High concordance between short and long read sequencing for genomics-based species identification and antimicrobial resistance

Jason D. Wittenbach, Paul Knysh, Grace Cox, Ian C. Herriott, Nicole Billings, Tim Farrell, Linnea Sahlberg, Cabell Maddux, Miriam H. Huntley

Day Zero Diagnostics, Boston, MA 02135

Introduction

Traditional laboratory techniques for the diagnosis of bacterial infections, consisting of species identification (ID) and antimicrobial susceptibility testing (AST), require time intensive culturing and phenotyping steps which can take days, delaying appropriate therapy during a critical time in patient care. The availability of high quality and low-cost rapid DNA sequencing -- as provided by recent advances in nanopore sequencing -- has the potential to transform infectious disease diagnosis with the use of whole genome sequencing (WGS). In previous work, we built a bioinformatics pipeline for species ID (Keynome ID) and a machine learning system for genomic AST prediction (Keynome *g*-AST) for performing these tasks from WGS inputs; when paired with our sample preparation technology this process provides pathogen ID and AST diagnosis from whole blood samples in hours instead of days.

Here we assess the differences in performance of these algorithms when Illumina short-read versus ONT long-read WGS data is used as input. Bacterial isolate strains across multiple species were selected based on phenotypic and genotypic diversity and genomic DNA was sequenced on both platforms. We report a high degree of concordance for ID (99.4%) and AST (97.7%) between the two sequencing platforms, demonstrating the suitability of ONT sequencing to support such applications.

However, it should be noted that on the *g*-AST task, for all but one of the small number of discordant predictions, the prediction from Illumina sequencing was correct and that from ONT was incorrect when compared to ground truth, suggesting further improvements in long-read accuracy would still be beneficial. We are currently assessing whether Guppy v5 or other rapid error correction methods could bridge this remaining gap.

Methods

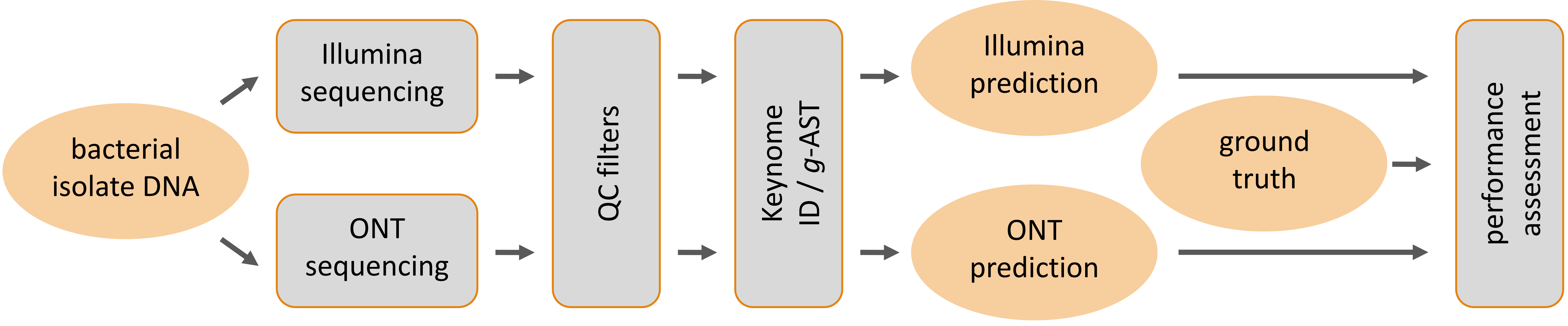
Sample selection: For ID we selected 3 - 5 bacterial strains from each of 50 species; for *g*-AST we selected 10 strains from each of 10 species. In each case, we developed a custom selection algorithm to select a diverse set of strains for each species – for ID it maximized phenotypic diversity with respect to antimicrobial resistance/susceptibility (AMR/S) profiles across all drugs for which AST results were available. For *g*-AST, the objective was extended to also maximize genomic diversity as well as the number of AST results available per strain.

Sample processing: Isolates of all strains were cultured and DNA extracted. This DNA was then prepped, sequenced, and base called on both Illumina (short read) and ONT (long read) platforms. Basic QC metrics to ensure sufficient sequencing yield, quality, and lack of contamination were computed and samples that failed to meet these criteria were reprocessed as appropriate.

Prediction and performance evaluation: For species ID, the Keynome ID algorithm predicted a single species for each sample which was compared to the phenotypically determined species. For *g*-AST, Keynome *g*-AST used the known species identity to make binary AMR/S predictions (susceptible versus not susceptible) for all drugs where the model had previously exhibited high performance.

Notes:

- Keynome *g*-AST models are trained on Illumina data, making this concordance task a critical assessment of the models' ability to generalize to long-read data.
- The Guppy v3 base caller was used for the ID assessment, while Guppy v4 with an additional error correction step (via *canu*) was used for AST assessment



Results: species ID

We analyzed 168 strains across 50 species, and found that paired species ID predictions from long- and short-read sequencing were 99.4% concordant (167/168 samples). The lone error came from a sample that was predicted to be *Enterobacter cloacae* with Illumina sequencing but *Enterobacter hormaechei* with ONT sequencing. These two species both belong to the "*Enterobacter cloacae* complex" or closely related organisms, making the predictions concordant at the complex level, though not at the species level.

