From Parts Lists to Ecosystem Networks: Development of a 2*nd* Generation Toolkit for Meta-Omics Analysis

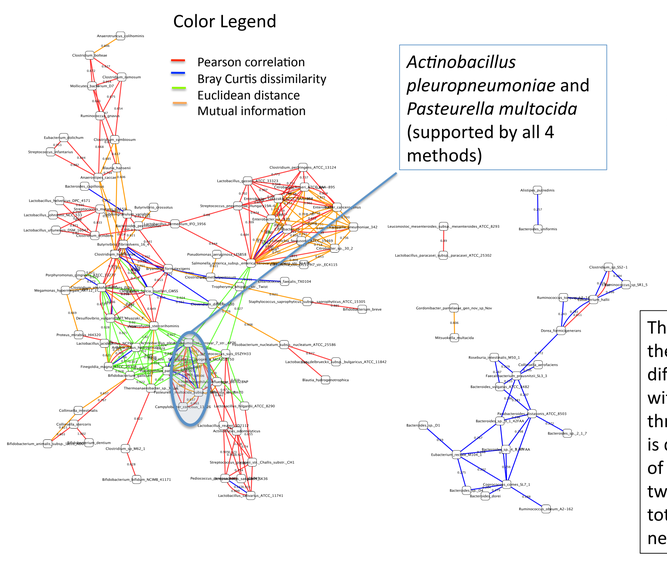
Karoline Faust, Gwen Falony, Falk Hildebrand, Youssef Darzi, Gipsi Lima-Mendez, Shujiro Okuda and Jeroen Raes

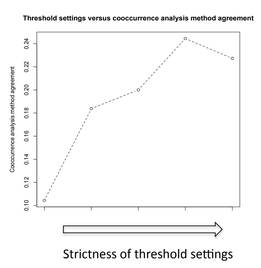
VIB – University of Brussels, Belgium (Contact: jeroen.raes@gmail.com)

Meta-omics (metagenomics, metatranscriptomics, metaproteomics) are powerful tools for the analysis of the human microbiota. Because of its complexity, meta-omics data has required the development of novel computational analysis tools to determine the functional and phylogenetic composition of the sampled community (Raes et al., Curr Opin Microbiol 2007). However, to go from a metagenomic 'parts list' (i.e. a bag of genes) to an initial understanding of the ecosystem structure and functioning (and its alteration in disease), current tools are not sufficient. Here, we present a set of novel approaches to extract species interaction and competition relationships, interpret metabolic changes and determine biologically relevant features from meta-omics data and apply these methods on intestinal microbiota data. Such methods greatly aid in the detection and interpretation of patterns in clinical omics datasets.

Prediction of species interactions with cooccurrence analysis

Disagreement between cooccurrence networks constructed with different methods





The plot above illustrates that the agreement between the different methods increases with increasingly strict threshold settings. Agreement is defined here as the number of links supported by at least two methods divided by the total number of links in the network.

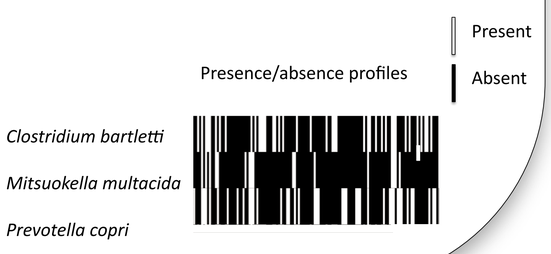
Patterns of co-presence and co-absence over multiple samples (from 16S or metagenomic studies) can be used to predict functional interactions between species. In practice, one computes all pairwise similarity or distance scores between species abundance profiles and then keeps only the pairs with scores above a selected threshold or p-value. The problem is the choice of an appropriate similarity/distance measure and its threshold. The network above unifies networks obtained with various measures. It demonstrates that the output of cooccurrence analysis is highly dependent on the selected measure. To avoid measure-specific biases, we only consider edges supported by more than one method for further analysis. Data from Qin et al. Nature 2010 (MetaHIT)

Going beyond binary cooccurrences

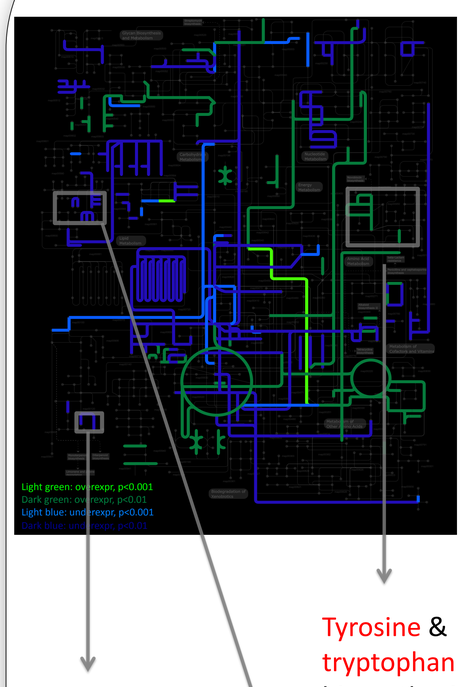
The cooccurrence analysis methods presented on the left can only predict two types of patterns: co-presence or mutual exclusion between two taxa.
We have now developed approaches that extend the number of rule types we can discover with cooccurrence analysis, allowing the discovery of more complex patterns comprising multiple species interactions and metadata information, such as:

- If species A is present and species B is absent, then species C is also present.
- If species A is present and the host is obese, then species B is also present.
- If species A, B and C are absent, then species D is present.

