



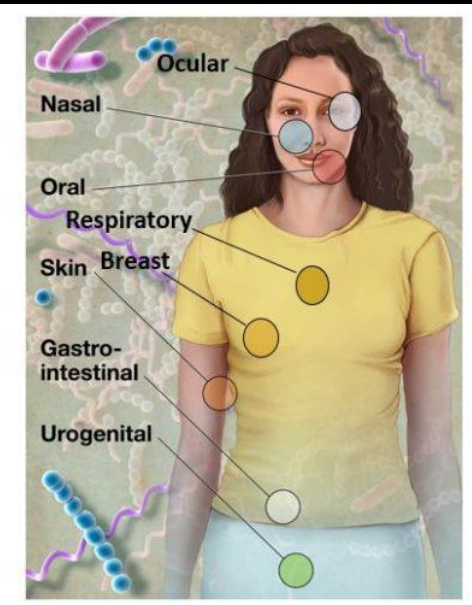
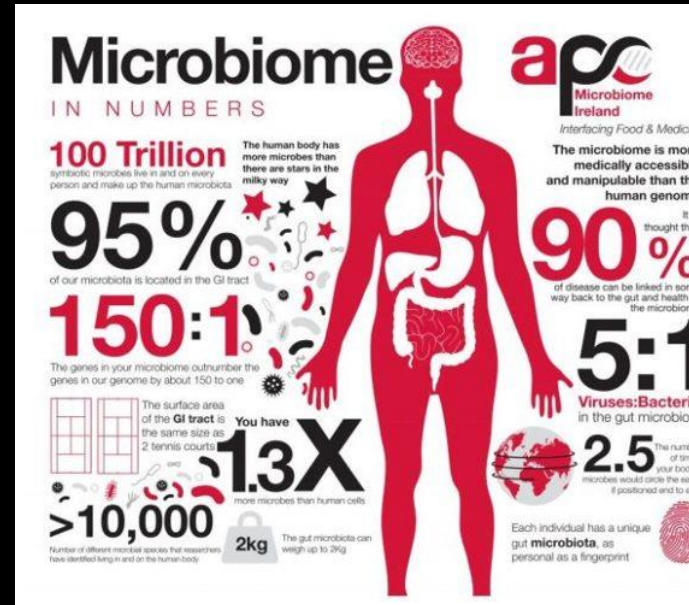
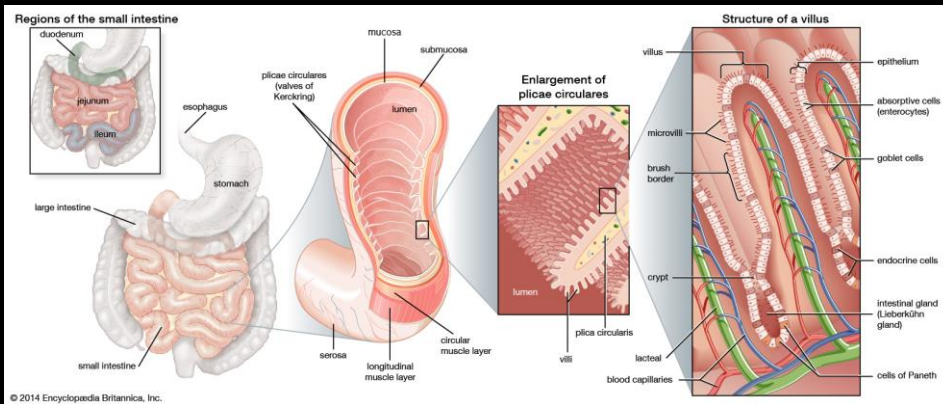
Gut Microbiome-Balance in Human Health and Disease

电子科大邹权课题组，杨凤龙

What is microbiome?

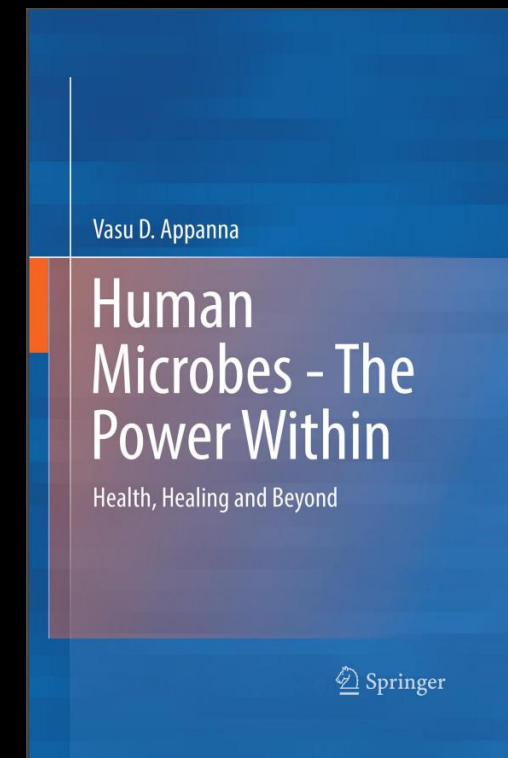
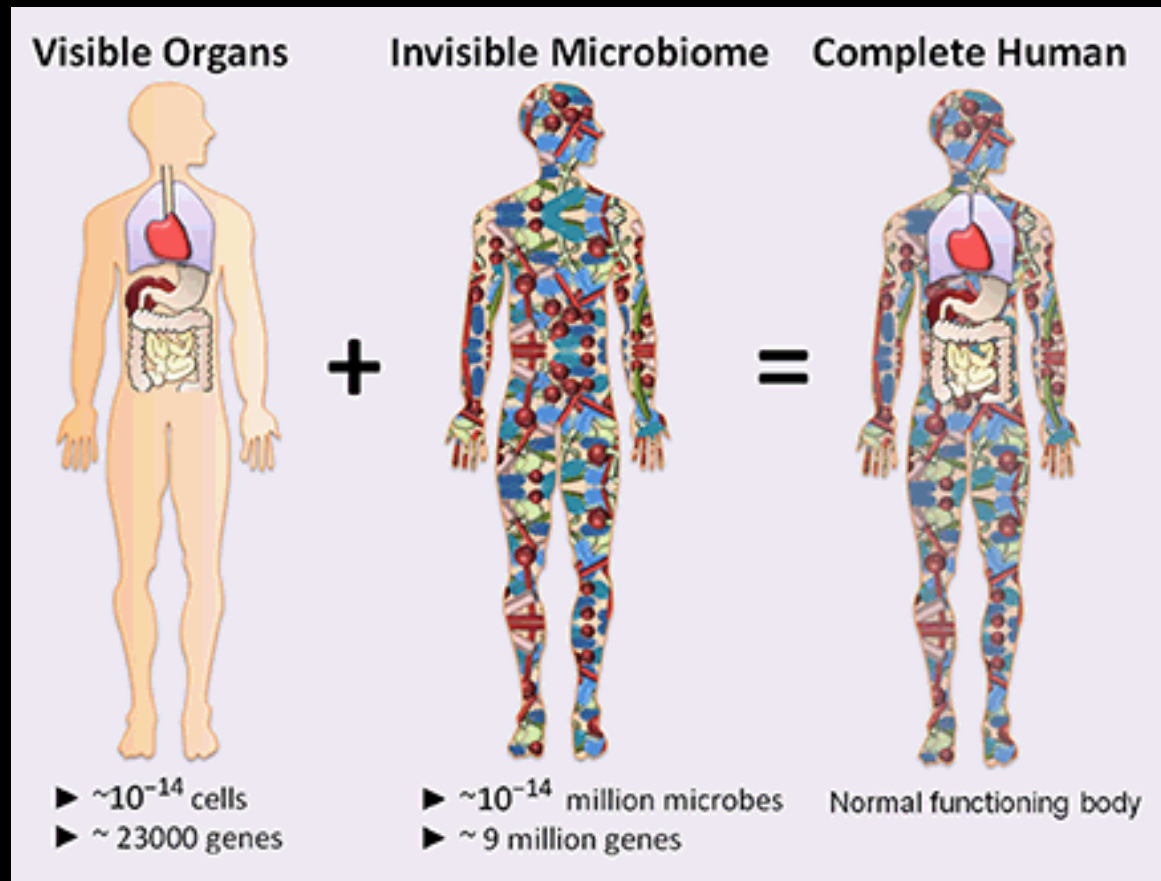
2

- The collection of all genomes of microbes in an ecosystem.



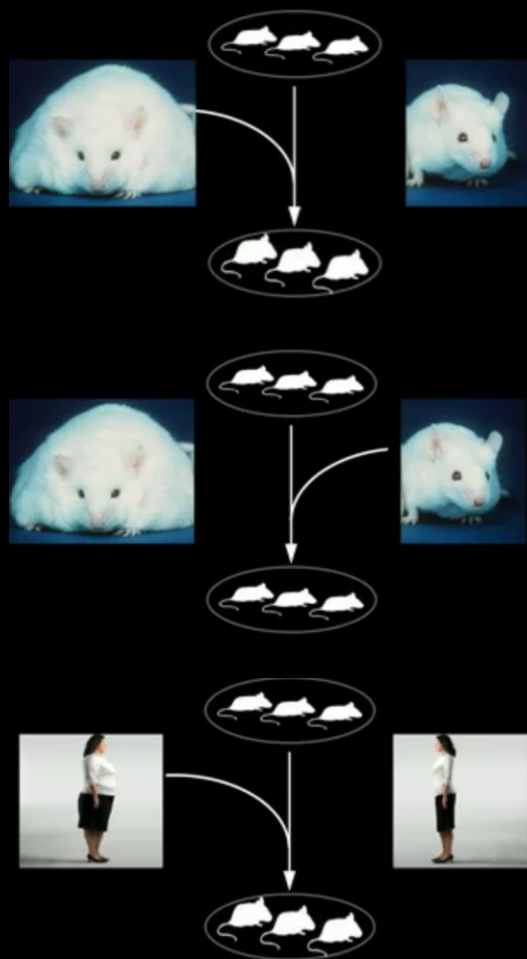
Who am I in a systematic view ?

3

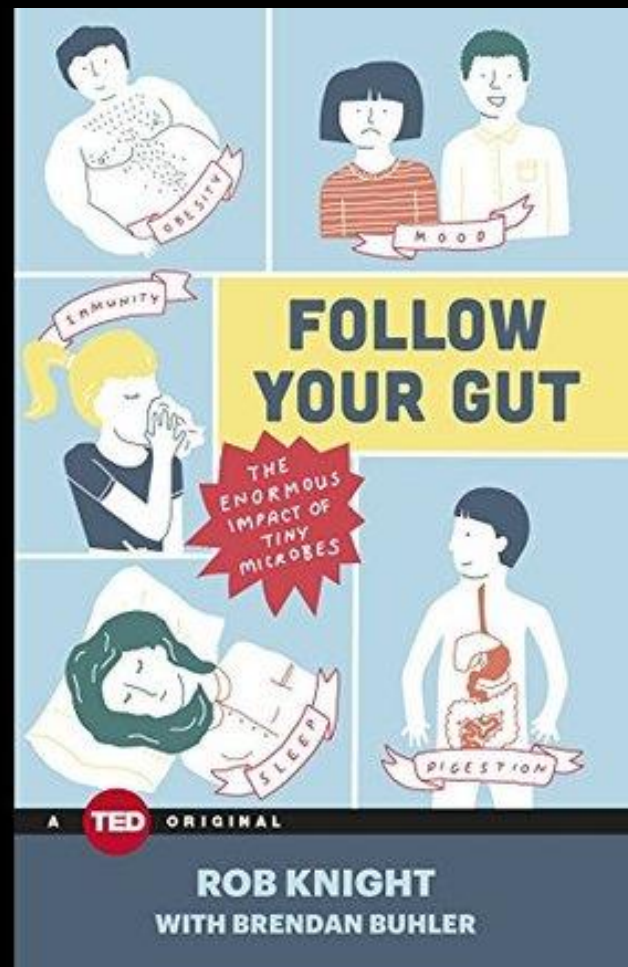


How our microbes make us who we are

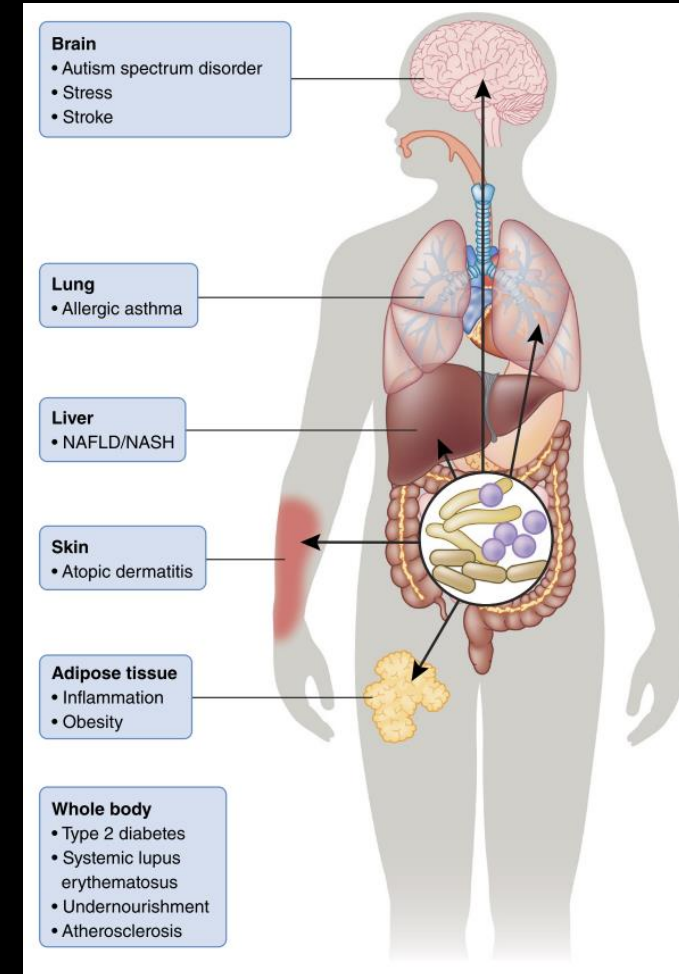
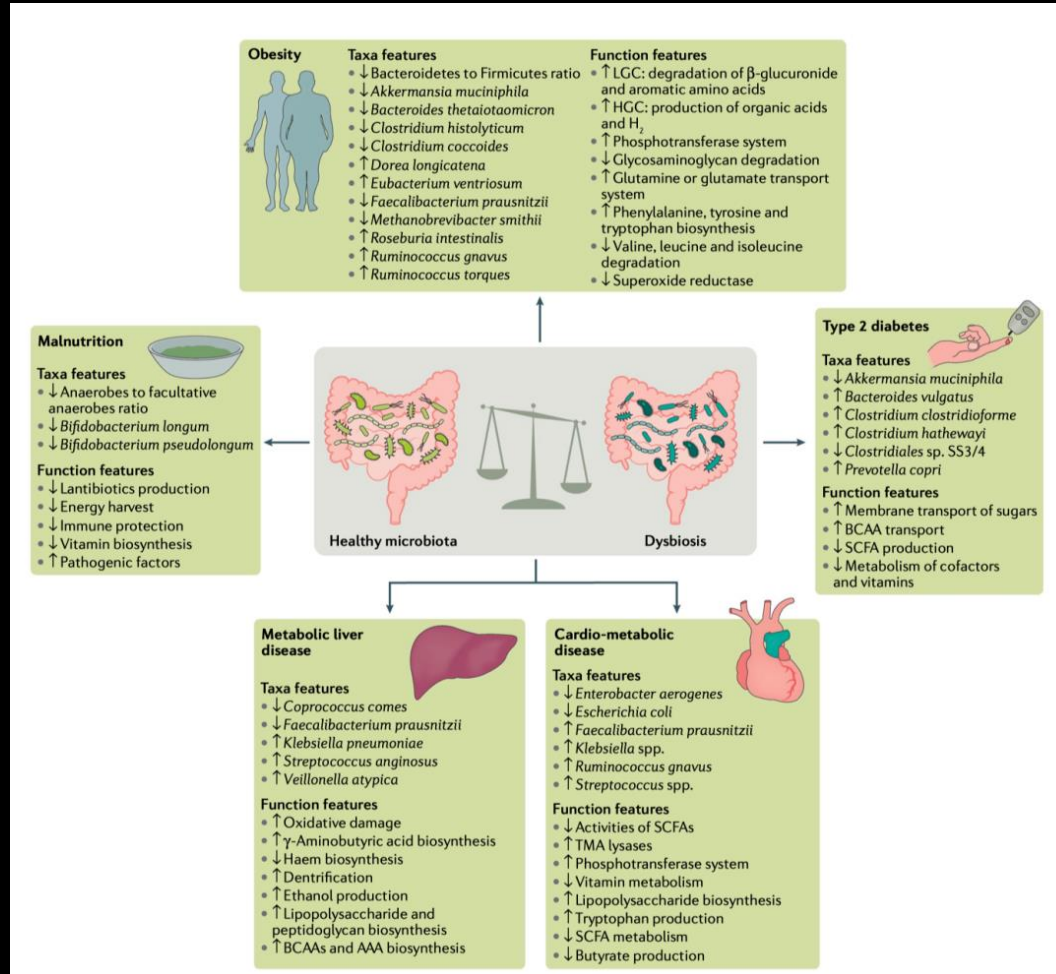
4



