Cameron et al., 2022

plastiC: A pipeline for recovery and characterization of plastid genomes from metagenomic datasets

Supplementary Materials

1. Model Training, Testing & Validation

1.1 Preparation of Training and Testing Datasets

Reference protein sequences were downloaded from complete plastid genomes available in RefSeq (n = 2633) using Entrez Direct (Kans, 2022) on 2022-11-2. 261 of these references were randomly removed and stored separately to be used as an independent testing set. The training set (n = 2372) was further refined to exclude any reference genomes that were smaller than the median plastid genome size of protists (94 kb) to ensure that outlier small genomes do not impact the training of the model.

diamond blastp was run on both training and testing datasets with UniRef100 to obtain KEGG annotations. These KEGG annotations were then used to calculate KEGG module completeness. For training sets, reference genome KEGG annotation counts were subsampled without replacement to create simulated examples of plastid genomes with lower levels of completeness ranging from 0 – 100% in increments of 5%. Test reference genomes were also subsampled to create simulated examples to validate the robustness of the completeness estimates ranging from 10 – 100% in increments of 10%.

1.2 Model Training, Cross-Validation & Testing

Scikit-learn was utilised to develop and validate machine learning models for estimating metagenomic plastid genome completeness. Specifically, Ada boosting, gradient boosting and random forest regressions were evaluated to determine the effectiveness for estimating completeness of plastid genomes. Training data were split 90 training set:10 test set. K-folds (n=5) cross-validation with shuffling was performed to cross-validate the model.

Each model was then tested on reference plastid genomes that were not used in the training set. In addition to the test plastid genome set, KEGG module completeness was predicted for a mitochondrial set (n = 142) to evaluate whether the model could accurately differentiate between different organellar genomes. Across the cross-validated model set, all three regression models were able to differentiate between plastid and mitochondrial completeness (*i.e.,* predict low completeness for mitochondria; high completeness for plastids; Table S2). However, the Ada boosting regression had the lowest plastid completeness estimate and highest mitochondrial estimate. In combination with the higher mean-squared error, this suggests that the Ada boosting regression model is not the best performing model for application in plastid completeness estimates.

In addition to the evaluation of the test plastid genomes when complete, this testing set was subsampled to lower completeness levels to examine the performance of the model with plastids of varying completeness. Predicted completeness values were compared to expected values at the subset levels to determine efficacy of the model on accurate estimation. Median prediction values and standard deviation was calculated for each iteration of the model produced through cross-validation (Table S3). A linear regression was performed (Figure S1; Table S4) on the predicted completeness compared to expected value and Pearson's correlation R^2 was calculated to identify similarity between predicted and expected completeness scores. All regressions and correlation coefficients were statistically significant ($p < 2e-16$) but the gradient boosting regression model showed the highest correlation between expected and predicted values. To conclude evaluation of the effectiveness of each model, differences between the expected and predicted values for each cross-validated model iteration were calculated (Figure S2). The random forest regression model had the best performing mean-squared error, highest completeness estimate for whole plastid genomes and lowest completeness estimates for whole mitochondrial genomes. However, it frequently overestimated completeness (median difference = -7.5). Based on the correlation between expected and predicted and median value of discrepancy between predicted and expected values (n = 0.37), the gradient boosting regression model was identified as the best-performing model for plastid genome completeness estimates.

2. Case Study: Lichen Metagenomes

Lichens are composite organisms composed of the symbiotic association between a primary fungal partner (mycobiont) and algal partner (photobiont). Chlorophyta (green alga) taxa are the photobiont in many lichen species suggesting that plastids should be present in lichen metagenomic samples. To test the effectiveness of *plastiC* on metagenomic data, lichen metagenomes were downloaded from the project accession PRJNA646656 (n = 13) in the European Nucleotide Archive. These samples were derived from 10 species of lichen spanning 6 genera (Table S5) which are all expected to have *Trebouxia*, a green algal genus, as their primary photobiont. Downloaded metagenomic datasets were filtered using *fastp* and human contamination was removed using *BMTagger*. Quality-controlled reads were assembled using *metaSPAdes*. These assemblies were used with *plastiC* to recover plastid genomes.

Plastid contigs were identified in all 13 samples using *Tiara* (Table S6). Metagenomic assemblies were binned using *metaBAT2* with the reduced bin size threshold of 50 kb. These bins were then searched to identify location of the identified plastid contigs based on contig identifiers. Bins that were composed of >90% plastid nucleotide were retained as probable plastid bins for further analysis. Of the 13 lichen metagenomes analysed, a single plastid bin was identified in 8 of them. For the remaining 5 samples, plastid contigs were not successfully binned and were retained in the unbinned portion with other sequences.

