1. **Aim**

*A good example is the best sermon*, claimed American polymath Benjamin Franklin. This worked-out example is intended as a step-by-step guide to assessing differential protein abundance from scratch using *iSanXoT*. To this end, we shall use PSM (peptide-spectrum match) data from a high-throughput, quantitative proteomics experiment based on liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS).

1. **Project description and starting files**

|  |  |
| --- | --- |
| Table 1. **Isotopic sample labeling.** | |
| **Sample Name** | **iTRAQ label** |
| Jurkat WT 1 | 113 |
| Jurkat WT 2 | 114 |
| Jurkat WT 3 | 115 |
| Jurkat WT 4 | 116 |
| Jurkat KO 1 | 117 |
| Jurkat KO 2 | 118 |
| Jurkat KO 3 | 119 |
| Jurkat KO 4 | 121 |

We shall be working with a simple model consisting of four wild-type (WT) and four knock-out (KO) Jurkat cell preparations which we would like to compare for differential protein abundance. The protein extracts obtained from these eight samples were subjected to tryptic protein digestion, after which the resulting peptides were isotopically labelled using the iTRAQ Reagent 8 plex kit[[1]](#footnote-1) as depicted in Table 1. Then the labeled peptide samples were pooled and partly separated into five fractions using high-pH reversed-phase chromatography[[2]](#footnote-2). LC-MS/MS analysis of both the unfractionated material and the five peptide fractions resulted in the six raw LC-MS/MS files that can be found in …\WOE\Raw\_files. These raw files were used as inputs to Protein Discoverer 2.1[[3]](#footnote-3) to generate, among others, the six *PSM.txt* files stored in …\WOE\PSM.txt\_files upon database searching against the FASTA human protein database that can be found in …\WOE\Database\_and\_category\_files. Please note that this protein database contains both true protein sequences and the corresponding inverted (decoy) sequences to enable false discovery rate (FDR) estimation for peptide identification.

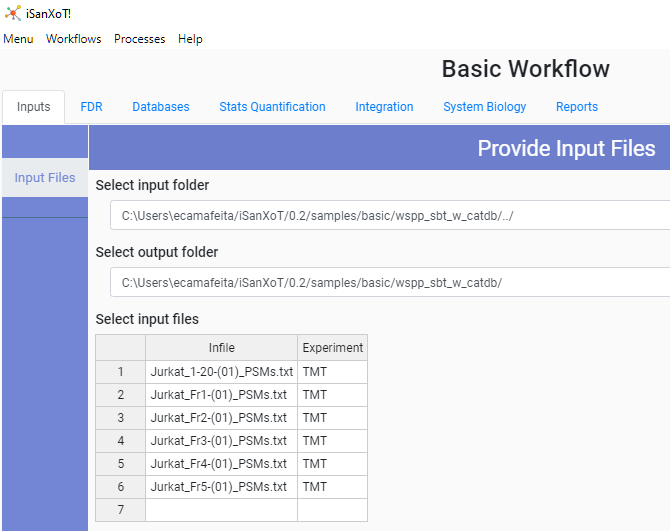
A close inspection of the *PSM.txt* files reveals that these plain text files hold 57 columns displaying information for every PSM obtained upon database searching; however, only a reduced subset of these data will be accessed by *iSanXoT,* as shown in Table 2.

|  |  |
| --- | --- |
| Table 2. **Data accessed by iSanXoT in the *PSMs.txt* files.** | |
| **Header** | **Description** |
| Spectrum File | Name of the raw LC-MS/MS file |
| First Scan | Scan number identifier |
| Sequence | Peptide amino acid sequence |
| Modifications | Unimod[[4]](#footnote-4) chemical or posttranslational modifications to peptide sequence |
| Charge | Peptide charge |
| XCorr | Cross-correlation value as provided by SEQUEST[[5]](#footnote-5) algorithm |
| MH+ [Da] | Measured monoisotopic protonated peptide mass in Da |
| Theo. MH+ [Da] | Theoretical monoisotopic protonated peptide mass in Da |
| DeltaM [ppm] | Difference between measured and theoretical monoisotopic mass in ppm |
| Protein Accessions | Accession codes for the proteins to which the peptide sequence is ascribed |
| 113-121 | Intensity of iTRAQ reporter ions 113-121 |

1. **Getting started: the Inputs menu**

Once the program has been successfully installed, executing the *isanxot.bat* script will bring you to the *iSanXoT* main page, where a number of predefined workflows are displayed. The link *WSPP-SBT sample with iSanXoT databases* found below the short description of the Basic Workflow will take you to the workflow *Inputs* window (Fig. 1), where the following information is provided:

* + *Input folder* specifies the location of the files containing PSM identification and quantification data (the *PSMs.txt* files in this example);
  + *Output folder* describes the path to the folder where *iSanXoT* output files will be stored. Selection of an *output folder* other than the *input folder* is strongly recommended;
  + *Select input files* indicates which of the *PSM.txt* files stored in the *input folder* must be considered by *iSanXoT* (all six files in this case). *PSMs.txt* file names are listed under *Infile*, while their experiment allocation is indicated in *Experiment*. All six *PSMs.txt* files originate from the same experiment (termed “TMT”) in our example, but larger projects may encompass several experiments (*e.g.* “TMT1”, “TMT2” and “TMT3”; or “Exp1”, “Exp2” and “Exp3”). Every time an *Input Folder* is selected (by clicking on *Choose folder*), *Infile* cells are automatically filled in with the names of every single file or subfolder therein; be aware that any file other than those containing PSM identification and quantitation data (*e.g.* *PSMs.txt* files) should be removed from *Infile*.

