Uncovering oral *Neisseria* tropism and persistence using metagenomic sequencing

Claudio Donati[1]*, Moreno Zolfo[2], Davide Albanese[1], Duy Tin Truong[2], Francesco Asnicar[2], Valerio Iebba[3], Duccio Cavalieri[4,5], Olivier Jousson[2], Carlotta De Filippo[5], Curtis Huttenhower[6,7] and Nicola Segata[2]*

Microbial epidemiology and population genomics have previously been carried out near-exclusively for organisms grown *in vitro*. Metagenomics helps to overcome this limitation, but it is still challenging to achieve strain-level characterization of microorganisms from culture-independent data with sufficient resolution for epidemiological modelling. Here, we have developed multiple complementary approaches that can be combined to profile and track individual microbial strains. To specifically profile highly recombinant neisseriae from oral metagenomes, we integrated four metagenomic analysis techniques: single nucleotide polymorphisms in the clade's core genome, DNA uptake sequence signatures, metagenomic multilocus sequence typing and strain-specific marker genes. We applied these tools to 520 oral metagenomes from the Human Microbiome Project, finding evidence of site tropism and temporal intra-subject strain retention. Although the opportunistic pathogen *Neisseria meningitidis* is enriched for colonization in the throat, *N. flavescens* and *N. subflava* populate the tongue dorsum, and *N. sicca*, *N. mucosa* and *N. elongata* the gingival plaque. The buccal mucosa appeared as an intermediate ecological niche between the plaque and the tongue. The resulting approaches to metagenomic strain profiling are generalizable and can be extended to other organisms and microbiomes across environments.

The oral microbiome is a complex ecological community with key roles in maintaining immune homeostasis[1–3], preventing pathogen invasion and colonization[4,5], and protecting against oral diseases[6,7]. Several members of the healthy oral microbiome are, however, opportunistic pathogens, with strains of highly variable pathogenic potential coexisting within the same species[8–12]. Characterizing the virulence potential of individual strains and their population genomics from sequence information is currently a challenge, in particular when using culture-free metagenomic sequencing. In addition to difficulties in reconstructing strain-specific genomes with metagenomic assembly, many oral bacteria exhibit complex recombination patterns[13], making metagenomic strain characterization and genotype tracking a very challenging task.

Neisseria is a clinically relevant genus that includes related species that are common and usually asymptomatic colonizers of the human oral cavity, where they account for roughly 10% of the bacterial population on the tongue[14]. Most species, including *Neisseria cinerea*, *Neisseria polysaccharea*, *Neisseria lactamica* and *Neisseria sicca* only rarely cause disease[15–17], while *Neisseria meningitidis* (the meningococcus) is one of the leading causes of sepsis and bacterial meningitis in young adults. The majority of cases of invasive disease are due to five hypervirulent lineages[18,19]. Recent epidemiological studies have used multilocus sequence typing (MLST)[18] and high-throughput 16S rRNA sequencing[14] to characterize their epidemiological patterns. Even though 16S rRNA sequencing can achieve high taxonomic resolution in appropriate computational settings[20], it is unable to detect strain specialization due to plasticity factors such as homologous recombination, which is particularly dramatic for *Neisseria* strains. For these reasons, little is known about the frequency, distribution, transmission potential, population structure and tissue tropism of many *Neisseria* species.

Here, we analyse 520 oral communities metagenomically sequenced by the Human Microbiome Project[21,22] to characterize human-associated neisseriae. First, using genomes from *Neisseria* isolates, we identified the conserved core genomic regions and built a map of genetic variability among related neisseriae. This approach can unambiguously distinguish these species in pure culture[23], but it has not been applied so far in metagenomics. Sequence alignment of metagenomes to the conserved core allowed the identification of genetic patterns characteristic of the different strains of neisseriae. This approach was coupled with the application of MLST in an assembly-free metagenomic context for strain-level community-wide profiling, with the use of species-specific markers for distinguishing closely related *Neisseria* strains, and short (12mer) sequence signatures characterizing specific *Neisseria* clades.

We applied these novel computational tools to unravel ecological and biogeographical organismal patterns using metagenomic samples. Although we focus here on *Neisseria* species, our approach is general and can be readily extended to other microbial populations with variable pathogenic potential sharing the same environments.

Results and discussion

Genome-wide phylogenetic analysis of neisseriae identifies a group of closely related species that colonize humans. Several species of neisseriae are known to colonize the mucosa of the oropharynx of healthy individuals[24]. Although 16S rRNA-based taxonomy is unable to distinguish closely related species[25], whole genome-based taxonomy has been shown to clearly classify the

[1]Computational Biology Unit, Research and Innovation Centre, Fondazione Edmund Mach, Via Edmund Mach 1, 38010 San Michele All'adige, Italy. [2]Centre for Integrative Biology, University of Trento, Via Sommarive 9, 38123 Trento, Italy. [3]Department of Public Health and Infectious Diseases, Institute Pasteur Cenci Bolognetti Foundation, Sapienza University of Rome, Piazzale Aldo Moro 5, 00185 Roma, Italy. [4]Department of Biology, University of Florence, Via Madonna del Piano 6, 50019 Sesto Fiorentino, Firenze, Italy. [5]Institute of Biometeorology, National Research Council (IBIMET-CNR), Via Caproni 8, 50145 Firenze, Italy. [6]Biostatistics Department, Harvard School of Public Health, Boston, Massachusetts 02115, USA. [7]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. *e-mail: claudio.donati@fmach.it; nicola.segata@unitn.it