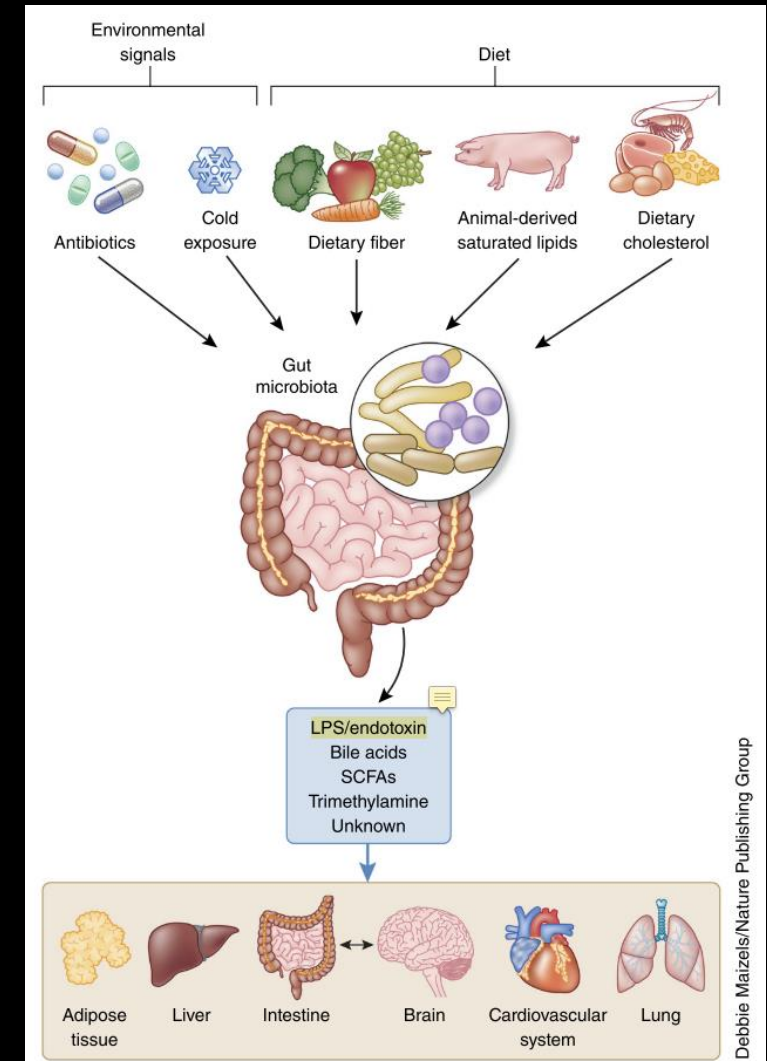
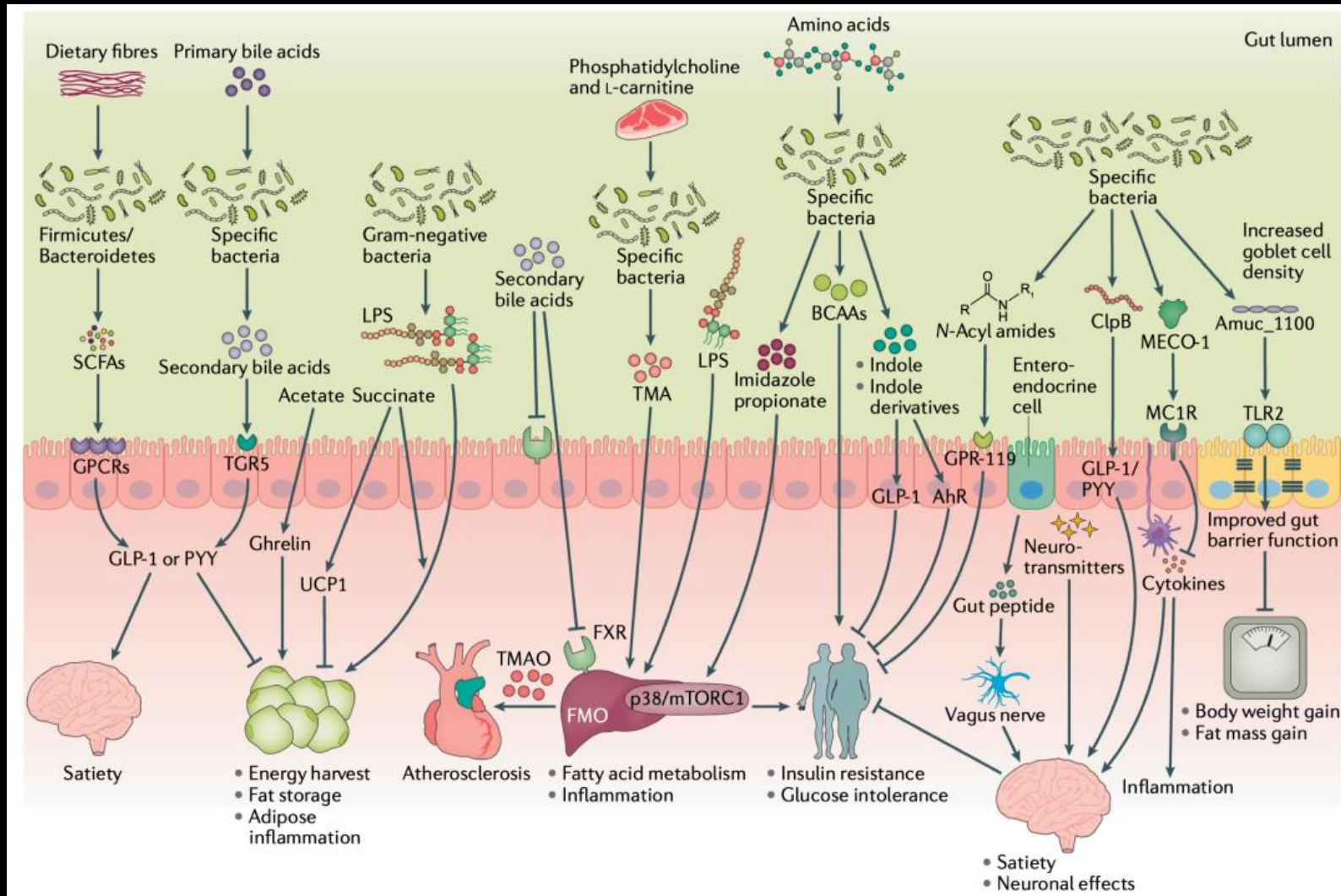
"Microbes are integrated into almost all aspects of our lives, redefining what it means to be human."



The gut microbiota is associated with various diseases in humans.

