HOUND: A NOVEL TOOL FOR AUTOMATED MAPPING OF GENOTYPE TO PHENOTYPE IN BACTERIAL GENOMES ASSEMBLED DE NOVO

CARLOS REDING\(^\ast\), NAPHTA SATAPOOMIN\(^1\), MATTHEW B. AVISON\(^\ast\)

\(^1\)University of Bristol School of Cellular and Molecular Medicine, University Walk, Bristol, BS8 1TD Bristol, UK.

\(^\ast\)Corresponding author(s): carlos.reding@bristol.ac.uk

Increasing evidence suggests that microbial species have a strong within species genetic heterogeneity. This can be problematic for the analysis of prokaryote genomes, which commonly relies on a reference genome to guide the assembly process; any difference between reference and sample genomes can introduce errors in the detection of small insertions, deletions, structural variations and even point mutations. This phenomenon jeopardises the genomic surveillance of antibiotic-resistant bacteria, triggering even a reproducibility crisis. Here we present Hound, an analysis pipeline that integrates publicly available tools to locally assemble prokaryote genomes \textit{de novo}, detect genes by similarity, and report the mutations found. Three features are exclusive to Hound: It reports relative gene copy number, retrieves sequences upstream the start codon to detect mutations in promoter regions, and, importantly, can merge contigs based on a user-given query sequence to reconstruct genes that are fragmented by the assembler. To demonstrate Hound, we screened through 5,032 bacterial whole-genome sequences isolated from farm animals and clinical patients using the amino acid sequence of \textit{bla}\textsubscript{TEM-1}, to predict resistance to amoxicillin-clavulanate. We believe this tool can facilitate the analysis of prokaryote species that currently lack a reference genome, and can be scaled up to build automated diagnostic systems.

I. INTRODUCTION

The advent of affordable genome sequencing has exposed the wide genetic heterogeneity that exists within microbial species [1]. With genome sizes that range between 2.69–2.92 Mb in \textit{Staphylococcus aureus}, or between 4.66–5.30 Mb for \textit{Escherichia coli}, it is not surprising that some begin to question the notion of \textit{species} [2, 3] or even \textit{clone} [4] in prokaryotes. This heterogeneity led to the concept of \textit{pan-genomes} [5], but it also exposes another, more technical problem: How to study the genome of prokaryotes without masking this genetic diversity?

Raw sequencing data are typically mapped onto a high-quality reference—whose sequence is known and resolved (i.e. circularised) [6, 7]—or databases containing them [8], to study the genetics of organisms from virus [9] to vertebrates [10] or plants [11]. The use of reference-mapped assemblies is used in comparative genomics [12], clinical microbiology [13], public health [14, 15], and even to inform policy making through the detection of specific mutations or phylogenetic analyses [16, 17]. Now, given the further reduction in sequencing costs, reference-mapped assemblies are increasingly used to predict antibiotic susceptibility in the clinic [18, 19]. And this can be problematic, given
the limitations of these type of assemblies to detect antibiotic-resistance genes. On the one hand, reads that cannot be mapped onto the reference genome, say, because they are plasmid-borne and not part of the chromosome, are excluded from the assembly. And this loss of data hinders the detection of antibiotic-resistance genes [19, 20]. On the other hand, the availability of reference genomes is skewed towards the most common pathogens [13, 16], limiting the study of other pathogens [21]. Consequently, the scope of tools like ResFinder [22], STARR [23], ARG-ANNOT [24], RAST [25], or ABRiicate [26] can be very limited. Particularly, because they rely on reference genomes to report the presence—or not—of antibiotic-resistance genes along with mutations in the coding sequence known to be associated with a resistant phenotype. As we show below, mapping sequencing data onto a reference genome can artificially modify the assembly [20]. This approach is not only limited for the study of other pathogens or the finding of novel, undocumented mutations; but of species that may have other biological or ecological importance where reference genomes and tools are scarce [27].