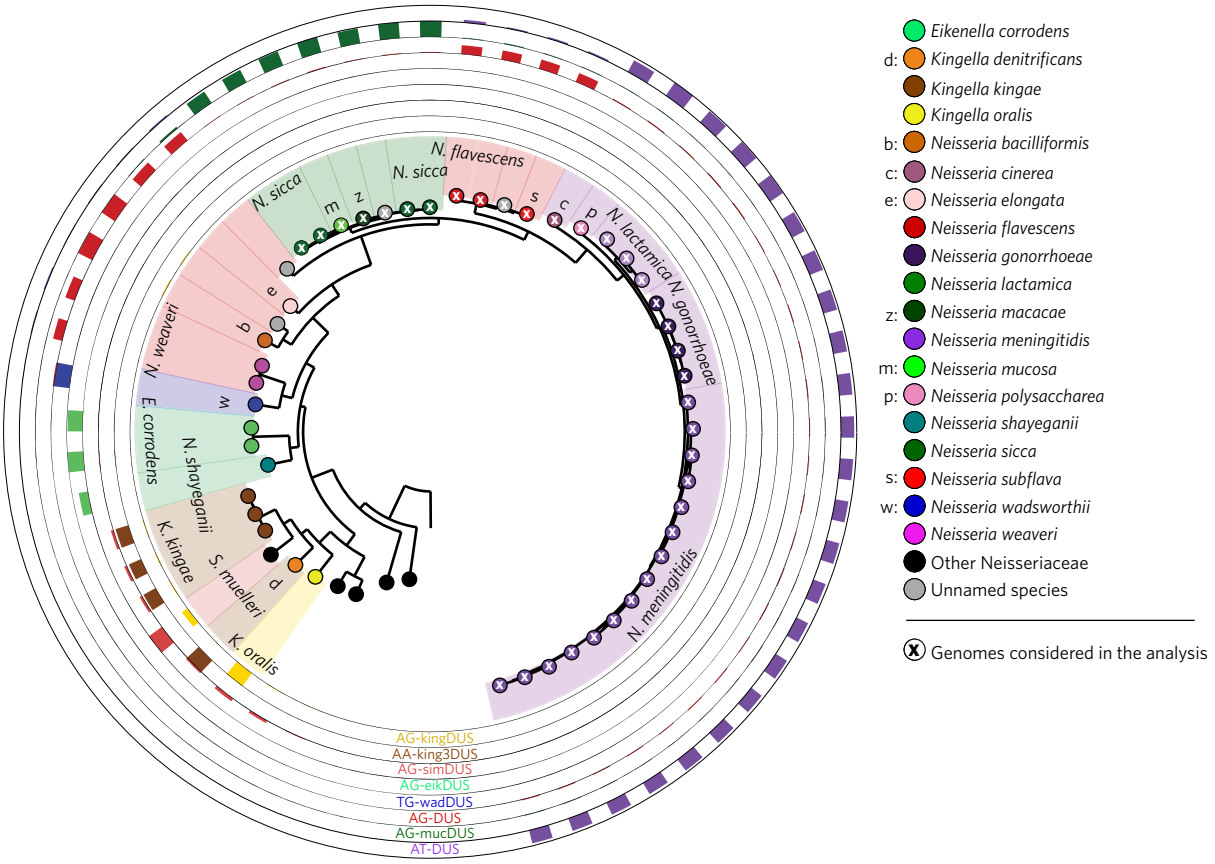

Figure 1 | The phylogenetic structure of the family Neisseriaceae identifies well-defined subtrees of closely related species. The phylogeny is constructed using all available genomes (non-draft genomes only for *N. meningitidis* and *N. gonorrhoeae*) by concatenating and aligning 400 conserved proteins automatically identified in the considered genomes[27]. *Neisseria* strains have been shown to encode in their genomes many copies of short genetic features (12mers called DUSs) that regulate genomic recombination and are subclade- or species-specific. The relative copy numbers of the eight previously identified[31] DUS dialects are reported as external circular bar plots, confirming the univocal association between them and specific subclades of this family. An 'X' in the node indicates a genome used for the core genome-based metagenomic analysis applied for the strain-level characterization. Single letters are used for species names that could not be overlaid to the tree because of lack of space. The rings represent the different DUSs detected. GraPhlAn was used to visualize the tree and associated information with a circular layout[70].

different species[23,26]. We retrieved all the 241 draft and final genomes belonging to the Neisseriaceae family (Supplementary Table 1) and reconstructed their phylogeny using a concatenated alignment of 400 conserved proteins[27]. The resulting phylogeny (rooted using *Chromobacterium violaceum*, a microorganism in the Neisseriales order but not in the Neisseriaceae family, Fig. 1) shows the occurrence of three major clusters of closely related species that are common colonizers of the oral mucosae.

The most basal cluster includes *N. sicca*, *N. mucosa* and *N. macacae*, the second cluster with intermediate branching includes *N. flavescens* and *N. subflava*, and the third most derived cluster includes *N. cinerea*, *N. polysaccharea*, *N. lactamica* and the opportunistic pathogen *N. meningitidis*. The latter cluster also includes *N. gonorrhoeae*, which is known to be closely related to the meningococcus at the genomic level[28]. More basally branching is a monophyletic clade including *N. elongata* and *N. bacilliformis*, which are also known to colonize humans. Overall, our phylogeny is consistent with other studies[23], although the order of some rather deep branches can still be further investigated[29]. However, the tree's most external clades (from *N. sicca* to *N. meningitidis*) have high statistical support and are highly consistent with other studies. For this reason, and because oral neisseriae are almost exclusively in this subtree, we restrict our metagenomic study to this set of strains (see Methods).

Other species such as *Kingella* and *Eikenella* have been identified to be related to the human neisseriae at the genus or family level and can sporadically be found in humans. The genomes of these groups are characterized by specific forms of the DNA uptake sequences (DUSs), defined as 12 bp sequences that are repeated thousands of times in the genomes of neisseriae (Supplementary Table 2), with higher frequency in the core regions[30], and that regulate the genomic integration of exogenous DNA by transformation[31] (Fig. 1). The available neisseriae spp. genomes and their reconstructed phylogeny constitute the base of our metagenomic strain-level investigation performed on the set of 520 shotgun metagenomic samples from the oral cavity sequenced by the Human Microbiome Project (HMP)[21,22] and its recent follow-up phase.

Genetic characterization of oral metagenomic data sets highlights the tissue tropism of commensal neisseriae. Our approach at strain level profiling is based on (1) identifying a conserved core genome (that is, the portion of the reference genome conserved in all strains) across the species of interest and (2) using this as a reference for identifying strain-specific variants in metagenomic samples. Given that for human neisseriae a sufficiently high number of genomes are available to define a robust core genome, this approach generalizes well for combinations of single nucleotide polymorphisms (SNPs) variants not directly captured by isolate sequencing. Specifically, for step (1) we aligned the complete genome of *N. meningitidis* MC58 against all other oral neisseriae genomes available in public databases (Supplementary

Table 1) and identified the positions in the core genome where at least one of the sequences had a nucleotide substitution. We then aligned the raw sequences of the oral HMP metagenomes against *N. meningitidis* MC58 and recorded the SNPs found in these positions in the metagenomes. In total, 25,202 SNPs were identified in this core genome (Supplementary Figs 1 and 2). Rarefaction analysis shows that this set provides an exhaustive estimate of the SNPs expected in these species (Supplementary Fig. 3). The use of only one genome as a reference dramatically speeds up the computational performances while simultaneously allowing the mapping of the other species given the limited sequence variation within the considered genomes (Supplementary Fig. 4). Using this procedure, we associated each metagenomic sample and reference genome with a list of SNPs that we used to identify the prevalent strains of *Neisseria* present in the former.

By performing a principal component analysis (PCA) of the SNP vectors of the oral metagenomes (Fig. 2), we investigated the intra- and inter-sample diversity of these bacteria among oral subsites of healthy subjects. Two groups of samples are clearly distinct (Fig. 2a), comprising samples from the tongue dorsum and gingival plaque, respectively (permutational multivariate analysis of variance (PERMANOVA) P value of $<1\times10^{-4}$; also Supplementary Fig. 5). A third group is composed of two samples from the throat that are closely related to the genomic sequences of *N. meningitidis*. Co-segregation of individual genomes with strains identified from metagenomes allows the transfer of taxonomic labels, highlighting how strains from different species are closely associated with samples from specific body sites. *N. flavescens* and *N. subflava* segregate with the tongue dorsum samples, while *N. sicca*, *N. macacae*, *N. mucosa* and two other *Neisseria* strains without assigned species names are closely related to gingival plaque samples (Fig. 2a). The buccal mucosa samples spread over a wide region between two clusters, either because they contain microorganisms from unknown species or because the sequences of more than one microorganism contributed to our data set. An additional small group was closely associated with one *N. cinerea* genome sequence, confirming the presence of this species in a reduced subset of oral samples.

The dozens of *N. meningitidis* genomes consistently cluster together with two of the three throat samples in which neisseriae were detected. These two samples derive from the same patient at different time points, and comparing the sequences from these two samples identifies only 41 differences out of 25,202 covered SNPs. This individual has thus been stably colonized by a specific *N. meningitidis* strain, supporting the ability of this procedure to discriminate the presence of closely related strains in metagenomic data sets. MetaPhlAn2 analysis[32] of all the other oral samples further confirmed the presence of *N. meningitidis* only in the throat samples, with no *N. meningitidis* specific markers present in the other available oral cavity locations, even at low abundance. Restricting the PCA analysis only to the mouth microbiome (gingival plaque, buccal mucosa and tongue dorsum, Fig. 2b), we found that the tongue dorsum samples can be further partitioned into two groups, one closely related to *N. flavescens* and *N. mucosa*, and the other to *N. subflava*. Our approach thus reveals a strong oral site tropism of commensal neisseriae. Interestingly, all the three main oral sites have similar abundances of neisseriae (ave/s.d. = 6.5 ± 7.9%, 6.4 ± 6.92% and 3.9 ± 3.92% for tongue dorsum, gingival plaque and buccal mucosa, respectively), suggesting that niche adaptation is occurring at the genetic rather than abundance level. A clustering obtained performing a discriminant analysis of principal component[33] using the Bayes information criterion to determine the optimal number of clusters (Supplementary Fig. 6) further refines this association between metagenomic samples and bacterial species, suggesting the presence of seven distinct clusters. Four of these clusters are associated with a single species (*N. cinerea*, *N. subflava*, *N. flavescens* and one undetermined Neisseria sp.) and

