

FUNCTIONAL AND TAXONOMIC BIOMARKERS FROM THE HUMAN GUT MICROBIOME IN LIVER CIRRHOSIS AND COLORECTAL CANCER - A MACHINE LEARNING SURVEY

Baptiste HENNECART^{1,*}, Sandy Frank KWAMOU NGAHA^{1,*}, Florian PLAZA OÑATE¹, Vadim PULLER¹, Thomas MONCION¹, Edi PRIFTI^{2,3} and Raynald de LAHONDES¹



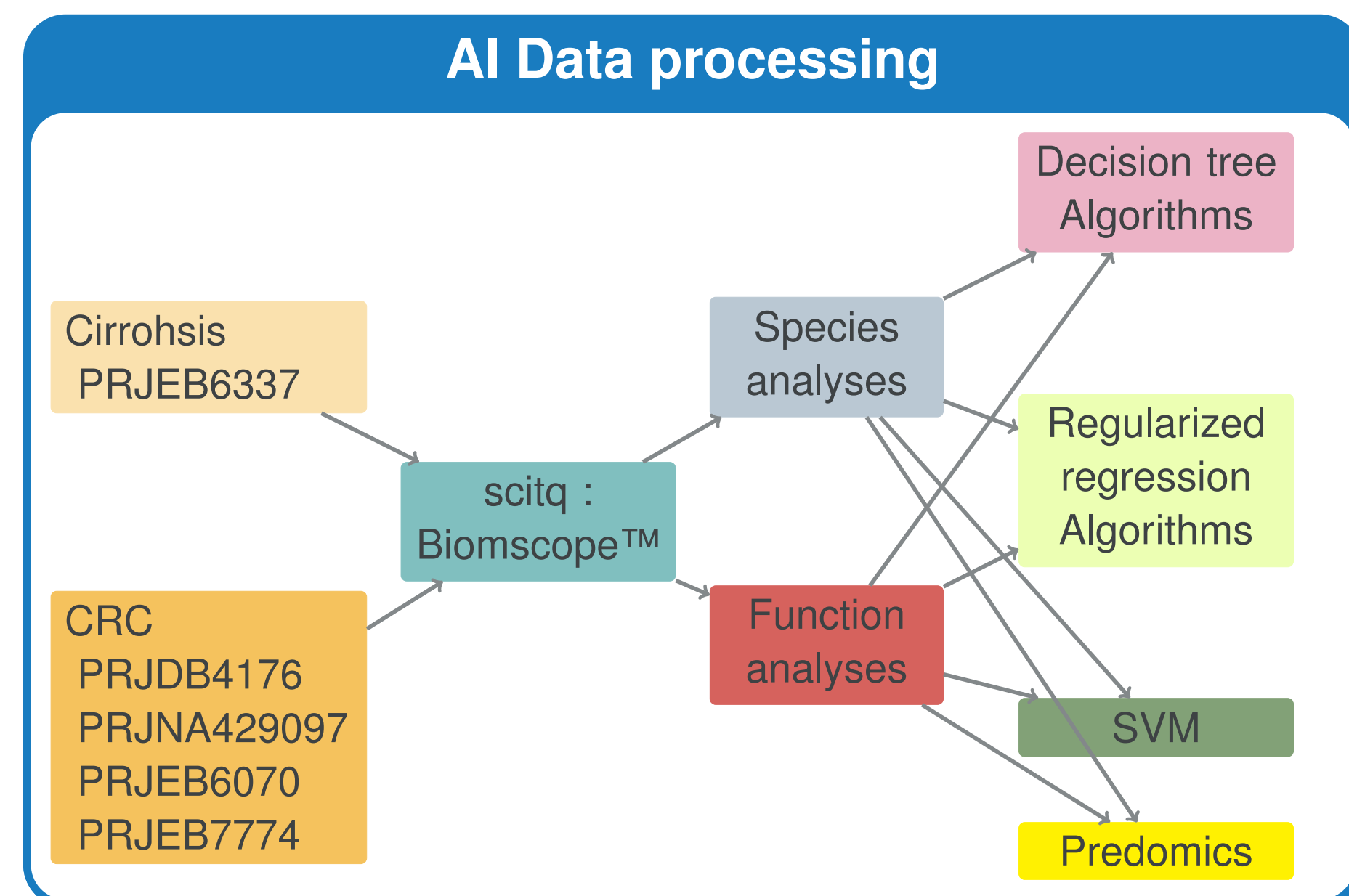
¹ GMT Science, Rouen

² IRD, Sorbonne Université, UMMISCO, Bondy

³ Sorbonne Université, INSERM, NutriOmics, AP-HP, Hôpital Pitié-Salpêtrière

Liver cirrhosis and colorectal cancer (CRC) are two pathologies associated with strong modifications in species composition of the human gut microbiome. Differential (Species) Abundance Analysis (DAA) is typically used to characterize these changes. AI methods are then applied to uncover specific signatures of pathologies. However, several changes in the microbiome composition cannot be detected by this approach (notably within a given species, strains may display different behaviors). We have re-analyzed different metagenomic studies using state-of-the-art DAA and machine learning classification on both taxonomic and functional quantification data. We compared both approaches in their ability to uncover meaningful signature for microbiome-based diagnosis. Two distinct meta-cohorts were used in liver cirrhosis and colorectal cancer. Leveraging gene-level analysis of sequencing data, we employed KEGG orthology (KO) data to quantify these functions using the Integrated Gene Catalog 2 (IGC2). IGC2 was also employed for DAA using its species annotation.

