Ng we find the input of hLGT is two-fold greater than mutation, consistent with the one

395   previous estimate performed on a similar selection of strains (Ezewudo et al. 2015).

396

397   Nm might be more diverse than Ng either Nm has a higher mutation rate, a greater rate of

398   hLGT or a higher effective population size. Several lines of evidence suggest that Nm has a

399   higher N_e. First, Nm has higher values of both R and θ, where R and θ are estimates of the rate

400   at which recombination initiates and the mutation rate, multiplied by N_e. Second, p_N/p_S is

401   lower in Nm in the fraction of the genome which does not seem to have undergone hLGT.

402   Third, LD declines faster in Nm and asymptotes at a lower level. However, this does not

403   preclude a role for either faster rates of mutation or recombination in the greater diversity in

404   Nm.

405

406   It is possible that the lower N_e in Ng is due to a bottleneck at the time when Ng was formed,

407   assuming that it is a derivative of Nm (Vazquez et al. 1993). Alternatively, it may be due to the

408   fact that Ng has a lower census population size. Currently ~10% of the human population is

409   asymptomatically infected with Nm (Claus et al. 2005; Yazdankhah et al. 2004), whereas levels

410   of Ng infection are thought to be very low – between 1 and 170 cases per 100,000 individuals in

411 Western Europe and America in 2017 (www.cdc.gov, ecdec.europa.eu). Hence, although there

412 seems to be a poor correlation between census and effective population size across species

413 (Bazin et al. 2006; Leffler et al. 2012; Lewontin 1974; Romiguier et al. 2014), we predict Nm to

414 have a much larger N_e than Ng, simply because it infects many more people.

415

416 In addition to the influence of hLGT we see the signature of recombination between strains of

417 the same species breaking down LD, since LD decreases with increasing distance between sites.

418 Similar patterns have been previously reported in both Nm (Budroni et al. 2011) and Ng

419 (Arnold et al. 2018) but these studies used different LD statistics and so it is hard to determine

420 what the comparative patterns are. The patterns are similar in the two species, but they are

421 consistent with a difference in N_e since the decay in LD is faster in Nm and asymptotes at a

422 slightly lower value. In both species the asymptote is above what is expected under free

423 recombination even taking into account sampling error and the fact that r^2 cannot be negative

424 (see above). The asymptote might be above this level for two reasons. First, there might be a

425 balance between drift and recombination. In a gene conversion model of recombination, a non-

426 zero asymptote is expected because once sites are further apart than the gene conversion tract

427 length, then increasing distance does not increase the rate of recombination. The asymptote is

428 then determined by a balance between drift increasing LD, and recombination breaking it

429 down. The second explanation is that there is population sub-structure in both species. It has

430 been argued, based on the phylogeny of strains that there is substructure in Nm (Budroni et al.

431 2011; Kong et al. 2013) and Ng (De Silva et al. 2016; Ezewudo et al. 2015; Grad et al. 2016; Lee

432 et al. 2018). In Nm it has been suggested that this structure arises because different sets of

433 strains have different restriction modification systems (Budroni et al. 2011). However, the

434 correspondence between clades of strains and these systems is not clear cut (Kong et al. 2013).

435

436 We find as others have found in some other species, that diversity varies across the genome in

437 Nm and Ng, and that this variation affects both synonymous and non-synonymous sites. This is

438 in large part driven by hLGT; regions of the genome with high rates of hLGT have high diversity.

439 However, when we focus on the part of the genome that is inferred not to have undergone

440 hLGT we find that levels of non-synonymous and synonymous diversity are correlated, but in a

441 manner which demonstrates that π_N/π_S declines with increasing π_S. A similar pattern has been

442 observed within the genomes of various eukaryotes (Castellano et al. 2018; Gossmann et al.

443 2011; Murray et al. 2017) as well as between eukaryotic species (Chen et al. 2017; Galtier 2016;

444 James et al. 2017). This pattern is consistent with an influence of linked selection on the

445 genome – regions of the genome with high levels of linked selection have low π_S, but relatively

446 high levels of π_N. Linked selection can influence diversity in two ways. First, both background

447 selection and genetic hitch-hiking can reduce the effective population size of a genomic region.

448 Second, hitch-hiking can lead to non-equilibrium dynamics which can affect the relative levels

449 of selected and neutral diversity; after a hitch-hiking event deleterious genetic diversity will

450 return to its equilibrium value faster than neutral diversity (Brandvain & Wright 2016; Do et al.

451 2015; Gordo & Dionisio 2005).

