

INTNet: a deep learning approach for integron identification and
classification
Supplementary Material

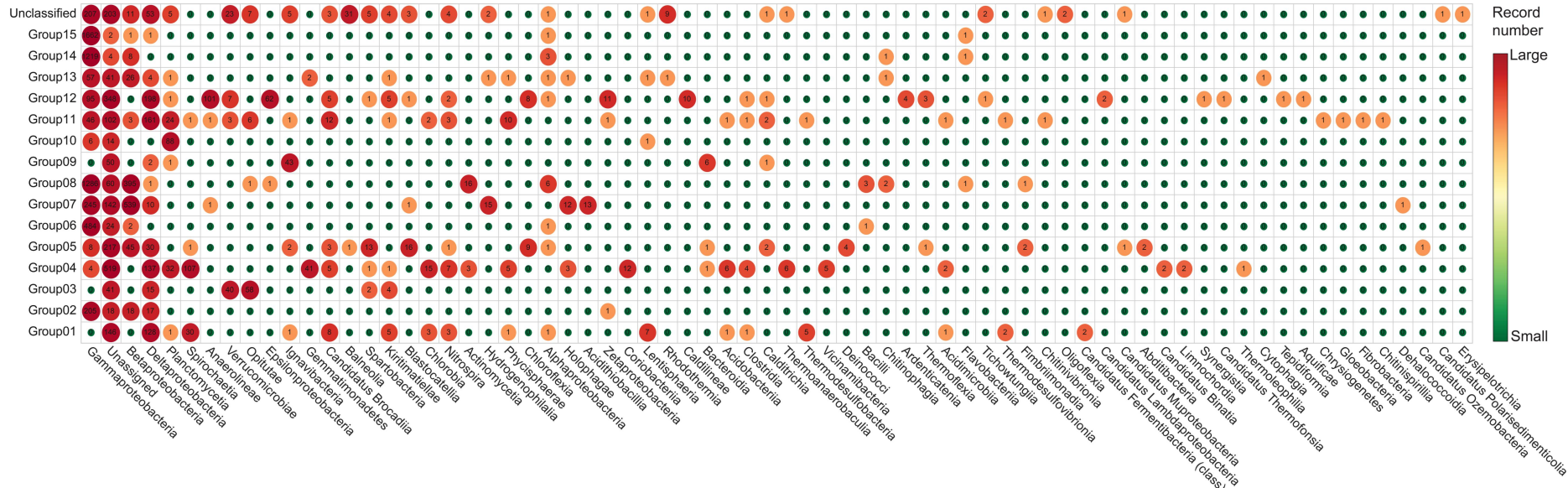
IntegronFinder testing parameters

IntegronFinder version 2.0.2 was used for integron identification. After receiving input DNA sequences, IntegronFinder annotated the CDS with Prodigal, which were analyzed via HMMer for IntI detection and Infernal for *attC* detection.

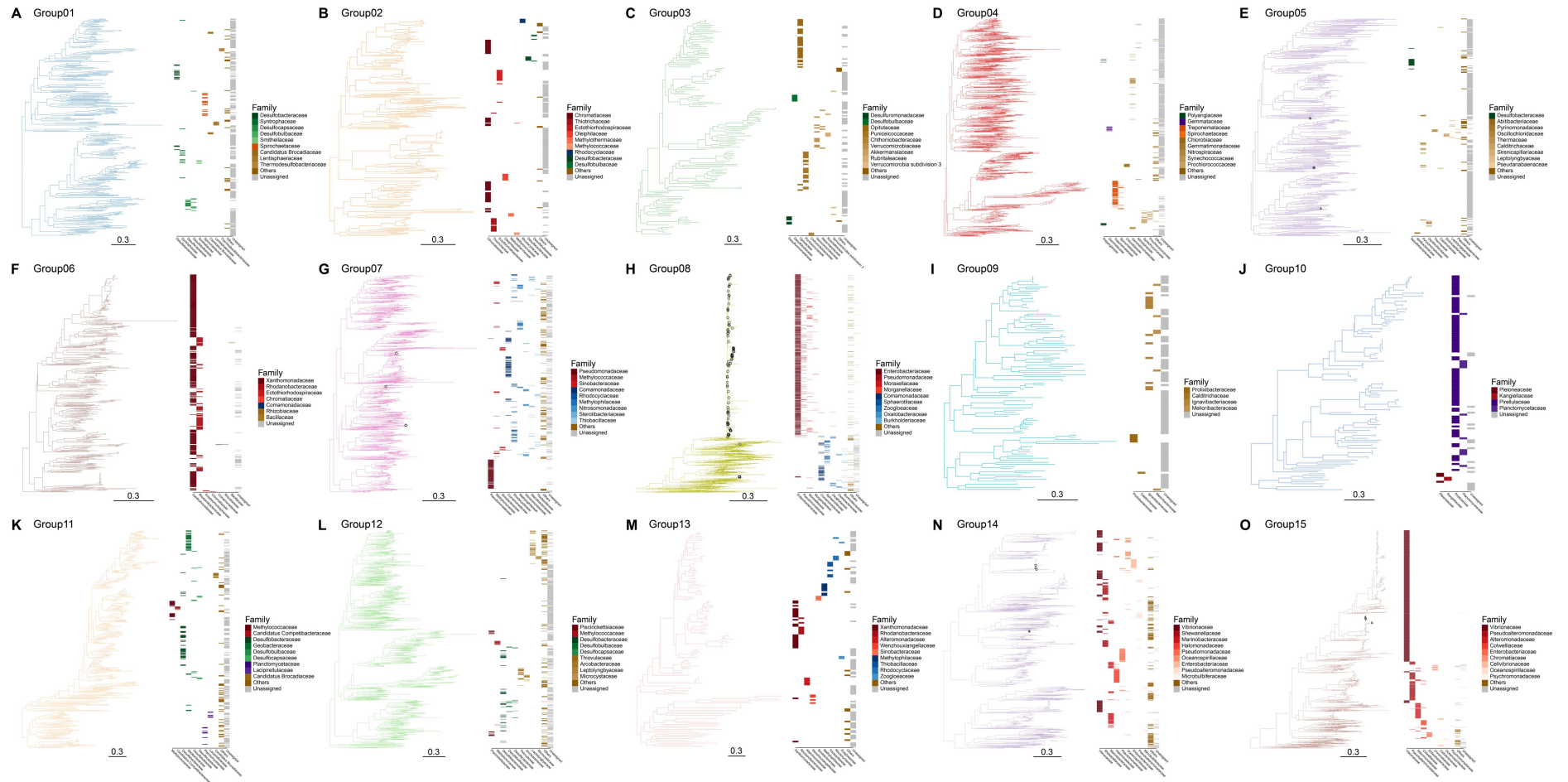
The detailed command to run IntegronFinder for testing was:

```
integron_finder --local-max --func-annot testing_data.fasta
```

