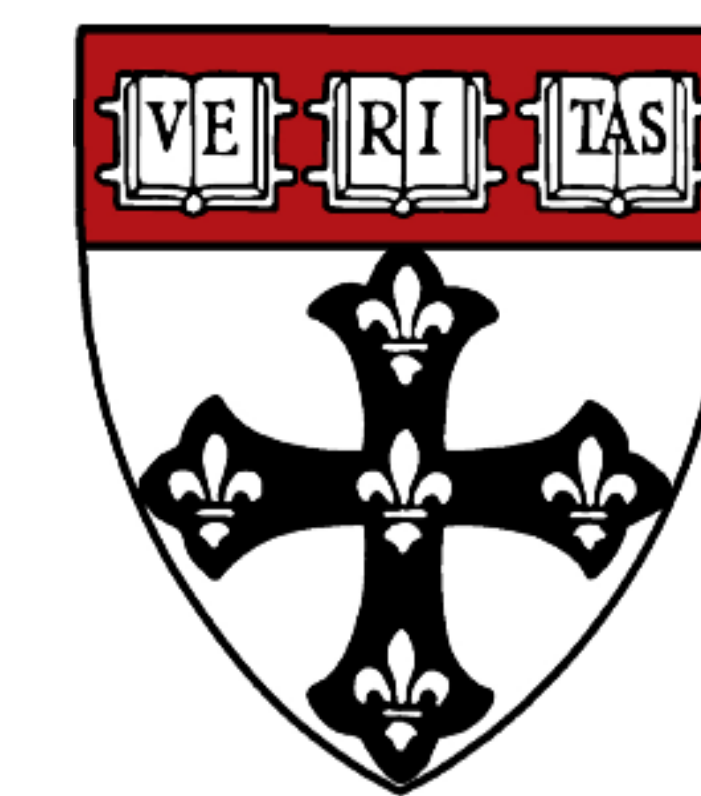




MaAsLin 3: Refining and extending generalized multivariable linear models for meta-omic association discovery



William A. Nickols^{1,2}, Jacob T. Nearing^{1,2,3}, Kelsey N. Thompson^{1,2,3}, Curtis Huttenhower^{1,2,3}

¹Department of Biostatistics, Harvard T.H. Chan School of Public Health

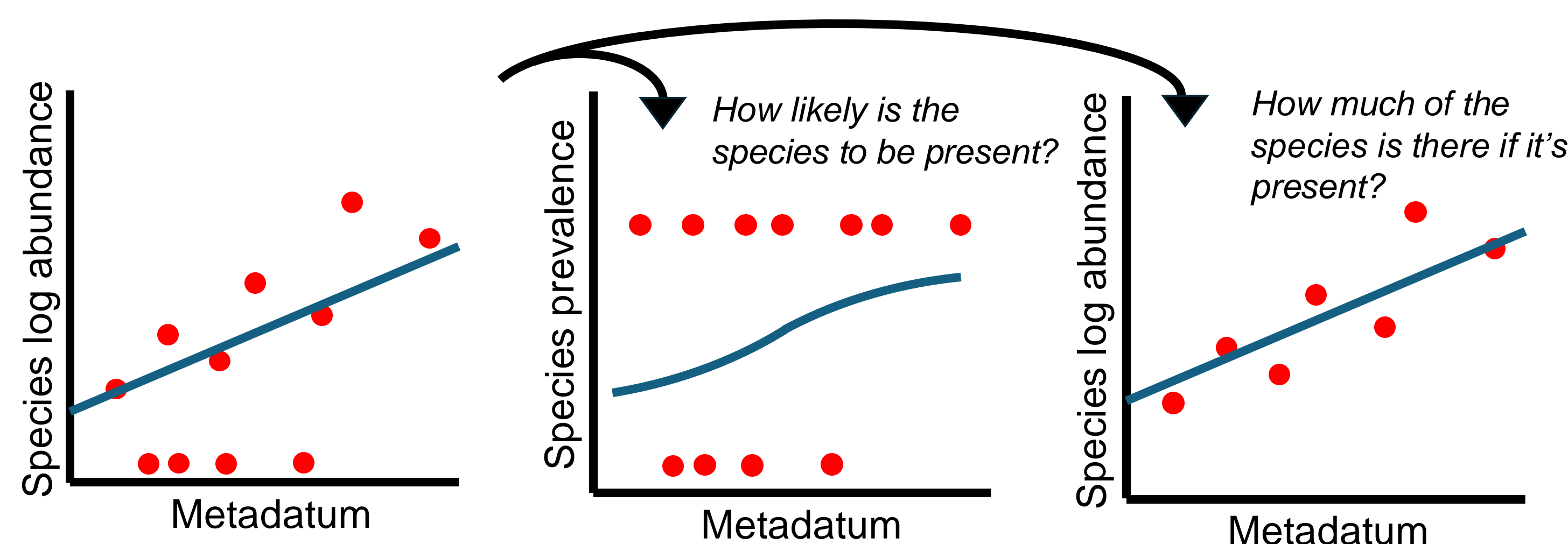
²Harvard Chan Microbiome in Public Health Center, Harvard T. H. Chan School of Public Health

