Systematic characterization and analysis of the taxonomic drivers of functional shifts in the human microbiome

Ohad Manor and Elhanan Borenstein

Comparative analyses of the human microbiome have identified both taxonomic and functional shifts that are associated with numerous diseases. To date, however, microbiome taxonomy and function have mostly been studied independently and the taxonomic drivers of functional imbalances have not been systematically identified. Here, **we present FishTaco, an analytical and computational framework that integrates taxonomic and functional comparative analyses to accurately quantify taxon-level contributions to disease-associated functional shifts**. Applying FishTaco to several large-scale metagenomic cohorts, we show that shifts in the microbiome's functional capacity can be traced back to specific taxa. Furthermore, the set of taxa driving functional shifts and their contribution levels vary markedly between functions. We additionally find that similar functional imbalances in different diseases are driven by both disease-specific and shared taxa. Such integrated analysis of microbiome ecological and functional dynamics can inform future microbiome-based therapy, pinpointing putative intervention targets for manipulating the microbiome's functional capacity.

Cell Host Microbe. 2017 Feb 8; 21(2): 254–267.

The method, illustrated in Figure 2 which is not helpful to me:

Step 1. Figure out how much each taxon contributes to a given gene abundance.

There are two ways to do this. Either (1) use reference genomes directly or (2) use their previously-published method ("Reconstructing the Genomic Content of Microbiome Taxa through Shotgun Metagenomic Deconvolution") which is basically a multiple linear regression:

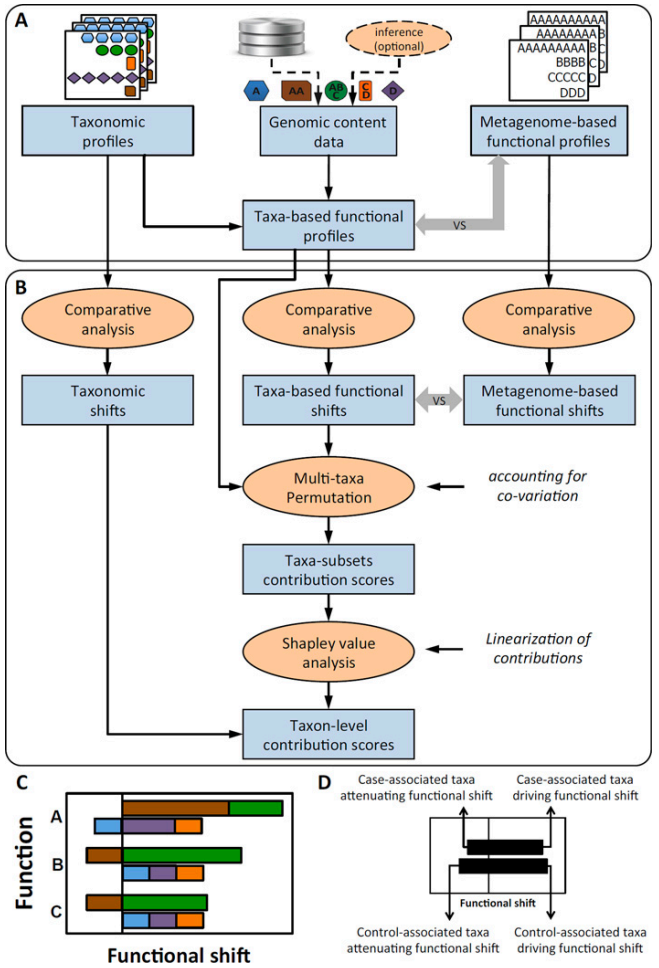
Formally, if a_k denotes the abundance of genome k in the community and e_{kj} denotes the prevalence of an element j in genome k (e.g., in terms of copy number or length in nucleotides), the total abundance of this element in the community can be represented as:

$$E_j = a_1 e_{1j} + a_2 e_{2j} + a_3 e_{3j} + \ldots + a_N e_{Nj}. \quad (1)$$

where you measure Ej (abundance of gene j) and all the a's (abundance of taxa), and estimate the e_{ij}'s (prevalence of gene j in taxon i).

They they solves this set of linear equations with non-negative elastic-net regularized linear regression

The contribution of each taxon to the gene abundance is each component of this function (i.e. it's a*e).

