

精准医学项目组网络会议

## Gut Microbiome-Balance in Human Health and Disease

电子科大邹权课题组,杨凤龙



## What is microbiome?

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







### Who am I in a systematic view ?



Deringer



## How our microbes make us who we are







https://www.ted.com/talks/rob\_knight\_how\_our\_microbes\_make\_us\_who\_we\_are



### The gut microbiota is associated with various diseases in humans.





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

Schroeder, et. al. 2016 Nature Medicine



### Microbial messengers regulate host metabolism



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

#### Schroeder, et. al., 2016, Nature Medicine



### Identification and validation of biomarker signature



7

Haller et. al., 2019 Nature Reviews



## Microbiota-changing interventions



Yong fan and Oluf Pedersen, 2020, Nature Reviews Microbiology



## Personalized Nutrition by Prediction of Glycemic Responses





### 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

#### Zeevi et al., 2015, Cell



## How to properly construct ML models?



Combined Algorithm Selection and Hyperparameter Optimization (CASH) problem



- 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 microbedisease associations inference



## mAML: an automated machine learning pipeline

mAML	Home	Metagenomics data	Example Result	Web Server	Download	l Help
DATABASE The Journel of Biological Databases and Currelian				<i>Databa</i> doi: 10.1093/databa D	<i>se,</i> 2020, 1–8 ase/baaa050 atabase tool	
Database tool						
mAML: an a with a micr disease clas	auton obior ssifica	nated mach ne reposito ation	ine learn ry for hu	ing pipe man	eline	
Fenglong Yang ar	nd Qua	n Zou*				
			http://lah	malah cn/	soft/mAn	Al lindex htm



### How microbiome data differs from data from other fields?



	Α	В	С	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 13

- 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

	A	В		С			E		F									
1	featureid	ERR1823587	ERR	1823588	ERR182	23589 E	ERR1823	590 EF	RR18235	91								
2	547	0.000		0.000		0.000	(	0.000	0.	000	species abu	Indar	ice: otu_ta	aDIe	e.x	IS		
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		Α	В	С	D	E	F		G		Н					
6	816	0.314	1	featureid	Domain	Phylum	Class	Order	Fami	ly Ge	nus	Species	3					
7	838	0.000	2	547	k_Bacter	ip_Prote	eoc_Gam	m o_En	tero f_En	terolg_	Enterobacter	S		Ta	xΟ	nomi	c ranks ta	x table xls
8	841	0.000	3	561	k_Bacter	ip_Prote	eoc_Gam	m o_En	tero f_En	terolg_	Escherichia	S		TG.				
9	848	0.000	4	570	<bacter< td=""><td>ip_Prote</td><td>ecc_Gam</td><td>m o_En</td><td>tero f_En</td><td>terolg_</td><td>Klebsiella</td><td>S</td><td></td><td></td><td></td><td></td><td></td><td></td></bacter<>	ip_Prote	ecc_Gam	m o_En	tero f_En	terolg_	Klebsiella	S						
10	904	0.000	5	724	k_Bacter	ip_Prote	ecc_Gam	m o_Pa	steu f_Pa	steur g_	Haemophilus	S						
11	906	0.000	6	816	K_Bacter	ID_FCB C	grc_Bact	erco_Ba	C	Α	В		С	D	E			
12	1263	0.047	0	941	K_Bacter	ip_FCB (	pc_Bacto	erco_Ba	C 1 #Sa	ampleID	Country		Disease.MESH.ID	BMI	Sex	<b>C</b>		
13	1301	0.000	q	848	Racter	in Fusor	bic Fuso	bao Eus	2 ERF	R182358	7 United States of	America	D006262	_	_			
14	1485	0.000	10	904	k Bacter	ip Terra	bc Firmi	iculo Ne	3 ERF	R182358	8 United States of	America	D006262					
15	1678	0.045	11	906	k Bacter	ip Terra	bc Firmi	iculo Ne	4 ERF	R182358	9 United States of	America	D006262		_			
16	1730	0.081	12	1263	k_Bacter	ip_Terra	bc_Firmi	icu o_Clo	5 ERM	182359	0 United States of /	America	D006262	_				
17	2172	0.000	13	1301	k_Bacter	ip_Terra	bc_Firmi	icu o_Ba		102250	United States of A     United States of A	America	D006262					
18	29465	0.000	14	1485	k_Bacter	ip_Terra	bcFirmi	icu o_Clo		1823250	2 United States of	America	D006262			S	amnla mar	ning file
19	33024	0.000	15	1678	k_Bacter	ip_Terra	bc_Actir	nolo_Ac		2182359	4 United States of	America	D006262			J	ampiemar	ping me.
20	33042	0.000	16	1/30	K_Bacter	ip_lerra	bc_Firm	icu o_Clo	10 ERF	182359	5 United States of	America	D006262			S	ample dat	a xls
21	35832	0.000	10	20/65	K_Archae	ip Torra	hc_lvietr		11 ERF	R182359	6 United States of	America	D006262					
22	39948	0.000	19	33024	k Bacter	in Terra	bc Firmi	iculo Ne	12 ERF	R182359	7 United States of	America	D006262					
23	40544	0.000	20	33042	k Bacter	ip Terra	bc Firmi	iculo Clo	13 ERF	R182359	8 United States of	America	D006262					
24	100883	0.000	21	35832	k_Bacter	ip_Prote	eoc_delta	a∕∈o_De	14 ERF	R182359	9 United States of	America	D006262					
25	102106	0.000	22	39948	k_Bacter	ip_Terra	bcFirmi	icu o_Ne	15 ERF	R182360	0 United States of	America	D006262					
26	189330	0.000	23	40544	k_Bacter	ip_Prote	eo c_Beta	pro_Bu	r 16 ERF	R182360	1 United States of	America	D000755	_				
27	216851	0.063	24	100883	k_Bacter	ip_Terra	bc_Firm	icu o_Ery	17 ERF	R182360	2 United States of	America	D000755					
28	239759	0.064	25	102106	k_Bacter	ip_Terra	bc_Actir		or 18 ERF	R182360	3 United States of	America	D000755		_			
20	200100	0.001	26	189330	K_Bacter	ip_lerra	bc_Firm		0 19 ERF	182360	4 United States of /	America	D000755	_	_			
			28	230750	k_Bacter	ip_rena	urc Bact	erco Ba		182360	6 United States of	America	D000755		_			
			20	2031031		ib_1006			22 ERF	2182360	7 United States of	America	D000755					
									23 ERF	R182360	8 United States of	America	D000755	_				