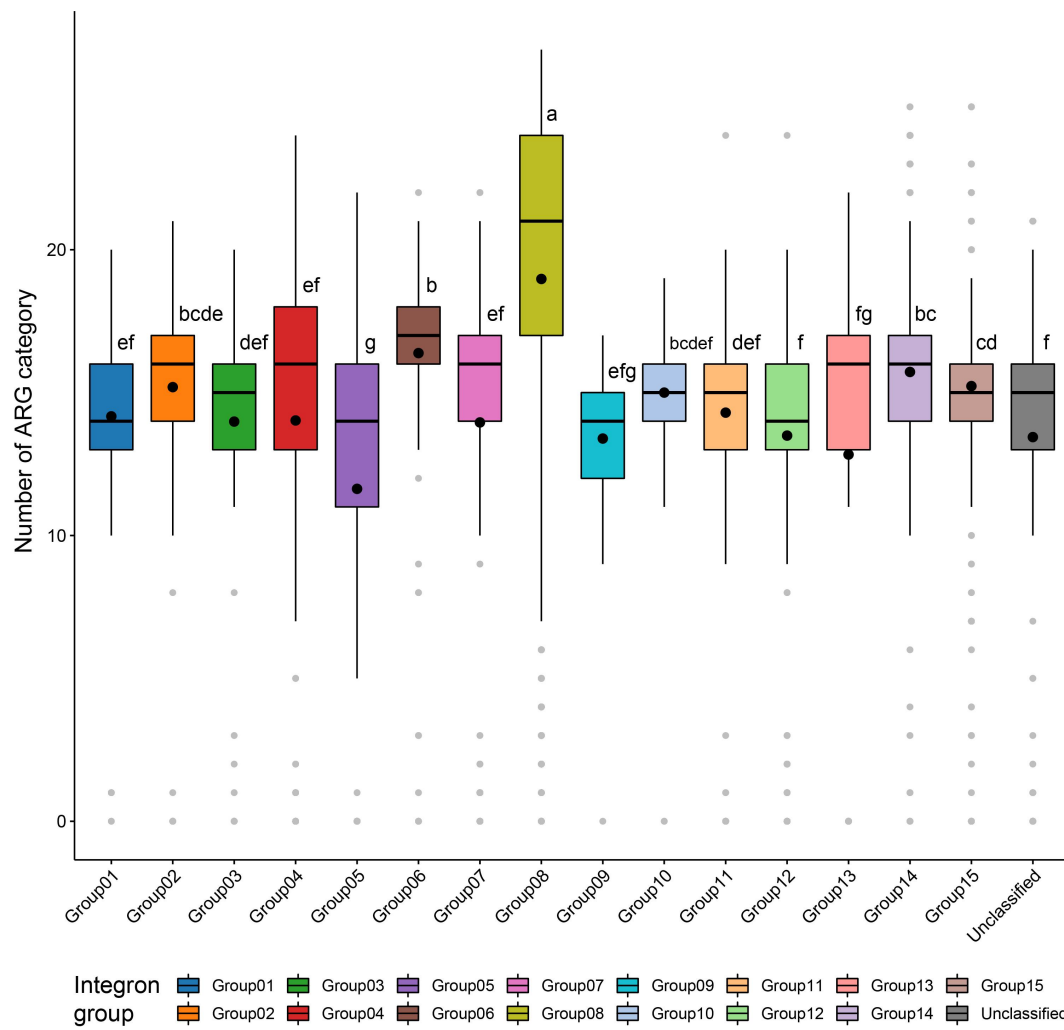
Supplementary Figures



Supplementary Fig. 1 | Bacterial hosts of integron. Circle size and color represent the rank of number of integron records per integron group and bacteria class. The number in each circle shows the absolute number of integron records.

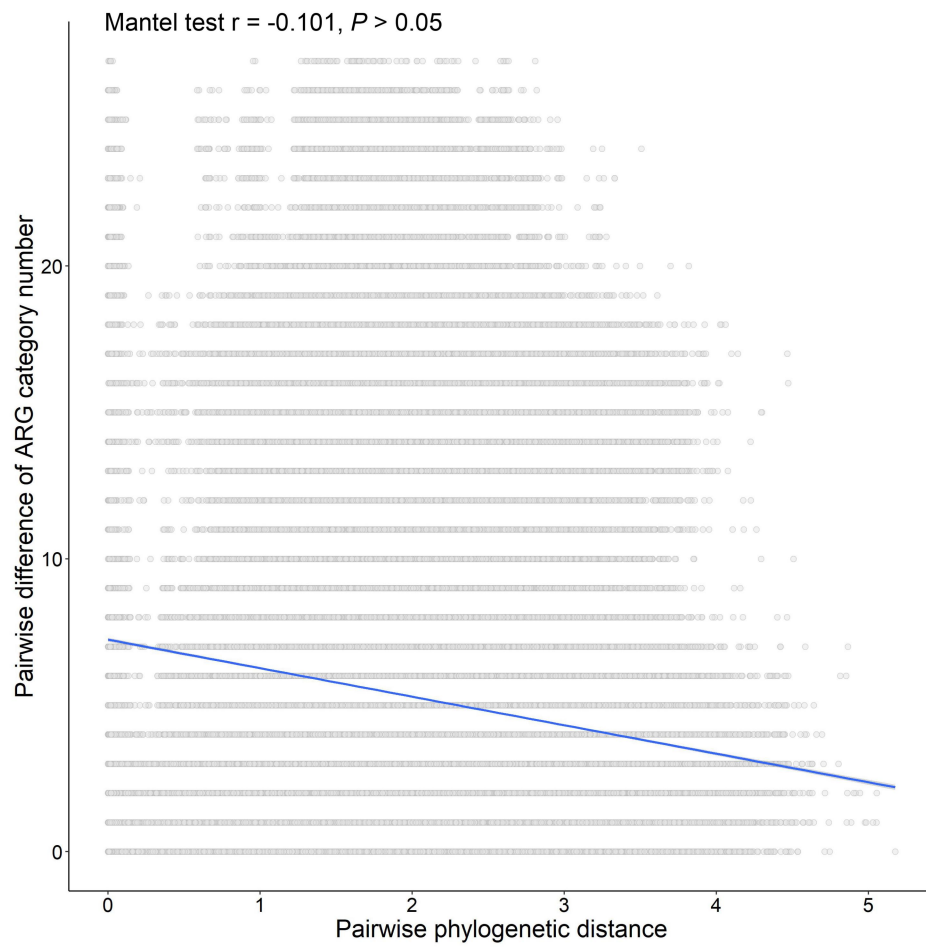


Supplementary Fig. 2 | Phylogenetic tree collection of integron. Subset trees were obtained from Fig. 1d, with the same branch and node colors. Bacterial family information is annotated for each sequence by the right heatmap. (A-O) Group01 to Group15.

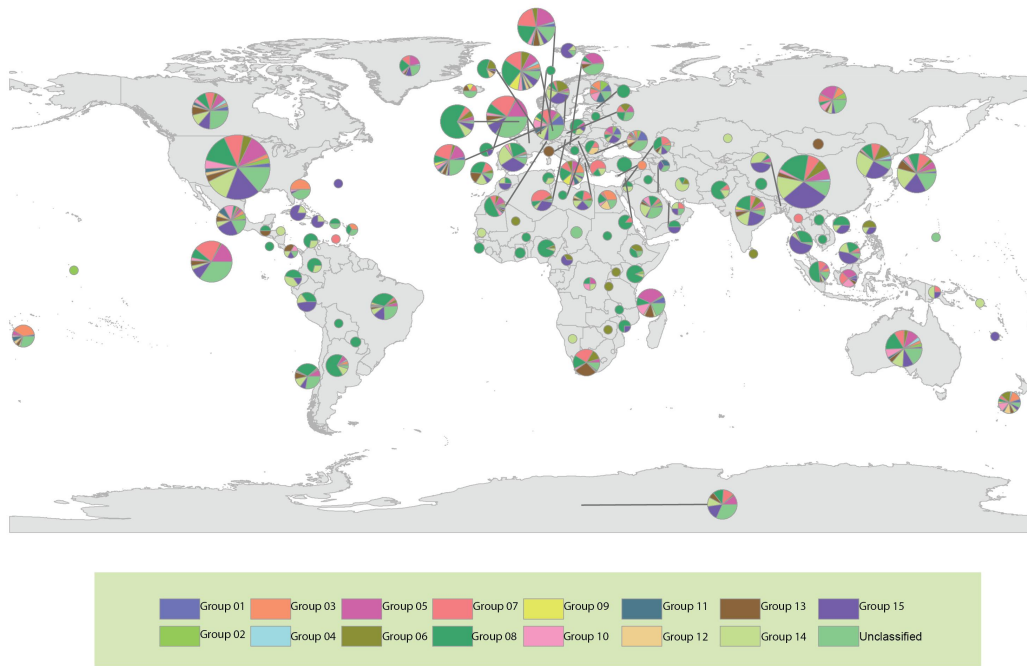


Supplementary Fig. 3 | Distribution of ARG category number per integron record. Boxplots are colored for different integron groups. Pairwise comparison is indicated by