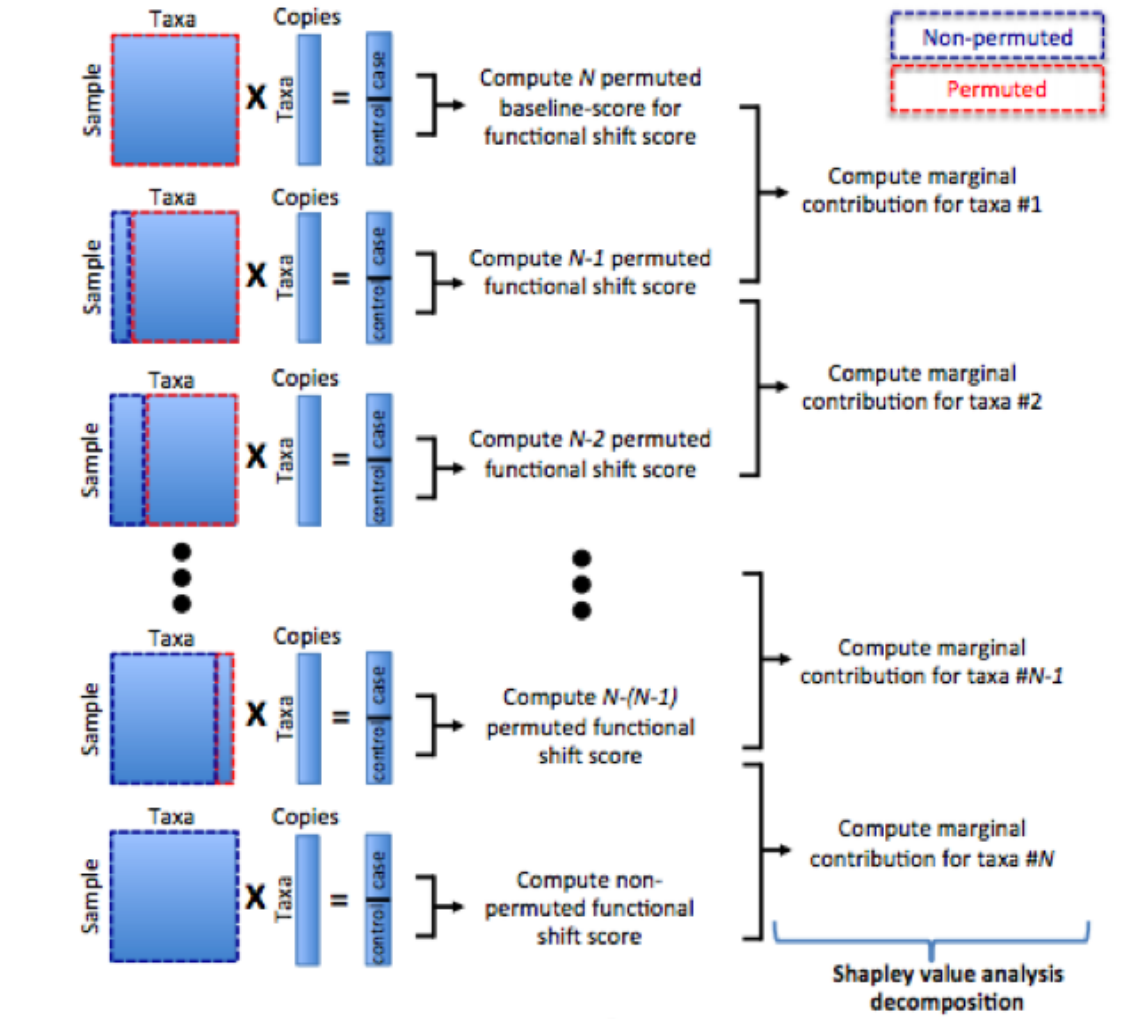


Step 2. Figure out **how much each taxon contributes to the shift in gene abundance** by permuting the taxa.

“compares the functional shifts observed in the original taxa-based functional profiles to the shifts observed when the relative abundances of a set of taxa are randomly permuted across sample”

They they use Shapley value analysis to determine the individual contribution of each taxon within this multi-taxa setting.

Shapley value is from game theory, from Wikipedia: “The setup is as follows: a coalition of players cooperates, and obtains a certain overall gain from that cooperation. Since some players may contribute more to the coalition than others or may possess different bargaining power (for example threatening to destroy the whole surplus), what final distribution of generated surplus among the players should arise in any particular game? Or phrased differently: how important is each player to the overall cooperation, and what payoff can he or she reasonably expect? The Shapley value provides one possible answer to this question.” Shapely value given by:

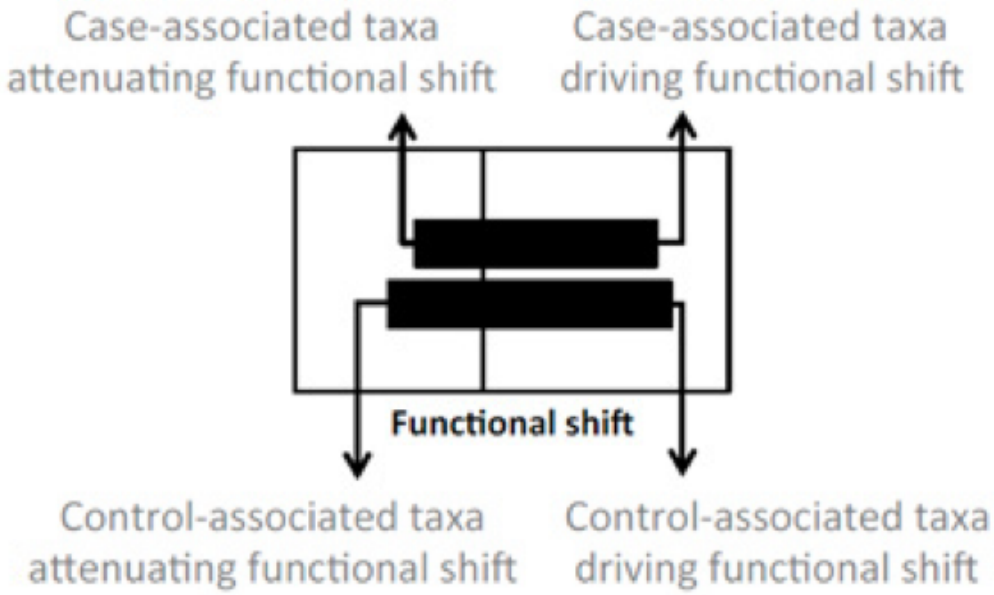


$$\phi_i(v) = \frac{1}{\text{number of players}} \sum_{\text{coalitions excluding } i} \frac{\text{marginal contribution of } i \text{ to coalition}}{\text{number of coalitions excluding } i \text{ of this size}}$$

Step 3. Visualize the contributions. There are four types of contributors (their description is in gray, mine is in black/colors)