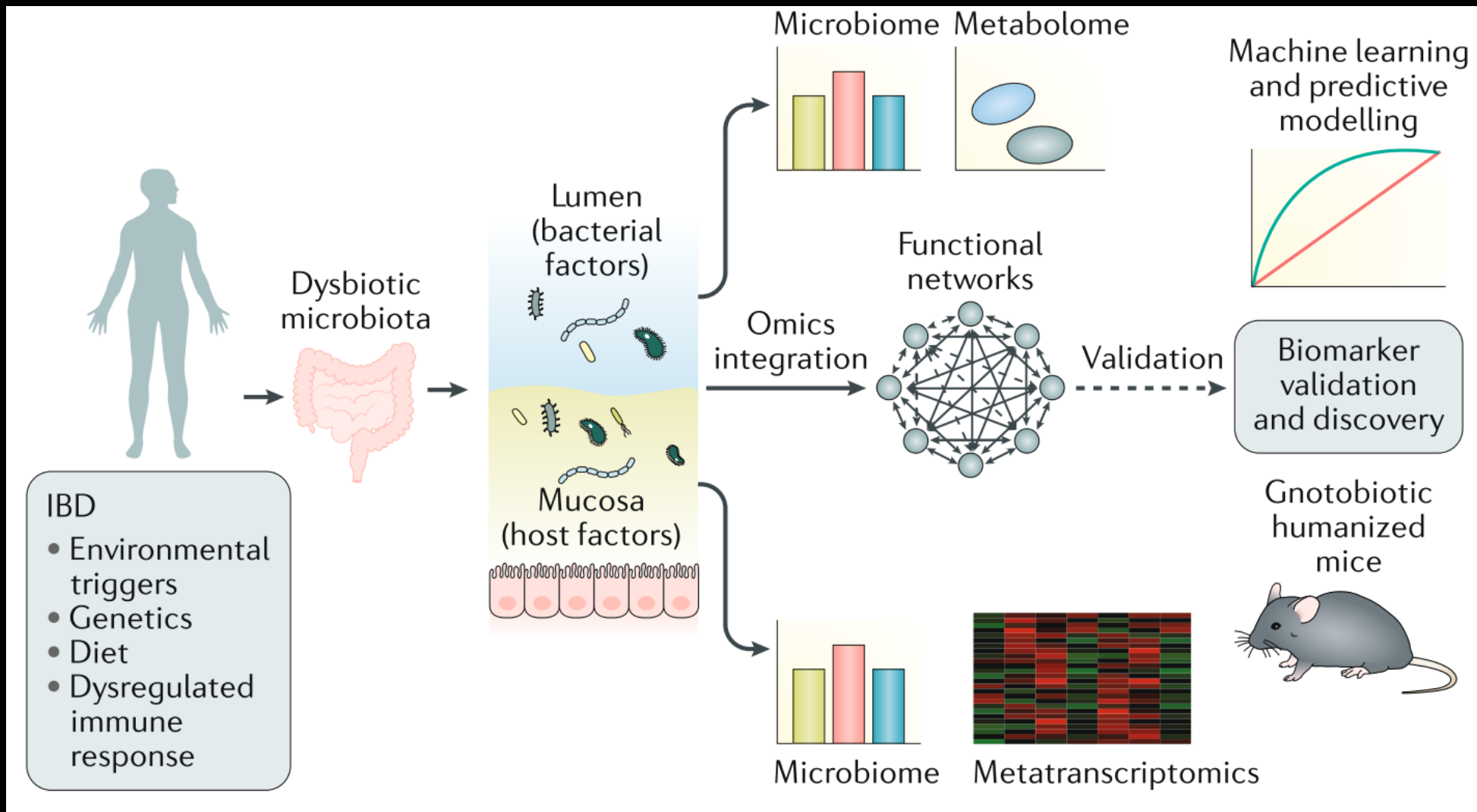


Microbial messengers regulate host metabolism

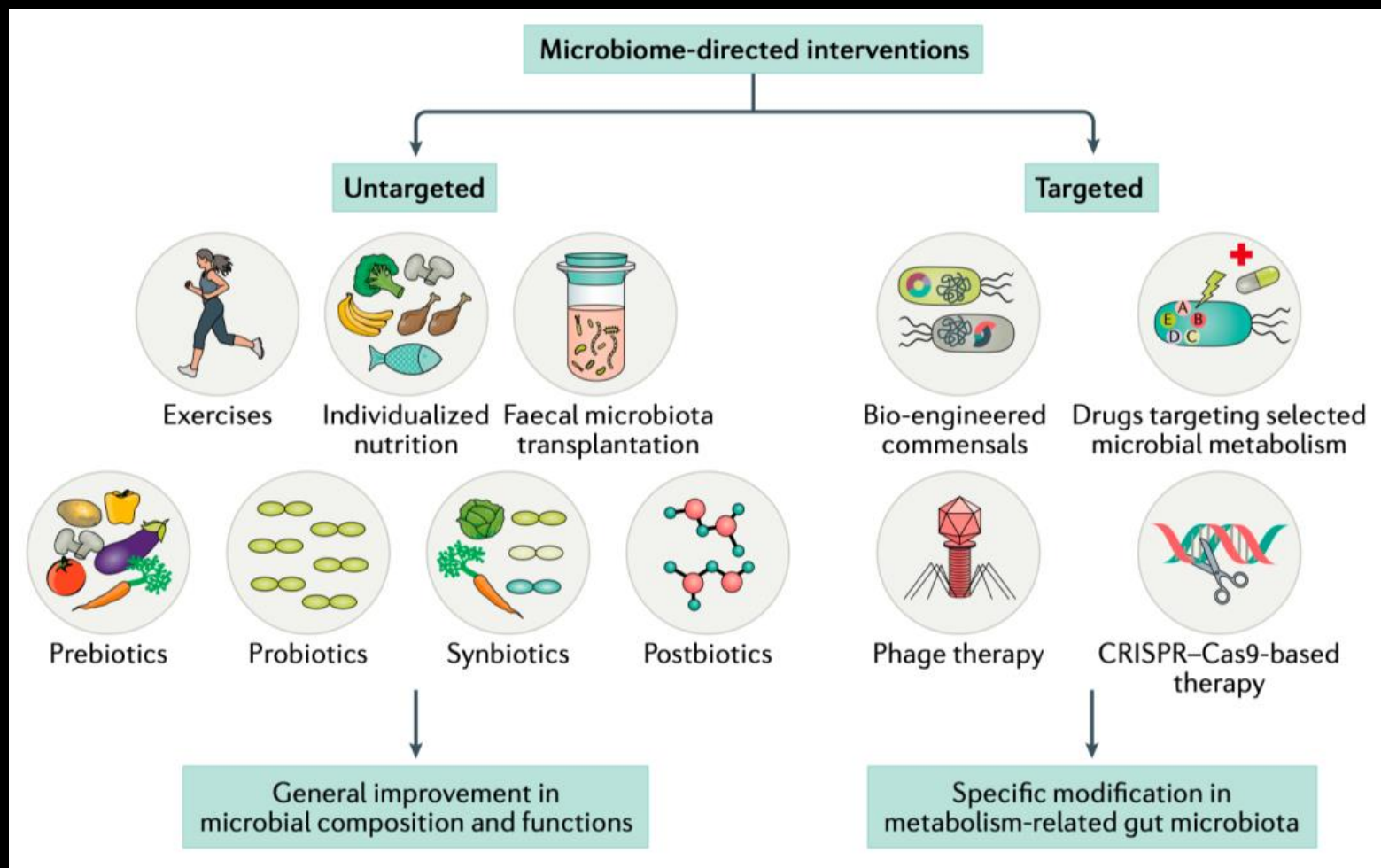


Identification and validation of biomarker signature

7



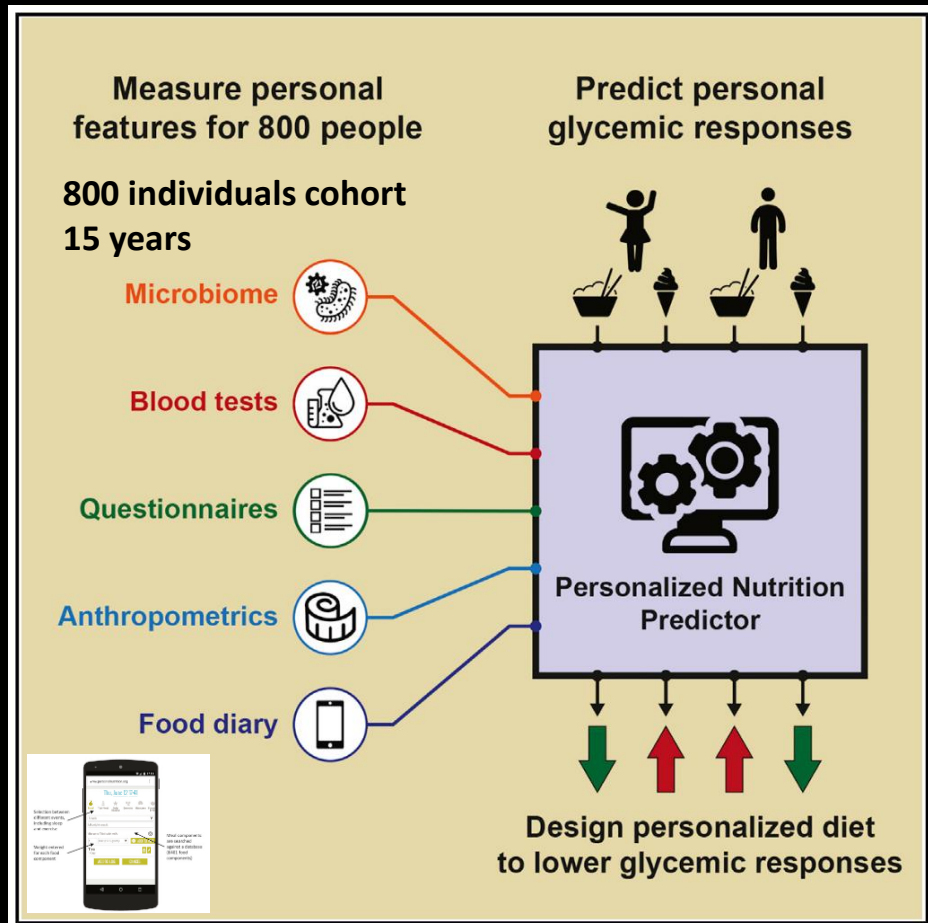
Microbiota-changing interventions



Yong fan and Oluf Pedersen, 2020, *Nature Reviews Microbiology*

Personalized Nutrition by Prediction of Glycemic Responses

9



Zeevi et al., 2015, *Cell*

Video Abstract



<https://www.daytwo.com/en/>



Eran Elinav & Eran Segal
Weizmann Institute of Science, Israel

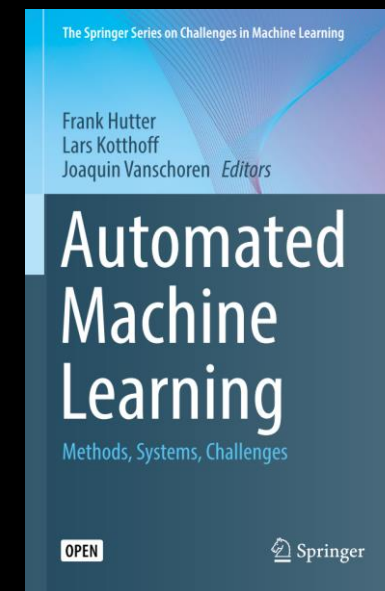
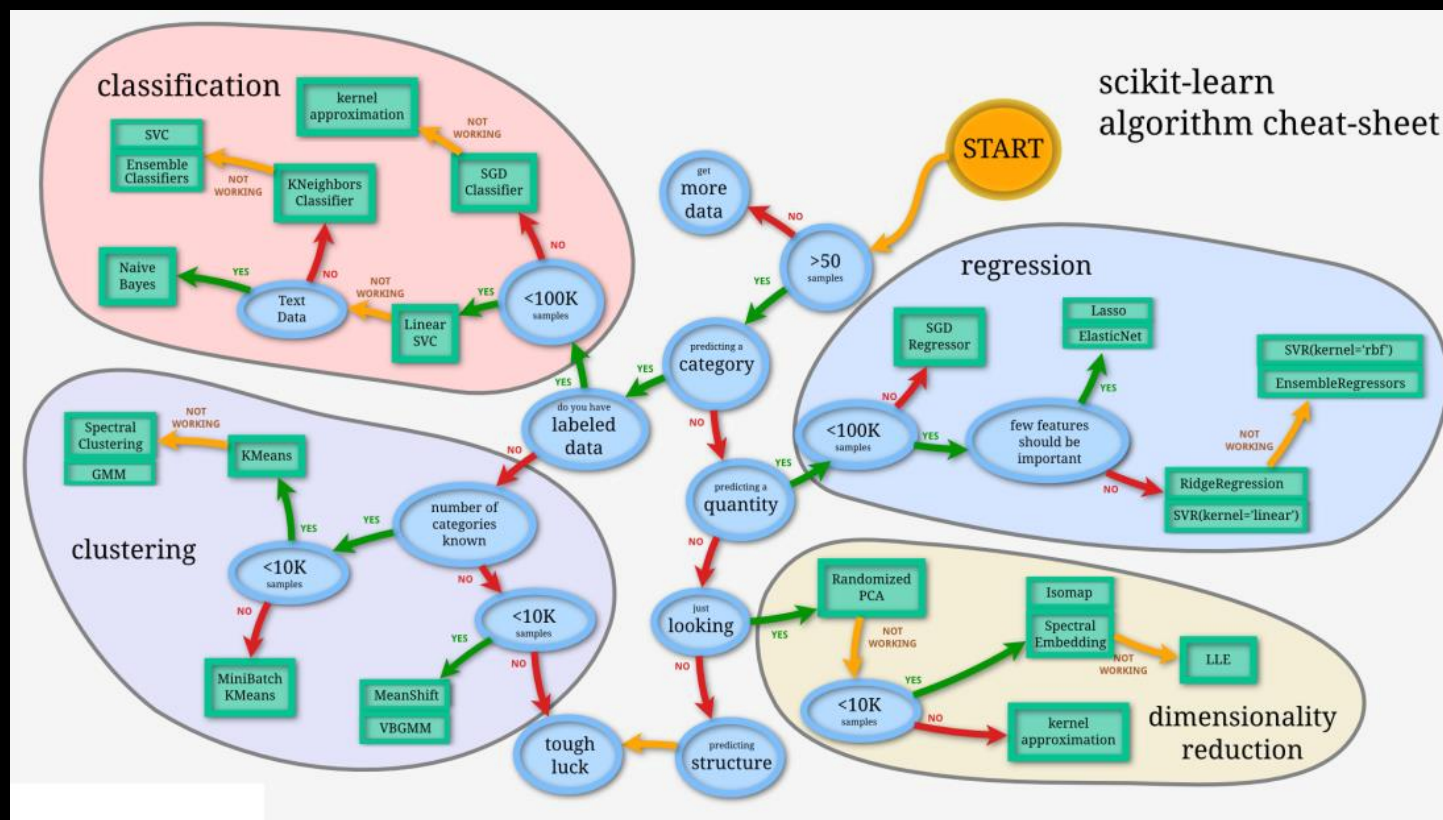
You receive your personalized nutrition app

DayTwo prepares unique personalized nutrition recommendations for each and every individual.

- Your personalized recommendations can be used for normalizing blood sugar levels throughout the day and night.
- We generate food recommendations specific to your body, like Top Meals for breakfast, lunch and dinner.

以色列威兹曼科学院2位Eran+科学计划+Cell10佳论文+创业者+4800万美元 → DayTwo

How to properly construct ML models?



Combined Algorithm Selection and Hyperparameter Optimization (CASH) problem



Outline

11

- How to automatically select the optimized model for microbiome data?
 1. mAML (an ML model-building pipeline and web-server)
 2. GMrepo ML repository
- Why balance-based disease predictions?
 1. Disease prediction models repository
 2. GBDAD (Gut Balance-Disease Association Database) for microbe-disease associations inference



mAML: an automated machine learning pipeline

12

mAML

[Home](#)

[Metagenomics data](#)

[Example Result](#)

[Web Server](#)

[Download](#)

[Help](#)



Database, 2020, 1–8

doi: 10.1093/database/baaa050

Database tool



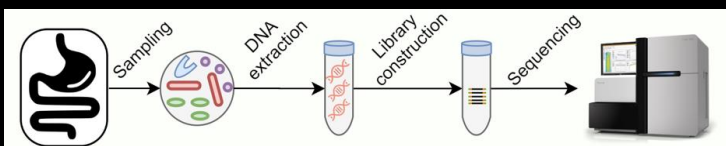
Database tool

mAML: an automated machine learning pipeline with a microbiome repository for human disease classification

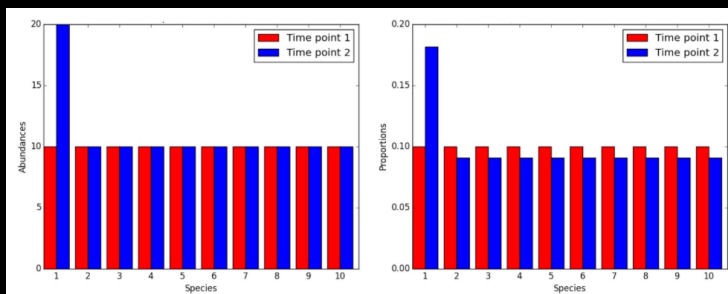
Fenglong Yang and Quan Zou*

<http://lab.malab.cn/soft/mAML/index.html>

How microbiome data differs from data from other fields?



	A	B	C	D	E	F
1	featureid	ERR1823587	ERR1823588	ERR1823589	ERR1823590	ERR1823591
2	547	0.000	0.000	0.000	0.000	0.000
3	561	0.139	0.000	0.000	0.000	0.000
4	570	0.027	0.000	0.000	0.000	0.047
5	724	0.000	0.000	0.000	0.000	0.000
6	816	0.314	0.547	0.415	0.395	0.290
7	838	0.000	0.000	0.000	0.000	0.000
8	841	0.000	0.000	0.034	0.000	0.028
9	848	0.000	0.000	0.000	0.000	0.000
10	904	0.000	0.020	0.000	0.000	0.000
11	906	0.000	0.000	0.000	0.000	0.000
12	1263	0.047	0.000	0.035	0.025	0.000
13	1301	0.000	0.000	0.000	0.000	0.000
14	1485	0.000	0.014	0.061	0.000	0.000
15	1678	0.045	0.000	0.059	0.120	0.031
16	1730	0.081	0.000	0.000	0.139	0.230
17	2172	0.000	0.000	0.000	0.000	0.000
18	29465	0.000	0.000	0.000	0.000	0.000
19	33024	0.000	0.000	0.000	0.000	0.000



- Microbiome data are high-dimensional low-sample size
- The OTU tables/microbiome abundance data are sparse
- Microbiome data are often unbalanced and biased
- **Compositionality**: proportions add to one, the change of a single microbe will also change the proportions of the remaining microbes.