## mAML pipeline for microbiome data



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



#### http://39.100.246.211:8050/Server

## Performance of mAML



Baseline datasets: <u>https://knights-lab.github.io/MLRepo/</u> <u>https://microbiomejournal.biomedcentral.com/articles/10.1186/2049-2618-1-11#Sec14</u>





### Steps for microbiome data preprocess

Edit the data	asets						
Collapse feat	ures	Order		¢	Transpo	se the feature i	matrix
Phenotype	Dise	None Domain			000755	× D006262	× •
Features can be	colla	Phylum			notype in	the sample me	tadata
and choose two	or mo	Class					
		Order					
		ганшу					
Edit the data	asets						
Collapse feat	Collapse features Order				Transpo	ose the feature	matrix
Phenotype	Dise	ase.MESH.ID 🗘	Labels	×	D000755	× D006262	× •
Features can b	Proj	ect.ID	ls. Select t	he ph	enotype in	the sample me	etadata
and choose two	Expe	eriment.type	this pheno	otype	t		
	Cour	eads.sequenced ntrv					
Filter low pr	Sex	2					
minimum pre	Host	tage					
	Rece	ent.antibiotics.use					
Features with	Dise	ase.MESH.ID	thresholds	within	a all classes	will be filtered	
			phi esnolu v	within	rair ciasses	win be filtered	

#### Feature selection

Methods	UnivariateFS	¢
	distal_DBA mRMR	
Score Fund	UnivariateFS	
No. of feat	None ures 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 dataset: [samples x Sample metadata: samples, label features] File name: sample data.xls File name: otu\_table.xls Feature abundance data (table or biom) index \$\$ 547 \$\$ 561 #SampleID \$ Disease.MESH.ID ERR1823589 ERR1823589 D006262 0 ERR1823610 ERR1823610 D000755 0 0 ERR1823593 0 0.1123 ERR1823593 D006262 ERR1823612 0.0071 ERR1823612 D000755 Edit the datasets ERR1823600 ERR1823600 D006262 0 0 ERR1823609 0 ERR1823609 D000755 Collapse features None ۵ ERR1823604 ERR1823604 D000755 0 0 Phenotype Disease.! Labels × D000755 × D006262 × ERR1823605 ERR1823605 D000755 0 0 Features can be collapsed into higher levels. Select the phenotype in the ERR1823613 0 0.0165 ERR1823613 D000755 sample metadata and choose two or more labels (groups) of this phenotype ERR1823595 ERR1823595 D006262 0 Filter low prevalence features ERR1823614 ERR1823614 D000755 0 ERR1823592 ERR1823592 D006262 minimum prevalence threshold 0.2 ERR1823597 ERR1823597 D006262 0 Features with prevalence lower than the threshold within all classes will be ERR1823607 ERR1823607 D000755 filtered ERR1823601 0 0 ERR1823601 D000755 Feature selection Feature selection method Methods UnivariateFS \$ Three methods for univariate feature selection are adopted here: Score Function Mutual information \$ 1. ANOVA F-value (ANOVA F-value between label/feature for No. of features to select 50 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). Sampling approach for the class imbalance problem Methods KMeansSMOTE ۵