1. **Validating peptide identification with FDR**

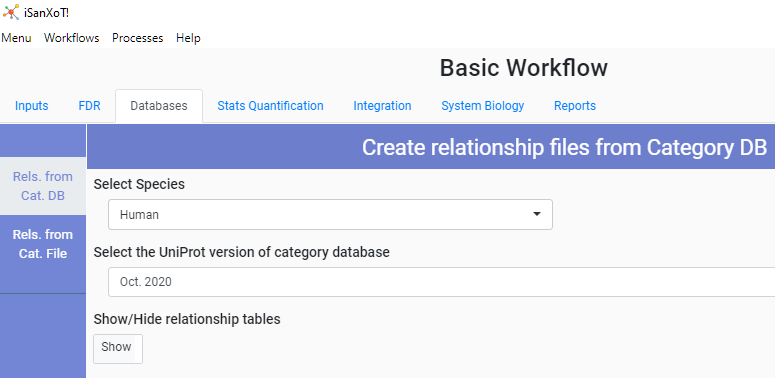
The next step is to validate the collection of PSMs that we have used as inputs in the previous Section. For that, *iSanXoT* relies on the probability ratio (pRatio) method[[6]](#footnote-6), an algorithm that calculates the probability of random peptide matching and provides the corresponding FDR for peptide identification. The *FDR* menu displays the following fields:

Figure 1. **The Basic Workflow Inputs window after loading the sample data.**

* *Experiment*, the aforementioned experiment allocation;
* *Threshold (ppm)* isthe postscoring mass filtering cutoff to be applied after using wide mass windows in the database search[[7]](#footnote-7), as was the case with these sample data. Threshold is actually equivalent to the relative deviation experimentally observed for precursor ions in a particular LC-MS/MS run (15 ppm in this case);
* *Isotopologue recovery* allows pRatio to recover precursor m/z values matching some 13C isotopologue when using wide mass windows in the LC-MS/MS acquisition7 that otherwise would remain unnoticed. For that, it must be indicated whether precursor m/z values should be tracked only around their experimental m/z (value = 1) or also ± 1 Th (value = 3) and ± 2 Th (value = 5, the one used here) away.
* *FDR* establishes the FDR cutoff for PSM validation. The value used here (0.01, *i.e.* 1% FDR) implies that one in every 100 validated PSMs is incorrect;
* *XCorr type* determines whether XCorr (SEQUEST5 cross correlation score) or the corrected XCorr (cXCorr) will be used by pRatio for FDR calculation;
* *Decoy label* is the tag attached to decoy protein identifiers in the concatenated protein database previously used for peptide identification (“INV\_” in this example).

1. **Relations matter: the Databases menu**

Figure 2. **The Basic Workflow Databases window in the WSPP-SBT sample with iSanXoT databases option.**

Two databases are necessary for *iSanXoT* to work: a protein sequence database and a protein-category relation database. The latter, a plain text file relating proteins to functional categories, is used by the Systems Biology Triangle (SBT) algorithm to estimate the experimental protein variance[[8]](#footnote-8). But don’t panic (yet): since you arrived here through the *WSPP-SBT sample with iSanXoT databases* link, your *Databases* menu should look like the one on Fig. 2. This means that you are having *iSanXoT* use its own databases downloaded from Uniprot[[9]](#footnote-9), and therefore all you have to do is to select a species (human in this case, since we are using Jurkat cell preparations) from the *Select Species* drop-down menu and a database version from the *Select version of Uniprot category database* menu.