Taxa which are **more abundant in cases** *and* which **don't have the gene**

Taxa which are **more abundant in cases** *and* which **have the gene**



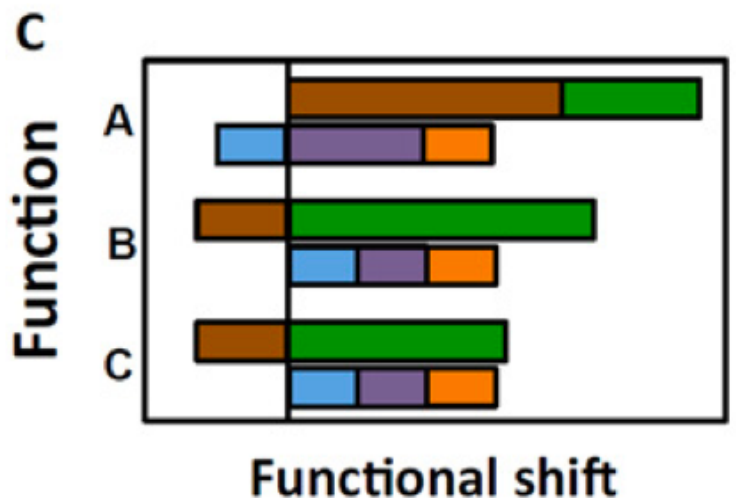
Taxa which are **more abundant in controls** and which **have the gene**

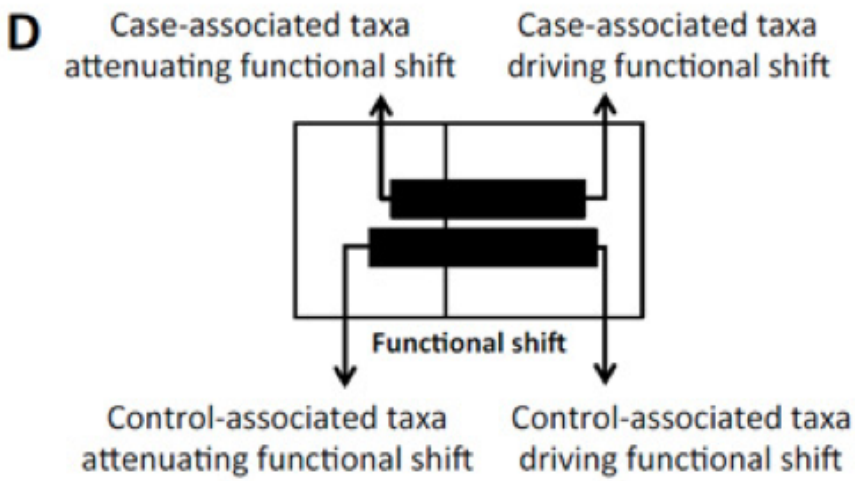
Taxa which are **more abundant in controls** and which **don't have the gene**

Gene is more abundant in controls

Gene is more abundant in cases

At the end of the day, in real data it looks like this (Fig 2C, left)





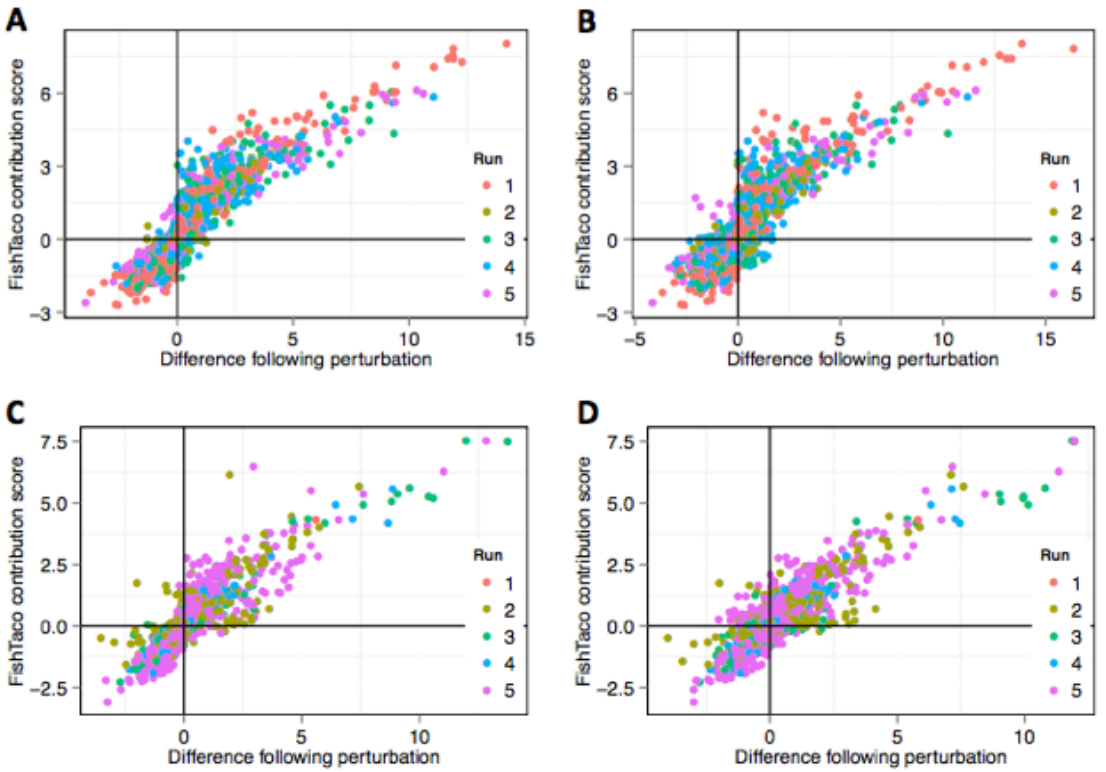
Gene is more abundant in controls **Gene is more abundant in cases**