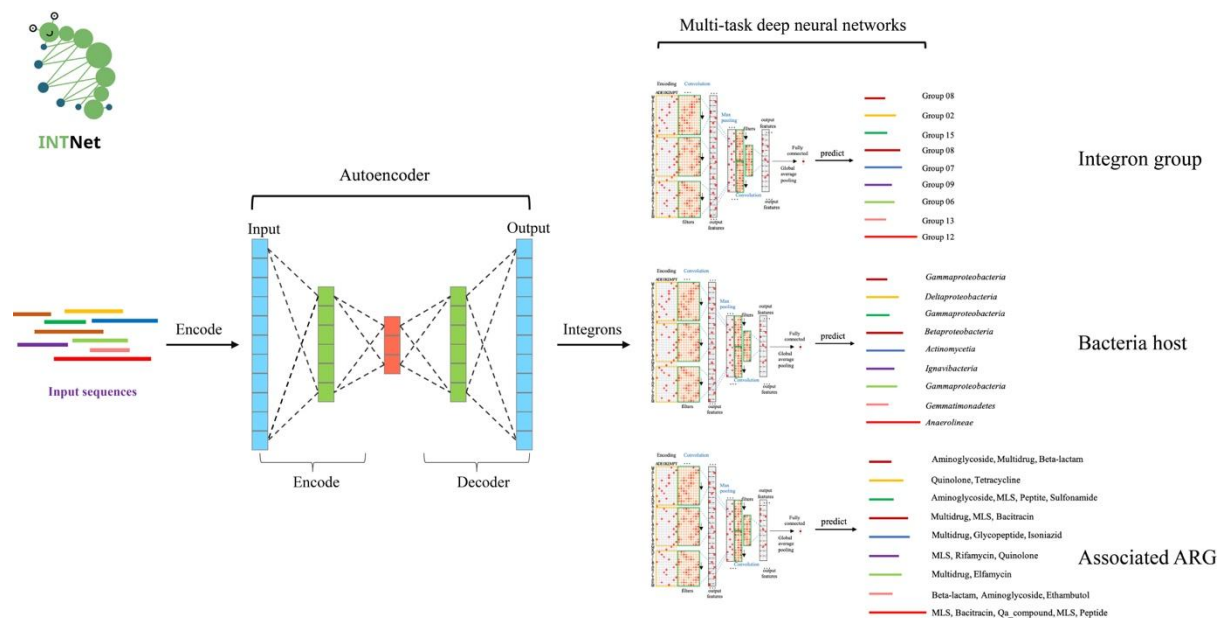
compact letter display.



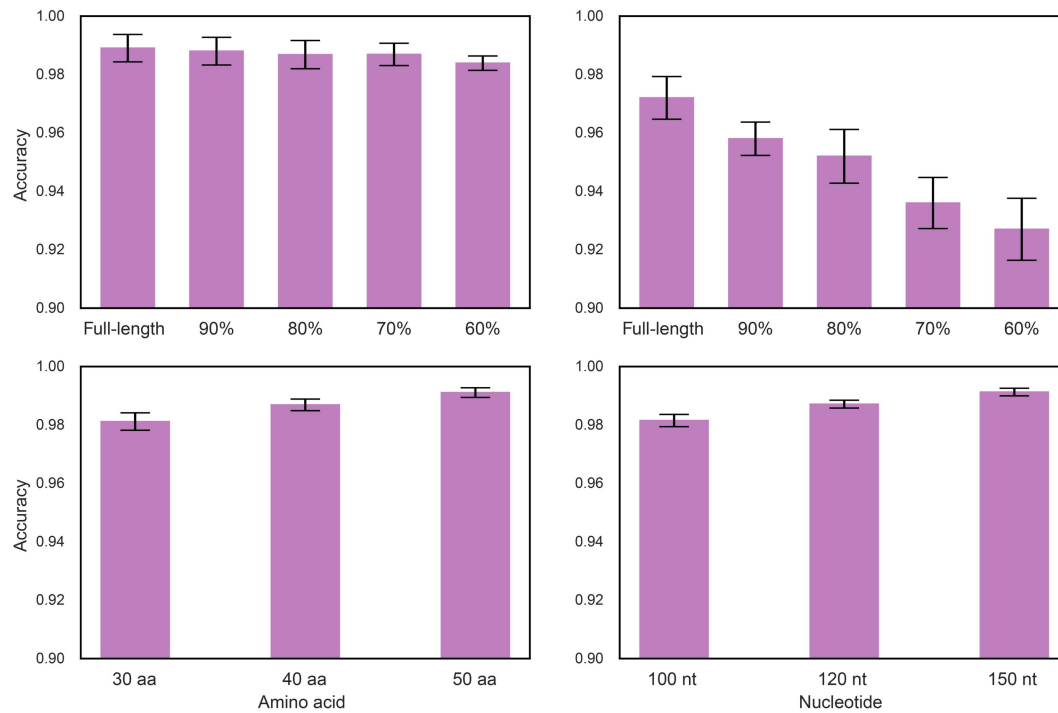
Supplementary Fig. 4 | Scatter plot for Mantel test between phylogenetic distance and difference of ARG category number. Measurements were calculated based on integron integrase phylogenetic tree. Mantel statistic r and associated P value, as well as linear regression line are shown.



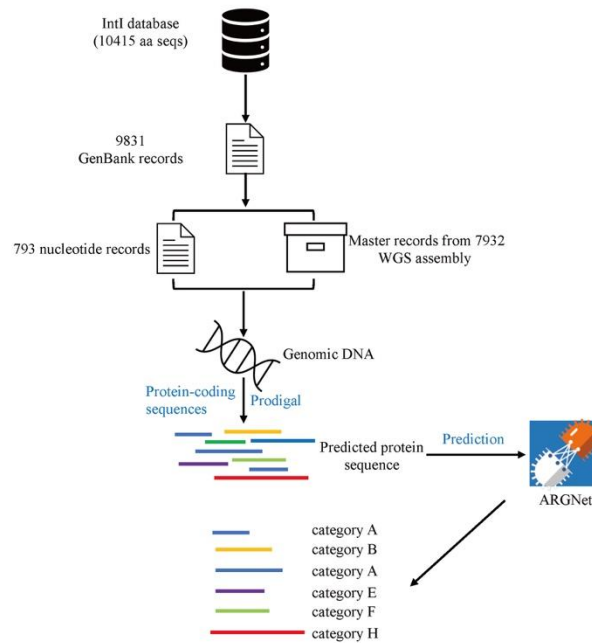
Supplementary Fig. 5 | Geographical distribution of integron. Proportion of integron records from different groups is shown in the pie chart for each country and region, with the pie chart diameter proportional to the number of records.



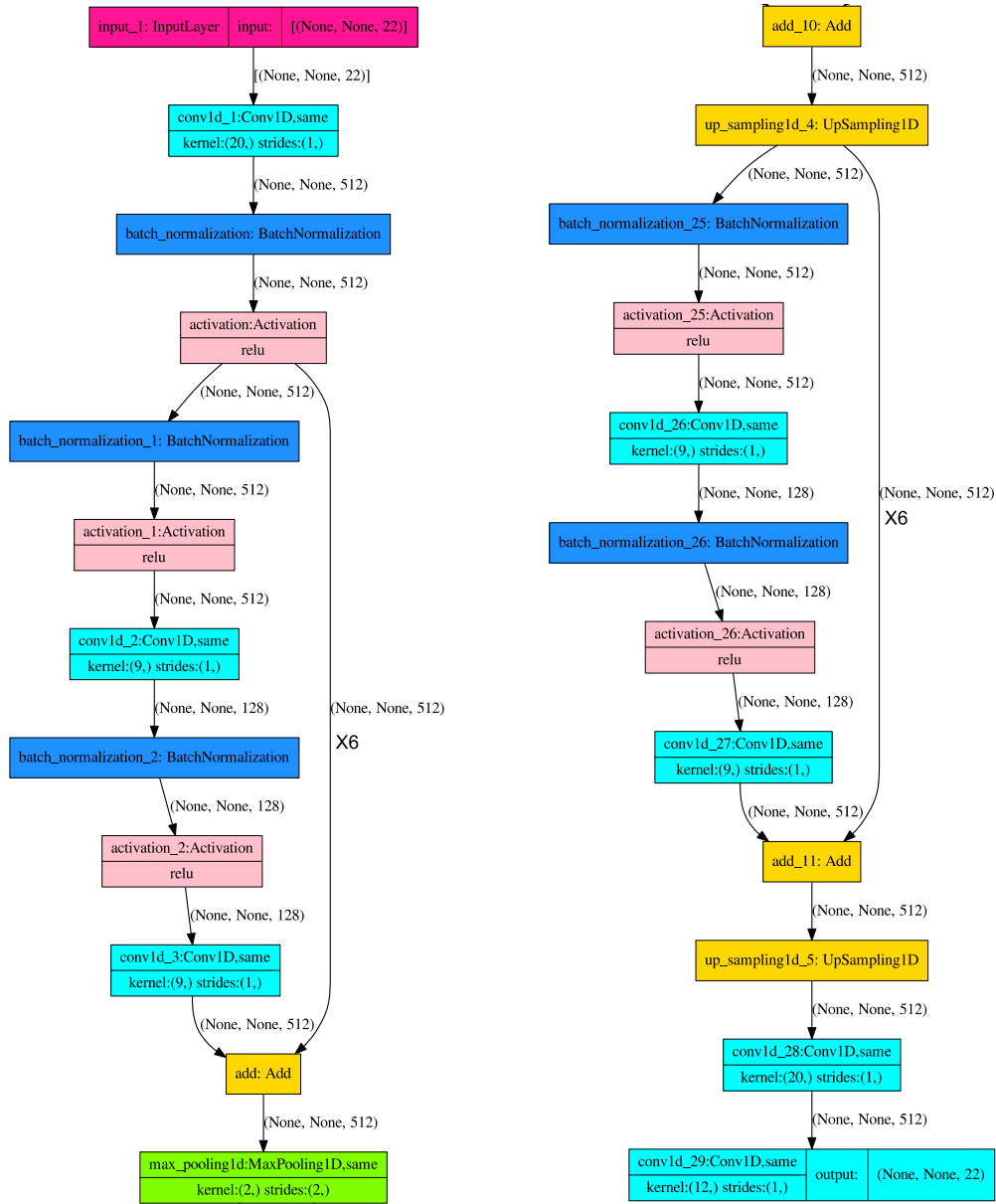
Supplementary Fig. 6 | Workflow of multitask multilabel deep learning model for integron (INTNet). From left to right, the first part is the input sequences, the second is the autoencoder model to identify integron sequences, the last part includes three CNNs to classify different aspects of integron.