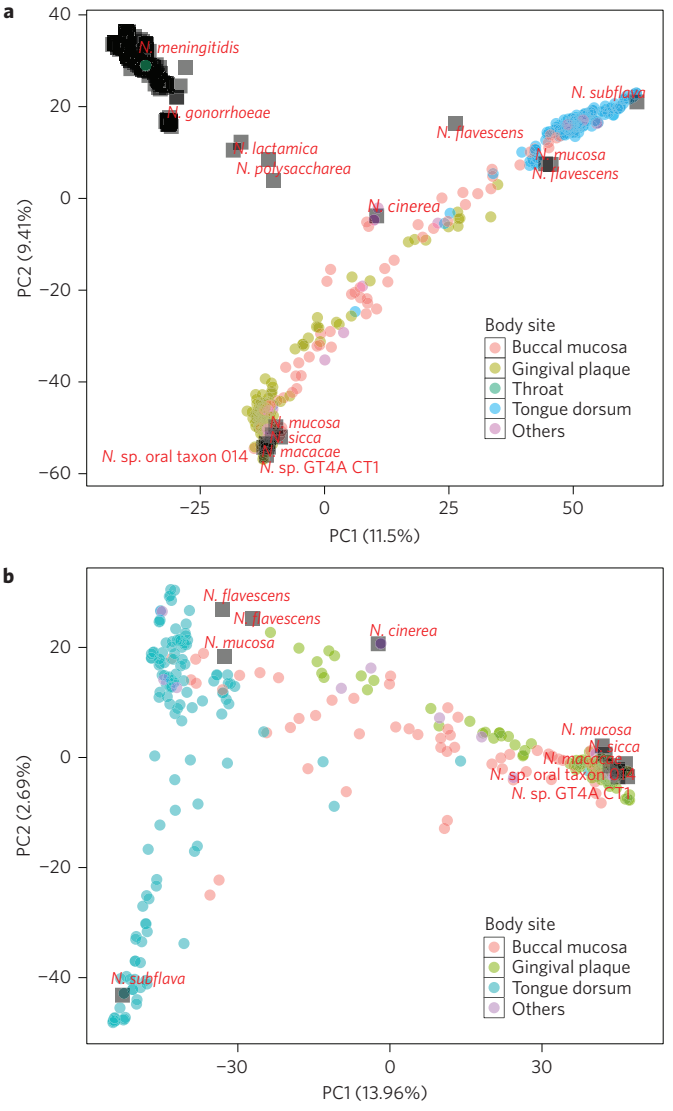

Figure 2 | Biogeographical patterns of oral *Neisseria* strain colonization identified using metagenomic SNP calling. a, SNP sample- and strain-specific patterns are used here to define a genetic distance matrix among *Neisseria* strains in the oral HMP metagenomic samples, which is then projected into two dimensions by PCA. In total, 347 *Neisseria* strains are reported from samples collected from the buccal mucosa (66), the gingival plaque (130), the tongue dorsum (133) and a few (16) other less comprehensively sampled locations (tonsils, throat, keratinized gingiva, hard palate and saliva). Dark squares indicate individual reference genomes. **b**, PCA plot restricted to the 345 samples from the mouth (that is, excluding two samples from the throat colonized by *N. meningitidis*).

were confirmed by MetaPhlAn2 analysis. One other cluster contains the genomes of *N. sicca*, *N. mucosa*, *N. macacae* and the undetermined *Neisseria* oral taxon 014 and *Neisseria* GT4A CT1, and two clusters are not associated with any sequenced Neisseria strains (Supplementary Fig. 7 and Supplementary Table 3).

Validation of the SNPs approach for mixed samples. Microbial communities can contain more than one strain of the same species, and our analysis may be limited by focusing on the single most abundant strain. We thus assessed, first, whether this occurs in real oral communities, and subsequently the effect of strain mixing in simulated data. To estimate the relative abundance of the dominant strain of neisseriae in the samples, for each

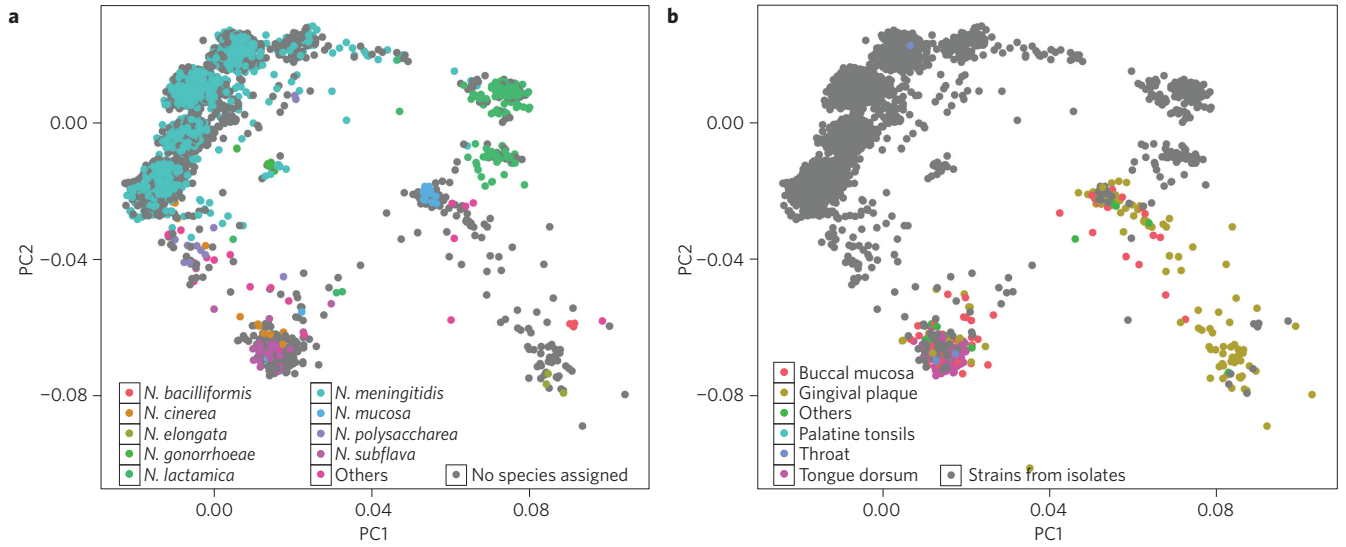

Figure 3 | Metagenomically inferred multilocus sequence types of oral neisseriae from the HMP data set segregate into three main body site clusters. **a**,**b**, PCoA of the concatenated alignment of 11,105 known and newly discovered MLST profiles. In **a**, points are coloured according to the most represented species in the sequence type in the MLST database. Grey points represent strains from isolates that could not be associated to any species. In **b**, points are coloured according to the most frequent body site in metagenomic data. Grey points represent strains from isolates (that is, non-metagenomic samples).

metagenome we computed the average fraction of reads supporting the dominant SNPs in polymorphic position. In all sampling sites, the dominant component contributed on average for more than 70% of the reads, from an average of 74% for samples from the buccal mucosa to an average of 90% in samples from the tongue dorsum (Supplementary Fig. 9).

To assess the impact of complex mixtures, we then generated 50 synthetic metagenomes with simulated Illumina sequencing noise from pairs of neisseriae genomes at four different relative abundances (59–41%, 67–33%, 85–15% and 97–3%) using Grinder[34]. We computed the fractions of these samples' SNPs profiles that were discordant with the corresponding position of the defined dominant and minor allele. We found that the percentage of discordant SNPs with the dominant allele decreased monotonically with increasing relative abundance of the latter from an average of 4.6% (dominant relative abundance of 59%) to an average of 0.22% (dominant relative abundance of 97%). Conversely, the fraction of discordant SNPs with the minority component increased from 10 to 14%, close to the average value (15%) for a randomly chosen genome (Supplementary Fig. 8).

Combining these two analyses, we progressively removed samples from our real metagenome collection in which the average contribution of the dominant strains was below a given threshold (60, 70, 80 and 90%). Using PCA, this provided a clear visualization of the population of communities dominated by strains near references versus those containing non-reference strains or mixtures (Supplementary Fig. 10). For example, most of the tongue dorsum samples were dominated by a single strain, while buccal mucosa communities consisted in some cases of a more complex mixture of strains detectable in this manner.

Short DUSs are easily detectable genetic features that may drive the ecology of oral species. DUSs are short genetic features (12 nucleotides) that regulate interspecies genomic recombination and transformation by limiting DNA exchange between different *Neisseria* species[31]. DUSs are unique to distinct sets of *Neisseria* species (Fig. 1), and by dictating the extent of homologous versus interspecies recombination probably have a key role in their ecology *in vivo*. DUSs also constitute an ideal tool for identifying *Neisseria* clades in unassembled sequences given the consistency of their distribution with respect to phylogenetic analysis (Fig. 1) and their short length.