³Infectious Disease and Microbiome Program, Broad Institute of MIT and Harvard



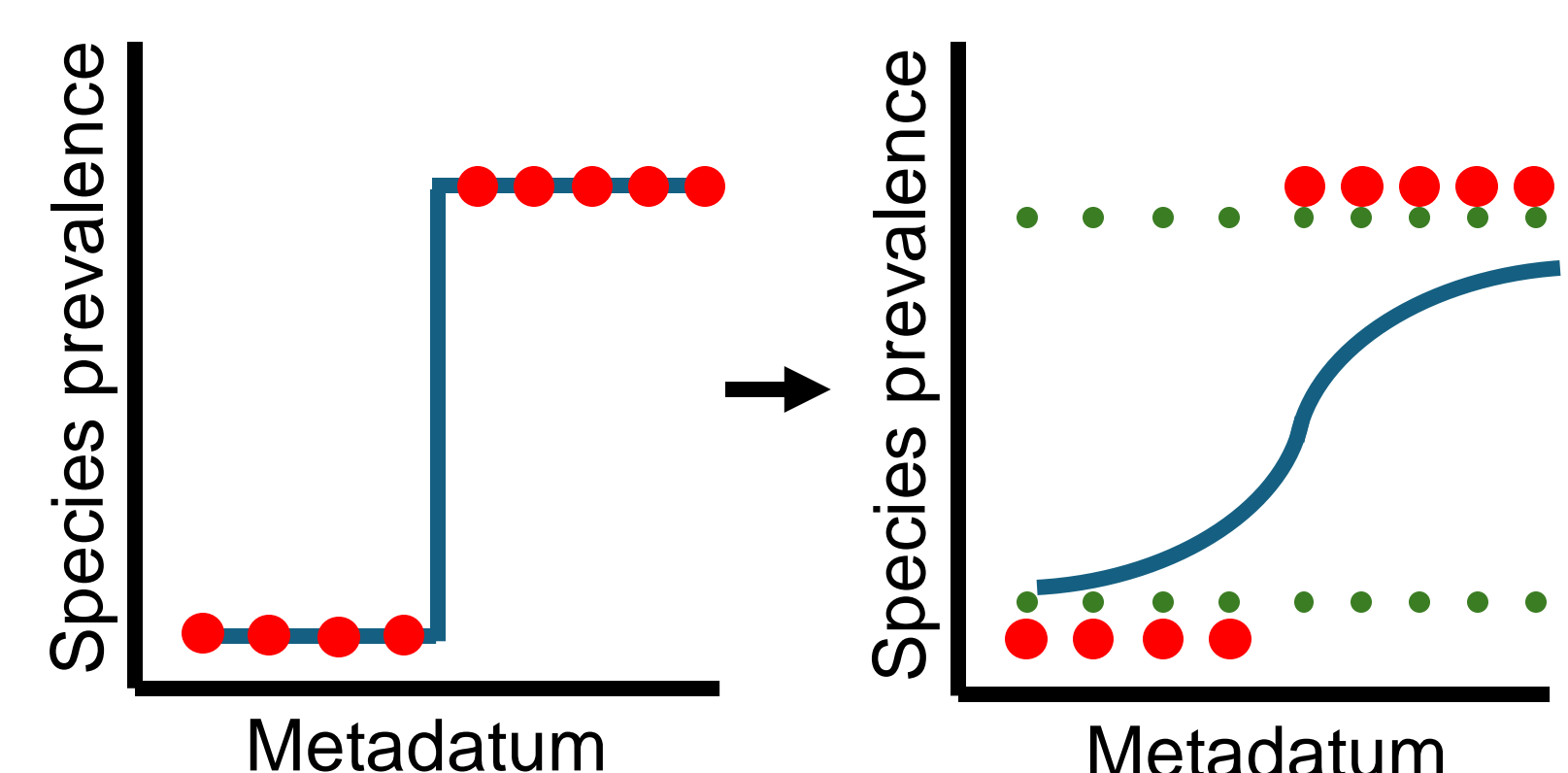
Microbiome analysis often involves differential abundance testing, determining how taxa abundances change with respect to a community phenotype. This testing is complicated by the fact that microbiome data are compositional, sparse, right-skewed, and high-dimensional. To address these challenges, MaAsLin 3 (Microbiome Multivariable Associations with Linear Models) simultaneously identifies both prevalence and abundance associations in a biologically motivated and statistically principled manner. Additionally, MaAsLin 3 enables more robust inference for abundance associations by accounting for compositionality with reference spike-ins or a median comparison procedure. Across a variety of simulations, MaAsLin 3 is more robust to the statistical properties of microbiome data than current state-of-the-art differential abundance methods. When applied to a large dataset of stool samples from an inflammatory bowel disease cohort, MaAsLin 3 indicates that the vast majority of so-called abundance associations are actually prevalence associations.