452

453 Nm and Ng occupy distinct niches and one might presume that they have undergone adaptive

454 evolution. Such adaptation might have been achieved through the acquisition of new genes,

455 and/or adaptation in their core genomes. We have tested for adaptive evolution in the core

456 genome using two approaches – a McDonald-Kreitman test in which numbers of non-

457 synonymous and synonymous substitutions are compared to numbers of non-synonymous and

458 synonymous polymorphisms (Eyre-Walker 2006; McDonald & Kreitman 1991). We find

459 significant evidence of adaptation when we compare the substitution data to the

460 polymorphism data of Nm, but no evidence if we use the polymorphism data of Ng. These

461 observations are consistent with a decrease in the N_e of Ng or an increase in Nm (Eyre-Walker

462 2002). The difference in N_e is consistent with the observation of higher diversity in Nm, lower

463 π_N/π_S, more rapid decay in LD and the lower asymptote in LD. However, it is difficult to resolve

464 whether Ng has undergone population size contraction or Nm population size expansion in the

465 past. Finally, it is tempting to estimate the fraction of substitutions fixed by adaptive evolution

466 as 1-1/FI – see (Eyre-Walker 2006). However, the simultaneous introduction of multiple

467 mutations by hLGT makes this estimate biased.

468

469   A central assumption in our analysis is that ClonalFrameML (Didelot & Wilson 2015) has

470   correctly identified regions of the genome that have undergone hLGT. The method identifies

471   the presence of hLGT from a clustering of mutations along an inferred clonal phylogeny; a

472   sudden burst of mutations along a branch in the phylogeny, that are spatially clustered

473   together in the genome are inferred to be due to hLGT. It will therefore be difficult for the

474   method to detect hLGT with relatively similar or short sequences. Furthermore, because we

475   have used a concatenation of protein coding sequences in our ClonalframeML analysis it may

476   be difficult to detect hLGT at the start and end of genes, because we will not have the flanking

477   sequences which provide additional support for hLGT. To investigate whether this latter effect

478   is important, we plotted the number of inferred hLGT events as a function of the distance from

479   the start or end of genes. We found that events are inferred slightly less often at the start/end

480   of genes, but the effect is not large (Figure 5).

481

482   The fact that ClonalFrameML has probably missed some hLGT events suggests that we may

483   have underestimated the input of variation from hLGT in both species – i.e. we have

484   underestimated r/m. However, an inability to correctly detect all hLGT events is unlikely to

485   explain the differences in the relative contribution of hLGT and mutation in the two species,

486   since both species have been treated identically. An inability to detect hLGT may however

487   explain why π_N and π_S are correlated even in the parts of the genome with no apparent hLGT

488   and hence there may be little or no variation in N_e across the genomes of Nm and Ng; there is

489   an expectation that π_N / π_S is likely to be lower amongst hLGT fragments because the

490   polymorphisms will be dominated by mutations that are fixed between species. Furthermore, it

491   is possible that all the variation that is shared between Nm and Ng is a consequence of hLGT

492   and we have not been able to identify all hLGT events.

493

494   **Conclusions**

495   We have investigated the diversity in Nm and Ng, and shown that Nm is more diverse then Ng.

496   Both species have acquired most of their variation through homologous lateral gene transfer.

497   Nm appears to have higher diversity in part due to it's higher effective population size. In both

498   species LD decays relatively slowly as a function of the distance between sites and there is

499   some evidence of adaptive evolution in the core genome of the two species.