Taxonomic source prediction was performed on the identified plastid bins. All plastid bins identified in the sample were attributed to *Trebouxia* which corresponds with expectations of the photobiont in these lichens being a trebouxoid green alga. Plastid bins ranged in estimated completeness from 10.77 to 96.39% and completeness was positively correlated to the bin span in these examples.

Table S1: Mean squared errors for trained regression models with k-fold cross-validation (n = 5; shuffled).

	Ada boosting regression	Gradient boosting regression	Random forest regression
CV1	0.0033	0.0003	0.0001
CV2	0.0034	0.0003	0.0001
CV3	0.0038	0.0003	0.0001
CV4	0.0044	0.0004	0.0001
CV5	0.0038	0.0003	0.0001
Mean	0.00374	0.00032	0.0001

Table S2: Median predicted completeness values ± standard deviation on whole plastid and mitochondrial reference genomes with k-fold cross validation (n = 5; shuffled) to ensure differentiation between organellar genome completeness scores.

	Ada boosting regression		Gradient boosting regression		Random forest regression	
	Plastids	Mitochondria	Plastids	Mitochondria	Plastids	Mitochondria
CV1	91.79 ± 6.36	5.6 ± 0.00	98.11 ± 7.27	0.43 ± 0.00	99.06 ± 6.97	0.06 ± 0.00
CV2	91.98 ± 6.42	5.40 ± 0.00	97.97 ± 7.49	0.52 ± 0.00	99.00 ± 7.17	0.05 ± 0.00
CV3	92.15 ± 6.43	4.83 ± 0.00	97.99 ± 7.11	0.47 ± 0.00	99.08 ± 6.97	0.05 ± 0.00
CV4	91.54 ± 6.31	6.81 ± 0.00	98.10 ± 7.60	0.45 ± 0.00	98.98 ± 7.21	0.05 ± 0.00
CV5	91.51 ± 6.55	5.49 ± 0.00	97.95 ± 7.39	0.44 ± 1.73	99.05 ± 7.09	0.05 ± 0.00
Mean	91.79 ± 6.41	5.62 ± 0.00	98.02 ± 7.36	0.45 ± 0.35	99.03 ± 7.08	0.053 ± 0.00

Table S3: Median predicted completeness values for plastid reference genomes in independent testing validation set (n = 261). Test set plastid references were subsampled down to 10% to evaluate the quality of completeness estimates provided to non-complete genomes.

Expected Completeness	Ada Boosting Regression	Gradient Boosting Regression	Random Forest Regression
100	92.56 ± 6.56	99.28 ± 7.41	100.00 ± 7.11
90	92.33 ± 6.58	95.78 ± 7.34	95.05 ± 6.72
80	85.07 ± 8.13	83.99 ± 8.39	87.50 ± 8.40
70	64.92 ± 7.85	69.63 ± 6.51	82.25 ± 7.95
60	49.21 ± 6.28	51.05 ± 5.22	60.00 ± 4.69
50	74.42 ± 7.65	68.87 ± 7.50	75.00 ± 7.07
40	49.04 ± 5.21	38.66 ± 5.03	57.10 ± 6.49
30	27.21 ± 3.88	36.73 ± 5.08	57.35 ± 8.76
20	28.68 ± 11.20	25.89 ± 6.73	34.80 ± 8.99
10	11.63 ± 4.71	8.34 ± 2.66	6.79 ± 6.18

Table S4: Linear regression and Pearson correlation as calculated comparing the predicted completeness to expected completeness on the independent test plastid genome estimations.

	Ada Boosting Regression	Gradient Boosting Regression	Random Forest Regression
Linear Equation	y = 0.88x + 9.19	y = 0.95x + 4.58	y = 0.87x + 16.85
Adjusted R^2	0.83	0.89	0.83
p-value (Linear regression)	<2.20e-16	<2.20e-16	<2.20e-16
Pearson’s Correlation	0.91	0.95	0.91
p-value (Pearson’s)	<2.20e-16	<2.20e-16	<2.20e-16