Microbiota data

14

	A	B	C	D	E	F
1	featureid	ERR1823587	ERR1823588	ERR1823589	ERR1823590	ERR1823591
2	547	0.000	0.000	0.000	0.000	0.000
3	561	0.139	0.000	0.000	0.000	0.000
4	570	0.027	0.000	0.000	0.000	0.047
5	724	0.000				

Species abundance: otu_table.xls

	A	B	C	D	E	F	G	H
1	featureid	Domain	Phylum	Class	Order	Family	Genus	Species
2	547	k_Bacteri	p_Proteoc	Gamm	o_Enterof	Enterot	g_Enterobacter	s_
3	561	k_Bacteri	p_Proteoc	Gamm	o_Enterof	Enterot	g_Escherichia	s_
4	570	k_Bacteri	p_Proteoc	Gamm	o_Enterof	Enterot	g_Klebsiella	s_
5	724	k_Bacteri	p_Proteoc	Gamm	o_Pasteuf	Pasteurg	Haemophilus	s_
6	816	k_Bacteri	p_FCB gr c	Bacter	o_Bac			
7	838	k_Bacteri	p_FCB gr c	Bacter	o_Bac			
8	841	k_Bacteri	p_Terrab	c_Firmicu	o_Clo			
9	848	k_Bacteri	p_Fusobac	Fusobac	o_Fus			
10	904	k_Bacteri	p_Terrab	c_Firmicu	o_Neg			
11	906	k_Bacteri	p_Terrab	c_Firmicu	o_Neg			
12	1263	k_Bacteri	p_Terrab	c_Firmicu	o_Clo			
13	1301	k_Bacteri	p_Terrab	c_Firmicu	o_Bac			
14	1485	k_Bacteri	p_Terrab	c_Firmicu	o_Clo			
15	1678	k_Bacteri	p_Terrab	c_Actinolo	Act			
16	1730	k_Bacteri	p_Terrab	c_Firmicu	o_Clo			
17	2172	k_Archae	p_Euryar	c_Methar	o_Met			
18	29465	k_Bacteri	p_Terrab	c_Firmicu	o_Neg			
19	33024	k_Bacteri	p_Terrab	c_Firmicu	o_Neg			
20	33042	k_Bacteri	p_Terrab	c_Firmicu	o_Clo			
21	35832	k_Bacteri	p_Proteoc	delta/eo	Del			
22	39948	k_Bacteri	p_Terrab	c_Firmicu	o_Neg			
23	40544	k_Bacteri	p_Proteoc	Betap	o_Bur			
24	100883	k_Bacteri	p_Terrab	c_Firmicu	o_Erys			
25	102106	k_Bacteri	p_Terrab	c_Actinolo	Cor			
26	189330	k_Bacteri	p_Terrab	c_Firmicu	o_Clo			
27	216851	k_Bacteri	p_Terrab	c_Firmicu	o_Clo			
28	239759	k_Bacteri	p_FCB gr c	Bacter	o_Bac			

Taxonomic ranks: tax_table.xls

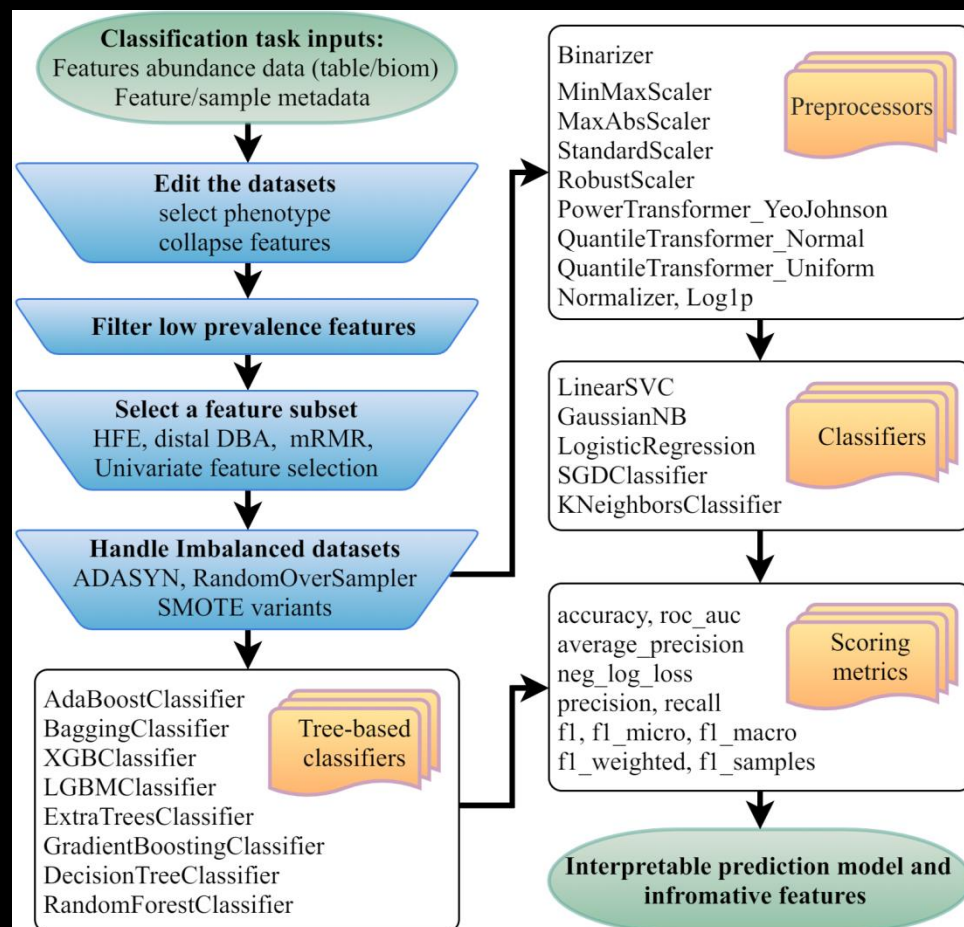
	A	B	C	D	E
1	#SampleID	Country	Disease.MESH.ID	BMI	Sex
2	ERR1823587	United States of America	D006262		
3	ERR1823588	United States of America	D006262		
4	ERR1823589	United States of America	D006262		
5	ERR1823590	United States of America	D006262		
6	ERR1823591	United States of America	D006262		
7	ERR1823592	United States of America	D006262		
8	ERR1823593	United States of America	D006262		
9	ERR1823594	United States of America	D006262		
10	ERR1823595	United States of America	D006262		
11	ERR1823596	United States of America	D006262		
12	ERR1823597	United States of America	D006262		
13	ERR1823598	United States of America	D006262		
14	ERR1823599	United States of America	D006262		
15	ERR1823600	United States of America	D006262		
16	ERR1823601	United States of America	D000755		
17	ERR1823602	United States of America	D000755		
18	ERR1823603	United States of America	D000755		
19	ERR1823604	United States of America	D000755		
20	ERR1823605	United States of America	D000755		
21	ERR1823606	United States of America	D000755		
22	ERR1823607	United States of America	D000755		
23	ERR1823608	United States of America	D000755		

Sample mapping file:
sample_data.xls



mAML pipeline for microbiome data

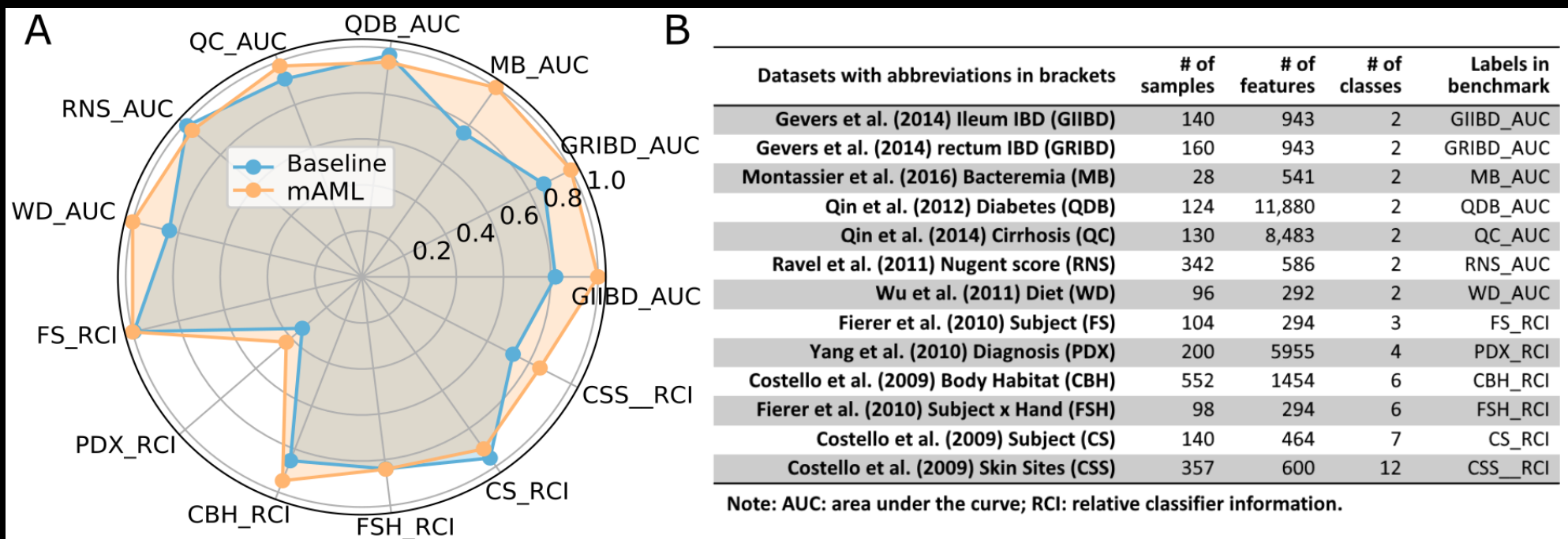
15



scikit-learn (13 classifiers and 10 feature preprocessing methods, and 4 feature subset methods)



<http://39.100.246.211:8050/Server>



Baseline datasets:

<https://knights-lab.github.io/MLRepo/>

<https://microbiomejournal.biomedcentral.com/articles/10.1186/2049-2618-1-11#Sec14>



Steps for microbiome data preprocess

17

Edit the datasets

Collapse features Order Transpose the feature matrix

Phenotype Disease.MESH.ID Labels x D000755 x D006262 x

Features can be collapsed into higher levels. Select the phenotype in the sample metadata and choose two or more labels (groups) of this phenotype

Order

Edit the datasets

Collapse features Order Transpose the feature matrix

Phenotype Disease.MESH.ID Labels x D000755 x D006262 x

Features can be collapsed into higher levels. Select the phenotype in the sample metadata and choose two or more labels (groups) of this phenotype

Filter low prevalence features

minimum prevalence threshold 0.2

Features with prevalence lower than the threshold within all classes will be filtered

Feature selection

Methods UnivariateFS

Score Function UnivariateFS

No. of features to select 50

mAML

Home Metagenomics data Example Result Web Server Download Help Contact

Build a predictive model

Upload datasets

Upload data use demo dataset Download demo dataset

Feature abundance data (table or biom)

Feature metadata Sample metadata

Edit the datasets

Collapse features None Transpose the feature matrix

Phenotype Disease.MESH.ID Labels x D000755 x D006262 x

Features can be collapsed into higher levels. Select the phenotype in the sample metadata and choose two or more labels (groups) of this phenotype

Filter low prevalence features

minimum prevalence threshold 0.2

Features with prevalence lower than the threshold within all classes will be filtered

Feature selection

Methods UnivariateFS

Score Function Mutual Information

No. of features to select 50

Sampling approach for the class imbalance problem

Methods KMeansSMOTE

Refresh Datasets preview

Feature dataset: [samples x features]
File name: otu_table.xls

index	547	561
ERR1823589	0	0
ERR1823610	0	0
ERR1823593	0	0.1123
ERR1823612	0.0071	0
ERR1823600	0	0
ERR1823609	0	0
ERR1823604	0	0
ERR1823605	0	0
ERR1823613	0	0.0165
ERR1823595	0	0
ERR1823614	0	0
ERR1823592	0	0
ERR1823597	0	0
ERR1823607	0	0
ERR1823601	0	0

Sample metadata: samples, label
File name: sample_data.xls

#SampleID	Disease.MESH.ID
ERR1823589	D006262
ERR1823610	D000755
ERR1823593	D006262
ERR1823612	D000755
ERR1823600	D006262
ERR1823609	D000755
ERR1823604	D000755
ERR1823605	D000755
ERR1823613	D000755
ERR1823595	D006262
ERR1823614	D000755
ERR1823592	D006262
ERR1823597	D006262
ERR1823607	D000755
ERR1823601	D000755

Feature selection method

Three methods for univariate feature selection are adopted here:

1. ANOVA F-value (ANOVA F-value between label/feature for classification tasks).
2. Mutual information (Mutual information for a discrete target).
3. Chi-squared stats (Chi-squared stats of non-negative features for classification tasks).



Combine algorithm selection and hyperparameter optimization

18

Select or edit the data preprocessing and classification methods

Data preprocessors Reset

StandardScaler()
RobustScaler(quantile_range=(25, 75), with_centering=False, with_scaling=True)
PowerTransformer(method='yeo-johnson', standardize=False)
QuantileTransformer(output_distribution='normal')
QuantileTransformer(output_distribution='uniform')
Normalizer()
FunctionTransformer(func=np.log1p, validate=False)

Format: one preprocessor on each individual row

Non-tree-based classifiers Reset

LinearSVC(random_state=RANDOM_STATE, dict(C=np.logspace(-4, 4, 3))
SGDClassifier(penalty='elasticnet', l1_ratio=0.15, random_state=0, dict(loss=['hinge', 'log', 'modified_huber', 'squared_hinge', 'perceptron'])

Format: one classifier and a corresponding dictionary of parameters separated by commas for each row

Tree-based classifiers Reset

DecisionTreeClassifier(random_state=0, dict(max_depth=list(map(int, np.logspace(2, 6, 5, base=2))))
BaggingClassifier(random_state=0, dict(n_estimators=list(map(int, np.linspace(5, 50, 10))))
GradientBoostingClassifier(), dict(learning_rate=[0.001, 0.01, 0.1, 0.2, 0.5])
AdaBoostClassifier(random_state=0, dict(learning_rate=[0.001, 0.01, 0.1, 0.2, 0.5])

Scalers selected

RobustScaler(copy=True, quantile_range=(25, 75), with_centering=False, with_scaling=True)
PowerTransformer(copy=True, method='yeo-johnson', standardize=False)
QuantileTransformer(copy=True, ignore_implicit_zeros=False, n_quantiles=1000, output_distribution='normal', random_state=None, subsample=100000)
QuantileTransformer(copy=True, ignore_implicit_zeros=False, n_quantiles=1000, output_distribution='uniform', random_state=None, subsample=100000)

Classifiers selected

KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski', metric_params=None, n_jobs=None, n_neighbors=5, p=2, weights='uniform'), ['n_neighbors': [5, 6, 8, 10]]
GaussianNB(priors=None, var_smoothing=1e-09, [])
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, l1_ratio=0.15, max_iter=100, multi_class='auto', n_jobs=None, penalty='elasticnet', random_state=0, solver='saga', tol=0.0001, verbose=0, warm_start=False), ['C': array([1.e-04, 1.e+00, 1.e+04])]
LinearSVC(C=1.0, class_weight=None, dual=True, fit_intercept=True,

Trees selected

DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None, max_features=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, presort=False, random_state=0, splitter='best'), ['max_depth': [4, 8, 16, 32, 64]]
BaggingClassifier(base_estimator=None, bootstrap=True, bootstrap_features=False, max_features=1.0, max_samples=1.0, n_estimators=10, n_jobs=None, oob_score=False, random_state=0, verbose=0,

Set parameters of grid search and nested cross-validation

scoring	accuracy	n_jobs	2
outer_cv	10	inner_cv	5

Fill in the email address to receive the results

Email address

Separate multiple addresses with commas. When the pipeline finished, all the results will be sent to your email address