500

501   **References**

502

503   Abrams AJ, Trees DL, and Nicholas RA. 2015. Complete Genome Sequences of Three Neisseria
504   gonorrhoeae Laboratory Reference Strains, Determined Using PacBio Single-Molecule
505   Real-Time Technology. *Genome Announc* 3. 10.1128/genomeA.01052-15
506   Arnold BJ, Gutmann MU, Grad YH, Sheppard SK, Corander J, Lipsitch M, and Hanage WP. 2018.
507   Weak Epistasis May Drive Adaptation in Recombining Bacteria. *Genetics* 208:1247-1260.
508   10.1534/genetics.117.300662
509   Bazin E, Glemin S, and Galtier N. 2006. Population size does not influence mitochondrial genetic
510   diversity in animals. *Science* 312:570-572.
511   Begun DJ, and Aquadro CF. 1992. Levels of naturally occuring DNA polymorphism correlate with
512   recombination rates in *D. melanogaster*. *Nature* 356:519-520.
513   Bennett JS, Jolley KA, Earle SG, Corton C, Bentley SD, Parkhill J, and Maiden MC. 2012. A
514   genomic approach to bacterial taxonomy: an examination and proposed reclassification
515   of species within the genus Neisseria. *Microbiology* 158:1570-1580.
516   10.1099/mic.0.056077-0
517   Bennett JS, Jolley KA, Sparling PF, Saunders NJ, Hart CA, Feavers IM, and Maiden MC. 2007.
518   Species status of Neisseria gonorrhoeae: evolutionary and epidemiological inferences
519   from multilocus sequence typing. *BMC Biol* 5:35. 10.1186/1741-7007-5-35
520   Bentley SD, Vernikos GS, Snyder LA, Churcher C, Arrowsmith C, Chillingworth T, Cronin A, Davis
521   PH, Holroyd NE, Jagels K, Maddison M, Moule S, Rabbinowitsch E, Sharp S, Unwin L,
522   Whitehead S, Quail MA, Achtman M, Barrell B, Saunders NJ, and Parkhill J. 2007.
523   Meningococcal genetic variation mechanisms viewed through comparative analysis of
524   serogroup C strain FAM18. *PLoS Genet* 3:e23. 10.1371/journal.pgen.0030023
525   Brandvain Y, and Wright SI. 2016. The Limits of Natural Selection in a Nonequilibrium World.
526   *Trends Genet* 32:201-210. 10.1016/j.tig.2016.01.004
527   Budroni S, Siena E, Hotopp JCD, Seib KL, Serruto D, Nofroni C, Comanducci M, Riley DR,
528   Daugherty SC, Angiuoli SV, Covacci A, Pizza M, Rappuoli R, Moxon ER, Tettelin H, and
529   Medini D. 2011. Neisseria meningitidis is structured in clades associated with restriction
530   modification systems that modulate homologous recombination. *Proceedings of the*
531   *National Academy of Sciences of the United States of America* 108:4494-4499.
532   10.1073/pnas.1019751108
533   Castellano D, James J, and Eyre-Walker A. 2018. Nearly neutral evolution across the Drosophila
534   melanogaster genome. *Mol Biol Evol* in press.
535   Charlesworth B, Morgan MT, and Charlesworth D. 1993. The effect of deleterious mutations on
536   neutral molecular variation. *Genetics* 134:1289-1303.
537   Charlesworth D, Charlesworth B, and Morgan MT. 1995. The pattern of neutral molecular
538   variation under the background selection model. *Genetics* 141:1619-1632.

539 Charlesworth J, and Eyre-Walker A. 2008. The McDonald-Kreitman test and slightly deleterious
540   mutations. *Mol Biol Evol* in press.
541 Chen J, Glemin S, and Lascoux M. 2017. Genetic Diversity and the Efficacy of Purifying Selection
542   across Plant and Animal Species. *Mol Biol Evol* 34:1417-1428. 10.1093/molbev/msx088
543 Chen X, and Zhang J. 2013. No gene-specific optimization of mutation rate in Escherichia coli.
544   *Mol Biol Evol* 30:1559-1562. 10.1093/molbev/mst060
545 Chung GT, Yoo JS, Oh HB, Lee YS, Cha SH, Kim SJ, and Yoo CK. 2008. Complete genome
546   sequence of Neisseria gonorrhoeae NCCP11945. *J Bacteriol* 190:6035-6036.
547   10.1128/JB.00566-08
548 Claus H, Maiden MC, Wilson DJ, McCarthy ND, Jolley KA, Urwin R, Hessler F, Frosch M, and
549   Vogel U. 2005. Genetic analysis of meningococci carried by children and young adults. *J
550   Infect Dis* 191:1263-1271. 10.1086/428590
551 Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F,
552   Wilczynski B, and de Hoon MJL. 2009. Biopython: freely available Python tools for
553   computational molecular biology and bioinformatics. *Bioinformatics* 25:1422-1423.
554   10.1093/bioinformatics/btp163
555 De Silva D, Peters J, Cole K, Cole MJ, Cresswell F, Dean G, Dave J, Thomas DR, Foster K, Waldram
556   A, Wilson DJ, Didelot X, Grad YH, Crook DW, Peto TEA, Walker AS, Paul J, and Eyre DW.
557   2016. Whole-genome sequencing to determine transmission of Neisseria gonorrhoeae:
558   an observational study. *Lancet Infectious Diseases* 16:1295-1303. 10.1016/S1473-
559   3099(16)30157-8
560 Didelot X, and Wilson DJ. 2015. ClonalFrameML: Efficient Inference of Recombination in Whole
561   Bacterial Genomes. *Plos Computational Biology* 11. ARTN e1004041
562 10.1371/journal.pcbi.1004041
563 Do R, Balick D, Li H, Adzhubei I, Sunyaev S, and Reich D. 2015. No evidence that selection has
564   been less effective at removing deleterious mutations in Europeans than in Africans. *Nat
565   Genet* 47:126-131. 10.1038/ng.3186
566 Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput.
567   *Nucleic Acids Res* 32:1792-1797. 10.1093/nar/gkh340
568 32/5/1792 [pii]
569 Eyre-Walker A. 2002. Changing effective population size and the McDonald-Kreitman test.
570   *Genetics* 162:2017-2024.
571 Eyre-Walker A. 2006. The genomic rate of adaptive evolution. *Trends Ecol Evol* 21:569-575.
572 Ezewudo MN, Joseph SJ, Castillo-Ramirez S, Dean D, Del Rio C, Didelot X, Dillon JA, Selden RF,
573   Shafer WM, Turingan RS, Unemo M, and Read TD. 2015. Population structure of
574   Neisseria gonorrhoeae based on whole genome data and its relationship with antibiotic
575   resistance. *PeerJ* 3:e806. 10.7717/peerj.806
576 Fay J, Wycoff GJ, and Wu C-I. 2001. Positive and negative selection on the human genome.
577   *Genetics* 158:1227-1234.
578 Feil EJ, Holmes EC, Bessen DE, Chan MS, Day NP, Enright MC, Goldstein R, Hood DW, Kalia A,
579   Moore CE, Zhou J, and Spratt BG. 2001. Recombination within natural populations of
580   pathogenic bacteria: short-term empirical estimates and long-term phylogenetic
581   consequences. *Proc Natl Acad Sci U S A* 98:182-187. 10.1073/pnas.98.1.182