By simply counting the occurrences of such 12mers in metagenomes, a dominant DUS was found in all samples (Supplementary Table 4). The number of occurrences ranges from an average of 5,600 (that is, 0.045% of reads) to an average of 181,000 (that is, 0.2% of the reads). Overlying the dominant DUS dialect in the SNPs-based PCA clustering of the oral and mouth microbiomes (Supplementary Fig. 11) and observing the DUS site-specific abundances (Supplementary Fig. 12), we obtained further confirmation of the subsite tropism discussed above.

Specifically, following the proposed nomenclature[31], AG-DUS was predominant in tongue samples, confirming that these are rich in *N. flavescens* and *N. subflava*. Gingival plaque samples (Supplementary Fig. 11b) also contain AG-mucDUS and AG-kingDUS, which are highly frequent in *N. macacae*, *N. sicca* and *N. mucosa*, and in *Kingella oralis* and *Simonsiella muelleri*, two species closely related to the oral neisseriae, respectively (Fig. 1). The three oral samples closest to *N. cinerea* in the PCA in Supplementary Fig. 11 are rich in AT-DUS, the most overrepresented repeat in *N. cinerea*[31], confirming the ability of the approach to identify neisseriae in metagenomic data.

MLST alleles reconstruction validates oral tropism. MLST has been widely applied on bacterial isolates[35], and its *in silico* variants can characterize strains from their genome sequences[36]. Interestingly, the same MLST scheme for neisseriae is effective on all known species of the genus. Here, we extend this tool by assembly-free allele reconstruction (see Methods) to type dominant strains in metagenomic samples. We identified 362 total sequence types, 26 of which occurred in more than one sample. Of these 26, 21 were previously unknown types[36].

MLST allelic profiles can link isolates with strict identity of (almost) all of the considered loci, as shown in Supplementary Fig. 13, where oral subsite-specific types are clustering in subbranches of the minimum spanning tree. By performing a principal coordinate analysis (PCoA) of both deposited types and newly identified ones using genetic distance, a discrete clustering structure arises and could be labelled using the available species information associated with the deposited types (Fig. 3a).

By cross-referencing species assignments (Fig. 3a) with oral site information (Fig. 3b) the *Neisseria* tropism is further and independently confirmed, with *N. meningitidis* present in the throat only,

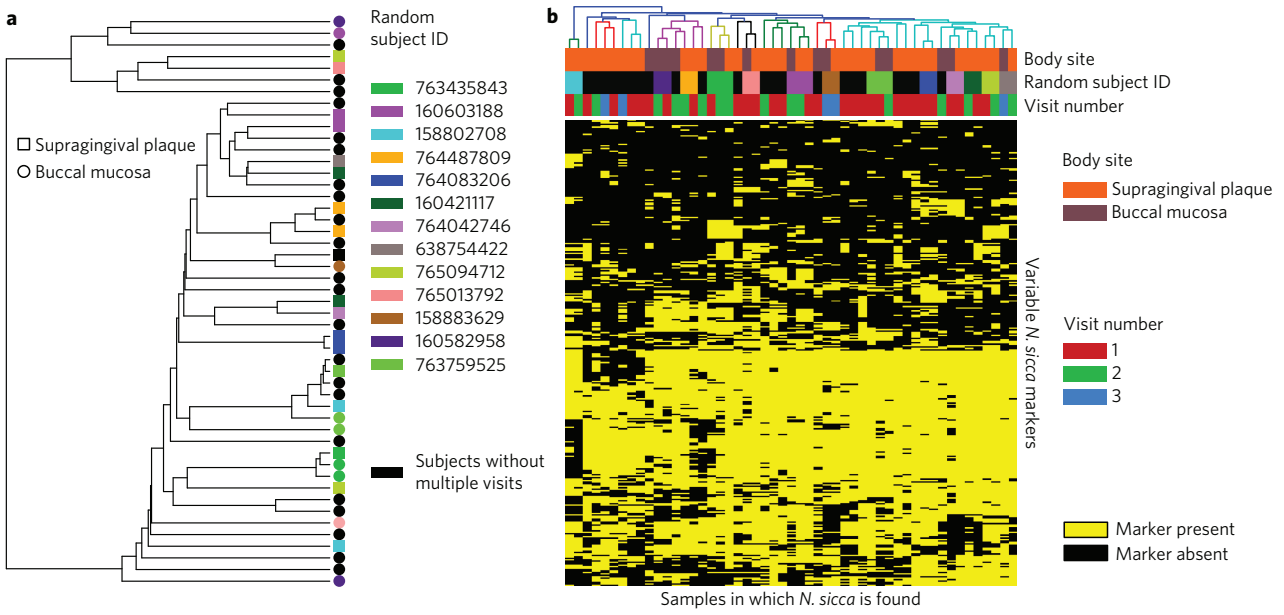

Figure 4 | Characterization of *N. sicca* strain clusters using genetic polymorphisms and genomic variants from multiple subjects and body site metagenomes. *N. sicca* was found in the 49 samples reported in the figure (total number of samples, 520). **a**, Phylogenetic tree of *N. sicca* strains from metagenomes built using the core genome SNP-based approach analysis (Fig. 2). Subjects with more than one sample in which *N. sicca* is detected are shown. **b**, Presence (or absence) in the metagenomes of each MetaPhlAn species-specific marker (see Methods) for *N. sicca*. Both panels share the same colour codes for subject IDs. The resulting patterns can be interpreted as strain-specific barcodes that can be used in an epidemiological framework. The hierarchical clustering induced by the presence/absence pattern of *N. sicca* markers (upper part of **b**) further highlights the clustering of samples in subject-specific strains.

N. subflava predominantly colonizing the tongue dorsum, and *N. mucosa* present on the plaque and the buccal mucosa. Although the MLST database does not include some species, such as *N. sicca* and *N. flavescens*, it instead contains three isolates from the *N. elongata* species that were not included in the core-genome-based genetic analysis, as the species was too divergent from the others. Several dozens of gingival plaque strains cluster around *N. elongata*, thus further refining the subsite-specialization of the *Neisseria* genus (Fig. 3b).

Genomic markers can track *Neisseria* strains across samples. To complement genetic and DUS-based profiling with genome-wide characterization, we employed unique combinations of species-specific markers (as described elsewere[37]) to assess the diversity of oral neisseriae below the level of species. Four species showed strong evidence (>33% of detected species-specific markers) of their presence in more than 50 samples (*N. flavescens*, 118; *N. elongata*, 70; *N. subflava*, 60; *N. sicca*, 52), and this allowed the taxonomic labelling of the samples in the genetic clustering to be improved (Supplementary Fig. 14). This analysis also detected other *Neisseria* species, including a sequenced oral isolate (*Neisseria* oral taxon 014), *K. oralis* (36 samples), *Kingella denitrificans* (six samples) and *Eikenella corrodens* (14 samples). The marker-based barcoding profiling further refined the resolution of the analysis. For example, strong clustering for *N. sicca* was detected, with strains diverging for very few markers belonging to samples from the same subject (at different time points or oral cavity locations, Fig. 4b), providing a robust strain-tracking mechanism for all the analysed species (Supplementary Figs 15–17 report all *Neisseria* species found in at least 15 samples).

Genetic and genomic analyses highlight stable inter-subject and longitudinal Neisseriaceae strain colonization. Despite many cross-sectional studies documenting the frequency of carriage of neisseriae, much less is known concerning the temporal stability of colonization by single strains. Again, longitudinal studies have concentrated on meningococcus, for which the same strain was shown to persist in the same subjects for several weeks or months[38,39]. Given the availability of multiple samples from the same patients taken at different collection visits (average time between visits of 219 days, s.d. of 69 days), we used the SNPs and genetic marker data to measure the propensity of carriers to maintain the same strains of *Neisseria* for extended periods of time.

We computed the number of mutations between the strains from distinct metagenomes, comparing samples drawn from different subjects with samples taken from the same subject and body site at different times. As shown in Fig. 5a, the distribution of the distances computed between samples taken from the same patients has a peak close to zero that is absent when uncorrelated samples are compared. In particular, 23% of those samples differ in less than 1% of the SNPs, corresponding to an estimated genetic distance between the *Neisseria* strains contained in the samples of 0.058 mutations per 100 bases. For comparison, from whole genome data, we estimated that the average distance between two uncorrelated strains of *N. meningitidis* is 3.18 mutations per 100 bases (Supplementary Fig. 4), in accordance with previous data[40], suggesting that in those samples the same strain persists between different visits. The percentage of samples within this cutoff reduces to 0.13% when different individuals are compared (Fig. 5b). On splitting the samples by location within the oral cavity (Supplementary Fig. 18), we find that strain persistence is especially high in tongue dorsum samples, and that numbers are too small to draw conclusions from other body sites.

