Published online 19 September 2016

doi: 10.1093/nar/gkw837

MetaMLST: multi-locus strain-level bacterial typing from metagenomic samples

Moreno Zolfo[1], Adrian Tett[1], Olivier Jousson[1], Claudio Donati[2] and Nicola Segata[1,*]

[1]Centre for Integrative Biology, University of Trento, Trento, TN 38123, Italy and [2]Computational Biology Unit, Research and Innovation Centre, Fondazione Edmund Mach, Via Edmund Mach 1, San Michele all'Adige 38010, Italy

Received July 16, 2016; Revised August 31, 2016; Accepted September 10, 2016

ABSTRACT

Metagenomic characterization of microbial communities has the potential to become a tool to identify pathogens in human samples. However, software tools able to extract strain-level typing information from metagenomic data are needed. Low-throughput molecular typing schema such as Multilocus Sequence Typing (MLST) are still widely used and provide a wealth of strain-level information that is currently not exploited by metagenomic methods. We introduce MetaMLST, a software tool that reconstructs the MLST loci of microorganisms present in microbial communities from metagenomic data. Tested on synthetic and spiked-in real metagenomes, the pipeline was able to reconstruct the MLST sequences with >98.5% accuracy at coverages as low as 1×. On real samples, the pipeline showed higher sensitivity than assembly-based approaches and it proved successful in identifying strains in epidemic outbreaks as well as in intestinal, skin and gastrointestinal microbiome samples.

INTRODUCTION

High resolution microbial strain identification and tracking is a key challenge both in a clinical and research settings. One of the most popular methods for strain level typing of microorganisms is multilocus sequence typing (MLST) (1), which is based on sequencing a small number of species-specific genomic loci (usually seven) which are known to be present in all strains of the given target species. Thanks to its simplicity and resolution, MLST approaches have been defined and adopted for many prokaryotic (2,3) and eukaryotic microbes (4). Databases of thousands of MLST profiles and sequences are now available for a large number of microbial species most of which are (opportunistic) pathogens (5,6).

The limiting factor that hampers the routine use of MLST in a clinical setting is the need to isolate and cultivate each bacterial species of interest. The MLST protocol (DNA extraction, polymerase chain reaction (PCR) amplification, purification and sequencing of the target loci) is also expensive and laborious. With the increasing throughput and decreasing cost of next generation sequencing technologies, the direct sequencing of the entire DNA content of a sample (metagenomics (7)) is rapidly becoming an effective approach for the characterization of complex microbial communities as it skips the time-consuming isolation and cultivation steps. Metagenomic datasets contain sequence information for all strains present in a given microbial community and can provide, in theory, typing data for all the species of interest within the sample.

A possible strategy to achieve metagenomic MLST typing involves the use of metagenomics assemblies, where the assembled metagenomics contigs are mapped against the MLST databases. However, this approach can only uncover the strains that are abundant (i.e. those that can have a sufficient depth to be metagenomically assembled). Moreover, metagenomics assembly is computationally demanding. Therefore, there is currently no method to easily and efficiently extract MLST loci from metagenomes.

To combine the effectiveness of the MLST approach with the ease of cultivation-free and high throughput metagenomics, we developed a novel computational pipeline for microbial typing called MetaMLST. Given a template database of MLST loci, for each species with an available MLST schema, MetaMLST performs an *in silico* consensus sequence reconstruction of the allelic profile of the microbial strains in a metagenomics sample. The reconstructed loci (profiles) are then identified by comparing them to a database of publicly available profiles maintained in PubMLST (5). New alleles, i.e. absent in the database, are determined by comparative sequence reconstruction using the sequence internal database as a template and given a confidence score. This mapping-based approach overcomes the computational limitations and lowers the limit of detection compared to metagenomic assembly. Following the downstream standard MLST pipeline, samples are then processed individually to determine the sequence types (ST) profiles and combined for epidemiological analyses, possibly integrating them with the large set of profiled strains available in public databases. The software is freely available

[*]To whom correspondence should be addressed. Tel: +39 461 285218; Fax: +39 461 283937; Email: nicola.segata@unitn.it

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

with supporting material at http://segatalab.cibio.unitn.it/tools/metamlst/.

MATERIALS AND METHODS

The pipeline MetaMLST processes metagenomics samples and identifies the most abundant MLST ST profiles of target species in each sample (Supplementary Figure S1). MetaMLST can also identify novel STs and alleles. The pipeline is organized in four steps: (i) retrieval of the available MLST data; (ii) mapping of the metagenomic reads against the retrieved reference sequences; (iii) detection of microbial targets and reconstruction of the sample-specific MLST loci; (iv) ST calling and downstream comparative analysis. Step (i) is routinely performed and made available at the MetaMLST website, and users can thus skip this first step and download the MetaMLST database directly.

Retrieval and curation of available MLST protocols and data

The available MLST sequences and profile were retrieved from the public archives of PubMLST (5) and mlst.net (6). Sequences were semi-manually curated to standardize the names of the loci across the database (species_locus_alleleID). Profiles and alleles were checked for consistency: profiles containing sequences of non-existing alleles, sequences of very low complexity (e.g. all-adenines), artefacts and sequences with inconsistent length (i.e. different in length by the majority of the alleles of the locus) were all removed from the internal database.

The retrieved MLST reference information was then organized in a SQLite 3 database. Additionally, the database can be expanded and personalized by the user starting from MLST sequences and profiles, respectively in FASTA format and Tab-separated format as described in the software repository.

Mapping phase

The pipeline accepts as input sequence files in FASTQ format which are subsequently mapped against all or part of the MLST loci present in the MetaMLST SQLite database. In this operation the pipeline extracts sequences of all loci present in the database and the MetaMLST-index module assembles a Bowtie2 index using the bowtie2-build tool. The metagenomes are then mapped against the index using Bowtie2 v. 2.2.6 (8). The resulting mapping file (in BAM format) is then processed by the downstream MetaMLST steps (below). For MetaMLST we recommend using Bowtie2 with local mapping using the following parameters: -a –no-unal -D 20 -R 3 -N 0 -L 20 -i S,1,0.50 (–very-sensitive-local).

MLST alleles reconstruction phase

The core of MetaMLST is the reconstruction of the MLST loci from the metagenomic reads. First, the algorithm considers the alignments of the reads against the internal database of MLST loci (mapping phase) to find the closest (or identical) reference MLST allele in the sample for each species of interest. This is done by computing the mapping score for each allele and selecting the allele that maximizes this score. The score of an allele k is defined as the average alignment score (as provided by Bowtie2) of the reads that map against the sequence of the allele k penalized by a value proportional to the number of reads that map against another allele other than k. The weight of the penalization is controlled by a parameter p. In our validation experiments, we assessed varying parameter p from 30 to 80 (the minimum score threshold that MetaMLST applies to select successful Bowtie2 alignments) and found it did not affect the selection of the best reference allele, thus we recommend 50 as a default value. Specifically, the tool defines the reference allele (*Ref_Allele*) as:

$$\mathrm{Ref\_Allele} = \mathrm{argma}x_k \left(\frac{\sum_{i=1}^{n_k} (S_{k,i})}{n_k} - p \cdot \frac{N - n_k}{n_k} \right)$$

where n_k is the number of reads aligning to that allele 'k', $S_{\mathrm{k,i}}$ is the Bowtie2 alignment score of the i-th read against that allele 'k' (AS:i:<N> field in the Bowtie2 SAM data format), N - n_k is the difference between the number of reads aligning to at least one allele of that locus (N) and the number of reads aligning to the specific allele k (n_k), and p is a penalty parameter regulating the impact of non-perfect matches (default: 50).

