A peer-reviewed version of this preprint was published in PeerJ on 27 June 2019.

View the peer-reviewed version (peerj.com/articles/7216), which is the preferred citable publication unless you specifically need to cite this preprint.

Vigué L, Eyre-Walker A. 2019. The comparative population genetics of Neisseria meningitidis and Neisseria gonorrhoeae. PeerJ 7:e7216 https://doi.org/10.7717/peerj. 7216

# The comparative population genetics of Neisseria meningitidis and Neisseria gonorrhoeae 

Lucile Vigue ${ }^{1}$, Adam Eyre-Walker ${ }^{\text {Corresp. } 2}$

${ }^{1}$ Ecole Polytechnique, Paris, France
2 School of Life Sciences, University of Sussex
Corresponding Author: Adam Eyre-Walker
Email address: a.c.eyre-walker@sussex.ac.uk
Neisseria meningitidis ( Nm ) and $N$. gonorrhoeae $(\mathrm{Ng})$ are closely related pathogenic bacteria. Using those genes found across 20 Nm and 15 Ng genomes we find that Nm is 7 x 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 both species. Several observations suggest that Nm has a higher effective population size than Ng ; it is more diverse, the ratio of nonsynonymous to synonymous polymorphism is lower, and LD declines more rapidly to a lower asymptote. 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 and found some evidence.

The comparative population genetics of Neisseria meningitidis and Neisseria gonorrhoeae

Lucile Vigue
Adam Eyre-Walker

School of Life Sciences
University of Sussex
Brighton
BN1 9QG
United Kingdom

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


#### Abstract

Neisseria meningitidis $(\mathrm{Nm})$ and N . gonorrhoeae $(\mathrm{Ng})$ are closely related pathogenic bacteria. Using those genes found across 20 Nm and 15 Ng genomes we find that Nm is 7 x 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 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 LD declines more rapidly to a lower asymptote. 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 and found some evidence.


## Introduction

The two closely related bacteria Neisseria meningitidis (Nm) and Neisseria gonorhoeae (Ng) are major human pathogens. Ng is the causative agent of the sexually transmitted disease gonorrhoeae which currently infects 106 million people each year worldwide (WHO 2012). When untreated, gonoccocal infections can result in long-term problems such as persistent urethritis, cervicitis, proctitis, pelvic inflammatory disease, infertility, first-trimester abortion, ectopic pregnancy and maternal death (WHO 2012). They also increase the risk of acquiring and transmitting HIV. In cases of pregnancy, Ng infections can cause severe damages to neonatal health (WHO 2012). In contrast, Nm is a human commensal infecting between $8-25 \%$ of the healthy human population (Claus et al. 2005; Yazdankhah et al. 2004), which only occasionally causes disease. However, it can cause meningococcal meningitidis and septicaemia with mortality rates that can reach $50 \%$ when untreated, and the global disease burden is estimated to be $\sim 500,000$ cases a year (Roberts 2008). Among the different micro-organisms that can cause meningitidis, it is regarded as one of the most important because of its ability to cause large epidemics

Here we consider several aspects of the population genetics of these bacterial species. The two species are sister taxa (Bennett et al. 2012), and Nm is known to be considerably more diverse than Ng within the genes that they share in common (Bennett et al. 2012; Bennett et al. 2007). There are several potential explanations for this. First, Ng might have a lower effective population size, either because it evolved from Nm and went through a bottleneck when the species was formed \{Vazquez, 1993 \#1591\}, or because it generally has a lower effective population size, possibly because it has a lower census population size. Second, Ng might have a lower mutation rate than Nm . Third, Ng might acquire less diversity through recombination than Nm . Both Ng and Nm are known to be naturally transformable, and it has been known for many years that both species acquire diversity, within their core genome, by homologous
recombination with genetically divergent strains (Spratt 1988; Spratt et al. 1989). We refer to this process as homologous lateral gene transfer (hLGT), to differentiate it from the acquisition of accessory genes by non-homologous lateral gene transfer (nhLGT) (however, note that the acquisition of new genes generally involves homologous recombination with flanking genes, so nhLGT will typically involve hLGT (Kong et al. 2013)). hLGT leads to mosaic genes, in which parts of the gene have been acquired from a different bacterial species. In fact, Nm and Ng were some of the first bacteria in which this form of recombination was demonstrated (Spratt 1988; Spratt et al. 1989). It has been estimated that Nm acquires SNPs through hLGT at a rate between 4 and 100x higher than via mutation (Feil et al. 2001; Hao et al. 2011; Kong et al. 2013; Vos \& Didelot 2009). In contrast this ratio has recently been estimated to be only about 2-fold in Ng (Ezewudo et al. 2015). It is unclear whether these ratios are significantly different.

The second question, we address is whether diversity varies across the core genome of the two species. Genetic diversity is known to vary across the genome of many species. This was originally demonstrated in Drosophila melanogaster by Begun and Aquadro (Begun \& Aquadro 1992) who showed that diversity was positively correlated to the rate of recombination. This was thought to be due to the effects of linked selection, in the form of genetic hitch-hiking (Maynard Smith \& Haigh 1974) and background selection (Charlesworth et al. 1993) depressing diversity in regions of the genome with low rates of recombination. Variation in diversity across the genome has been demonstrated in many other species including the bacterium Escherichia coli (Maddamsetti et al. 2015; Martincorena et al. 2012). The reasons for this variation remain unclear (Chen \& Zhang 2013; Maddamsetti et al. 2015; Martincorena \& Luscombe 2013).

The final question we address is whether Nm and Ng have undergone adaptive evolution. Nm and Ng inhabit different niches and one presumes they have undergone adaptive evolution to allow them to do this. Some, may be much of this adaptation might be through the acquisition of new genes via nhLGT, but there might also be adaptation in the core genome. Two recent analyses using the $d_{N} / d_{s}$ test on the core genome have found limited evidence for adaptive

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

88 89 conservative. Here we apply two additional tests.