### Combine algorithm selection and hyperparameter optimization

18

Select or edit the data preprocessing and classification	Scalers selected	Result output files and description Download all the files
Data preprocessors Reset	RobustScaler(copy=True, quantile_range=(25, 75), with_centering=False, with_scaling=True) PowerTransformer(copy=True, method='yeo-johnson',	Result files         Description           Huttenhower 2012, HMPBS 5.cov/filer_0.2, prevalence.csv         *filter_0.2, prevalence.csv: filtered features with prevalence lower than 0.2           Huttenhower 2012, HMPBS 5.cov/mmr (resture report.csv)         *mmr (resture seport.csv: feature selection report with mRMR
StandardScaler() RobustScaler(quantile_range=(25, 75), with_centering=False) PowerTransformer(inethod='yeo_johnson; standardize=False) QuantileTransformer(output_distribution='normal') QuantileTransformer(output_distribution='uniform')	standardize=haise) QuantileTansformer(copy=True, ignore_implicit_zeros=Faise, n_quantiles=1000, output_distribution='normal', random_state=None, subsample=1000000) QuantileTransformer(copy=True, ignore_implicit_zeros=Faise,	Set parameters of grid search and nested cross-validation  Litter/wer2012_HMPB5_cov_logitid/gression.hpertuned Cov/Liptid/HDB5_cov_Logitid/gression.hpertuned Cov/Liptid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov_Logitid/HDB5_cov
Normalizer() FunctionTransformer(func=np.log1p, validate=False)	n_quantiles=1000, output_distribution='uniform', random_state=None, subcomple=1000001	Scoring     accuracy     Injour     Z       Huttenhower2012_HMPBS.5.cx/LogiticRegression/hpertuned.roc.curve.sy     *roc.curve.sy: roc.curve.data       outer_cv     10     inner_cv     5
Non-tree-based classifiers	Classifiers selected	Huttenhower2012_HMPBS.5_cxvlg_stats.cxv       "log_stats.cxv trief summary of the results"         Fill in the email address to receive the results       "functionTransformer_LogisticRegression.model.z" final optimized model
LinearSVC(random_state=RANDOM_STATE), dict(C=np.logspace(-4, 4, 3)) SGDC(lassifier(penalty=elasticnet; I1_ratio=0.15, random_state=0), dictdoss={hinge; 'log; 'modified_huber;'squared_hinge; 'perceptron') Format: one classifier and a corresponding dictionary of parameters seperated by commas for each row	KNeighborsClassifier(algorithm="auto; leaf_size=30, metric='minkowski', metric_params=None.n_jobs=None.n_neighbors=5, p=2, weights='uniform}, [m_neighbors: [5, 6, 8, 10] Gaussiankl@priors=None.var_smoothing=1=0-09, [] LogisticRegression(C=1.0, class_weight=None, dual=False, ftj.terccpt_scaling=1, 11_ratio=0.15, max_iter=100, multi_class="auto; n_jobs=None, penalty="elasticnet; random_state=0.3oby="saga; toi=0.0001 verbose=0, warm_state="saga", ioi=0.001 verbose=0, warm_state="saga", ioi=0.001 verbose=0, warm_state=1.5ee, [C: array[[1=04, 1=04, 1=04]]] LinearSVC(C=1.0, class_weight=None, dual=True, ft_intercept=True,	Email address       126.com         Separate multiple addresses with commas. When the pipeline finished, all the results will be sent to your email address       Image: Classification report of the best model         Confirmed parameter settings       Image: Classification report of the best model
Tree-based classifiers Reset	Trees selected	Press the button to run or cancel the pipeline
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, nplimspace(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])	DecksionTreeClassifier(class_weight=None, criterion='gini', max_depth=None, min_inpurity_decrease=0.0, min_inpurity_split=None, min_samples_laef=1, min_samples_split=2, min_weight_fraction_laef=0.0, presort=False, random_state=0, splitter='best}, ['max_depth': [4, 8, 16, 32, 64]] BaggingClassifier('base_estimator=None, bootstrap=True, bootstrap_fractures=False, max_features=1.0, max_samples=1.0, n_estimators=10, n_lobs=None, cods coore=False, random_state=0, verbose=0,	Run Subject to the set of the set
38		