Here we sought to build a pipeline to analyse prokaryotes genomes assembled de novo, without using a reference to guide the assembly process. De novo assemblies lack most of the limitations mentioned above, but can also introduce others. Particularly, the fragmentation of genes—whose sequences are split across multiple contigs by the assembler [19]. Hound implements an algorithm to re-purpose a query sequence as a local reference to detect and merge the relevant contigs, so that its sequence can be reconstructed unambiguously. Another issue we tackle is that resistance is not only given by the presence, or otherwise, of specific genes. The over-expression of genes that protect against one antibiotic, whether through mutations in the promoter or increase in relative copy number, can provide resistance to multiple drugs. For example, amoxicillin-resistant *Escherichia coli* are most commonly resistant due to the production the TEM β-lactamase enzyme, encoded by the mobile gene *bla*TEM [28]. Amoxicillin-clavulanate will kill amoxicillin-resistance *E. coli* because clavulanate inhibits TEM, and this combination as been widely used in human [29, 30] and veterinary medicine [31]. However, *E. coli* can become resistant to amoxicillin-clavulanate by over-producing TEM due to promoter mutations [29] or increased gene copy number [32]. Therefore, we built into Hound the capability to retrieve sequences beyond a gene's coding sequence and include the promoter, as well as the relative gene copy number, to allow the detection of such variants with our pipeline.

II. RESULTS

**Pipeline overview.** Hound integrates tools widely-used to assemble Nanopore and Illumina reads de novo, and screen the resulting assemblies for user-given query sequences, into a single tool. Hound supports nucleotide and amino acid sequences, but we suggest the query to be an amino acid sequence where possible to avoid variations introduced by synonymous mutations. Our pipeline is modular as Figure 1 illustrates to allow performing only a subset of the tasks, and relies on SPAdes [33] as its backend assembler due to its combination of speed, accuracy, and support for sequencing data from multiple
Figure 1. Description of the Hound pipeline for the analysis of prokaryote genomes assembled de novo. While Hound can be run in a single step, the user is given three steps of granularity. 1) The first step is to assemble the quality-filtered FASTQ files and depth of coverage data generated. During this step, each assembly is converted into a BLAST database to facilitate downstream analyses. Note that once assembled, there is no need to repeat the step—whence the choice of granularity. 2) In the second step, Hound will search by similarity any user-given target(s) in the assemblies generated in 1). The sequence(s) must be in FASTA format and preferably be the amino acid sequence to avoid variations introduced by synonymous mutations. Files with multiple entries are supported. To compute the relative gene copy number (RCN), Hound will use a number of house-keeping genes to compute a baseline depth of coverage. We used four but there is no limit in number of house-keeping genes. If the identity of the sequence found in the assembly with respect to the target is above 90%, and the contigs harbouring subsets of the query have...
overlapping common sequences, Hound will shortlist the contigs for its contig-merging routine. Here, after sorting them first by identity and length, Hound will iteratively run pairwise alignments between the first pair with overlapping coordinates, discard one of the two overlapping sequences to avoid introducing duplications, and compare the resulting sequence to the user-given query. Hound will process the next contig in the list and add it to the previous merged sequence until the reconstructed sequence has, at least, equal length to the query which will have at least $\text{MIN_ID}$ identity. Note all these contigs will have overlapping coordinates that are located at the boundaries of the contigs, thus, only genes truncated by the assembler and not by insertion sequences—mobile elements—will be included in this analysis.

At this point, once the query sequence can be reconstructed, the assembly is re-written with the new contig name being a concatenation of all the founding contigs (i.e. >NODE_1+NODE_43+NODE_24). If coverage data exists, the coordinates used earlier by the merging routine are applied to this dataset to preserve coverage in the new, merged contig.

