Bactabolize: A tool for high-throughput generation of bacterial strain-specific metabolic models

Authors

Ben Vezina[1+*], Stephen C. Watts[1+], Jane Hawkey[1], Helena B. Cooper[1], Louise M. Judd[1], Adam Jenney[2], Jonathan M. Monk[3], Kathryn E. Holt[1,4], Kelly L. Wyres[1*]

[1] Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne, Victoria, Australia

[2] Microbiology Unit, Alfred Health, Melbourne, Victoria, Australia

[3] Department of Bioengineering, University of California, San Diego, CA, United States of America

[4] Department of Infection Biology, London School of Hygiene and Tropical Medicine, London, UK

[+] These authors contributed equally

[*] Corresponding authors (benjamin.vezina@monash.edu, kelly.wyres@monash.edu)

Abstract

Metabolic capacity can vary substantially within a bacterial species, leading to ecological niche separation, as well as differences in virulence and antimicrobial susceptibility. Genome-scale metabolic models are useful tools for studying the metabolic potential of individuals, and with the rapid expansion of genomic sequencing there is a wealth of data that can be leveraged for comparative analysis. However, there exist few tools to construct strain-specific metabolic models at scale.

Here we describe Bactabolize (github.com/kelwyres/Bactabolize), a reference-based tool which rapidly produces strain-specific metabolic models and growth phenotype predictions. We describe a pan reference model for the priority antimicrobial-resistant pathogen, *Klebsiella pneumoniae* (github.com/kelwyres/KpSC-pan-metabolic-model), and a quality control framework for using draft genome assemblies as input for Bactabolize.

The Bactabolize-derived model for *K. pneumoniae* reference strain KPPR1 outperformed the CarveMe-derived model across ≥201 substrate and ≥1220 knockout mutant growth predictions. Novel draft genomes passing our systematically-defined quality control criteria resulted in models with a high degree of completeness (≥99% genes and reactions captured) and high accuracy (mean 0.97, n=10).

We anticipate the tools and framework described herein will facilitate large-scale metabolic modelling analyses that broaden our understanding of diversity within bacterial species and inform novel control strategies for priority pathogens.

Introduction

Bacteria exhibit metabolic diversity and can utilise a broad range of substrates for growth. It has become clear amongst pathogens that there is an intertwined relationship between metabolism and nutrient usage with virulence and antimicrobial resistance (1-7). Comparative analyses of metabolic profiles (e.g. substrate usage) are key to fully understanding these relationships. Traditionally, these profiles have been assessed via phenotypic growth on a limited number of substrates, such as those used to delineate between species (8-10) which form the basis of a number of commercial products for species identification. However, these methods are not sufficiently discriminatory for in-depth comparisons within species, and alternative approaches such as the Omnilog Phenotype MicroArray system (Biolog) are too expensive and/or labour intensive for application to large numbers of isolates. Similarly, probing of essential metabolism-associated genes via transposon mutant libraries (e.g. to identify novel virulence factors and therapeutic targets) (4, 11, 12) cannot be easily scaled across diverse bacterial populations.

Genome-scale metabolic models or metabolic reconstructions are a computational approach to analysing the metabolic potential of an organism, within which the entire biochemical network is represented as a stoichiometric matrix (13). Metabolic models are constructed programmatically, but typically informed and at least partially validated using phenotypic growth data (14-16). Once constructed, they can be run through simulations and analysed under various contexts, such as *in silico* growth experiments (Flux Balance Analysis [FBA]) to predict substrate usage profiles (17), evaluate the impact of single gene knockouts on growth (14, 18), and identify metabolic chokepoints for drug targets (19), among others. Traditionally, metabolic models are strain-specific (i.e. each model represents a unique individual http://bigg.ucsd.edu/models) and may not be applicable to other isolates due to unrepresented genetic diversity.

We recently described 37 curated strain-specific models for the *Klebsiella pneumoniae* Species Complex (*Kp*SC) (14) comprised of *K. pneumoniae* and its close relatives (20). These organisms are a common cause of healthcare-associated infections world-wide, and among the World Health Organization's priority antimicrobial resistant pathogens (21). *Kp*SC are highly diverse and gene content can differ substantially between strains (22, 23). Accordingly, our models varied in terms of gene and reaction content, resulting in variable growth substrate usage profiles and metabolic redundancy (14). Similar variation has also been described in other key bacterial pathogens e.g. *Escherichia coli* (24)*, Salmonella enterica* (25), *Staphylococcus aureus* (26) and *Pseudomonas aeruginosa* (27). This is highly relevant to the use of metabolic models for the exploration of virulence and antimicrobial resistance, and for the identification of novel drug targets. Therefore, such works should seek to include multiple strain-specific models, and in some cases 100s-1000s of models may be required to accurately represent population diversity (22, 28, 29).

There are several open source tools currently available that can rapidly produce strain-specific metabolic models, including CarveMe (30), ModelSEED (31) and KBase (32) (see the recent review by Mendoza and colleagues for comparative descriptions (33)), as well as a recently published modelling and analysis pipeline, ChiMera, which leverages CarveMe for model construction (34). In their systematic analysis Mendoza *et al.* indicated CarveMe and ModelSEED to be of particular interest for large-scale studies due to their speed and model quality (33). Like KBase, ModelSEED is a web interface application, limiting its utility for high-throughput analysis of 100s – 1000s of bacterial genomes. CarveMe is a command line application; it is open source but is dependent on commercial solvers such as CPLEX (free for academic use). However, its use of a universal reference model may limit specificity of strain-specific models (35), and result in overestimation of model genes. These limitations can be overcome by manual curation of the output models, but such curation is highly labour intensive and not suitable for high-throughput analyses. Furthermore, the CarveMe database (BiGG universal_model) appears to be no longer

actively maintained, meaning that there is no opportunity to integrate novel structural and/or biochemical data as these become available in the literature (as discussed in COBRA community forums).

Here, we present Bactabolize (available at https://github.com/kelwyres/Bactabolize), an easy-to-use tool which allows scalable production of strain-specific draft metabolic models and prediction of growth phenotypes. Bactabolize builds upon the reference-based model reconstruction approach described by Norsigian *et al*. (35), leveraging the COBRApy framework (36) and BiGG nomenclature (37). We present a pan-metabolic reference model for the *Kp*SC (derived from our 37 curated strain-specific models (14)), and describe an exemplar quality control framework for the application of Bactabolize to *Kp*SC draft genome assemblies. We show that Bactabolize can rapidly produce strain-specific models from draft genomes with a high degree of completeness (as compared to models generated from completed genome assemblies), resulting in highly accurate growth predictions that match or exceed the accuracy of models from CarveMe and manual curation efforts.

Results

Description of Bactabolize

Bactabolize is written in Python 3 and utilises the metabolic modelling library COBRApy (36). Bactabolize has four main commands:

i) Draft model generation (*draft_model* command), which generates a strain-specific draft metabolic reconstruction ('model') using the approach outlined previously (35), and uses gap-filling to identify any missing reactions required to simulate growth in the user-specified conditions
ii) Patching incomplete models (*patch_model* command) by the addition of missing reactions e.g. those identified by the automated gap-filling process
iii) Substrate usage analysis via Flux Balance Analysis (FBA) (*fba* command) to predict growth outcomes for a specified range of substrates supported by the model(s)
iv) **Fig. 1**).

Additional processing scripts are provided alongside Bactabolize to improve model metadata annotation (improve_model_annotations.py), convert models generated using KBase and ModelSEED to Bactabolize/BiGG-compatible format (SEED_to_BiGG_model_convert.sh), generate network graph files from models (model_to_network_graph.py) and merging output FBA profiles (merge_fba_profiles_longtable.sh).

For draft model construction, Bactabolize requires users to provide an input assembly (annotated or unannotated FASTA or Genbank format respectively), a reference model (JSON format) and the corresponding reference sequence data (gene and protein sequences in two separate multi-fasta files or a single Genbank annotation in a .gbk file) (**Figure S1**). If the input assembly is unannotated, Bactabolize will identify coding sequences using Prodigal (38) but will otherwise honour the existing coding sequence (CDS) notations and optionally use Prodigal to search for additional CDS. Draft genome-scale metabolic models are output in both SMBL v3.1 (39) and JSON formats (one pair of files for each independent strain-specific model), along with an optional MEMOTE quality report (40). Bactabolize will identify orthologs in the input genome(s) compared to the reference sequence data using Bi-directional BLAST (41) Best Hits (BBH) (42) using BLAST+ (35). Users can parameterise the ortholog finding settings (coverage and identity thresholds) for BBH. Alternatively, there is the option of using protein similarity to identify orthologs instead of identity.

Once a draft model has been constructed, it is validated via a simulated growth experiment on user-input choice of media and atmosphere (aerobic or anerobic). Predefined media include BG11 (Gibco), M9 + glucose (35), nutrient media (43), Luria-Bertani (LB) (43), Tryptic Soy (TSA) (43), TSA + sheep blood (43), LB as specified by the CarveMe developers (30), Chemically Defined Medium (CDM)-like (33), Plantarum Minimal Medium (PMM) PMM5-like (33) and PMM7-like (33). Users can also define custom media as Bactabolize supports several complex media ingredients, including peptone (peptic digest of bovine and porcine tissue) (44-46), tryptone (pancreatic digest of casein) (44, 46, 47), soy peptone/soytone (digest of soymeal) (44, 46, 48, 49), yeast extract (50-55) and beef extract (44, 46). If the model fails to simulate growth, gap-filling is performed to indicate missing reactions. Users can add these reactions to a patch JSON file and optionally use the *patch_model* command to correct the model (**Figure S2**). Bactabolize uses a conservative gap-filling approach that only adds the minimum number of reactions to enable growth under the chosen conditions. We recommend testing the models in minimal media and atmosphere expected to support growth for all isolates of the species of interest, unless the user has access to matched phenotypic data demonstrating growth for individual isolates in specific conditions. Aggressive gap-filling will effectively homogenise the models and should be avoided if the goal is to understand the underlying strain diversity.

Substrate usage analysis (the *fba* command) is performed iteratively for each possible carbon, nitrogen, sulfur and phosphor substrate supported by the model(s) (**Figure S3**), by replacing the default substrate in the user specified growth medium (specified in the fba_spec JSON file). For example, in M9 media the default substrates are glucose (carbon), ammonia (nitrogen), sulphate (sulfur) and phosphate (phosphor). Each substrate can be tested in aerobic and/or anaerobic conditions. Growth prediction output is recorded in a tab delimited file (one per strain). The merge_fba_profiles_longtable.sh helper script will combine the outputs for multiple strains into a single file for downstream analysis.

The growth impacts of single-gene knockout mutations can be simulated via the *sgk* command (**Figure S4**). Bactabolize will iterate through every gene in the model, temporarily removing it and its associated reactions (unless they are also associated with another gene) and running FBA to simulate growth in the user-specified conditions. The output is comparable to single-gene knockout studies such as transposon mutagenesis and can be used to probe gene essentiality.

We recorded the time required for Bactabolize to build draft models and performed 1692 independent growth predictions for each of 35 *Kp*SC genomes (tested in triplicate) on a high-performance computing cluster (Intel(R) Xeon(R) Platinum 8260 CPU @ 2.40GHz and 340 GB of requested memory on a CentOS Linux release 7.9.2009 environment). The mean CPU time required for model construction was 98.41 seconds (range 83.72 - 112.55 seconds), while the mean CPU time for growth predictions was 88.79 seconds (range 85.15 - 103.37 seconds). On a standard consumer laptop (Intel(R) Core(TM) i5-8365U CPU @ 1.60GHz and 15 GB of memory on Windows Subsystem for Linux (WSL1) environment), the mean CPU time for model construction was 87.27 seconds (range 76.133 – 102.694 seconds), while growth predictions took 82.19 seconds (range 72.24 – 103.37 seconds).

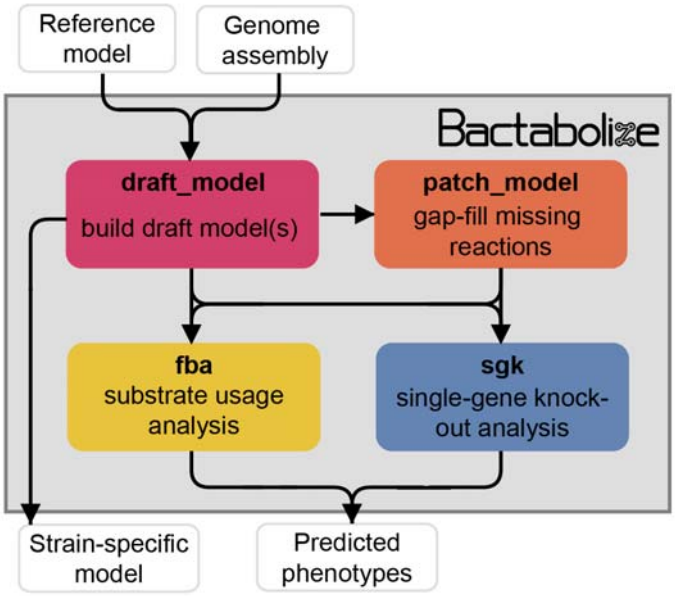


Figure 1: Simplified overview of Bactabolize's main commands. In pink is the *draft_model* command, which builds a draft strain-specific metabolic model using an input reference model and an input target assembly (approach adapted from (35)). If the model fails to simulate growth, Bactabolize will attempt automated gap-filling and produce a model patch file. The *patch_model* command (orange) allows the addition of missing reactions to produce a valid draft model that can simulate growth in a user-specified growth environment. A functioning model can be passed to the *fba* command (yellow), which performs Flux Balance Analysis to simulate growth in the user specified conditions, across all carbon, nitrogen, phosphorus and sulfur metabolite sources supported by the model under aerobic and anerobic conditions. The *sgk* command (blue) shows the Single Gene Knockout analysis, which outputs a predicted phenotype. User inputs and outputs are shown in white boxes while Bactabolize commands are shown inside the grey box.

