

Подходы машинного обучения к
определению новых бактериальных
патогенов по данным NGS.

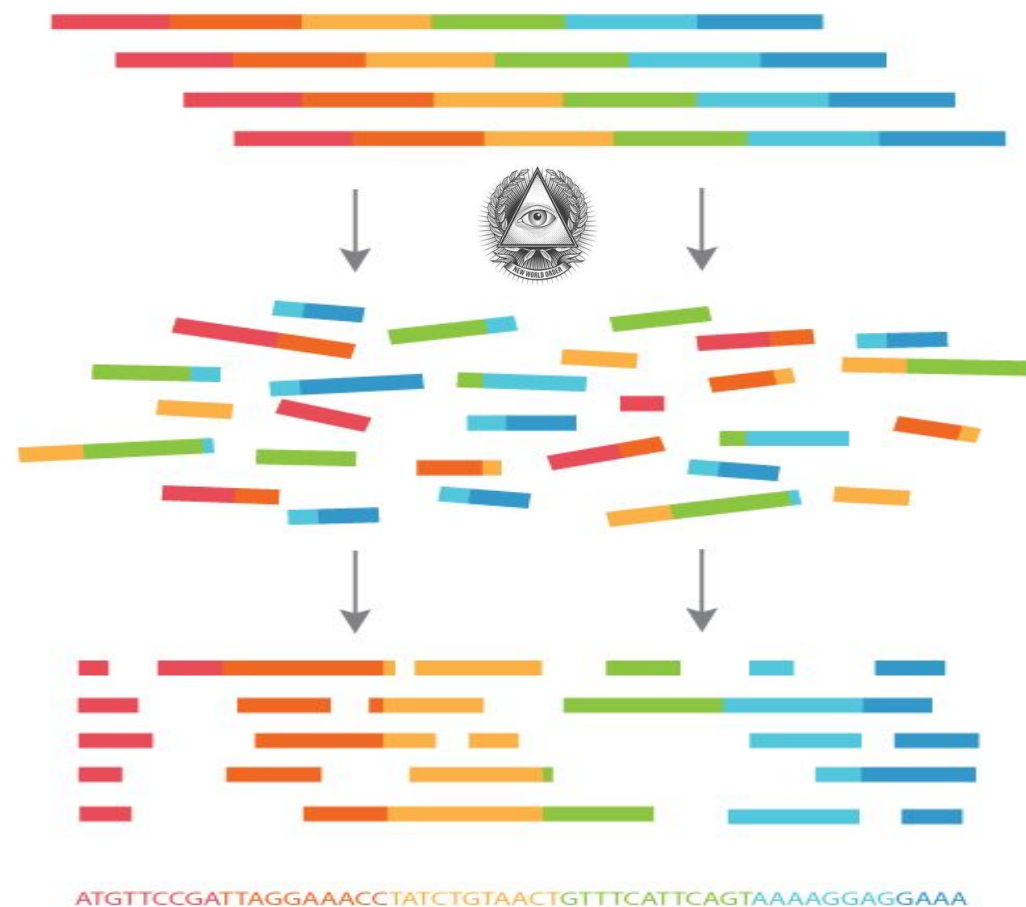
Предсказание патогенности бактерии
по данным NGS

Руководитель –
Константин Зайцев
Великий (Университет

Цель работы



Использование подходов машинного обучения для создания предиктора патогенности бактерии по данным NGS-секвенирования, путем обучения на большом количестве штаммов с известной патогенностью и извлечением правильных признаков из данных NGS-секвенирования



Данные

- Для обучения использовались данные из таких баз как например

JGI **IMG/M**
INTEGRATED MICROBIAL GENOMES & MICROBIOME SAMPLES

Quick Genome Search:

My Analysis Carts**: 0 [Genomes](#) | 0 [Scaffolds](#) | 0 [Functions](#) | 0 [Genes](#)

Home Find Genomes Find Genes Find Functions Compare Genomes OMICS My IMG Data Marts Help

IMG Content

Datasets	JGI	All
Bacteria	11752	53718
Archaea	471	1280
Eukarya	77	267
Plasmids	1	1193
Viruses	0	7924
Genome Fragments	1	1193
Metagenome	4659	6045
Cell Enrichments	611	611
Single Particle Sorts	1344	1344
Metatranscriptome	1723	1746
Total Datasets	75321	
Last Datasets Added On:		
Genome	2017-07-17	
Metagenome	2017-07-10	
Project Map Metagenome Projects Map System Requirements Microbial Genomics & Metagenomics Workshop		

The **Integrated Microbial Genomes (IMG)** system serves as a community resource for analysis and annotation of genome and metagenome datasets in a comprehensive comparative context. The **IMG data warehouse** integrates genome and metagenome datasets provided by IMG users with a comprehensive set of publicly available genome and metagenome datasets.