METHOD

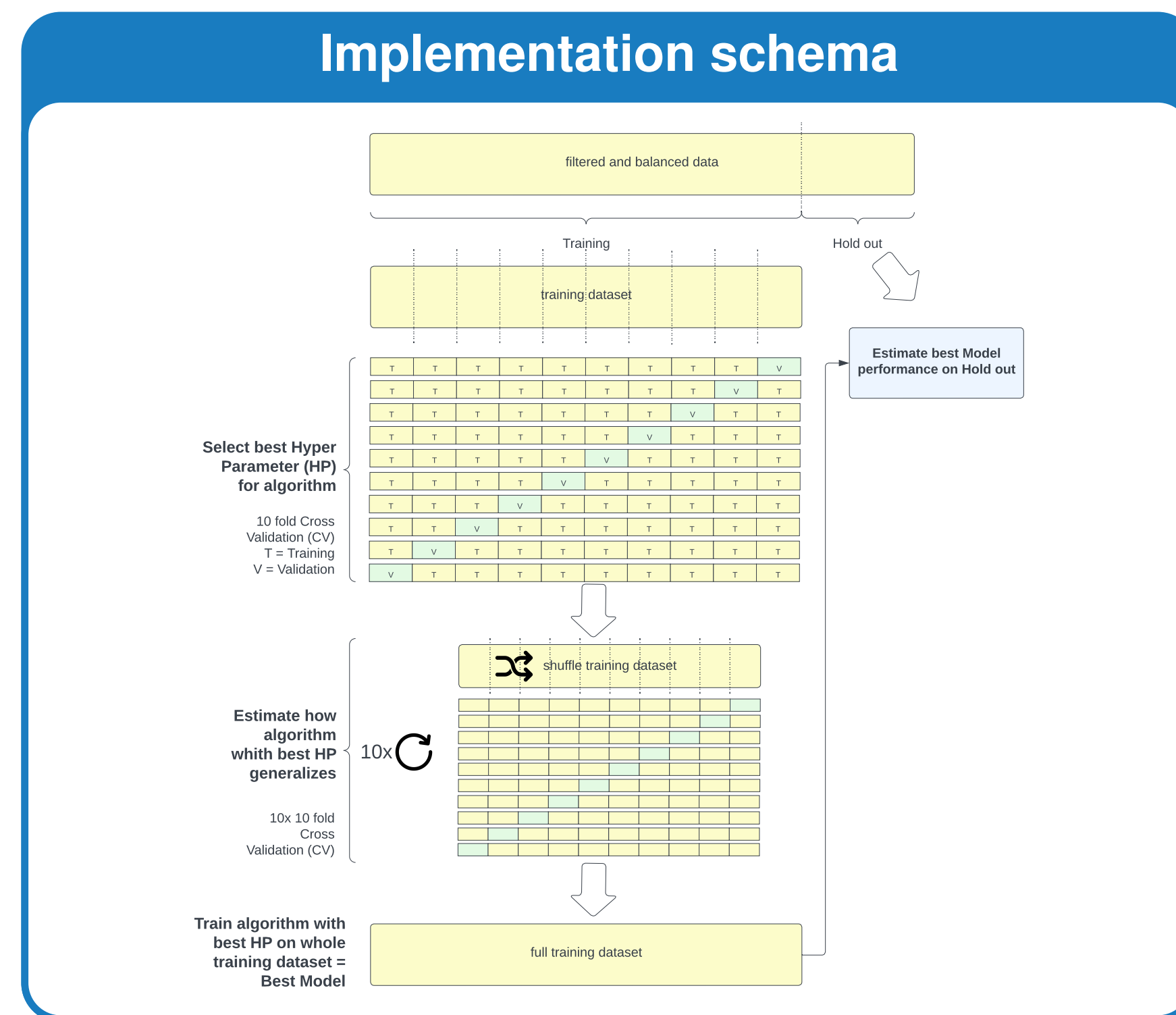


scitq (opensource orchestration): <https://github.com/gmtscience/scitq>
 Biomscope™: proprietary analysis pipeline, GMT Science, MD IVD, CE marked
 Predomics (AI algorithm): Prifti 2020 (doi: 10.1093/gigascience/giaa010)

