

THE REFERENCE GENOME PROJECT

Pascale Gaudet for the Reference Genome Group of the Gene Ontology Consortium: The Reference Genome project is overseen by Pascale Gaudet (dictyBase), Rex Chisholm (dictyBase) and includes these representatives from the curatorial staff: Tanya Berardini (TAIR), Emily Dimmer (GOA), Stacia R. Engel (SGD), Petra Fey (dictyBase), David P. Hill (MGI), Doug Howe (ZFIN), Jim Hu (E.coliWiki), Rachael Huntley (GOA), Varsha K. Khodiyar (UCL), Ranjana Kishore (WormBase), Donghui Li (TAIR), Ruth C. Lovering (UCL), Fiona McCarthy (AgBase), Li Ni (MGI), Victoria Petri (RGD), Deborah A. Siegel (EcoliWiki), Susan Tweedie (FlyBase), Kimberly Van Auken (WormBase), and Valerie Wood (GeneDB)—as well as the following computational staff representatives: Siddhartha Basu (dictyBase), Seth Carbon (BBOP), Mary Dolan (MGI), and Christopher J. Mungall (BBOP)—those establishing the protein families to be annotated: Kara Dolinski (PPOD), Michael S. Livstone (PPOD), and Paul Thomas (PANTHER)—and, the four PIs of the GO Consortium: Michael Ashburner (FlyBase), Judith A. Blake (MGI), J. Michael Cherry (SGD), and Suzanna E. Lewis (BBOP).

Complete functional annotation of genomes is a powerful tool for researchers; however, such annotation is a time-consuming task limited by the availability of experimental data. The function of genes for which there is no experimental data can often be predicted via comparison to related, annotated sequences of known function. We describe here the Reference Genome project, an effort from the Gene Ontology (GO) Consortium to fully annotate twelve genomes to rigorous standards: human, plus eleven organisms that are important models in biomedical research, including mouse, fly, zebrafish, yeast and *E. coli*. To achieve this, we examine existing experimentally based annotations in a phylogenetic context in order to infer the function(s) of ancestral proteins and propagate these annotations to their descendants. This endeavor faces many difficult challenges, such as: the determination and provision of reference protein sets for each genome; the identification of gene families for curation; the application of consistent best practices for annotation; the development of methodologies for evaluating progress towards our goal; and the development of software tools to support this effort.

Annotated genomes are greatly valuable to the research community and will provide the basis for using sequence similarity to annotate further genomes. An overview of the project as well as links to all resources described below can be found at <http://geneontology.org/GO.refgenome.shtml>

This work is supported by NHGRI grant #HG002273 (Gene Ontology Consortium) and NIGMS #GM081084-01A1 (Phylogenetic tree building and annotation software development).