582   Galtier N. 2016. Adaptive Protein Evolution in Animals and the Effective Population Size
583   Hypothesis. *PLoS Genet* 12:e1005774. 10.1371/journal.pgen.1005774
584   Gojobori J, Tang H, Akey JM, and Wu CI. 2007. Adaptive evolution in humans revealed by the
585   negative correlation between polymorphism and fixation pahses of evolution. *Proc Natl*
586   *Acad Sci USA* 104:3907-3912.
587   Gordo I, and Dionisio F. 2005. Nonequilibrium model for estimating parameters of deleterious
588   mutations. *Phys Rev E Stat Nonlin Soft Matter Phys* 71:031907.
589   10.1103/PhysRevE.71.031907
590   Gossmann TI, Woolfit M, and Eyre-Walker A. 2011. Quantifying the variation in the effective
591   population size within a genome. *Genetics* 189:1389-1402.
592   10.1534/genetics.111.132654
593   Grad YH, Harris SR, Kirkcaldy RD, Green AG, Marks DS, Bentley SD, Trees D, and Lipsitch M.
594   2016. Genomic Epidemiology of Gonococcal Resistance to Extended-Spectrum
595   Cephalosporins, Macrolides, and Fluoroquinolones in the United States, 2000-2013.
596   *Journal of Infectious Diseases* 214:1579-1587. 10.1093/infdis/jiw420
597   Hao W, Ma JH, Warren K, Tsang RS, Low DE, Jamieson FB, and Alexander DC. 2011. Extensive
598   genomic variation within clonal complexes of Neisseria meningitidis. *Genome Biol Evol*
599   3:1406-1418. 10.1093/gbe/evr119
600   Hill WG, and Robertson A. 1968. Linkage disequilibrium in finite populations. *Theoret Appl*
601   *Genet* 38:226-231.
602   James J, Castellano D, and Eyre-Walker A. 2017. DNA sequence diversity and the efficiency of
603   natural selection in animal mitochondrial DNA. *Heredity (Edinb)* 118:88-95.
604   10.1038/hdy.2016.108
605   Joseph B, Schneiker-Bekel S, Schramm-Gluck A, Blom J, Claus H, Linke B, Schwarz RF, Becker A,
606   Goesmann A, Frosch M, and Schoen C. 2010. Comparative genome biology of a
607   serogroup B carriage and disease strain supports a polygenic nature of meningococcal
608   virulence. *J Bacteriol* 192:5363-5377. 10.1128/JB.00883-10
609   Joseph B, Schwarz RF, Linke B, Blom J, Becker A, Claus H, Goesmann A, Frosch M, Muller T,
610   Vogel U, and Schoen C. 2011. Virulence Evolution of the Human Pathogen Neisseria
611   meningitidis by Recombination in the Core and Accessory Genome. *PLoS ONE* 6. ARTN
612   e18441
613   10.1371/journal.pone.0018441
614   Kong Y, Ma JH, Warren K, Tsang RS, Low DE, Jamieson FB, Alexander DC, and Hao W. 2013.
615   Homologous recombination drives both sequence diversity and gene content variation
616   in Neisseria meningitidis. *Genome Biol Evol* 5:1611-1627. 10.1093/gbe/evt116
617   Lee RS, Seemann T, Heffernan H, Kwong JC, da Silva AG, Carter GP, Woodhouse R, Dyet KH,
618   Bulach DM, Stinear TP, Howden BP, and Williamson DA. 2018. Genomic epidemiology
619   and antimicrobial resistance of Neisseria gonorrhoeae in New Zealand. *Journal of*
620   *Antimicrobial Chemotherapy* 73:353-364. 10.1093/jac/dkx405
621   Leffler EM, Bullaughey K, Matute DR, Meyer WK, Segurel L, Venkat A, Andolfatto P, and
622   Przeworski M. 2012. Revisiting an old riddle: what determines genetic diversity levels
623   within species? *PLoS biology* 10:e1001388. 10.1371/journal.pbio.1001388
624   Lewontin RC. 1974. *The genetic basis of evolutionary change*. New York: Columbia University
625   Press.