*Kp*SC pan-metabolic reference model

We constructed a species complex-specific pan-metabolic reference model by combining a collection of 37 manually curated models for which we have previously demonstrated high accuracy (range 88.3%–96.8% for prediction of 94 distinct growth phenotypes (14)). These models represent a diverse collection of *Kp*SC (14) (including at least one each of the seven major taxa in the complex; *K. pneumoniae, Klebsiella variicola subsp variicola, Klebsiella variicola subsp tropica, Klebsiella quasipneumoniae subsp quasipneumoniae, Klebsiella quasipneumoniae subsp similipneumoniae, Klebsiella quaisivariicola, Klebsiella africana*). The combined pan-model, known as *Kp*SC-pan v1, comprises a total of 1265 distinct genes, 2319 reactions and 1696 metabolites, and is available at github.com/kelwyres/KpSC-pan-metabolic-model.

Performance comparison

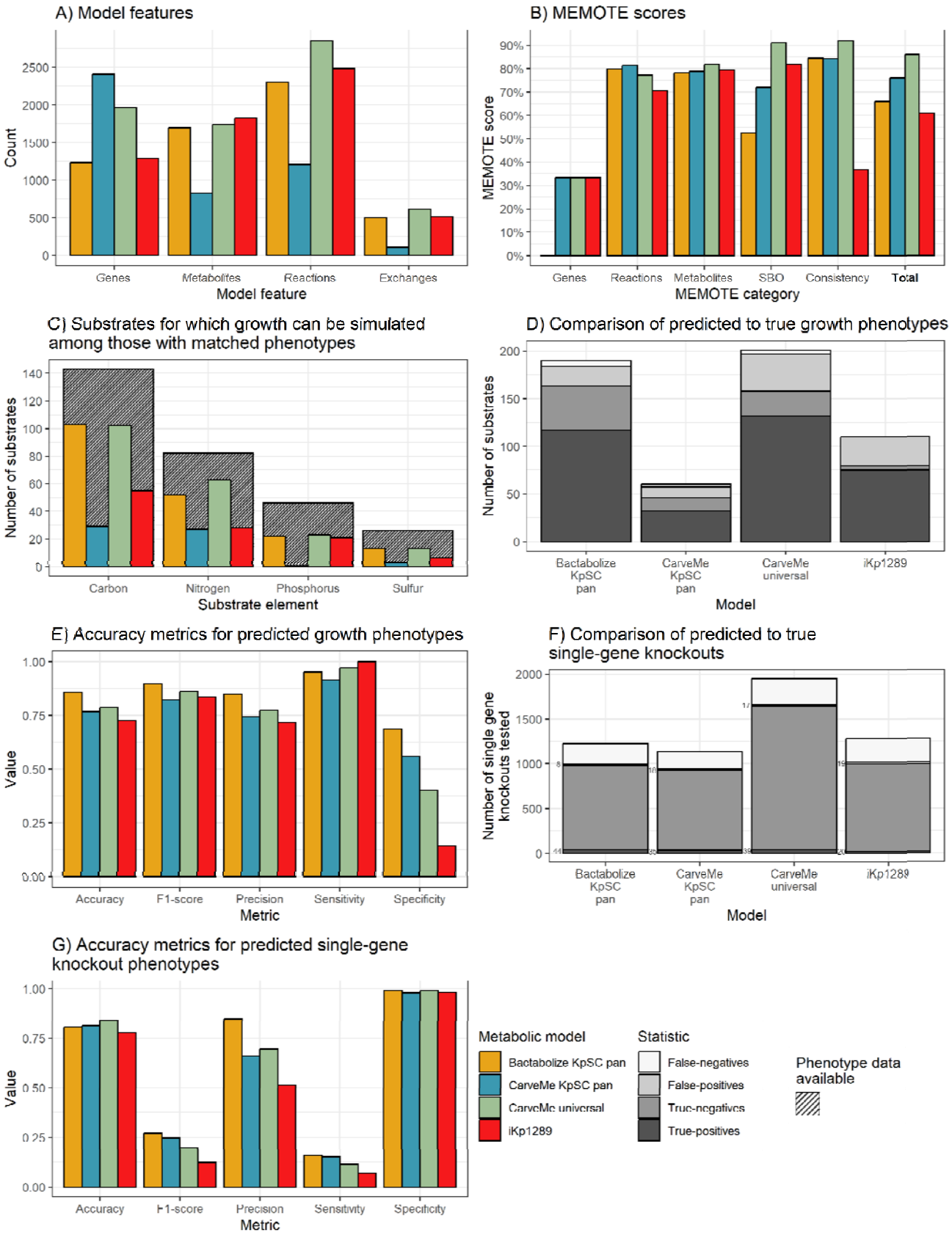
We compared the output and performance of Bactabolize to CarveMe (30) and a manually curated metabolic reconstruction of *K. pneumoniae* strain KPPR1 (also known as VK055 and ATCC 43816, metabolic model named iKp1289) (15). This isolate was chosen as there is a completed genome sequence (Genbank accession: CP009208), single-source growth phenotype (15) and single-gene knockout growth essentiality data available (56). Draft models were built using; i) Bactabolize with the *Kp*SC pan v1 reference; ii) CarveMe, with its universal reference model (CarveMe universal); and iii) CarveMe, with *Kp*SC-pan v1 reference (CarveMe *Kp*SC pan). Importantly, neither *K. pneumoniae* KPPR1 nor its genetic lineage (7 gene multi-locus sequence type, ST493), are represented in the *Kp*SC pan-reference model, meaning these benchmarking comparisons were on equal footing.

The Bactabolize draft model captured a comparable number of genes and reactions (n = 1233 and 2307, respectively) to the manually curated model (n = 1289 and 2484, respectively) but fewer than the CarveMe universal model (n = 1960 and 2857)

Fig. 2A). In contrast, the number of metabolites represented in the Bactabolize and CarveMe universal models were similar (1696 vs 1737) and both were lower than the number represented in iKp1289 (n = 1827). The CarveMe *Kp*SC pan model method captured considerably more genes than any of the other models (n = 2407), but these were associated with many fewer unique reactions and metabolites (1206 and 825, respectively). Upon further investigation we determined that this method resulted in the over prescription of gene reaction rules (GPRs) to multiple reactions (mean 2.2 GPRs per reaction when compared to Bactabolize using the same pan reference model: 1.94 GPRs per reaction; and CarveMe Universal: 2.12 GPRs per reaction). MEMOTE scores, (produced by the MEMOTE report (40)) indicate the quality of the model metadata annotations, with the scores ranging between 0 – 100%. These provide a measure of model portability and the level of connected databases available to support the metabolite, reaction and genetic information represented in the model, but bear no reflection on model accuracy. Bactabolize performs on the lower end, with CarveMe universal performing the best (**Fig. 2B**). However, Bactabolize using the *Kp*SC-pan model outperforms the model propagation mode of CarveMe using the same reference model (**Fig. 2B**). Work is ongoing to improve the annotations in the *Kp*SC-pan reference model, to improve large-scale model propagation.

We assessed the performance of each model for *in silico* prediction of growth phenotypes compared to the previously published experimental data (15). Accuracy, sensitivity, specificity, precision and F1 scores were calculated (57). Note that the specific set of growth substrates and gene knockouts that can be simulated is determined by the sets of genes and metabolites captured by each model and is therefore model-dependent (**Data S1 and S2**). Among those with matched experimental phenotype data, the Bactabolize and CarveMe universal models were able to predict growth for a greater number of carbon, nitrogen, phosphorous and sulfur substrates than both the iKp1289 model and the CarveMe *Kp*SC pan models (**Fig. 2C**, **Data S1**). While the CarveMe universal model had the highest number of true-positive growth predictions overall, it also had a comparably high number of false-positive predictions (**Fig. 2D**). In contrast, the Bactabolize model had fewer false-positive predictions, resulting in the highest overall accuracy metrics (**Fig. 2E, Data S1**). Similarly, while the CarveMe universal model resulted in the highest absolute number of true-positive gene essentiality predictions, driving a high accuracy, the Bactabolize model was associated with the greatest overall precision, sensitivity, and specificity (**Figs. 2F & 2G**).

236



237

Figure 2 (previous page): *K. pneumoniae* KPPR1 metabolic model benchmarking comparisons. **A)** Counts of model features; genes, metabolites and reactions captured by each model. Exchanges refers to number of exchange reactions, a subset of reactions involved in substrate uptake, which determine the number of distinct growth substrates for which phenotypes can be predicted with the model. **B)** MEMOTE scores indicating the richness of annotations and metadata for metabolic model features according to database outlinks. SBO refers to score of Systems Biology Ontology (SBO), a controlled vocabulary for systems biology. Consistency refers to the score of stoichiometric consistency and chemical formulae annotation. Total refers to total MEMOTE score, as a combination of all previous scores, and is shown in bold. **C)** Counts of carbon, nitrogen, phosphorus and sulfur growth substrates that can be simulated by models and for which matched phenotypes were available for comparison (15). Hatched columns indicate the total number of substrates for which phenotypic data for *K. pneumoniae* KPPR1 were described (15). **D)** and **E)** Accuracy metrics for predicted to true phenotypes for the growth substrates shown in D and E, respectively False-negatives, true-negatives, false-positives and true-positives are coloured as shown in legend. **F)** and **G)** Accuracy metrics for the KPPR1 single-gene knockout mutant library described in (56) shown in F and G, respectively. Numbers of true positives and false positives are shown to the left of the respective columns.

Quality control framework for input genome assemblies

There are now thousands of bacterial genomes available in public databases, the majority of which are in draft form. If we are to use these data for high-throughput metabolic modelling studies, it is essential to evaluate the expected model accuracies and understand the minimum input genome quality requirements. Here we performed a systematic analysis leveraging our published curated *Kp*SC models (n=37, (14)), which were generated using completed genome sequences and were therefore considered to represent 'complete' models. We randomly subsampled the corresponding Illumina read sets to various depths (10 – 100x, increments of 10) in triplicate and generated draft assemblies that were passed to Bactabolize for generation of draft metabolic models (**Data S3**). Due to low read depth (≤30x), two isolates were removed from this analysis. Additionally, ten 10x depth read samples failed to produce assemblies, leaving 1040 draft genomes for analysis. The resulting draft metabolic models were compared to the complete models to; i) determine the proportions of complete model genes and reactions captured in the draft models; and ii) compare 846 *in silico* aerobic growth predictions in M9 minimal media, where growth on 266 carbon, 153 nitrogen, 59 phosphorus and 25 sulfur sources were examined. Substrates containing multiple elements were tested as sole sources of each element independently and in combination, e.g. 1,5-Diaminopentane was tested as a sole carbon, sole nitrogen and sole carbon plus nitrogen source.

As expected, assembly quality generally increased with increasing sequencing depth i.e. assemblies generated from higher depth read sets were associated with higher N50 values, fewer contigs and fewer assembly graph dead-ends, although the rate of improvement drastically declined beyond 40-50x depth (**Figure S5, Data S3**). We noted that it was rare for draft models to capture 100% of model genes and reactions (just 420 of all 1040 draft assemblies were associated with models that captured 100% model genes) (**Data S3, Figure S6**), even when using the highest quality draft genomes. However, ≥99% of genes and reactions were commonly captured, which plateaued from 40x depth onwards (**Figure S6**). Therefore, we sought to evaluate whether ≥99% model capture would produce functionally accurate models.

We used FBA to simulate substrate growth profiles for the 40x depth assemblies, representing a sequencing depth that can be routinely achieved with standard Illumina library preparations. All but one assembly triplicate set (isolate SB4767 98% gene capture, 99% reaction capture) captured ≥99% but ≤100% model genes and/or reactions. The substrate growth profiles were then compared to those of the complete models. The vast majority of draft models produced accurate growth predictions; 102 of 108 models resulted in predictions with 100% concordance to those from the corresponding complete models. Three models for *K. quasipneumoniae similipneumoniae* isolate SB164 resulted in predictions with a mean of 99.8% concordance. The remaining three models were for isolate SB4767 and resulted in mean of 80.4% concordance. Notably, these models were those representing <99% gene capture. Together, these data suggest that draft

models capturing ≥99% of the complete model genes/reactions generate highly accurate growth predictions and that these capture rates can be readily achieved from draft genome assemblies.