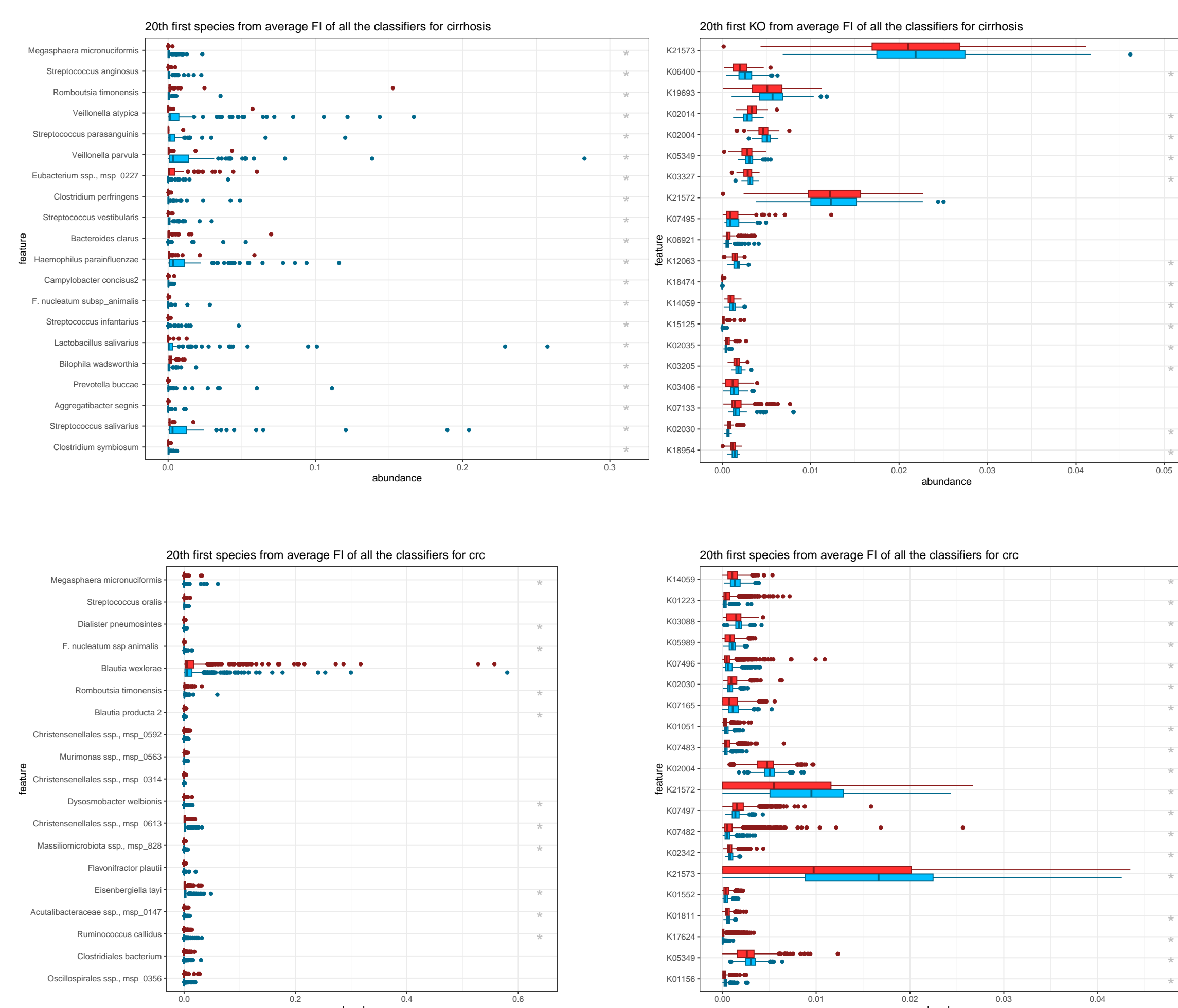
MACHINE LEARNING ALGORITHMS

Class	Algorithm	Description
Decision tree	Random Forest XGBoost	Combine multiple decision trees to make predictions. - Automatic feature selection - Not easy to interpret and very complex.
Regularized regression	LASSO Elastic Net	Like classic regression but some penalties terms of L1 and/or L2 norms are added to the loss function - Reduce the number of variables used by the model. - More interpretable than decision tree-based algorithms
SVM	SVM	SVMs do not provide direct interpretability and less interpretable than decision tree-based algorithms. - Ability to handle non-linear classification problems.
