Biological modules associated with prophage density in pathogenic and commensal *Escherichia coli*

Tarcisio José Domingos Coutinho1, Glória Regina Franco2 e Francisco Pereira Lobo1

1 - Departamento de Biologia Geral – ICB/UFMG, 2 - Departamento de Bioquímica e Imunologia – ICB/UFMG

Integrated phages (prophages) play an important role in the genomic diversification and fitness cost to the infected host, since they are major contributors to the diversity of bacterial gene repertories but may also lead to host death. Prophages able to propagate horizontally may allow bacteria to adapt to many ecological niches through horizontal transfer of biological modules. Our evolutionary interest in prophages consist in understanding the role of adaptive genes that reach bacterial genomes through phage integration and their contributions to the complex antagonistic and mutualistic prophage-bacteria interactions. We used comparative genomics and statistical analyses to study the influence of genes carried by prophages in *Escherichia coli* pathogenic and commensal lineages. We downloaded 33 and 17 complete genomes of pathogenic and non-pathogenic *E. coli* strains, respectively, from the National Center for Biotechnology Information (NCBI). We used PHASTER (http://phaster.ca/) to analyze bacterial genomes in order to find prophages (intact, incomplete or questionable), Rstudio (https://www.rstudio.com/) for graphics and KOMODO2 (https://www.komodo.cnptia.embrapa.br/) for correlation analyses. We detected 470 phages in *E. coli* genomes, 359 in pathogenic and 111 in non-pathogenic lineages. For each genome, we calculated the phage density (number of phages divided by genome length). We observed that pathogenic *E. coli* contain a significantly higher number of prophages when compared with non-pathogenic strains (Wilcoxon test, p-value = 0.009283). In order to detect potential modules associated with phage density, we searched for Gene Ontology (GO) terms whose frequency increases or decreases with prophage density in the two groups of *E. coli* analyzed. We selected 225 GO terms (115/110 terms with Pearson correlation > 0.5 and < -0.5 respectively) in pathogenic lineages and compared their correlation values with the ones found in non-pathogenic lineages. We found several of these correlated terms to be exclusive of pathogenic lineages (e.g. catalase, urea, hemolysis, parasitism, urease). Other terms presented a correlation higher than 0.5 in pathogenic lineages, but virtually no correlation in non-pathogenic ones that are related to metabolic pathways and symbiosis (e.g. nickel cation binding, cysteine-type peptidase activity, metallochaperone activity, DNA replication initiation, modification by symbiont of host morphology or physiology). Together, our analyses suggest that phages carry several biological modules that favor a pathogenic phenotype in *E. coli* lineages, and also others that may affect their metabolism, with both classes favoring bacterial fitness and evolution.