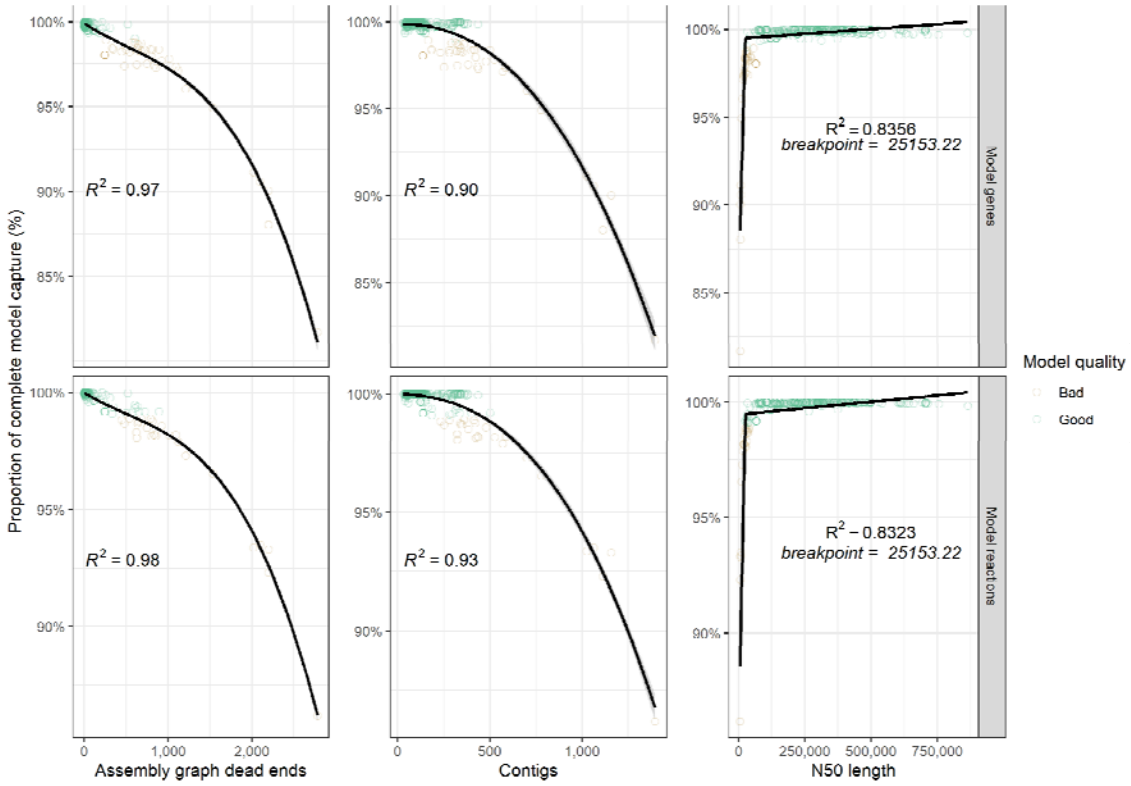


Figure 3: Scatterplots showing distribution of best performing assembly metrics 'assembly graph dead ends', 'contigs' and 'N50' against model feature capture (genes and reactions). Each point represents the mean values from a single genome (technical triplicate) and is coloured by model quality. 'Good' models capture ≥99% of the model metric as compared to the corresponding complete model (shown at each facet), 'Bad' models capture <99%. Cubic polynomial line plotted for assembly 'graph dead ends', 'contigs', while a segmented linear model was plotted for 'N50'. R^2 is shown on each panel.

We investigated the relationships between assembly quality metrics and model gene/reaction capture in more detail. Variation in assembly graph dead-ends accounted for the greatest amount of variation in model capture, closely followed by raw contig counts (cubic polynomial fit, R^2 of ≥0.98 for graph dead-ends, R^2 of ≥0.9 for contig count). A segmented linear model was fitted to N50 length (R^2 ≥ 0.83), producing a breakpoint at 25153 bp (**Fig. 3**).

To further explore the optimum thresholds for assembly metrics, we tallied the number of draft assemblies resulting in ≥99% and <99% gene and reaction capture at increasing graph dead-end and contig count count-offs, and decreasing N50 cut-offs. Draft models that captured ≥99% of the complete model genes/reactions were considered 'good' models, whereas draft models that captured <99% of complete model genes/reactions were considered 'bad' models. The optimum threshold for assembly graph dead end was determined to be ≤200. At this value, 94.44% of 'good' models were captured, and 0% 'bad' models. The optimum threshold for contig counts was determined as ≤130 contigs at which 67.92% of 'good' and 0% 'bad' models were captured (**Fig. 4**). The optimum threshold for N50 was determined to be ≥65000, at which 94.97% of 'good' and 1.71% of 'bad' models were captured. The assembly graph dead-end threshold results in comparatively higher sensitivity (i.e. a higher proportion of 'good' models pass the threshold) than contig count and comparatively better specificity (i.e. lower proportion of 'bad' models pass the

threshold) than N50, but the underlying metric information is not universally available because many isolate genomes are deposited in public databases only as assemblies without the associated assembly graph. We therefore recommend a three-tier approach, whereby the assembly graph dead-end criterion is preferenced if available, followed by N50 and then contig count.

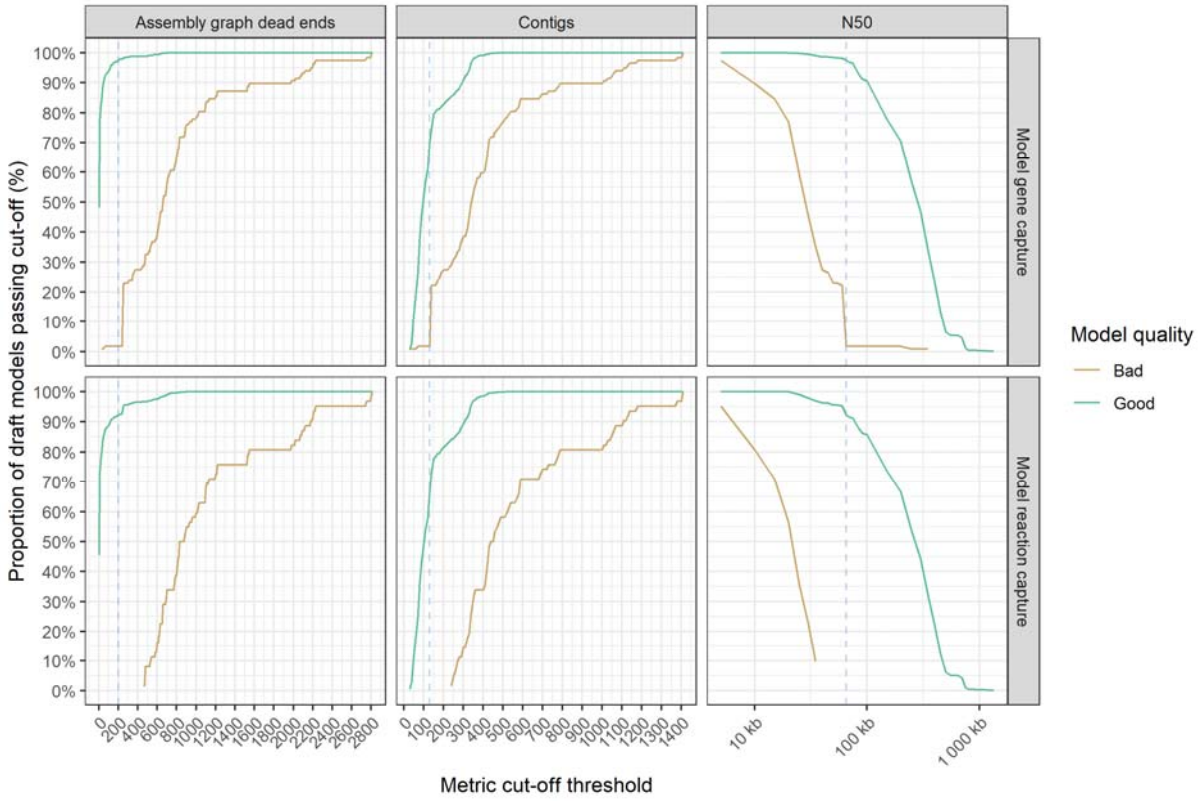


Figure 4: Line graphs showing the impact of assembly metric cut-off thresholds on model feature capture (n = 1040). 'Good' models which captured ≥99% of model features are shown in green, while 'bad' models captured <99% model features are shown in gold. The blue dotted line shows the metric cut-off thresholds, to minimize the number of models that capture <99% model features and maximise models that capture ≥99%. Metric cut-off statistics are calculated in intervals of 10 for assembly graph dead ends and contigs, and every 5000 for N50.

Impact of gap-filling models

Of the 901 draft genome assemblies which passed our QC criteria (≤200 assembly graph dead ends), 23 of the resulting draft models failed to simulate growth in M9 minimal media with glucose (despite capturing ≥99% of the genes and reactions in the corresponding complete models). It is expected that all *Kp*SC models should be able to simulate growth on M9 media with glucose as a sole carbon source, as this central metabolism is universal amongst *Kp*SC. To replace missing, critical reactions required for growth on M9 with glucose, we investigated model gap-filling using the *patch_model* command of Bactabolize. We then assessed the accuracy of the gap-filled models for prediction of growth on the full range of substrates, as compared to the predictions from the corresponding complete models.

Gap filling added 1 – 3 missing reactions to each model, with a median of one, fully restoring biomass production in M9 media with glucose in all but two of the 23 failed models. The missing

reactions appeared to be random genes across these 23 genomes, likely due to missing information in these assemblies.

Substrate usage predictions from the 21 successfully gap-filled models were highly accurate, with 18/21 having a prediction concordance of ≥99% across all 846 growth conditions (12/21 had 100% concordance) (**Figure S7**). We therefore conclude that models generated for genome assemblies passing our QC criteria, which have been gap-filled to successfully simulate growth on minimal media plus glucose, are suitable for the prediction of growth across a range of substrates.

Predictive accuracy of draft models

We assessed the accuracy of Bactabolize for the construction of draft models for 10 novel *Kp*SC clinical isolates, representing five of the major taxa in the complex. We included five isolates for which the associated STs were represented in the *Kp*SC-pan v1 model and five isolates with STs that were not represented. Whole genome sequence data were generated on the Illumina platform and draft assemblies generated *de novo*. The resultant assemblies had 0-4 graph dead-ends, N50s of 151958-388486 bp and 83-187 contigs (**Data S4**), within the tiered threshold values.

FBA was performed, and the predicted growth profiles compared to matched phenotypic growth data for 16 carbon sources derived from Vitek GN ID cards. Though the number of tested carbon sources was limited, all were associated with high accuracy metrics (**Fig. 5**, **Data S4**). As expected, models for isolates with STs represented in the *Kp*SC-pan v1 reference performed slightly better (mean accuracy = 0.98) than those for non-represented STs (mean accuracy = 0.95).

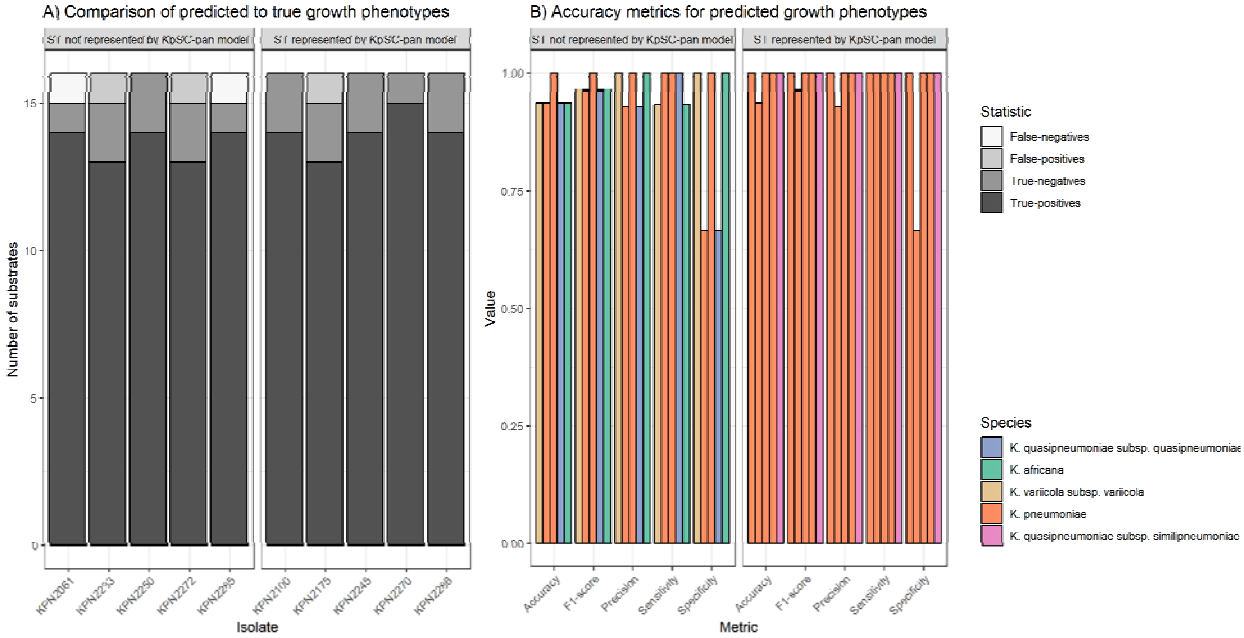


Figure 5: A) Comparisons of predicted to true phenotypes for 16 carbon source substrates. False-negatives, true-negatives, false-positives and true-positives are coloured as shown in legend. Each column represents a different isolate, separated by ST representation in the *Kp*SC-pan model. **B)** Accuracy metrics for predicted vs phenotypic growth comparisons shown in A. Each column represents a different isolate, coloured by taxa and separated by ST representation in the *Kp*SC-pan v1 model.

Discussion

In this work we described Bactabolize, a pipeline for rapid and scalable production of accurate bacterial strain-specific metabolic models and growth phenotype predictions. We describe a pan-reference model for the *Kp*SC and demonstrate that a draft strain-specific model generated *de novo* via Bactabolize using the *Kp*SC-pan v1 reference was highly accurate for growth phenotype prediction (85.79% accuracy for substrate usage across 190 substrates, and 80.57% for gene essentiality across 1220 genes). Importantly, we also described a quality control framework for the use of draft genome assemblies as input for metabolic reconstructions. We used a systematic analysis to; i) evaluate the proportion of gene and reaction capture compared to the corresponding 'completed' models; ii) define quality control thresholds for input assemblies (three tier approach for *Kp*SC; ≤200 assembly graph dead ends, followed by ≥65000 N50, followed by ≤130 contigs); and iii) estimate the accuracy of the resultant growth predictions. While the quality control thresholds and accuracy estimates are specific to *Kp*SC, the conceptual framework can be applied to any organism and is essential to support the confident application of metabolic modelling for large-scale genome datasets. We appreciate that assembly graphs may not be available for dead end count, e.g. for draft genome assemblies accessed via public repositories, however we encourage users to include this information in their quality control procedures wherever possible (e.g. using the recently published counter tool available at https://github.com/rrwick/GFA-dead-end-counter), because these counts represent a direct reflection of the completeness of the genome assembly. In contrast, contig counts and N50 are influenced by biological features such as repeat copy numbers as well as the underlying sequence data quality e.g. a bacterial genome harbouring many insertion sequence insertions will result in a draft assembly with a high number of contigs regardless of the sequence data quality and completeness.