The corresponding analysis using genomic (that is, unique markers) rather than genetic information (that is, core SNPs) confirms subject-specific strain persistence (Fig. 5c,d), which is consistent for all species. The only partial exception is *N. flavescens*, which shows a high degree of variability in the genomic markers. Together, these complementary approaches to track strains across samples

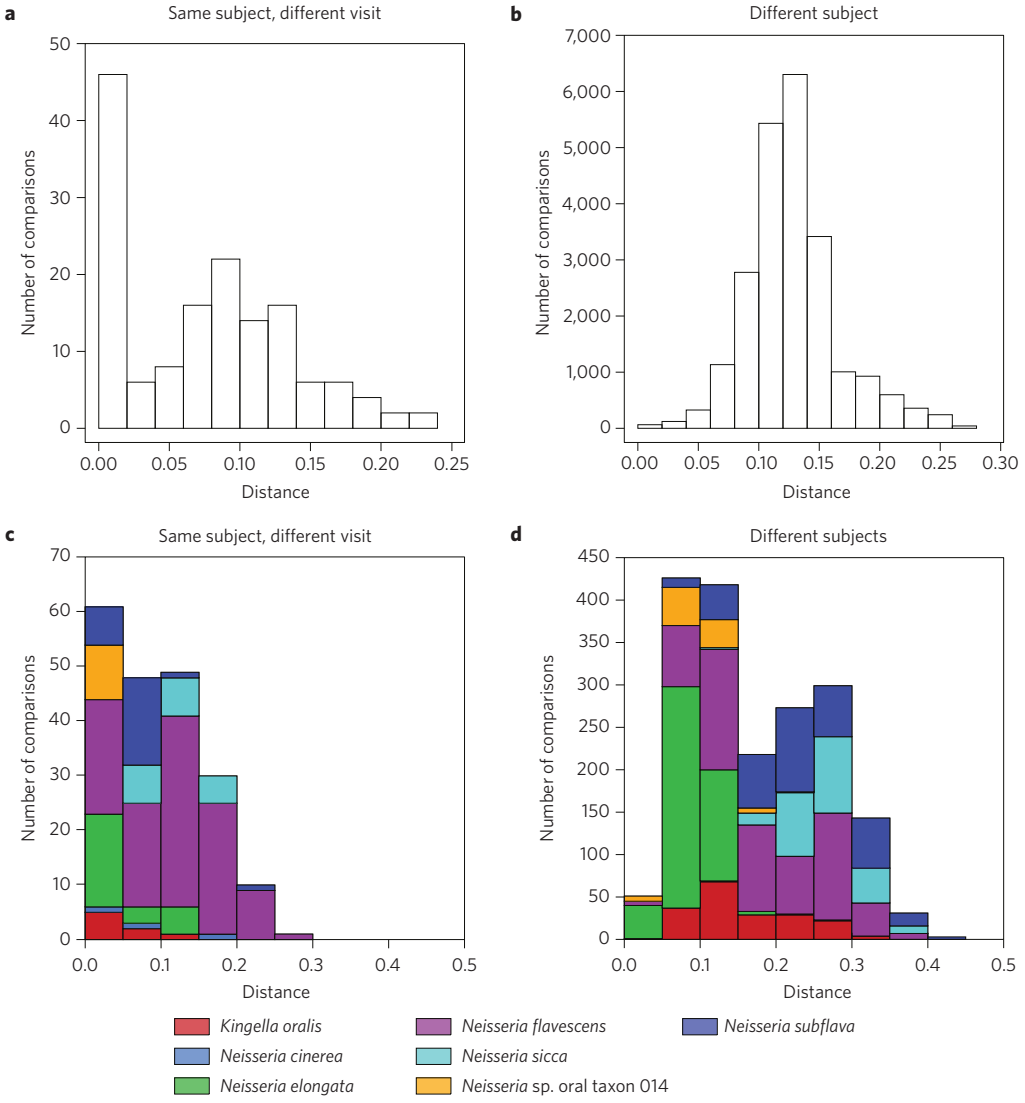

Figure 5 | Persistent subject-specific, site-specific and strain-specific colonization by neisseriae identified using longitudinal metagenomic core genome SNP and genomic marker analyses. a, Distribution of the fraction of discordant SNPs in 148 pairwise comparisons between samples from the same subject, same body site and different visit. **b**, As in **a**, but for 22,748 pairwise comparisons between samples from different patients and body sites. **c**, Distribution of the fraction of discordant markers between samples from the same subject, same body site and different visits. **d**, As in **c**, but between samples from different subjects and body sites.

provide a reliable means to study the stability of microorganisms associated with the human host.

Conclusions

Using hundreds of metagenomic samples spanning 242 subjects in multiple oral sites, we have shown that the genomic variability of microbial species can be exploited to identify the presence of microorganisms of interest with strain-level resolution. Applying multiple novel metagenomic analysis approaches tools (see Methods and a graphical summary in Supplementary Fig. 19), we found that different species of *Neisseria* colonize different sites within the oral cavity and oropharynx of the same subjects, showing close associations between species and tissues, which represent unique microbial ecological niches with specific physico-chemical conditions. The association between species and tissue was confirmed by measuring the frequency of DUS sequences[30] and by reconstructing the MLST profiles of the dominant clones in the samples. Taken together, these results suggest that the diversification of this group of species has been driven by clonal expansion accompanied by accumulation of adaptive mutations and genomic modifications specific to the microenvironments in the oral cavity.

The developed methods allowed us to trace the presence of the same strains for extended periods in longitudinal data sets. Persistence of SNPs patterns and markers demonstrated that, in most cases, the same individual and the same body site have been colonized by one specific strain over the time spanned by our data set. This result provides both an internal consistency check of our approach, confirming that strain-level resolution can be attained and exploited for strain tracking across samples, as well as novel estimates about the duration of colonization of human neisseriae and their turnover time.

The presented tools are a first step toward a culture-independent framework for analysing targeted members of the human microbiome for comparative and population genomics. These methods provide one of the first means to investigate microbiome strains with a resolution comparable to what is routinely achieved with single-isolate sequencing. Microbial community analyses based on existing SNP-targeted approaches[41], for example, are not taxon-specific, and instead profile the overall genetic variability of the

microbiome, and it is challenging to achieve sufficient coverage of target organisms using assembly-based approaches[42–44]. For neisseriae, in particular, binning of contigs[45,46] is especially difficult because of the high recombination rate. As a result, despite notable results from metagenomic assembly[47], this approach is usually targeted at discovering new members of the microbiome rather than performing comparative cross-sectional analyses (with only a few exceptions[48]). Other strain-level tools[32,49–51] have been developed for the detection of organisms, but they lack the ability to reconstruct their full sequence or have never been applied to oral samples[51]. The methods developed here are designed to exploit several complementary features of the microbial community, thus capturing both whole-population and taxon-specific information and minimizing biases associated with any single approach.

These analysis strategies are based on a number of assumptions. For metagenome analysis using genetic variation, by selecting only the dominant SNP in each position, we implicitly assumed that a single strain within the target species is dominant in each sample. Previous analysis based on the presence of non-core genomic markers has shown that this assumption does not hold true in all cases[52]. To overcome this limitation, our approach could be generalized using probability estimates of the dominant and alternative alleles. Another assumption is that for each target species, a number of genome sequences commensurate with its phenotypic variability and the complexity of its population structure are available. It is known that the genomic variability within species can be large[8,9,11,12], and that this variability determines the existence of strains with a wide range of pathogenic or metabolic potentials coexisting in the same species (for example, pathogenic and commensal *Escherichia coli*[53,54]). Although strain-to-strain variability has been extensively studied in the context of pathogen genomics, it remains confined to isolate sequencing that is biased towards cultivable and pathogenic organisms. One classical example of what we might miss is the enhanced risk of gastric cancer in individuals colonized by strains of *Helicobacter pylori* expressing the cag pathogenicity island[55]. This island plays a central role in gastric carcinogenesis, encoding a Type IV secretion system that delivers the oncoprotein CagA into the cytoplasm of host cells and is present in the genome of some, but not all, the circulating strains of *H. pylori*[56]. Regional differences in the distribution of cag-positive and cag-negative strains might contribute to the diversity of the epidemiology of gastric cancer[57], but these differences would not be revealed by metagenomic studies unless a sufficient number of reference genomes were available. Increasing the number and quality of sequenced microbes will greatly improve our ability to analyse metagenomic data.