## Materials and methods

## Dataset

Fifteen genomes of Neisseria gonorrhoeae (FA 1090, NCCP11945, MS11, FA19, FA6140, 35/02, 32867, 34530, 34769, FDAARGOS 204, FDAARGOS 205, FDAARGOS 207, NCTC13799, NCTC13798, NCTC13800) and 20 genomes of Neisseria meningitidis (MC58, Z2491, FAM18, 053442, alpha14, 8013, alpha710, WUE 2594, G2136, M01-240149, M04-240196, H44/76, M01240355, NZ-05/33, 510612, NM3686, M7124, NM3682, NM3683, L91543) were downloaded from Genbank. From these all protein coding sequences were extracted. We retained those coding sequences that started NTG, terminated with TAA, TAG or TGA and had a length that is a multiple of 3 . We identified orthologs using reciprocal BLAST, with an e value threshold of 0.00001; i.e. each protein coding gene in each genome was BLASTed against the genes of FA1090, and then the best hit was BLASTed back onto the original genome, and retaining only those hits in which the original query sequence was the best hit. Similar selections of genes were obtained using alternative starting genomes. The protein sequences of the orthologs were aligned using MUSCLE (Edgar 2004). We selected genes where the alignments meet these criteria: the number of gaps is lower than $1 \%$ of the length of the sequence and the total number of nucleotides in gaps is lower than 10\% of the total number of nucleotides in the sequence. Sequences with internal stop codons were removed. This resulted in a dataset of 1145 genes belonging to the core genome of both Neisseria gonorrhoeae and Neisseria meningitidis. We used the BioPython Phylo library (Cock et al. 2009) to estimate a phylogeny of the strains based on the core genome alignment.

## Analyses

We detected hLGT using ClonalFrameML (Didelot \& Wilson 2015) run on a concatenation of the protein coding sequences from the core genome of both species. For some analyses we masked those regions inferred to be due to hLGT, in the strains affected.

We investigated whether linkage disequilibrium (LD) declines with the distance between sites by measuring the LD between all pairs of polymorphisms within each gene. We measured LD
using the $r^{2}$ statistic $\{$ Hill, 1968 \#320\}. LD values were then assigned to bins based on the distance between the two sites - 10bp bins between 1-100bp, a bin from 101-200bp and then 200bp bins between 201-800bp. We took the average LD and distance between sites for each bin in manner which weighted each gene equally - we estimated the average LD and distance for pairs of sites in each bin for each gene and then averaged those values across genes. To estimate the approximate half-life of LD we found the distance between sites that gave approximately half the LD between the LD for the 1-10bp bin and the asymptotic value of the LD. Because $r^{2}$ is constrained to be positive the expected value of $r^{2}$ is greater than zero even when there is no LD. To calculate the expected value we simulated two bi-allelic loci with mutations at particular frequencies; we generated the expected frequencies of the four haplotypes expected under free recombination, and then sampled $N$ chromosomes with those expected haplotype frequencies using a multi-nomial distribution. We repeated this 10,000 times. We found that the expected value of $r^{2}$ is independent of the allele frequencies.

To investigate the relationship between the non-synonymous, $\pi_{N}$, and synonymous, $\pi_{\mathrm{S}}$, nucleotide diversity in detail we used a variation of the method of James et al. (James et al. 2017) to combine data from different genes. We split the synonymous polymorphisms into two groups according to whether they were in an odd or even numbered codon and then used the two groups to estimate two synonymous nucleotide diversities that have independent sampling errors, $\pi_{\mathrm{s} 1}$ and $\pi_{\mathrm{s} 2}$. One of these, say $\pi_{\mathrm{s} 1}$, is used to rank and group genes, and the other is averaged across genes in the group to give an unbiased estimate of $\pi_{s}$ for the group. $\pi_{N}$ is also averaged across the genes in the group.

To investigate the diversity around sites that are fixed between Nm and Ng for different alleles we focused on genes that had at least one synonymous polymorphism and one fixed difference between the two species. For each fixed difference, we identified all the synonymous polymorphisms that were less than 1 kb away and we grouped them by windows of 100 bp . Because background selection can potentially lead to a lower dip in diversity around fixed nonsynonymous mutations we normalised the diversity around fixed synonymous and non-
synonymous substitutions by dividing the number of synonymous polymorphisms in a particular window by the total number of synonymous polymorphisms in the gene, multiplied by the window size over the gene length.

## Results

## Recombination and mutation

We are interested in how genetic variation is generated and distributed in the two Neisseria species $N$. meningitidis ( Nm ) and $N$. gonorrhoeae ( Ng ). Although, the presence and absence of genes in the strains of the two species is an important aspect of this problem, here we focus on the genetic variation that is present in the core genome that is common to both species. Using reciprocal BLAST we identified 1145 genes present across the 15 genomes of Ng and 20 genomes of Nm that we analysed. The total length of this core genome is 1.1 MB long. We find that Nm is $\sim 7.6$ fold more diverse than Ng consistent with previous qualitative reports (Bennett et al. 2012; Bennett et al. 2007). The difference in diversity is more apparent at synonymous ( $\sim 8.9$ fold) than non-synonymous ( $\sim 5.5$ fold) sites (Table 1), a pattern we return to later. The two species share a modest amount of diversity; 35\% of all polymorphisms in Ng are shared with Nm , and $4.5 \%$ of those in Nm are shared with Ng .