Once the reference allele for each locus of each MLST-target species in the sample has been identified, the pipeline reconstructs the sample-specific sequence of the loci basing on the reference allele. A consensus sequence is built starting from the reads aligning to the locus, and using a majority rule to determine the nucleotide in each position. Specifically, the aligned reads, grouped by species and locus, pass through the samtools-mpileup tool v 0.1.19 (9) (primary and secondary alignments are considered equally). Then, if the reads do not cover the full length of the locus, the sequence of the reference allele is used to infer the nucleotidic sequence for the positions at zero coverage (i.e. with no aligning-reads) and the corresponding non-perfect confidence C_k for the locus k is set to:

$$C_k = 1 - \frac{B_k}{L_k}$$

where: B_k is the number of uncovered position in the locus k and L_k is the length of the locus k.

The MetaMLST.py script outputs the list of the species detected in the sample, the reconstructed sequences of their MLST loci, and for each locus a confidence score and the percentage of single nucleotide variants (SNV) from the closest reference over the whole locus length. Selectable thresholds allow the user to filter Bowtie2 alignments below a certain length (min_length) or score (minscore) and above a certain amount of SNVs with respect to the closest reference allele (max_xM). Min_length was set to 90nt in order to remove shorter alignments (considering original Illumina reads of 100 bps). The default values for max_xM and minscore (respectively 5 and 80) were chosen based on a grid search strategy (values in {3,5,8,10} for max_xM and {40,60,80,100} for minscore) performed on synthetic data maximizing the percentage identity of the predictions compared to the original genomes (Supplementary Table S1).

Merging and comparing profiles

MetaMLST then analyzes the upstream-generated data and assigns an ST to each sample (MetaMLST-merge.py script). By comparing the reconstructed sequences with the database, we identify putative new alleles not already present in the MLST repositories. New alleles are assigned if they differ by no more than three SNVs compared to any of the reference alleles. This conservative user-selectable threshold corresponds to the allelic intra-locus sequence divergence observed for 75% of the alleles (Supplementary Figure S2): samples containing new alleles exceeding this value are discarded from the downstream analysis, as these loci are likely to belong to a different organism. If the same new allele occurs in more than one sample, the pipeline is able to track it and assigns the same identifier.

In MLST, each referenced ST is defined by a combination of alleles; MetaMLST can identify both reference profiles and novel profiles (i.e. containing one or more new alleles, or a new combination of existing ones). If the same combination of alleles occurs in more than one sample, MetaMLST-merge.py tracks it and assigns the same ST to all of the identical profiles, allowing the user to track these new types across multiple samples.

Construction of the synthetic and semi-synthetic datasets used for validation

For the synthetic metagenome validation, 12 metagenomes were generated from reference genomes (20 million reads, 50 microbial species each). The 50 abundance values were randomly sampled from a lognormal distribution and normalized to produce the relative abundances with an Illumina error model (10). We assigned the top *n* relative abundances to the *n* MLST-target microbes ($4 \leq n \leq 5$, Supplementary Table S2), and the remaining to non-MLST-target species. The reconstruction step of MetaMLST was run on the metagenomes and the reconstructed sequences were extracted and mapped with BLAST (11) against the reference genomes used to compose the corresponding metagenome. For every metagenome, the identity score was calculated for each locus of each target species.

To generate the semi-synthetic metagenomes used to validate the pipeline's results at various coverages, we selected six target species and six metagenomes from the Human Microbiome Project (HMP) which resulted negative for those species when analyzed with MetaPhlAn 2 (12,13) (an application specifically designed for the detection, but not for the typing, of microbial species in metagenomes). We generated a set of synthetic reads at various coverages (Supplementary Table S4) and merged them with the corresponding metagenome. MetaMLST.py was run on the resulting samples and the reconstructed sequences were mapped with BLAST as for the synthetic validation mentioned above.

Post-processing and visualization analysis

Minimum Spanning Trees were obtained with PHYLOViZ v. 1.1 using the eBURST Full-MST algorithm (14,15). The trees for phylogenetic analysis were elaborated with RAXML v8 (16) using (-m GTRCAT -p 1234), and then plotted with GraPhlAn v. 1.1 (17), using the metadata associated with each sample. The principal component analysis (PCA) plots were computed using the scikit-learn package (18) providing the multi-loci concatenated alignments built using MUSCLE v3.8.31 (19) as arrays of binary features.

For the analysis on the ancient *Helicobacter pylori* metagenomic reads from 'Otzi The Iceman', we mapped the available reads from Maixner *et al.* (20) against the database of *H. pylori* alleles from PubMLST (5). Before mapping human reads were removed using Bowtie2 (8) against the human genome (hg19) in end-to-end mode and selecting only the non-aligning reads against the hg19 index (–un-bz2 option). MetaMLST was then applied on the alignments. We carried out the PCA analysis as mentioned above on the only sample that passed the strict thresholds applied by MetaMLST (sample C0057 from the original study), which had a confidence score of 95.94% according to the pipeline.

Code and data availability

MetaMLST is Open Source, released in Bitbucket and freely available with supporting material, tutorials, and scripts at: http://segatalab.cibio.unitn.it/tools/metamlst. The synthetic metagenomes used in the method's validation are available under BioProject PRJNA339720.

RESULTS

The MetaMLST pipeline for strain-level typing

MetaMLST is a software pipeline that processes metagenomic reads to provide a cultivation-free version of the MLST approach (1) (Supplementary Figure S1). For each microbial species with an available MLST schema, the pipeline is able to identify and track the dominant strain of that species in a complex microbiome sample. MetaMLST adopts the set of loci (typically between 5 and 10) designed in the organism-specific MLST protocols and uses the sequence variants (alleles) of these loci that are available from public databases (5,6).

In the first step of the pipeline, we collect the entire MLST sequence repertoire from public sources removing duplicated or chimeric sequences (see 'Materials and Methods' section) and we provide the user with the output of this initial pre-processing. Next, reads from the metagenomic samples are aligned against the MLST sequence database using Bowtie2 (8). A microbial species is considered detected if all its MLST loci are found in the sample (see 'Materials and Methods' section): partially detected species are thus discarded from the downstream analysis. From the alignments we build a consensus sequence for each locus using a simple majority rule. If low coverage prevents the determination of the nucleotide sequence at a limited number of positions, the closest allele from the reference database is used to guide the definition of the consensus sequence. This only occurs when these positions are at the terminal parts of the sequence (i.e. the first or last regions of each locus). These terminal parts are typically highly conserved, but a confidence score reflecting the potential uncertainty introduced by this procedure is provided by the tool for each reconstructed locus (see 'Materials and Methods' section). The stringency threshold of this score can be set by the user to limit the fraction of reconstructed nucleotides permitted. The reconstructed loci are then processed as per the standard MLST procedure by assigning a MLST sequence-type (ST) (5) to each detected strain (see 'Materials and Methods' section). If the tool detects a set of alleles that do not correspond to known STs in the reference database, new alleles and STs are defined. These new alleles and profiles can be included in the internal database, to be used in future analyses or to cross compare different metagenomics samples.

