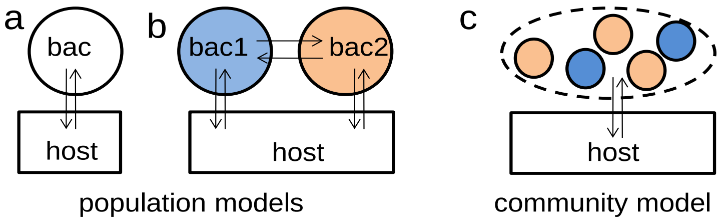
## Use Case 6: Mapping Transcriptome Data to Metabolic Models of Gut Microbiota

A. Douglas (Cornell) with J. Chaston (Brigham Young University), A. Moya (University of Valencia), G. Thomas (University of York), A Heddi (INSA Lyon), and B. Barker (CAC)

***Motivation:*** The gut microbiota plays a central role in maintaining the gut health of humans and other animals. The health benefits of the microbiota vary with the composition of the microbiota, offering an excellent opportunity for microbial therapies to support health and resolve clinical conditions. The goal of this study is to identify the microbial traits and interactions with the host gut that promote host nutritional health. We will use multi-compartment metabolic modeling: first, to investigate how metabolic exchange among the bacteria and with the host influence (and is influenced by) the abundance and persistence of bacteria in the gut; and, second, how the metabolic activities of the bacteria shape nutrient acquisition by the gut, with consequent effects on host nutritional health. Candidate bacterial genes contributing to host nutritional health will be identified by metagenome-wide association studies (requiring the deposition and analysis of ~10TB of genomic sequence data), and metabolic modeling of the systems will elucidate mechanisms and provide strategies for the rational design of microbial therapies to improve human nutritional health. The experimental designs comprise a step-wise increase in bacterial complexity from mono-associations through di-associations and 5-member associations to the unmanipulated microbiota (see figure).

*Structure of multi-compartment metabolic models [bacteria (bac) are Acetobacter (blue) and Lactobacillus (red)] with transport reactions indicated by arrows. The inputs of each metabolic system are defined by the nutrient content of the food.*



This gradual increase in experimental complexity correlates with an increase in data and general computational complexity, which are difficult to estimate precisely up-front. Though particular parts of computation will be preferentially carried out by individual project members, data sharing between collaborators will be critical during all stages of the project.

***Activity:*** Some of the modeling algorithms that will be applied include standard FBA, (dynamic) OptCom, and FALCON [1-3]; note, FALCON was developed by Brandon Barker, a CAC staff member who will be working on the project. We will need to modify or combine aspects of existing algorithms to make the most out of our data and model structure. FALCON and OptCom have a sufficient time complexity that bursting will be necessary, but as our algorithms and models are not finalized, the exact order of computation time needed is unknown. Much of the work will be possible to do, particularly in the development phase, on one or two virtual machines.

Researchers will be able to remotely mount datasets stored in the cloud for non-IO intensive modeling or analysis (for instance, using the secure SSHFS). More intensive data processing can be performed directly in the federated cloud, taking advantage of the scalable resources available. The scale of the metabolic modeling analyses and simulations depend critically on the extent and complexity of datasets obtained during the course of the project, making purchasing of hardware and storage up-front highly undesirable.

Hosting of a VM image that includes software used in simulation and analysis will be available in the cloud for at least 5 years, allowing interested scientists to easily access and deploy a copy of the same system, software, and data used during the project. After 5 years, the VM can be archived at, for example, CAC's Archival Storage with a Globus front end [4]. This would make it simple for any interested researcher (even those without InCommon credentials) to download and use the AWS-compatible VM, which could then be used either on Amazon Web Services, their desktop, or a private KVM-compatible cloud. This has the additional benefit of mitigating the issues that prevent reproducibility in the computational sciences [5].

## References

[1] Orth, J., Thiele, I. & Palsson, O. (2010). What is flux balance analysis? *Nature Biotechnology*, 28(3), 245–48. Retrieved from: http://www.ncbi.nlm.nih.gov/pubmed/20212490.

[2] Zomorrodi, A., Islam, M. & Maranas, C. (2014). d-OptCom: Dynamic multi-level and multi-objective metabolic modeling of microbial communities. *ACS Synthetic Biology* 3(4), 247–57. Retrieved from: http://dx.doi.org/10.1021/sb4001307.

[3] Barker, B. et al. (2014l). A robust and efficient method for estimating enzyme complex abundance and metabolic flux from expression data. Retrieved from: http://arxiv.org/abs/1404.4755.

[4] Foster, I. ( 2005). Globus Toolkit Version 4: Software for Service-Oriented Systems. In *Network and Parallel Computing SE-2*, Lecture Notes in Computer Science, eds. H. Reed & W. Jiang. Springer Berlin Heidelberg, 2–13. Retrieved from: http://dx.doi.org/10.1007/11577188_2..

[5] Peng, R. (2011). Reproducible research in computational science. *Science*, 334(6060), 1226–27. Retrieved from: http://www.sciencemag.org/content/334/6060/1226.abstract.