**How to use omics data for reconstruction of metabolic models**

**FastCore** List of reactions present

**GIMME** List with scores (or expression values) for each reaction. If no expression data, score should be -1 (will be considered ‘expressed’)

**IMAT** List with tri-valued scores (or expression values) for each reaction. If no expression data, score should be negative (will be considered moderately expressed)

**CORDA** 3 lists: high, medium and negative confidence reactions

**tINIT** List with scores for each reaction. Optional: list of metabolites that should be present

**1. Assess presence/absence (PA) of genes in transcriptomics samples:**

a) Determine PA call for a gene in a sample (based on <https://doi.org/10.1371/journal.pcbi.1007185>):

- Assess distribution of all genes in all samples of the dataset;

- If gene expression in the sample is lower than the first quartile, gene is considered off (PA = 0)

- If gene expression in the sample is greater than the third quartile, gene is considered on (PA = 1)

- For the other cases, the distribution of the gene across all samples is assessed:

- If gene expression in the sample is lower than the medium, gene is considered off (PA=0)

- If gene expression in the sample is greater than the medium, gene is considered on (PA=1)

- For each gene, the average of the PA calls across the samples is calculated, obtaining the *Presence Ratio*

b) Determine final transcriptomics score for a gene. Presence ratios from both transcriptomics types (RNAseq and Microarray) are averaged.

b.1) If there are also proteomics data to compare it to:

|  |  |  |
| --- | --- | --- |
| **RNAseq Presence Ratio** | **Microarray Presence Ratio** | **Transcriptomics Score** |
| mean ratio >= 0.75 | | 2 |
| 0.5 < mean ratio < 0.75 | | 1 |
| mean ratio <= 0.5 | | 0 |
| NA | ratio >= 0.8 | 2 |
| NA | 0.5 < ratio < 0.8 | 1 |
| NA | ratio <= 0.5 | 0 |
| ratio >= 0.8 | NA | 2 |
| 0.5 < ratio < 0.8 | NA | 1 |
| ratio <= 0.5 | NA | 0 |
| NA | NA | -1 |

\*NAs are due to: no experimental data for the gene(s)

b.2) If there are not proteomics data to compare it to, these will be the final gene scores:

|  |  |  |
| --- | --- | --- |
| **RNAseq**  **Presence Ratio** | **Microarray Presence Ratio** | **Transcriptomics Score** |
| ratio >= 0.75 | ratio >= 0.75 | 3 |
| mean ratio <= 0.5 | | 0 |
| Other cases | | 1 |
| NA | ratio >= 0.8 | 2 |
| NA | 0.5 < ratio < 0.8 | 1 |
| NA | ratio <= 0.5 | 0 |
| ratio >= 0.8 | NA | 2 |
| 0.5 < ratio < 0.8 | NA | 1 |
| ratio <= 0.5 | NA | 0 |
| NA | NA | -1 |

\*NAs are due to: no experimental data for the gene(s)

**2. [If available] Assess presence/absence of genes in proteomics samples (protein copy numbers):**

a) Determine PA call for a gene in a sample (based on <https://doi.org/10.1371/journal.pcbi.1007185>):

- Assess distribution of all genes in all samples of the dataset;

- If gene expression in the sample is lower than the first quartile, gene is considered off (PA = 0)

- If gene expression in the sample is greater than the third quartile, gene is considered on (PA = 1)

- For the other cases, the distribution of the gene across all samples is assessed:

- If gene expression in the sample is lower than the medium, gene is considered off (PA=0)

- If gene expression in the sample is greater than the medium, gene is considered on (PA=1)

- For each gene, the average of the PA calls across the samples is calculated, obtaining the *Presence Ratio*

b) Determine final proteomics score for a gene. This only applies if data from only one dataset is available. If more than one dataset is available, it was done similarly to point *1.b.1)*.

|  |  |
| --- | --- |
| **Presence Ratio** | **Proteomics Score** |
| ratio >= 0.8 | 2 |
| 0.5 < ratio < 0.8 | 1 |
| ratio <= 0.5 | 0 |
| NA | -1 |

\*NAs are due to: no experimental data for the gene(s)

**4. Assess Final gene scores, if both transcriptomics and proteomics are available:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Transcriptomics** | **Proteomics** | **Obs.** | **Final Gene Scores** |
| 2 | 2 |  | 3 |
| 2 | 1 |  | 2 |
| 2 | 0 | If all prote- samples are 0  Else | 0  1 |
| 2 | -1 |  | 1 |
| 1 | 2 |  | 2 |
| 1 | 1 |  | 1 |
| 1 | 0 |  | 0 |
| 1 | -1 |  | 0 |
| 0 | 2 | If all prote- samples are 0  Else | 0  1 |
| 0 | 1 |  | 0 |
| 0 | 0 |  | 0 |
| 0 | -1 |  | 0 |
| -1 | 2 |  | 1 |
| -1 | 1 |  | 0 |
| -1 | 0 |  | 0 |
| -1 | -1 |  | -1 |

\*-1 are due to: no experimental data for the gene(s)/enzyme(s)

**3. GPR rules: from genes to reactions – Gene Mapping**

To get the reaction scores from the genes, Gene-Protein-Reaction (GPR) rules are used. When a reaction is catalyzed by a multimeric enzyme complex (*AND* rule), the minimum expression sets the activity of the reaction. When a reaction is catalyzed by iso-enzymes (*OR* rule), the highest expression is considered to set the activity of the reaction. Examples:

- Gene A OR Gene B: the one with higher expression will set the reaction’s expression

- Gene A AND Gene B: the one with lower expression will set the reaction’s expression

- (Gene A AND Gene B) OR Gene C: the higher expression between gene C and the minimum expression between gene A and gene B will set the reaction’s expression. Example 1: for A=1, B=2 and C=0, the reaction expression will be 1. Example 2: for A=0, B=1 and C=2, the reaction expression will be 2.

The GPR rules are present in the general model to use as base for the reconstruction.

**4. Assess Final presence of reactions**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Final Gene Scores** | **FastCore** | **GIMME** | **IMAT** | **CORDA** | **tINIT** |
| 3 | Present | 2 | 2 | High | 20 |
| 2 | Present | 2 | 2 | High | 15 |
| 1 | Present | 1 | 1 | Medium | 10 |
| 0 | - | 0 | 0 | Negative | -8 |
| -1 | - | -1 | -1 | - | -2 |
| NA | - | -1 | -1 | - | -2 |

\*-1 are due to: no experimental data for the gene(s)/enzyme(s) of that reaction

\*NAs are due to reaction with no GPR (i.e., no gene or enzyme associated).