It is well known that Nm and Ng undergo substantial levels of homologous recombination with divergent strains, possibly from other species of bacteria. This leads both to the acquisition of new genes, but also to the acquisition of parts of genes that are already present in the genome; we refer to these processes as non-homologous (nhLGT) and homologous lateral gene transfer (hLGT) respectively. To quantify the role that hLGT plays in the acquisition of diversity in the core genome we ran ClonalFrameML (Didelot \& Wilson 2015) (Didelot \& Wilson 2015). The method estimates the ratio of the rate at which recombination tracts initiate $(R)$ and the rate of mutation $(\theta)$, both multiplied by twice the effective population size, $N_{e}$, along with the average recombination tract length, $\delta$, and the proportion of sites that differ between the imported and resident sequences, $v$. Estimates of these parameters are given in Table 2. The overall effect of
recombination relative to mutation can be estimated as $R \delta v / \theta=r / m$. In Nm we find that recombination introduces $6.43(95 \% \mathrm{Cl}=6.16$ to 6.71$)$ times more variation than mutation, whereas in Ng it introduces $1.97(1.76,2.19)$ times as much. These estimates are similar to some of those obtained previously. In Nm the r/m ratio has been estimated to be 5.37(Hao et al. 2011), 6.71 (Vos \& Didelot 2009), 16.4 (Kong et al. 2013) and 100 (Feil et al. 2001). So, our estimate is similar to the first two estimates, but substantially lower than the last two estimates. Both of these latter estimates were obtained from very closely related strains and hence may reflect the value of $r / m$ before natural selection has had an opportunity to operate. In Ng it has been estimated that 2.2 x as much variation is introduced by recombination (Ezewudo et al. 2015), which is very similar to our estimate. The estimates of $r / m$ mean that ~87\% of all polymorphisms in Nm are a consequence of recombination, whereas in Ng it is $66 \%$. The difference between the two species in the influence of recombination is largely driven by a difference in the ratio of the rate at which recombination initiates versus the mutation rate $(R / \theta)$, since although the tract lengths are on average slightly longer in Nm, they introduce slightly less variation than Ng (Table 2).

ClonalframeML estimates the ratio of $R$ and $\theta$ but not their absolute values. However, we can estimate the absolute value as follows. We note that the nucleotide diversity is due to the input of mutation and the input of recombination: i.e. $\pi=\theta+R \delta v$. If we note that ClonalframeML gives us an estimate of $R / \theta$ we can rewrite this equation as $\pi=\theta+\theta \delta \cup R / \theta$, from which we can estimate $\theta=\pi /(1+\delta \cup R / \theta)$. Estimates of $R$ and $\theta$ are given in Table 2. From this it is evident that the nucleotide diversity is higher in Nm both because of a 3 -fold greater mutational input, and a 9-fold greater rate at which recombination tracts initiate in Nm.

The parameters $R$ and $\theta$ are the rates of recombination initiation and mutation, multiplied by the effective population size. Hence a simple reason why both parameters are higher in Nm might simply be that Nm has a higher $N_{e}$ than Ng . To test this idea, we masked all sequences that were identified as due to hLGT by ClonalframeML and estimated the levels of nonsynonymous and synonymous diversity. Under a model in which synonymous mutations are
neutral and non-synonymous mutations are deleterious, but drawn from some distribution, we expect $\pi_{N} / \pi_{\mathrm{S}}$ to be lower in species with high $N_{e}$; this is because selection is more effective in species with higher $N_{e}$ and hence the proportion of mutations that are effectively neutral is lower \{Ohta, 1972 \#342;Ohta, 1977 \#35;Ohta, 1992 \#459\}. This is what we find $-\pi_{\mathrm{N}} / \pi_{\mathrm{S}}=0.095$ ( $\mathrm{SE}=0.0023$ ) in Nm versus 0.23 ( 0.014 ) in Ng. These are significantly different to each other (normal test $\mathrm{z}=9.5, \mathrm{p}<0.001$ ).

As we described above, Nm and Ng share a modest amount of genetic variation. It is of some interest whether this is a consequence of hLGT or the inheritance of genetic variation from their common ancestor. If we exclude those sequences inferred to be due to hLGT we find that the two species still share a modest amount of genetic variation $-15.5 \%$ of all Ng polymorphisms are shared with Nm and $2.4 \%$ of Nm polymorphisms are shared with Ng , approximately half of all shared polymorphisms in each case.

## Linkage disequilibrium

Homologous recombination can both increase and decrease linkage disequilibrium (LD); homologous recombination with divergent strains, of the sort detected by ClonalFrameML, generates LD because it simultaneously introduces many mutations that are initially linked to each other. However, homologous recombination amongst a set of closely related strains breaks-up LD. To investigate how these two forces play out, we calculated the LD between all pairs of sites in the genome and plotted these as a function of the distance between sites. As expected we observe a decline in LD with distance (Figure 1A). Both species show similar patterns with LD declining rapidly; in Nm the approximate half-life is 30bp and in Ng it is 100bp. The decline could be due to two processes. If most hLGT fragments tend to be short, with decreasing numbers of long fragments, then LD will be greater between closely linked sites. However, we also expect a decline due to recombination between closely related strains, and in fact we observe a decline even when we focus on those parts of the genome which do not appear to have undergone hLGT (Figure 1B).

In both species LD asymptotes above zero, i.e. above the value you would expect under free recombination. The non-zero asymptote could be due to one of three reasons - statistical bias, population substructure and a balance between genetic drift and recombination. The statistical bias arises because our measure of LD, $r^{2}$ (Hill \& Robertson 1968), cannot be negative, so positive values of $r^{2}$ are expected even if there is no LD if sample sizes are small; for sample sizes of 15 and 20 the expected value of $r^{2}$ is 0.079 and 0.050 respectively (see materials and methods), so the asymptote is clearly above this level. Both, Nm and Ng have been shown to have some level of population structure so this is the likely to be part of the explanation (Budroni et al. 2011; Joseph et al. 2011). However, the slower decay in LD, and higher asymptote in Ng , is consistent with Ng having a smaller $N_{e}$ than Nm - i.e. the non-zero asymptote might in part be caused by a balance between genetic drift creating LD, and recombination breaking it down.