	Species	Total strains analyzed	Illumina-ONT concordant strains (% total)
1	*Acinetobacter baumannii*	3	3 (100%)
2	*Acinetobacter ursingii*	3	3 (100%)
3	*Citrobacter freundii*	3	3 (100%)
4	*Citrobacter koseri*	4	4 (100%)
5	*Enterobacter aerogenes*	3	3 (100%)
6	*Enterobacter cloacae* complex[5]	4	3 (75%)
7	*Enterococcus avium*	3	3 (100%)
8	*Enterococcus casseliflavus*	3	3 (100%)
9	*Enterococcus faecalis*	4	4 (100%)
10	*Enterococcus faecium*	3	3 (100%)
11	*Enterococcus gallinarum*	4	4 (100%)
12	*Enterococcus raffinosus*	3	3 (100%)
13	*Escherichia coli*	3	3 (100%)
14	*Haemophilus influenzae*	3	3 (100%)
15	*Klebsiella oxytoca*	3	3 (100%)
16	*Klebsiella pneumoniae*	3	3 (100%)
17	*Listeria monocytogenes*	3	3 (100%)
18	*Morganella morganii*	3	3 (100%)
19	*Pantoea agglomerans*	5	5 (100%)
20	*Pasteurella multocida*	3	3 (100%)
21	*Propionibacterium acnes*	3	3 (100%)
22	*Proteus mirabilis*	3	3 (100%)
23	*Pseudomonas aeruginosa*	3	3 (100%)
24	*Pseudomonas putida*	4	4 (100%)
25	*Raoultella ornithinolytica*	3	3 (100%)
26	*Salmonella enterica*	4	4 (100%)
27	*Serratia liquefaciens*	3	3 (100%)
28	*Serratia marcescens*	4	4 (100%)
29	*Staphylococcus aureus*	4	4 (100%)
30	*Staphylococcus capitis*	4	4 (100%)
31	*Staphylococcus caprae*	3	3 (100%)
32	*Staphylococcus epidermidis*	4	4 (100%)
33	*Staphylococcus haemolyticus*	5	5 (100%)
34	*Staphylococcus hominis*	3	3 (100%)
35	*Staphylococcus lugdunensis*	4	4 (100%)
36	*Staphylococcus saprophyticus*	3	3 (100%)
37	*Staphylococcus simulans*	4	4 (100%)
38	*Staphylococcus warneri*	3	3 (100%)
39	*Stenotrophomonas maltophilia*	3	3 (100%)
40	*Streptococcus agalactiae*	3	3 (100%)
41	*Streptococcus anginosus*	3	3 (100%)
42	*Streptococcus constellatus*	3	3 (100%)
43	*Streptococcus dysgalactiae*	3	3 (100%)
44	*Streptococcus intermedius*	3	3 (100%)
45	*Streptococcus mutans*	3	3 (100%)
46	*Streptococcus parasanguinis*	3	3 (100%)
47	*Streptococcus pneumoniae*	4	4 (100%)
48	*Streptococcus pyogenes*	3	3 (100%)
49	*Streptococcus salivarius*	4	4 (100%)
50	*Streptococcus sanguinis*	3	3 (100%)

Results: *g*-AST

We analyzed *g*-AST predictions from 35 models for species-drug combinations across 9 species and 15 unique antimicrobial agents, making predictions on 10 strains per species for all drugs where a high-performing model was available. This resulted in a total of 350 unique predictions, which were 97.7% (342/350 predictions) concordant between long- and short-read sequencing data.

Discordant predictions did show some tendency to cluster by species-drug combination, indicating that some of the discordance might be related to specific models failing to generalize. Though, when compared to ground truth, 7 or the 8 discordant predictions showed correct Illumina predictions and incorrect ONT predictions.

per species results

Species	Strains analyzed	Models analyzed	Total predictions	Number concordant	Percent concordant
Acinetobacter baumannii	10	3	30	28	93.3%
Enterobacter cloacae	10	1	10	10	100.0%
Enterococcus faecalis	10	5	50	49	98.0%
Enterococcus faecium	10	3	30	30	100.0%
Escherichia coli	10	8	80	80	100.0%
Klebsiella pneumoniae	10	5	50	46	92.0%
Staphylococcus aureus	10	7	70	69	98.6%
Streptococcus agalactiae	10	2	20	20	100.0%
Streptococcus pneumoniae	10	1	10	10	100.0%
Total	**90**	**35**	**350**	**342**	**97.7%**

discordant samples

Strain ID	Strain Species	Drug	Phenotype AST	Illumina Prediction	ONT Prediction	Illumina Accurate	ONT Accurate
AB-3	*A. baumannii*	TMP/SMX	NS	NS	S	YES	NO
AB-8	*A. baumannii*	TMP/SMX	NS	NS	S	YES	NO
EFS-11	*E. faecalis*	tetracycline	NS	S	NS	NO	YES
KP-4	*K. pneumoniae*	ceftriaxone	NS	NS	S	YES	NO
KP-4	*K. pneumoniae*	meropenem	NS	NS	S	YES	NO
KP-5	*K. pneumoniae*	meropenem	NS	NS	S	YES	NO
KP-7	*K. pneumoniae*	meropenem	NS	NS	S	YES	NO
SAU-8	*S. aureus*	tetracycline	S	S	NS	YES	NO

DAYZERO DIAGNOSTICS