626   Maddamsetti R, Hatcher PJ, Cruveiller S, Medigue C, Barrick JE, and Lenski RE. 2015.
627   Synonymous Genetic Variation in Natural Isolates of Escherichia coli Does Not Predict
628   Where Synonymous Substitutions Occur in a Long-Term Experiment. *Mol Biol Evol*
629   32:2897-2904. 10.1093/molbev/msv161
630   Martincorena I, and Luscombe NM. 2013. Non-random mutation: The evolution of targeted
631   hypermutation and hypomutation. *BioEssays : news and reviews in molecular, cellular*
632   *and developmental biology* 35:123-130. 10.1002/bies.201200150
633   Martincorena I, Seshasayee AS, and Luscombe NM. 2012. Evidence of non-random mutation
634   rates suggests an evolutionary risk management strategy. *Nature* 485:95-98.
635   10.1038/nature10995
636   Maynard Smith J, and Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genet Res*
637   23:23-35.
638   McDonald JH, and Kreitman M. 1991. Adaptive evolution at the *Adh* locus in *Drosophila*. *Nature*
639   351:652-654.
640   Murray GGR, Soares AER, Novak BJ, Schaefer NK, Cahill JA, Baker AJ, Demboski JR, Doll A, Da
641   Fonseca RR, Fulton TL, Gilbert MTP, Heintzman PD, Letts B, McIntosh G, O'Connell BL,
642   Peck M, Pipes ML, Rice ES, Santos KM, Sohrweide AG, Vohr SH, Corbett-Detig RB, Green
643   RE, and Shapiro B. 2017. Natural selection shaped the rise and fall of passenger pigeon
644   genomic diversity. *Science* 358:951-954. 10.1126/science.aao0960
645   Ohta T. 1972. Population size and rate of evolution. *J Mol Evol* 1:305-314.
646   Ohta T. 1977. Extension of the neutral mutation drift hypothesis. In: Kimura M, ed. *Molecular*
647   *Evolution and Polymorphism*. Mishima: National Institute of Genetics, 148-167.
648   Ohta T. 1992. The nearly neutral theory of molecular evolution. *Ann Rev Ecol Syst* 23:263-286.
649   Parkhill J, Achtman M, James KD, Bentley SD, Churcher C, Klee SR, Morelli G, Basham D, Brown
650   D, Chillingworth T, Davies RM, Davis P, Devlin K, Feltwell T, Hamlin N, Holroyd S, Jagels
651   K, Leather S, Moule S, Mungall K, Quail MA, Rajandream MA, Rutherford KM, Simmonds
652   M, Skelton J, Whitehead S, Spratt BG, and Barrell BG. 2000. Complete DNA sequence of
653   a serogroup A strain of Neisseria meningitidis Z2491. *Nature* 404:502-506.
654   10.1038/35006655
655   Peng J, Yang L, Yang F, Yang J, Yan Y, Nie H, Zhang X, Xiong Z, Jiang Y, Cheng F, Xu X, Chen S, Sun
656   L, Li W, Shen Y, Shao Z, Liang X, Xu J, and Jin Q. 2008. Characterization of ST-4821
657   complex, a unique Neisseria meningitidis clone. *Genomics* 91:78-87.
658   10.1016/j.ygeno.2007.10.004
659   Roberts L. 2008. An ill wind, bringing meningitis. *Science* 320:1710-1715.
660   Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, Chenuil A, Chiari Y, Dernat R, Duret L,
661   Faivre N, Loire E, Lourenco JM, Nabholz B, Roux C, Tsagkogeorga G, Weber AAT, Weinert
662   LA, Belkhir K, Bierne N, Glemin S, and Galtier N. 2014. Comparative population genomics
663   in animals uncovers the determinants of genetic diversity. *Nature* 515:261-U243.
664   10.1038/nature13685
665   Rusniok C, Vallenet D, Floquet S, Ewles H, Mouze-Soulama C, Brown D, Lajus A, Buchrieser C,
666   Medigue C, Glaser P, and Pelicic V. 2009. NeMeSys: a biological resource for narrowing
667   the gap between sequence and function in the human pathogen Neisseria meningitidis.
668   *Genome Biol* 10:R110. 10.1186/gb-2009-10-10-r110

