Exercise on enrichment and pathways

In an industrial laboratory, several rounds of random mutagenesis and selection were carried out on a strain to increase a specific metabolic capacity. To analyze the differences between the evolved strain and the wild type, gene expression was assessed using microarrays for both strains under the same condition, with two replicates for each strain. The data is available for download as 'Sclav_data.xlsx'. A t-test was used to evaluate which genes were significantly overexpressed in the evolved strain compared to the wild type, the result of which is displayed in column K of the first sheet. Besides the general functional annotations of the genes, EC numbers have also been predicted (when possible) to allow linking the metabolic capacities of the strain to the expression data. Note that most genes are not enzyme-coding or encode enzymes of unknown function, and therefore do not have an EC number assigned.

The second sheet of the Excel file contains a conversion table to link EC numbers to KEGG pathways (see http://www.genome.jp/kegg-bin/show_pathway?map00520 for an example pathway), and the third sheet contains a set of 50 metabolic pathways from the KEGG database that can be used to test whether any of them are overrepresented in the set of significantly overexpressed genes.

Write a script to perform a Fisher's exact test for each of the metabolic pathways in the 'Pathways' sheet. For the calculations, use the `fisher_exact()` method from Scipy (http://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.fisher_exact.html) . For each pathway, you will compare the number of genes in the set of overexpressed genes with the number of genes in the rest of the genome, based on the EC annotations of the genes. With an initial alpha value of 0.05, use the Bonferroni method to correct your p-value for multiple testing.
Write an output function for your script that provides a tab-delimited table with KEGG pathway names and obtained p-values.
Which pathways are significantly overrepresented in the genes that are overexpressed in the evolved strain, compared to the wild type?

Have a look at the scientific paper at http://onlinelibrary.wiley.com/doi/10.1111/j.1751-7915.2010.00226.x/full that describes the biological context of the experiment and see if the results make sense. Do your analyses reveal any new insights compared to what was described in the paper?

The Bonferroni correction can be somewhat conservative, in the sense that it tries to protect strictly from getting false positives to the expense of sometimes getting a relatively high number of false negatives. Based on the context provided by the paper, do you see any candidate pathways that are likely to have a real signal based on biological interpretation, yet which did not reach the adjusted threshold due to the Bonferroni correction? I.e., are there candidates for false negatives? What would be an alternative (less conservative) way to correct for multiple testing?

Optional: For those pathways with p < alpha, plot the expression values on the corresponding EC numbers in a pathway map visualization using iPath. See http://pathways.embl.de/help.html for instructions.