## GMrepo ML repository



http://39.100.246.211:8050/Dataset

#### Download all the GM repo metagenome datasets

Datasets	# of sam	ples # of ge	nus, # of spe	cies # of cl	asses Classification task
D00006787	7242	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.



- 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 microbedisease associations inference



## 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



Healthy microbiota

Healthy microbiota

 $\overline{\Delta}$ 

0

 $\overline{\Delta}$ 

Dysbiosis

 $\Delta \Downarrow$ 

inference

Dorea formicigenerans (28118)

Odoribacter splanchnicus (39486)

evidence

predicted association

∘↓

 $\Delta$ 

## What is GutBalance?



Balance-based data preparation

Disease risk prediction modeling

(1)

4

New data



 $\theta$  matrix

 $|\theta_{ij}|$ 



## Balance datasets and LR models Repository



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 metage	enome	Amplicon					
Datasets (MeSH ID)	Nr. sample	Nr. s balances	Classification task	SbpMatrix	BalanceMatrix	Files4ML	Moldels
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)	<b>a</b>
D000755	28	39	Anemia Sickle Cell(14), Healthy(14)	csv, svg	feature, label	zip(.arff, .libsvm, .csv)	<b></b>
D001172	168	58	Arthritis Rheumatoid(96), Healthy(72)	csv, svg	feature, label	zip(.arff, .libsvm, .csv)	

#### http://39.100.246.211:8051/balance/models/

24

## **Disease Risk Prediction**



Upload data file, select diseases and predict risks

Shotgun metagenome	Amplicon ← 1				
Upload dataset				<b>∠</b> Use demo	dataset 2
		Drag and [	Drop or Select Files		3
Select diseases				□Use all d	liseases 4
X D003920 Diabetes_Mel	litus	matory_Bowel_Diseases			5 × -
Make prediction	€-6				
Export				•	
\$	SampleID	÷	Disease	- ← 8	Risk Probability
	g → filter data				
	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

25



# Gut Balance-Disease Association Database (GBDAD)

LR	CC	bef	fi	cie	nts			micr	obe-d	isea	se as	sociat	ion	scor	e
	5	2			×	)		MicroPhel     Streptococcal phaye     Liver disease	noDB Disease : Strepto	Streptococcal	me <u>Browse</u> C pharyngitis <b>ngitis</b>	), Search © Tool	ビ Submission	E Doc & Downloa	d
	Ţ		2					Herpes simplex infec	tion Ann	notation Infla	mmation of the throat du	e to Streptococcus pyoge	nes. [NCI]		
	_					3		Measles Rubella	Associa	ation				Download Table	
				J	$\frown$	K		Cardiovascular disea	ce > No. Ass	sociation ID	Microbe	Disease	Microbe type	Association score	
					()			Respiratory system of	lisease > 1 ME	BP00003639 Arci	nobacterium haemolyticum	Streptococcal pharyngitis	Bacteria	0.94	
		-		5	$\smile$	-		Skin disease	> 2 ME	BP000001754	Hsv (usually type 1)	Streptococcal pharyngitis	Virus	0.88	
				C.	-	⋗		Parotid disease	> 3 ME	BP00000402	Cytomegalo virus	Streptococcal pharyngitis	Virus	0.84	
				• •		•		Sweat gland disease	> 5 ME	BP00002411 Cc	rynebacterium diphtheriae	Streptococcal pharyngitis	Bacteria	0.84	
			7		-			Viral exanthem	> 6 ME	BP00000430	Hiv	Streptococcal pharyngitis	Virus	0.8	
	\$							Colorectal carcinoma	a > 7 MB	BP00002475	Streptococcus pyogenes	Streptococcal pharyngitis	Bacteria	0.8	
hotgun metagenome	•	Amplicon										G	BDA	١D	
Balance ID	≑ BD	≎MeSH ID	≑ DD	≑ LRcoef	≑MeSH Heading	≎ DCID	≑ Diseases Category	<pre>@NumeratorSp</pre>	<pre>@Denominator1Sp</pre>	≑ Denominator	©Numerator 2Sp Evidence	<pre>@Denominator1 Ev</pre>	ridence	<pre> @Denominator2</pre>	2 Eviden
filter data					Int	f									
261299_28117_33038	1	D015212	111	-0.675	Inflammatory Bowel Diseases	C06	Digestive System Diseases	Intestinibacter bartlettii	Alistipes putredinis	Ruminococcu gnavus	s	Cirrhosis of liv Cystic fibrosis_ Autism_0.44; Per developmental di otherwise specif Crohn's disease_	er1.31; -0.44; vasive sorder - not ied_0.44; -0.44	Allergy_0.35; colitis_0.35; Atheroscleros fibrosis_0.00 cholangitis of liver_0.70 developmental otherwise spe Autism_0.35;	; Ulcera ; Asthma sis_0.35 ; Scler •0.35; C 0; Perva L disord scified_ Ischemi
308626940	1	D015212	111	1.123	Inflammatory Bowel Diseases	C06	Digestive System Diseases	Streptococcus thermophilus	Phascolarctobacterium succinatutens			Pulmonary tuberc	ulosis_0.67		
304_1211813	1	D015212	111	0.68	Inflammatory Bowel	C06	Digestive System Diseases	Streptococcus salivarius	Alistipes ihumii		Autism0.44; Cirrhosis of liver_0.87; Pervasive developmental disorden = act	Systemic lupus erythematosus0	.67		
					DISCUSCS						otherwise specified0.4	4			