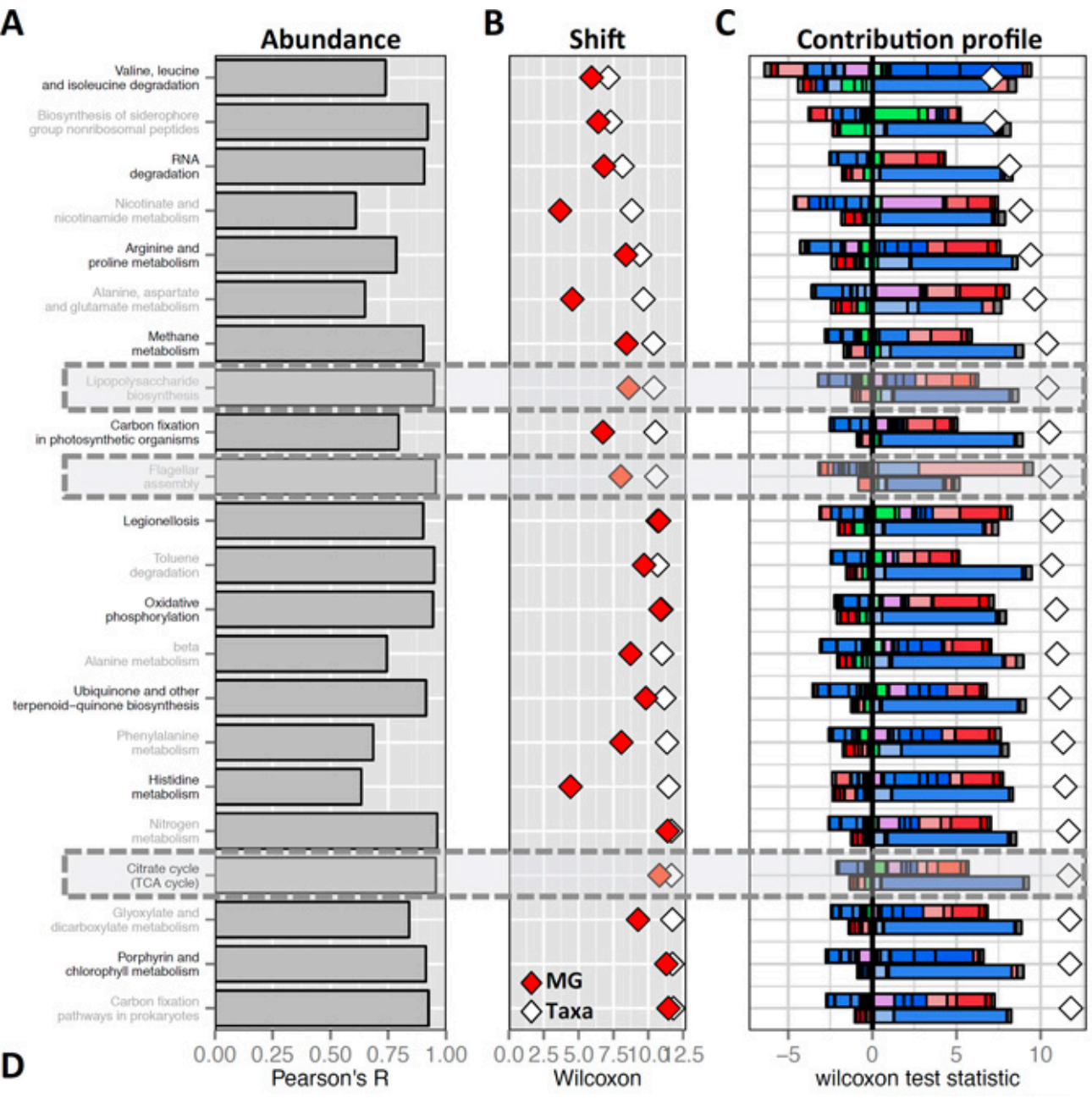
They did some simulations to check their FishTaco score with “real” contributions to shifts.
They did 5 runs of this simulation (the different colors)

A and B: using “simulated communities comprised of synthetic species”
C and D: using reference genomes

x-axis = actual taxon-level impact on functional shift; y-axis = fishtaco score

left column = impact was calculated by eliminating the taxon’s compositional shift
right column = impact was calculated by removing the taxon from the community





Next, test their method on tongue and cheek HMP metagenomics samples (n=106 and 107).

A. Correlation between taxa-based and metagenome-based functional profiles (focusing on 22 functions that were enriched in tongue). Taxa-based was determined by getting taxonomic composition with MetaPhlAn and then estimating functions from IMG reference genomes. Metagenome-based was done using HUMAnN. (median Pearson's correlation across functions R=0.91)

B. Functional shift score calculated from metagenomics (MG, red) and taxa-based functional profiles (Taxa, white). Wilcoxon test statistic was used as the shift score. Spearman correlation of shift scores $\rho=0.65$, $p<10^{-5}$