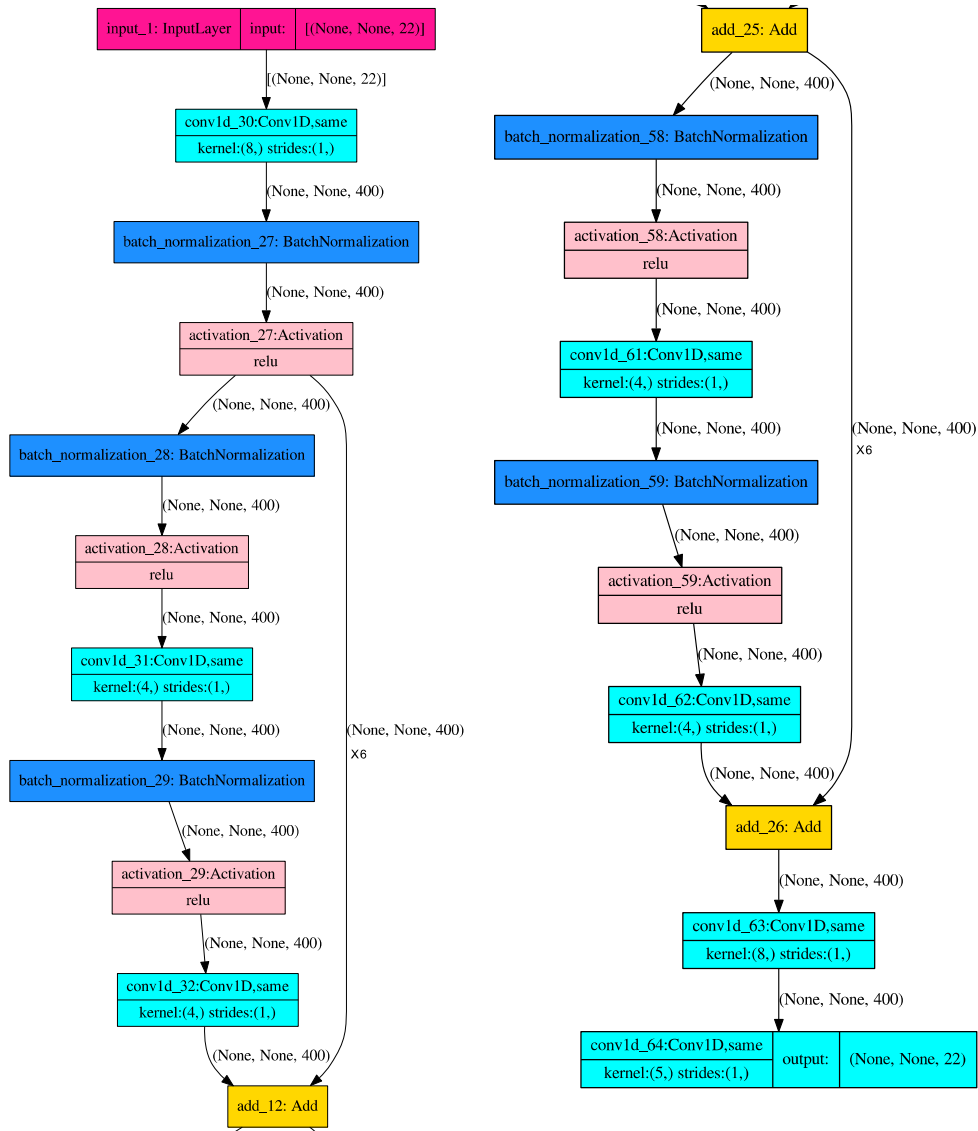
Supplementary Fig. 7 | The five-fold cross-validation of long-sequence and short-read CNN classifiers of integron group. The upper part is the results of long-sequence CNN, the lower part is short-read CNN. The left side is for amino acid sequences and the right side is for nucleotide sequences. The purple bars and error bars represent the mean values and standard deviation values of five-fold cross-validation for different testing dataset.



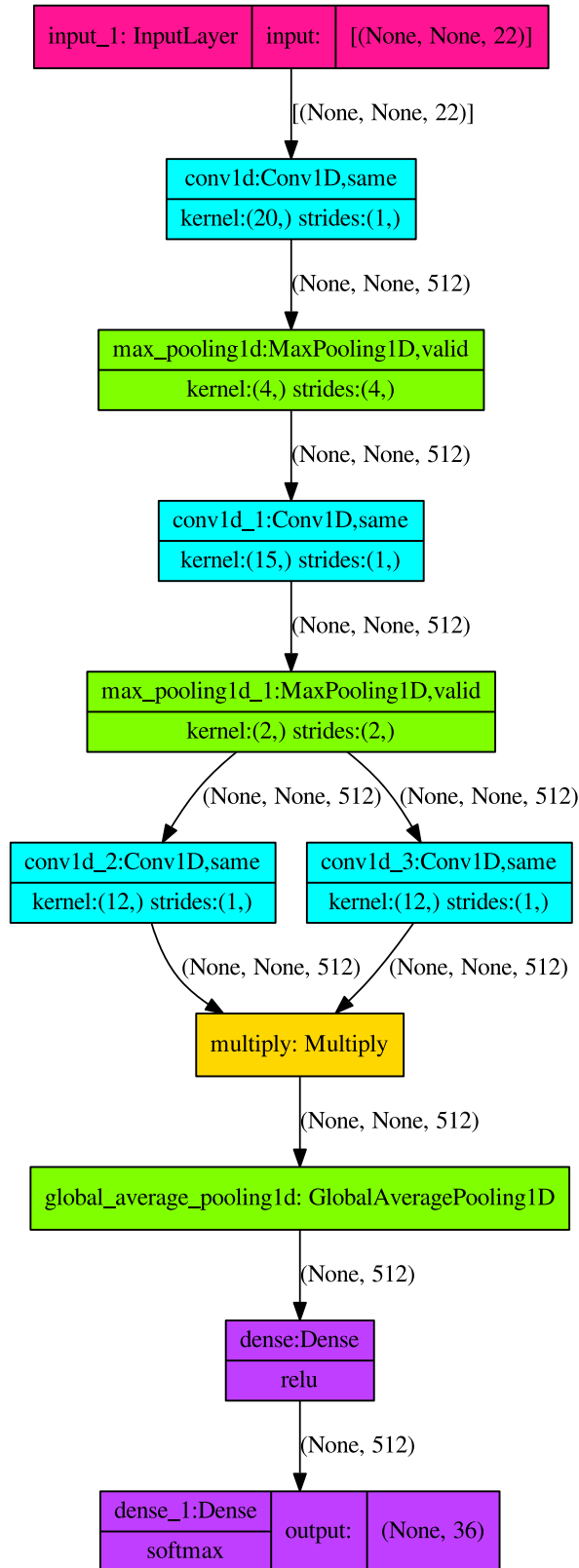
Supplementary Fig. 8 | Data cleaning workflow of integron associated ARGs. GenBank records were downloaded for 9,831 of 10,415 integron protein sequences (the remaining 584 do not have GenBank records). After de-duplication, 9,831 GenBank records were assigned into 793 nucleotide records and 7,932 WGS master records. Protein-coding sequences for DNA sequence of each record were predicted by Prodigal, which were used for ARG prediction by ARGNet.



