Identification and functional annotation of taxonomically-restricted genes in bacteria

Genes create families — lineages that stretch through time, all the way back to a founding member. This ancestor gene multiplied into many copies and morphed a bit with each new generation. For most of the last 40 years, scientists assumed that this was the primary mechanism how new genes were born — they simply arose from copies of existing old genes (*duplication*). The old genes have to continuously work to maintain functions required by an organism (*selective pressure*), and the new copies became free to evolve novel properties as a response to different environmental demands (*neofunctionalization*).

However, recently it has been shown that some genes are born outside such 'family circles'. These genes do not share common ancestry (*homology*) with any known gene from other organisms. Therefore, they are named orphan/ORFan genes or more precisely, Taxonomically-Restricted Genes (TRG). The origin of TRGs in scientific literature has been often described as *mysterious*, *enigmatic* or *unknown*. Although most of these genes have not been characterized, recent studies have shown that TRGs are important for specific, organismal properties such as: limb regeneration in salamanders, sociality in the honey bee and suppression of human cancer.

Current studies aiming at large-scale identification of TRG genes usually overestimate the number of TRG genes due to incomplete set of considered in research sequences or lack of precision of computational methods used for detection of homologs. Consequently, thousands of genes previously proposed as orphans turned out to have distant homologs in other organisms and thus, lost their uniqueness. Therefore, in the presented project we will use a whole arsenal of bioinformatics tools (including our own solutions) to identify and characterize TRG genes in a representative group of organisms that includes over 40 thousand bacterial strains (i.e. ~140 million genes).

The ultimate goal of the project is to look for answers to the essential questions in biology of TRG in bacteria: - *How spread are the TRG genes on different levels of bacterial taxonomy classification (genera, species and strains)? - Is it possible to define biological factors (e.g., specific environmental conditions) that influence orphan emergence? - To what extent the diversification of gene content reflects the ecological needs of different taxa and emergence of new features (e.g., antibiotic resistance). - Are TRG genes more often found in certain bacterial groups (e.g. pathogenic bacteria)? - What are the predominant molecular genetic mechanisms controlling orphan appearance, turnover, and fixation? - What is the function of TRG genes?*

Preliminary results of our analyses of more than 6 thousand bacterial strains show that every proteome contains on average 3% unique TRG genes. Interestingly, the highest numbers of TRG genes are present in pathogenic bacteria. The unquestionable champion of this comparison is *Prevotella copri* strain related to rheumatoid arthritis - it contains over 40% genes (2025 out of 4835) that have no recognisable homologues outside its genome, even in other strains of the same species!

A comparative genome analysis of genes conserved in pathogenic bacteria but not in their eukaryotic hosts may offer insights into the drug therapy approaches of the near future. We performed a pilot study on full set of proteins from *S. aureus MRSA225* strain, which is resistant to most antibiotics and has been associated with the recently largest outbreak of infections. We found in this strain 39 specific genes that produce proteins directly associated with pathogenicity (e.g., toxins, factors suppressing the immunological response, proteins involved in iron acquisition through capturing heme from human hemoglobin).

As the final product of this project we will create a web-based and publicly available encyclopedia of TRG genes in bacteria. The portal will offer card-based information interface for each of the predicted TRG genes. We expect that the portal will provide reference and inspiration for molecular biologist to study function of TRG genes in various strains, species and groups of bacteria. In addition, information stored in the portal will allow for identification of sequences that can be used as molecular markers in metagenomic, environmental or medical studies.