☒ Confirmed parameter settings

Press the button to run or cancel the pipeline

Run

Result output files and description Download all the files

Result files

- Huttenhower2012_HMPBS.5.csv: filter_0.2_prevalence.csv
- Huttenhower2012_HMPBS.5.csv: mmmr_features_report.csv
- Huttenhower2012_HMPBS.5.csv: mmmr_sel_features.csv
- Huttenhower2012_HMPBS.5.csv: all_cv_results.csv
- Huttenhower2012_HMPBS.5.csv: LogisticRegression.hyper_tuned.PyCM_report.html
- Huttenhower2012_HMPBS.5.csv: LogisticRegression.hyper_tuned.confusion_matrix.csv
- Huttenhower2012_HMPBS.5.csv: LogisticRegression.hyper_tuned.confusion_matrix.svg
- Huttenhower2012_HMPBS.5.csv: LogisticRegression.hyper_tuned.classification_report.csv
- Huttenhower2012_HMPBS.5.csv: LogisticRegression.hyper_tuned.classification_report.svg
- Huttenhower2012_HMPBS.5.csv: LogisticRegression.hyper_tuned.roc_curve.csv
- Huttenhower2012_HMPBS.5.csv: LogisticRegression.hyper_tuned.roc_curve.svg
- Huttenhower2012_HMPBS.5.csv: LogisticRegression.hyper_tuned.coef.csv
- Huttenhower2012_HMPBS.5.csv: LogisticRegression.hyper_tuned.coef.svg
- Huttenhower2012_HMPBS.5.csv: log_stats.csv
- Huttenhower2012_HMPBS.5.csv: FunctionTransformer_LogisticRegression.model.z

Description

- *filter_0.2_prevalence.csv: filtered features with prevalence lower than 0.2
- *mmmr_features_report.csv: feature selection report with mRMR
- *mmmr_sel_features.csv: feature subset selected with mRMR
- *all_cv_results.csv: detailed results of GridSearchCV containing train and test scores
- *PyCM_report.html: overall statistics parameters for model evaluation
- *confusion_matrix.csv: confusion matrix of the best model
- *confusion_matrix.svg: confusion matrix plot
- *classification_report.csv: classification metrics (precision, recall, f1-score)
- *classification_report.svg: classification metrics plot
- *roc_curve.csv: roc curve data
- *roc_curve.svg: roc curve plot
- *LogisticRegression.hyper_tuned.coef.csv: coefficient of the features
- *LogisticRegression.hyper_tuned.coef.svg: coefficient of the features (the higher the more informative)
- *log_stats.csv: brief summary of the results
- *FunctionTransformer_LogisticRegression.model.z: final optimized model

Classification report of the best model

A Confusion matrix

Accuracy: 0.87

	Anterior nares	Buccal mucosa	Stool	Supragingival plaque	Tongue dorsum
Anterior nares	15	0	0	0	0
Buccal mucosa	2	141	0	0	0
Stool	1	1	124	0	0
Supragingival plaque	0	1	1	100	0
Tongue dorsum	1	15	0	0	100

B Classification report

	f1-score	precision	recall
weighted avg	0.87	0.87	0.87
macro avg	0.87	0.87	0.87
accuracy	0.87	0.87	0.87
Tongue dorsum	0.87	0.87	0.87
Supragingival plaque	0.87	0.87	0.87
Stool	0.87	0.87	0.87
Buccal mucosa	0.87	0.87	0.87
Anterior nares	0.87	0.87	0.87

C ROC curve

True Positive Rate

False Positive Rate

D Coefficients of the features

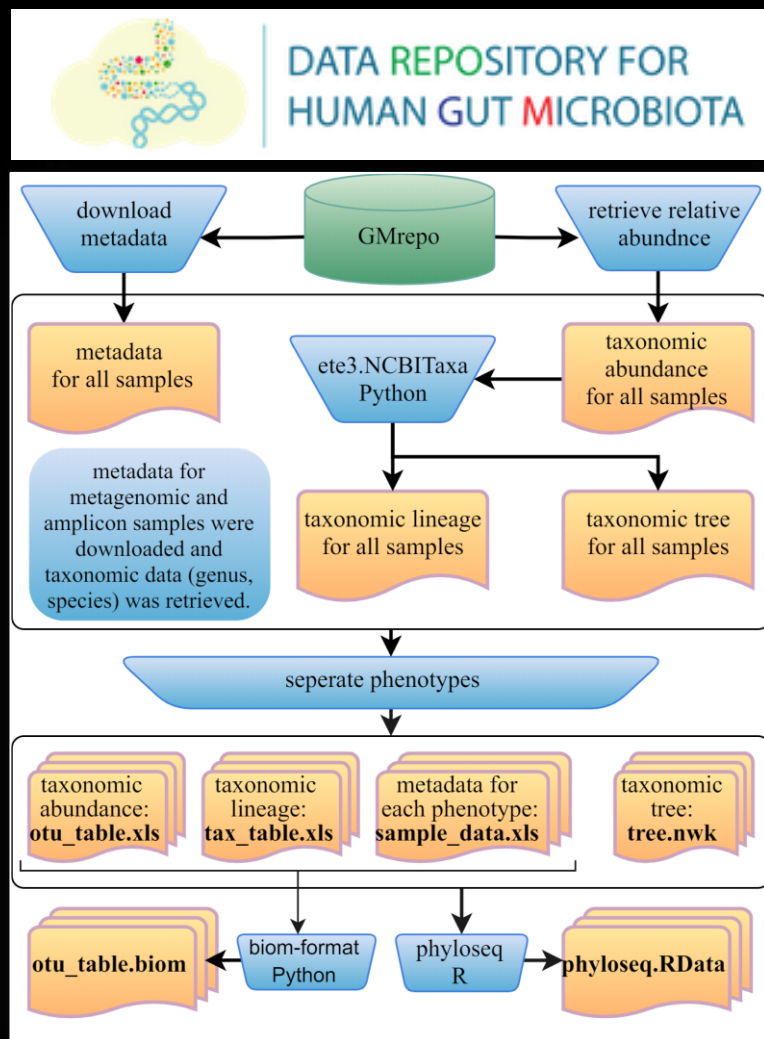
Anterior nares, Buccal mucosa, Stool, Supragingival plaque, Tongue dorsum

Visualization outputs generated by the mAML pipeline. (A) Confusion matrix. (B) Classification report. (C) ROC curve. (D) Coefficients of the top 20 important features. Note that in case of decision tree based models, feature importances bar plot will be provided instead of feature's coefficient plot.



GMrepo ML repository

19



Download all the GMrepo metagenome datasets

Datasets	# of samples	# of genus	# of species	# of classes	Classification task
D000067877	242	61	146	2	Autism_Spectrum_Disorder(55), Healthy(187)
D000236	696	46	120	2	Adenoma(220), Healthy(476)
D000755	28	37	70	2	Anemia_Sickle_Cell(14), Healthy(14)
D001172	168	51	106	2	Arthritis_Rheumatoid(96), Healthy(72)
D001289	202	56	124	2	Attention_Deficit_Disorder_with_Hyperactivity(15), Healthy(187)
D001327	272	65	146	2	Autoimmune_Diseases(85), Healthy(187)
D001714	214	57	131	2	Bipolar_Disorder(27), Healthy(187)
D002318	204	54	123	2	Cardiovascular_Diseases(17), Healthy(187)
D002446	205	58	124	2	Celiac_Disease(18), Healthy(187)
D003015	215	63	149	2	Clostridium_Infections(28), Healthy(187)
D003093	1066	78	229	2	Colitis_Ulcerative(451), Healthy(615)
D003248	252	63	145	2	Constipation(65), Healthy(187)
D003424	859	60	186	2	Crohn_Disease(227), Healthy(632)

GMrepo ML repository (GMrepo Machine Learning repository) was developed to facilitate the application of mAML, which expands the microbiome learning repository related with human disease.

Totally, the repository involves **12,429 metagenomic samples** covering 49 disease phenotypes and **38,643 amplicon samples** referring to **71 disease phenotypes**.

<http://39.100.246.211:8050/Dataset>



Outline

20

- How to automatically select the optimized model for microbiome data?
 1. mAML (an ML model-building pipeline and web-server)
 2. GMrepo ML repository
- Why balance-based disease predictions?
 1. Disease prediction models repository
 2. GBDAD (Gut Balance-Disease Association Database) for microbe-disease associations inference



GutBalance

21



GutBalance

[Home](#) [Models Repository](#) [Risk Prediction](#) [GBDAD](#) [Tutorial](#)

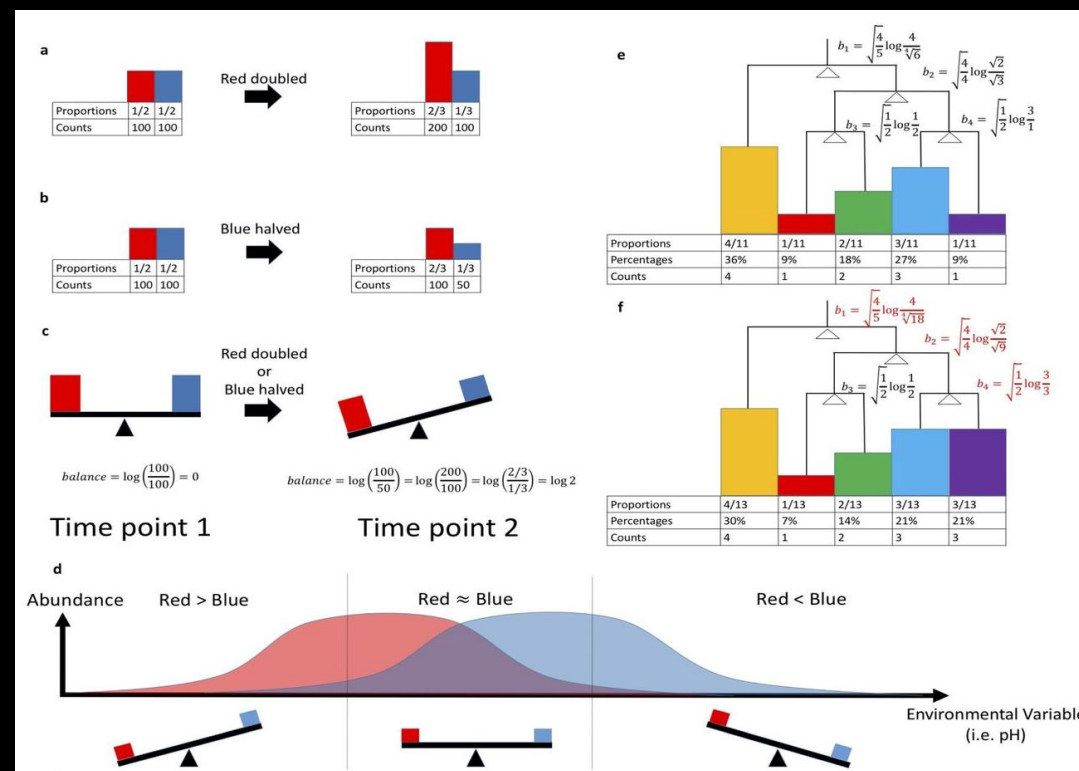
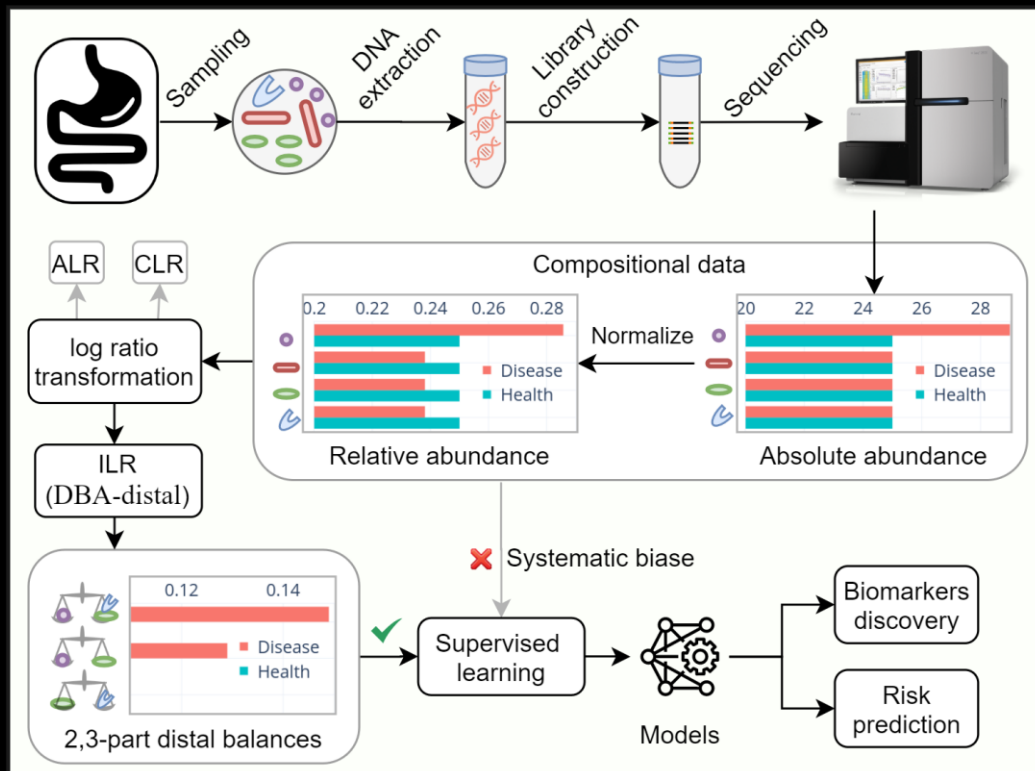
GutBalance: a server for the human gut microbiome-based disease prediction and biomarker discovery with compositionality addressed

"Everything you see exists together in a delicate balance. As king, you need to understand that balance and respect all the creatures, from the crawling ant to the leaping antelope."