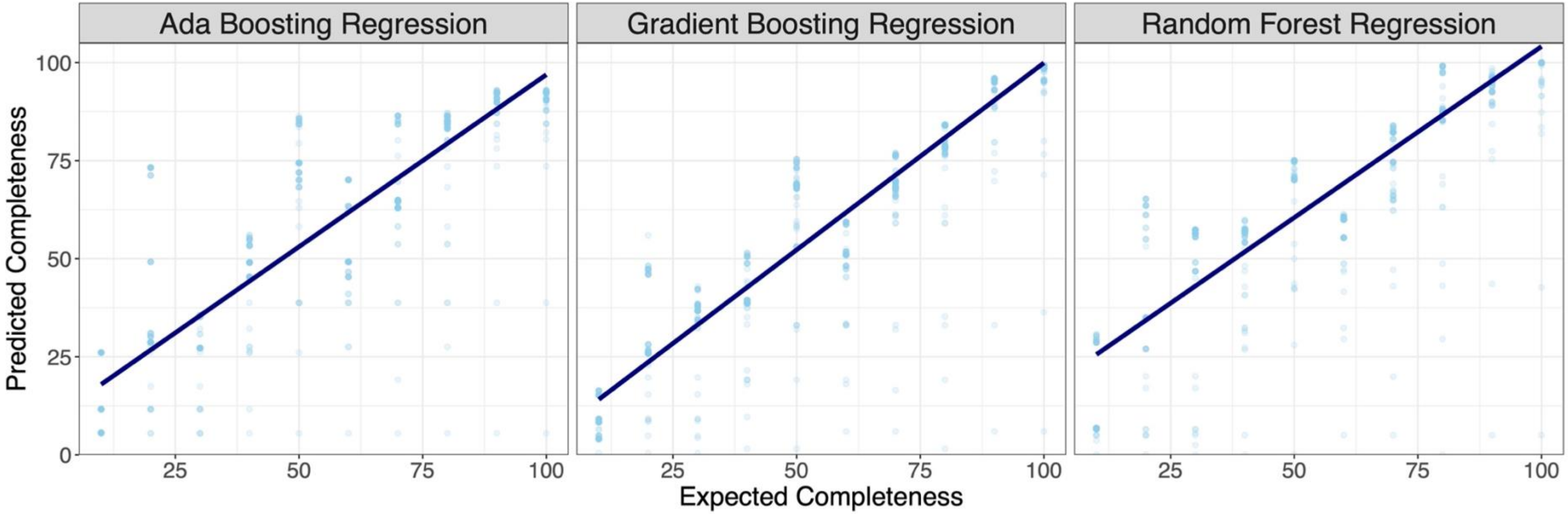

Figure S1: Predicted completeness estimates on test plastid genomes (not used in model training) with three different regression models: Ada boosting, gradient boosting and random forest. The resulting linear equation on linear regression of predicted ~ expected values is plotted to demonstrate the predictive performance of the different models.

