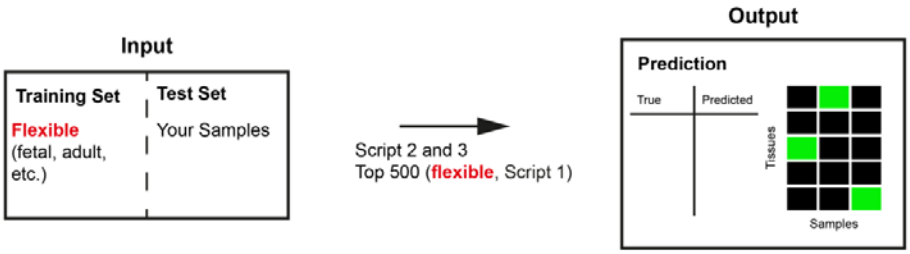
## What is KeyGenes?

KeyGenes is an algorithm to predict the identity and determines identity scores of queried samples (test set) to a provided group of samples (training set). It uses transcriptional profiles of the queried data (test set) and matches them to sets of transcriptional profiles of organs or cell types (training set). KeyGenes uses a 10-fold cross validation on the basis of a LASSO (Least Absolute Shrinkage and Selection Operator) regression available in the R package “glmnet” (Friedman et al., 2010).

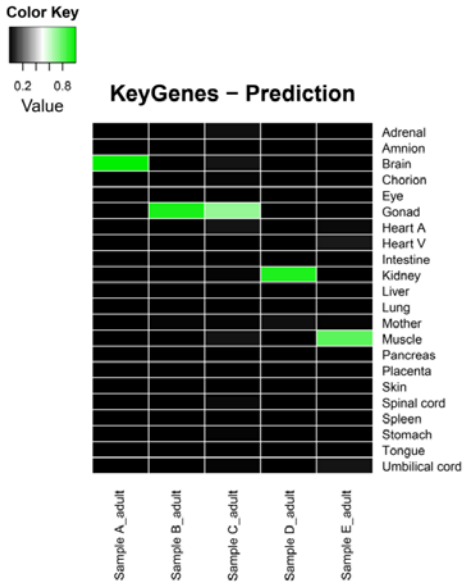



Information about the different “fixed” training sets provided as a headstart as well as the instructions how to use either the Web App on “fixed” training sets or the R scripts on “fixed” or “flexible” training sets can be found on http://www.keygenes.nl/ (“How to use KeyGenes”). The R scripts and the different available “fixed” training sets (with associated files), can be downloaded from http://www.keygenes.nl/.

## What do I get from KeyGenes?

The output from KeyGenes consists of four files:

1. A PDF file (KeyGenes_Heatmap.pdf) with a heatmap containing the identity scores (between 0 and 1) of the queried samples matched to the samples included in the training set.



2. A text file (KeyGenes_Matrix.txt) containing a matrix with the identity scores (between 0 and 1) of the queried samples matched to the samples included in the training set.

| | **Sample A_adult** | **Sample B_adult** | **Sample C_adult** | **Sample D_adult** | **Sample E_adult** |
|---|---|---|---|---|---|
| **Adrenal** | 0.0000 | 0.0016 | 0.0342 | 0.0001 | 0.0062 |
| **Amnion** | 0.0001 | 0.0041 | 0.0093 | 0.0006 | 0.0060 |
| **Brain** | 0.9968 | 0.0014 | 0.0410 | 0.0003 | 0.0007 |
| **Chorion** | 0.0005 | 0.0034 | 0.0142 | 0.0004 | 0.0094 |
| **Eye** | 0.0002 | 0.0010 | 0.0005 | 0.0000 | 0.0000 |
| **Gonad** | 0.0000 | 0.9447 | 0.7024 | 0.0001 | 0.0023 |
| **Heart A** | 0.0000 | 0.0057 | 0.0426 | 0.0001 | 0.0262 |
| **Heart V** | 0.0000 | 0.0001 | 0.0002 | 0.0000 | 0.0511 |
| **Intestine** | 0.0000 | 0.0008 | 0.0075 | 0.0000 | 0.0004 |
| **Kidney** | 0.0000 | 0.0009 | 0.0143 | 0.0000 | 0.0008 |
| **Liver** | 0.0003 | 0.0034 | 0.0024 | 0.9961 | 0.0014 |
| **Lung** | 0.0000 | 0.0001 | 0.0006 | 0.0000 | 0.0005 |
| **Mother** | 0.0005 | 0.0116 | 0.0120 | 0.0010 | 0.0193 |
| **Muscle** | 0.0000 | 0.0029 | 0.0449 | 0.0000 | 0.8185 |
| **Pancreas** | 0.0000 | 0.0001 | 0.0018 | 0.0000 | 0.0001 |
| **Placenta** | 0.0006 | 0.0022 | 0.0055 | 0.0001 | 0.0049 |
| **Skin** | 0.0001 | 0.0000 | 0.0008 | 0.0000 | 0.0001 |
| **Spinal cord** | 0.0001 | 0.0048 | 0.0220 | 0.0000 | 0.0007 |
| **Spleen** | 0.0000 | 0.0014 | 0.0070 | 0.0010 | 0.0008 |
| **Stomach** | 0.0000 | 0.0015 | 0.0148 | 0.0000 | 0.0003 |
| **Tongue** | 0.0000 | 0.0000 | 0.0088 | 0.0000 | 0.0029 |
| **Umbilical cord** | 0.0006 | 0.0081 | 0.0130 | 0.0001 | 0.0475 |

