Diagnosing IBD from the fecal microbiome

Eliseo Papa, Michael Docktor, Christopher Smillie, Sarah Weber, Sarah P. Preheim, Dirk Gevers, Georgia Giannoukos, Dawn Ciulla, Diana Tabbaa, Jay Ingram, David B Schauer, Doyle V Ward, Joshua R Korzenik, Ramnik J Xavier, **Athos Bousvaros, Eric J Alm**

Eliseo Papa, Ph.D
Harvard–MIT Health Sciences & Technology
elipapa@alum.mit.edu
@elipapa

Why pediatric IBD?

Inflammation of autoimmune origin which presents with confusing symptoms

To date, there is no known causative agent. Rather the etiology of IBD is believed to be multifactorial: environmental, genetic and microbial.

Diagnosis is challenging in children, which leads to a **systematic diagnostic delay**

Delay can have significant negative impact on growth and quality of life

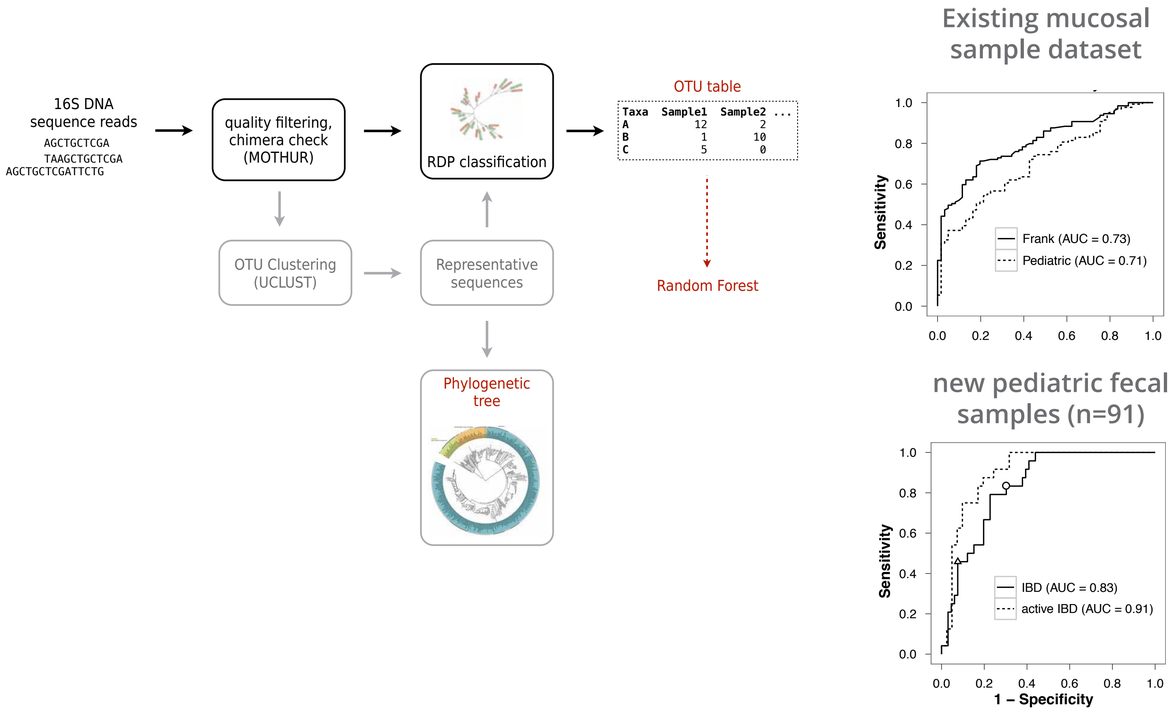
SLiME: Synthetic Learning in Microbial Ecology

Microbial differences exist between healthy and IBD patients

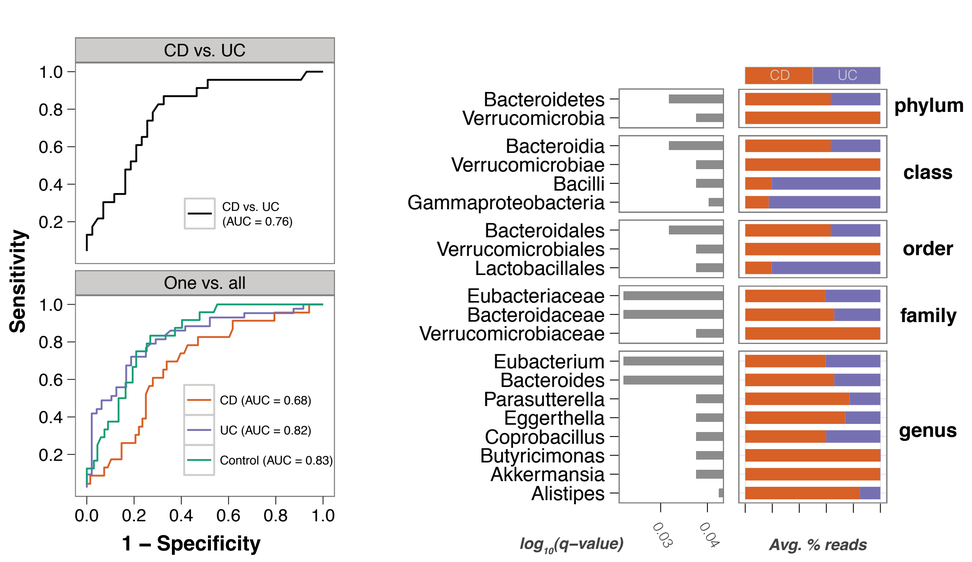
Machine learning algorithms use example data to learn and discover structure (a combination of **features**) in datasets, in order to classify samples into distinct **labels**