669   Sattath S, Elyashiv E, Kolodny O, Rinott Y, and Sella G. 2011. Pervasive Adaptive Protein
670   Evolution Apparent in Diversity Patterns around Amino Acid Substitutions in Drosophila
671   simulans. *PLoS Genetics* 7. ARTN e1001302
672   10.1371/journal.pgen.1001302
673   Schoen C, Blom J, Claus H, Schramm-Gluck A, Brandt P, Muller T, Goesmann A, Joseph B,
674   Konietzny S, Kurzai O, Schmitt C, Friedrich T, Linke B, Vogel U, and Frosch M. 2008.
675   Whole-genome comparison of disease and carriage strains provides insights into
676   virulence evolution in Neisseria meningitidis. *Proc Natl Acad Sci U S A* 105:3473-3478.
677   10.1073/pnas.0800151105
678   Schoen C, Weber-Lehmann J, Blom J, Joseph B, Goesmann A, Strittmatter A, and Frosch M.
679   2011. Whole-genome sequence of the transformable Neisseria meningitidis serogroup A
680   strain WUE2594. *J Bacteriol* 193:2064-2065. 10.1128/JB.00084-11
681   Spratt BG. 1988. Hybrid penicillin-binding proteins in penicillin-resistant strains of Neisseria
682   gonorrhoeae. *Nature* 332:173-176. 10.1038/332173a0
683   Spratt BG, Zhang QY, Jones DM, Hutchison A, Brannigan JA, and Dowson CG. 1989. Recruitment
684   of a penicillin-binding protein gene from Neisseria flavescens during the emergence of
685   penicillin resistance in Neisseria meningitidis. *Proc Natl Acad Sci U S A* 86:8988-8992.
686   Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA
687   polymorphism. *Genetics* 123:585-595.
688   Tettelin H, Saunders NJ, Heidelberg J, Jeffries AC, Nelson KE, Eisen JA, Ketchum KA, Hood DW,
689   Peden JF, Dodson RJ, Nelson WC, Gwinn ML, DeBoy R, Peterson JD, Hickey EK, Haft DH,
690   Salzberg SL, White O, Fleischmann RD, Dougherty BA, Mason T, Ciecko A, Parksey DS,
691   Blair E, Cittone H, Clark EB, Cotton MD, Utterback TR, Khouri H, Qin H, Vamathevan J,
692   Gill J, Scarlato V, Masignani V, Pizza M, Grandi G, Sun L, Smith HO, Fraser CM, Moxon ER,
693   Rappuoli R, and Venter JC. 2000. Complete genome sequence of Neisseria meningitidis
694   serogroup B strain MC58. *Science* 287:1809-1815.
695   Vazquez JA, Delafuente L, Berron S, Orourke M, Smith NH, Zhou JJ, and Spratt BG. 1993.
696   Ecological Separation and Genetic Isolation of Neisseria-Gonorrhoeae and Neisseria-
697   Meningitidis. *Current Biology* 3:567-572. Doi 10.1016/0960-9822(93)90001-5
698   Vos M, and Didelot X. 2009. A comparison of homologous recombination rates in bacteria and
699   archaea. *ISME J* 3:199-208. ismej200893 [pii]
700   10.1038/ismej.2008.93
701   Welch JJ, Eyre-Walker A, and Waxman D. 2008. Divergence and Polymorphism Under the
702   Nearly Neutral Theory of Molecular Evolution. *J Mol Evol* 67:418-426. 10.1007/s00239-
703   008-9146-9
704   WHO. 2012. Global action plan to control the spread and impact of antimicrobial resistance in
705   *Neisseria gonorrhoeae*.
706   Yazdankhah SP, Kriz P, Tzanakaki G, Kremastinou J, Kalmusova J, Musilek M, Alvestad T, Jolley
707   KA, Wilson DJ, McCarthy ND, Caugant DA, and Maiden MC. 2004. Distribution of
708   serogroups and genotypes among disease-associated and carried isolates of Neisseria
709   meningitidis from the Czech Republic, Greece, and Norway. *J Clin Microbiol* 42:5146-
710   5153. 10.1128/JCM.42.11.5146-5153.2004

711   Yu D, Jin Y, Yin ZQ, Ren HG, Zhou W, Liang L, and Yue JJ. 2014. A Genome-Wide Identification of
712   Genes Undergoing Recombination and Positive Selection in Neisseria. *Biomed Research*
713   *International*. Artn 815672
714   10.1155/2014/815672
715   Zhang Y, Yang J, Xu L, Zhu Y, Liu B, Shao Z, Zhang X, and Jin Q. 2014. Complete Genome
716   Sequence of Neisseria meningitidis Serogroup A Strain NMA510612, Isolated from a
717   Patient with Bacterial Meningitis in China. *Genome Announc* 2.
718   10.1128/genomeA.00360-14
719
720

Nucleotide diversity estimates across all sites in the core genome (pi) and at 0-fold non-synonymous sites (pi_N) and 4-fold synonymous sites (pi_S).

	π	π_S	π_N
N. gonorrhoeae	0.0029 (0.0008)	0.007 (0.002)	0.0014 (0.0004)
N. meningitidis	0.022 (0.007)	0.06 (0.02)	0.007 (0.002)
Ratio	7.6	8.6	5.0

4 **Table 1.** Diversity estimates across all sites in the core genome (π) and at 0-fold non-

5 synonymous sites (π_N) and 4-fold synonymous sites (π_S).