Improved association testing

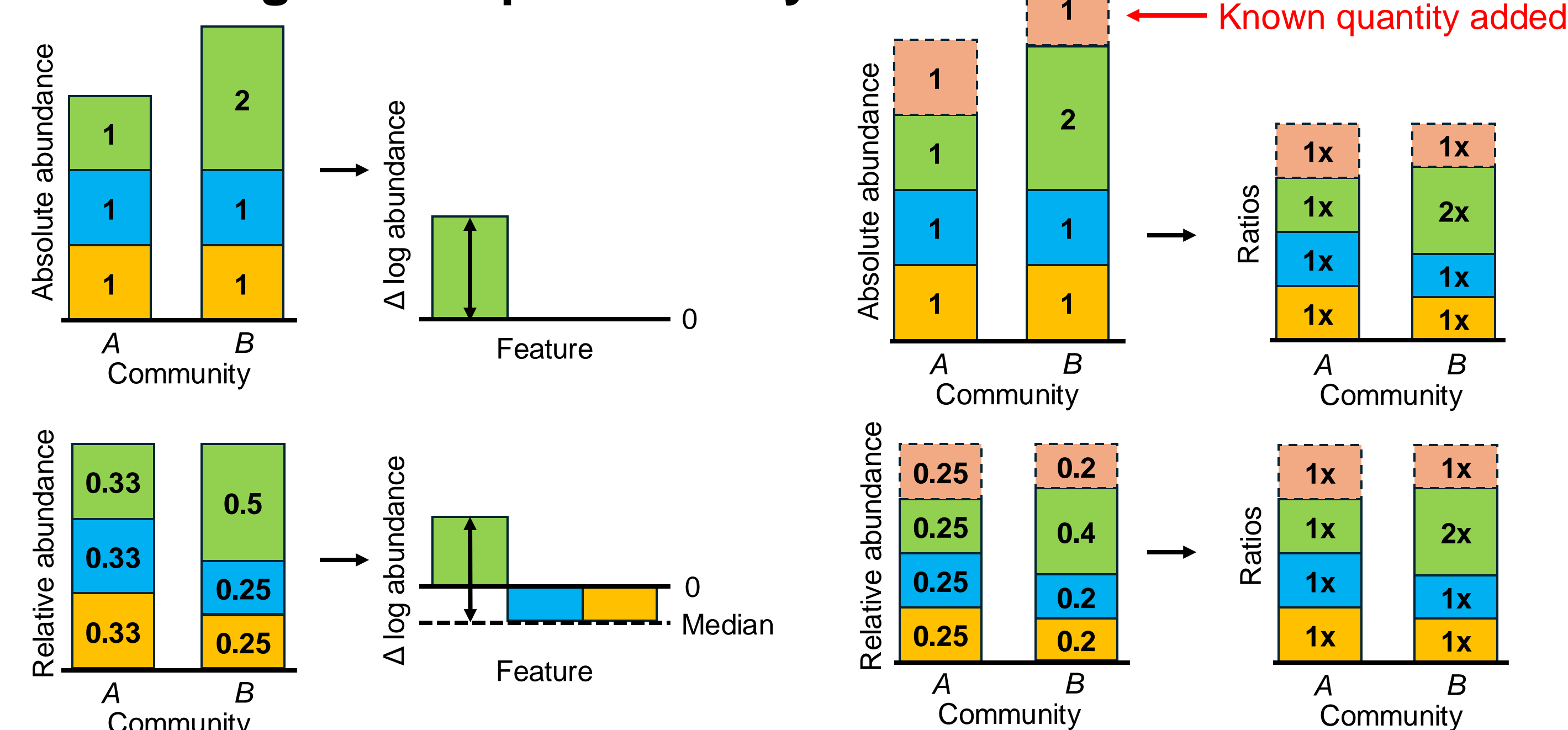


MaAsLin 3 separates prevalence testing (determining how likely a taxon is to be present) from abundance testing (determining how much of a taxon is present if it is present). Multivariable logistic and linear models allow users to control for a wide range of parameters.

Logistic regression runs the risk that a taxon's presence can be linearly separable (left). MaAsLin 3 avoids this