Can machine learning applied to microbiota data be used as a diagnostic test?

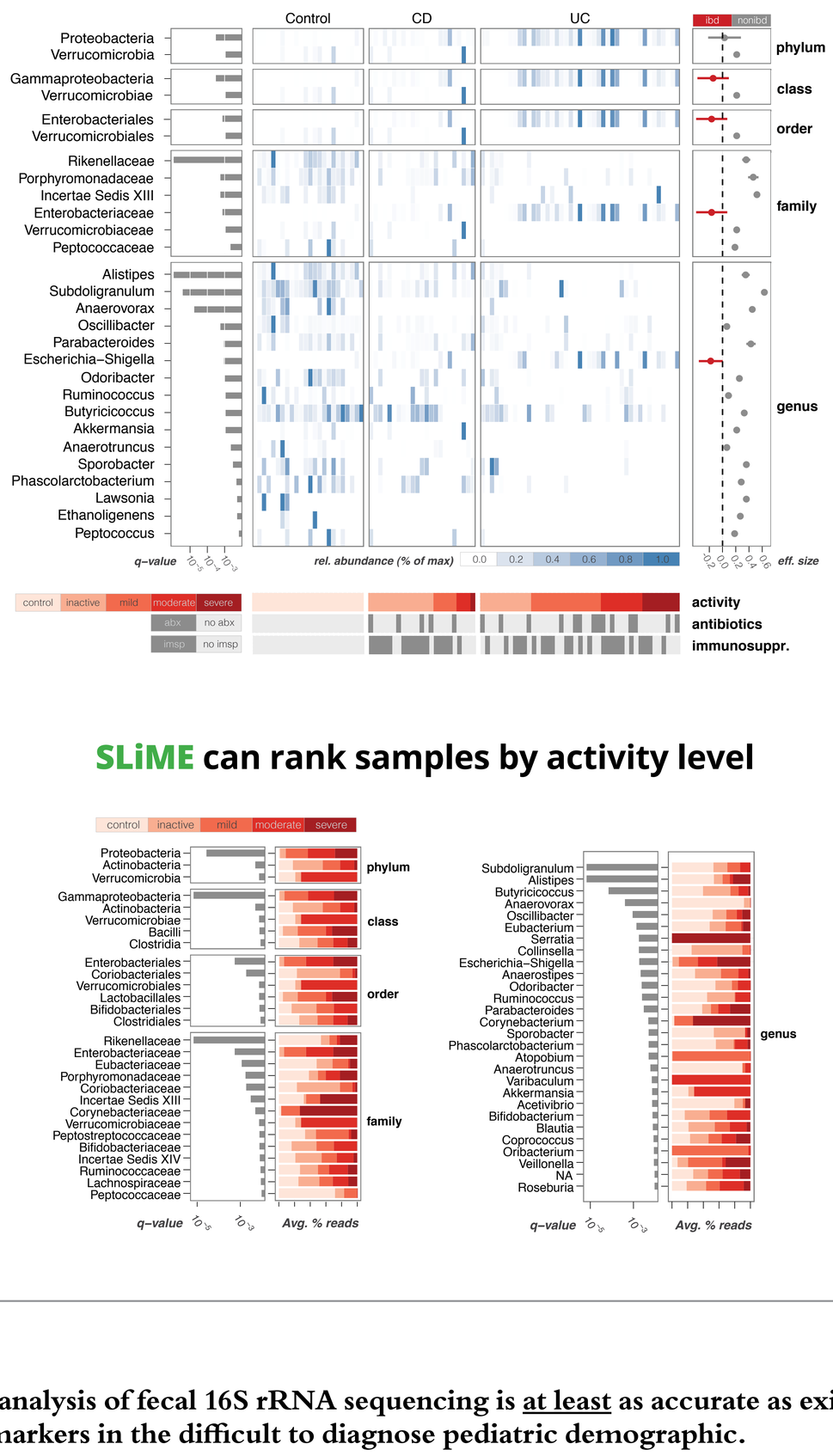
SLiME distinguishes IBD and non-IBD patients