IMG provides users with tools ([IMG UI Map](#)) for analyzing publicly available genome datasets and metagenome datasets ([Nucleic Acids Research, October 13, 2016](#)).

[IMG Statistics](#) [Data Usage Policy](#)

Sequenced at:	Isolates		SAGs		MAGs	
	JGI	All	JGI	All	JGI	All
Bacteria	6058	47154	1813	2186	3854	4347
Archaea	206	772	186	282	79	226
Eukarya	76	266	0	0	1	1
Viruses	0	7854	0	44	0	0

(Only data sets with GOLD metadata were counted.)

Combined assembly data sets were excluded in the following metagenome and metatranscriptome table statistics.

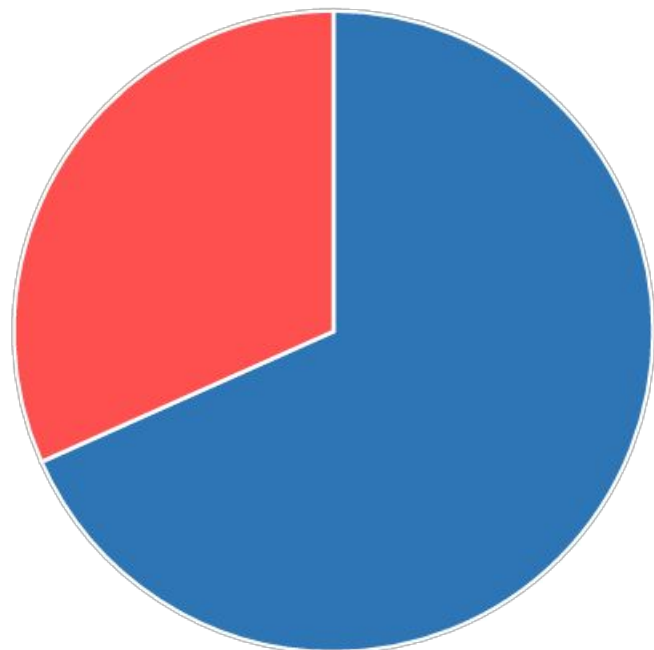
Metagenome Metatranscriptome

Engineered	JGI	ALL	Environmental	JGI	ALL	Host-associated	JGI	ALL
Dispersed	20	55	...	21	21	...	0	00