**Reference-mapped vs de novo assemblies.** The first step in the screening of our 5,032 farm animal and human clinical samples, is the assembly *de novo* of the reads. A first look at the output of the pipeline reveals the assemblies seldom have the same size (Figure 2A). Now, while *E. coli* is by far the most abundant species in our dataset, as identified by Kraken 2 [37], the dataset also contains isolates of *Klebsiella pneumoniae* (15.7% of the total), and one isolate of *Pseudomonas aeruginosa* (<1%) and *Salmonella enterica* (<1%) particularly across the human clinical isolates. This means multiple species can be monitored depending on each use case.

When we removed all non-*E. coli* from the dataset and compared the assembly sizes, we found the variation with respect to the available reference genomes—K-12 MG1655 and 0157:h7 Sakai—to be substantial as the histogram in Figure 2A shows, with most assemblies having sizes in-between these references. These two genomes are the only ones validated by the National Center for Biotechnology Information (NCBI) and flagged as reference genomes accordingly. The variability that we observed in genome sizes is consistent with the aforementioned notion of pan-genomes. Next, we compared reference-mapped assemblies using to different *E. coli* genomes using `--assemble --reference $\text{REF_GENOME}` with their respective reference—beyond the above MG1655 and 0157:h7 genomes we also used that for enterotoxigenic *E. coli* (ETEC)—as well as the *de novo* assemblies, using pairwise alignments with MUMmer [38].

When comparing the *de novo* assemblies to the references, this alignment revealed signatures of small deletions, insertions, and repeated regions that were different with each reference used (Figure 2B). However, the signatures vanished when we used reference-mapped assemblies (Figure 2C). This means the use of references discards or includes details from the final assembly that can ultimately alter its size—unique and robust when assembled *de novo*—and help explain the inconsistent results [39] that occur when comparing different tools.
**Figure 2. Variability of within-species genome size is not captured by the canonical use of reference genomes.** A) Histogram (top) and distribution (bottom) of genome sizes across 3562 farm and clinical E. coli isolates. The canonical size of the reference wild-type (K-12 substr. MG1655, genome GCF_000005845.2 in the NCBI) and Shiga toxin-producing strain (0157:H7 str. Sakai, genome GCA_000008865.2 in the NCBI) are marked at approx. 4.64 Mb and 5.39 Mb in red-toned vertical lines. B–C) Dotplots of one representative isolate to visualise the genome-genome sequence alignment between *de novo* (B) or reference-mapped (C) assemblies and three different known genomes: MG1655, O157:H7, and enterotoxigenic E. coli (ETEC). Sequences present in these references, but not in our isolate, are noted by a horizontal gap in the dotplot, whereas the converse is noted by a vertical gap. Note that in C) the reported assembly size of our isolate, in the y–axis, varies depending on the choice of reference–size is constant in those genomes assembled *de novo*.

**Monitoring blaTEM-1 in farm and clinical isolates.** We used this tool to screen through 5,032 whole-genome sequencing (WGS) datasets, from farm animals (2,494) and human clinical isolates (2,538), to detect *blaTEM-1*, its promoter region, and relative copy number. Among the output files generated by Hound, there is a figure with the multiple alignment and phylogeny of the sequences, containing the aforementioned metadata depending upon the flags provided, that can be produced with the flag --plot $FILENAME. This is an exemplar of how this tool can improve the prediction and, therefore, the surveillance of antibiotic resistance with complex phenotypes that are notoriously difficult to predict from mutations in the coding sequence alone.
The result shows 39.16% (n=994) of clinical isolates were bluTEM-1 positive compared to 51.84% (n=1,283) in those from farm animals (Figures 3B and C). It is noteworthy to mention that human clinical isolates came from routine surveillance of gram-negative bacteria and include multiple species beyond E. coli that less commonly carry bluTEM-1.
whereas those from farm animals were *E. coli* isolates sequenced because of their resistant phenotype [40]. Interestingly however, only 22.52% (n=289) of the isolates from farmed animals harboured 2 or more copies of the gene. This contrast with data from the clinical isolates, where 39.33% of the *bla*TEM-1 positive isolates (n=391) harboured 2 or more copies. Since increased gene copy number is associated with amoxicillin-clavulanate resistance [32], this would fit with a more intense use of this combination to treat humans. Moreover, as Figures 2A and B illustrate, in some isolates *bla*TEM-1 number is in the hundreds. While their occurrence is rare—n=6 in farm isolates, n=31 in clinical isolates—it suggests the circulation of one or more plasmids with very high copy number.