Validation on synthetic and semi-synthetic metagenomes

We first evaluated the accuracy of MetaMLST by identifying the STs from a set of 12 synthetic metagenomes which comprised, in total, 240 million reads (Supplementary Table S2). Each synthetic metagenome contained 50 species, 5 of which represented target species present in the MLST reference databases. The relative abundances of the species followed a lognormal distribution with the abundance of the target species being randomly selected (see 'Materials and Methods' section and Supplementary Table S2). Even though some target strains were only present at low coverages, MetaMLST correctly identified more than 96% of the target STs (Figure 1A , Supplementary Table S3). For example, *Salmonella enterica* was correctly recognized even when at low abundance in the synthetic metagenome (0.87% corresponding to 174k reads and a coverage of 3.7×, Supplementary Tables S2 and 3). On the other hand, in this validation, MetaMLST was always run for all species with an available MLST scheme and it never detected and/or typed species not present in the synthetic sample, thus achieving a 100% specificity.

To further validate the performances of the pipeline we generated semi-synthetic metagenomes by mixing real microbiome samples and synthetic reads taken at variable coverages (from 0.5 to 25×, see Supplementary Table S4). Synthetically generated reads from six MLST-trackable organisms, each representative of a body site sampled by HMP, were merged at various coverages with six HMP samples from the respective body sites (see 'Materials and Methods' section). As expected, the average accuracy of reconstruction by MetaMLST shows a coverage-dependent behavior, but even at a coverage as low as 3× the accuracy reaches >99.85% (Figure 1B and Supplementary Table S5).

As well as bacterial typing, standard MLST approaches are also used to type fungi (4,21,22). To ascertain if MetaMLST can successfully track fungal strains we applied it to both synthetic and semi-synthetic metagenomes containing *Candida albicans*. Even at low coverage MetaMLST showed high accuracy in detecting and typing *C. albicans* (Figure 1B). However, we noticed a decreased accuracy of reconstruction in a subset of *C. albicans* strains, due to the divergence of the two alleles of its diploid genome (e.g. AAT1a and ZWF1b loci). Thus, we recommend MetaMLST only for typing of haploid organisms.

Altogether, our validation showed that MetaMLST can successfully type bacterial and fungal strains with high accuracy in complex metagenomics datasets, even at low coverages. Additionally, no nucleotide errors were detected for coverages >5× in our validations. Importantly, this validation test demonstrated that the confidence scores provided by MetaMLST are highly precise i.e. a high score correlated with an accurate performance whereas a low score suggests that the results should be interpreted with caution (Supplementary Figure S3). Overall, these findings indicate that MetaMLST provides an easy and accurate pipeline for the typing of microbial species from complex metagenomes.

Comparing MetaMLST against the assembly-based approach

To further validate the pipeline we compared the results of MetaMLST with assembled metagenomes available from the HMP (23). Assembly is the only other available approach to identify MLST STs from metagenomes (24) although it is substantially more demanding (both in terms of memory and computational time) and requires higher sequence coverages of the target organisms. We selected four bacterial species commonly found at high abundancies in the HMP skin and stool datasets (*Staphylococcus aureus*, *Propionibacterium acnes*, *Staphylococcus epidermidis* and *E. coli*). The assemblies of these organisms were downloaded from the HMP website and mapped using BLAST (11) against the MLST reference sequences from PubMLST (5) to identify their MLST profile. We then compared the results against the profiles ascribed by MetaMLST applied to the raw reads. MetaMLST successfully identified the target species in all of the 31 metagenomes tested. In contrast to these alleles that were successfully reconstructed by MetaMLST, the metagenomic assembly approach failed to assign an ST to the target organism in 16 cases (i.e. failure to detect all the MLST loci, presumably due to a lack of sufficient coverage) (Figure 1C). Newer metagenomic assemblers like metaSPAdes (25) and MegaHIT (26) may provide higher quality assemblies, but assembly will always have a minimum required coverage that is higher than consensus sequence estimation via mapping. Where STs were assigned both by MetaMLST and assembly-based analyses they were found to be in perfect agreement except in three cases (four loci) where the consensus sequence was different (Supplementary Table S6). Specifically, when comparing the sequences of all the loci for those three cases, the number of single nucleotide differences was generally very low: 2 out of 3003 nucleotides for *S. epidermidis*, 1 out of 4253 for *P. acnes* and 9 out of 2954 for *E. coli*. Additionally, all the differences in the four loci of these three cases were located in positions where multiple different nucleotide choices were supported by the mapping reads suggesting the presence of more than one strain at comparable abundances (Supplementary Figure S4). This confirms that, in a minority of the cases (here 4 out of 233 loci), the presence of multiple strains at very similar abundance can produce non-perfect ST calls. Therefore, MetaMLST demonstrated that on the same metagenomics datasets it can equal and even surpass the performance of an assembly-based approach at a fraction of the time effort and of the coverage often required to perform a metagenomic assembly (27) (Supplementary Table S6).