Recombination rate estimates obtained from ClonalFrameML along with their 95% confidence intervals.

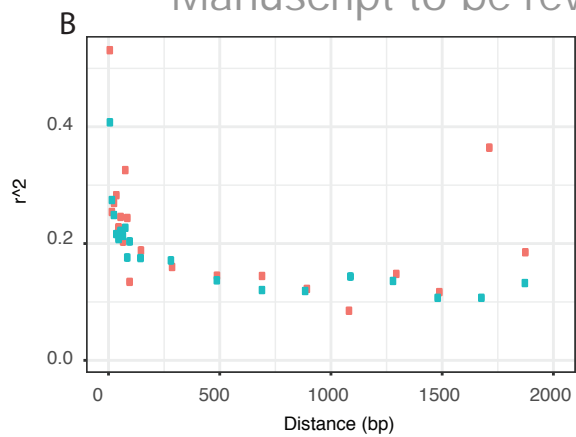
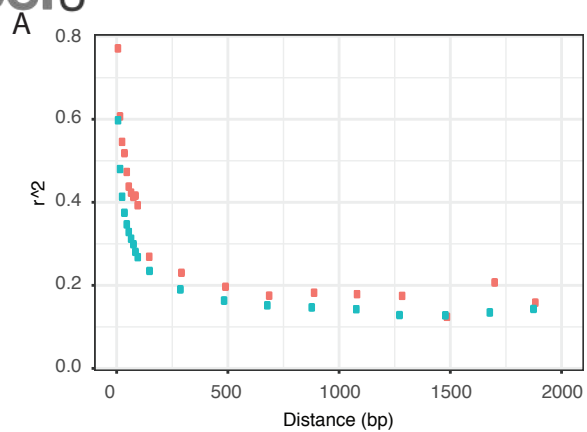
Given is the rate at which recombination tracts initiate (R) relative to the rate of mutation (theta), both multiplied by the effective population size, the average length of recombination tracts (delta) and the proportion of sites that differ to the resident sequence (mu), along with the rate at which sites change due to recombination relative to mutation (r/m)

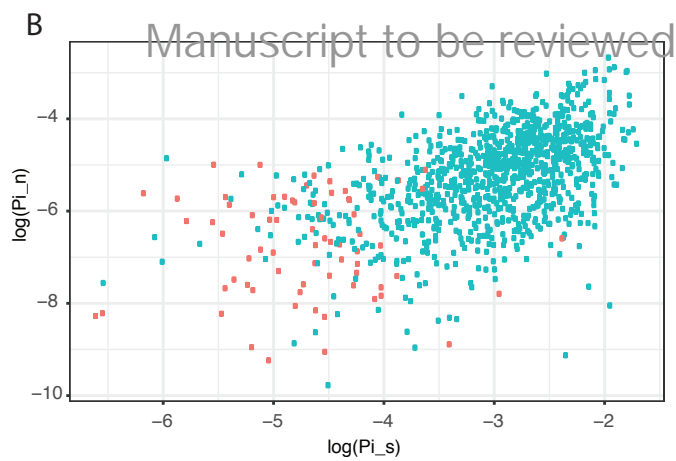
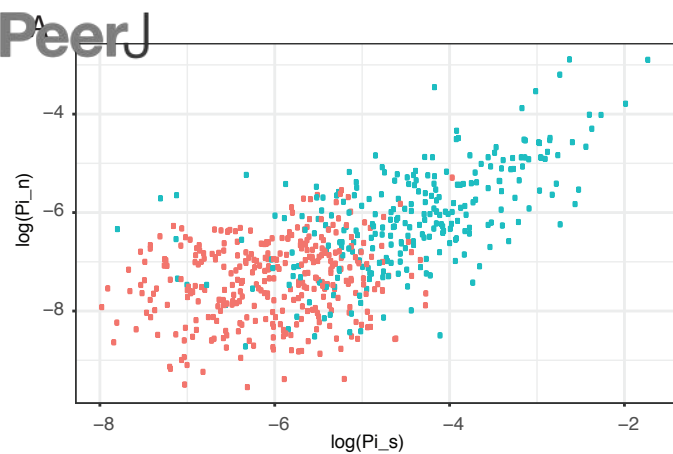
Species	R/θ	δ	ν (%)	r/m	θ (x 10⁻³)	R (x 10⁻⁴)
Ng	0.41 (0.39, 0.43)	70 (67, 72)	6.9 (6.8, 7.1)	2.0 (1.8, 2.2)	1.0 (0.8, 1.2)	4.0 (3.0, 5.0)
Nm	1.2 (1.2, 1.3)	99 (98, 100)	5.3 (5.3, 5.4)	6.4 (6.2, 6.7)	3.0 (2.5, 3.5)	36 (30, 44)
Ratio (Nm/Ng)	3.0	1.4	0.77	3.2	3.0	9.3