Bactabolize's reference-based reconstruction approach is reductive, meaning the resultant draft models will comprise only the genes, reactions and metabolites present in the reference, or a subset thereof, and will not include novel reactions unless they are manually identified and curated by the user. This is an important caveat that should be considered carefully for application of Bactabolize to large genome data sets, particularly for genetically diverse organisms such as those in the *Kp*SC. The use of a pan-reference derived from multiple curated strain-specific models results in greater representation of the population diversity and partially alleviates the shortcomings of the reference-based approach. However, draft models constructed for strains with a corresponding lineage represented in the reference are likely more accurate. Our analysis indicated that a draft *Kp*SC model generated by Bactabolize with the *Kp*SC pan v1 reference was equally or more accurate than the current gold standard automated approach, CarveMe with universal model (30), and outperformed a manually curated model (15). The latter was constructed using the KBase pipeline (32), which uses RAST to annotate the sequences with Enzyme Commission numbers. It has been demonstrated several times that the Enzyme Commission scheme has systematic errors (58, 59), leading to a loss in accuracy when compared to the ortholog identification methods used in CarveMe and Bactabolize.

The CarveMe universal reference model captures greater diversity than the *Kp*SC pan v1 reference, which resulted in a comparatively greater number of genes, reactions and metabolites in the corresponding CarveMe draft model, and ability to simulate growth outcomes for a greater number of distinct substrates (**Figure 2**). Overall, CarveMe (with the universal model) performed extremely well, with high numbers of true-positive growth predictions. However, these were also accompanied by comparatively higher numbers of false-positive predictions, which resulted in a lower overall accuracy score for substrate usage analysis compared to Bactabolize with the *Kp*SC-pan v1 reference (**Figure 2**), and comparatively lower sensitivity and specificity for the gene essentiality analysis. False-positive predictions may indicate that the relevant metabolic machinery

are present in the cell but were not active during the growth experiments (e.g. due to lack of gene expression). In this regard, false-positives are not always a sign of model inaccuracy. However, false-positive predictions can also occur from incorrect gene annotations e.g. due to reduced specificity of ortholog assignment resulting from the use of the universal model without manual curation. Given a key objective here is to facilitate high-throughput analysis for large numbers of genomes, it is not feasible to expect that all models will be manually curated, and therefore we believe that identifying fewer genes with lower overall error rates provides greater confidence in the resulting draft models. We also note that the CarveMe universal reference model is no longer being actively maintained, but in contrast, user defined species- (or genera-) specific references can be iteratively curated and updated to incorporate new knowledge and data as they become available. Accordingly, the accuracy of models derived from such references is expected to continually improve.

Bactabolize and the *Kp*SC pan v1 model are freely available under open source licenses and satisfy the four features of the FAIR research principles (findability, accessibility, interoperability and reusability) (60). In addition to the *Kp*SC pan-reference described here, a pan-reference model has been described previously for *Salmonella enterica* (representing 410 strains (25)). We are actively working to expand and improve the *Kp*SC pan reference model and welcome similar efforts to generate high quality references for other organisms. Together these resources will facilitate population wide metabolic analyses for global priority pathogens, which can be used to understand how they transmit, cause disease and evolve drug-resistance, and to identify novel therapeutic targets.

Methods

Bactabolize pipeline

Bactabolize utilises the existing metabolic modelling library COBRApy (36) and Python 3 (61). All code is freely available and open source at GitHub (www.github.com/kelwyres/Bactabolize) under a GNU General Public License v3.0. Users should additionally cite COBRApy (36) if Bactabolize is used.

Klebsiella pneumoniae Species Complex-pan metabolic model

The 37 metabolic models from a previous study (14) were combined with the iY1228 model using the create_master_model.py script (available at 10.6084/m9.figshare.21728717). Briefly, all GPRs from the iYL1228 model and the associated sequences were included, as well as new GPRs identified from the 36 additional strains by manual curation following comparison to the matched phenotype data (as described in (14)). Additionally, orthologous sequence variants with <75% nucleotide identity to gene sequences associated with these gene reaction rules (GPRs) were added if there was phenotype data supporting the reaction. The biomass reaction was updated, removing the metabolites udpgalur_c and udpgal_c as their production was strain-specific.

Metadata annotations were improved using the improve_model_annotations.py script (also available in the Bactabolize code repository) resulting in the *Kp*SC_pan v1 used in this study, available at www.github.com/kelwyres/KpSC-pan-metabolic-model.

Draft model generation

Bactabolize draft models were generated using the *draft_model* command in Bactabolize v1 with the *Kp*SC-pan v1 model as a reference, and the following options:

--min_coverage 25 --min_pident 80 --media M9 --atmosphere aerobic

CarveMe draft models were generated firstly using the universal reference with the following commands: '-g M9 -i M9'. The --universe-file mode was also used, so the *Kp*SC-pan model could be used as a reference, with the previously described command.

Speed calculations

Bactabolize *draft_model* and *fba* commands were timed via a script using the date +%s.%N command run before and after command on the MASSIVE computing cluster (Intel(R) Xeon(R) Platinum 8260 CPU @ 2.40GHz and 340 GB of memory, CentOS Linux release 7.9.2009 environment). Speed tests were also performed on a standard consumer laptop with the following hardware: Intel(R) Core(TM) i5-8365U CPU @ 1.60GHz and 15 GB of memory on Windows Subsystem for Linux (WSL1) environment.

Performance comparison

The genome of *K. pneumoniae* KPPR1 was obtained from Genbank under the accession: CP009208, and draft metabolic models were generated using Bactabolize and CarveMe as described above. The previously described, manually curated model for KPPR1 (iKp1289) was also included for comparison (15). The following KPPR1 phenotype data were retrieved from published studies: BIOLOG Phenotypic Microarray data (15) and single gene knockout data inferred from the outputs of a TraDIS transposon mutagenesis library (56).

A list of BIOLOG growth substrates for plates PM1, PM2A, PM3B, and PM4A (62) were converted where possible to BiGG and SEED IDs by manual search of the BiGG (bigg.ucsd.edu) and SEED websites (https://modelseed.org/biochem/compounds). An updated BiGG to SEED dictionary can be found in **Data S1**. A total of 143 of 190 carbon, 82 of 95 nitrogen, 46 of 59 phosphor and 26 of 35 sulfur substrates were successfully matched to BiGG and SEED IDs (**Data S1**). These growth data were compared to *in silico* predictions generated via FBA using the *fba* command from Bactabolize to optimise the biomass objective function with the following options:

--fba_spec_name m9 --fba_open_value -20

Gene essentiality was inferred from single gene knockout growth predictions using the *sgk* command from Bactabolize with the following options to mirror the growth conditions of the TraDIS library (LB media grown aerobically):

--media_type lb --atmosphere aerobic

In all cases, an objective value cut-off of $\geq 10^{-4}$ was used to indicate binarised growth as per previous studies (14, 63).

In silico predictions were compared to matched phenotype data and the following accuracy metrics were calculated:

$$Precision = \frac{TP}{TP + FP}$$

$$Sensitivity/recall = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$F1 - score = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity}$$

Quality control framework

Illumina read sets (250 bp paired end) and completed genome sequences for 37 *Kp*SC isolates were described previously (14). Here we randomly subsampled the Illumina reads at various depths (10 – 100, by increments of 10) using rasusa version 0.3.0 (64) in technical triplicate. Reads were then trimmed using TrimGalore version 0.5.0 (65) and assembled *de novo* with Unicycler version 0.4.7 (66), default parameters. Assembly statistics and assembly graph dead ends were calculated using the GFA-dead-end-counter version 1.0.0 (https://github.com/rrwick/GFA-dead-end-counter) (67). Draft metabolic models were generated with Bactabolize using the *Kp*SC-pan v1 reference, and growth substrate profiles were predicted as described above. We compared the outputs from models generated for draft genome assemblies to those generated for the corresponding completed genomes. Where necessary models were gap-filled via the *patch_model* command.

Predictive accuracy of draft models

Novel growth phenotype data were generated for 10 *Kp*SC clinical isolates from our in house collection using the VITEK 2 GN ID card system as described previously (14). Briefly, isolates were grown on Tryptic Soy (OXOID) agar plates overnight at 37°C, then analysed using VITEK 2 GN ID cards (bioMérieux) and read on the VITEK 2 Compact (bioMérieux) as per manufacturer's instructions using software version 8.0. DNA was extracted for whole-genome sequencing via Genfind v3 extraction kit, library preparation performed using Nextera Flex (Illumina) using ¼ reagents. Paired-end read data (300 bp) were generated on an Illumina NovaSeq6000 SP v1.0 and have been deposited in the European Nucleotide Archive under Bioproject PRJNA777643 (individual read accession numbers are given in **Data S4**). Draft genome assemblies were generated with Unicycler, and draft metabolic models and growth predictions were generated with Bactabolize as described above.

Statistics and visualisation

Statistical analysis and graphical visualisation were performed using R version 4.0.3 [23], RStudio version 1.3.1093 (68), with the following software packages: tidyverse version 1.3.1 (69), viridis version 0.5.1 (70), RColorBrewer version 1.1-2 (71), ggpubr version 0.4.0 (72) ggpmisc version 0.4.4 (73), aplot version 0.1.6 (74), colorspace version 2.0-2 (75), ggpattern version 0.4.3-3 (76), ggtext version 0.1.1 (77) and glue version 1.4.2 (78).

Linear regression analysis was performed in R using the lm function in R and a third degree polynomial model was fitted to plots with the following equation: y ~ poly(x, 3, raw = TRUE). The segmented linear model was fitted using segmented version 1.6-2 (79).

All code used to generate results can be found as supplemental material, https://github.com/kelwyres/Bactabolize and on Figshare (10.6084/m9.figshare.21728717).

Logo

The Bactabolize logo was constructed in Inkscape version 1.0.1 (80). The font used is Proportional TFB (81) and Element (82).

Author contributions

Conceptualization: BV, JH, JMM, KEH, KLW

Methodology: SCW, BV, LMJ, JH, JMM, KLW

Software: SCW, BV

Validation: BV, HBC, KLW

Formal analysis: BV, KLW

Investigation: BV, KLW

Resources: AJ

Writing - Original Draft: BV, KLW

Writing - Review & Editing: All authors

Visualization: BV, KLW

Supervision: JMM, KEH, KLW

Project administration: KLW

Funding acquisition: JMM, KEH, KLW

All authors contributed to and approve of the manuscript in its current form.

Acknowledgements

We thank Sylvain Brisse for the isolates used in this study and review of the manuscript.

References

1. Su Q, Guan T, He Y, Lv H. Siderophore Biosynthesis Governs the Virulence of Uropathogenic *Escherichia coli* by Coordinately Modulating the Differential Metabolism. Journal of Proteome Research. 2016;15(4):1323-32.
2. Wu Y, Meng Y, Qian L, Ding B, Han H, Chen H, et al. The Vancomycin Resistance-Associated Regulatory System VraSR Modulates Biofilm Formation of *Staphylococcus epidermidis* in an ica-Dependent Manner. mSphere.6(5):e00641-21.
3. Mir M, Prisic S, Kang C-M, Lun S, Guo H, Murry JP, et al. Mycobacterial gene cuvA is required for optimal nutrient utilization and virulence. Infection and immunity. 2014;82(10):4104-17.
4. Vornhagen J, Sun Y, Breen P, Forsyth V, Zhao L, Mobley HLT, et al. The *Klebsiella pneumoniae* citrate synthase gene, gltA, influences site specific fitness during infection. PLOS Pathogens. 2019;15(8):e1008010.
5. Eberl C, Weiss AS, Jochum LM, Durai Raj AC, Ring D, Hussain S, et al. *E. coli* enhance colonization resistance against *Salmonella* Typhimurium by competing for galactitol, a context-dependent limiting carbon source. Cell Host & Microbe. 2021.
6. Jenior ML, Dickenson ME, Papin JA. Genome-scale metabolic modeling reveals increased reliance on valine catabolism in clinical isolates of *Klebsiella pneumoniae*. bioRxiv. 2021:2021.09.08.459555.