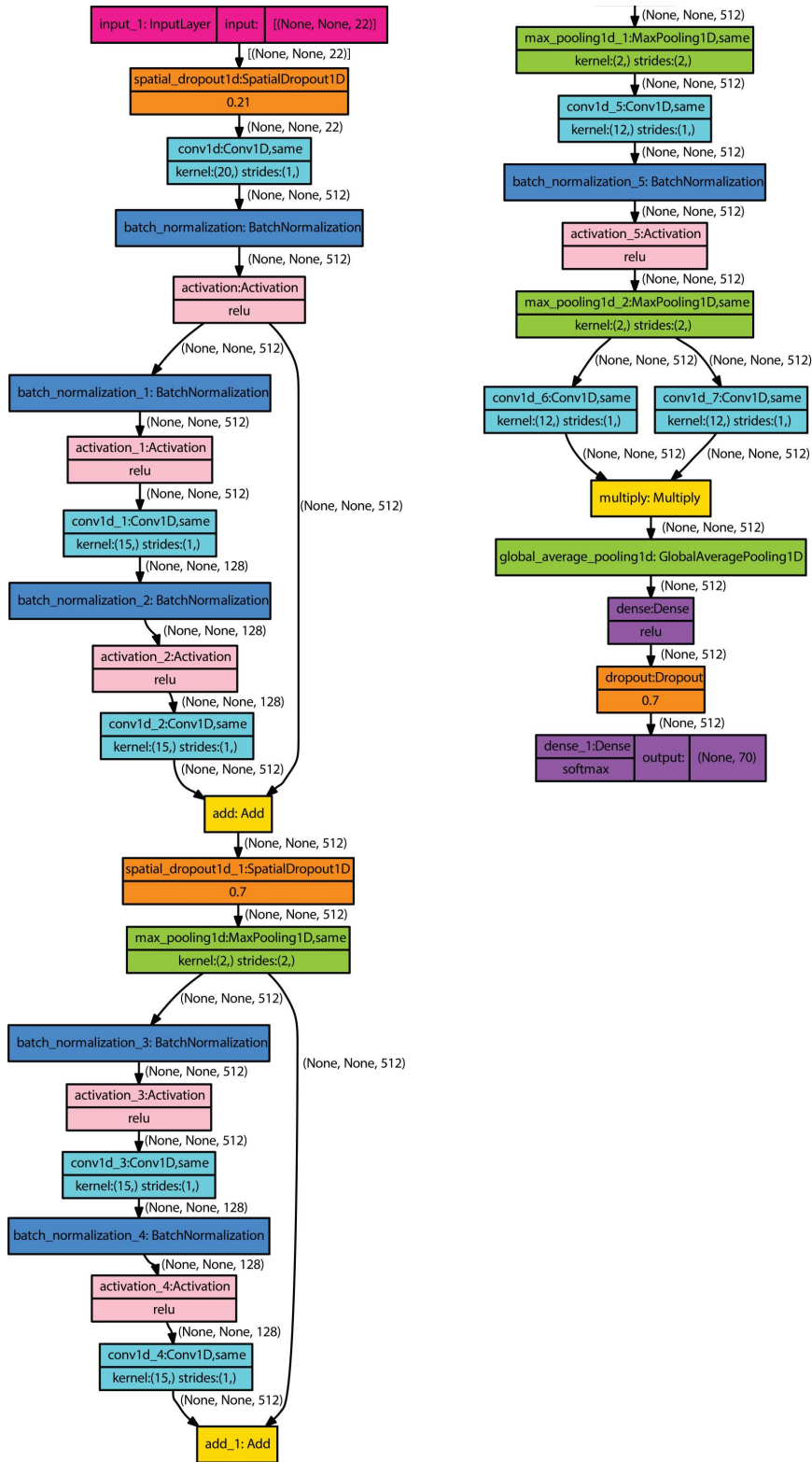
Supplementary Fig. 9 | Structure and parameters of autoencoder in INENet-L. The left part is encoder, and the right part is decoder. Different types of layers are represented in different colors. The shape information and parameters of each layer are shown in parentheses. “None” means any dimension in batch size or the length dimension.



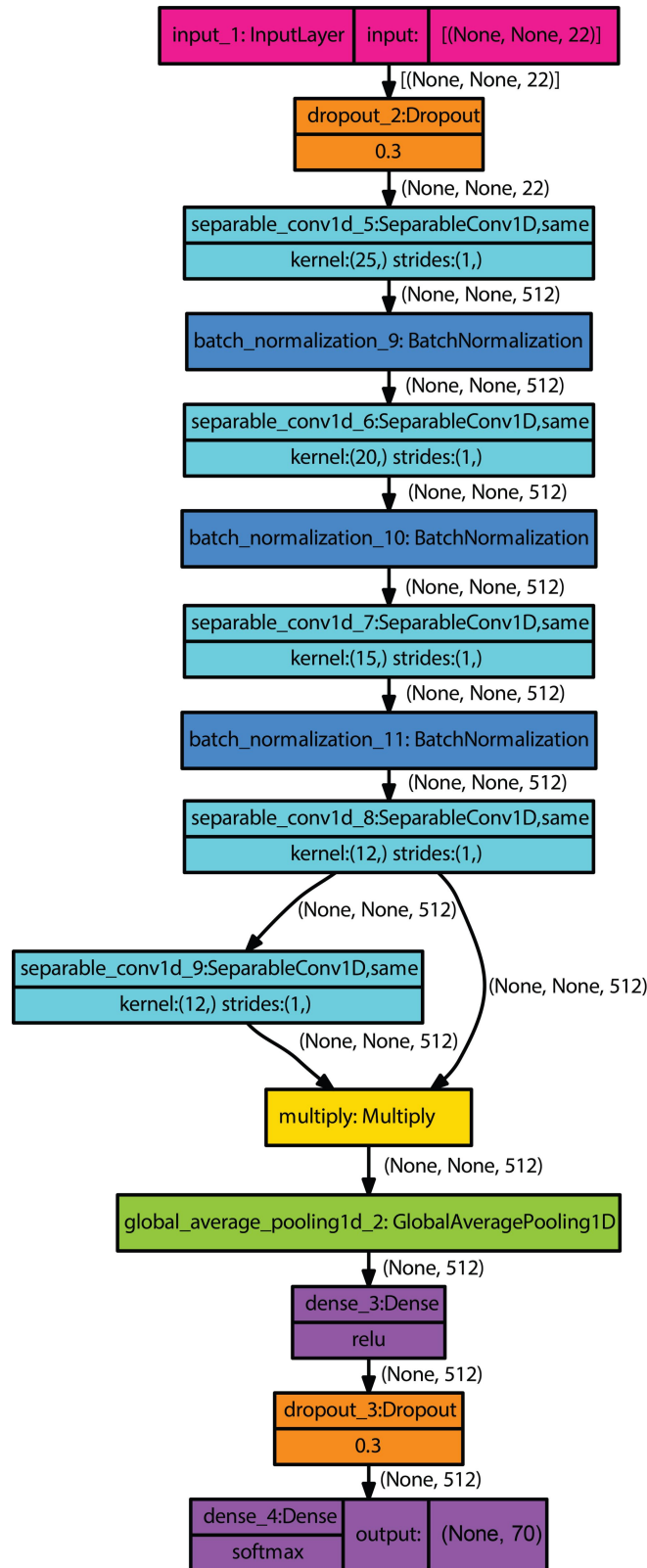
Supplementary Fig. 10 | Structure and parameters of autoencoder in INTNet-S. The left part is encoder, and the right part is decoder. Different types of layers are represented in different colors. The shape information and parameters of each layer are shown in parentheses. “None” means any dimension in batch size or the length dimension.