Example:
In the presence of *Clostridium bartletti* and the absence of *Mitsuokella multacida*, *Prevotella copri* is also absent.



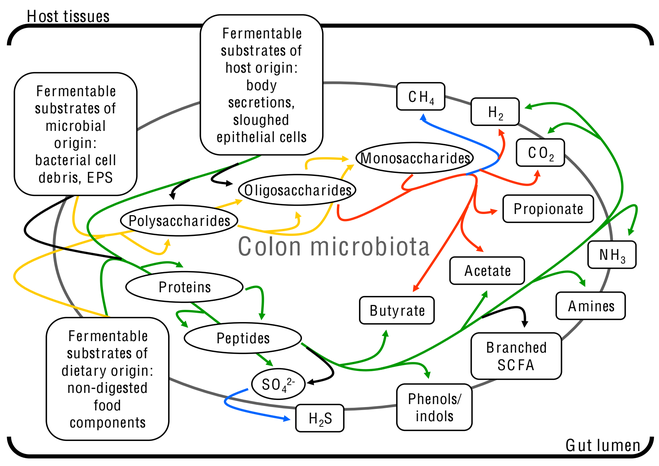
Integration of multiple meta-omics datatypes



Together, metagenomics, -transcriptomics, -proteomics and –metabolomics provide a complete overview of the parts list, the active processes and their metabolic result in microbial ecosystems (Raes and Bork, Nat Rev Microbiol 2009). However, the complexity of such datasets and the ecosystem at hand makes joint interpretation a challenging task. We are developing methods to integrate these multiple datatypes based on metabolic network mapping and pathway analysis. Here, we show the result of such approaches in the analysis of the metaproteome and –metabolome in a Crohn's twin study (collaboration with Janet Jansson, Claire Fraser-Liggett, Robert Hettich), in which we link metabolites overrepresented in Crohn's patients to overexpressed proteins in the metaproteome.

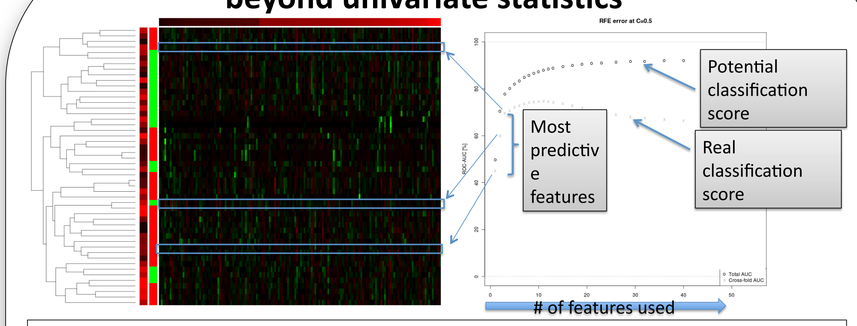
Functional characterization of the human gut ecosystem: implementation of a colon-specific metagenome analysis framework

In order to facilitate analyses of gut metagenomic datasets, we are currently developing a dynamic gut specific annotation and interpretation tool providing an instant overview of the major metabolic fluxes in the human colon ecosystem. This model combines both literature and database information concerning all saccharolytic and proteolytic enzyme systems and fermentation pathways that have been identified within the large-intestinal microbiota to this date. Unlike many currently available metabolic models that originated from and still evolve around eukaryotic cell processes, the present model will be restricted to colon microbiota-specific metabolic pathways, focussing on trophic interactions between large-intestinal microorganisms - including cross-feeding of fermentation products and partially hydrolyzed substrates. Evolving around identified and characterized bacterial clusters composing the core microbiota within the colon ecosystem, it will integrate all relevant data available concerning colon fermentation processes originating from both *in vitro* and *in vivo* trials, including recent 16SrRNA gene-based studies and classical descriptive metagenome analyses. Currently, the model consists of 63 metabolic modules, covering core saccharolytic and proteolytic fermentation processes.



■, Energy metabolism (EM) - Core processes (23 modules);
■, EM - Protein degradation (32 mod.); ■, EM - Other (5 mod.); ■, EM - Carbohydrate degradation (3 mod.)

Biomarker detection in metagenomics – beyond univariate statistics



In many medical studies the ultimate goal is the identification of a biomarker that is clearly associated to a health state of the patient. While in microarray studies more sophisticated methods for biomarker discovery have been established, in metagenomic studies the method of choice is currently mostly univariate statistics. We are developing a suite of tools based on multivariate statistic and machine learning approaches to identify diagnostic markers (both species and genes) from human microbiome data. In the right figure we show the potential of one multivariate biomarker discovery algorithm applied on the MetaHIT dataset (Inter99 cohort, collaboration with a.o. Dusko Ehrlich, Wang Jun, Oluf Pedersen, Peer Bork) in which a feature selection approach detected three functional modules that, in combination, distinguish between obese and lean patients with 70% precision. Only one of these features is significant in a univariate analysis (corrected p-val = 0.046) demonstrating the power of this class of methods.

References
1. J. Raes, K.U. Foerstner and P. Bork (2007), "Get the most out of your metagenome: computational analysis of environmental sequence data", Curr. Opin. Microbiol., vol. 10, pp. 490-498.
2. J. Qin, R. Li, J. Raes et al. (2010), "A human gut microbial gene catalogue established by metagenomic sequencing", Nature, vol. 464, pp. 59-65.
3. J. Raes and P. Bork (2008), "Molecular eco-systems biology: towards an understanding of community function", Nat. Rev. Microbiol., vol. 6, pp. 693-699.