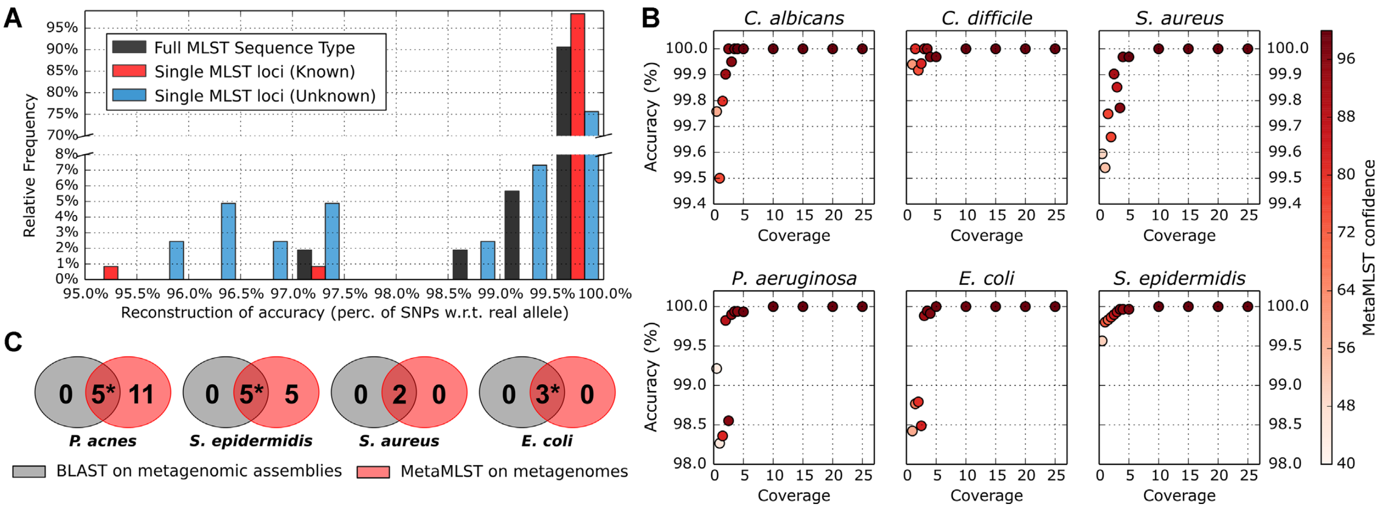

Figure 1. Application of MetaMLST on synthetic and semi-synthetic metagenomes highlights the high accuracy of the approach. (**A**) Frequency histogram for the reconstruction accuracy of MLST profiles (black) and of single MLST loci (blue and red) reconstructed by MetaMLST on synthetic metagenomes. We used a total of 12 synthetic metagenomes (2 Gbps depth each, 100 nt read length), sampled from reference genomes using an Illumina error model (10). The accuracy (i.e. percentage of identity to the reference genomes) was computed for known MLST alleles (i.e. present in existing MLST databases, in red) and for unknown MLST alleles (blue). We consider a sequence type (ST) correctly identified when >99% of the sequence is identical to the corresponding reference genomes. (**B**) Reconstruction accuracy of MLST loci from semi-synthetic datasets at increasing sequence coverages. Color-intensity represents the average of the confidence scores attributed by MetaMLST to the reconstruction of each locus (see 'Materials and Methods' section). (**C**) Number of known ST detected by MetaMLST and by metagenomic assembly in the samples. STs detected by both approaches are in the intersections and marked with an asterisk when one of the predictions is in disagreement. We used here a subset of HMP samples (i.e. samples from anterior nares, retroauricolar crease and stool) for which a metagenomic assembly was run successfully (23). The two methods disagreed for one case in *Propionibacterium acnes*, *Staphylococcus epidermidis* and *Escherichia coli* (marked with an asterisk) and agreed in all other cases. However, MetaMLST often (16 out of 31 cases) identifies more targets compared to metagenomic assembly.

MetaMLST accurately detects *E. coli* sequence types in complex gut metagenomes

We further investigated the performance of MetaMLST by applying it to a set of 531 gut microbiome samples. This set included samples from Type-II Diabetes affected Chinese patients (including the healthy controls) (28), from patients infected by the Shiga toxin-producing *E. coli* O104:H4 from the 2011 German outbreak (29) and healthy subjects from the HMP dataset (23). We focused on *E. coli* as this is one of the most abundant species in the human gut. MetaMLST was able to reconstruct 78 MLST *E. coli* profiles, of which 31 were known and 47 novel STs. Cross-referencing the STs inferred by MetaMLST with the available metadata, we identified common STs across the different metagenomics samples using PCA (Figure 2A) and minimum spanning trees (Figure 2B). The analysis across datasets highlighted a high prevalence of a subset of *E. coli* STs belonging to the ST-complex 10, a common group of commensal *E. coli* STs comprising the majority of Group A *E. coli* (30,31) (Figure 2B). In total 19 different samples showed the presence of a ST included in the complex, of which 10 were classified as *E. coli* ST10.

Importantly, MetaMLST was also able to track the pathogenic agent responsible for the 2011 *E. coli* outbreak in Germany correctly identifying the strain as ST-678, thus confirming the results of Loman *et al.* (29) that were obtained by looking at the MLST loci on the manually curated metagenomic assemblies. In total we detected *E. coli* ST-678 in six samples from the outbreak dataset. For two of the outbreak-positive samples as determined by qPCR (29) we detect a non-pathogenic ST due to its presence at higher abundances than ST-678 as highlighted elsewhere (32) This confirms that MetaMLST detects the most dominant strain for a given species in each sample. MetaMLST also detected ST-678 in one subject enrolled in the Chinese diabetes study that did not present with symptoms of intestinal infection (Figure 2B). Exploring further, by mapping the metagenomic reads against the genome of the ST-678 outbreak isolate (32,33), we found that the *E. coli* ST-678 identified in the Chinese subject did not possess the Shiga-toxin genes responsible for pathogenicity. Therefore, despite the usual enterohemorrhagic phenotype of this ST and its known presence only in patients affected by acute gastroenteritis, hemorrhagic colitis or hemolytic-uremic syndrome according to the MLST records ((5) from http://mlst.warwick.ac.uk/mlst/ as of April 2016) and other typing approaches (34,35), strain ST-678 can also appear as an asymptomatic gut colonizer when the Shiga-toxin gene is not present.

MetaMLST applied to the skin microbiome highlights microbial body-site type and subject specificity

We next applied MetaMLST on an extended set of human skin metagenomic samples which included all those from the HMP (23) and from Oh *et al.* (36) (473 total samples, see Supplementary Table S7). Both datasets included samples taken from the same subjects at different body-sites. Focusing on one of the most prevalent skin inhabitants, *S. epidermidis*, 100 metagenomes resulted positive with MetaMLST identifying 79 different ST of which 60 were putatively novel. The phylogenetic tree built on the reconstructed alle-