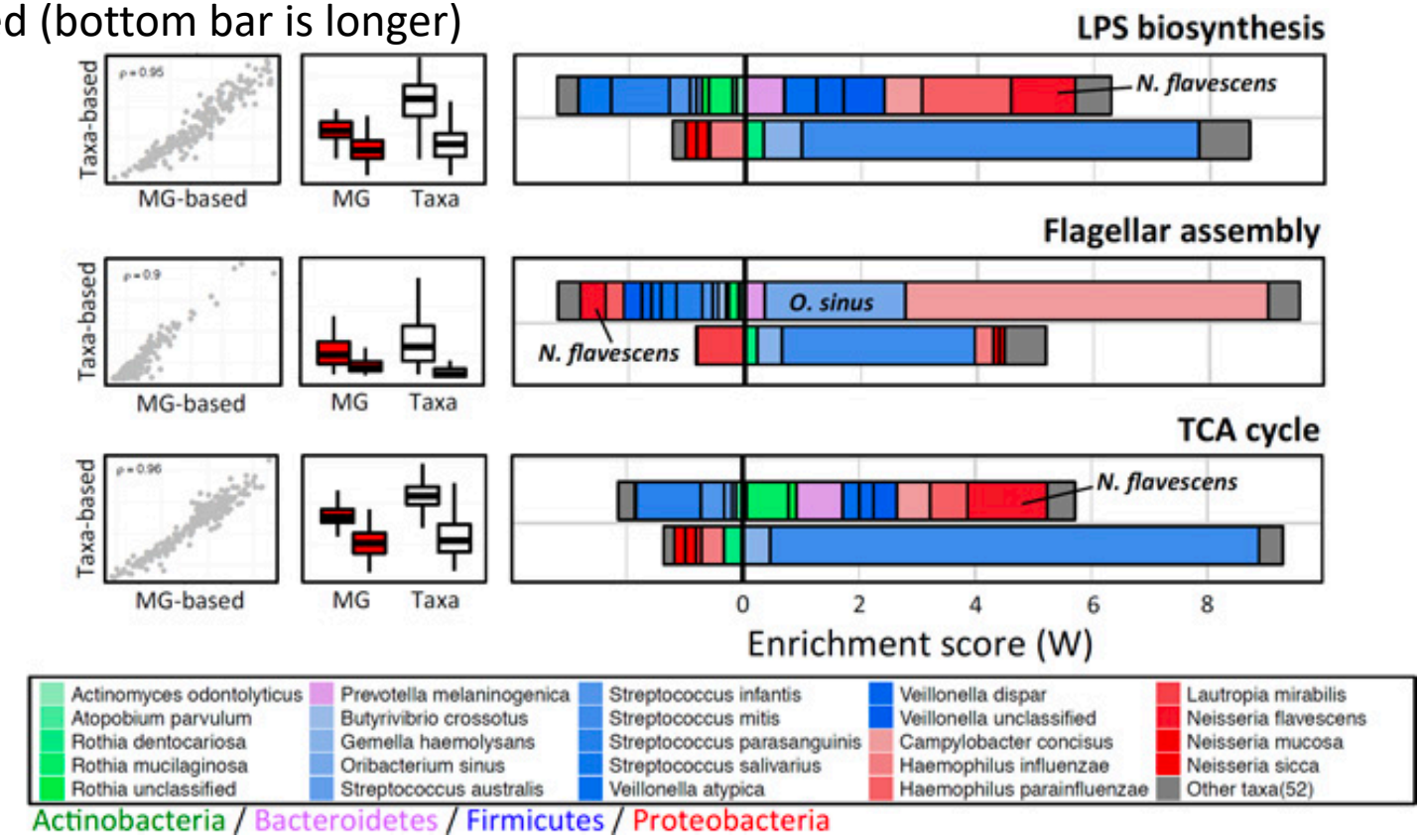
"These findings confirm that our taxa-based functional profiles capture the underlying functional composition of the metagenomes"

C. FishTaco taxon-level shift contribution profile for each tongue-enriched function. White boxes is the same as in panel B. Grey boxes are "pathways of interest" – see next slide

Zooming in on a few functional pathways of interest (grey boxes in above slide).

Check out how *Neisseria flavescens* drives the observed shift in one function (i.e. is on the right side of the plot, contributes to increased abundance of the function) while attenuating the shift in another (i.e. on the left side, contributes to decreased abundance of the function). Also, different species of same phylum have different impacts on functional shifts (look at Proteobacteria, red, and how some of them are on the right side [contribute to increase function] and others are on the left side [contribute to decreased function] even for the same pathway).

More interestingly: with this analysis you can see that some functions are driven by tongue-associated taxa whereas others by cheek-associated. e.g. flagellar assembly driven by tongue-associated (top bar is longer), but TCA cycle driven by cheek-associated (bottom bar is longer)