## How to infer microbe-disease associations based on GBDAD?







## Literature supports for association inferences

28

Bala 🐨 Ba	alance Deg 斗 Disease 🔻	Disease Degree 🔻 LRo		Disease 🔻	Category 🖤	Diseases Cate 🔻	NumeratorS 🔻	Denominat 🔻	Denominat 🔻	NumeratorE 🔍	Denominator1E 🛛 🐨	Denominat 🔻	Associations infered	rightness i inference	References 🗸	Related descriptions	Full text links or DOI
853_39491	5 D013167	53	0.29	4 Spandylitis Ankylasing	C05	Muscukskeletal Diseases	Faecalbacterium provenitzi	Eubacterium rectale		Ankylosing spondylitis 0.23			Eukacterium rectale	unknown		<u>.</u>	
165179_218538	3 D015179	85	-0.23	1 Colorectal Neoplasms	C05	Digestive System Diseases	Prevotella capri	Dalster invise		Colorectal cancer0.40	Colorectal cancer0.52		Prevotella casri	unknown			
853_39485	2 (2003093	109	0.03	1. Colitis Ulcerative	C05	Diaestive System Diseases	Faecalbacterium pravanitzi	Eubacterium elizera		Ulcerative colitic-1.13			Eutocterium elizens	right	Well-known probletic role of $\ensuremath{\textit{Eularcterium}}$ genus for the treatment of UC		
729_33043	2 (2003424	96	-0.09	5 Crohn Disease	C05	Dizestive System Diseases	Haemophike parainfluenzae	Caprococcus eutoctus			Crohn's disease0.50		Haemophike parait/kenzae	unknown	18 of 26 infere	ences with literature s	upport
853_1680	2 (2003424	96	0.17	8 Crohn Disease	C05	Dizestive System Diseases	Faecalbacterium pravanitzi	Bficbbacterium adolescentis		Crohn's disease3.63	Crohn's disease0.79		Blicbbacterium adolescentis	right	Dysbices of the faecal microbiota in patients with Crohn's deease and their unaffected relatives	Methods Focusing on families with at least three members affected with CD, faccal sorteler of 69 patients with CD, the their unaffected relatives and SD anched controls were subjected to community (ingentiting of the produminant microbiota using densitying anders get electrophonesis (DGCE). To analyze the DGCE profiles, BioNamerics offsere environments in a training anders get electrophonesis (DGCE). To analyze the DGCE profiles, BioNamerics software	nd DOI: 10.1136/aut.2010.223263
39485_45851	2 (2003922	37	0.71	9 Diabetes Melitus Type 1	C18	Nutritional and Metaboli: Diseases	Eubacterium elizens	Buturivibrio crossotus		Type 1 diabetes melitus0.45			Butyrivitnio crossotus	may right	Bacterial Bile Metabolising Gene Abundance in Crohn's Ulcerative Colitis and Type 2 Diabetes Metagenomes	[7: Plaundzi Jahr represents however less than 1% of al the bits genes found in cluster 1 in our analysis of the Metal-Mit the main sources of bits breig from <i>Resolution interestica</i> (2006, <i>Psychiatricitances Lessonium</i> (106) and <i>Resolution</i> crossouts (200, Interestingly, cluster 1 is largely composed of bakyrate-producing bacteria, also positively associated with hereits).	data. http://ournalusios.org/isionae/ article?id=10.1371/ournal.come. 01.15175
818_28116	2 D000973	52	0.98	5 Hypertension	C14	Cardiovascular Diseases	Bacteroides thetaiotaomicron	Bacteroides ovatus		Hypertension0.50			Bacteroides ovatus,-	right	Surface glycan-binding proteins are essential for cereal beta-glucan utilization by the human gut symbiont Bacteroides ovatus	Here, we prevent combined biophysical structural, and in vivo characterization of the two SGBPs encoded by the Biocheroldis output mixed-Hiskoga Pulgara utilization toos (PMLGRL), thereby elucidation that his virial in the metabo of this biolational detancy ennel polyanchroteka in particular, molecular insiste paired through several crystalbargachic this biolational detancy ennel polyanchroteka in particular, molecular insiste paired through several crystalbargachic	fem https://inkapipoer.com/article/ 10.1007/c0018-019-03115-3
821_853	2 0065626	63	-0.18	Non-alcoholic Fatty Liver Disease	C06	Digestive System Diseases	Bacteroides volgetos	Faecaltacterium procentzi			Non-alcoholic fatty liver disease0.23		Becteroides vulgetus,-	may right	Characteristics of intestinal bacteria with fatty liver diseases and cimbols	In an animal tauk, CSTRUK MC mice were feld a high-fat der (PHD) freely for 10 weeks. Those who developed metaboli diordens, including distortios, were reanded as responses, and those with odd not were required as non-responders. The the aut microbiota from responders and non-responders were transmitted to germ-free (GF) mice that were divided	titos//www.sciencedirect.com/ science/article/si/51/005208119 3222318