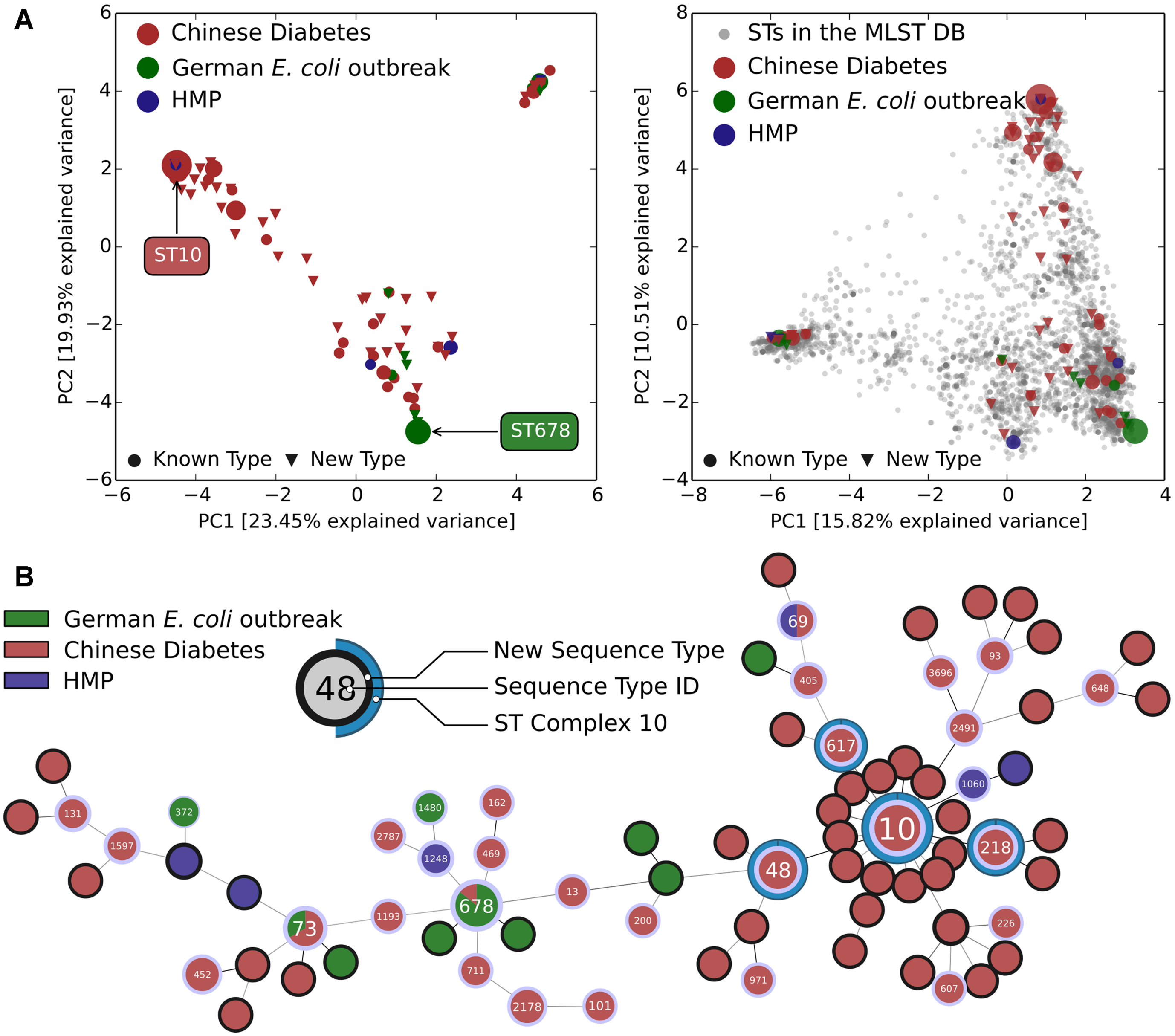

Figure 2. MetaMLST typing *Escherichia coli* in the human gut microbiome (**A**) PCA on reconstructed loci-sequences by MetaMLST (see 'Materials and Methods' section) without (left) and with (right) the publicly available sequences for *E. coli*. Samples are colored by dataset and scaled in size according to the abundance of the ST. Circles represent known STs, while triangles represent new types. The pathogenic type ST-678 associated with the German 2011 outbreak and the benign type ST-10 are highlighted and tracked by MetaMLST. (**B**) Minimum Spanning Tree on the ST, computed with PHYLOViZ full MST algorithm (see 'Materials and Methods' section) and colored by dataset. The size scales with the abundance of each ST and the color is proportional to the contribution of each dataset to that ST. New types detected by MetaMLST are circled in black. Members of the non-pathogenic Clonal Complex 10 are circled in blue to highlight its high prevalence in the Chinese population.

les (using RAxML (16), Figure 3A) showed an association between phylogenetic subtrees and body site type (moist, sebaceous and mixed). Minimum Spanning Tree analysis (15) also shows a clear separation based on body-site (Figure 3B). This segregation confirms the observation by the authors of one of the datasets included in the analysis (36) and further strengthens the idea of *S. epidermidis* body-site specificity. Samples associated with the toe nails have been reported previously (36) to show a mixed—neither fully-sebaceous, nor fully-moist—aggregation in the human microbiome. Interestingly, in our dataset all the toe-nails-associated *S. epidermidis* types were closer to moist ST, both according to the phylogeny and the Minimum Spanning Tree.

We then extended our analysis to two other commonly skin-associated microbes: *P. acnes* and *S. aureus*. We found these organisms STs are highly subject-specific and that these subject specific types are common at different body locations, (Figure 4). For example, for *S. aureus* we found that samples from subjects sh01 (nine samples) and hv10 (eight samples) either belonged to the same ST or were phylogenetically very closely related (Figure 4A). These types

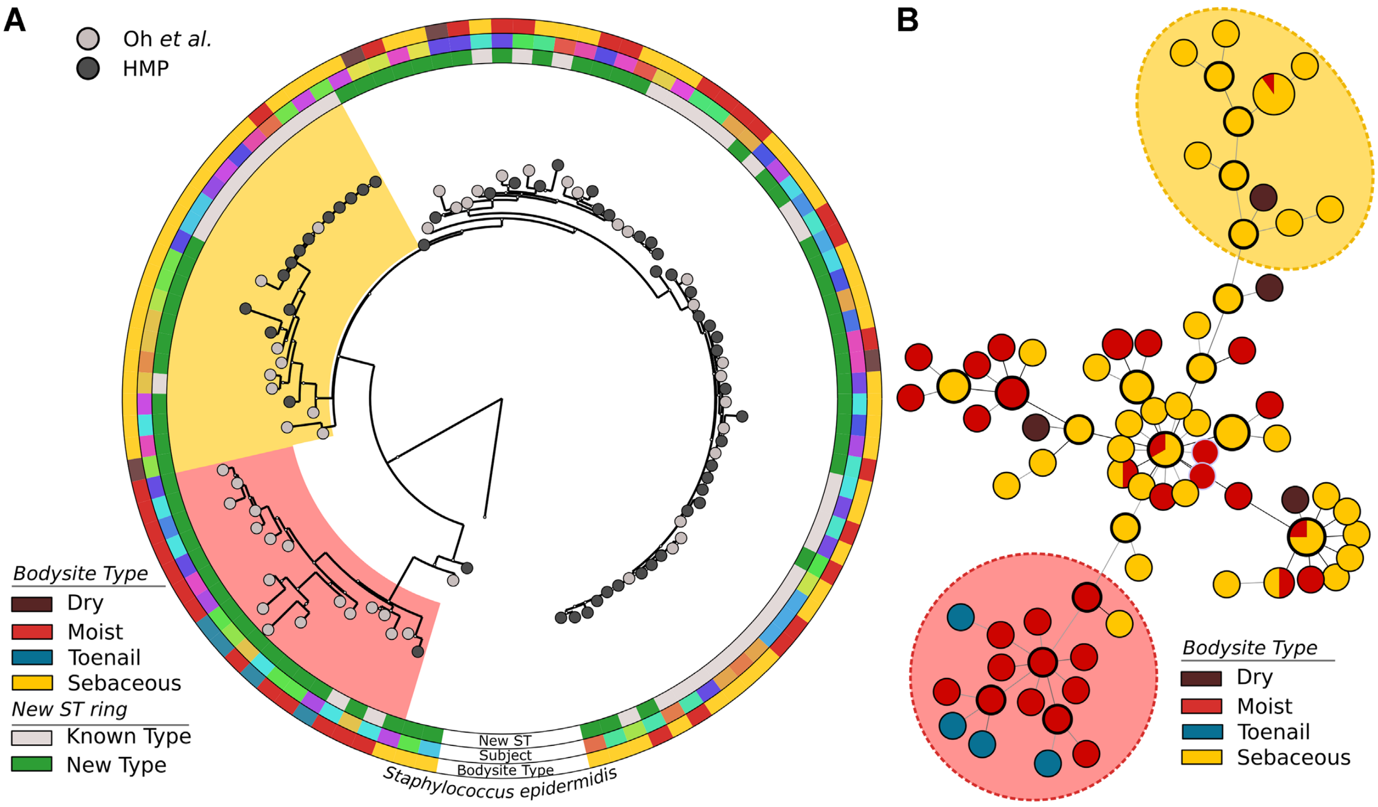

Figure 3. MetaMLST applied to *Staphylococcus epidermidis* on 473 metagenomic samples. (**A**) Phylogenetic tree on the concatenated loci reconstructed by MetaMLST. Oh *et al*. (light gray) and HMP (dark gray) samples are associated with their metadata: Body-site type (outer ring) and Subject ID (intermediate ring). New STs are marked in green in the inner ring. The 'moist' and 'sebaceous' subtrees are highlighted in red and yellow, respectively. (**B**) Minimum Spanning Tree on the ST, computed with PHYLOViZ full MST algorithm and colored by body-site type. Each node represents an ST and the size scales with the abundance of that ST in the dataset. STs detected in more than one body-site type and their relative proportion at each site is indicated in the inner node represented as a pie chart. Two groups of mainly-moist and mainly-sebaceous associated STs are visible at the top and bottom edges of the MST.