Using culture-free methods for the strain-level tracking of commensal organisms will be a first step towards mapping their role and interactions in the microbiome. For neisseriae, being able to monitor their distribution at the species and strain level could, for instance, shed more light on the development of natural immunity[58,59] and the impact of containment strategies based on vaccination[23,60]. It could also help identify factors that might predispose to disease[61] and provide a means for early detection of the clonal waves that often precede disease outbreaks[62]. Expanding the analysis to other species and data sets will be instrumental in exploiting fully the wealth of information provided by current and upcoming sequencing technologies and to define the role of individual strains in the context of the colonizing microbiome.

Methods

Genome-wide phylogenetic analysis of neisseriae. All 241 available reference genomes belonging to the Neisseriaceae family as of 27 July 2014 were downloaded from RefSeq and NCBI[63]. For each genome, the automatic downloading procedure retrieved the genomic sequence, the coding gene repertoire and the protein sequences. The proteomes were fed to PhyloPhlAn (ref. 27) to efficiently reconstruct a whole-genome phylogeny of the Neisseriaceae family using *Chromobacterium violaceum* ATCC 12472 (the Chromobacteriaceae family according to the NCBI taxonomy) as outgroup. In brief, PhyloPhlAn identifies homologues of the 400 most universally conserved bacterial proteins, which are then extracted, aligned, concatenated and used to build the phylogenetic tree reported in Fig. 1a. All the genomes from *N. meningitidis* and *N. gonorrhoeae* (186 and 19, respectively) were employed in a first version of the tree. However, given that for both species strictly monophyletic and low-diversity clades were reconstructed, we decided to improve the readability of the tree by using, in the final version shown in Fig. 1, only those genomes from the two species without gaps in the sequence (15 for *N. meningitidis* and 4 for *N. gonorrhoeae*). As also discussed elsewhere[31], our analysis confirms that the strain labelled as *N. mucosa* C102 is confidently rooted inside the *N. flavescens*/*N. subflava* subtree and is not related to the *N. mucosa* ATCC 25996 strain. Given the additional available evidence based on the 16S rRNA gene[31] and the analysis of the genome of the ATCC 25996 strain[26], we thus considered the C102 strain as a *Neisseria* genome without confident species assignment (Fig. 1).

Description of the metagenomic data set from the oral cavity used in the study. We considered here the shotgun metagenomic samples from the oral cavity produced by the HMP[21,22]. The samples from the first wave of HMP sequencing included samples from 242 healthy subjects, with 107 samples from the buccal mucosa, 1 from the hard palate, 6 from the gingiva and the tonsils, 5 from the saliva, 7 from the subgingival plaque, 115 from the supragingival plaque, 7 from the throat and 122 from the tongue. We also considered the samples produced in a second wave of sequencing from the HMP (available at http://www.hmpdacc.org/HMIWGS/healthy/) for a total of 520 oral metagenomes. Up to three time points (average temporal separation of 219 days) are available for each subject. The numbers of samples from each body site and each visit are reported in Supplementary Table 5.

DUS copy number detection from genomes and metagenomes. The eight 12-mer DUSs previously identified from 23 Neisseriaceae genomes[31] were used as templates for screening all the 241 genomes now available. To account for potentially unknown DUSs, the screening pipeline also quantified the copy numbers of all 12mers with a single nucleotide mismatch from one of the eight template DUSs. The eight most abundant DUSs (accordingly to the maximum value across all genomes) were exactly those previously identified[31], and the downstream analysis thus considered only these. The same counting pipeline was applied to all the metagenomes by directly processing the FASTQ files, finding up to 1,320,232 occurrences of a DUS (specifically AG-DUS) in a metagenome (specifically SRS017007 from the tongue dorsum). It is worth noticing that the total number of DUSs identified in metagenomes is underestimated by ~10%, given that only 88% of the 100-nt-long reads are accessible by full-length matching of 12mer sequences.

Genome alignments and SNPs identification. We used Nucmer[64] to compute the pairwise genome alignments of the genome sequence of *N. meningitidis* strain MC58 against all sequences of *N. cinerea*, *N. flavescens*, *N. gonorrhoeae*, *N. lactamica*, *N. macacae*, *N. meningitidis*, *N. mucosa*, *N. polysaccharea*, *N. sicca* and *N. subflava*. Two Neisseria isolates identified as *Neisseria* sp. oral taxon 014 and *Neisseria* sp. oral taxon 020, which were closely related to oral human neisseriae (Supplementary Fig. 1 and Supplementary Table 1), were also included in the alignment phase. The alignments were post-processed using tools of the MUMmer software suite to identify the SNPs in the pairwise alignments and the coordinates of the aligned regions. Specifically, for each pairwise alignment, SNPs were identified using the command 'show-snps' (parameters: –ClrHI). Similarly, for each pairwise alignment the coordinates of the aligned regions were determined using the command 'show-coords' (parameters: –B). These data were collected using custom scripts to identify the coordinates on the reference genome of those regions that could be aligned univocally against all other genomes (core genome of 245,993 total bases, of which 96,495 aligned univocally in all sequences, Supplementary Figs 1 and 2). Within these regions we identified the coordinates of the positions that were variable in at least one pairwise alignment. Finally, by parsing the pairwise alignments, for each genome we extracted the alleles corresponding to the variable position, building a matrix of SNPs that was used in downstream analysis.

Schematically, the algorithm can be decomposed into the following steps. For each genome: (1) carry out pairwise alignment against the reference genome using nucmer; (2) identify the aligned regions using show-coords; and (3) identify variable positions (SNPs) using show-snps. Subsequently, the core genome is determined and a SNPs matrix is computed based on the upstream data.

To assess the completeness of the SNP data set and its ability to provide an exhaustive compendium of the SNPs that are to be expected in this group of organisms, we performed a rarefaction analysis by sampling a growing number of genomes and counting the number of sites that were polymorphic for each of the samples. We found that the number of SNPs grows logarithmically with increasing number of genomes (Supplementary Fig. 3), similarly to what has been found previously in other bacteria[65]. This result was consistent with the predicted number of SNPs in a single homogeneous population in a model based on coalescent theory[66], where each new genome contributes a number of novel SNPs proportional to the time to the last common ancestor with an individual already in the sample[65]. In this model, the number of new SNPs introduced by new sequences drops rapidly

with increasing number of genomes, in proportion to 1/*N*, where *N* is the number of genomes already sequenced. From the logarithmic fit to the data we extrapolated that, with 209 genomes already in the data set, one additional genome would contribute only 23 new SNPs, representing less than 0.1% of SNPs and suggesting that the size of the data set is sufficient to attain strain-level resolution in this group of organisms.

Metagenome alignments and SNPs identification. We aligned the raw sequences of the WGS metagenomic data set to the reference complete genome of the *N. meningitidis* MC58 strain using the BowTie2 sequence aligner (local alignment, -D 20 -R 3 -N 0 -L 20 -i S,1,0.50)[67] to also align reads from more distantly related genomes (Supplementary Fig. 4). Although the efficiency of the alignment can be dependent on how close the organisms in the samples are with respect to the reference strain, we do not expect this to have a significant impact on the results when species having a comparable genetic distance with the reference are considered, as in our case (for example, Fig. 2a). For each position of the core genome that was determined to be variable by genome alignment, we determined the majority consensus base found in each metagenome. The coverage of the variable positions varied widely between samples and sampling sites, from an average of 11× in the buccal mucosa samples to an average of 150× in the tongue dorsum samples (Supplementary Fig. 20). This is predominantly due to the substantially different fraction of human DNA contamination between oral sites (average of 82% for the buccal mucosa but only 19% for the tongue dorsum) rather than to differences in read depth or average relative abundance (6.5% for the tongue dorsum and 3.9% for the buccal mucosa). Given that due to statistical fluctuations (Supplementary Table 6) not all positions were covered in each metagenome, this procedure yielded for each metagenome a set of SNPs of variable length. Selecting samples covering at least 20,000 SNPs and choosing only those positions that were covered in all samples defined a final data set of 5,578 positions.