--The Lion King movie quote

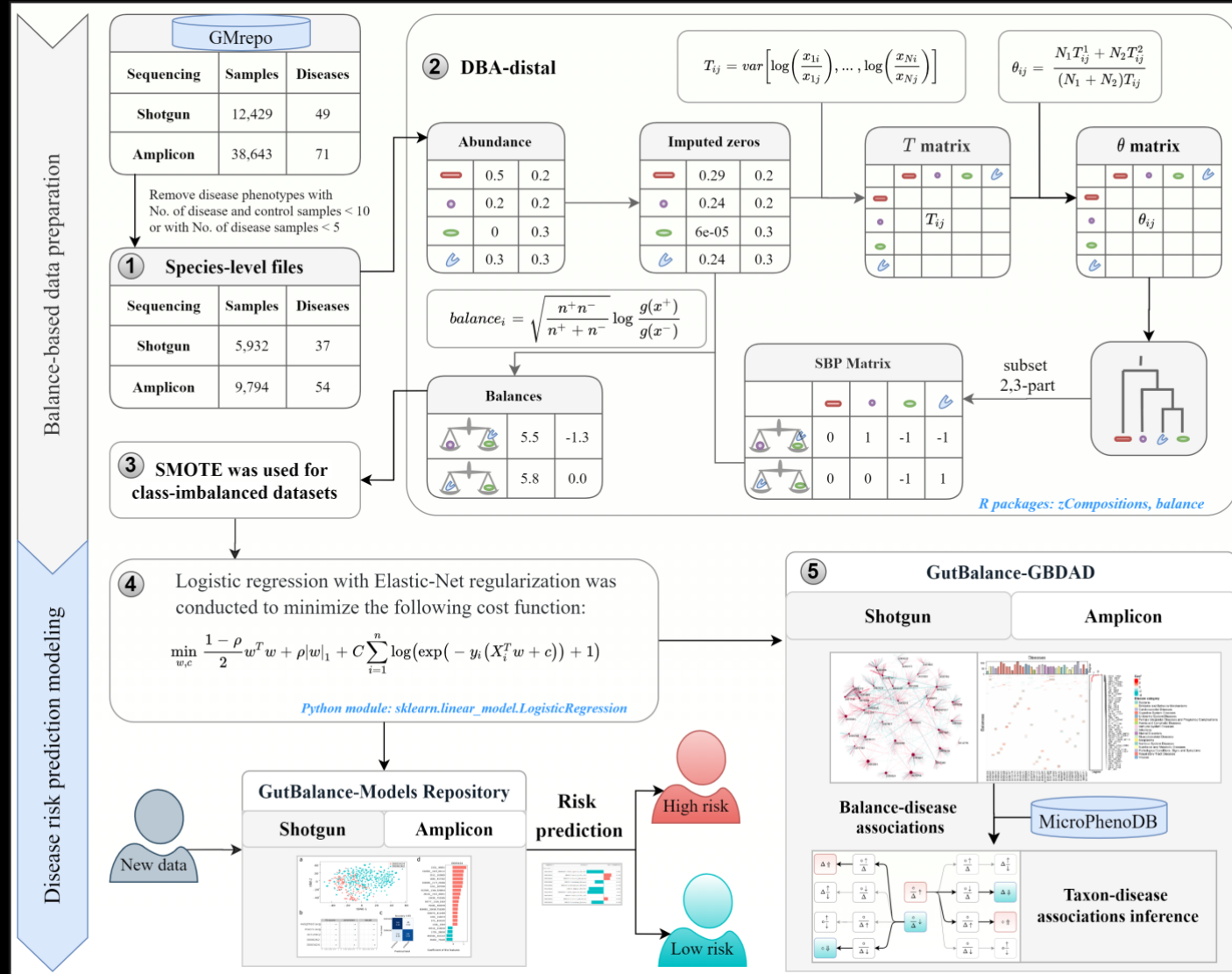
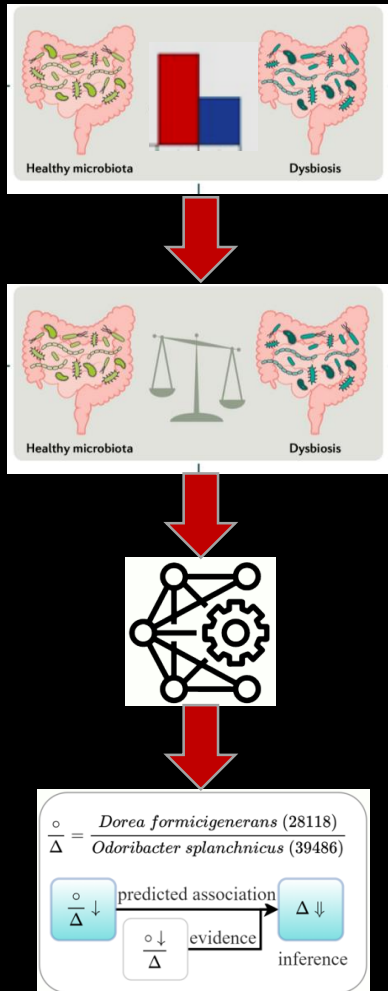
<http://lab.malab.cn/soft/GutBalance>

Why Balance?



James T. Morton et. al., 2017, *MSystems*

What is GutBalance?

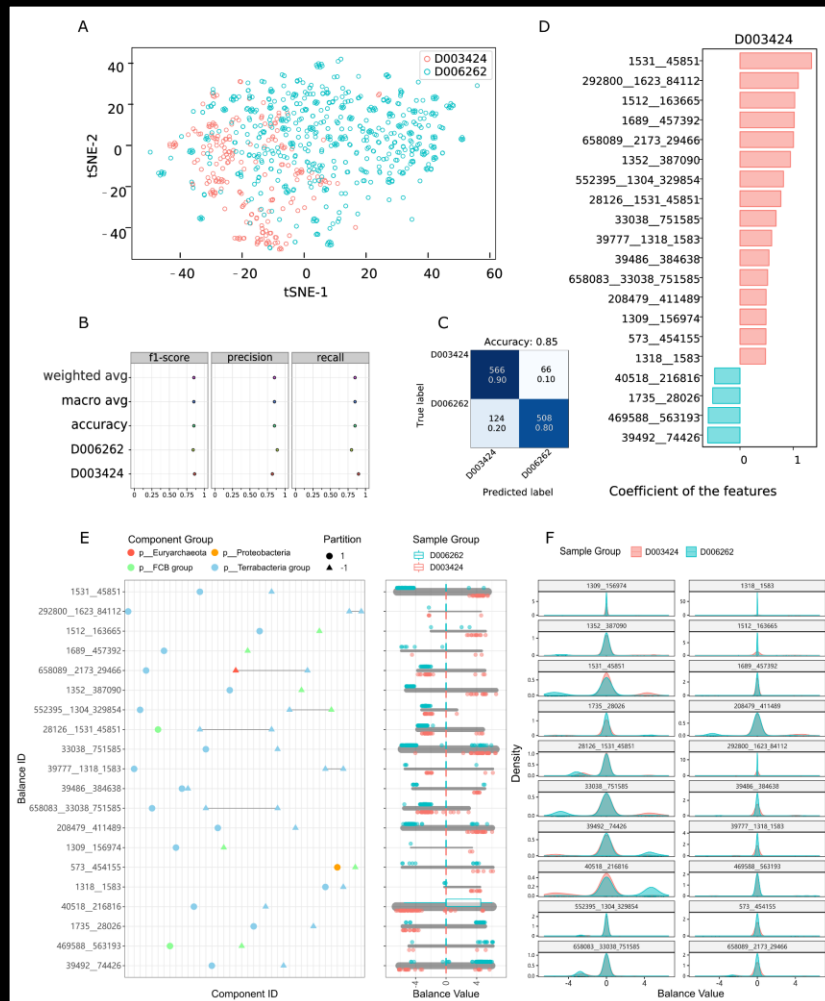




Balance datasets and LR models Repository

24

The GutBalance models repository contains **91** balances-based Logistic Regression (LR) models for human gut microbiome-associated diseases prediction, **37** for shotgun metagenomic data, and **54** for amplicon data.



Shotgun metagenome		Amplicon					
Datasets (MeSH ID)	Nr. samples	Nr. balances	Classification task	SbpMatrix	BalanceMatrix	Files4ML	Models
D000067877	242	79	Autism Spectrum Disorder(55), Healthy(187)	csv, svg	feature, label	zip(.arff, .libsvm, .csv)	
D000236	696	69	Adenoma(220), Healthy(476)	csv, svg	feature, label	zip(.arff, .libsvm, .csv)	
D000755	28	39	Anemia Sickle Cell(14), Healthy(14)	csv, svg	feature, label	zip(.arff, .libsvm, .csv)	
D001172	168	58	Arthritis Rheumatoid(96), Healthy(72)	csv, svg	feature, label	zip(.arff, .libsvm, .csv)	

<http://39.100.246.211:8051/balance/models/>



Disease Risk Prediction

25

GutBalance Home Models Repository Risk Prediction GBDAD Tutorial

Gut microbiota-related disease risk prediction

Upload a species composition file with species represented by NCBI taxonomic IDs, select one or multiple diseases to make risk probabilities of the chosen diseases.

	A	B	C	D	E	F
1	featureid	ERR1823587	ERR1823588	ERR1823589	ERR1823590	ERR1823591
2	547	0.000	0.000	0.000	0.000	0.000
3	561	0.139	0.000	0.000	0.000	0.000
4	570	0.027	0.000	0.000	0.000	0.047
5	724	0.000	0.000	0.000	0.000	0.000
6	816	0.314	0.547	0.415	0.395	0.290
7	838	0.000	0.000	0.000	0.000	0.000
8	841	0.000	0.000	0.034	0.000	0.028
9	848	0.000	0.000	0.000	0.000	0.000
10	904	0.000	0.020	0.000	0.000	0.000
11	906	0.000	0.000	0.000	0.000	0.000
12	1263	0.047	0.000	0.035	0.025	0.000
13	1301	0.000	0.000	0.000	0.000	0.000
14	1485	0.000	0.014	0.061	0.000	0.000
15	1678	0.045	0.000	0.059	0.120	0.031
16	1730	0.081	0.000	0.000	0.139	0.230
17	2172	0.000	0.000	0.000	0.000	0.000
18	29465	0.000	0.000	0.000	0.000	0.000
19	33024	0.000	0.000	0.000	0.000	0.000

Shotgun metagenome **Amplicon** ← 1

Upload dataset ☒ Use demo dataset 2

Drag and Drop or Select Files 3

Select diseases ☐ Use all diseases 4

D003920 Diabetes_Mellitus D015212 Inflammatory_Bowel_Diseases 5

Make prediction ← 6

Export

SampleID	Disease	Risk Probability
ERR2238694	D003920 Diabetes_Mellitus	0.734
ERR2238694	D015212 Inflammatory_Bowel_Diseases	0.724
ERR2238647	D015212 Inflammatory_Bowel_Diseases	0.394
ERR2238738	D003920 Diabetes_Mellitus	0.266
ERR2238648	D015212 Inflammatory_Bowel_Diseases	0.172
ERR2238687	D015212 Inflammatory_Bowel_Diseases	0.155

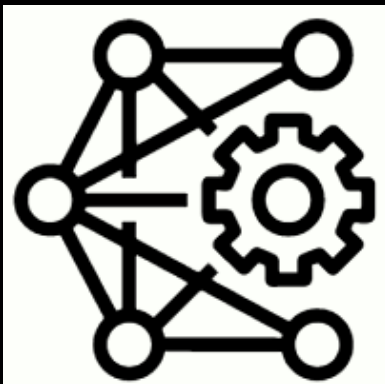
Upload data file, select diseases and predict risks



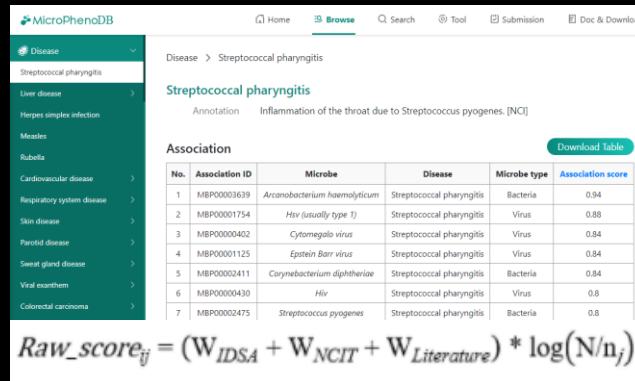
Gut Balance-Disease Association Database (GBDAD)