## Diversity across the genome

Nucleotide diversity is known to vary across the genomes of many organisms. This is largely thought to be driven by variation in the mutation rate or variation in the effects of linked selection. However, in bacteria, and particularly Nm and Ng , it could also be due to variation in the frequency of hLGT. All of these processes are expected to affect synonymous and nonsynonymous diversity to greater or lesser extents, and indeed we observe a positive correlation between non-synonymous and synonymous diversity, demonstrating that both vary across the genome in concert. At least part of this pattern is driven by hLGT because genes with hLGT show higher $\pi_{N}$ and $\pi_{s}$ values than genes without any evidence of hLGT (Figure 2).

However, to investigate whether there is also variation in the effective population size we removed sequences inferred to be due to hLGT by ClonalFrameML from our data. This reduces our data substantially and so to reduce statistical sampling issues we used the method of James et al. (2016) to combine data from different genes. We find, when we do this, that $\pi_{N}$ and $\pi_{\mathrm{s}}$ are still significantly correlated suggesting the correlation between them is not just driven by hLGT ( Ng slope $=0.23, \mathrm{p}<0.001$; Nm slope $=0.53, \mathrm{p}<0.001$ )(Figure 3). The remaining correlation could
be due to variation in the mutation rate or variation in the effects of linked selection. We can test whether there is variation in the effects of linked selection by considering the slope between $\log \left(\pi_{N}\right)$ and $\log \left(\pi_{s}\right)$. Under a model in which there is no variation in linked selection then the slope of this relationship is expected to be one, and if there is variation in linked selection the slope if expected to be less than one \{Galtier, 2016 \#1452;Welch, 2008 \#892\}. Linked selection has two consequences. First, it increases the stochasticity in allele frequencies. For example, the spread of advantageous mutation or the elimination of deleterious genetic variation, removes linked genetic diversity; whether a linked mutation survives either process is a random process depending on whether the advantageous or deleterious mutation occurs in linkage with the target mutation. This can be thought of as reduction in the effective population size. Second, genetic hitch-hiking leads to non-equilibrium dynamics. After a selective sweep genetic diversity will recover, but this happens faster for deleterious than neutral mutations (Brandvain \& Wright 2016; Do et al. 2015; Gordo \& Dionisio 2005). In both cases we expect a negative correlation between $\pi_{\mathrm{N}} / \pi_{\mathrm{s}}$ and $\pi_{\mathrm{s}}$, which manifests itself in a positive correlation between $\log \left(\pi_{N}\right)$ and $\log \left(\pi_{s}\right)$ but with a slope of less than one (James et al. 2017). We find that the slope of the relationship between $\log \left(\pi_{N}\right)$ and $\log \left(\pi_{\mathrm{s}}\right)$ is $0.23(0.052)$ and 0.59 ( 0.070 ) for Ng and $N m$ respectively, in both cases significantly less than one ( $p<0.001$ ); i.e. $\pi_{N}$ increases as $\pi_{S}$ increases but not as fast. The slopes are significantly different to each other (t-test, $\mathrm{p}<0.001$ ).

## Adaptive evolution

Nm and Ng are ecologically quite different and one presumes the two species have undergone adaptation to live in their respective environments. Some, may be most of this adaptation will have come about through the acquisition of whole genes through nhLGT. However, some of the adaptation may have occurred within the core genome of the two species. To investigate whether there has been adaptation in the core genome we used two approaches. First, we used the McDonald-Kreitman (McDonald \& Kreitman 1991) approach to estimate the rate of adaptive evolution (Fay et al. 2001). In this method the numbers of non-synonymous and synonymous substitutions (i.e. differences between the two species, $d_{N}$ and $d_{S}$ respectively) are compared to the numbers of non-synonymous and synonymous polymorphisms ( $p_{N}$ and $p_{S}$
respectively). Under a neutral model in which mutations are either neutral or strongly deleterious we expect $d_{N} / d_{S}=p_{N} / p_{S}$. In contrast if there are slightly deleterious nonsynonymous mutations we expect $d_{N} / d_{S}<p_{N} / p_{S}$, and if there are some advantageous mutations we expect $d_{N} / d_{S}>p_{N} / p_{S}$. Summing $d_{N}, d_{S}, p_{N}$ and $p_{S}$ we calculate the fixation index $F I=d_{N} p_{S} / d_{S}$ $p_{N}$ (Gojobori et al. 2007); adaptive is indicated if $F I>1$.

We find that our estimate of FI differs if we use the polymorphism data of Nm or Ng ; using the SNP data of Nm we estimate that $F /$ is significantly greater than one suggesting adaptive evolution has occurred ( $F I=1.51$ with $95 \%$ Cis $=1.41$ and 1.61), but if we use the SNP data of Ng , our estimate is significantly less than one ( $0.92(0.83,0.99)$ ). Estimates less than one can occur if there are slightly deleterious mutations (SDMs) segregating, but even if we restrict our analysis to common polymorphisms, which should remove many of the SDMs (Charlesworth \& Eyre-Walker 2008; Fay et al. 2001), we find that the FI<1 using the SNP data of Ng (using SNPs with allele frequencies above $15 \%, \mathrm{FI}=0.78(0.78,0.88)$ ). An explanation for why FI differs between the two species is that either Nm has undergone population expansion, or Ng has undergone contraction. If there are slightly deleterious mutations then population size expansion leads to an overestimate of FI whereas contraction leads to an underestimate (EyreWalker 2002; McDonald \& Kreitman 1991). As we argue above, a simple explanation for why Nm is more diverse than Ng is that Nm has a higher Ne . We find no evidence of expansion or contraction amongst the current strains - Tajima's D, a measure of a skew in the site frequency spectrum away from what we expect for neutral mutations in a stationary population size is close to zero and not significantly different to zero in both species in the regions of the genome that have no evidence of hLGT ( -0.073 and -0.093 in Nm and Ng respectively), consistent with previous analyses in Nm (Joseph et al. 2011). However, the expansion or contraction in either Nm or Ng could have occurred sometime in the past.