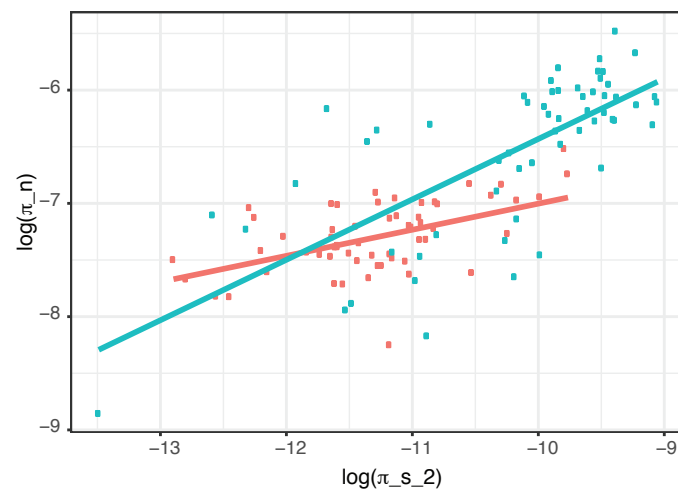
3
4   **Table 2.** Recombination rate estimates obtained from ClonalFrameML along with their 95%
5   confidence intervals.

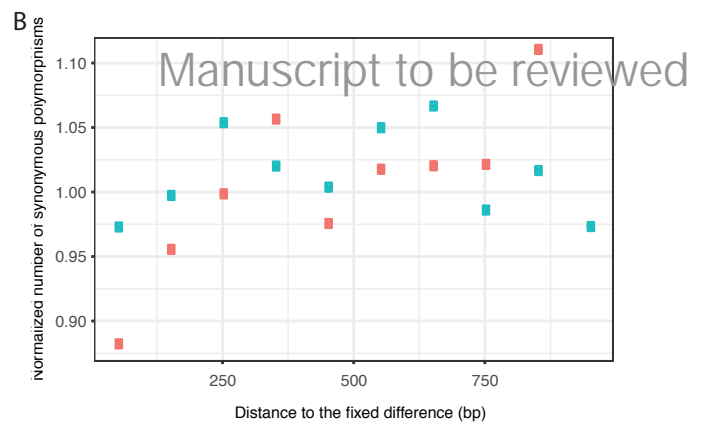
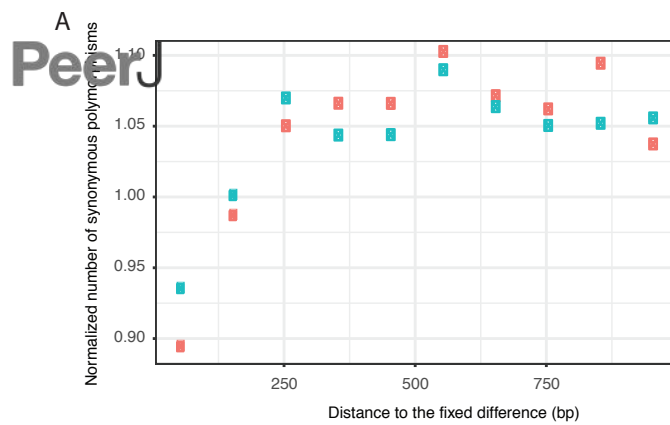
Decay in linkage disequilibrium with the distance between sites.

Linkage disequilibrium, as measured by r^2 (Hill & Robertson 1968), between pairs of polymorphic sites as a function of the distance between sites for (A) all sites and (B) for those sites not inferred to have undergone hLGT. Each point represents the average r^2 between all pairs of points separated by a certain distance in bins of 10bp between 0 and 100bp, a bin of 101 to 200bp and then bins of 200bp upto 800bp. Nm in green, Ng in red.

low

Correlation between non-synonymous and synonymous diversity across the genome.

he correlation between the log of the non-synonymous nucleotide diversity and the log of the synonymous diversity for core genes in A) Nm and B) Ng. Points in green are genes with evidence of hLGT and red are those genes without evidence of hLGT. Note that some genes are excluded because they have either no non-synonymous or synonymous diversity.

Correlation between non-synonymous and synonymous diversity excluding regions with evidence of hLGT.

he correlation between the log of the non-synonymous nucleotide diversity plotted and the log of the synonymous diversity for regions of the genome that have not undergone hLGT. Green is Nm, red is Ng. Also shown are the lines of best fit.

Synonymous diversity around sites fixed for either non-synonymous or synonymous substitutions.

Average synonymous diversity in A) Nm and B) Ng around sites that are fixed for either a non-synonymous (red) or synonymous (green) substitution between Nm and Ng.

hLGT tracts at the start and end of genes,

The number of sequences inferred to be due to hLGT in both species as a function of the distance from the A) start and B) end of genes, where the distance was the proportion of the gene length from the start and end