by augmenting the data with a 0 and a 1 pseudocount at each data point's covariate values and performing weighted regression (right).



Accounting for compositionality



Compare fold changes against the median

- No experimental modification
- Gives absolute changes when <50% features change

Compute ratios to spike-in

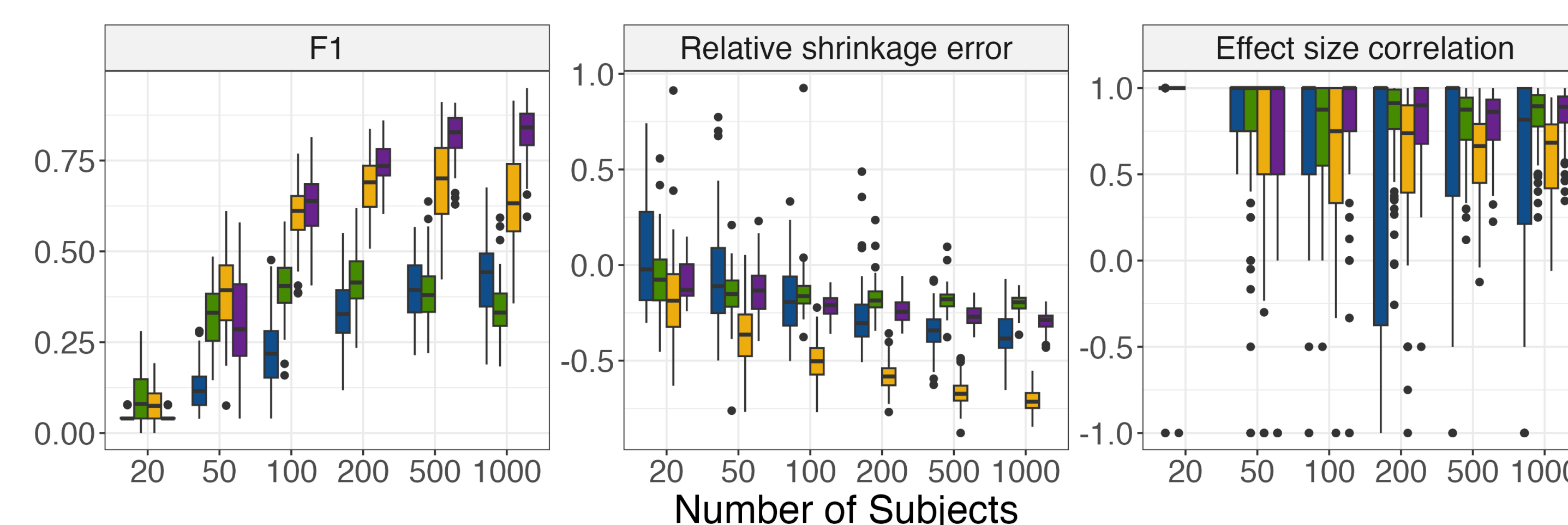
- No assumptions about few species changing
- Requires experimental protocol modification