26

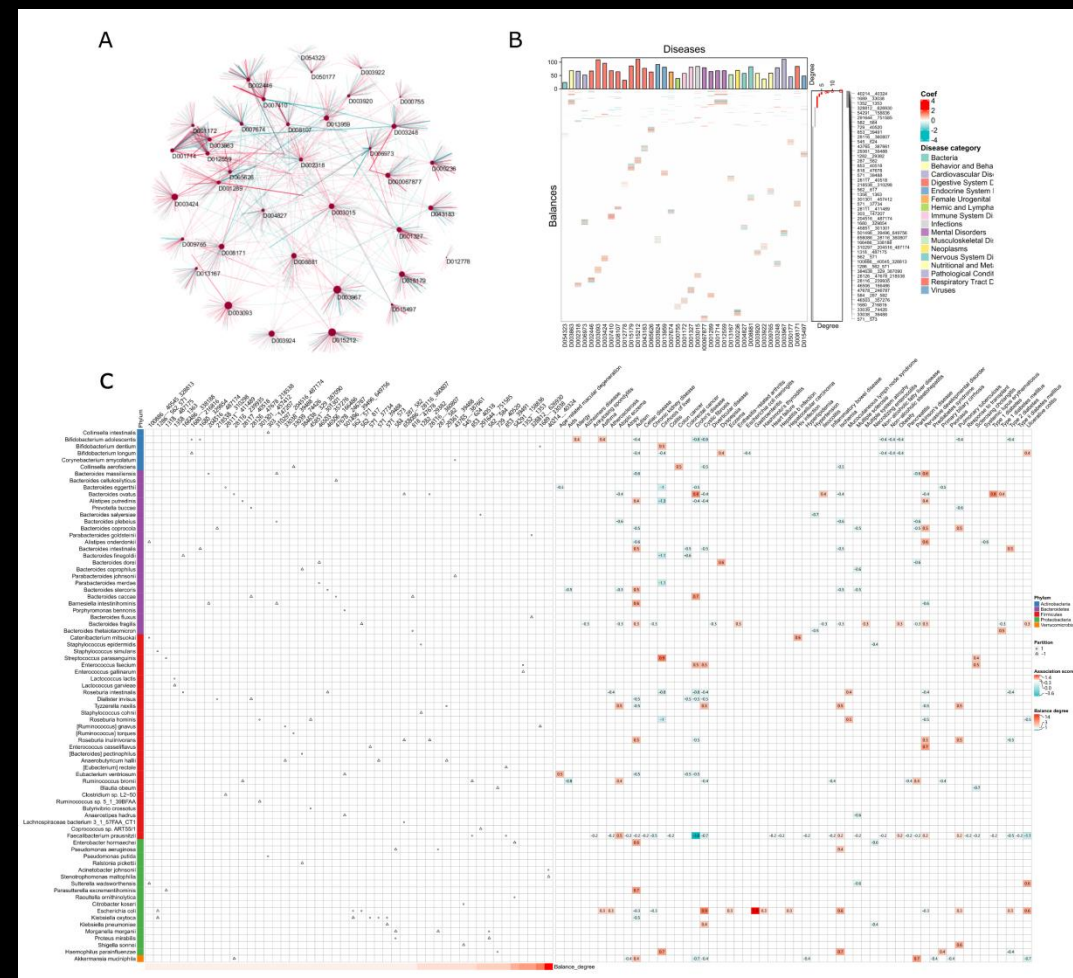
LR coefficients



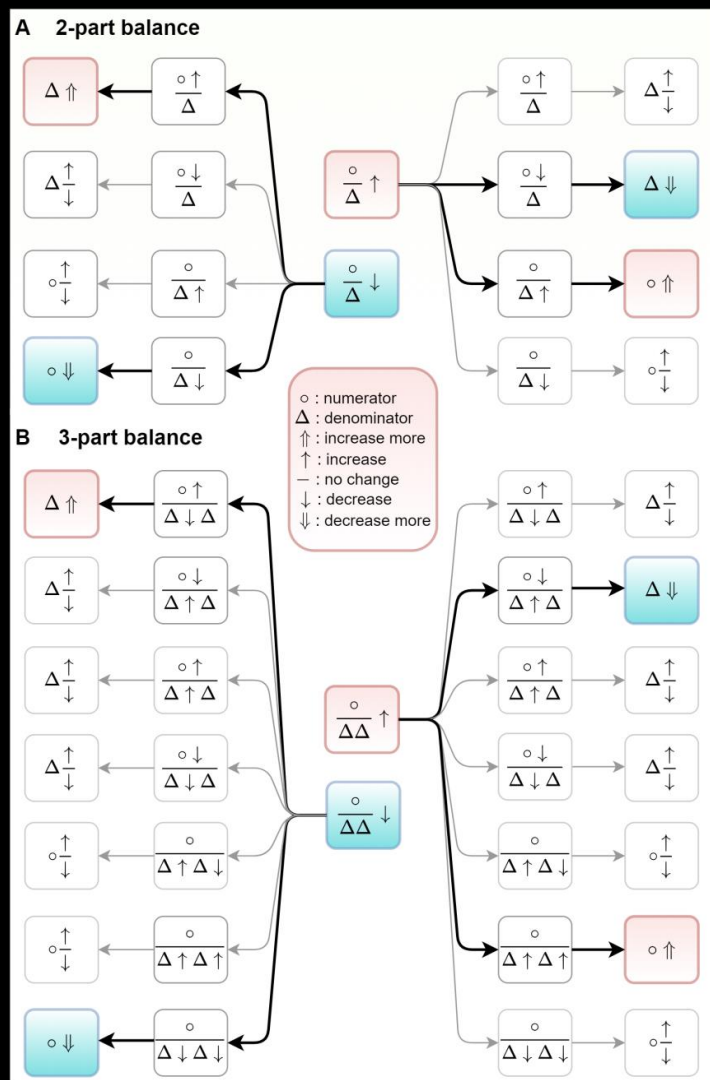
microbe-disease association score



GBDAD											
Shotgun metagenome	Amplicon										
Balance ID	BD ID	MeSH ID	DO	LRcoef	MeSH Heading	DCID	Diseases Category	NumeratorSp	Denominator1Sp	Denominator2Sp	Evidence
filter data... Inf											
261299_28117_33038	1	D015212	111	-0.675	Inflammatory Bowel Diseases	C06	Digestive System Diseases	Intestinibacter bartlettii	Alistipes putredinis	Ruminococcus gnavus	Cirrhosis of liver_-1.31; Cystic fibrosis_-0.44; Autism_0.44; Pervasive developmental disorder - not otherwise specified_0.44; Crohn's disease_-0.44
1308_626940	1	D015212	111	1.123	Inflammatory Bowel Diseases	C06	Digestive System Diseases	Streptococcus thermophilus	Phascolarctobacterium succinatutens		Pulmonary tuberculosis_0.67
1304_1211813	1	D015212	111	0.68	Inflammatory Bowel Diseases	C06	Digestive System Diseases	Streptococcus salivarius	Alistipes ihumii		Autism_-0.44; Cirrhosis of liver_0.87; Pervasive developmental disorder - not otherwise specified_-0.44
39492_40520	1	D015212	111	-0.097	Inflammatory Bowel Diseases	C06	Digestive System Diseases	Eubacterium siraeum	Blautia obeum		Autism_0.67 Sclerosing cholangitis_-0.67

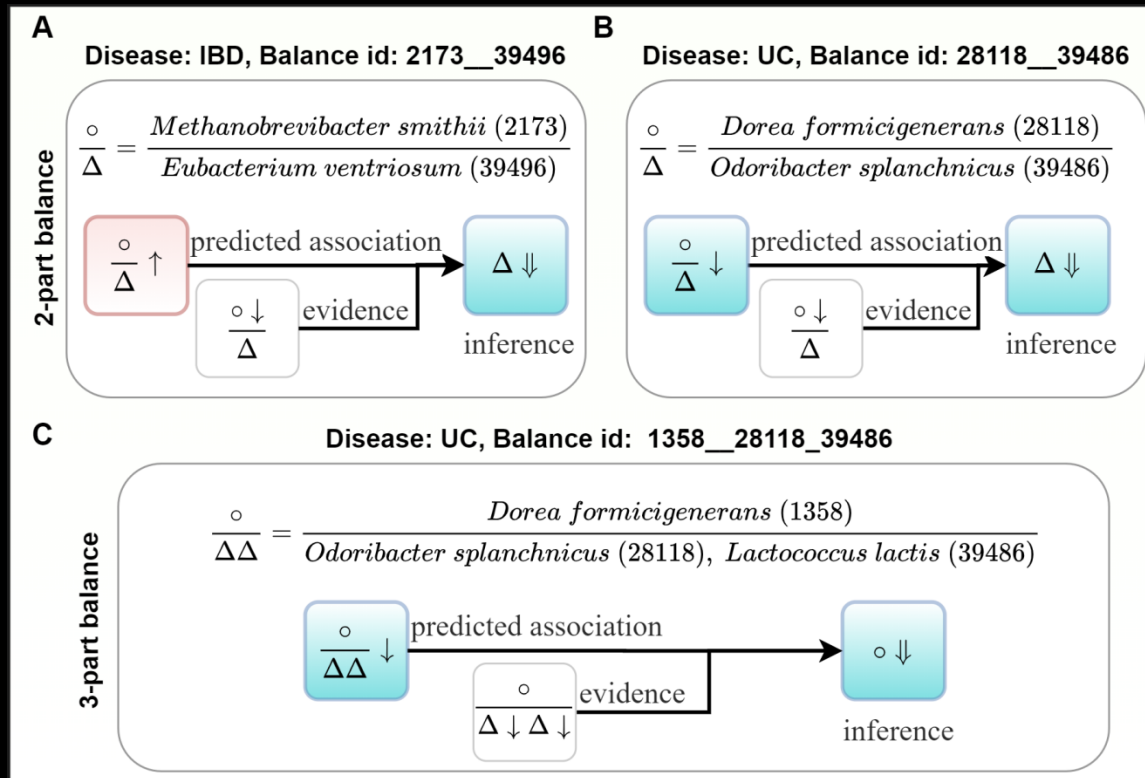


How to infer microbe-disease associations based on GBDAD?



$$balance_i = \sqrt{\frac{n^+ n^-}{n^+ + n^-}} \log \frac{g(x^+)}{g(x^-)}$$

SBP Matrix				
	—	○	●	↕
	0	1	-1	-1
	0	0	-1	1





Literature supports for association inferences

28

Ref	Balance Deg	Disease	Disease Degree	LROR	Disease	Category	Disease Cate	NumeratorI	DenominatorI	DenominatorE	DenominatorE	Associations infer	Rightness inference	References	Related descriptions	Full text links or DOI		
853_39481	5	D013167	53	0.294	Spondylo Ankylosing	C05	Musculoskeletal Diseases	<i>Faecalibacterium prausnitzii</i>	<i>Escherichia coli</i>			Ankylosing spondylitis,-023	<i>Escherichia coli</i> ,-	unknown	-	-		
105179_238538	3	D015179	86	-0.231	Colorectal Neoplasms	C06	Digestive System Diseases	<i>Prevotella copri</i>	<i>Deliber invasi</i>			Colorectal cancer,-048	Colorectal cancer,-052	<i>Prevotella copri</i> ,-	unknown	-		
853_39485	2	D003303	109	0.031	Colitis Ulcerative	C06	Digestive System Diseases	<i>Faecalibacterium prausnitzii</i>	<i>Escherichia coli</i>			Ulcerative colitis,-113	<i>Escherichia coli</i> ,-	right	Well-known probiotic role of <i>Escherichia coli</i> for the treatment of UC	-		
725_33043	2	D003424	96	-0.095	Crohn Disease	C06	Digestive System Diseases	<i>Haemophilus parainfluenzae</i>	<i>Campylobacter</i>			Crohn's disease,-050	<i>Haemophilus parainfluenzae</i> ,-	unknown	-	-		
853_3980	2	D003424	96	0.178	Crohn Disease	C06	Digestive System Diseases	<i>Faecalibacterium prausnitzii</i>	<i>Blautia obeum</i>			Crohn's disease,-363	Crohn's disease,-079	<i>Blautia obeum</i> ,-	right	Dysbiosis of the faecal microbiota in patients with Crohn's disease and their unaffected relatives	DOI: 10.1136/gut.2010.223263	
39485_45851	2	D003922	37	0.719	Diabetes Mellitus Type 1	C18	Nutritional and Metabolic Diseases	<i>Escherichia coli</i>	<i>Butyrivibrio</i>			Type 1 diabetes mellitus,-045	<i>Butyrivibrio</i> ,-	may right	Bacterial 16S Metabolism Gene Abundance in Crohn's, Ulcerative Colitis and Type 2 Diabetes Metagenomes	-	https://doi.org/10.1177/1422006710381111	
818_28116	2	D009973	52	0.985	Hypertension	C14	Cardiovascular Diseases	<i>Bacteroides</i>	<i>Bacteroides</i>			Hypertension,-050	<i>Bacteroides</i> ,-	right	Surface glycocalyx-binding proteins are essential for cerebral beta-glucan utilization by the human gut symbiont <i>Bacteroides</i>	-	https://doi.org/10.1186/1475287511111111	
821_853	2	D009626	63	-0.181	Non-alcoholic Fatty Liver Disease	C06	Digestive System Diseases	<i>Bacteroides</i>	<i>Faecalibacterium prausnitzii</i>			Non-alcoholic fatty liver disease,-023	<i>Bacteroides</i> ,-	may right	Characteristics of intestinal bacteria with fatty liver diseases and cirrhosis	-	https://doi.org/10.1186/1475287511111111	
23816_36726	2	D009765	59	0.034	Obesity	C18	Nutritional and Metabolic Diseases	<i>Blautia obeum</i>	<i>Bacteroides</i>			Obesity,-044	<i>Bacteroides</i> ,-	right	<i>Bacteroides vulgatus</i> and <i>Bacteroides</i> distal Reduce Gut Microbial Lipopolysaccharide Production and Inhibit Atherosclerosis	-	https://doi.org/10.1186/1475287511111111	
1080_105179	1	D001172	58	0.333	Arthritis Rheumatoid	C20	Immune System Diseases	<i>Blautia obeum</i>	<i>Prevotella copri</i>			New-onset untreated rheumatoid arthritis,040	<i>Blautia obeum</i> ,-	unknown	-	-		
28118_39485	1	D003303	109	-0.688	Colitis Ulcerative	C06	Digestive System Diseases	<i>Oribacterium</i>	<i>Oribacterium</i>			Ulcerative colitis,-056	<i>Oribacterium</i> ,-	right	As explained in the manuscript	-		
40545_338188	1	D003303	109	-0.042	Colitis Ulcerative	C06	Digestive System Diseases	<i>Subdoligranulum</i>	<i>Bacteroides</i>			Ulcerative colitis,056	<i>Bacteroides</i> ,-	unknown	-	-		
817_39482	1	D003303	109	0.053	Colitis Ulcerative	C06	Digestive System Diseases	<i>Bacteroides</i>	<i>Escherichia coli</i>			Ulcerative colitis,033	<i>Escherichia coli</i> ,-	right	Well-known probiotic role of <i>Escherichia coli</i> for the treatment of UC	-		
457402_105179_238538	1	D015179	86	-0.083	Colorectal Neoplasms	C06	Digestive System Diseases	<i>Escherichia coli</i>	<i>Prevotella copri</i>			Colorectal cancer,-048	Colorectal cancer,-052	<i>Escherichia coli</i> sp. 2.1.31,-	right	Well-known increased level of <i>Escherichia coli</i> in the feces of patients with colorectal cancer	-	
1351_354946	1	D015179	86	-0.272	Colorectal Neoplasms	C06	Digestive System Diseases	<i>Enterococcus</i>	<i>Hungateella</i>			Colorectal cancer,048	<i>Hungateella</i> ,-	unknown	-	-		
605900_853_1080	1	D003424	96	-0.041	Crohn Disease	C06	Digestive System Diseases	<i>Lachnospiraceae</i>	<i>Faecalibacterium prausnitzii</i>			Crohn's disease,-363	Crohn's disease,-079	<i>Lachnospiraceae</i> ,-	right	Butyrate-producing bacteria supplemented in vitro to Crohn's disease patient microbiota increased butyrate production and enhanced intestinal epithelial barrier integrity	https://doi.org/10.1186/1475287511111111	
28111_33039	1	D003424	96	0.017	Crohn Disease	C06	Digestive System Diseases	<i>Bacteroides</i>	<i>Ruminococcus</i>			Crohn's disease,-050	<i>Ruminococcus</i> ,-	unknown	-	-		
40503_47678	1	D003424	96	0.037	Crohn Disease	C06	Digestive System Diseases	<i>Parabacteroides</i>	<i>Bacteroides</i>			Crohn's disease,067	<i>Parabacteroides</i> ,-	right	Beyond metagenomics, metatranscriptomics illuminates microbiome functionality in BD	-	DOI: 10.1038/nature18181	
39491_495988	1	D003922	37	0.079	Diabetes Mellitus Type 1	C18	Nutritional and Metabolic Diseases	<i>Escherichia coli</i>	<i>Bacteroides</i>			Type 1 diabetes mellitus,-037	<i>Bacteroides</i> sp. 2.1.22,-	unknown	-	-		
328813_420932	1	D009973	52	0.216	Hypertension	C14	Cardiovascular Diseases	<i>Alkales</i>	<i>Alkales</i>			Hypertension,052	<i>Alkales</i> ,-	right	The Genus <i>Alkales</i> : Gut Bacteria With Emerging Implications to Inflammation, Cancer, and Mental Health	-	https://doi.org/10.1186/1475287511111111	
40519_333067	1	D009973	52	0.123	Hypertension	C14	Cardiovascular Diseases	<i>Ruminococcus</i>	<i>Blautia obeum</i>			Hypertension,-050	<i>Blautia obeum</i> ,-	right	Gut microbiota in atherosclerosis focus on trimethylamine N-oxide	-	https://doi.org/10.1186/1475287511111111	
39498_40518	1	D015212	111	-0.197	Inflammatory Bowel Diseases	C06	Digestive System Diseases	<i>Akkermansia</i>	<i>Ruminococcus</i>			Inflammatory bowel disease,-041	<i>Akkermansia</i> ,-	may right	Perspectives and Prospects	-	https://doi.org/10.1186/1475287511111111	
2173_39496	1	D015212	111	0.803	Inflammatory Bowel Diseases	C06	Digestive System Diseases	<i>Methanobrevibacter</i>	<i>Escherichia coli</i>			Inflammatory bowel disease,-050	<i>Escherichia coli</i> ,-	right	As explained in the manuscript	-		
1623_310862	1	D009626	63	-0.419	Non-alcoholic Fatty Liver Disease	C06	Digestive System Diseases	<i>Lactobacillus</i>	<i>Proteus</i>			Non-alcoholic fatty liver disease,056	<i>Proteus</i> ,-	right	The Molecular and Mechanistic Insights Based on Gut-Liver Axis: Nutritional Target for Non-Alcoholic Fatty Liver Disease (NAFLD) Improvement	-	https://doi.org/10.1186/1475287511111111	
1680_2173	1	D009765	59	0.169	Obesity	C18	Nutritional and Metabolic Diseases	<i>Blautia obeum</i>	<i>Methanobrevibacter</i>			Obesity,-040	Obesity,050	<i>Methanobrevibacter</i> ,-	right	Obesity-associated gut microbiota is enriched in <i>Lactobacillus reuteri</i> and <i>Blautia obeum</i> in <i>Blautia obeum</i> animals and <i>Methanobrevibacter</i> animal	-	https://doi.org/10.1186/1475287511111111
817_347287	1	D009765	59	-0.298	Obesity	C18	Nutritional and Metabolic Diseases	<i>Bacteroides</i>	<i>Collinsella</i>			Obesity,033	<i>Collinsella</i> ,-	may right	Bioaugmentation of a Common Product of Food Processing by a Human Gut Bacterium	-	https://doi.org/10.1186/1475287511111111	