A second approach is to look for a hallmark signature of adaptive evolution, a dip in genetic diversity around advantageous mutations (Sattath et al. 2011). We find that synonymous diversity is lower close to sites that are fixed for different nucleotides in the two species, and
this dip is significantly greater for non-synonymous than synonymous fixed differences when considering diversity in Nm ( $\mathrm{p}<0.001$ for distances 1-100bp, 101-200bp and 201-300bp), consistent with a proportion of non-synonymous mutations being fixed by positive adaptive evolution; a similar pattern is not evident in Ng , possibly because it is less diverse. There is however an alternative explanation for the greater dip around non-synonymous substitutions; if the strength of background selection varies across the genome, then regions with high levels of background selection will have low diversity but will tend to also fix slightly deleterious nonsynonymous mutations. To investigate whether there is evidence of this we considered whether the ratio of the non-synonymous to synonymous substitution rates, $\log \left(d_{N} / d_{S}\right)$, was correlated to synonymous diversity, $\log \left(\pi_{s}\right)$. We again use the method of James et al. (James et al. 2017) to combine data from different genes and find that a strong negative correlation in Nm (slope $=-0.10, \mathrm{p}=0.018$ ) but not in Ng (slope $=-0.001, \mathrm{p}=0.97$ ). This suggests that background selection might be a factor in Nm. To take into account the potential variation in background selection we normalised the data from each gene by dividing the number of synonymous SNPs in each window in each gene by the total number of synonymous SNPs in the gene, multiplied by the window length over the gene length; in other words we normalized the diversity across each gene by the average diversity in the gene. This will account for variation in background selection at a gene level, but not at a sub-gene level. The normalised data show a greater dip in diversity for fixed non-synonymous than synonymous substitutions in both Nm (combining t-test results from the 3 closest points, $p<0.001$ ), and NG $(p=0.0024)$ (figure 4) although the differences are not large.

## Discussion

We have investigated several aspects of the population genetics of the two bacteria Neisseria meningitidis and $N$. gonorrhoeae. We find, as others have (Bennett et al. 2012; Bennett et al. 2007), that Nm is substantially more diverse than Ng , but that the two species share a moderate amount of diversity in the genes that they have in common. This shared diversity could have been a consequence of ancestral polymorphism that has been inherited by both species, or due to hLGT transferring variation between the two. We find a substantial fraction is
indeed due to hLGT, since if we remove the fraction of the genome that appears to have undergone hLGT, the fraction of shared polymorphism drops considerably. However, there is some diversity that appears to have been inherited from the ancestor.

In both species we find that most of their genetic diversity has been acquired by recombination, rather than by mutation. In Nm we estimate that the total input from hLGT is 6 -fold greater than from mutation; this is in line with the estimates of Hao et al. (Hao et al. 2011) and Vos and Didelot (Vos \& Didelot 2009), but lower than two other estimates (Feil et al. 2001; Kong et al. 2013). Both of these high estimates were derived by considering very closely related strains. If hLGT events are on average more deleterious than single nucleotide changes then we expect $r / m$ estimates to be greater for more closely related strains, because natural selection has had more opportunity to remove the deleterious mutations in distantly related strains. This has the implication that $r / m$ may be far higher amongst newly arising mutations than often thought. In Ng we find the input of hLGT is 2 -fold greater than mutation, consistent with the one previous estimate performed on a similar selection of strains (Ezewudo et al. 2015).

Nm is more diverse than Ng which could be due to either a higher mutation rate, a greater rate of hLGT or a higher effective population size. Several lines of evidence suggest that Nm has a higher $N_{e}$. First, Nm has higher values of both $R$ and $\theta$, where $R$ and $\theta$ are estimates of the rate at which recombination initiates and the mutation rate, multiplied by $N_{e}$. Second, $p_{N} / p_{S}$ is lower in Nm in the fraction of the genome which does not seem to have undergone hLGT. Third, LD declines faster in Nm and asymptotes at a lower level. However, this does not preclude a role for either faster rates of mutation or recombination in the greater diversity in Nm.

It is possible that the lower $N_{e}$ in Ng is due to a bottleneck at the time when Ng was formed, assuming that it is a derivative of Nm (Vazquez et al. 1993). Alternatively, it may be due to the fact that Ng has a lower census population size. Currently $5-20 \%$ of the human population is asymptomatically infected with Nm (Claus et al. 2005; Yazdankhah et al. 2004), whereas levels
of Ng infection are thought to be very low. Hence, although there seems to be a poor correlation between census and effective population size across species (Bazin et al. 2006; Leffler et al. 2012; Lewontin 1974; Romiguier et al. 2014), we predict Nm to have a much larger $N_{e}$ than Ng , simply because it infects many more people.

In addition to the influence of hLGT we see the signature of recombination between strains of the same species breaking down LD, since LD decreases with increasing distance between sites. The patterns are similar in the two species, but they are consistent with a difference in $N_{e}$ since the decay in LD is faster in Nm and asymptotes at a slightly lower value. In both species the asymptote is above what is expected under free recombination even taking into account sampling error and the fact that $r^{2}$ cannot be negative (see above). The asymptote might be above this level for two reasons. First, there might be a balance between drift and recombination. In a gene conversion model of recombination, a non-zero asymptote is expected because once sites are further apart than the gene conversion tract length, then increasing distance does not increase the rate of recombination. The asymptote is then determined by a balance between drift increasing LD, and recombination breaking it down. The second explanation is that there is population sub-structure in both species. It has been argued, based on the phylogeny of strains that there is substructure in Nm (Budroni et al. 2011; Kong et al. 2013) and Ng (De Silva et al. 2016; Ezewudo et al. 2015; Grad et al. 2016; Lee et al. 2018). In Nm it has been suggested that this structure arises because different sets of strains have different restriction modification systems (Budroni et al. 2011). However, the correspondence between clades of strains and these systems is not very clear cut (Kong et al. 2013).