New analysis options

| Feature | Explanation | Use case |
|-----------------------------------|---|--|
| General mixed model specification | Any valid lme4 formula can be specified | An interaction term can evaluate the difference in a treatment's effect between two populations. |
| Omnibus testing | For a covariate with multiple categories, this is an ANOVA-style test of whether the different levels have different coefficients. | In a study with participants from multiple countries, this could test whether species abundances differ among countries. |
| Level-versus-level differences | For a covariate with ordered levels, this is a test for differences between consecutive levels. | In a study of participants with different stages of colorectal cancer, this could test whether species abundances differ between consecutive stages. |
| Contrast testing | This is a general test for a linear combination of coefficients against a constant (typically used for testing whether two coefficients are different). | In a study with samples from different soil types, this could test whether species abundances differ between each pair of soil types. |
| Feature-specific covariates | For a covariate that differs per-feature, this includes the relevant covariate in each feature's model. | In a metatranscriptomic study, this controls for a gene's DNA abundance when regressing the RNA abundance. |

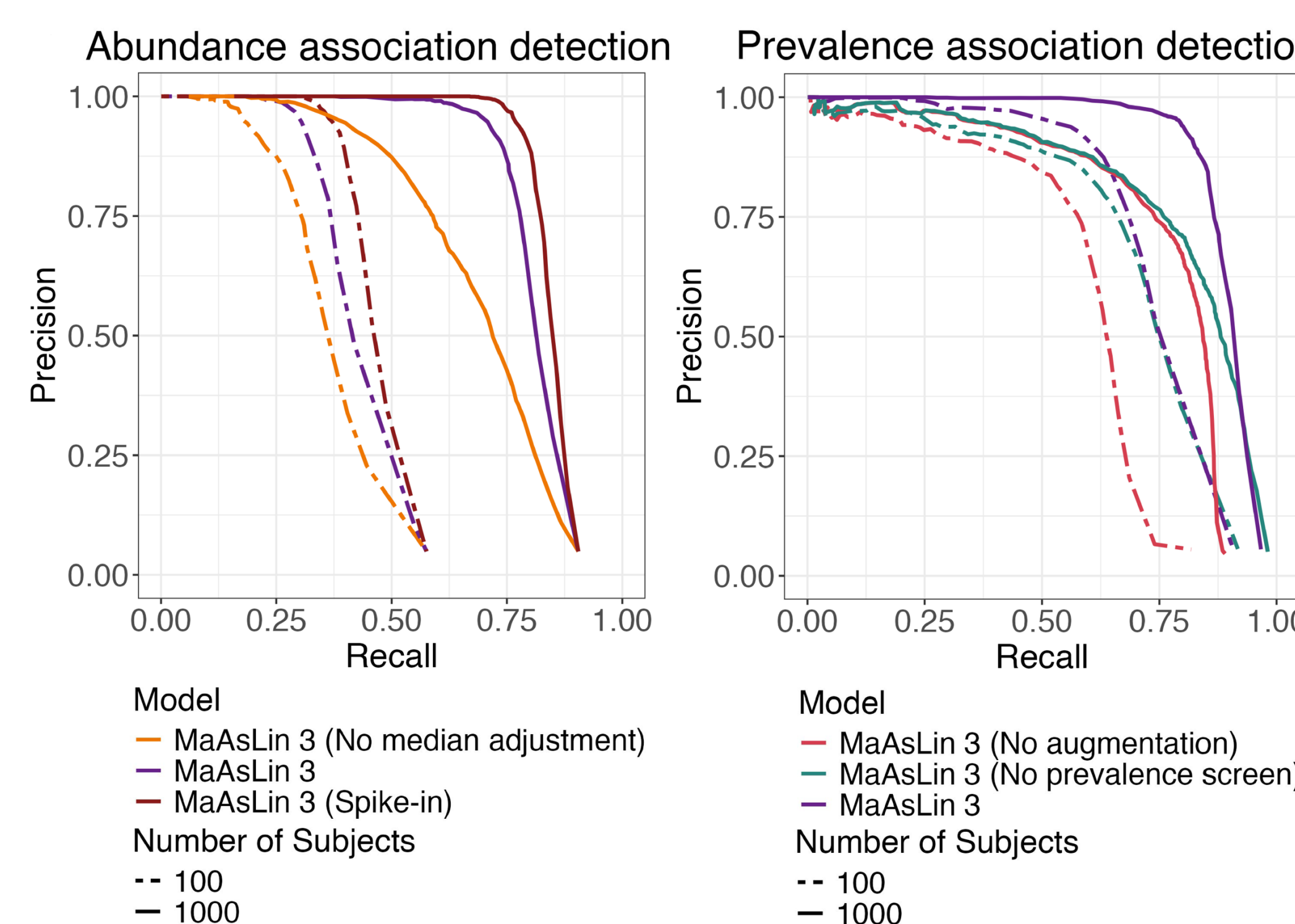
Synthetic evaluation

Synthetic datasets were created using SparseDOSSA 2 with 100 microbes, 5 categorical or continuous metadata variables, and included prevalence or abundance effects for 10% of feature-metadata pairs. The included fold changes (abundance) or log-odds changes (prevalence) were chosen uniformly from 2.5 to 5. All differential abundance tools were run on each dataset, and associations were false discovery rate corrected. MaAsLin 3 (with its default data augmentation and median comparison) outperformed other differential abundance methods. F1: harmonic mean of precision and recall. Relative shrinkage error: $(|Fit\ coefficient| - |True\ coefficient|) / |True\ coefficient|$. Effect size correlation: Correlation between fit and true coefficients.