Predomics	Terbeam - bin - ter - bininter - terinter	Focus on using simple and explainable machine learning models

ML IMPLEMENTATION PROTOCOL



SIGNATURES FOR CIRRHOSIS (TOP) & CRC (BOTTOM)



20th most important features according to their average importance across all the classifiers.

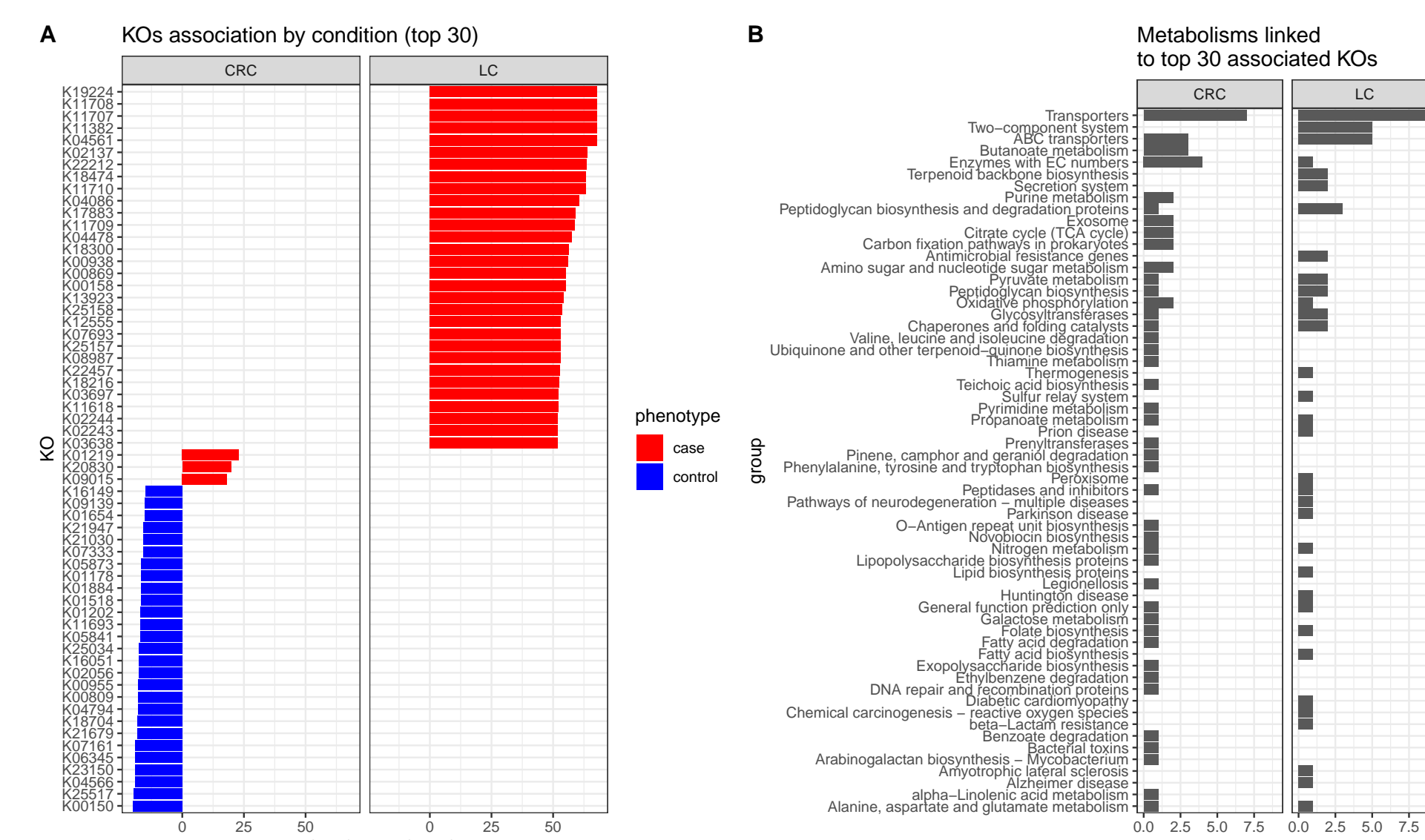
The * indicates those for which there is a significant difference in mean abundance using the non-parametric Wilcoxon test, with Benjamini Hochberg multiple test correction and a significance level of 0.05.