could be identified at moist, sebaceous and dry body sites. The same trend could be seen for *P. acnes*, in particular ST 70 was highly conserved and prevalent on a single subject, while other STs (1, 2, 3, 4) were highly prevalent on multiple subjects, even across different datasets. New STs could be detected in both analyses, and often represented minor variations of known STs, with few nucleotides changing from reference alleles. The identification of highly prevalent STs, the segregation of clusters of strains according to the body-site type, as well as the conservation of the same ST in different body sites, all highlight the potentialities of MetaMLST in tracking strains among subjects in clinical or epidemiological settings.

MetaMLST identifies sequence types of pathogens from archaeological samples

We extended our analysis to provide evidences of the usefulness of MetaMLST also in contexts where microbial cultivation, and thus a traditional MLST analysis, is not possible, such as in an archaeological setting. We applied MetaMLST to the metagenomic samples originally extracted from the stomach of 'Ötzi the Iceman', a 5300-year-old mummy found in a melting glacier in northern Italy (20). Analyzing the original metagenomes with MetaMLST, we were able to demonstrate the presence of *H. pylori*, confirming the results of Maixner *et al*. (20), and reconstructed its MLST profile. MetaMLST assigned a new ST to the *H. pylori* strain present in the Iceman's stomach; the ST was however phylogenetically very close to other European STs available in publicly available datasets (5). By comparing the reconstructed profiles with public available MLST types, the ancient *H. pylori* is placed at the boundaries of the European and Asian clusters (Figure 5 and Supplementary Figure S5). The closest non-European STs belong to the hpAsia2 structure population, similarly to what reported by Maixner and colleagues in their original analysis of the dataset (20). MetaMLST reconstructed the MLST loci of the ancient *H. pylori* with a confidence score of 96%, providing evidences about the possibility to efficiently identify and study ancient microbes from archaeological samples with MetaMLST.

DISCUSSION

Identifying and tracking microbial strains is an important task for several biomedical settings including pathogen detection and disease outbreak characterization (37). MLST is one of the most successful approaches in microbial identification and tracking (1), and it has been extensively applied in the last twenty years (38,39). MLST provides strain-level resolution, it is highly reproducible across different labora-

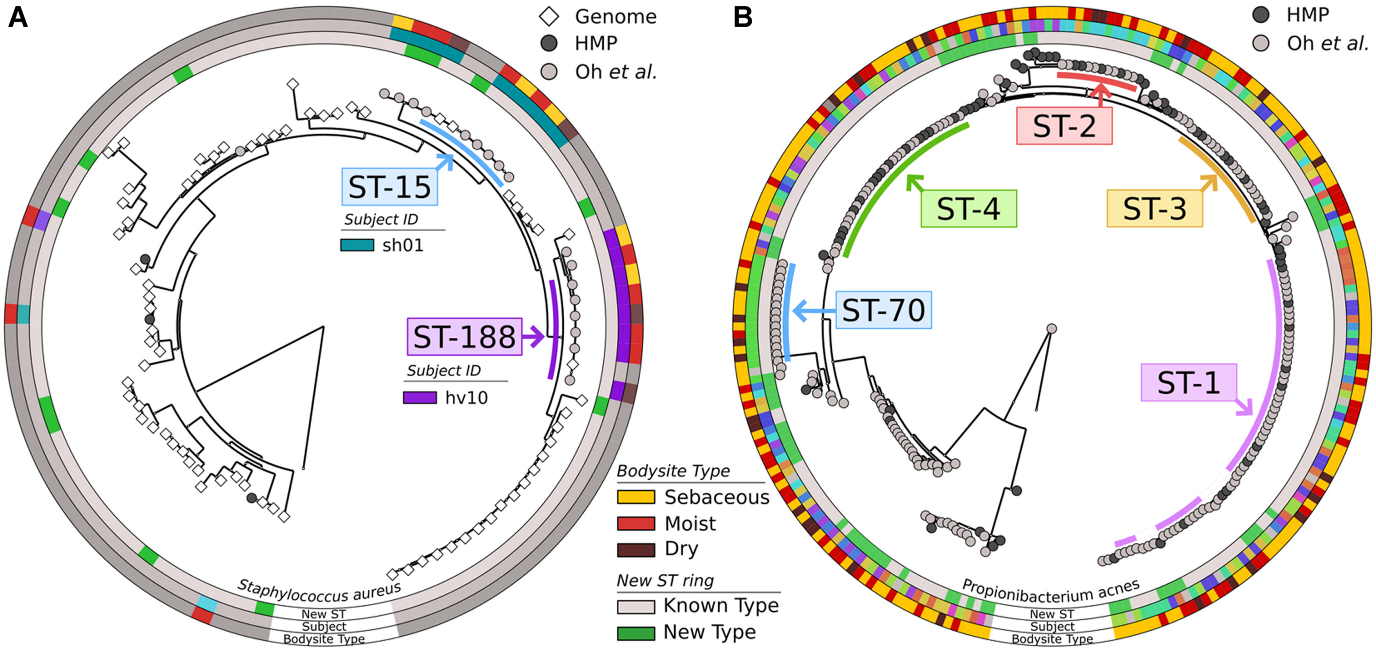

Figure 4. MetaMLST applied to *Staphylococcus aureus* (**A**) and *Propionibacterium acnes* (**B**) on 473 metagenomic samples. Phylogenetic tree on the concatenated loci reconstructed by MetaMLST. Oh *et al.* (light gray) and HMP (dark gray) samples are associated with their metadata: Subject ID (intermediate ring) and body-site type (outer ring). New STs are marked in green in the inner ring. *S. aureus* tree (A) was computed together with available reference genomes. Particularly, subjects sh01 (nine samples, light blue) and hv10 (eight samples, purple) were colonized with either the same or very closely related *S. aureus*. In *P. acnes* we show instead both STs that are highly conserved in one subject (ST-70, blue arc) as well as STs that are highly prevalent across different subjects (STs 2, 3, 4 and 1). Prevalence in the cohort and occurrence within each subject are reported in Supplementary Table S8.