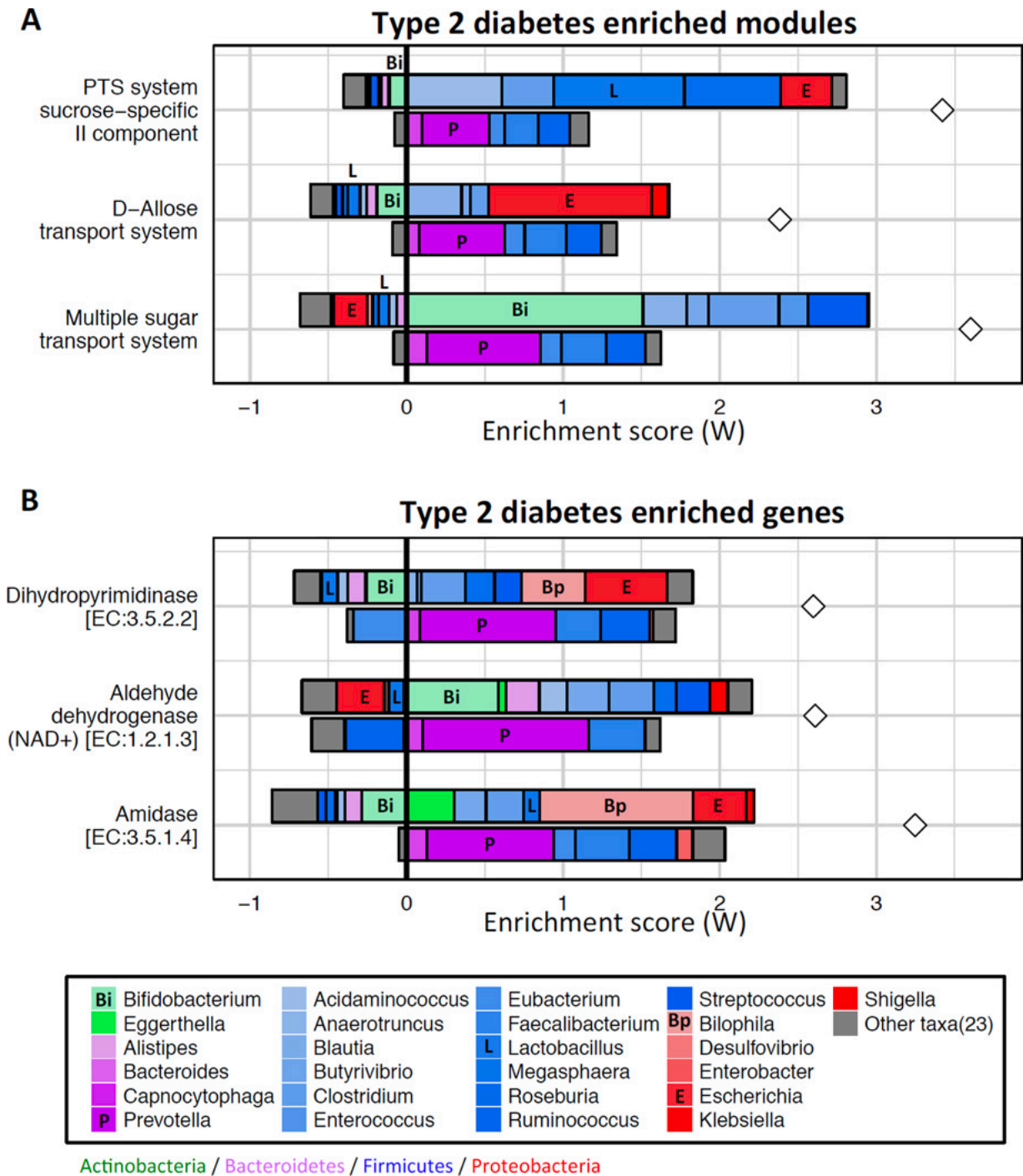
Other parts of the figure:

Scatter plots: each dot is a sample, x-axis = abundance of that pathway based on metagenome and y-axis = based on their taxa-based model. Spearman correlation shown.

Boxplots: left box = tongue, right box = cheek. Red = metagenomics-based function abundance, white = taxa-based.

Next, use fishtaco to look into T2D. Data from Qin et al 2013 cohort, metagenomics data. Used genus-level taxonomic abundances (bc species-level wasn't reported), coupled with IMG reference genomes for each genus to make their taxa-based functional profiles. Also compared these taxa-based profiles with the profiles from the metagenomics data (median pearson's R=0.76).

They found a bunch of results that matched results from other papers, and did some sanity-checks to check that patterns they expect wrt abundance of bugs and associated functions matched what was found in the metagenomics.