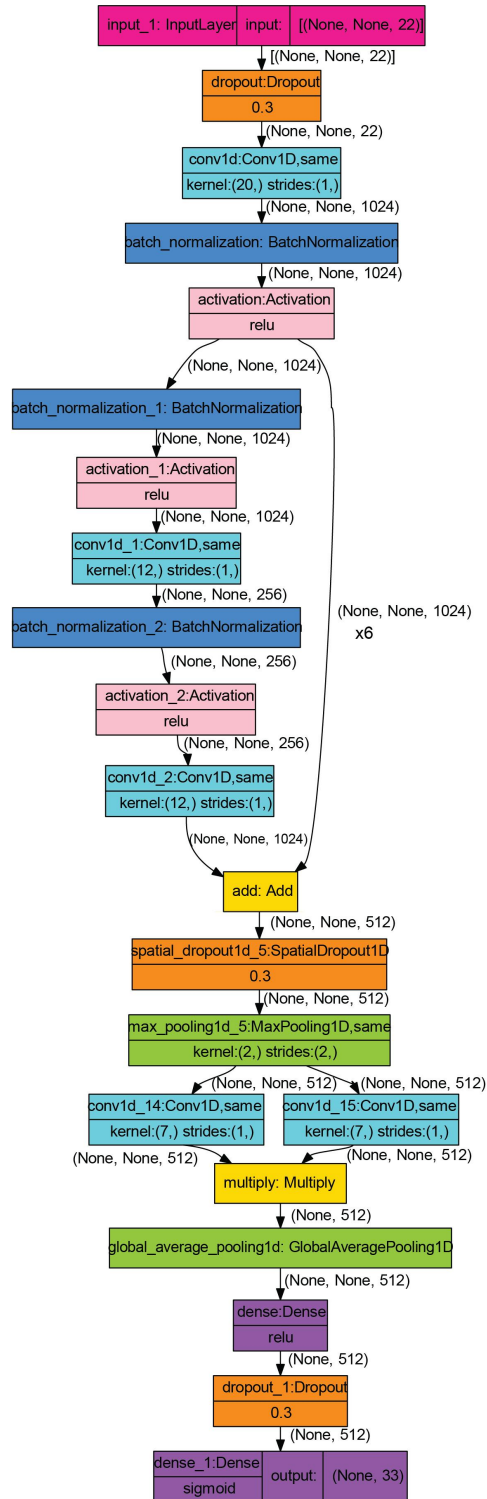
Supplementary Fig. 11 | Structure and parameters of CNN classifier of IntI in INTNet-L. Different types of layers are represented in different colors. The shape information and parameters of each layer are shown in parentheses. “None” means any dimension in batch size or the length dimension.



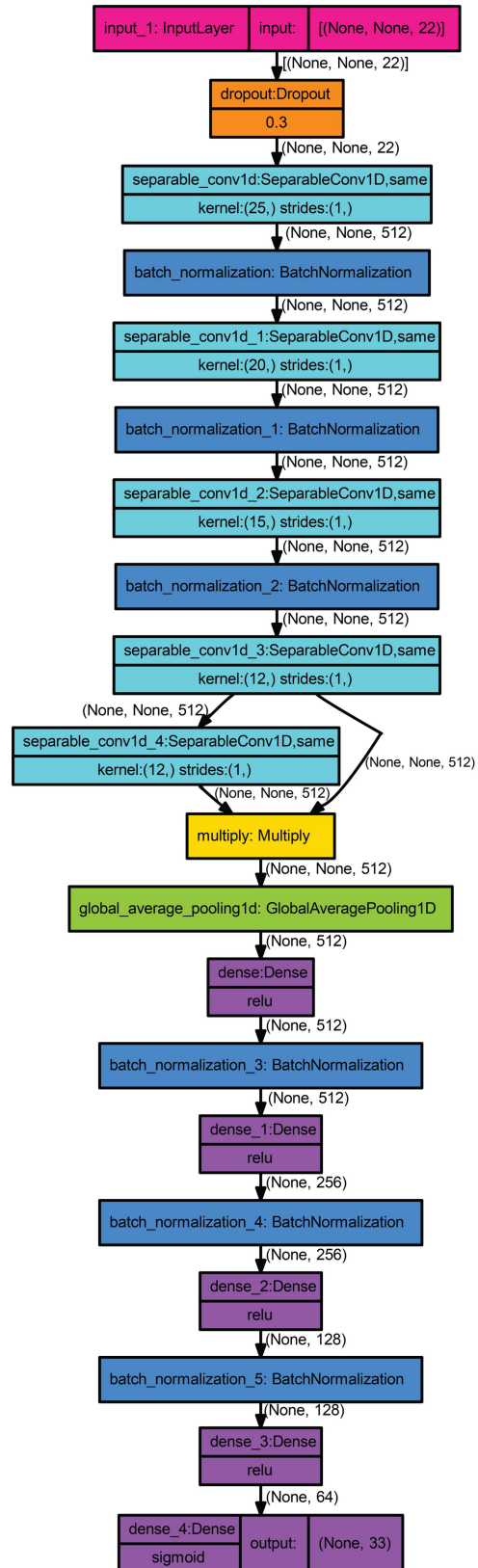
Supplementary Fig. 13 | Structure and parameters of CNN classifier of long-sequence model for identification of bacterial hosts of integrons. Different types of layers are represented in different colors. The shape information and parameters of each layer are shown in parentheses. “None” means any dimension in batch size or the length dimension.



Supplementary Fig. 14 | Structure and parameters of CNN classifier of short-sequence model for identification of bacterial hosts of integrons. Different types of layers are represented in different colors. The shape information and parameters of each layer are shown in parentheses. “None” means any dimension in batch size or the length dimension.



Supplementary Fig. 15 | Structure and parameters of multilabel CNN classifier of long-sequence model for associated ARGs of integrons. Different types of layers are represented in different colors. The shape information and parameters of each layer are shown in parentheses. “None” means any dimension in batch size or the length dimension.



Supplementary Fig. 16 | Structure and parameters of multilabel CNN classifier of short-sequence model for associated ARGs of integrons. Different types of layers are represented in different colors. The shape information and parameters of each layer are shown in parentheses. “None” means any dimension in batch size or the length dimension.

Supplementary Tables

Supplementary Table 1 | 33 categories of associated ARGs of integron predicted by ARGNet.

Aminocoumarin	Aminoglycoside	Bacitracin	Beta-lactam	Bleomycin	chloramphenicol
elfamycin	ethambutol	fosfomycin	fosmidomycin	fusidic_acid	glycopeptide
isoniazid	kasugamycin	MLS	Multidrug	Mupirocin	Nitrofurantoin
Nitroimidazole	Peptide	Pleuromutilin	Polymyxin	Pyrazinamide	Qa_compound
Quinolone	Rifamycin	Streptothricin	Sulfonamide	Tetracenomycin	Tetracycline
Triclosan	Trimethoprim	Tunicamycin			

Supplementary Table 2 | Long amino acid sequence testing results of multilabel CNN for associated ARGs of integron.