7. Hudson AW, Barnes AJ, Bray AS, Zafar MA. *Klebsiella pneumoniae* L-Fucose metabolism promotes gastrointestinal colonization and modulates its virulence determinants. bioRxiv. 2022:2022.05.18.492588.
8. Rodrigues C, Passet V, Rakotondrasoa A, Diallo TA, Criscuolo A, Brisse S. Description of *Klebsiella africanensis* sp. nov., *Klebsiella variicola* subsp. *tropicalensis* subsp. nov. and *Klebsiella variicola* subsp. *variicola* subsp. nov. Res Microbiol. 2019;170(3):165-70.
9. Blin C, Passet V, Touchon M, Rocha EPC, Brisse S. Metabolic diversity of the emerging pathogenic lineages of *Klebsiella pneumoniae*. Environmental Microbiology. 2017;19(5):1881-98.
10. Brisse S, Fevre C, Passet V, Issenhuth-Jeanjean S, Tournebize R, Diancourt L, et al. Virulent Clones of *Klebsiella pneumoniae*: Identification and Evolutionary Scenario Based on Genomic and Phenotypic Characterization. PLOS ONE. 2009;4(3):e4982.
11. Mobegi FM, van Hijum SAFT, Burghout P, Bootsma HJ, de Vries SPW, van der Gaast-de Jongh CE, et al. From microbial gene essentiality to novel antimicrobial drug targets. BMC Genomics. 2014;15(1):958.
12. Hogan AM, Scoffone VC, Makarov V, Gislason AS, Tesfu H, Stietz MS, et al. Competitive Fitness of Essential Gene Knockdowns Reveals a Broad-Spectrum Antibacterial Inhibitor of the Cell Division Protein FtsZ. Antimicrobial Agents and Chemotherapy. 2018;62(12):e01231-18.
13. Edwards JS, Palsson BO. Systems Properties of the Haemophilus influenzaeRd Metabolic Genotype. Journal of Biological Chemistry. 1999;274(25):17410-6.
14. Hawkey J, Vezina B, Monk JM, Judd LM, Harshegyi T, López-Fernández S, et al. A curated collection of *Klebsiella* metabolic models reveals variable substrate usage and gene essentiality. Genome Research. 2022.
15. Henry CS, Rotman E, Lathem WW, Tyo KEJ, Hauser AR, Mandel MJ. Generation and Validation of the iKp1289 Metabolic Model for *Klebsiella pneumoniae* KPPR1. The Journal of Infectious Diseases. 2017;215(suppl_1):S37-S43.
16. Liao Y-C, Huang T-W, Chen F-C, Charusanti P, Hong JSJ, Chang H-Y, et al. An Experimentally Validated Genome-Scale Metabolic Reconstruction of *Klebsiella pneumoniae* MGH 78578, *i*YL1228. Journal of Bacteriology. 2011;193(7):1710-7.
17. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nature Biotechnology. 2010;28(3):245-8.
18. Stanway RR, Bushell E, Chiappino-Pepe A, Roques M, Sanderson T, Franke-Fayard B, et al. Genome-Scale Identification of Essential Metabolic Processes for Targeting the Plasmodium Liver Stage. Cell. 2019;179(5):1112-28.e26.
19. Ramos PIP, Fernández Do Porto D, Lanzarotti E, Sosa EJ, Burguener G, Pardo AM, et al. An integrative, multi-omics approach towards the prioritization of *Klebsiella pneumoniae* drug targets. Scientific Reports. 2018;8(1):10755.
20. Wyres KL, Lam MMC, Holt KE. Population genomics of *Klebsiella pneumoniae*. Nature Reviews Microbiology. 2020;18(6):344-59.
21. WHO. WHO publishes list of bacteria for which new antibiotics are urgently needed 2017 [Available from: https://www.who.int/news/item/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed.
22. Gorrie CL, Mirčeta M, Wick RR, Judd LM, Lam MMC, Gomi R, et al. Genomic dissection of *Klebsiella pneumoniae* infections in hospital patients reveals insights into an opportunistic pathogen. Nature Communications. 2022;13(1):3017.
23. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, et al. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance *Klebsiella pneumoniae*, an urgent threat to public health. Proceedings of the National Academy of Sciences. 2015;112(27):E3574.
24. Monk JM, Charusanti P, Aziz RK, Lerman JA, Premyodhin N, Orth JD, et al. Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. Proceedings of the National Academy of Sciences. 2013;110(50):20338.
25. Seif Y, Kavvas E, Lachance J-C, Yurkovich JT, Nuccio S-P, Fang X, et al. Genome-scale metabolic reconstructions of multiple *Salmonella* strains reveal serovar-specific metabolic traits. Nature Communications. 2018;9(1):3771.

26. Bosi E, Monk JM, Aziz RK, Fondi M, Nizet V, Palsson BØ. Comparative genome-scale modelling of *Staphylococcus aureus* strains identifies strain-specific metabolic capabilities linked to pathogenicity. Proceedings of the National Academy of Sciences. 2016;113(26):E3801-E9.
27. Bartell JA, Blazier AS, Yen P, Thøgersen JC, Jelsbak L, Goldberg JB, et al. Reconstruction of the metabolic network of Pseudomonas aeruginosa to interrogate virulence factor synthesis. Nature Communications. 2017;8(1):14631.
28. Croucher NJ, Coupland PG, Stevenson AE, Callendrello A, Bentley SD, Hanage WP. Diversification of bacterial genome content through distinct mechanisms over different timescales. Nature Communications. 2014;5(1):5471.
29. Cummins EA, Hall RJ, Connor C, McInerney JO, McNally A. Distinct evolutionary trajectories in the *Escherichia coli* pangenome occur within sequence types. Microbial Genomics. 2022;8(11).
30. Machado D, Andrejev S, Tramontano M, Patil KR. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Res. 2018;46(15):7542-53.
31. Seaver SMD, Liu F, Zhang Q, Jeffryes J, Faria JP, Edirisinghe JN, et al. The ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. Nucleic Acids Res. 2021;49(D1):D575-D88.
32. Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, et al. KBase: The United States Department of Energy Systems Biology Knowledgebase. Nature Biotechnology. 2018;36(7):566-9.
33. Mendoza SN, Olivier BG, Molenaar D, Teusink B. A systematic assessment of current genome-scale metabolic reconstruction tools. Genome Biology. 2019;20(1):158.
34. Tamasco G, Kumar M, Zengler K, Silva-Rocha R, da Silva RR. ChiMera: an easy to use pipeline for bacterial genome based metabolic network reconstruction, evaluation and visualization. BMC Bioinformatics. 2022;23(1):512.
35. Norsigian CJ, Fang X, Seif Y, Monk JM, Palsson BO. A workflow for generating multi-strain genome-scale metabolic models of prokaryotes. Nature Protocols. 2020;15(1):1-14.
36. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC Systems Biology. 2013;7(1):74.
37. Schellenberger J, Park JO, Conrad TM, Palsson BØ. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. BMC Bioinformatics. 2010;11(1):213.
38. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11(1):119.
39. Keating SM, Waltemath D, König M, Zhang F, Dräger A, Chaouiya C, et al. SBML Level 3: an extensible format for the exchange and reuse of biological models. Molecular Systems Biology. 2020;16(8):e9110.
40. Lieven C, Beber ME, Olivier BG, Bergmann FT, Ataman M, Babaei P, et al. MEMOTE for standardized genome-scale metabolic model testing. Nature Biotechnology. 2020;38(3):272-6.
41. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10(1):421.
42. Hernández-Salmerón JE, Moreno-Hagelsieb G. Progress in quickly finding orthologs as reciprocal best hits: comparing blast, last, diamond and MMseqs2. BMC Genomics. 2020;21(1):741.
43. BD. Difco™ & BBL™ Manual, 2nd Edition. In: BD, editor. 2009.
44. Biosciences BD. BD Bionutrients Technical Manual: BD Biosciences – Advanced Bioprocessing. BD Biosciences; 2015.
45. Loginova LI, Manuilova VP, Tolstikov VP. Content of free amino acids in peptone and the dynamics of their consumption in the microbiological synthesis of dextran. Pharmaceutical Chemistry Journal. 1974;8(4):249-51.
46. ThermoFisherScientific. Technical guide to peptones, supplements, and feeds: Enhancing performance of mammalian and microbial bioprocesses. ThermoFisherScientific; 2019.

47. Clausen E, Gildberg A, Raa J. Preparation and testing of an autolysate of fish viscera as growth substrate for bacteria. Applied and Environmental Microbiology. 1985;50(6):1556-7.
48. Hagely KB, Palmquist D, Bilyeu KD. Classification of Distinct Seed Carbohydrate Profiles in Soybean. Journal of Agricultural and Food Chemistry. 2013;61(5):1105-11.
49. Choct M, Dersjant-Li Y, McLeish J, Peisker M. Soy Oligosaccharides and Soluble Non-starch Polysaccharides: A Review of Digestion, Nutritive and Anti-nutritive Effects in Pigs and Poultry. Asian-Australasian Journal of Animal Sciences. 2010;23.
50. Tomé D. Yeast Extracts: Nutritional and Flavoring Food Ingredients. ACS Food Science & Technology. 2021;1(4):487-94.
51. Plata MR, Koch C, Wechselberger P, Herwig C, Lendl B. Determination of carbohydrates present in *Saccharomyces cerevisiae* using mid-infrared spectroscopy and partial least squares regression. Anal Bioanal Chem. 2013;405(25):8241-50.
52. Liu Y, Huang G, Lv M. Extraction, characterization and antioxidant activities of mannan from yeast cell wall. International Journal of Biological Macromolecules. 2018;118:952-6.
53. Blagović B. Lipid Composition of Brewer's Yeast. Food Technology and Biotechnology. 2001;39:175-81.
54. Blagović B, Mesarić M, Marić V, Rupčić J. Characterization of lipid components in the whole cells and plasma membranes of baker's Yeast. Croatica Chemica Acta. 2005;78:479-84.
55. Avramia I, Amariei S. Spent Brewer’s Yeast as a Source of Insoluble β-Glucans. International Journal of Molecular Sciences. 2021;22(2).
56. Short Francesca L, Di Sario G, Reichmann Nathalie T, Kleanthous C, Parkhill J, Taylor Peter W, et al. Genomic Profiling Reveals Distinct Routes To Complement Resistance in *Klebsiella pneumoniae*. Infection and Immunity. 2020;88(8):e00043-20.
57. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv. 2020;arXiv:2010.16061.
58. Green ML, Karp PD. Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers. Nucleic Acids Res. 2005;33(13):4035-9.
59. Rembeza E, Engqvist MKM. Experimental and computational investigation of enzyme functional annotations uncovers misannotation in the EC 1.1.3.15 enzyme class. PLOS Computational Biology. 2021;17(9):e1009446.
60. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016;3(1):160018.
61. Van Rossum G, & Drake, F. L. Python 3 Reference Manual. Scotts Valley, CA2009.
62. BIOLOG. Phenotype MicroArrays. In: BIOLOG, editor. 2020.
63. Norsigian CJ, Attia H, Szubin R, Yassin AS, Palsson BØ, Aziz RK, et al. Comparative Genome-Scale Metabolic Modeling of Metallo-Beta-Lactamase–Producing Multidrug-Resistant *Klebsiella pneumoniae* Clinical Isolates. Frontiers in Cellular and Infection Microbiology. 2019;9(161).
64. Hall MB. Rasusa: Randomly subsample sequencing reads to a specified coverage 2019 [Available from: https://github.com/mbhall88/rasusa.
65. Krueger F. Trim Galore GitHub2012 [Available from: https://github.com/FelixKrueger/TrimGalore.
66. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLOS Computational Biology. 2017;13(6):e1005595.
67. Wick RR. Dead-end count for QC of short-read assemblies. 2023.
68. RStudio-Team. RStudio: Integrated Development for R. Boston, MA RStudio; 2020.
69. Wickham H, Averick M, Bryan J, Chang W, McGowan LDA, François R, et al. Welcome to the Tidyverse. Journal of Open Source Software. 2019;4(43):1686.
70. Garnier S. viridis: Default Color Maps from 'matplotlib'. 2018.
71. Neuwirth E. RColorBrewer: ColorBrewer Palettes. 1.1-2 ed2014.
72. Kassambara A. ggpubr: 'ggplot2' Based Publication Ready Plots. 0.5.0 ed2022.
73. Aphalo PJ, Slowikowski K, Mouksassi S. ggpmisc: Miscellaneous Extensions to 'ggplot2'. 0.4.1 ed2021.
74. Yu G. aplot: Decorate a 'ggplot' with Associated Information. 0.0.6 ed2020.

75. Zeileis A, Fisher JC, Hornik K, Ihaka R, McWhite CD, Murrell P, et al. colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes. Journal of Statistical Software. 2020;96(1):1 - 49.
76. F.C. M, Davis TL. Package ‘ggpattern’. 2022.
77. Wilke CO. ggtext: Improved Text Rendering Support for 'ggplot2'. 0.1.1 ed2020.
78. Hester J, Bryan J, RStudio. glue: Interpreted String Literals. 1.6.2 ed2022.
79. Muggeo VMR. segmented: Regression Models with Break-Points / Change-Points (with Possibly Random Effects) Estimation. 1.6-2 ed2022.
80. The-Inkscape-Team. Inkscape. 1.0.1 (3bc2e813f5, 2020-09-07) ed2020.
81. zanatlija. Proportional TFB. 2012.
82. weknow. Element. 2015.

Supplementary figures

Figure S1

Flow diagram showing the overview of the draft_model module from Bactabolize, which produces draft metabolic models. Input and output files are shown in light pink while Bactabolize processes are shown in dark pink. Third-party dependencies are indicated within the white boxes.