Among the 20th most important KEGG Orthologies (KOs), transporter metabolism is the most represented in both meta-cohorts (n = 7 KOs in LC and n = 5 in CRC). Replication and repair metabolism is also largely represented in the CRC meta-cohort (n = 5).

CONCLUSION

Pathology	Species	Functions
Cirrhosis	- excellent overall performance - confirm the trend of oral species as markers (<i>Veillonella</i> , <i>Streptococcus</i>) - for regulatory consideration, too complex algorithms should be put aside	- superseded by species - except for precision, which promotes the idea of a double parameter test - functional trend is an enrichment of functions related to transporters and peptidoglycan biosynthesis and degradation proteins metabolisms
CRC	- performance largely below that of cirrhosis, notably for precision - sensitivity reach acceptable level (0.88) - signature include known involved species such as <i>F. nucleatum</i> or opportunist species (<i>D. pneumosintes</i>)	- performance was overall unsatisfactory notably for sensitivity (recall) - still exceed precision performance of species signatures - calls for enhancement of method like more initial variable filtering or better stratification of patients or pathology. - functional trend is an enrichment of functions related to transporters and butanoate metabolisms

FUNCTIONAL BIOMARKERS DISCOVERY



A. 30th most associated KOs by enlarged meta-cohort (PRJNA510445, PRJNA510445, PRJEB12449, PRJEB27928, PRJNA758208, PRJNA429097, PRJNA763023, PRJNA763023, for CRC, in addition to cohorts cited in "Method"). q-values and coefficients calculated by MaAsLin2. B. Metabolisms linked to top 30 positively associated KOs to cirrhosis and CRC

PERFORMANCE OF ML CLASSIFIERS ON INDEPENDENT DATA

Data	Type	Algorithm	AUC	Recall	Precision
Cirrhosis	Species	RF	0.98	0.92	0.89
		SVM	0.94	0.80	0.87
		LASSO	0.95	0.92	0.82
		ENET	0.98	0.92	0.87
		XGB	0.96	0.92	0.92
	Functions	Predomics	0.94	0.92	0.86
		RF	0.95	0.81	0.94
		SVM	0.50	0.24	0.71
		LASSO	0.51	0.43	0.47
		ENET	0.50	0.43	0.47
CRC	Species	XGB	0.92	0.81	0.85
		Predomics	0.91	0.85	0.81
		RF	0.68	0.87	0.51
		SVM	0.63	0.52	0.48
		LASSO	0.68	0.86	0.51
	Functions	ENET	0.70	0.79	0.52
		XGB	0.72	0.84	0.53
		RF	0.57	0.10	0.57
		SVM	0.74	0.33	0.54
		LASSO	0.75	0.26	0.57
ENET	0.78	0.10	0.50		
XGB	0.75	0.26	0.67		