Acid Mine Drainage, Whalefall

Классификация бактерий из выборки

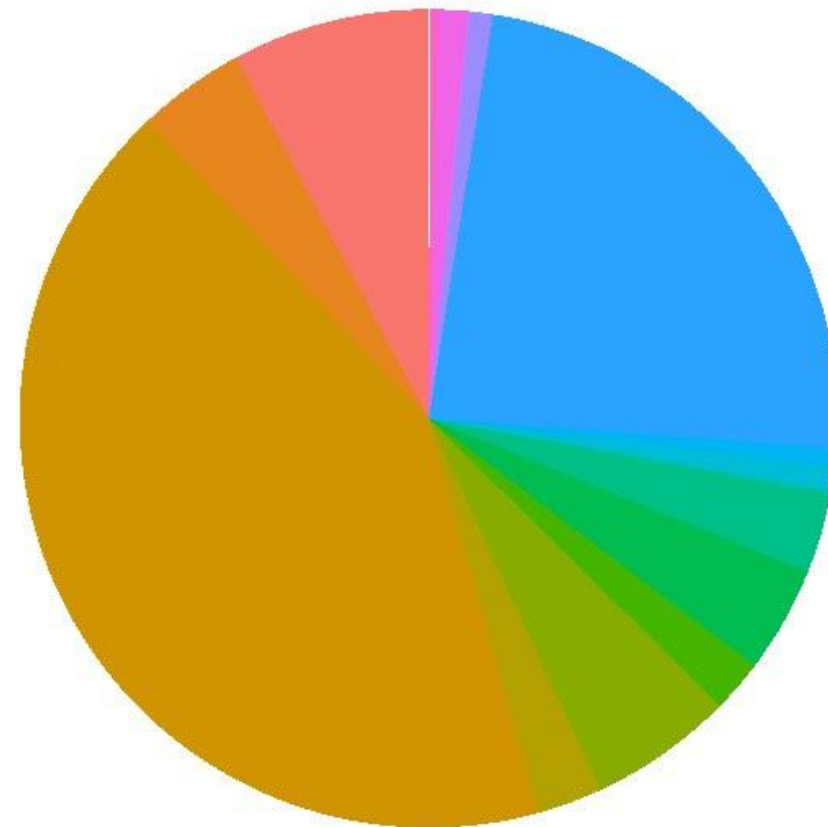
NP 418
P 902
1320



Доля патогенов
в выборке

ФЕНОТИП

Непатоген
Патоген



Классы бактерий
в полной выборке

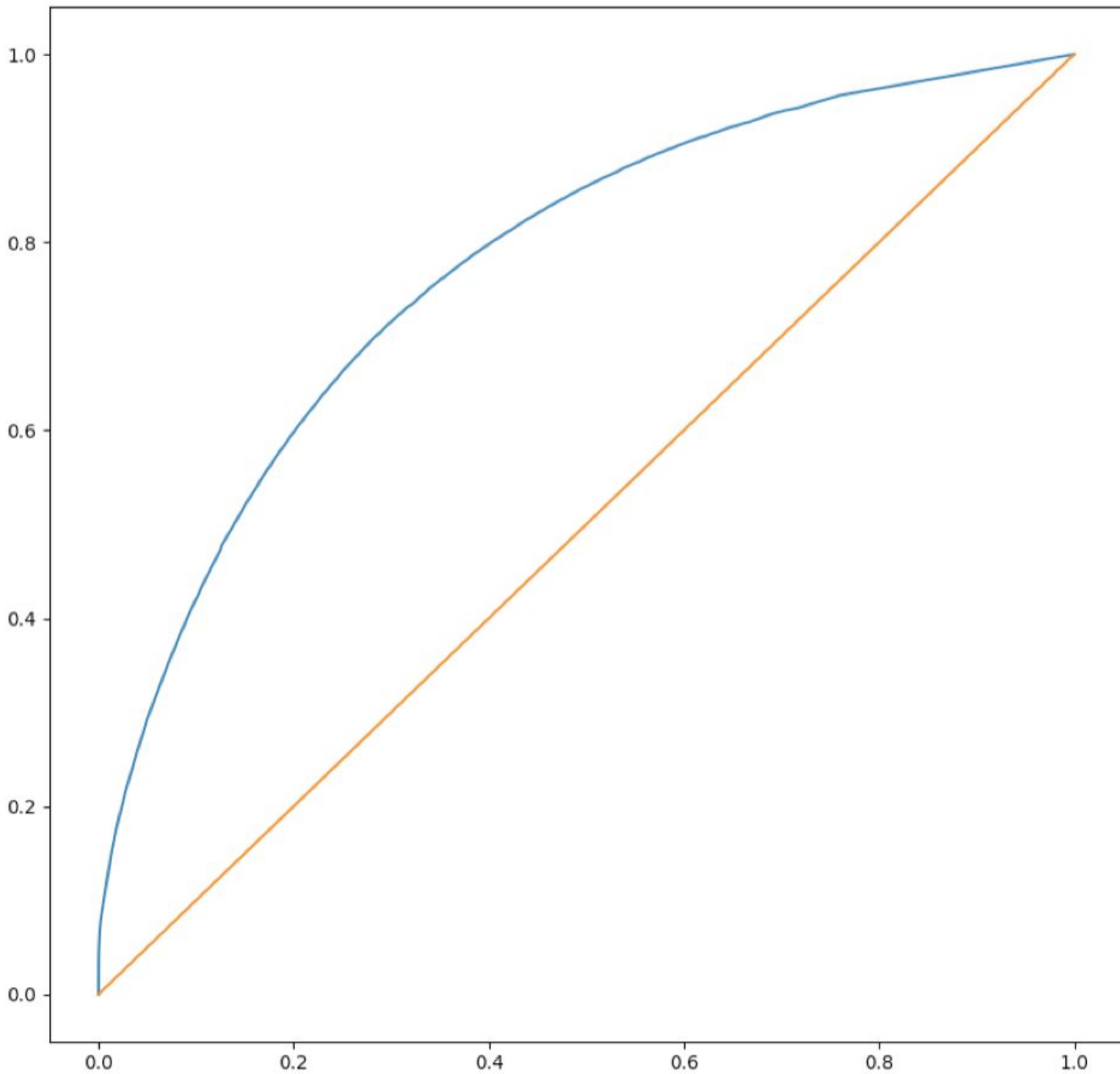
Класс

Actinobacteria
Alphaproteobacteria
Bacilli
Bacteroidia
Betaproteobacteria
Chlamydia
Clostridia
Epsilonproteobacteria
Erysipelotrichia
Flavobacteriia
Fusobacteriia
Gammaproteobacteria
Mollicutes
Negativicutes
Spirochaetia
Tissierellia
unclassified

Модели

Данные анализировали при помощи нескольких подходов –

- 1) Обучение сверточной нейронной сети для классификации отдельных ридов
- 2) Логистическая регрессия для классификации полных бактериальных геномов
- 3) Random Forest для классификации как отдельных ридов, так и полных бактериальных геномов



Результаты

AUC-ROC = 0.78

Логистическая регрессия для бактерий