ARG	Full length			90%			80%			70%			60%			Support
	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score	
Aminoglycoside	0.99	1	1	0.99	1	1	0.99	1	1	0.99	0.98	0.99	0.99	0.97	0.98	1406
Beta-lactam	0.99	1	1	0.99	1	1	0.99	1	0.99	0.99	0.99	0.99	0.99	0.98	0.98	1404
Multidrug	0.99	1	1	0.99	1	1	0.99	1	0.99	0.99	1	0.99	0.99	1	0.99	1404
MLS	0.99	1	0.99	0.99	1	0.99	0.99	1	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1399
Peptide	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.98	0.99	0.96	0.98	1399
Quinolone	0.99	1	0.99	0.99	1	0.99	0.99	1	0.99	0.99	0.99	0.99	0.99	0.97	0.98	1398
Tetracycline	0.99	1	0.99	0.99	1	0.99	0.99	1	0.99	0.99	0.98	0.98	0.99	0.97	0.98	1398
Bacitracin	0.98	0.99	0.99	0.98	0.99	0.99	0.98	0.98	0.98	0.98	0.95	0.97	0.98	0.93	0.95	1393
Rifamycin	0.98	0.99	0.99	0.98	0.99	0.99	0.98	0.98	0.98	0.98	0.95	0.97	0.98	0.93	0.96	1393
Glycopeptide	0.96	0.98	0.97	0.96	0.98	0.97	0.96	0.97	0.97	0.96	0.93	0.95	0.95	0.89	0.92	1356
Aminocoumarin	0.88	0.88	0.88	0.87	0.86	0.87	0.87	0.83	0.85	0.86	0.78	0.82	0.85	0.77	0.81	1192
Elfamycin	0.88	0.89	0.88	0.86	0.92	0.89	0.85	0.94	0.89	0.85	0.93	0.89	0.84	0.91	0.87	1175
Sulfonamide	0.84	0.86	0.85	0.82	0.88	0.85	0.80	0.87	0.84	0.80	0.83	0.82	0.79	0.82	0.81	1094
Pleuromutilin	0.81	0.83	0.82	0.79	0.85	0.82	0.77	0.85	0.81	0.77	0.83	0.80	0.73	0.79	0.76	1019
Ethambutol	0.68	0.82	0.75	0.68	0.82	0.75	0.67	0.79	0.72	0.62	0.75	0.68	0.60	0.75	0.67	756
Chloramphenicol	0.57	0.63	0.60	0.53	0.61	0.57	0.54	0.63	0.58	0.50	0.54	0.52	0.51	0.58	0.54	649
Fosfomycin	0.57	0.64	0.61	0.56	0.67	0.61	0.55	0.67	0.60	0.53	0.61	0.57	0.51	0.59	0.54	593
Triclosan	0.60	0.67	0.63	0.56	0.63	0.59	0.53	0.65	0.58	0.55	0.63	0.58	0.50	0.58	0.54	516
Polymyxin	0.45	0.68	0.54	0.43	0.67	0.52	0.42	0.64	0.51	0.39	0.63	0.48	0.40	0.58	0.48	461
Trimethoprim	0.45	0.64	0.53	0.45	0.65	0.54	0.44	0.59	0.50	0.46	0.56	0.51	0.41	0.53	0.46	294
Mupirocin	0.23	0.51	0.32	0.22	0.54	0.31	0.19	0.41	0.26	0.19	0.41	0.26	0.18	0.38	0.24	268

Isoniazid	0.33	0.53	0.40	0.32	0.46	0.38	0.28	0.36	0.31	0.27	0.33	0.30	0.23	0.31	0.26	203
Nitrofurantoin	0.48	0.81	0.60	0.48	0.77	0.59	0.51	0.72	0.60	0.47	0.66	0.55	0.44	0.68	0.53	199
qa_compound	0.56	0.82	0.67	0.53	0.72	0.61	0.49	0.59	0.54	0.44	0.54	0.48	0.43	0.55	0.48	185
Kasugamycin	0.78	0.88	0.82	0.75	0.86	0.80	0.71	0.78	0.74	0.67	0.79	0.73	0.60	0.77	0.67	179
Nitroimidazole	0.84	0.87	0.86	0.85	0.91	0.88	0.85	0.85	0.85	0.79	0.82	0.81	0.71	0.75	0.73	175
Fosmidomycin	0.49	0.85	0.62	0.46	0.86	0.60	0.47	0.71	0.57	0.43	0.67	0.53	0.44	0.71	0.55	105
Streptothricin	0.13	0.77	0.22	0.10	0.73	0.18	0.09	0.54	0.15	0.13	0.81	0.22	0.09	0.62	0.15	26
Fusidic_acid	0.02	0.38	0.03	0.01	0.23	0.02	0.02	0.38	0.03	0.01	0.31	0.02	0.02	0.46	0.03	13
Bleomycin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Pyrazinamide	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Micro average	0.82	0.91	0.86	0.81	0.91	0.86	0.80	0.90	0.85	0.79	0.88	0.83	0.78	0.86	0.82	23055
Weighted average	0.87	0.91	0.89	0.87	0.91	0.89	0.86	0.90	0.88	0.85	0.88	0.86	0.85	0.86	0.85	23055
Samples average	0.83	0.91	0.86	0.82	0.91	0.85	0.81	0.90	0.84	0.81	0.88	0.83	0.79	0.86	0.81	23055

The column named “support” represented the number of each label in the testing dataset.

Supplementary Table 3 | Long nucleotide sequence testing results of multilabel CNN for associated ARGs of integron.

ARG	Full length			90%			80%			70%			60%			Support
	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score	
Aminoglycoside	0.99	1	1	0.99	1	1	0.99	0.97	0.98	0.99	0.98	0.99	0.99	0.97	0.98	1406
Beta-lactam	0.99	1	1	0.99	1	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1404
Multidrug	0.99	1	1	0.99	1	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1404
MLS	0.99	1	0.99	0.99	1	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1399
Peptide	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.96	0.97	0.99	0.98	0.98	0.99	0.96	0.97	1399
Quinolone	0.99	1	0.99	0.99	1	0.99	0.99	0.98	0.98	0.99	0.98	0.98	0.99	0.98	0.98	1398
Tetracycline	0.99	1	0.99	0.99	1	0.99	0.99	0.98	0.98	0.99	0.98	0.98	0.99	0.98	0.98	1398
Bacitracin	0.98	0.99	0.99	0.98	0.98	0.98	0.98	0.93	0.95	0.98	0.95	0.96	0.98	0.93	0.95	1393
Rifamycin	0.98	0.99	0.99	0.98	0.99	0.99	0.98	0.94	0.96	0.98	0.96	0.97	0.98	0.94	0.96	1393
Glycopeptide	0.96	0.98	0.97	0.96	0.98	0.97	0.96	0.89	0.92	0.95	0.94	0.94	0.96	0.89	0.92	1356
Aminocoumarin	0.88	0.87	0.88	0.87	0.88	0.87	0.85	0.77	0.81	0.86	0.78	0.82	0.85	0.77	0.81	1192
Elfamycin	0.88	0.89	0.89	0.85	0.93	0.89	0.84	0.90	0.87	0.85	0.92	0.88	0.84	0.90	0.87	1175
Sulfonamide	0.84	0.86	0.85	0.81	0.88	0.85	0.80	0.82	0.81	0.81	0.84	0.82	0.80	0.82	0.81	1094
Pleuromutilin	0.81	0.83	0.82	0.79	0.86	0.83	0.74	0.80	0.77	0.76	0.82	0.79	0.74	0.80	0.77	1019
Ethambutol	0.69	0.82	0.75	0.68	0.81	0.74	0.59	0.77	0.67	0.62	0.79	0.69	0.59	0.77	0.67	756
Chloramphenicol	0.58	0.63	0.60	0.53	0.61	0.57	0.51	0.57	0.54	0.52	0.56	0.54	0.51	0.57	0.54	649
Fosfomycin	0.57	0.64	0.61	0.56	0.69	0.62	0.52	0.63	0.57	0.52	0.61	0.56	0.52	0.63	0.57	593
Triclosan	0.60	0.68	0.64	0.57	0.65	0.61	0.51	0.60	0.55	0.54	0.62	0.58	0.51	0.60	0.55	516
Polymyxin	0.45	0.68	0.54	0.44	0.69	0.54	0.41	0.60	0.49	0.41	0.65	0.50	0.41	0.60	0.49	461
Trimethoprim	0.45	0.64	0.53	0.44	0.63	0.52	0.45	0.52	0.48	0.45	0.56	0.50	0.45	0.52	0.48	294
Mupirocin	0.22	0.51	0.31	0.21	0.49	0.30	0.18	0.39	0.24	0.19	0.41	0.26	0.18	0.39	0.24	268
Isoniazid	0.33	0.54	0.41	0.33	0.47	0.38	0.24	0.30	0.27	0.26	0.30	0.28	0.24	0.30	0.27	203
Nitrofurantoin	0.48	0.81	0.60	0.50	0.79	0.61	0.42	0.68	0.52	0.46	0.70	0.56	0.42	0.68	0.52	199
qa_compound	0.56	0.81	0.67	0.53	0.76	0.62	0.45	0.56	0.50	0.47	0.63	0.54	0.45	0.56	0.50	185
Kasugamycin	0.77	0.88	0.82	0.77	0.88	0.82	0.67	0.84	0.75	0.66	0.82	0.73	0.67	0.84	0.75	179

Nitroimidazole	0.84	0.87	0.86	0.82	0.89	0.85	0.77	0.82	0.80	0.79	0.83	0.81	0.77	0.82	0.80	175
Fosmidomycin	0.49	0.87	0.63	0.46	0.80	0.59	0.43	0.67	0.52	0.47	0.72	0.57	0.43	0.67	0.52	105
Streptothricin	0.13	0.77	0.22	0.13	0.85	0.22	0.09	0.58	0.16	0.10	0.65	0.18	0.09	0.58	0.16	26
Fusidic_acid	0.02	0.38	0.03	0.02	0.46	0.04	0.02	0.46	0.04	0.02	0.38	0.03	0.02	0.46	0.04	13
Bleomycin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Pyrazinamide	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Micro average	0.82	0.91	0.86	0.81	0.91	0.86	0.78	0.87	0.82	0.79	0.88	0.83	0.78	0.87	0.82	23055
Weighted average	0.87	0.91	0.89	0.87	0.91	0.89	0.85	0.87	0.86	0.85	0.88	0.86	0.85	0.87	0.86	23055
Samples average	0.83	0.91	0.86	0.82	0.91	0.85	0.79	0.86	0.81	0.80	0.88	0.83	0.79	0.86	0.81	23055

The column named “support” represented the number of each label in the testing dataset.