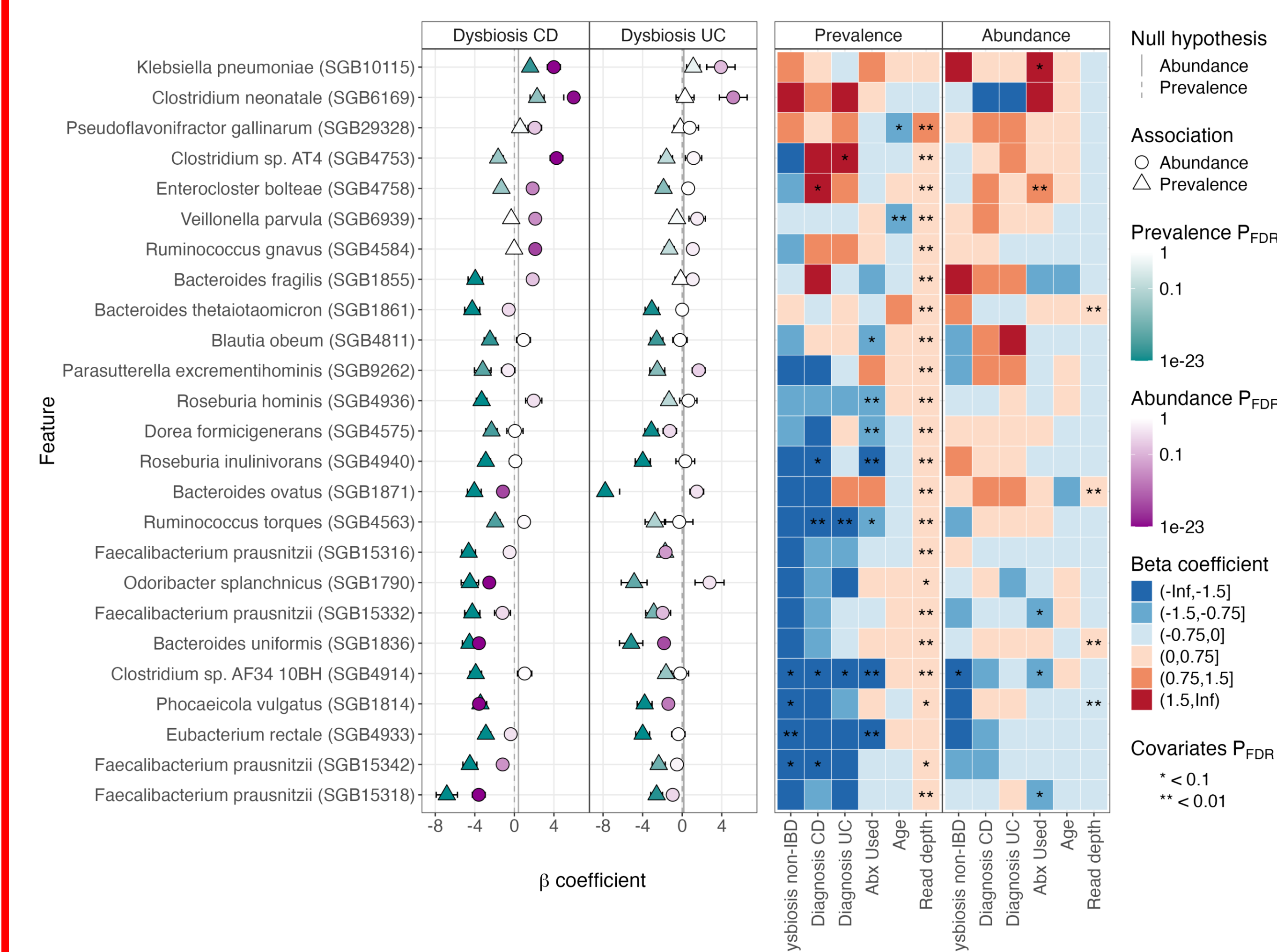
For the abundance associations, at any given recall level, precision was improved with the median comparison or spike-in normalization to handle compositionality.

For the prevalence associations, at any given recall level, precision was improved with the data augmentation scheme and a prevalence screen to avoid abundance effects appearing as prevalence effects.