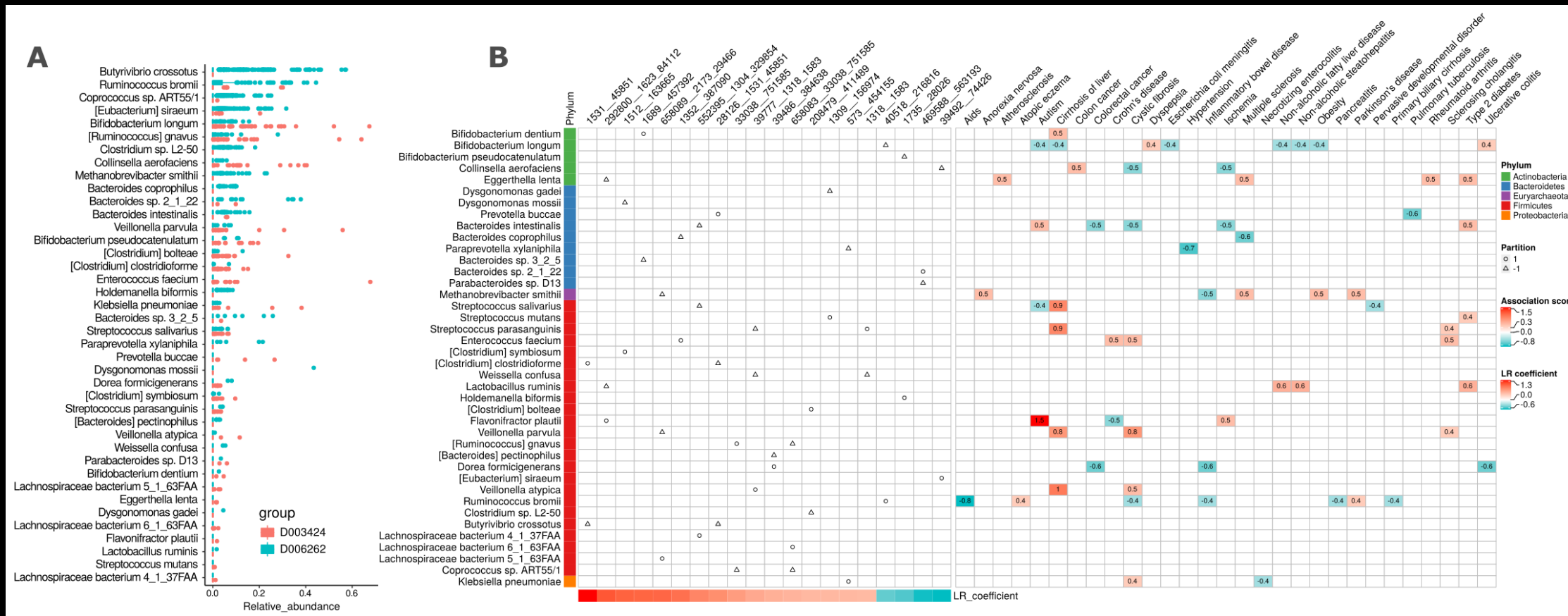
18 of 26 inferences with literature support

18 of 26 inferences with literature support

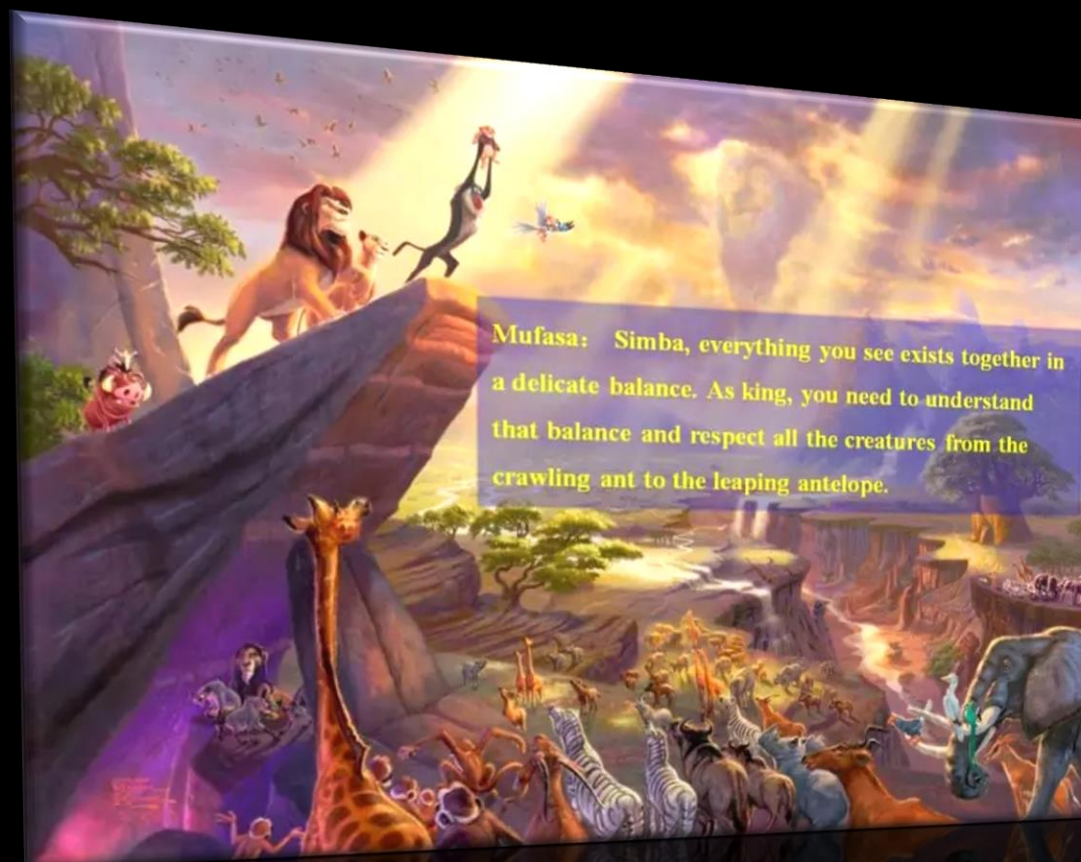


Microbe-disease association inference from species abundance

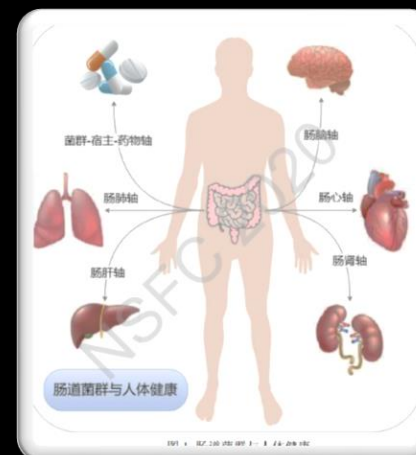
29



Balance: 中庸之道



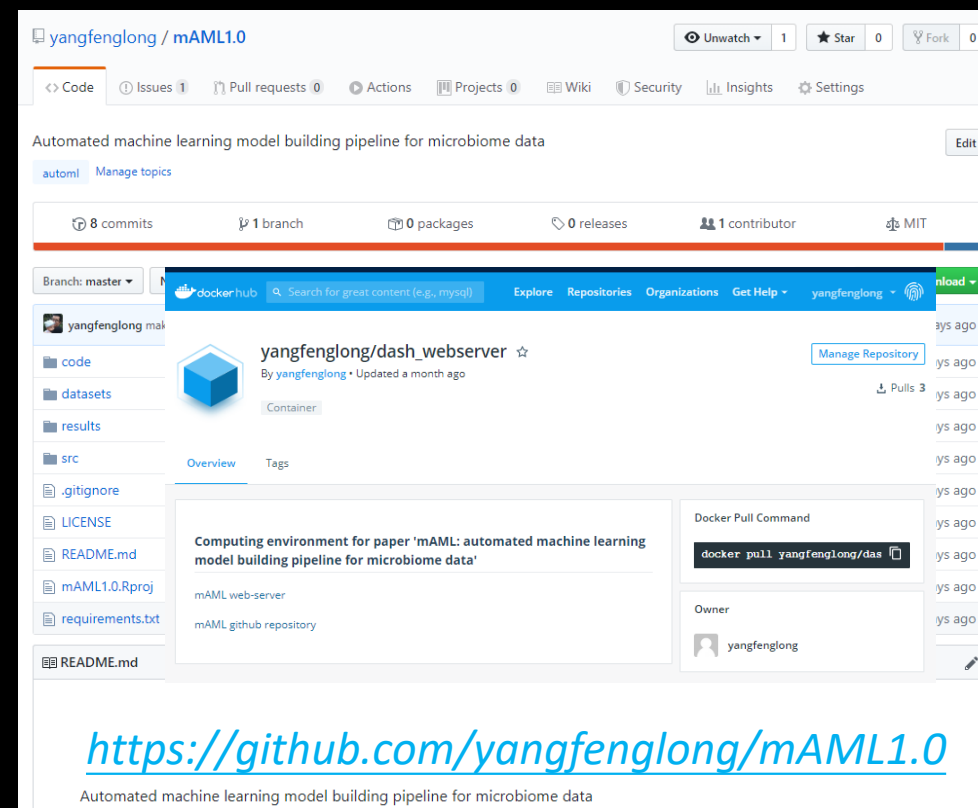
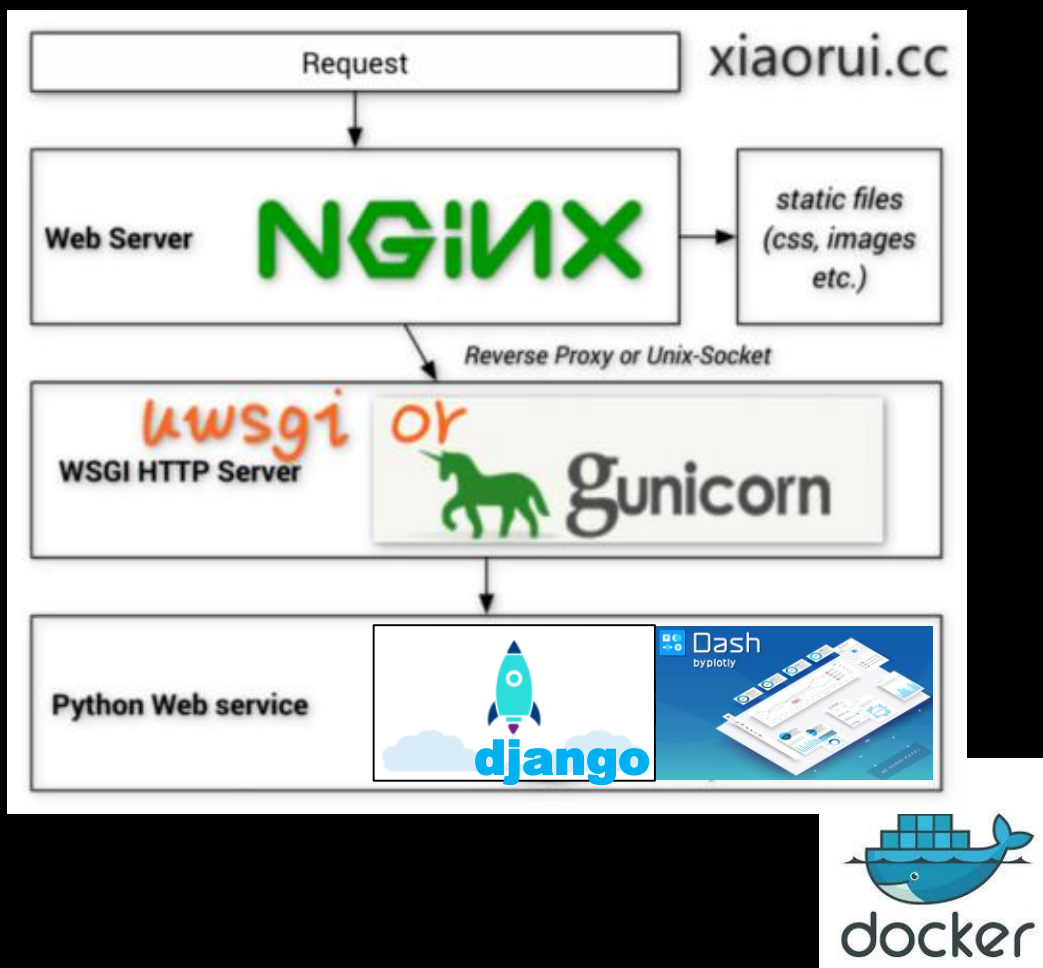
东晋葛洪《肘后备急方》，也就是启发屠呦呦先生提取青蒿素的古代医方，当中记载：用大粪久放之后发酵及沉淀，得到上清液，治疗重度肠道疾病。





How the web-server was build?

31



Thanks

Wu, S. *et al.* GMrepo: a database of curated and consistently annotated human gut metagenomes. **Nucleic Acids Res** (2020)



Yao G, *et al.* MicroPhenoDB associates metagenomic data with pathogenic microbes, microbial core genes, and human disease phenotypes. **BiorXiv** (2020)