SLiME classifies CD and UC patients



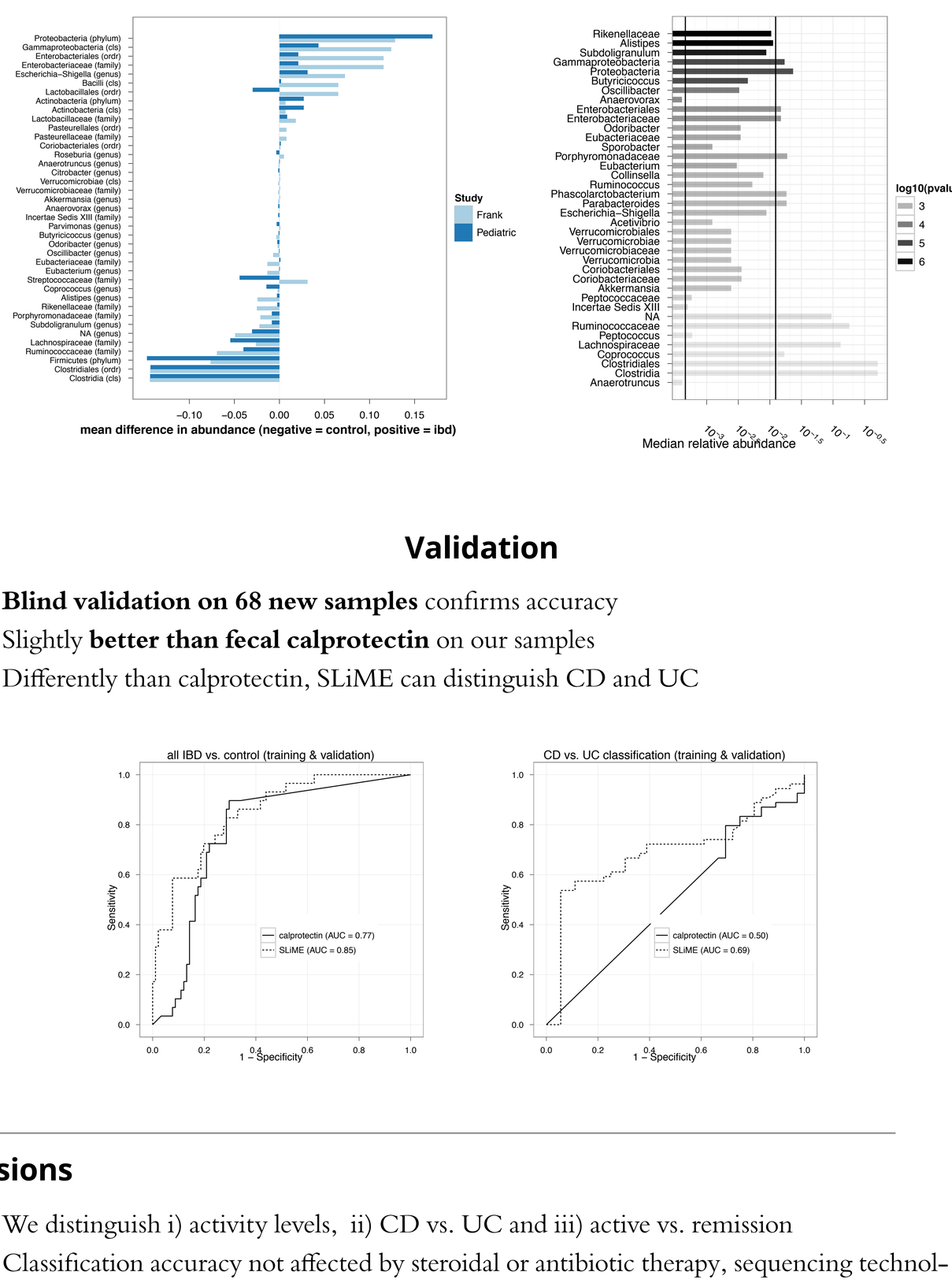
SLiME identifies taxa correlated with disease status



Mucosal samples vs. fecal samples

Taxa abundances in mucosal samples and fecal samples correlate well, however different taxa were important in the two sets as features determining the outcome of the classifier.

Deeper sequencing - by increasing the number of available features - improves the accuracy of the classifier.



SLiME can rank samples by activity level

Validation

Blind validation on 68 new samples confirms accuracy

Slightly **better than fecal calprotectin** on our samples

Differently than calprotectin, SLiME can distinguish CD and UC

Conclusions

SLiME analysis of fecal 16S rRNA sequencing is <u>at least</u> as accurate as existing fecal biomarkers in the difficult to diagnose pediatric demographic.

We demonstrate how powerful algorithms from the field of machine learning can be used to analyze complex microbial community data

We identify specific bacterial taxa associated with disease activity

We distinguish i) activity levels, ii) CD vs. UC and iii) active vs. remission

Classification accuracy not affected by steroidal or antibiotic therapy, sequencing technology or individual-to-individual variation

Clinical feasibility in the primary care setting will depend on cost of sequencing