3. A text file (KeyGenes_Prediction.txt) with the queried samples and the sample in the training set with the highest identity score.

| **True Tissue** | **Predicted Tissue** |
|---|---|
| Sample A_adult | Brain |
| Sample B_adult | Gonad |
| Sample C_adult | Gonad |
| Sample D_adult | Kidney |
| Sample E_adult | Muscle |

4. A text file (KeyGenes_Classifier.txt) containing the list of classifier genes per sample calculated from the training set used to determine the identity scores (between 0 and 1) of the queried samples matched to the samples included in the training set.

| | | | | |
|---|---|---|---|---|
| **Adrenal** | ENSG00000105398 | ENSG00000109132 | ENSG00000136931 | ENSG00000147256 |
| **Amnion** | ENSG00000000005 | ENSG00000078399 | ENSG00000094755 | ENSG00000115221 |
| **Brain** | ENSG00000120068 | ENSG00000125462 | ENSG00000130287 | ENSG00000170370 |
| **Chorion** | ENSG00000079689 | ENSG00000094755 | ENSG00000106366 | ENSG00000114854 |
| **Eye** | ENSG00000080166 | ENSG00000122592 | ENSG00000137273 | ENSG00000147655 |
| **Gonad** | ENSG00000104435 | ENSG00000115596 | ENSG00000143355 | ENSG00000143954 |
| **Heart A** | ENSG00000174429 | ENSG00000242349 | | |
| **Heart V** | ENSG00000129991 | ENSG00000160808 | | |
| **Intestine** | ENSG00000112818 | ENSG00000173702 | ENSG00000181541 | ENSG00000233041 |
| **Kidney** | ENSG00000074803 | ENSG00000075891 | ENSG00000116218 | ENSG00000128713 |
| **Liver** | ENSG00000055957 | ENSG00000077274 | ENSG00000091513 | ENSG00000132855 |
| **Lung** | ENSG00000164265 | | | |
| **Mother** | ENSG00000106511 | ENSG00000117472 | ENSG00000164825 | ENSG00000166426 |
| **Muscle** | ENSG00000000005 | ENSG00000122180 | ENSG00000133055 | ENSG00000180818 |
| **Pancreas** | ENSG00000114204 | ENSG00000120057 | ENSG00000130675 | ENSG00000137731 |
| **Placenta** | ENSG00000164707 | ENSG00000170498 | ENSG00000243130 | |
| **Skin** | ENSG00000092607 | ENSG00000121742 | ENSG00000152785 | ENSG00000167768 |
| **Spinal cord** | ENSG00000052344 | ENSG00000120068 | ENSG00000171532 | ENSG00000177551 |
| **Spleen** | ENSG00000073754 | ENSG00000133135 | ENSG00000136931 | ENSG00000183072 |
| **Stomach** | ENSG00000066405 | ENSG00000131668 | ENSG00000134812 | ENSG00000182333 |
| **Tongue** | ENSG00000122180 | ENSG00000131668 | ENSG00000133055 | ENSG00000169469 |
| **Umbilical cord** | ENSG00000052850 | ENSG00000077279 | ENSG00000120149 | ENSG00000156076 |

### References:

Friedman J., Hastie T., and Tibshirani R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*. 33(1): 1-22.