We find as others have found in some other species, that diversity varies across the genome in Nm and Ng , and that this variation affects both synonymous and non-synonymous sites. This is in large part driven by hLGT; regions of the genome with high rates of hLGT have high diversity. However, when we focus on the part of the genome that is inferred not to have undergone hLGT we find that levels of non-synonymous and synonymous diversity are correlated, but in a manner which demonstrates that $\pi_{N} / \pi_{\mathrm{S}}$ declines with increasing $\pi_{\mathrm{S}}$. A similar pattern has been
observed across the genomes of various eukaryotes (Castellano et al. 2018; Gossmann et al. 2011; Murray et al. 2017) as well as between various eukaryotic species (Chen et al. 2017; Galtier 2016; James et al. 2017). This pattern is consistent with an influence of linked selection on the genome - regions of the genome with high levels of linked selection have low $\pi_{\mathrm{s}}$, but relatively high levels of $\pi_{\mathrm{N}}$. Linked selection can influence diversity in two ways. First, both background selection and genetic hitch-hiking can reduce the effective population size of a genomic region. Second, hitch-hiking can lead to non-equilibrium dynamics which can affect the relative levels of selected and neutral diversity; after a hitch-hiking event deleterious genetic diversity will return to its equilibrium value faster than neutral diversity (Brandvain \& Wright 2016; Do et al. 2015; Gordo \& Dionisio 2005).

Nm and Ng occupy distinct niches and one might presume that they have undergone adaptive evolution. Such adaptation might have been achieved through the acquisition of new genes, and/or adaptation in their core genomes. We have tested for adaptive evolution using two approaches - a McDonald-Kreitman test in which numbers of non-synonymous and synonymous substitutions are compared to numbers of non-synonymous and synonymous polymorphisms. We find significant evidence of adaptation when we compare the substitution data to the polymorphism data of Nm , but no evidence if we use the polymorphism data of Ng . These observations are consistent with a decrease in the $N_{e}$ of Ng or an increase in Nm (EyreWalker 2002). The difference in $N_{e}$ is consistent with the observation of higher diversity in Nm, lower $\pi_{N} / \pi_{\mathrm{s}}$, more rapid decay in LD and the lower asymptote in LD. However, it is difficult to resolve whether Ng has undergone population size contraction or Nm population size expansion in the past.

It is tempting to attempt to estimate the fraction of substitutions fixed by adaptive evolution as 1-1/FI - see (refs) However, the simultaneous introduction of multiple mutations by hLGT makes this estimate biased.

A central assumption in our analysis is that ClonalFrameML (Didelot \& Wilson 2015) has correctly identified regions of the genome that have undergone hLGT. The method identifies the presence of hLGT from a clustering of mutations along an inferred clonal phylogeny; a sudden burst of mutations along a branch in the phylogeny, that are spatially clustered together in the genome are inferred to be due to hLGT. It is clear therefore that it will be difficult for the method to detect hLGT with relatively similar or short sequences. Furthermore, because we have used a concatenation of protein coding sequences in our ClonalframeML analysis it may be difficult to detect hLGT at the start and end of genes, because we will not have the flanking sequences which provide additional support for hLGT. To investigate whether this latter effect is significant, we plotted the number of inferred hLGT events as a function of the distance from the start or end of genes. We found that events are inferred slightly less often at the start/end of genes, but the effect is not large (Figure 5).

The fact that ClonalFrameML has probably missed some hLGT events suggests that we may have underestimated the input of variation from hLGT in both species - i.e. we have underestimated $R / \theta$. However, an inability to correctly detect all hLGT events is unlikely to explain the differences in the relative contribution of hLGT and mutation in the two species, since both species have been treated identically. An inability to detect hLGT may however explain why $\pi_{N}$ and $\pi_{\mathrm{S}}$ are correlated even in the parts of the genome with no apparent hLGT and hence there may be little or no variation in $N_{e}$ across the genomes of Nm and Ng . Furthermore, it is possible that all the variation that is shared between Nm and Ng is a consequence of hLGT and we have not been able to identify all hLGT events.

## References

Bazin E, Glemin S, and Galtier N. 2006. Population size does not influence mitochondrial genetic diversity in animals. Science 312:570-572.
Begun DJ, and Aquadro CF. 1992. Levels of naturally occuring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature 356:519-520.

Bennett JS, Jolley KA, Earle SG, Corton C, Bentley SD, Parkhill J, and Maiden MC. 2012. A genomic approach to bacterial taxonomy: an examination and proposed reclassification of species within the genus Neisseria. Microbiology 158:1570-1580. 10.1099/mic.0.056077-0

Bennett JS, Jolley KA, Sparling PF, Saunders NJ, Hart CA, Feavers IM, and Maiden MC. 2007. Species status of Neisseria gonorrhoeae: evolutionary and epidemiological inferences from multilocus sequence typing. BMC Biol 5:35. 10.1186/1741-7007-5-35
Brandvain Y, and Wright SI. 2016. The Limits of Natural Selection in a Nonequilibrium World. Trends Genet 32:201-210. 10.1016/j.tig.2016.01.004
Budroni S, Siena E, Hotopp JCD, Seib KL, Serruto D, Nofroni C, Comanducci M, Riley DR, Daugherty SC, Angiuoli SV, Covacci A, Pizza M, Rappuoli R, Moxon ER, Tettelin H, and Medini D. 2011. Neisseria meningitidis is structured in clades associated with restriction modification systems that modulate homologous recombination. Proceedings of the National Academy of Sciences of the United States of America 108:4494-4499. 10.1073/pnas. 1019751108

Castellano D, James J, and Eyre-Walker A. 2018. Nearly neutral evolution across the Drosophila melanogaster genome. Mol Biol Evol in press.
Charlesworth B, Morgan MT, and Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289-1303.
Charlesworth J, and Eyre-Walker A. 2008. The McDonald-Kreitman test and slightly deleterious mutations. Mol Biol Evol in press.
Chen J, Glemin S, and Lascoux M. 2017. Genetic Diversity and the Efficacy of Purifying Selection across Plant and Animal Species. Mol Biol Evol 34:1417-1428. 10.1093/molbev/msx088
Chen X, and Zhang J. 2013. No gene-specific optimization of mutation rate in Escherichia coli. Mol Biol Evol 30:1559-1562. 10.1093/molbev/mst060
Claus H, Maiden MC, Wilson DJ, McCarthy ND, Jolley KA, Urwin R, Hessler F, Frosch M, and Vogel U. 2005. Genetic analysis of meningococci carried by children and young adults. J Infect Dis 191:1263-1271. 10.1086/428590
Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, and de Hoon MJL. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422-1423. 10.1093/bioinformatics/btp163