216816_357276	2 (2009765	59	0.03	4 Obestv	C18	Nutritional and Metabolic Diseases	Blickbacterium knaum	Bacteroides dorei		Obesity0.44			Bacteroides darei	right	Becteroides vulgatus and Becteroides dorel Reduce Gut Microbial Upopolyseccharide Production and Inhibit Atherosclerosis	Caven that gat microbiotiz-derived LVB and settemic endotoxiems are involved in the oriset and progression of not on adverscription but also many provident disorders such as inflummation boxel disease, buelty and related metabalic deeses; and nonalcoholic steatoheastits, our findings suggest that Biscteroides treatment may serve as a novel and	https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/ https://www.ahaio.emaik.cro/doi/
1680_165179	1 D001172	58	0.33	3 Arthritis Rheumatoid	C20	Immune System Diseases	Blicbbacterium acblescentis	Prevotella capri			New-onset untreated rheumatoid arthrits_0.40		Břiddacterium addescentis_+	unknown		-	
28118_39496	1 (2003093	109	-0.68	8 Colitis Ulcerative	C05	Dizestive System Diseases	Odmbacter galanchnicus	Dorea Tormicigenerans			Ulcerative colitis0.56		Odmbacter galanchnicus	right	As explained in the manuscript		
40545_338188	1 0003093	109	-0.04	2 Colitis Ulcerative	C05	Digestive System Diseases	Sutterella wadsworthereis	Recteroides Anegoldii		Ulcerative colitis_0.56			Bacteroides finegoldL+	unknown			
817_39492	1 D003093	109	0.05	3 Colitis Ulcerative	C05	Dizestive System Diseases	Bacteroides fragilis	Eubacterium skaeum		Ulcerative colitis_0.33			Eukacterium skaeum	right	Well-known probatic role of <i>Eulacterium</i> genus for the treatment of UC		
457402_105179.218538	1 D015179	85	-0.08	3 Colorectal Neoplasms	C05	Digestive System Diseases	Eutacterium ga 3.1.31	Prevatella capri	Dialister invisus		Colorectal cancer0.40	Colorectal cancer0.52	Eukacterium ga 3.1.31	right	Well-known increased level of <i>Eubactenium</i> genus in the feces of patients with colorectal cancer		
1351_154046	1 D015179	86	-0.27	2 Colorectal Neoplasms	C05	Dizestive System Diseases	Enterococcue faecale	Hunzaitella hathewayi		Colorectal cancer_0.48			Hungatella hathewayi_+	unknown		-	
605950_853_1680	1 D003424	96	-0.04	1 Crohn Disease	C05	Digestive System Diseases	Lachnopiraceae bacterium 3.1.46FAA	Faecalbacterium procentzi	Bifidobacterium adolescentis		Crohn's disease3.63	Crohn's deesse 0.79	Lachnospiaceae bacterium 3.1.45FAA	right	Butyrate-producing bacteria supplemented in vitro to Crohn's deese patient microbiota increased butyrate production and enhanced intestina epithelial barrier integrity	With Biohannics bandmatching-tool, band-classes were identified over all profiles for group-specific DGGE profiles of Raminococcacese and Lachongacese as described earlier. The decreased abandrese of Firmicistes backgroup the families Raminococcacese (also referred as clostridal cluster IV) and Lachongainacese (also referred as clostridal or	g to <u>https://www.comc.org/article/P</u> uster <u>MC-00097586</u>
28111_33039	1 (2003424	96	0.01	7 Crohn Disease	C05	Digestive System Diseases	Bacteroides eggerthal	Ruminococcus torques		Crohn's disease0.50			Ruminococcus torques	unknown			
46503_47678	1 0003424	96	0.03	7 Crohn Disease	C05	Digestive System Diseases	Parabacteroides merche	Bacteroides caccae			Crohn's disease_0.67		Parabacteroides merche_+	right	Bevond metagenomics, metatranscriptomics illuminates microbiome functionality in IBD	With the addition of transmittance analysis, the investigators found that whereas the abundance of most bacterial gas correleted well with their relative contribution to the transmittanian part (when averaged accross all samples from all individuality, on an individual sample level, some species (such as Arabacteroides mender) had an extremely high corre	ies lation DOt 10.1038/wgastro.2018.15
39491_469588	1 (2003922	37	0.07	9 Diabetes Mellitus Type 1	C18	Nutritional and Metabolic Diseases	Eubacterium rectale	Bacteroides ga 2.1.22		Type 1 diabetes melitur0.37			Bacteroides ga 2.1.22	unknown		-	