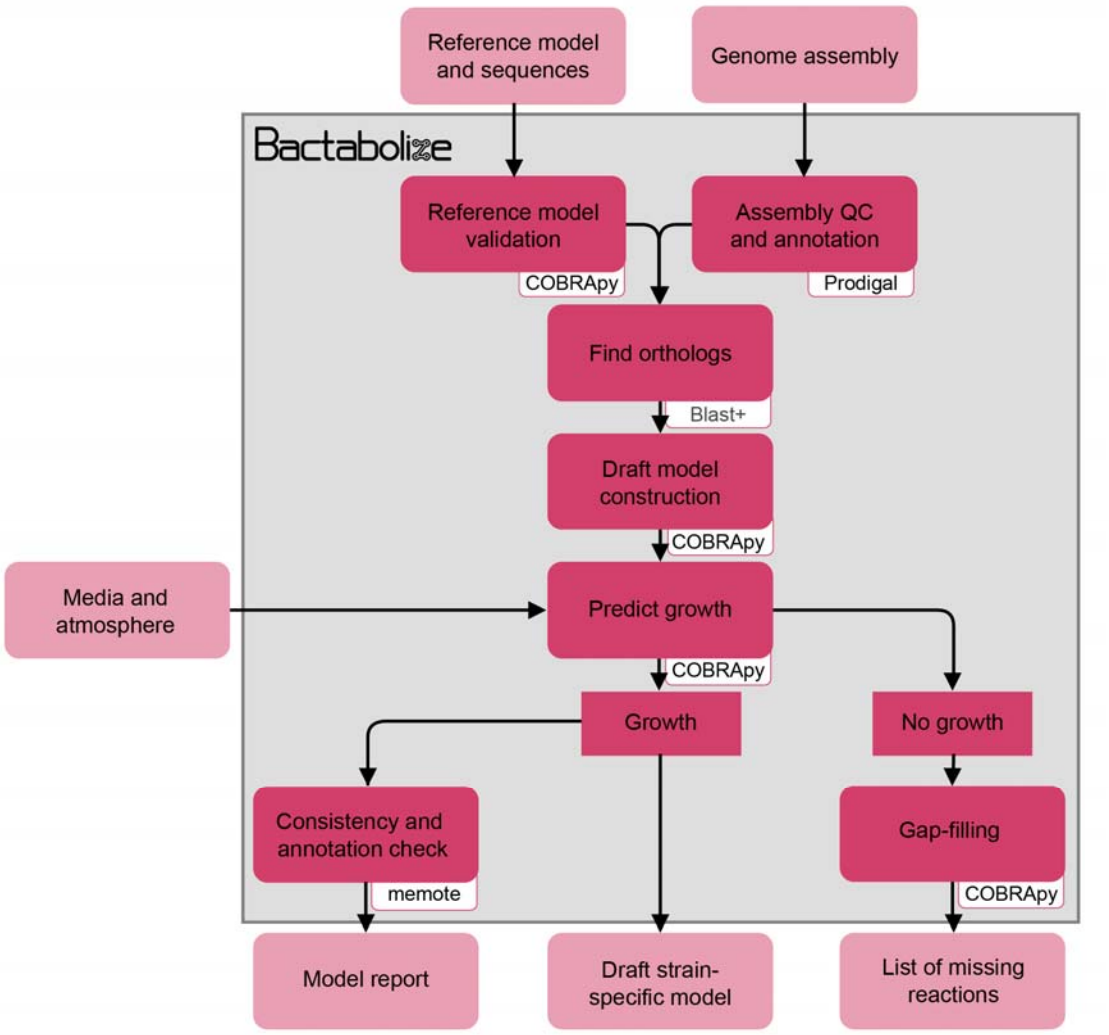


Figure S2

Flow diagram showing the overview of the patch_model module from Bactabolize, which patches metabolic models that do not simulate growth. Input and output files are shown in light orange while Bactabolize processes are shown in dark orange. Third-party dependencies are indicated within the white boxes.

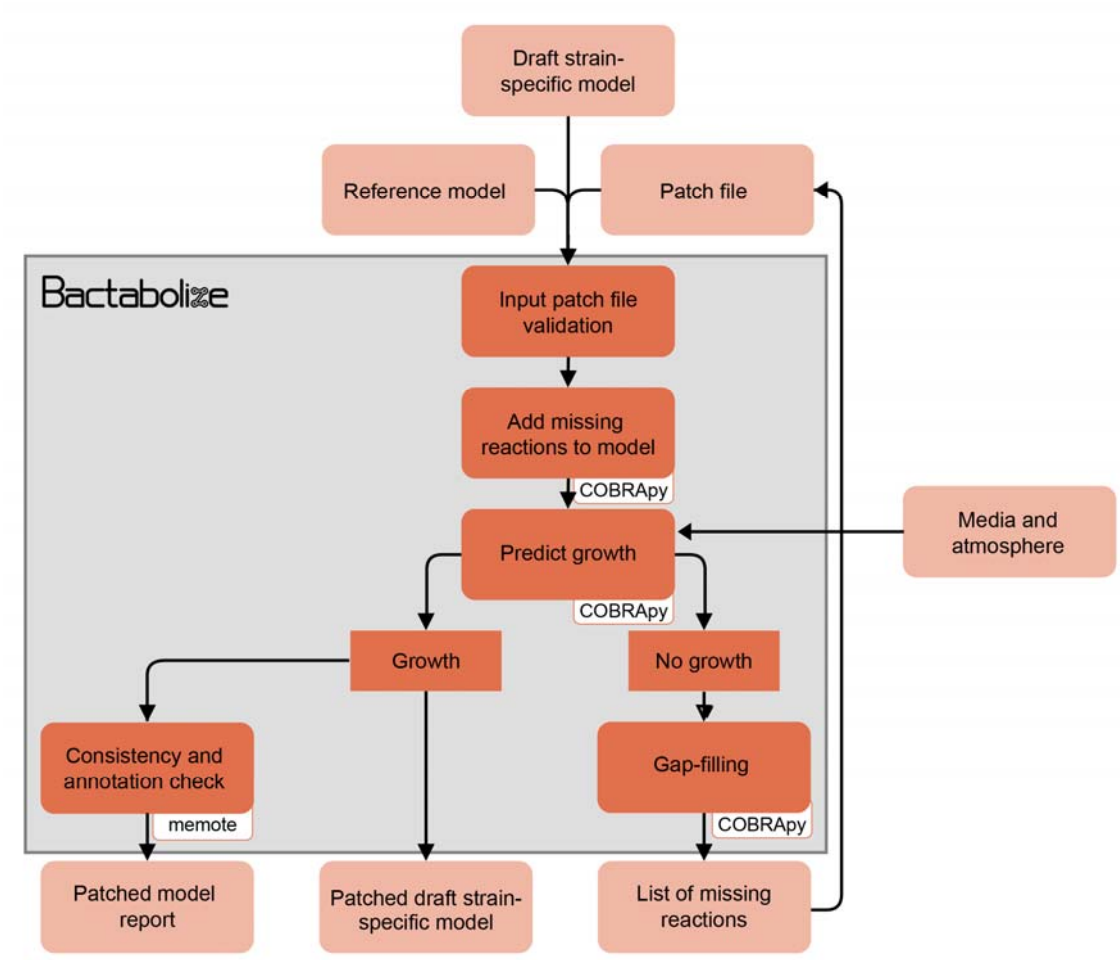


Figure S3

Flow diagram showing the overview of the fba module from Bactabolize, which performs growth simulations using Flux Balance Analysis. Input and output files are shown in light yellow while Bactabolize processes are shown in dark yellow. Third-party dependencies are indicated within white boxes. C, carbon; N, nitrogen, P, phosphorus; S, sulphur; O_2, oxygen.

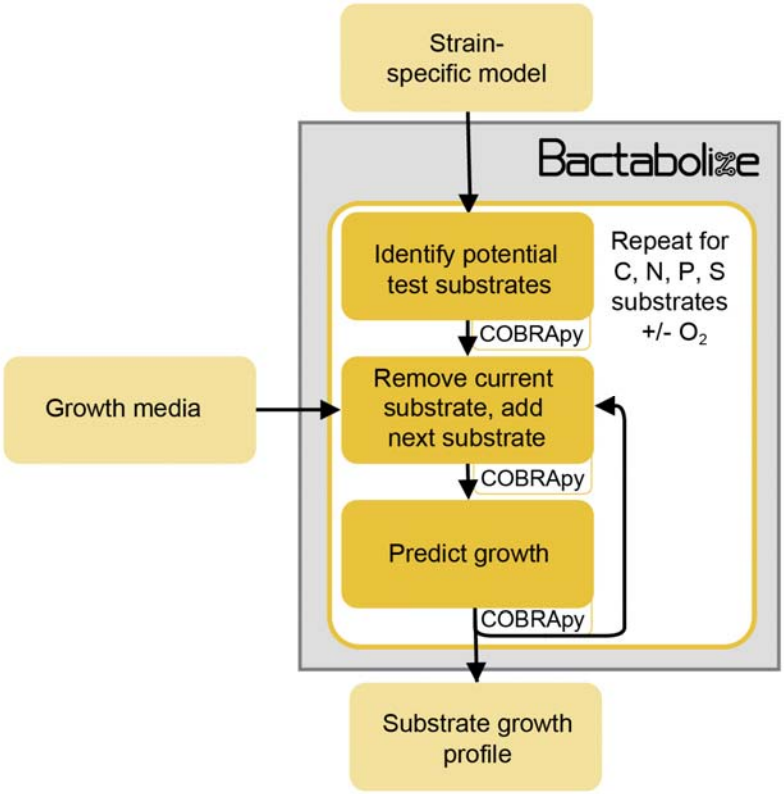
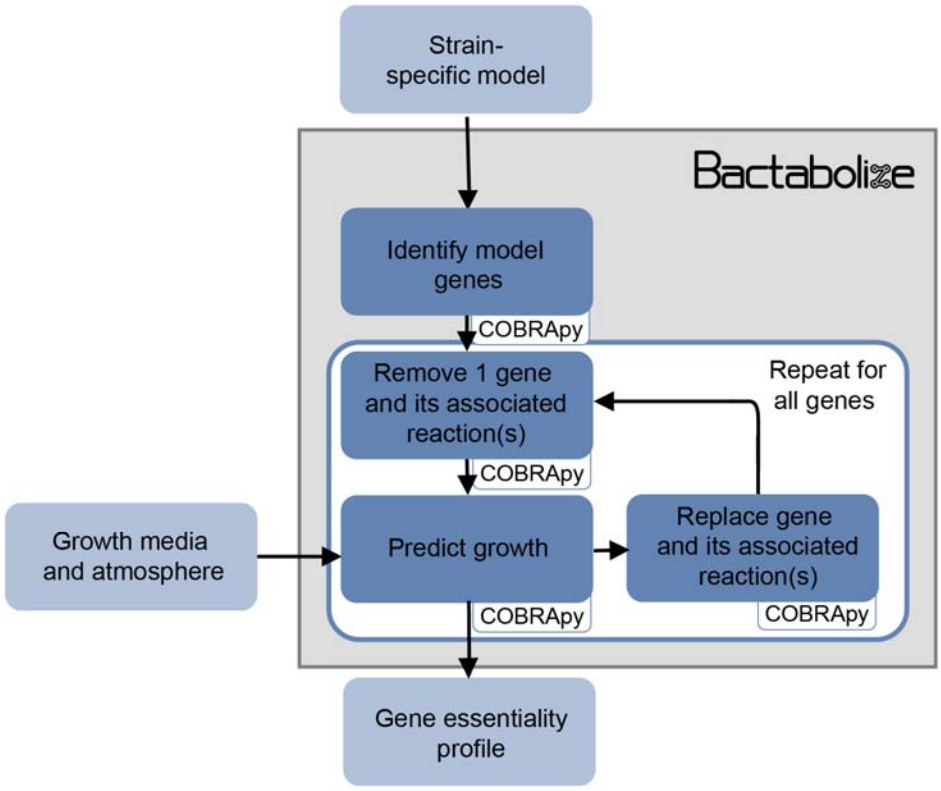


Figure S4

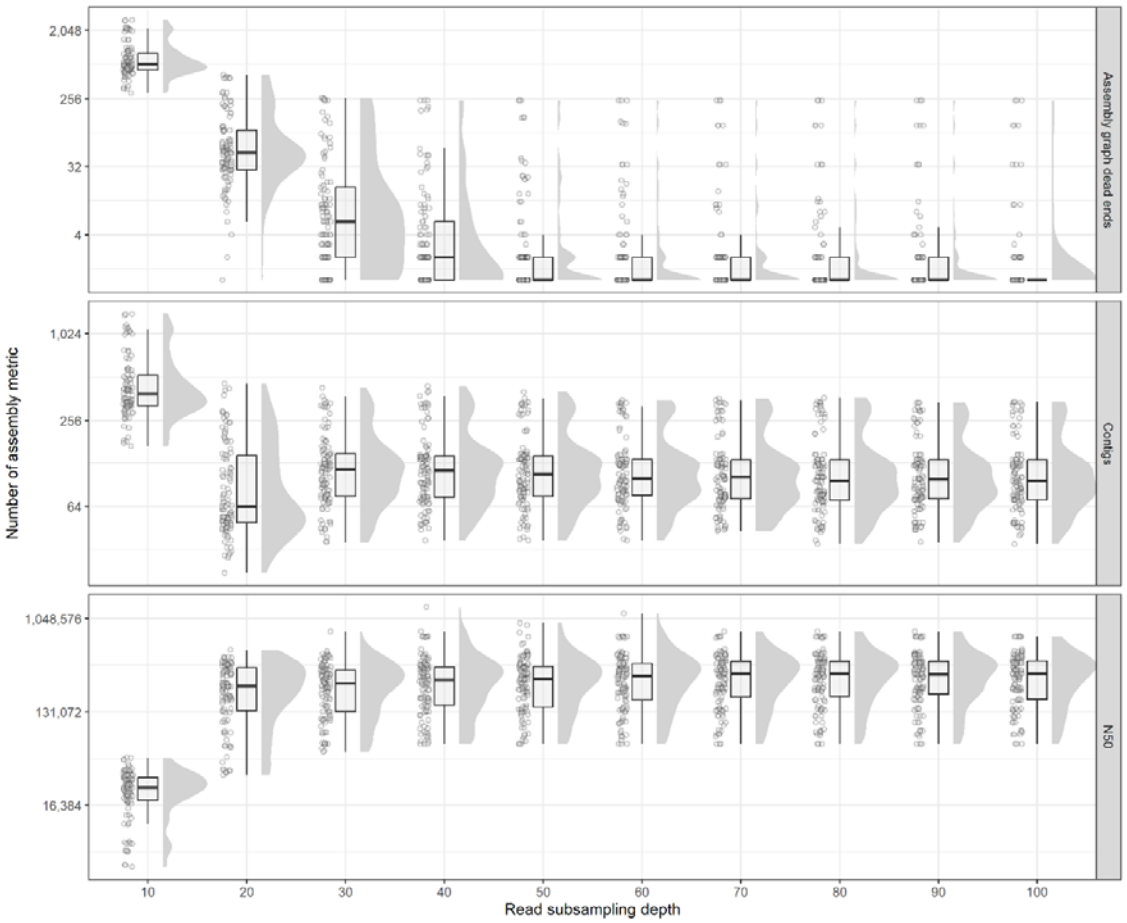
Flow diagram showing the overview of the sgk module from Bactabolize, which performs Single Gene Knockout analysis. Input and output files are shown in light blue while Bactabolize processes are shown in dark blue. Third-party dependencies are indicated within white boxes.



335 **Figure S5**

336 Raincloud plot showing distributions of assembly metrics across various read subsampling depths
337 (10x increments).