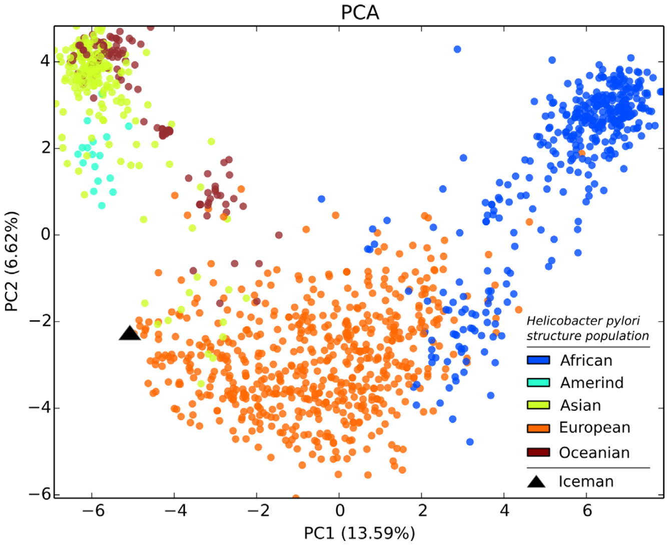

Figure 5. PCA plot of the public available MLST types for *Helicobacter pylori*, colored by aggregated structure population. The concatenated MLST-loci sequences of those *H. pylori* isolates that could be associated with a structure population in the PubMLST database (5) were analyzed in PCA space. The colors represent continental groups of structure populations; the black triangle indicates the Iceman metagenomically-inferred MLST type. The ancient *H. pylori* reconstructed loci are at the boundaries between European (orange) and Asian (yellow green) types. Values in brackets of PC1 and PC2 represent the amount of explained variance of the PCA analysis.

tories and large databases for the most relevant human (opportunistic) pathogens are available (5,6). This approach remains however very time-consuming especially due the need of isolate and cultivate each target species separately.

Here we present MetaMLST, a fast approach for microbial strain tracking and identification applicable directly to complex metagenomic samples taken directly from a given environment. MetaMLST takes advantage of the publicly available information on MLST alleles and profiles, which can be used to guide epidemiological and clinical investigation. Importantly, the tool can be extended to recently proposed MLST protocols comprising more loci (e.g. eMLST (40) and rMLST (41)). We validated MetaMLST on semi-synthetic metagenomes, and were able to identify all microbial targets with >95% identity (>99.85% with spiked-in isolates in real metagenomes at 3× coverage), substantially improving the performances achievable with a metagenomics assemblies-based analysis. Compared to assembly-free strain profiling (32,42–44) our method is the first estimating MLST profiles and it enables rapid analysis of metagenomics samples. MetaMLST processed the metagenomes at an average speed of ∼35 000 reads/sec/CPU which makes it suitable for the analysis of very large metagenomic datasets. Moreover, the method takes advantage of the publicly available MLST sequences, profiles and isolates, which are orders of magnitude more available than reference genomes. MetaMLST is designed to identify the most abundant strain of each trackable (i.e. MLST referenced) microbial species, which is normally the case in traditional MLST analysis, where the typing is performed on a single isolated microbial colony. Strains from the same species are rarely comparably abundant in a sample, but MetaMLST would provide noisy results in those cases. The MetaMLST method, however, is faster and potentially cheaper in the long term, as it does not require the time consuming isolation of each microbe.

We applied MetaMLST to hundreds of gastrointestinal, oral and skin metagenomes to highlight the potentialities of the approach. Using this method we have already shown body site specificity and tropism in oral *Neisseria spp.* (45), and here we see a similar ecological behavior for skin associated *S. epidermidis* populations, in agreement with previous studies (36). Moreover, MetaMLST provided insights on the pathogenic *E. coli* strain (ST-678) when applied to gut metagenomes from patients of the Germany 2011 outbreak, and showed the longitudinal persistence of uropathogenic and necrotizing enterocolitis-related *E. coli* strains among preterm-born infants (46). Interestingly, the pipeline was able to identify an ancient *H. pylori* strain from the metagenomes of a 5300 years old mummy, extracted from Maixner *et al.*, confirming the result of the original work (20). These results strongly confirm the effectiveness of MetaMLST, which allows accurate microbial strain typing without the need of isolation and cultivation, directly from metagenomic samples.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement [PCIG13-GA-2013-618833 to N.S.]; Centre for Integrative Biology, University of Trento startup funds (to N.S.); MIUR 'Futuro in Ricerca' [RBFR13EWWI_001 to N.S.]; Leo Pharma Foundation (to N.S.); Fondazione CARITRO Fellowship Rif.Int.2013.0239 (to N.S.); Fondazione CARITRO Fellowship Rif.int.2014.0325 (to A.T.); 'Terme di Comano' (to N.S.). Funding for open access charge: MIUR 'Futuro in Ricerca' [RBFR13EWWI_001].
Conflict of interest statement. None declared.

REFERENCES

1. Maiden,M.C., Bygraves,J.A., Feil,E., Morelli,G., Russell,J.E., Urwin,R., Zhang,Q., Zhou,J., Zurth,K., Caugant,D.A. *et al.* (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 3140–3145.
2. Zhang,R., Gu,D.X., Huang,Y.L., Chan,E.W., Chen,G.X. and Chen,S. (2016) Comparative genetic characterization of Enteroaggregative Escherichia coli strains recovered from clinical and non-clinical settings. *Sci. Rep.*, **6**, e24321.
3. Fazeli,H., Sadighian,H., Esfahani,B.N. and Pourmand,M.R. (2015) Genetic characterization of Pseudomonas aeruginosa-resistant isolates at the university teaching hospital in Iran. *Adv. Biomed. Res.*, **4**, e156.
4. Bougnoux,M.E., Tavanti,A., Bouchier,C., Gow,N.A., Magnier,A., Davidson,A.D., Maiden,M.C., D'Enfert,C. and Odds,F.C. (2003) Collaborative consensus for optimized multilocus sequence typing of Candida albicans. *J. Clin. Microbiol.*, **41**, 5265–5266.
5. Jolley,K.A. and Maiden,M.C. (2010) BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*, **11**, e595.
6. Aanensen,D.M. and Spratt,B.G. (2005) The multilocus sequence typing network: mlst.net. *Nucleic Acids Res.*, **33**, W728–W733.
7. Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D., Paulsen,I., Nelson,K.E., Nelson,W. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.
8. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
9. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
10. McElroy,K.E., Luciani,F. and Thomas,T. (2012) GemSIM: general, error-model based simulator of next-generation sequencing data. *BMC Genomics*, **13**, e74.
11. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
12. Segata,N., Waldron,L., Ballarini,A., Narasimhan,V., Jousson,O. and Huttenhower,C. (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods*, **9**, 811–814.
13. Truong,D.T., Franzosa,E.A., Tickle,T.L., Scholz,M., Weingart,G., Pasolli,E., Tett,A., Huttenhower,C. and Segata,N. (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods*, **12**, 902–903.
14. Francisco,A.P., Vaz,C., Monteiro,P.T., Melo-Cristino,J., Ramirez,M. and Carrico,J.A. (2012) PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. *BMC Bioinformatics*, **13**, e87.
15. Francisco,A.P., Bugalho,M., Ramirez,M. and Carrico,J.A. (2009) Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. *BMC Bioinformatics*, **10**, e152.
16. Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.
17. Asnicar,F., Weingart,G., Tickle,T.L., Huttenhower,C. and Segata,N. (2015) Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ*, **3**, e1029.
18. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R. and Dubourg,V. (2011) Scikit-learn: machine learning in python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
19. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
20. Maixner,F., Krause-Kyora,B., Turaev,D., Herbig,A., Hoopmann,M.R., Hallows,J.L., Kusebauch,U., Vigl,E.E., Malfertheiner,P., Megraud,F. *et al.* (2016) The 5300-year-old Helicobacter pylori genome of the Iceman. *Science*, **351**, 162–165.
21. Dodgson,A.R., Pujol,C., Denning,D.W., Soll,D.R. and Fox,A.J. (2003) Multilocus sequence typing of Candida glabrata reveals geographically enriched clades. *J. Clin. Microbiol.*, **41**, 5709–5717.
22. Bain,J.M., Tavanti,A., Davidson,A.D., Jacobsen,M.D., Shaw,D., Gow,N.A. and Odds,F.C. (2007) Multilocus sequence typing of the pathogenic fungus Aspergillus fumigatus. *J. Clin. Microbiol.*, **45**, 1469–1477.
23. H.M.P. Consortium (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214.
24. Howe,A. and Chain,P.S.G. (2015) Challenges and opportunities in understanding microbial communities with metagenome assembly (accompanied by IPython Notebook tutorial). *Front. Microbiol.*, **6**, e678
25. Bankevich,A., Nurk,S., Antipov,D., Gurevich,A.A., Dvorkin,M., Kulikov,A.S., Lesin,V.M., Nikolenko,S.I., Pham,S., Prjibelski,A.D. *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, **19**, 455–477.
26. Li,D., Luo,R., Liu,C.M., Leung,C.M., Ting,H.F., Sadakane,K., Yamashita,H. and Lam,T.W. (2016) MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods*, **102**, 3–11.
27. Desai,A., Marwah,V.S., Yadav,A., Jha,V., Dhaygude,K., Bangar,U., Kulkarni,V. and Jere,A. (2013) Identification of optimum sequencing depth especially for de novo genome assembly of small genomes using next generation sequencing data. *PLoS One*, **8**, e60204.