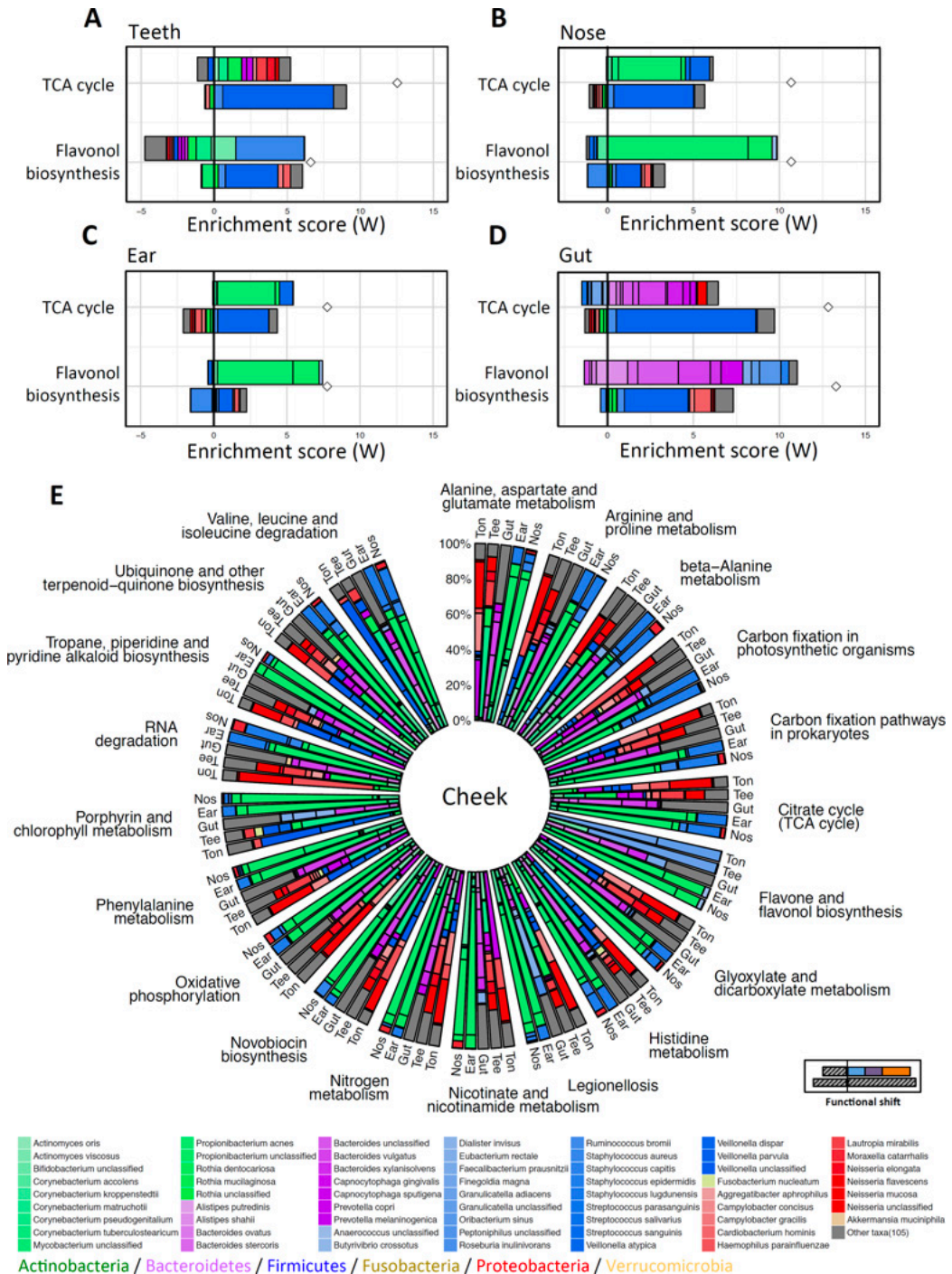


Now they're trying to understand a bit more about the general behavior of these functional shifts: are they always driven by the same taxa, or not?

"We again found that in each of these body sites, different functions had markedly different shift contribution profiles (Figure 5A–D). Importantly, however, not only did different functions show differences in their taxon-level contribution profiles in each site, but the same function often exhibited substantially different contribution profiles across the different body sites (with less variability between the two skin sites; Figure 5A–D)."

A-D: cheek vs. each of these other sites, each panel is a different body site comparison

E: cheek vs each of these other sites, grouped by the enriched function. Just showing the top right part of the fishtaco plot (i.e. taxa that are more abundant in the non-cheek site and that have the gene). Each bar is different = each functional shift is driven by different taxa in the different body sites.