338



339

340

341

342

343

344

345

346

347

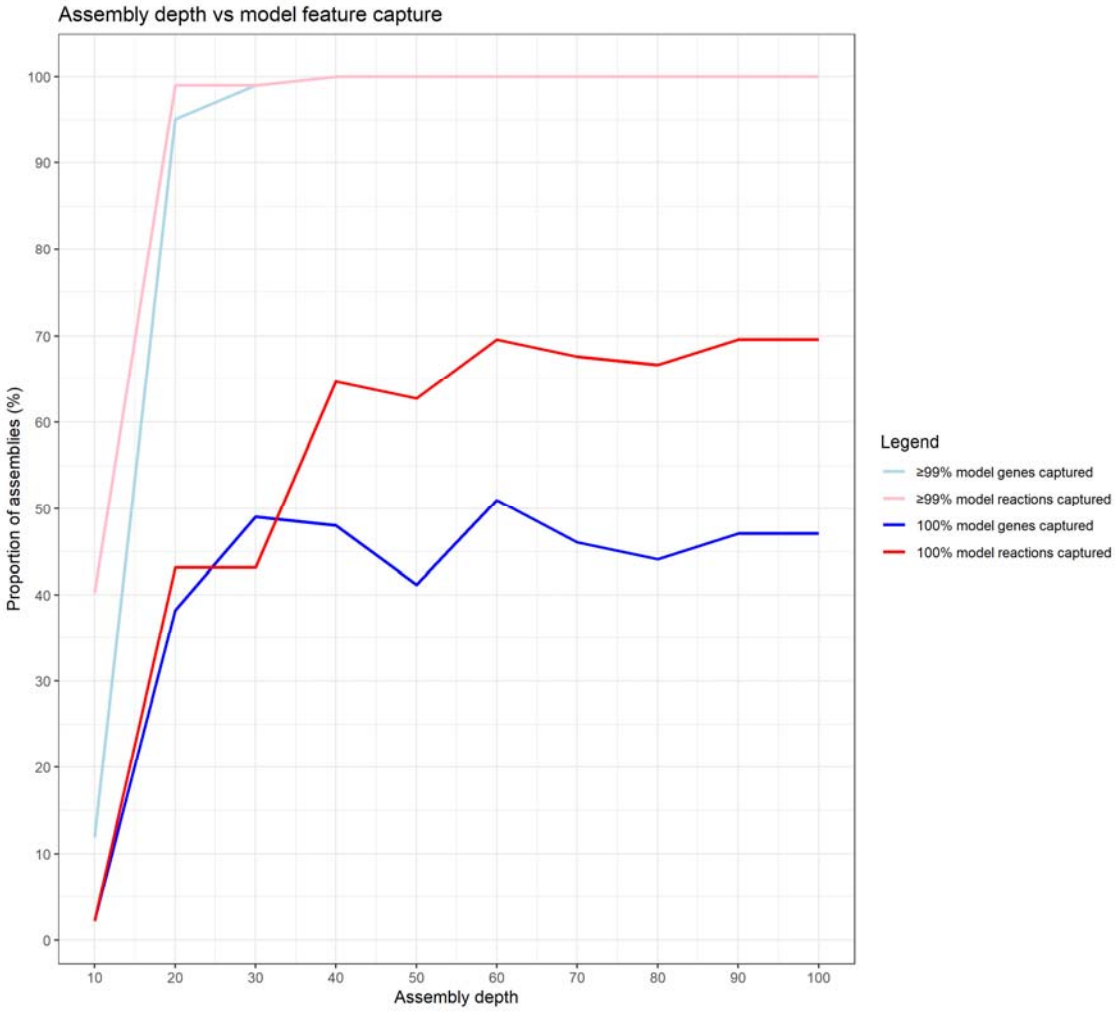
348

349

350

Figure S6

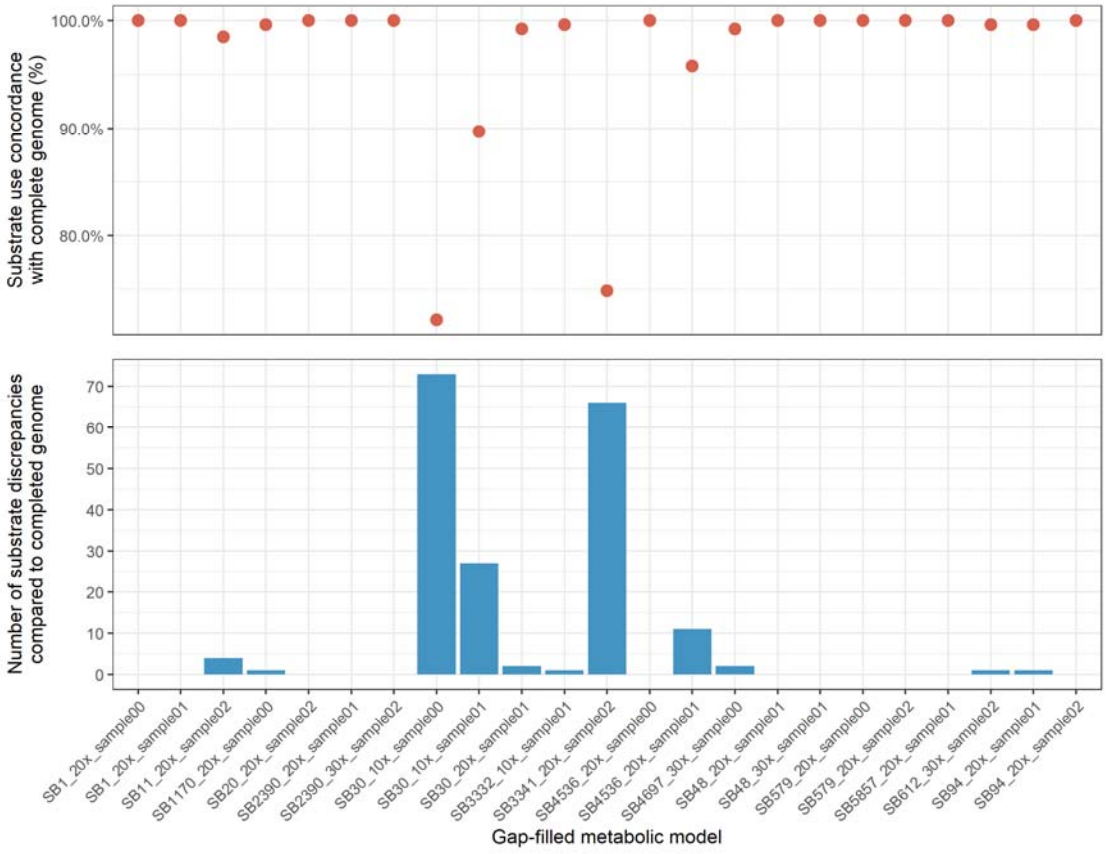
Line graph showing the capture of model features of draft assemblies (short read only) at various depths, compared to the corresponding completed genome (long-read + short read assemblies).



365 ## Figure S7

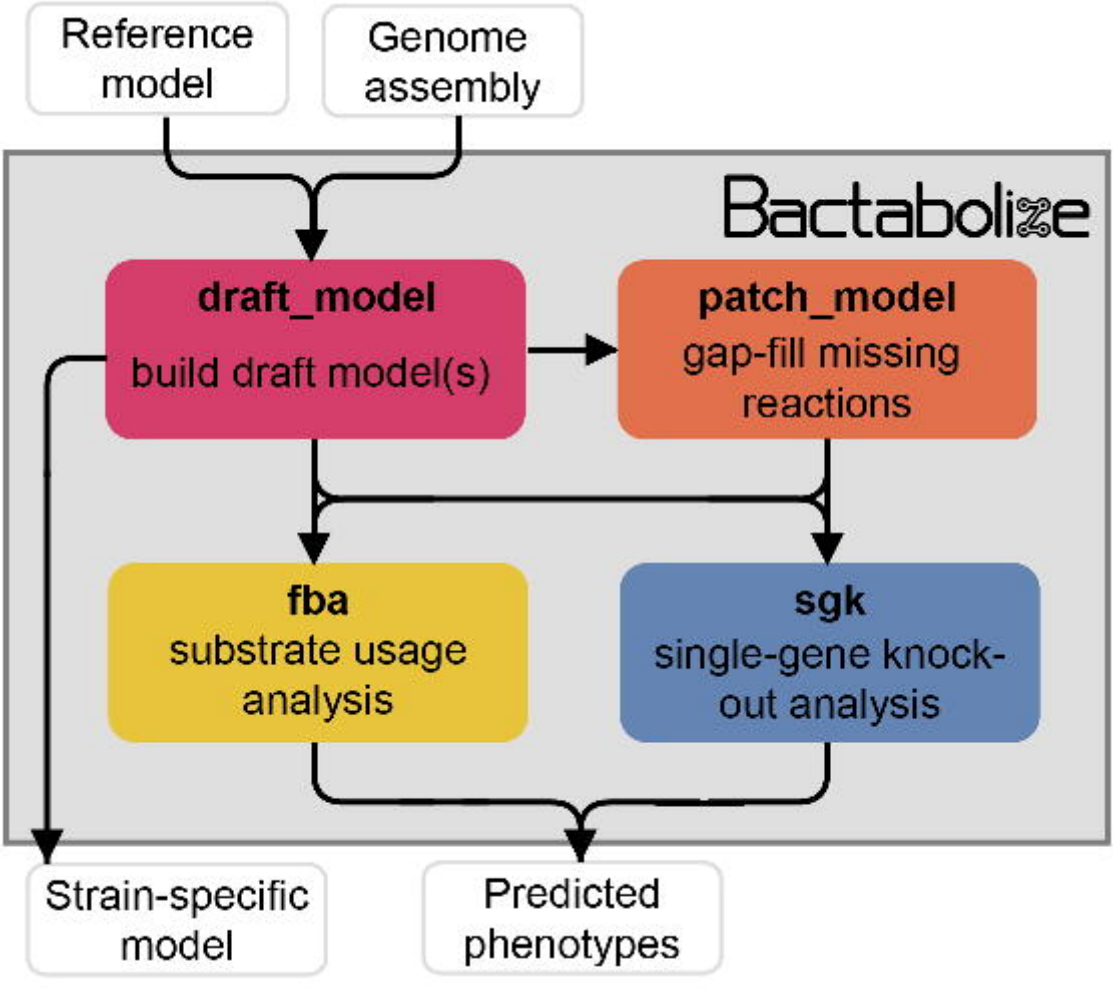
366 Faceted graphs showing the number of substrate usage (fba module) discrepancies of gap-filled
367 models (patch_model module) which initially did not produce biomass (models which failed to
368 simulate growth). The dots indicate percentage concordance with the completed genome model,
369 while the columns indicate number of substrates with discrepancies (no simulated growth in
370 patched model, but growth in completed genome model).