PCA of the allelic profiles of both metagenomes and genomes was performed using the function *dudi.pca* of the *ade4* package of R v3.0.2.

Distances between the profiles of SNPs were computed as the fraction of discordant SNPs using the function *dist.dna* (with the option *model* = '*raw*') from the *ape* package of R. PERMANOVA *P* values were computed using the *adonis* function of the *vegan* package of R.

Discriminant analysis of principal components was performed using the functions *find.clusters* and *dapc* of the *adegenet* package of R. The number of clusters was determined using the Bayes information criterion as described in ref. 33.

Synthetic metagenomes. Synthetic metagenomes were generated using Grinder using a fourth-degree polynomial error model for Illumina reads[34] with the parameters -total_reads 500000 -read_dist 101 -mutation_dist poly4 3e-3 3.3e-8 -fastq_output 1 -qual_levels 30 10 -insert_dist 396 normal 27 -exclude_chars N -diversity 2. The four values of relative abundance were obtained using the parameters -abundance_model powerlaw 0.5, 1.0, 2.5 and 5.0, respectively. The average coverage of the most abundant species was between 14× and 24× (assuming an average length of the genomes of 2 Mb) and for the least abundant species was between 11× and 1×.

Metagenomic MLST analysis. We performed an *in silico* MLST analysis from metagenomes by mapping the metagenomic samples against all known MLST sequences for neisseriae (4,606 total sequences, 7 loci from 7 genes: abcZ, adk, aroE, fumC, gdh, pdhC and pgm) available from PubMLST.org (ref. 36). The mapping was performed with BowTie2 sequence aligner (local alignment, with parameters -D 20 -R 3 -N 0 -L 20 -i S,1,0.50)[67]. The obtained alignment was then used to: (1) identify the closest reference variant (allele) for each MLST locus and (2) construct a consensus sequence for each locus, using the best-matching reference allele detected at step (1) as a blueprint to fill gaps from the reads. A nucleotide-by-nucleotide majority rule approach was used to define each nucleotide in the sequence of each locus in the case of overlapping reads. This operation was performed for each sample independently. Using custom scripts, we then analysed all the resulting sequences together to assemble an allelic profile for each sample, considering duplicates and repetitions in the allelic profiles. We identified known profiles with respect to the publicly available MLST typing table, and we observed potentially new sequence types (and alleles) for neisseriae. Isolates from PubMLST database[68] were used to associate species to the known allelic profiles. The version of the pipeline used for this *in silico* MLST analysis has been implemented and is available with additional information and tutorials at http://segatalab.cibio.unitn.it/tools/metamlst/.

The concatenated alignment of the seven loci was used to compute a matrix of the distances between sequence types, using the raw number of substitutions as a measure of distance using the function *dist.dna* of the *ape* software package R v3.0.2. A PCoA of the distance matrix was performed using the function *cmdscale* in R.

Strain-level MetaPhlAn marker profiling. MetaPhlAn (ref. 37) is a taxonomic profiling tool for shotgun metagenomic data based on clade-specific markers. In version 2.0 (ref. 32) it employs up to 200 such markers that are species-specific, meaning that (1) those genes are core genes for the clade and (2) those genes are not present in any other clade. Markers are then used to infer relative abundances based on their average coverage across clades. For those species with a relatively small number of sequenced genomes (including all Neisseriaceae species except for *N. meningitidis* and *N. gonorrhoeae*), the markers are, however, not guaranteed to respect the 'coreness' property and can thus occasionally miss in samples containing the target species. In these cases, the pattern of marker presence/absence for a specific species provides a strain barcoding system that can be used to track strains across samples and test the genomic identity of strains in different samples[69]. For this Article, we performed such a barcoding method on all samples in which a *Neisseria* species was present at a relative abundance of at least 1% and at least 33% of the markers for that species were consistently present.

To extract the markers for different *Neisseria* species, it was sufficient to process the MetaPhlAn2 database file (available in the db_v2 subfolder of the MetaPhlAn2 repository https://bitbucket.org/biobakery/metaphlan2/src) with the BowTie2 package, with scripts using the following syntax:

$ bowtie2-inspect db_v20/mpa–v20_m200 > db_v20/markers.fasta

$ python mpa3src/extract_markers.py--mpa_pkl db_v20/mpa_v20_m200.pkl--clade s__Neisseria_meningitidis--ofn_markers db_v20/s__Neisseria_meningitidis.markers.fasta--ifn_markers db_v20/markers.fasta

This example is specific for *N. meningitidis*, but markers can be extracted for each of the *Neisseria* species using the '—clade' option with the following general rule: 's__Neisseria_speciesname'.

Accepted 19 April 2016;

References

1. Novak, N., Haberstok, J., Bieber, T. & Allam, J. P. The immune privilege of the oral mucosa. *Trends Mol. Med.* **14,** 191–198 (2008).
2. Feller, L. *et al.* Oral mucosal immunity. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.* **116,** 576–583 (2013).
3. Allam, J. P. *et al.* Toll-like receptor 4 ligation enforces tolerogenic properties of oral mucosal Langerhans cells. *J. Allerg. Clin. Immunol.* **121,** 368–374 (2008).
4. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490,** 55–60 (2012).
5. Tlaskalova-Hogenova, H. *et al.* The role of gut microbiota (commensal bacteria) and the mucosal barrier in the pathogenesis of inflammatory and autoimmune diseases and cancer: contribution of germ-free and gnotobiotic animal models of human diseases. *Cell. Mol. Immunol.* **8,** 110–120 (2011).
6. Nascimento, M. M., Gordan, V. V., Garvan, C. W., Browngardt, C. M. & Burne, R. A. Correlations of oral bacterial arginine and urea catabolism with caries experience. *Oral Microbiol. Immunol.* **24,** 89–95 (2009).
7. Burton, J. P., Chilcott, C. N., Moore, C. J., Speiser, G. & Tagg, J. R. A preliminary study of the effect of probiotic *Streptococcus salivarius* K12 on oral malodour parameters. *J. Appl. Microbiol.* **100,** 754–764 (2006).
8. Napimoga, M. H., Hofling, J. F., Klein, M. I., Kamiya, R. U. & Goncalves, R. B. Transmission, diversity and virulence factors of *Streptococcus mutans* genotypes. *J. Oral Sci.* **47,** 59–64 (2005).
9. De Chiara, M. *et al.* Genome sequencing of disease and carriage isolates of nontypeable *Haemophilus influenzae* identifies discrete population structure. *Proc. Natl Acad. Sci. USA* **111,** 5439–5444 (2014).
10. Kadioglu, A., Weiser, J. N., Paton, J. C. & Andrew, P. W. The role of *Streptococcus pneumoniae* virulence factors in host respiratory colonization and disease. *Nature Rev. Microbiol.* **6,** 288–301 (2008).
11. Napimoga, M. H. *et al.* Genotypic diversity and virulence traits of *Streptococcus mutans* in caries-free and caries-active individuals. *J. Med. Microbiol.* **53,** 697–703 (2004).
12. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial 'pan-genome'. *Proc. Natl Acad. Sci. USA* **102,** 13950–13955 (2005).
13. Hanage, W. P., Fraser, C. & Spratt, B. G. The impact of homologous recombination on the generation of diversity in bacteria. *J. Theor. Biol.* **239,** 210–219 (2006).
14. Segata, N. *et al.* Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biol.* **13,** R42 (2012).
15. Snyder, L. A. & Saunders, N. J. The majority of genes in the pathogenic *Neisseria* species are present in non-pathogenic *Neisseria lactamica*, including those designated as 'virulence genes'. *BMC Genomics* **7,** 128 (2006).
16. Knapp, J. S., Totten, P., Mulks, M. & Minshew, B. Characterization of *Neisseria cinerea*, a nonpathogenic species isolated on Martin–Lewis medium selective for pathogenic Neisseria spp. *J. Clin. Microbiol.* **19,** 63–67 (1984).
17. Johnson, A. P. The pathogenic potential of commensal species of Neisseria. *J. Clin. Pathol.* **36,** 213–223 (1983).
18. Maiden, M. C. *et al.* Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl Acad. Sci. USA* **95,** 3140–3145 (1998).

