Journal of Glycomics & Lipidomics

Extended Abstract

# Comparative genomics: From genome sequences to genome biology

**Michael Y. Galperin**

## Abstract

Genome biology aims at using the complete genome sequences to reconstruct all metabolic and signaling pathways that could operate in the target organisms and identify the likely regulatory hubs and potential drug targets. Such analysis requires comprehensive functional annotation of all proteins encoded in each sequenced genome. Standard sequence analysis typically fails to provide (confident) functional assignment for at least a third of the genes even in the relatively small prokaryotic genomes. As a result, comparative genomics has to deal with the constantly growing numbers of "hypothetical" proteins whose functions remain unknown. This talk will discuss using comparative genomics to improve our understanding of microbial metabolic and signaling pathways, including some recent examples of identification of "missing" enzymes and prediction of alternative enzyme variants. It will show that the number of truly enigmatic "conserved hypothetical" proteins is relatively small, particularly in the reduced genomes of pathogenic bacteria, which suggests that most of their cellular functions are already accounted for. In contrast, the number of uncharacterized genes in free-living organisms remains quite large and their functions remain obscure. Our current hypothesis is that many of these genes have “house-cleaning” function, which is almost as important as house-keeping, particularly for aerobic bacteria and for eukaryotic cells. We shall also briefly discuss how comparative genomics could be used for identification of priority targets for future research and the challenges in characterization of their functions.

The genus Aspergillus is one of the best studied genera of filamentous fungi, largely because of the medical (A. fumigatus, A. terreus), food spoilage (A. flavus, A. parasiticus), and industrial (A. niger, A. aculeatus, A. oryzae) relevance of some of its species, in addition to the fundamental studies in the model fungus A. nidulans that have contributed broadly to our understanding of eukaryotic cell biology and molecular processes. Aspergilli can grow in a wide range of niches, mainly in soils and on dead matter, and some are also capable of colonizing living animal or plant hosts and, in total, approximately 350 species have been identified in this genus. The broad relevance and economic importance of the genus has pushed it to the forefront of fungal research, with one of the largest academic and industrial research communities

Aspergillus species are characterized by the unifying feature of the “aspergillum,” an asexual reproductive structure. The aspergilli form a broad monophyletic group, but show large taxonomic divergence with respect to morphology and phylogenetic distance. Genome sequences for three aspergilli were among the first to be reported from filamentous fungi and were soon followed by an additional five genomes. This has resulted in many genomic, comparative genomic, and post-genomic studies covering a wide variety of topics largely due to the size of the Aspergillus research community. These studies were facilitated by genome resources for this genus, such as CADRE and AspGD in which gene curation and functional annotation of reference species were combined with synteny and orthology analysis. The inclusion of these genomes in MycoCosm enabled comparison to sister and more distant genera. These studies also revealed substantial genomic variations between these species and raised questions about the evolution of various aspects of fungal biology within the genus.

In this study, ten novel genome sequences of the genus Aspergillus were generated, namely A. luchuensis, A. aculeatus, A. brasiliensis, A. carbonarius, A. glaucus, A. sydowii, A. tubingensis, A. versicolor, A. wentii, and A. zonatus. These species were chosen primarily to provide better coverage of the whole genus, to complement the already available genome sequences of A. clavatus, A. fischeri, A. flavus, A. fumigatus, A. nidulans, A. niger, A. oryzae, and A. terreus, and to allow more detailed data mining of the industrially relevant section Nigri (A. luchuensis, A. aculeatus, A. brasiliensis, A. carbonarius, A. niger, A. tubingensis). Additional species from the section Nidulantes were included because of the high divergence of the genome sequence of A. nidulans from the other Aspergillus genomes, and A. sydowii because of its marine life-style and being a pathogen of Gorgonian corals. We demonstrate that this combined set of genomes provides a highly valuable dataset for comparative and functional genomics. This study was performed as a global consortium effort with different researchers addressing different topics as subgroups of the consortium. Where possible, experimental data were generated to examine inferences from the genomic differences and to provide an unprecedented comparative analysis of variation and functional specialization within a fungal genus.

Draft genome assemblies and annotations were generated for three coral species: Galaxea fascicularis (Complexa), Fungia sp., and Goniastrea aspera (Robusta). Whilst phylogenetic analyses strongly support a deep split between Complexa and Robusta, synteny analyses reveal a high level of gene order conservation between all corals, but not between corals and sea anemones or between sea anemones. HOX-related gene

clusters are, however, well preserved across all of these combinations. Differences between species are apparent in the distribution and numbers of protein domains and an apparent correlation between number of HSP20 proteins and stress tolerance.

Uniquely amongst animals, a complete histidine biosynthesis pathway is present in robust corals but not in complex corals or sea anemones. This pathway appears to be ancestral, and its retention in the robust coral lineage has important implications for coral nutrition and symbiosis.

**This work is partly presented at 2nd International Conference on Big Data Analysis and Data Mining 30-December 01, 2015 San Antonio, USA**

Michael Y. Galperin
NCBI, National Institutes of Health, USA, E-mail: galperin@ncbi.nlm.nih.gov