BST281: Genomic Data Manipulation, Spring 2017

Final project example

**Scalable metabolic reconstruction for metagenomic data and the
human microbiome**

Sahar Abubucker, Nicola Segata, Johannes Goll, Alyxandria Schubert, Beltran Rodriguez-Mueller, Jeremy Zucker, the Human Microbiome Project Metabolic Reconstruction team, the Human Microbiome Consortium, Patrick D. Schloss, Dirk Gevers, Makedonka Mitreva, Curtis Huttenhower

Microbial communities carry out the majority of biochemical activity on the planet, and they play integral roles in metabolism and immune homeostasis in the human microbiome2. Shotgun metagenomic data assaying these communities typically comprise short reads from hundreds of organisms, however, and it can be challenging to assemble these sequences comparably to single-organism genomes. Here, we propose an alternative means to determine the functional and metabolic potential of a microbial community metagenome. We infer the pathways present or absent within a community and their relative abundances directly from short reads, allowing a profile of community metabolism and biomolecular functionality to be reconstructed in lieu of metagenome assembly. We validated this methodology using a collection of four synthetic metagenomes, determining the presence and abundance of functional modules and pathways with high accuracy. Finally, we performed metabolic reconstruction on 741 samples drawn from 7 main body sites on 103 individuals as part of the Human Microbiome Project (HMP)3, demonstrating the scalability of our methodology and the critical importance of microbial metabolism in the human microbiome.

Our methodology, the HMP Unified Metabolic Analysis Network (HUMAnN; **Fig. 1**), begins with quality trimmed and filtered Illumina sequence reads. These are mapped to orthologous gene families using accelerated blastx, and the abundance of each gene family is determined by the number of read hits normalized by the quality of each hit (inverse p-value) and the length of the gene. Gene families are assigned to pathways using MinPath4 (maximum parsimony), with subsequent taxonomic limitation to remove pathways discordant with the organismal distribution of the community. Smoothing of these count data is performed numerically (using the Witten-Bell method) and biologically (by allowing for a small number of missing genes in an otherwise well-covered pathway). Finally, each pathway is assigned a coverage (presence/absence) score and a relative abundance. We have primarily analyzed KEGG1 modules (small ~5-20 gene pathways), but the approach is generally applicable to any set of orthologous gene families and to any catalog of functional pathways.



**Figure 1: Overview of the HUMAnN method for metabolic and functional reconstruction from metagenomic data.** The HMP Unified Metabolic Analysis Network (HUMAnN) software recovers the presence, absence, and abundance of microbial gene families and pathways from metagenomic data. Cleaned short DNA reads are aligned to the KEGG Orthology1 (or any other characterized sequence database) using accelerated translated BLAST. Gene family abundances are calculated as weighted sums of the alignments from each read, normalized by gene length and alignment quality. Pathway reconstruction is performed using a maximum parsimony approach followed by taxonomic limitation (to remove false positive pathway identifications) and gap filling (to account for rare genes in abundant pathways). The resulting output is a set of matrices of pathway coverages (presence/absence) and abundances, as analyzed here for the seven primary body sites of the Human Microbiome Project2.

Applying HUMAnN to data from the HMP has provided striking insights linking the microbes present in the human microbiome with their function, metabolism, and host interactions. Specifically, previous studies have found that no organisms as analyzed by 16S taxonomic marker sequences are present in all body sites or individuals. Conversely, we find that 19 of 220 pathways are confidently present in every HMP sample, and over twice as many (53) are present in at least 90% of samples. This demonstrates a degree of functional consistency that is lacking at the organismal level - who's there varies, but what they're doing is more constant. Conversely, the relative abundances of most pathways vary among body sites but not among individuals (**Fig. 2**); 263 of 1,110 pathway/site combinations, for example, are significantly differentially abundant in exactly one body site. This is due in part to very low functional variability at each body site across different hosts, suggesting that community function is strongly dictated by microbial environment and less strongly by the host. Critically, this does not yet speak to host genetics, environment, or disease, as the HMP comprises a normal baseline of healthy individuals; each of these represents an additional area for future studies of microbial community function.

In order to quantitatively validate our approach, we synthesized four mock communities using artificial reads from a maq Illumina error model: two low complexity (20 organisms) and two high (100), two with equal abundances and two with lognormal distributions. Figure 1 demonstrates the significant correlation of HUMAnN's inferences with true abundances in the high-complexity, lognormally distributed community; likewise, our pathway coverage pAUC at an 0.1 FPR was 0.75 (versus naive 0.56). Results in the simpler three communities were comparable, as was the correlation of inferred single gene family abundances (ρ=0.86). Together, this suggests that our methodology improves over current simpler approaches and can recover accurate functional coverage and abundance profiles from experimental datasets such as the HMP.

An open source implementation of our methodology is available online at http://huttenhower.sph.harvard.edu/humann, which we hope will enable future results comparing microbial functionality in the healthy HMP population with disease and with disparate human and environmental microbial communities.



**Figure 1: HUMAnN performance on synthetic and human microbiome data.** **A)** Accuracy of the HUMAnN algorithm compared to a naive best-BLAST-hit approach. Includes reconstructed and gold standard relative abundances of 242 KEGG1 functional modules in a synthetic metagenome comprising 100 KEGG manually curated single-organism genomes with lognormally distributed abundances. **B)** Functional modules with relative abundance significantly differential in at least one of the seven body sites analyzed using 741 shotgun metagenomic samples from the Human Microbiome Project. The KEGG functional hierarchy is colored by the body site in which each module or pathway is significantly overrepresented.

**References**

1. Kanehisa, M. et al. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* (2013).

2. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207-214 (2012).

3. A framework for human microbiome research. *Nature* **486**, 215-221 (2012) PMC:3377744.

4. Ye, Y. & Doak, T.G. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Comput Biol* **5**, e1000465 (2009) PMC:2714467.