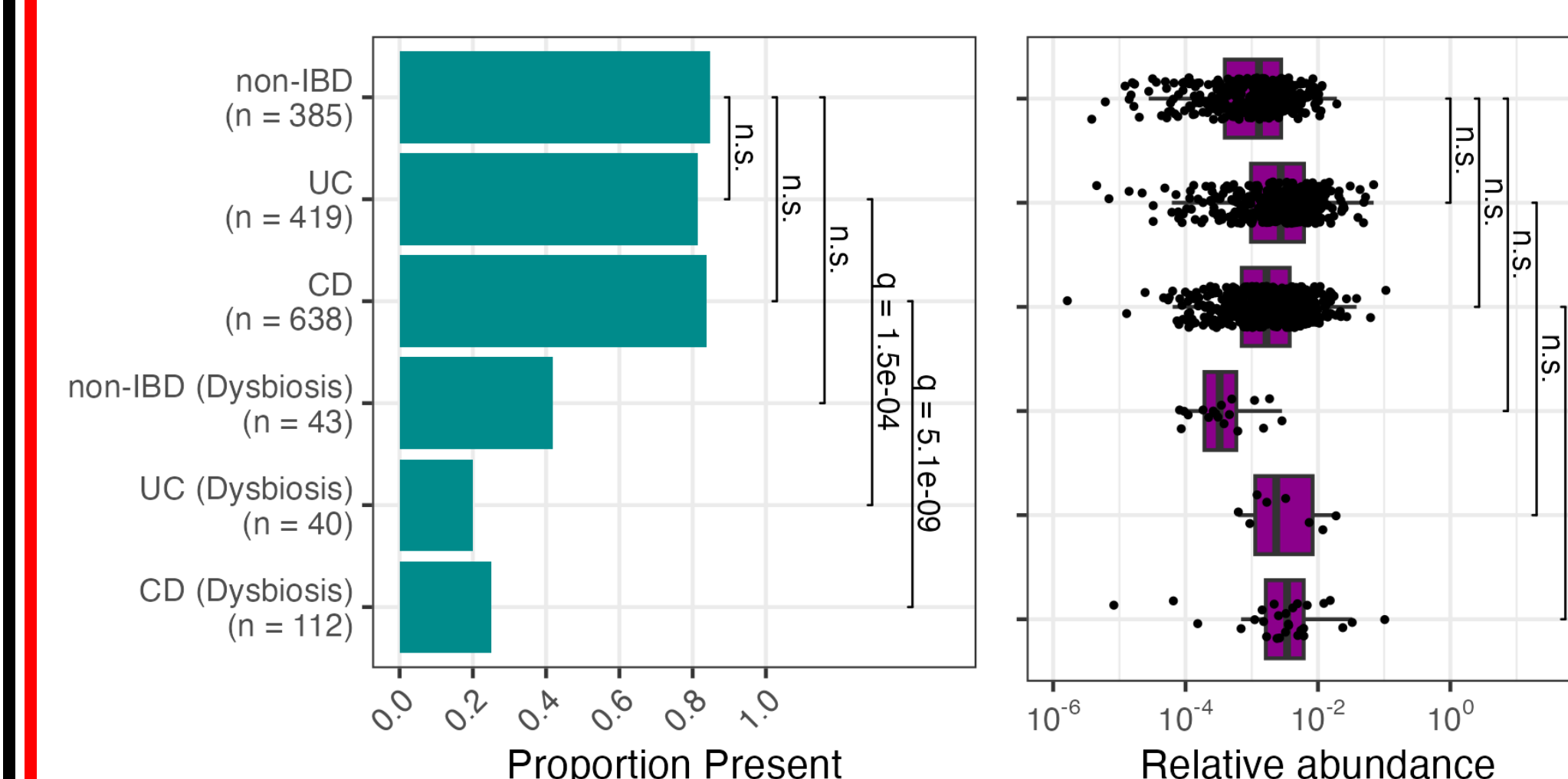


IBD association testing

Using MetaPhlAn 4, taxonomic profiles were constructed from 1637 metagenomic samples in the Inflammatory Bowel Disease Multi-omics Database. We analyzed microbial associations with Crohn's Disease, Ulcerative Colitis, and dysbiosis while controlling for antibiotic use, age, and read depth. Repeated sampling was accounted for with participant-specific random intercepts.



Of the resulting 372 associations, 287 (77%) were prevalence associations, suggesting that many previously reported abundance associations are actually prevalence associations. Consistent with previous reports of decreased diversity in stool samples from patients with inflammatory bowel diseases, 59% of significant abundance associations and 89% of significant prevalence associations had negative coefficients.



An association of interest was the prevalence-only reduction of *Dysosmobacter welbionis* during CD and UC dysbiosis. *D. welbionis* is a recently isolated human gut commensal associated with metabolic disorders in humans and prevention of diet-induced obesity in mice.

Acknowledgements

The work was supported by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health (R24DK110499) to C.H. and the National Institute of Allergy and Infectious Diseases (U19AI110820) to D. Rasko (to C.H.) and the Crohn's and Colitis Foundation Early Career grant (K.N.T.).

<https://huttenhower.sph.harvard.edu>