329813_626932	1 D006973	52	0.21	5 Hypertension	C]4	Cardiovascular Diseases	Alistipes onderdonki	Alitipes inditinctus			Hypertension_0.52		Alistipes anderdankiL+	right	The Genus Alitabes: Gut Bacteria With Emerging Implications to Inflammation, Cancer, and Mental Health	Table 3. Summary of studies reporting the experimental or observational associations between Allstates gas and variou non-communicable diseases Bottom line: Hypertension-Pathogenic.	s <u>https://www.nchinim.nih.gov/p</u> mc/acticles/PMC7296073/
40519_333367	1 D005973	52	0.12	3 Hypertension	C14	Cardovascular Diseases	Ruminococcus calibles	Clostrickum agavagiforme		Hypertension0.50			Clostricium assaraciforme	right	Gut microbiota in atherosclerosis focus on trimethylamine N-oxide	It use suggeted that car microbiota promoted hypertension progression via regulating IP associated homones, such, sectorial, documents and nonspirationine. Animal target should will be added that MAAO and enhanced IP of hypertension via cannot affect IP of normatensive animals it was explained that TMAAO existed an indirect effect on hypertension via the sectorial document of the sector o	6 M http://minelbzary.wiey.com/d ci/hik/10.1111/arom.13038
39488_40518	1 0015212	111	-0.19	7 Diseases	C06	Digestive System Diseases	Anaerobutyricum halli	Ruminococcus bromii			Inflammatory bowel disease0.41		Anaerobutyricum hallE-	may right	Perspectives and Prospects	As we have learned more about the put microbioms, several of the dominant anaerdic species have been identified candidates to be therapeutic agents or the next generation of probotic organisms. These were mentioned in Chao. 5, include Facetalitectrium anaerdized. Amendbartcom Tall, Robebus geneis and Akhemanism mucipathe.	6 https://link.apringer.com/chapte and <u>2/10.1007/978-3-030-43246-</u> <u>1.12</u>
2173_39496	1 D015212	111	0.80	Inflammatory Bowel Diseases	C05	Dizestive System Diseases	Methanckrevikacte r smithi	Eubacterium ventriceum		Inflammatory bowel decase0.50			Eukacterium ventriasum	right	As explained in the manuscript		
1623_102862	1 D065626	63	-0.41	9 Non-alcoholic Fatty Liver Disease	C05	Digestive System Diseases	Lactobacillus ruminis	Proteus penneri		Non-alcoholic fatty liver disease_0.55			Proteus permeri_+	right	The Molecular and Mechanistic Insights Based on Gut-Liver Axis- Nutritional Target for Non-Alcoholic Fatty Liver Disease (NAFLD) Improvement	A chical research with 300 statests by Bernes et al reveals that the casculating levels of intrestrylemme-N-casid ETM is a novel inductor of metabolic synohrone and XM-LD Abhadh TMANO is surfresting in the excitosited in the low, intrestrylemme ITM etc. Abhadh TMANO abhadh TMANO is generated from gat bacteria L-camitine, cholie, or betaine have been the major substrate VMLR evolution.	AO) A), <u>https://www.mdpi.com/1422-</u> s.for <u>0067/21/3/3056/htm</u>
1680_2173	1 D009765	59	0.16	9 Obesty	C18	Nutritional and Metabolic Diseases	Blicbbacterium acblescentis	Methanotrevibact er snithi		Obesky0.40	Obesky_0.50		Methandrevbacter smithL-	right	Obesty-associated gut microbiota is enriched in Lactobacillus reuteri and degleted in Billidobacterium animalis and Methanobrevibacter smithil	In GPCR, & animale (adds ratio 1078)=0.08; 95% confidence interval (C0.039–1.01; P=0.05%) and M anithii (OR=107; C1.039–037; P=0.03) were associated with normal weight whereas (actubac/like resteri (OR=1.78; 95% C1.103–3.10; P=0.04); was accusted with observe.	Manual an com/articles
817_147207	1 0009765	59	-0.29	6 Obesity	C18	Nutritional and Metabolic Diseases	Bacteroides fragilis	Colineela Intestinala		Obesty_0.33			Colliselb intestinals_+	may right	Bioremediation of a Common Product of Food Processing by a Human Gut Bacterium	The abundance of Califeelit is positively correlated with insulin levels in overweight and obere pregnent women (Com- Aranoo et al., 2010). Califeelit is decreased in humans comunitia a high protein/See saar dee for weight loss (Maker al., 2011) and in huge 2 debetes positives comunities a light fiber det (Califeel et al., 2016).	e http://www.sciencedirect.com/ et acience/article/bid/51931312819 204583