Multicopy plasmids with hundreds of copies are not unheard of [41]). Using the flag

```
--promoter --cutoff 250
```

we used Hound to retrieve the promoter of *bla*TEM-1 and flag any mutation found, annotated in the figure produced as black dots which are more common in isolates from humans as Figures 3B and C illustrate. Mutations found by Hound with respect to the query sequence used, as well as relative copy number and other metadata can be exported into a spreadsheet by using the flag

```
--summary $SPREADSHEET
```

which can then be parsed to highlight isolates with specific mutations. An illustrative spreadsheet is included as a supplementary table.

A useful feature of Hound is its ability to populate iteratively the same assemblies to look for different genes, allowing the simultaneous detection of genes, calculation of their relative copy number, or re-analysis of prior data. Figure 3C shows two heatmaps illustrating the detection of different β-lactamases, nitroreductases, and efflux pumps in both farm and clinical isolates as well as any mutations found in their coding sequence and relative copy number.

III. DISCUSSION

An increasingly problematic issue, particularly in the detection of antimicrobial-resistance genes, is the lack of reproducibility [42, 43]. The choice of reference genomes is typically opaque to the user beyond the species, being notoriously difficult to underpin the exact assembly used. Along with the variety of existing pipelines that yield inconsistent results between laboratories [39], this problem led to the suggestion of standardised, ISO-certified pipelines [39]. Here we argue that they still fail to detect antibiotic-resistant variants driven by the over-expression of enzymes like β-lactams given their inability to account for variations in gene expression and copy number. Now, beyond the detection of antibiotic-resistant mutants, if we found different genetic signatures when comparing our assemblies to different references it is not unreasonable to think this will also be the case for other reference-mapped assemblies. Thus, using a standardised reference is unlikely to avoid this problem—exacerbated by the scarcity of tools to analyse *de novo* assemblies.

Hound is a step towards facilitating the analysis of these assemblies, not only by addressing a key limitation—gene fragmentation—but also by reducing the knowledge and technological burdens. The fact that we assembled and analysed more than 5,000 genomes on an 8-core, 16GB RAM laptop over the span of 8–9 days shows the potential
for Hound to be implemented in larger and more powerful infrastructures for monitoring and diagnostic purposes. Now, Hound has some limitations. For example, it cannot report whether a gene of interest sits within a genomic island, but it can be used to detect whether genes associated to such islands are in the same contig as the gene of interest—complementing other bespoke analyses. Another limitation is that it currently supports short-read Illumina sequencing data only due to its availability during the development of this tool. However, given our use of SPAdes as backend assembler it is possible to add support for long-read nanopore and PacBio sequencing data in future releases.

The use of de novo assemblies means Hound is not only agnostic with respect to what genes can be monitored. It is also independent of the microbial species analysed—not possible with the use of reference genomes. With the majority of prokaryote diversity still being unknown and unsequenced [44, 45], we believe that Hound can be a useful tool to study non-model microorganisms that lack any reference—and help build them iteratively thanks to its contig-merging routine when sequencing costs in other platforms increase.