De Silva D, Peters J, Cole K, Cole MJ, Cresswell F, Dean G, Dave J, Thomas DR, Foster K, Waldram A, Wilson DJ, Didelot X, Grad YH, Crook DW, Peto TEA, Walker AS, Paul J, and Eyre DW. 2016. Whole-genome sequencing to determine transmission of Neisseria gonorrhoeae: an observational study. Lancet Infectious Diseases 16:1295-1303. 10.1016/S1473-3099(16)30157-8
Didelot X, and Wilson DJ. 2015. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. Plos Computational Biology 11. ARTN e1004041
10.1371/journal.pcbi. 1004041

Do R, Balick D, Li H, Adzhubei I, Sunyaev S, and Reich D. 2015. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. Nat Genet 47:126-131. 10.1038/ng. 3186

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797. 10.1093/nar/gkh340
32/5/1792 [pii]
Eyre-Walker A. 2002. Changing effective population size and the McDonald-Kreitman test. Genetics 162:2017-2024.
Ezewudo MN, Joseph SJ, Castillo-Ramirez S, Dean D, Del Rio C, Didelot X, Dillon JA, Selden RF, Shafer WM, Turingan RS, Unemo M, and Read TD. 2015. Population structure of Neisseria gonorrhoeae based on whole genome data and its relationship with antibiotic resistance. PeerJ 3:e806. 10.7717/peerj. 806
Fay J, Wycoff GJ, and Wu C-I. 2001. Positive and negative selection on the human genome. Genetics 158:1227-1234.
Feil EJ, Holmes EC, Bessen DE, Chan MS, Day NP, Enright MC, Goldstein R, Hood DW, Kalia A, Moore CE, Zhou J, and Spratt BG. 2001. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc Natl Acad Sci U S A 98:182-187. 10.1073/pnas.98.1.182
Galtier N. 2016. Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. PLoS Genet 12:e1005774. 10.1371/journal.pgen. 1005774
Gojobori J, Tang H, Akey JM, and Wu CI. 2007. Adaptive evolution in humans revealed by the negative correlation between polymorphism and fixation pahses of evolution. Proc Natl Acad Sci USA 104:3907-3912.
Gordo I, and Dionisio F. 2005. Nonequilibrium model for estimating parameters of deleterious mutations. Phys Rev E Stat Nonlin Soft Matter Phys 71:031907. 10.1103/PhysRevE.71.031907

Gossmann TI, Woolfit M, and Eyre-Walker A. 2011. Quantifying the variation in the effective population size within a genome. Genetics 189:1389-1402.
10.1534/genetics.111.132654

Grad YH, Harris SR, Kirkcaldy RD, Green AG, Marks DS, Bentley SD, Trees D, and Lipsitch M. 2016. Genomic Epidemiology of Gonococcal Resistance to Extended-Spectrum Cephalosporins, Macrolides, and Fluoroquinolones in the United States, 2000-2013. Journal of Infectious Diseases 214:1579-1587. 10.1093/infdis/jiw420
Hao W, Ma JH, Warren K, Tsang RS, Low DE, Jamieson FB, and Alexander DC. 2011. Extensive genomic variation within clonal complexes of Neisseria meningitidis. Genome Biol Evol 3:1406-1418. 10.1093/gbe/evr119
Hill WG, and Robertson A. 1968. Linkage disequilibrium in finite populations. Theoret Appl Genet 38:226-231.
James J, Castellano D, and Eyre-Walker A. 2017. DNA sequence diversity and the efficiency of natural selection in animal mitochondrial DNA. Heredity (Edinb) 118:88-95. 10.1038/hdy. 2016.108

Joseph B, Schwarz RF, Linke B, Blom J, Becker A, Claus H, Goesmann A, Frosch M, Muller T, Vogel U, and Schoen C. 2011. Virulence Evolution of the Human Pathogen Neisseria meningitidis by Recombination in the Core and Accessory Genome. PLoS ONE 6. ARTN e18441
10.1371/journal.pone. 0018441

Kong Y, Ma JH, Warren K, Tsang RS, Low DE, Jamieson FB, Alexander DC, and Hao W. 2013. Homologous recombination drives both sequence diversity and gene content variation in Neisseria meningitidis. Genome Biol Evol 5:1611-1627. 10.1093/gbe/evt116
Lee RS, Seemann T, Heffernan H, Kwong JC, da Silva AG, Carter GP, Woodhouse R, Dyet KH, Bulach DM, Stinear TP, Howden BP, and Williamson DA. 2018. Genomic epidemiology and antimicrobial resistance of Neisseria gonorrhoeae in New Zealand. Journal of Antimicrobial Chemotherapy 73:353-364. 10.1093/jac/dkx405
Leffler EM, Bullaughey K, Matute DR, Meyer WK, Segurel L, Venkat A, Andolfatto P, and Przeworski M. 2012. Revisiting an old riddle: what determines genetic diversity levels within species? PLoS biology 10:e1001388. 10.1371/journal.pbio. 1001388
Lewontin RC. 1974. The genetic basis of evolutionary change. New York: Columbia University Press.
Maddamsetti R, Hatcher PJ, Cruveiller S, Medigue C, Barrick JE, and Lenski RE. 2015. Synonymous Genetic Variation in Natural Isolates of Escherichia coli Does Not Predict Where Synonymous Substitutions Occur in a Long-Term Experiment. Mol Biol Evol 32:2897-2904. 10.1093/molbev/msv161
Martincorena I, and Luscombe NM. 2013. Non-random mutation: The evolution of targeted hypermutation and hypomutation. BioEssays : news and reviews in molecular, cellular and developmental biology 35:123-130. 10.1002/bies. 201200150
Martincorena I, Seshasayee AS, and Luscombe NM. 2012. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. Nature 485:95-98. 10.1038/nature10995