### Microbe-disease association inference from species abundance









东晋葛洪《肘后备 急方》,也就是启 发屠呦呦先生提取 青莺的古代医方, 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?





🛛 yangfeng	long / mAML1.0	)				•	O Unwatch ▼ 1	★ Star	0 ¥ Fork 0
<> Code	ן Issues 🚺 👘 Pt	Ill requests 0	Actions	Projects 0	🔳 Wiki	C Security	🔟 Insights	🔅 Settings	
Automated n automi Mana	nachine learning r Ige topics	nodel building p	pipeline for	microbiome d	ata				Edit
🕞 8 con	nmits	0 1 branch	🗇 <b>0</b> pa	ackages	🛇 0 relea	ses	🎎 1 contribute	or	a <u>∱</u> a MIT
Branch: master	✓ I dockert	ub ۹ Search for g	reat content (e.g	g., mysql) Ex	plore Reposito	ories Organiz	ations Get Help -	r yangfenglor	nload <del>+</del>
yangfengl	ong mak	yangfenglo By yangfenglong	<b>ng/dash_v</b> • Updated a mor	webserver ជំ				Manage Re	pository Jys ago Pulls 3 Jys ago
results	Overview	Container							iys ago iys ago
<ul><li>gitignore</li><li>LICENSE</li></ul>	6	-	6		d maakina la	un in a	Docker Pull Comn	nand	ys ago ys ago
README.r	nd model I	uilding pipeline f	for paper 'm	ne data'	a machine lea	arning	docker pull y	angfenglong/da	ys ago
requireme	mAML was mAML git	b-server					Owner yangfenglo	ng	iys ago
README.r	nd						6.2		ø

#### https://github.com/yangfenglong/mAML1.0

Automated machine learning model building pipeline for microbiome data





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



DATA REPOSITORY FOR HUMAN GUT MICROBIOTA

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