IV. METHODS

Genome sequencing and assembly. Genomic DNA libraries from isolate bacteria are prepared using the Nextera XT Library Prep Kit (Illumina, San Diego, USA) following the manufacturer's protocol with the following modifications: Input DNA is increase 2-fold, and PCR elongation time is increased to 45 seconds. DNA quantification and library preparation are carried out on a Hamilton Microlab STAR automated liquid handling system (Hamilton Bonaduz AG, Switzerland), and the libraries sequenced on a Illumina NovaSeq 6000 (Illumina, San Diego, USA) using a 250 paired-end protocol by MicrobesNG. Reads are adapter-trimmed using Trimmomatic 0.30 [46] with a sliding window quality cutoff of Q15.

The flag --preprocess reads.zip processes the reads provider by MicrobesNG to create the directory structure required by Hound, with paired reads being stored in $DIR/reads/, and assemblies in $DIR/assemblies/de_novo/ for reads assembled de novo or $DIR/assemblies/reference-mapped/ for those mapped to a reference. This will depend on whether --assembly --de-novo or --assembly --reference $REF_GENOME have been passed. Hound then assembles de novo the resulting reads using SPAdes with the --isolate flag and k-mer size of 127, given the sequencing platform and protocol. For reference-mapped assemblies, Hound aligns the reads to a user-given reference using the Burrows-Wheeler Aligner and SAMtools with standard parameters. The coverage depth for all assemblies is then calculated with SAMtools if the flags --coverage and --hk-genes $HK_GENES are given, the baseline depth being the median coverage depth of all loci included in $HK_GENES, faster than computing the median coverage of the whole assembly to avoid any bias introduced by plasmid carriage. The relative copy number (RCN) is then calculated as the coverage depth of all loci in $TARGET_GENES divided by the baseline coverage [47].
Indexing of assemblies. The resulting assemblies are indexed using `makeblastdb` from BLAST+, with flags `-parse_seqids` and `-dbtype nucl`, to facilitate the search of the sequences in file $TARGET_GENES$. The search is run with `blastn`, which uses nucleotide sequences, or `tblastn` depending on whether the flag `--nucl` is passed to Hound. Without this flag, Hound assumes that the file $TARGET_GENES$ contains amino acid sequences and will therefore use `tblastn`.

Multiple alignment and phylogeny. When the flag `--phylo` is passed, Hound will use muscle 3.80 to align the target gene sequences found in all assemblies given its accuracy and speed [48]. Penalties for the introduction and extension of gaps are pre-set with a value of $-9950.0$ to avoid excessive fragmentation of the alignment. This alignment is then used by Hound to generate a phylogeny with PhyML [49] with seed 100100 for repeatability.

REFERENCES

(2011).
612 (2012).
affordable mycobacterial diagnosis with whole-genome sequencing: a prospective
15. Hendriksen, R. S., Bortolaia, V., Tate, H., Tyson, G. H., Aarestrup, F. M. & McDermott, P. F. Using genomics to track global antimicrobial resistance. Front. Public
Health 7, 242 (2019).
16. Sichtig, H., Minogue, T., Yan, Y., et al. FDA-ARGOS is a database with public quality-
17. GLASS whole-genome sequencing for surveillance of antimicrobial resistance World
prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a ret-
20. Valiente-Mullor, C., Beamud, B., Ansari, I., et al. One is not enough: on the effects
of reference genome for the mapping and subsequent analyses of short-reads. PLoS
22. Zankari, E., Hasman, H., Cosentino, S., et al. Identification of acquired antimicro-
23. De Man, T. J. & Limbago, B. M. SSTAR, a stand-alone easy-to-use antimicrobial
resistance gene predictor. mSphere 1, 10–1128 (2016).
bioinformatic tool to discover antibiotic resistance genes in bacterial genomes.


**Competing interests.** The authors have declared no competing interests.

**Acknowledgements.** This work was funded by an Arwain DGC grant from the Welsh government and FARM-SASFE (BBSRC grant BB/T004592/1) grant awarded to MA.

**Code availability statement** A Python implementation of Hound can be found in https://gitlab.com/rc-reding/software/