19. Urwin, R. *et al.* Distribution of surface protein variants among hyperinvasive meningococci: implications for vaccine design. *Infect. Immun.* **72,** 5955–5962 (2004).
20. Eren, A. M., Borisy, G. G., Huse, S. M. & Mark Welch, J. L. Oligotyping analysis of the human oral microbiome. *Proc. Natl Acad. Sci. USA* **111,** E2875–E2884 (2014).
21. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–214 (2012).
22. The Human Microbiome Project Consortium. A framework for human microbiome research. *Nature* **486,** 215–221 (2012).
23. Muzzi, A., Mora, M., Pizza, M., Rappuoli, R. & Donati, C. Conservation of meningococcal antigens in the genus *Neisseria*. *mBio* **4,** e00163–e00113 (2013).
24. Knapp, J. S. & Hook, E. W. III. Prevalence and persistence of *Neisseria cinerea* and other *Neisseria* spp. in adults. *J. Clin. Microbiol.* **26,** 896–900 (1988).
25. Bennett, J. S. *et al.* A genomic approach to bacterial taxonomy: an examination and proposed reclassification of species within the genus *Neisseria*. *Microbiology* **158,** 1570–1580 (2012).
26. Marri, P. R. *et al.* Genome sequencing reveals widespread virulence gene exchange among human Neisseria species. *PLoS ONE* **5,** e11835 (2010).
27. Segata, N., Bornigen, D., Morgan, X. C. & Huttenhower, C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature Commun.* **4,** 2304 (2013).
28. Tinsley, C. R. & Nassif, X. Analysis of the genetic differences between *Neisseria meningitidis* and *Neisseria gonorrhoeae*: two closely related bacteria expressing two different pathogenicities. *Proc. Natl Acad. Sci. USA* **93,** 11109–11114 (1996).
29. Bennett, J. S., Watkins, E. R., Jolley, K. A., Harrison, O. B. & Maiden, M. C. Identifying Neisseria species by use of the 50S ribosomal protein L6 (rplF) gene. *J. Clin. Microbiol.* **52,** 1375–1381 (2014).
30. Treangen, T. J., Ambur, O. H., Tonjum, T. & Rocha, E. The impact of the neisserial DNA uptake sequences on genome evolution and stability. *Genome Biol.* **9,** R60 (2008).
31. Frye, S. A., Nilsen, M., Tonjum, T. & Ambur, O. H. Dialects of the DNA uptake sequence in Neisseriaceae. *PLoS Genet.* **9,** e1003458 (2013).
32. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods* **12,** 902–903 (2015).
33. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* **11,** 94 (2010).
34. Angly, F. E., Willner, D., Rohwer, F., Hugenholtz, P. & Tyson, G. W. Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res.* **40,** e94 (2012).
35. Maiden, M. C. Multilocus sequence typing of bacteria. *Ann. Rev. Microbiol.* **60,** 561–588 (2006).
36. Jolley, K. A. & Maiden, M. C. BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11,** 595 (2010).
37. Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods* **9,** 811–814 (2012).
38. Jones, G. R. *et al.* Dynamics of carriage of *Neisseria meningitidis* in a group of military recruits: subtype stability and specificity of the immune response following colonization. *J. Infect. Dis.* **178,** 451–459 (1998).
39. Glitza, I. C. *et al.* Longitudinal study of meningococcal carrier rates in teenagers. *Int. J. Hygiene Environ. Health* **211,** 263–272 (2008).
40. Jolley, K. A., Wilson, D. J., Kriz, P., McVean, G. & Maiden, M. C. The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in *Neisseria meningitidis*. *Mol. Biol. Evol.* **22,** 562–569 (2005).
41. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493,** 45–50 (2013).
42. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31,** 1674–1676 (2015).
43. Namiki, T., Hachiya, T., Tanaka, H. & Sakakibara, Y. MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. *Nucleic Acids Res.* **40,** e155 (2012).
44. Peng, Y., Leung, H. C., Yiu, S. M. & Chin, F. Y. Meta-IDBA: a *de novo* assembler for metagenomic data. *Bioinformatics* **27,** i94–i101 (2011).
45. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nature Methods* **11,** 1144–1146 (2014).
46. Imelfort, M. *et al.* GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ* **2,** e603 (2014).
47. Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523,** 208–211 (2015).
48. Loman, N. J. *et al.* A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4. *Jama* **309,** 1502–1510 (2013).
49. Hong, C. *et al.* PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome* **2,** 1–15 (2014).
50. Francis, O. E. *et al.* Pathoscope: species identification and strain attribution with unassembled sequencing data. *Genome Res.* **23,** 1721–1729 (2013).
51. Scholz, M. *et al.* Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nature Methods* **13,** 435–438 (2016).
52. Greenblum, S., Carr, R. & Borenstein, E. Extensive strain-level copy-number variation across human gut microbiome species. *Cell* **160,** 583–594 (2015).
53. Kaper, J. B., Nataro, J. P. & Mobley, H. L. Pathogenic *Escherichia coli*. *Nature Rev. Microbiol.* **2,** 123–140 (2004).
54. Rasko, D. A. *et al.* The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J. Bacteriol.* **190,** 6881–6893 (2008).
55. Parsonnet, J., Friedman, G. D., Orentreich, N. & Vogelman, H. Risk for gastric cancer in people with CagA positive or CagA negative *Helicobacter pylori* infection. *Gut* **40,** 297–301 (1997).
56. Hatakeyama, M. *Helicobacter pylori* CagA and gastric cancer: a paradigm for hit-and-run carcinogenesis. *Cell Host Microbe* **15,** 306–316 (2014).
57. Nguyen, L. T., Uchida, T., Murakami, K., Fujioka, T. & Moriyama, M. *Helicobacter pylori* virulence and the diversity of gastric cancer in Asia. *J. Med. Microbiol.* **57,** 1445–1453 (2008).
58. Goldschneider, I., Gotschlich, E. C. & Artenstein, M. S. Human immunity to the meningococcus. II. Development of natural immunity. *J. Exp. Med.* **129,** 1327–1348 (1969).
59. Gold, R., Goldschneider, I., Lepow, M. L., Draper, T. F. & Randolph, M. Carriage of *Neisseria meningitidis* and *Neisseria lactamica* in infants and children. *J. Infect. Dis.* **137,** 112–121 (1978).
60. Maiden, M. C. & Frosch, M. Can we, should we, eradicate the meningococcus? *Vaccine* **30**(Suppl 2), B52–56 (2012).
61. Moir, J. W. Meningitis in adolescents: the role of commensal microbiota. *Trends Microbiol.* **23,** 181–182 (2015).
62. Leimkugel, J. *et al.* Clonal waves of Neisseria colonisation and disease in the African meningitis belt: eight-year longitudinal study in northern Ghana. *PLoS Med.* **4,** e101 (2007).
63. Pruitt, K. D., Tatusova, T., Brown, G. R. & Maglott, D. R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40,** D130–D135 (2012).
64. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5,** R12 (2004).
65. Muzzi, A. & Donati, C. Population genetics and evolution of the pan-genome of *Streptococcus pneumoniae*. *Int. J. Med. Microbiol.* **301,** 619–622 (2011).
66. Hein, J., Schierup, M. & Wiuf, C. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory* (Oxford Univ. Press, 2004).
67. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (2012).
68. Jolley, K. A., Chan, M. S. & Maiden, M. C. mlstdbNet—distributed multi-locus sequence typing (MLST) databases. *BMC Bioinformatics* **5,** 86 (2004).
69. Franzosa, E. A. *et al.* Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl Acad. Sci. USA* **111,** E2329–2338 (2014).
70. Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C. & Segata, N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ* **3,** e1029 (2015).

Acknowledgements

The authors acknowledge the Human Microbiome Project Consortium and the generous participation of individuals from the Saint Louis (MO) and Houston (TX) areas who made the Human Microbiome Project possible. This work was supported in part by NIH grants R01HG005969 and U54DE023798, NSF grant DBI-1053486 and Army Research Office grant W911NF-11-1-0473 to C.H., and a European Union FP7 Marie-Curie grant (PCIG13-618833), MIUR grant FIR RBFR13EWWI, Fondazione Caritro grant Rif. Int.2013.0239, CIBIO Start-up funds and Terme di Comano grants to N.S.

Author contributions

C.D. and N.S. conceived the study, implemented the software and performed the analyses. M.Z., D.A., D.T.T., F.A., V.I. and C.H. contributed to the analyses. D.C., O.J., C.D.F. and C.H. provided feedback and contributed to the writing. C.D., C.H. and N.S. wrote the manuscript.

Competing interests

The authors declare no competing financial interests.