Supplementary Table 4 | Short amino acid sequence testing results of multilabel CNN for associated ARGs of integron.

ARG	30 aa			40 aa			50 aa			Support
	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score	
Aminoglycoside	0.99	1	1	0.99	1	1	0.99	1	1	7030
Beta-lactam	0.99	1	1	0.99	1	0.99	0.99	1	0.99	7020
Multidrug	0.99	1	1	0.99	1	1	0.99	1	1	7020
MLS	0.99	1	0.99	0.99	1	0.99	0.99	1	0.99	6995
Peptide	0.99	1	0.99	0.99	1	0.99	0.99	1	0.99	6995
Quinolone	0.99	1	0.99	0.99	1	0.99	0.99	1	0.99	6990
Tetracycline	0.99	1	0.99	0.99	1	0.99	0.99	1	0.99	6990
Bacitracin	0.98	1	0.99	0.98	1	0.99	0.98	1	0.99	6965
Rifamycin	0.98	1	0.99	0.98	1	0.99	0.98	1	0.99	6965
Glycopeptide	0.96	1	0.98	0.96	1	0.98	0.96	1	0.98	6780
Aminocoumarin	0.88	0.95	0.91	0.88	0.94	0.91	0.88	0.94	0.91	5960
Elfamycin	0.87	0.94	0.90	0.87	0.95	0.91	0.87	0.95	0.91	5875
Sulfonamide	0.84	0.91	0.87	0.84	0.90	0.87	0.84	0.90	0.87	5470
Pleuromutilin	0.82	0.86	0.84	0.84	0.88	0.85	0.84	0.88	0.85	5095
Ethambutol	0.69	0.75	0.72	0.71	0.76	0.73	0.71	0.76	0.73	3780
Chloramphenicol	0.69	0.49	0.58	0.71	0.49	0.58	0.71	0.49	0.58	3245
Fosfomycin	0.73	0.50	0.60	0.74	0.51	0.60	0.74	0.51	0.60	2965
Triclosan	0.68	0.55	0.61	0.70	0.54	0.61	0.70	0.54	0.61	2580
Polymyxin	0.80	0.50	0.62	0.82	0.51	0.63	0.82	0.51	0.63	2305
Trimethoprim	0.73	0.58	0.64	0.75	0.58	0.65	0.75	0.58	0.65	1470
Mupirocin	0.51	0.09	0.16	0.51	0.11	0.18	0.51	0.11	0.18	1340
Isoniazid	0.51	0.33	0.40	0.54	0.34	0.42	0.54	0.34	0.42	1015
Nitrofurantoin	0.75	0.80	0.78	0.77	0.81	0.78	0.77	0.81	0.78	995
qa_compound	0.70	0.87	0.77	0.71	0.89	0.79	0.71	0.89	0.79	925
Kasugamycin	0.84	1	0.91	0.85	1	0.92	0.85	1	0.92	895

Nitroimidazole	0.84	1	0.91	0.85	1	0.92	0.85	1	0.92	875
Fosmidomycin	0	0	0	0	0	0	0	0	0	525
Streptothricin	0	0	0	0	0	0	0	0	0	130
Fusidic_acid	1	0.03	0.06	1	0.03	0.06	1	0.03	0.06	65
Bleomycin	0	0	0	0	0	0	0	0	0	10
Pyrazinamide	0	0	0	0	0	0	0	0	0	5
Micro average	0.91	0.90	0.90	0.92	0.90	0.91	0.92	0.90	0.91	115275
Weighted average	0.90	0.90	0.89	0.90	0.90	0.90	0.90	0.90	0.90	115275
Samples average	0.92	0.90	0.90	0.92	0.90	0.90	0.92	0.90	0.90	115275

The column named “support” represented the number of each label in the testing dataset.

Supplementary Table 5 | Short nucleotide sequence testing results of multilabel CNN for associated ARGs of integron.

ARG	100 nt			120 nt			150 nt			Support
	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score	
Aminoglycoside	0.99	1	1	0.99	1	1	0.99	1	1	6730
Beta-lactam	0.99	1	1	0.99	1	1	0.99	1	1	6718
Multidrug	0.99	1	1	0.99	1	1	0.99	1	1	6716
MLS	0.99	1	0.99	0.99	1	0.99	0.99	1	0.99	6694
Peptide	0.99	1	0.99	0.99	1	0.99	0.99	1	0.99	6694
Quinolone	0.99	1	0.99	0.99	1	0.99	0.99	1	0.99	6690
Tetracycline	0.99	1	0.99	0.99	1	0.99	0.99	1	0.99	6690
Bacitracin	0.98	1	0.99	0.98	1	0.99	0.98	1	0.99	6667
Rifamycin	0.98	1	0.99	0.98	1	0.99	0.98	1	0.99	6666
Glycopeptide	0.96	1	0.98	0.96	1	0.98	0.96	1	0.98	6492
Aminocoumarin	0.84	1	0.91	0.84	1	0.91	0.84	1	0.91	5685
Elfamycin	0.84	1	0.91	0.84	1	0.91	0.84	1	0.91	5674
Sulfonamide	0.77	1	0.87	0.77	1	0.87	0.77	1	0.87	5256
Pleuromutilin	0.73	1	0.84	0.73	1	0.84	0.73	1	0.84	4920
Ethambutol	0.89	0.03	0.06	0.89	0.03	0.06	0.89	0.03	0.06	3620
Chloramphenicol	0.48	0.97	0.64	0.48	0.97	0.64	0.48	0.97	0.64	3120
Fosfomycin	0.57	0.03	0.05	0.57	0.03	0.05	0.57	0.03	0.05	2850
Triclosan	1	0.02	0.04	1	0.02	0.04	1	0.02	0.04	2489
Polymyxin	0.92	0.04	0.08	0.92	0.04	0.08	0.92	0.04	0.08	2211
Trimethoprim	0.79	0.01	0.02	0.79	0.01	0.02	0.79	0.01	0.02	1426
Mupirocin	0.77	0.02	0.04	0.77	0.02	0.04	0.77	0.02	0.04	1261
Isoniazid	1	0.03	0.06	1	0.03	0.06	1	0.03	0.06	984
Nitrofurantoin	1	0.01	0.03	1	0.01	0.03	1	0.01	0.03	955
qa_compound	1	0	0.01	1	0	0.01	1	0	0.01	916
Kasugamycin	1	0.06	0.11	1	0.06	0.11	1	0.06	0.11	893

Nitroimidazole	1	0.01	0.02	1	0.01	0.02	1	0.01	0.02	873
Fosmidomycin	0	0	0	0	0	0	0	0	0	523
Streptothricin	0	0	0	0	0	0	0	0	0	130
Fusidic_acid	0	0	0	0	0	0	0	0	0	63
Bleomycin	0	0	0	0	0	0	0	0	0	8
Pyrazinamide	0	0	0	0	0	0	0	0	0	3
Micro average	0.90	0.83	0.86	0.90	0.83	0.86	0.90	0.83	0.86	110617
Weighted average	0.91	0.83	0.80	0.91	0.83	0.80	0.91	0.83	0.80	110617
Samples average	0.90	0.85	0.86	0.90	0.85	0.86	0.90	0.85	0.86	110617

The column named “support” represented the number of each label in the testing dataset.