28. Qin,J., Li,Y., Cai,Z., Li,S., Zhu,J., Zhang,F., Liang,S., Zhang,W., Guan,Y., Shen,D. *et al.* (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, **490**, 55–60.
29. Loman,N.J., Constantinidou,C., Christner,M., Rohde,H., Chan,J.Z., Quick,J., Weir,J.C., Quince,C., Smith,G.P., Betley,J.R. *et al.* (2013) A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104:H4. *JAMA*, **309**, 1502–1510.
30. Gordon,D.M., Clermont,O., Tolley,H. and Denamur,E. (2008) Assigning Escherichia coli strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method. *Environ. Microbiol.*, **10**, 2484–2496.
31. Wirth,T., Falush,D., Lan,R., Colles,F., Mensa,P., Wieler,L.H., Karch,H., Reeves,P.R., Maiden,M.C. and Ochman,H. (2006) Sex and virulence in Escherichia coli: an evolutionary perspective. *Mol. Microbiol.*, **60**, 1136–1151.
32. Scholz,M., Ward,D.V., Pasolli,E., Tolio,T., Zolfo,M., Asnicar,F., Truong,D.T., Tett,A., Morrow,A.L. and Segata,N. (2016) Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat. Methods*, **13**, 435–438.
33. Ahmed,S.A., Awosika,J., Baldwin,C., Bishop-Lilly,K.A., Biswas,B., Broomall,S., Chain,P.S., Chertkov,O., Chokoshvili,O., Coyne,S. *et al.* (2012) Genomic comparison of Escherichia coli O104:H4 isolates from 2009 and 2011 reveals plasmid, and prophage heterogeneity, including shiga toxin encoding phage stx2. *PLoS One*, **7**, e48228.
34. Monecke,S., Mariani-Kurkdjian,P., Bingen,E., Weill,F.X., Baliere,C., Slickers,P. and Ehricht,R. (2011) Presence of enterohemorrhagic Escherichia coli ST678/O104:H4 in France prior to 2011. *Appl. Eviron. Microbiol.*, **77**, 8784–8786.
35. Touchon,M., Hoede,C., Tenaillon,O., Barbe,V., Baeriswyl,S., Bidet,P., Bingen,E., Bonacorsi,S., Bouchier,C., Bouvet,O. *et al.* (2009) Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS Genet.*, **5**, e1000344.
36. Oh,J., Byrd,A.L., Deming,C., Conlan,S., Kong,H.H. and Segre,J.A. (2014) Biogeography and individuality shape function in the human skin metagenome. *Nature*, **514**, 59–64.
37. Karch,H., Denamur,E., Dobrindt,U., Finlay,B.B., Hengge,R., Johannes,L., Ron,E.Z., Tønjum,T., Sansonetti,P.J. and Vicente,M. (2012) The enemy within us: lessons from the 2011 European Escherichia coli O104:H4 outbreak. *EMBO Mol. Med.*, **4**, 841–848.
38. Maiden,M.C. (2006) Multilocus sequence typing of bacteria. *Annu. Rev. Microbiol.*, **60**, 561–588.
39. Perez-Losada,M., Cabezas,P., Castro-Nallar,E. and Crandall,K.A. (2013) Pathogen typing in the genomics era: MLST and the future of molecular epidemiology. *Infect. Genet. Evol.*, **16**, 38–53.
40. Zhang,J., Kong,Y., Ruan,Z., Huang,J., Song,T., Song,J., Jiang,Y., Yu,Y. and Xie,X. (2014) Correlation between Ureaplasma subgroup 2 and genitourinary tract disease outcomes revealed by an expanded multilocus sequence typing (eMLST) scheme. *PLoS One*, **9**, e104347.
41. Jolley,K.A., Bliss,C.M., Bennett,J.S., Bratcher,H.B., Brehony,C., Colles,F.M., Wimalarathna,H., Harrison,O.B., Sheppard,S.K., Cody,A.J. *et al.* (2012) Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology*, **158**, 1005–1015.
42. Francis,O.E., Bendall,M., Manimaran,S., Hong,C., Clement,N.L., Castro-Nallar,E., Snell,Q., Schaalje,G.B., Clement,M.J., Crandall,K.A. *et al.* (2013) Pathoscope: species identification and strain attribution with unassembled sequencing data. *Genome Res.*, **23**, 1721–1729.
43. Ahn,T.H., Chai,J. and Pan,C. (2015) Sigma: strain-level inference of genomes from metagenomic analysis for biosurveillance. *Bioinformatics*, **31**, 170–177.
44. Franzosa,E.A., Huang,K., Meadow,J.F., Gevers,D., Lemon,K.P., Bohannan,B.J. and Huttenhower,C. (2015) Identifying personal microbiomes using metagenomic codes. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E2930–E2938.
45. Donati,C., Zolfo,M., Albanese,D., Truong,D.T., Asnicar,F., Iebba,V., Cavalieri,D., Jousson,O., De Filippo,C. and Huttenhower,C. (2016) Uncovering oral Neisseria tropism and persistence using metagenomic sequencing. *Nat. Microbiol.*, **1**, e16070.
46. Ward,D.V., Scholz,M., Zolfo,M., Taft,D.H., Schibler,K.R., Tett,A., Segata,N. and Morrow,A.L. (2016) Metagenomic sequencing with strain-level resolution implicates uropathogenic E. coli in necrotizing enterocolitis and mortality in preterm infants. *Cell Rep.*, **14**, 2912–2924.