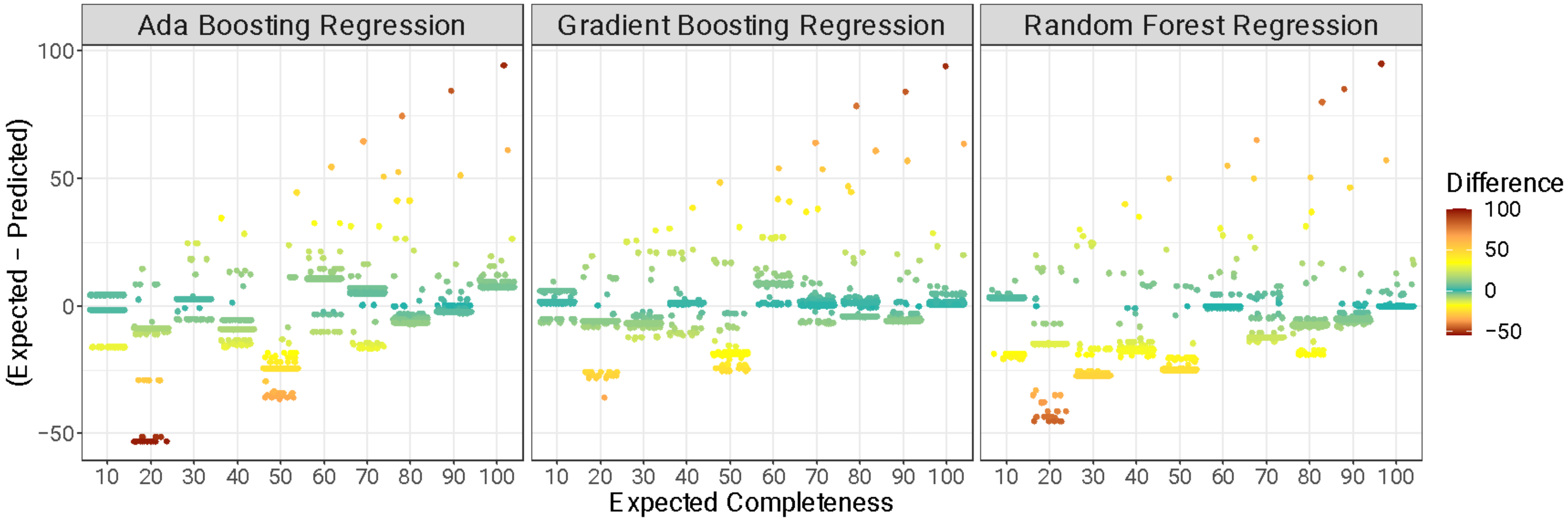

Figure S2: Differences in expected completeness of subsampled plastid reference genomes and predicted completeness. Positive values in difference indicate underestimation of plastid predicted completeness while negative values represent overestimation. Gradient boosting regression consistently had the smallest median discrepancy between predicted values and expected value (0.37), while random forest had the largest discrepancy frequently resulting in the overestimation of predicted plastid completeness (median = -7.5). Ada boosting regression performed moderately (median = -2.12).

plastiC: A pipeline for recovery and characterization of plastid genomes from metagenomic datasets

Table S5: Sample accession information for PRJNA646656.

Run Accession	Sample Accession	Secondary Study Accession	Scientific Name
SRR12240187	SAMN15548970	SRP272267	*Xanthoparmelia chlorchroa*
SRR12240188	SAMN15548969	SRP272267	*Xanthoparmelia chlorochroa*
SRR12240177	SAMN15548974	SRP272267	*Xanthoparmelia maricopensis*
SRR12240174	SAMN15548977	SRP272267	*Xanthoparmelia neocumberlandia*
SRR12240175	SAMN15548976	SRP272267	*Xanthoparmelia neocumberlandia*
SRR12240180	SAMN15548971	SRP272267	*Xanthopermelia chlorochoa*
SRR12240179	SAMN15548972	SRP272267	*Xanthopermelia mexicana*
SRR12240178	SAMN15548973	SRP272267	*Xanthopermelia plittii*
SRR12240185	SAMN15548980	SRP272267	*Mobergia calculiformis*
SRR12240183	SAMN15548982	SRP272267	*Physcia biziana*
SRR12240182	SAMN15548983	SRP272267	*Physciella chloantha*
SRR12240181	SAMN15548984	SRP272267	*Rinodina sp*
SRR12240184	SAMN15548981	SRP272267	*Oxernella safavidorum*

Table S6: Lichen metagenome assembly and binning information. Metagenomes were assembled into contigs and these assemblies were used to identify plastid contigs using *Tiara* and for binning with *metaBAT2*. Probable plastid bins were identified based on the distribution and location of plastid contigs within bins, with a threshold of >90% to be retained for downstream analyses.

Run Accession	Total Contigs	Plastid Contigs	Total Bins	Probable Plastid Bins
SRR12240187	190210	12	15	1
SRR12240188	115748	12	18	1
SRR12240177	364371	20	23	1
SRR12240174	253456	108	23	0
SRR12240175	171912	93	13	1
SRR12240180	164812	14	20	1
SRR12240179	253280	3	24	1
SRR12240178	190261	81	17	1
SRR12240185	129614	505	28	1
SRR12240183	330121	64	27	0
SRR12240182	650113	31	26	0
SRR12240181	170759	207	28	0
SRR12240184	455310	587	22	0

Table S7: Probable plastid bin characteristics including bin span, total contig count and number of plastid contigs. Completeness estimates were performed based on KEGG module coverage and gradient boosting regression, and taxonomic association of plastid genomes performed with *CAT*.

Run Accession	Bin Span	Total Contig Count	Number of Plastid Contigs	Completeness Estimate	Taxonomic Association
SRR12240187	231882	1	1	96.05	*Trebouxia*
SRR12240188	231734	2	2	96.05	*Trebouxia*
SRR12240177	238137	10	10	95.85	*Trebouxia*
SRR12240174	N/A	N/A	N/A	N/A	N/A
SRR12240175	121980	15	14	33.37	*Trebouxia*
SRR12240180	254207	5	5	96.05	*Trebouxia*
SRR12240179	258520	1	1	96.39	*Trebouxia*
SRR12240178	64547	11	10	10.77	*Trebouxia*
SRR12240185	77893	9	8	19.08	*Trebouxia*
SRR12240183	N/A	N/A	N/A	N/A	N/A
SRR12240182	N/A	N/A	N/A	N/A	N/A
SRR12240181	N/A	N/A	N/A	N/A	N/A
SRR12240184	N/A	N/A	N/A	N/A	N/A