Maynard Smith J, and Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet Res 23:23-35.
McDonald JH, and Kreitman M. 1991. Adaptive evolution at the Adh locus in Drosophila. Nature 351:652-654.
Murray GGR, Soares AER, Novak BJ, Schaefer NK, Cahill JA, Baker AJ, Demboski JR, Doll A, Da Fonseca RR, Fulton TL, Gilbert MTP, Heintzman PD, Letts B, McIntosh G, O'Connell BL, Peck M, Pipes ML, Rice ES, Santos KM, Sohrweide AG, Vohr SH, Corbett-Detig RB, Green RE, and Shapiro B. 2017. Natural selection shaped the rise and fall of passenger pigeon genomic diversity. Science 358:951-954. 10.1126/science.aao0960
Roberts L. 2008. An ill wind, bringing meningitis. Science 320:1710-1715.
Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, Chenuil A, Chiari Y, Dernat R, Duret L, Faivre N, Loire E, Lourenco JM, Nabholz B, Roux C, Tsagkogeorga G, Weber AAT, Weinert LA, Belkhir K, Bierne N, Glemin S, and Galtier N. 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature 515:261-U243. 10.1038/nature13685

Sattath S, Elyashiv E, Kolodny O, Rinott Y, and Sella G. 2011. Pervasive Adaptive Protein Evolution Apparent in Diversity Patterns around Amino Acid Substitutions in Drosophila simulans. PLoS Genetics 7. ARTN e1001302
10.1371/journal.pgen. 1001302

Spratt BG. 1988. Hybrid penicillin-binding proteins in penicillin-resistant strains of Neisseria gonorrhoeae. Nature 332:173-176. 10.1038/332173a0

Spratt BG, Zhang QY, Jones DM, Hutchison A, Brannigan JA, and Dowson CG. 1989. Recruitment of a penicillin-binding protein gene from Neisseria flavescens during the emergence of penicillin resistance in Neisseria meningitidis. Proc Natl Acad Sci U S A 86:8988-8992. Vazquez JA, Delafuente L, Berron S, Orourke M, Smith NH, Zhou JJ, and Spratt BG. 1993. Ecological Separation and Genetic Isolation of Neisseria-Gonorrhoeae and NeisseriaMeningitidis. Current Biology 3:567-572. Doi 10.1016/0960-9822(93)90001-5
Vos M, and Didelot X. 2009. A comparison of homologous recombination rates in bacteria and archaea. ISME J 3:199-208. ismej200893 [pii]
10.1038/ismej. 2008.93

WHO. 2012. Global action plan to control the spread and impact of antimicrobial resistance in Neisseria gonorrhoeae.
Yazdankhah SP, Kriz P, Tzanakaki G, Kremastinou J, Kalmusova J, Musilek M, Alvestad T, Jolley KA, Wilson DJ, McCarthy ND, Caugant DA, and Maiden MC. 2004. Distribution of serogroups and genotypes among disease-associated and carried isolates of Neisseria meningitidis from the Czech Republic, Greece, and Norway. J Clin Microbiol 42:51465153. 10.1128/JCM.42.11.5146-5153.2004

Yu D, Jin Y, Yin ZQ, Ren HG, Zhou W, Liang L, and Yue JJ. 2014. A Genome-Wide Identification of Genes Undergoing Recombination and Positive Selection in Neisseria. Biomed Research International. Artn 815672
10.1155/2014/815672

## Table $\mathbf{1}_{\text {(on next page) }}$

Nucleotide diversity estimates across all sites in the core genome (pi) and at 0-fold nonsynonymous sites ( $\mathrm{pi}_{\mathrm{N}}$ ) and 4 -fold synonymous sites ( $\mathrm{pi}_{\mathrm{s}}$ ).

1

|  | $\pi$ | $\pi_{\mathrm{S}}$ | $\pi_{\mathrm{N}}$ |
| :--- | :--- | :--- | :--- |
| 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 | synonymous sites ( $\pi_{\mathrm{N}}$ ) and 4-fold synonymous sites $\left(\pi_{\mathrm{s}}\right)$.

## Table 2 (on next page)

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)

1

| Species | $R / \theta$ | $\delta$ | $v(\%)$ | $r / m$ | $\theta\left(\times 10^{-3}\right)$ | $R\left(\times 10^{-4}\right)$ |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| Ng | $0.41(0.39$, <br> $0.43)$ | $70(67,72)$ | $6.9(6.8$, <br> $7.1)$ | $2.0(1.8$, <br> $2.2)$ | $1.0(0.8,1.2)$ | $4.0(3.0,5.0)$ |
| Nm | $1.2(1.2$, <br> $1.3)$ | $99(98$, <br> $100)$ | $5.3(5.3$, <br> $5.4)$ | $6.4(6.2$, <br> $6.7)$ | $3.0(2.5,3.5)$ | $36(30,44)$ |
| Ratio <br> $(\mathrm{Nm} / \mathrm{Ng})$ | 3.0 | 1.4 | 0.77 | 3.2 | 3.0 | 9.3 |

4 Table 2. Recombination rate estimates obtained from ClonalFrameML along with their 95\%

confidence intervals.

Figure $\mathbf{1}_{\text {(on next page) }}$
Linkage disequilibrium 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.


Figure 2 (on next page)
Variation in diversity across the genome.

The 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.



## Figure $\mathbf{3}_{\text {(on next page) }}$

Variation in diversity across the non-hLGT fraction of the genome.

The 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.


## Figure 4 (on next page)

Testing for adaptive evolution.

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 .

B)


Figure 5(on next page)
No strong bias in the detection of hLGT

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.