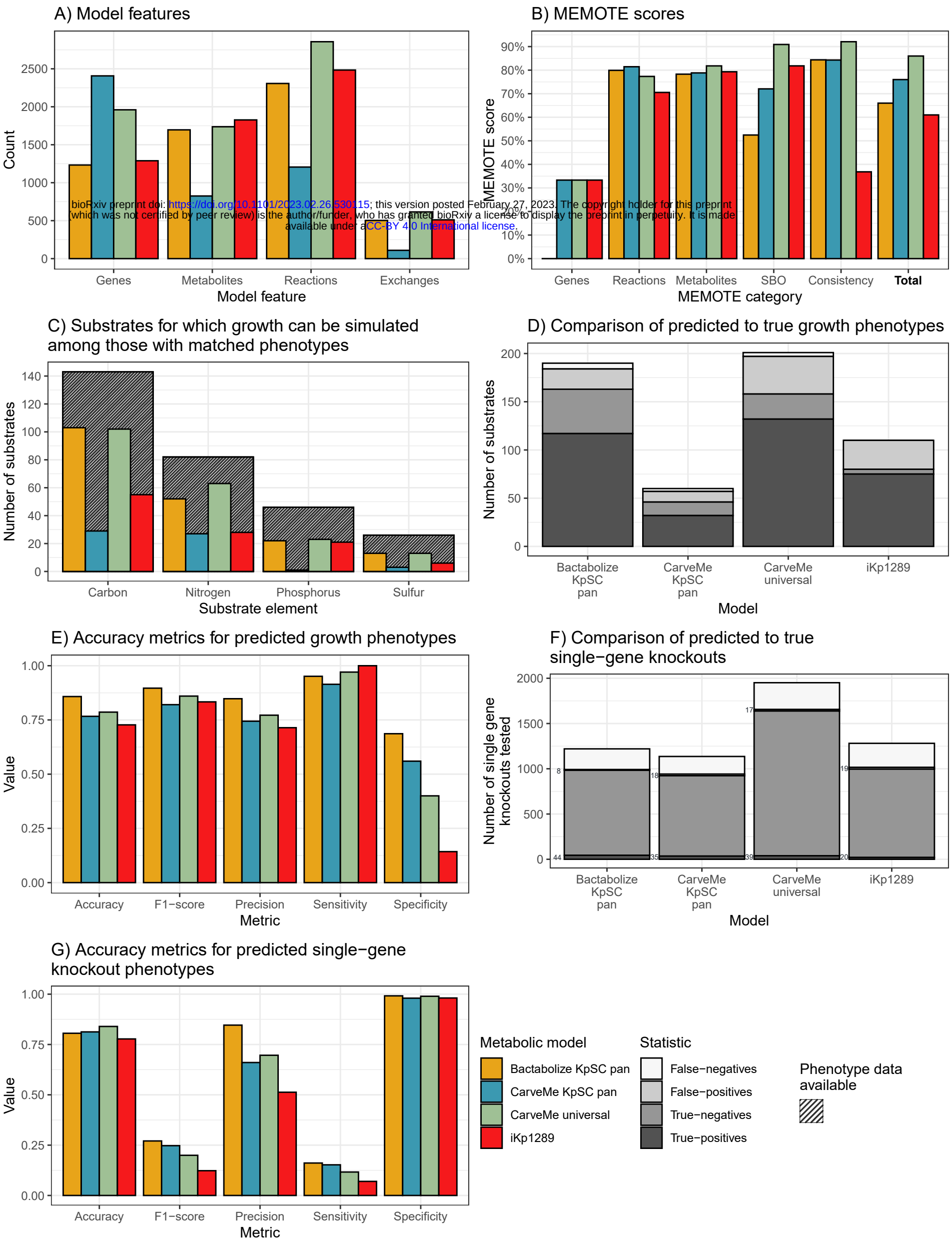
371

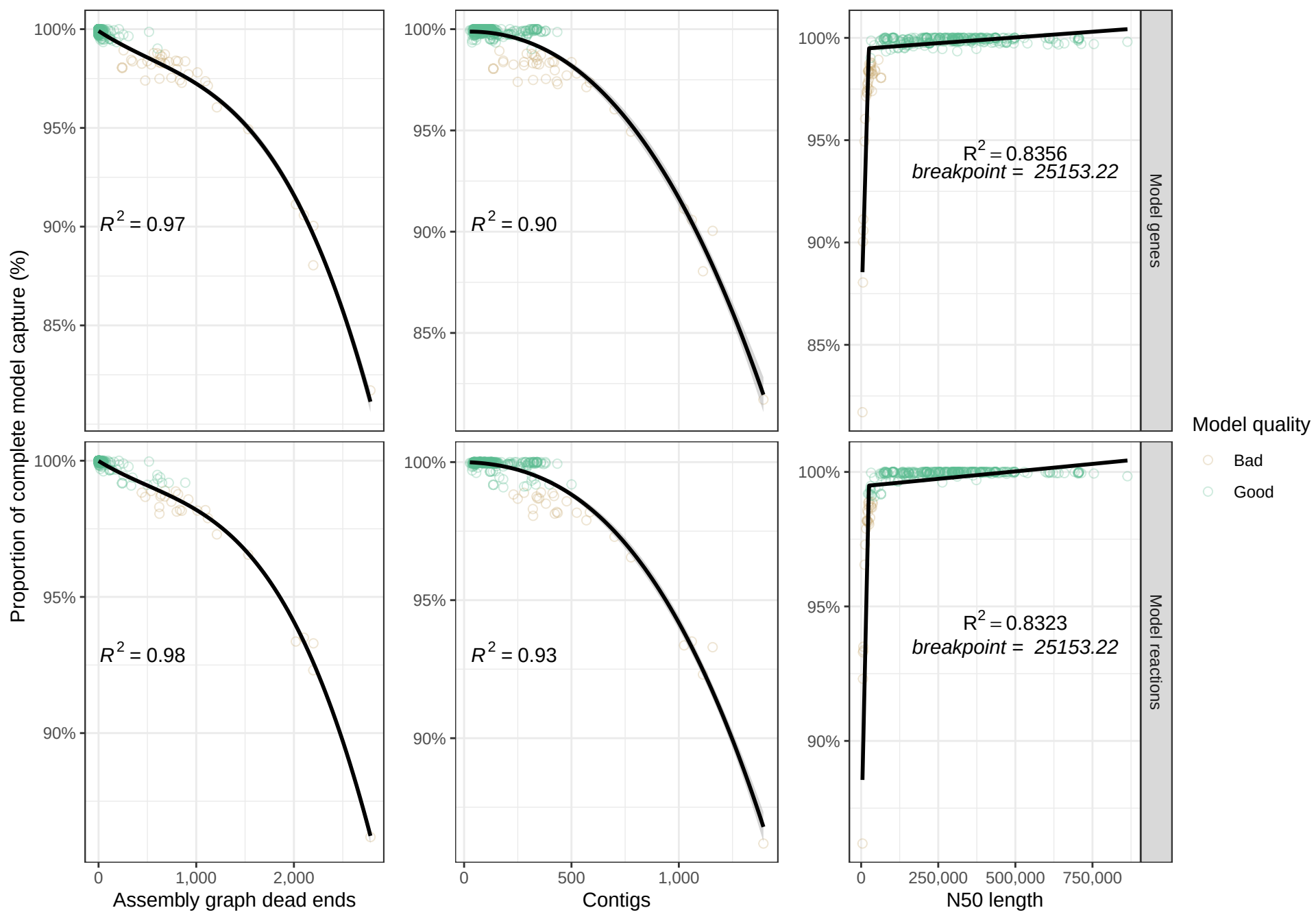


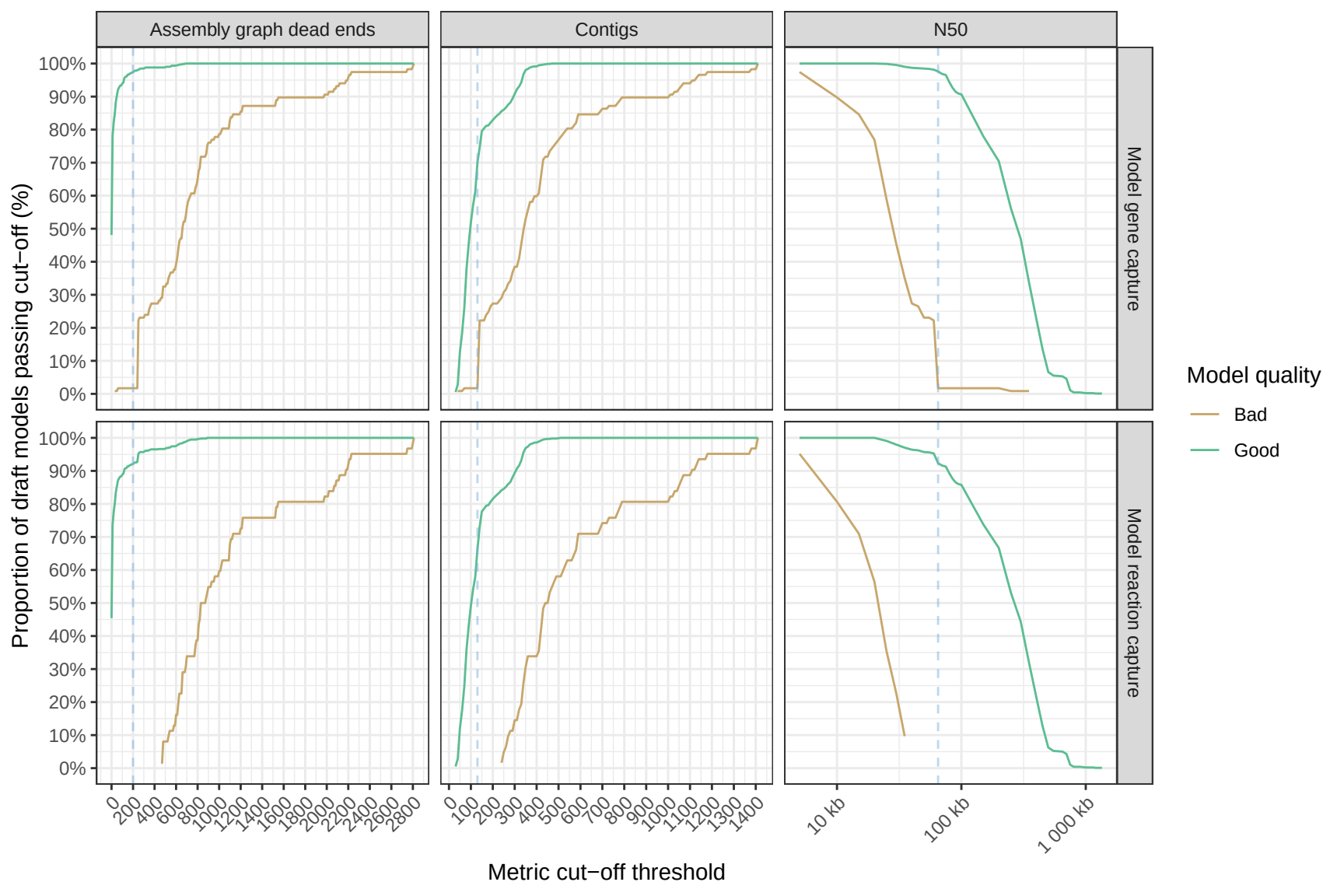
372

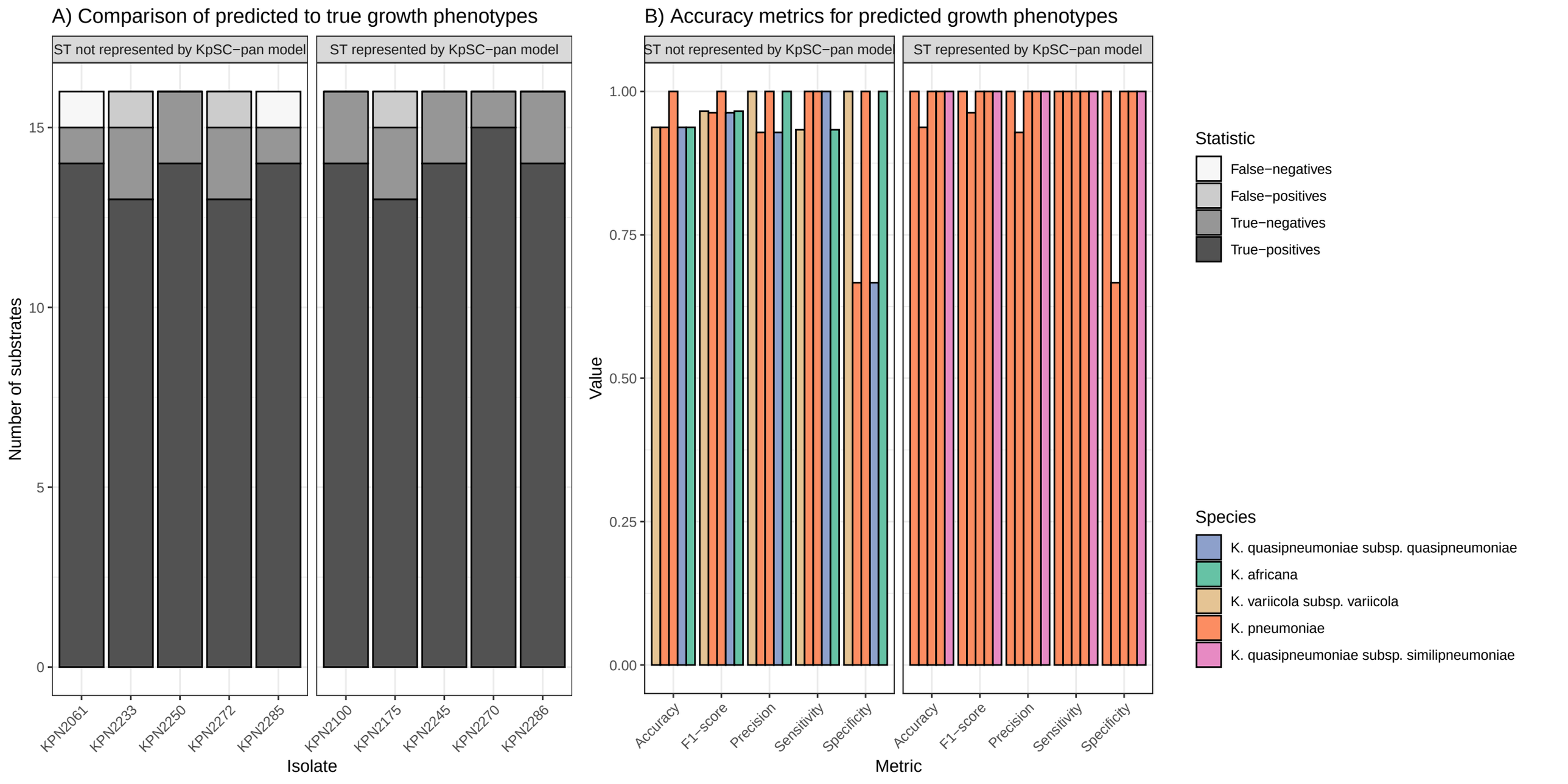
373


Reference model
Genome assembly
Bactabolize
draft_model
build draft model(s)
patch_model
gap-fill missing reactions
fba
substrate usage analysis
sgk
single-gene knock-out analysis
Strain-specific model
Predicted phenotypes




Proportion of complete model capture (%)
100%
95%
90%
85%
R^2 = 0.97
R^2 = 0.90
R^2 = 0.98
R^2 = 0.93
R^2 = 0.8356
breakpoint = 25153.22
R^2 = 0.8323
breakpoint = 25153.22
Model genes
Model reactions
0
1,000
2,000
500
250,000
500,000
750,000
Assembly graph dead ends
Contigs
N50 length
Model quality
Bad
Good


Assembly graph dead ends
Contigs
N50
Model gene capture
Model reaction capture
Proportion of draft models passing cut−off (%)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
2400
2600
2800
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
10 kb
100 kb
1 000 kb
Metric cut−off threshold
Model quality
Bad
Good


A) Comparison of predicted to true growth phenotypes
ST not represented by KpSC−pan model
ST represented by KpSC−pan model
Number of substrates
0
5
10
15
KPN2061
KPN2233
KPN2250
KPN2272
KPN2285
KPN2100
KPN2175
KPN2245
KPN2270
KPN2286
Isolate
B) Accuracy metrics for predicted growth phenotypes
ST not represented by KpSC−pan model
ST represented by KpSC−pan model
Value
0.00
0.25
0.50
0.75
1.00
Accuracy
F1−score
Precision
Sensitivity
Specificity
Metric
Statistic
False−negatives
False−positives
True−negatives
True−positives
Species
K. quasipneumoniae subsp. quasipneumoniae
K. africana
K. variicola subsp. variicola
K. pneumoniae
K. quasipneumoniae subsp. similipneumoniae