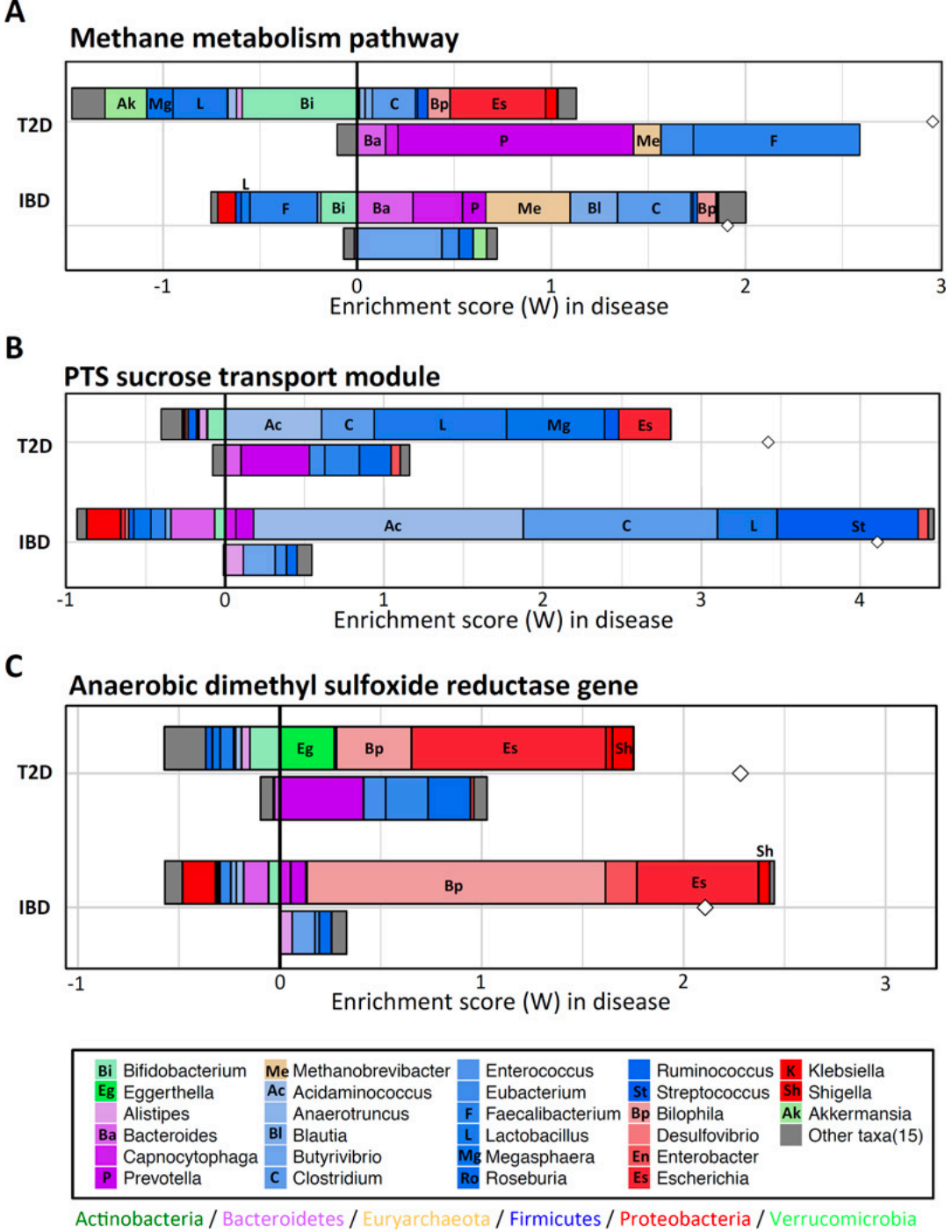


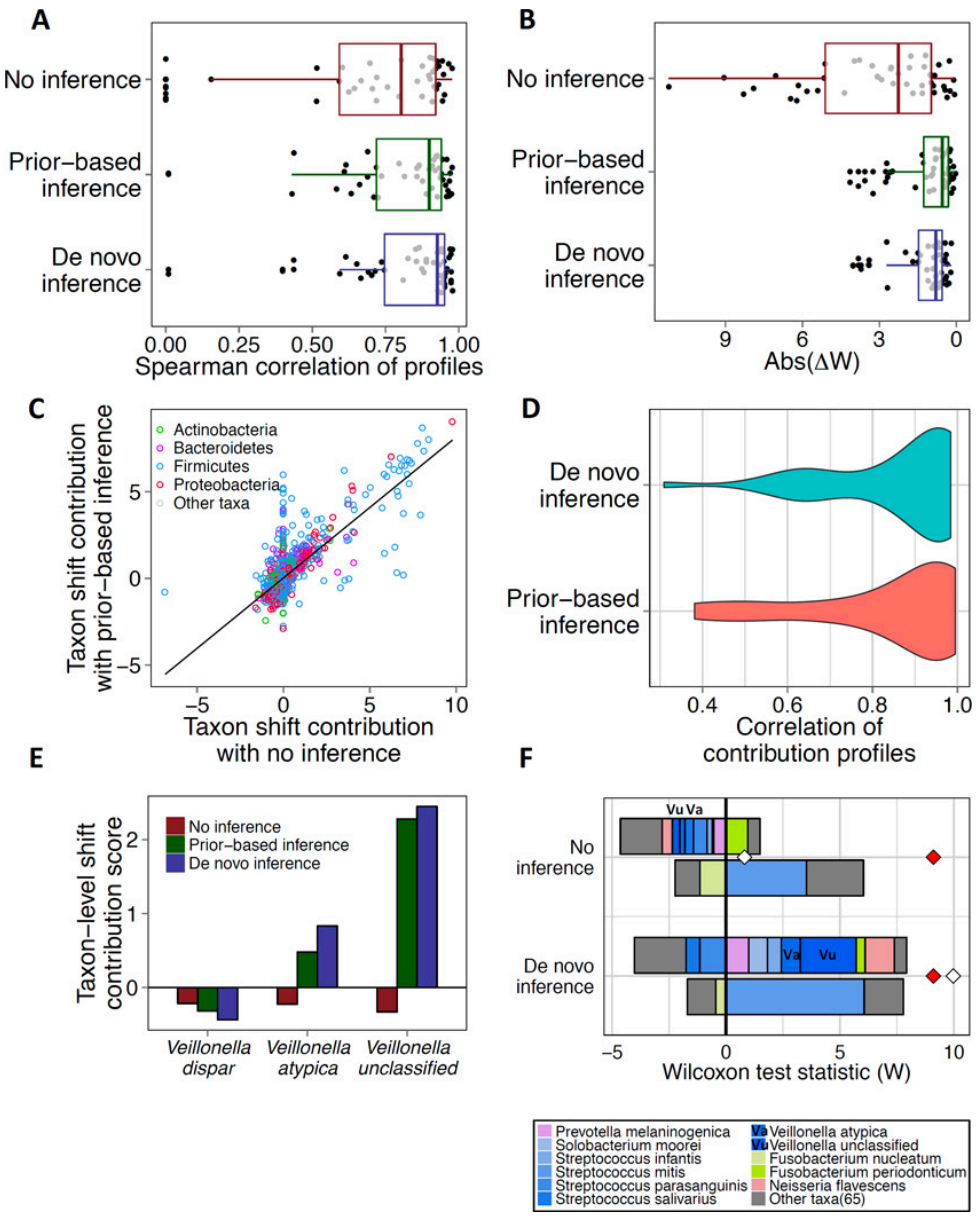
Next did the same thing, but compared taxa contributions to shifts that were common in two diseases (each box, top bars are the fishtaco plot for T2D data and bottom bars are for IBD data)

Methane metabolism is enriched in both T2D and IBD. Can find disease-specific drivers of the function enrichment (e.g. E. coli, in red top right bar, is a T2D-specific driver of this enrichment)

PTS sucrose transport also enriched in both diseases. Shift mostly driven by Firmicutes, just like above some taxa are shared and some are specific to one of the diseases.

C. is mostly them showing that they can also do this at the gene level.





Now they're comparing FishTaco when you do step 1 (infer functional profile of each individual taxa) with only reference genomes or with their previously published genomic content inference method.

no inference = only use genomes
prior-based = use reference genomes as a starting point (I guess to help solve the linear equations? Dunno.)
de novo = use their genomic content inference method

A. correlation with function abundances from metagenomics
B. difference in the "functional shift score"
C. each dot is the contribution of a single species (colored by phyla) to a single pathway.
D. correlations btw genome-based functional shift profiles in HMP vs. the other two ways of getting them
E. a few taxa had different contributions if you used genomes to infer their functional profiles (red) vs. if you inferred it from the data (green and blue).
F. shift profiles for flavonoid biosynthesis. top: if you only use genomes to infer functional content, you don't see an enrichment in this function. bottom: if you do their genomic content inference, the taxa-based "shift" matches the metagenomics-based one (red and white boxes), and you get different contributions