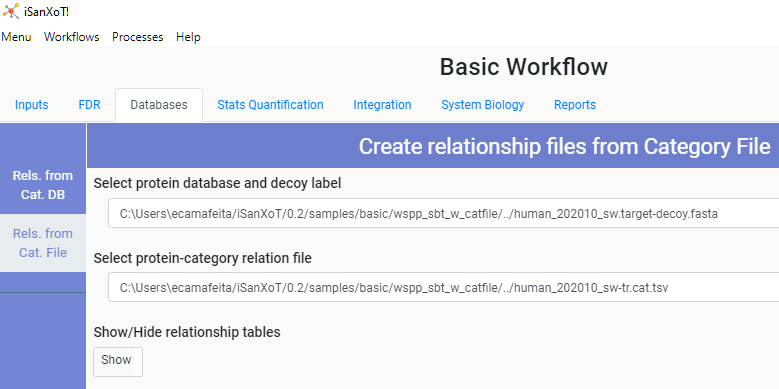
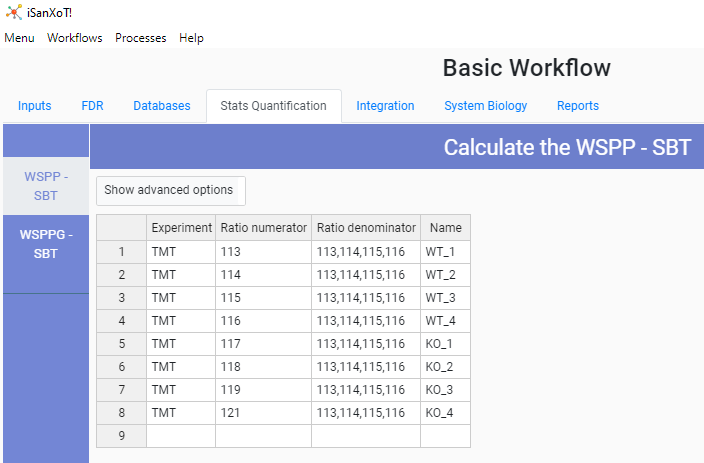
 Nevertheless, if for some reason you want to use your own databases, this is indeed possible. Select *Menu* on the top left corner and afterwards *Main Page* from the drop-down menu. Then follow the link *WSPP-SBT sample with user-provided* *databases* showing in the *Basic Workflow* field. In the *Databases* menu, both the *Select Species* and the *Select version of Uniprot category database* fields are now empty; but if you activate the *Rels. From Cat. File* option on the left, the *Protein database* and *Protein-category relation file* fields will display the path to a protein database and a protein-category relation file, respectively, stored in your hard drive (Fig. 3).

Figure 3. **The Basic Workflow Databases window in the WSPP-SBT sample with user-provided databases option.**

Figure 4. **The Basic Workflow Stats Quantification window.**

1. **Quantification of samples via the WSPP/SBT statistical model**

The first step towards evaluating differential protein abundance between our four WT and four KO samples is to have the WSPP (weighted spectrum, peptide and protein) statistical model and the SBT algorithm carry out the scan -> peptide, peptide -> protein and protein -> category integrations[[10]](#footnote-10) for every sample. This is achieved via the table shown in the *Stats Quantification* menu, where the following fields have been completed (Fig. 4):

* *Experiment*: once more, the experiment allocation (remember that all samples belong to the same experiment in this example);
* *Ratio numerator* specifies which sample quantitation values (each of the eight iTRAQ channels in our case) make up the numerator for the calculation of log2-ratio values with every PSM included in the *PSMs.txt* files. These log2 ratios will be used for the starting scan -> peptide integration;
* *Ratio denominator* specifies which sample quantitation values make up the denominator for the calculation of log2-ratio values with every PSM included in the *PSMs.txt* files. In our example, we use the average value of the four iTRAQ channels corresponding to WT samples (113-116; note the use of commas without spaces), meaning that each of the four WT and four KO individual samples will be compared with the average of the four WT samples.
* *Name* designates the integration name. The integration results will be saved to a folder with this name.

1. **Aggregate statistics: the *Integration* menu**

So far, we have instructed *iSanXoT* to compare each of the eight samples with the average of WT samples, which will enable us to observe, for instance, how differential protein abundance varies across samples. Now we can move further to attain a global comparison between the four KO and the four WT samples. For that, we shall use the *Integration* menu to:

1. Generate two sets of merged protein relation and protein data files: one corresponding to the four KO replicates and one from the four WT replicates;
2. Prepare the relation and data files corresponding to the KO/WT ratio and carry out the statistical calculations.

Step (i) is accomplished via the *NorCombine* submenu (Fig. 5):

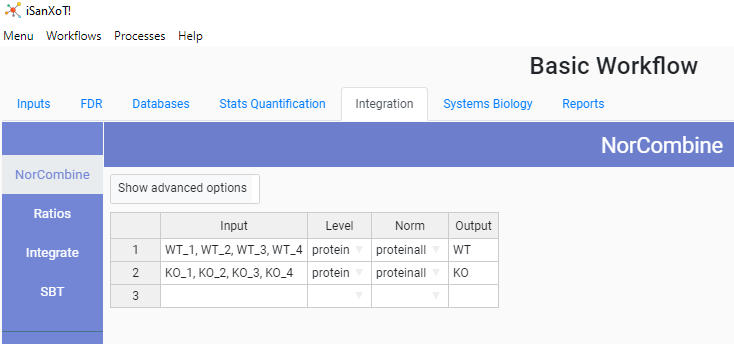
* *Input* indicates the names of the integration folders from which the merged protein relation and protein data files will be prepared (WT\_1-WT\_4 on one side and KO\_1-KO\_4 on the other side; note the use of commas without spaces);
* *Level* indicates whether the combination of relation and data files should be made at the peptide, protein (as is the case here) or the category level;

Figure 5. **The NorCombine submenu in the Basic Workflow Integration window.**

* *Norm* specifies the normalization scheme (“proteinall” here) to be used with *level*;
* *Output* is the name of the folder where the merged relation and data files will be saved. In our example we have a folder for the WT files and another folder for the KO files.

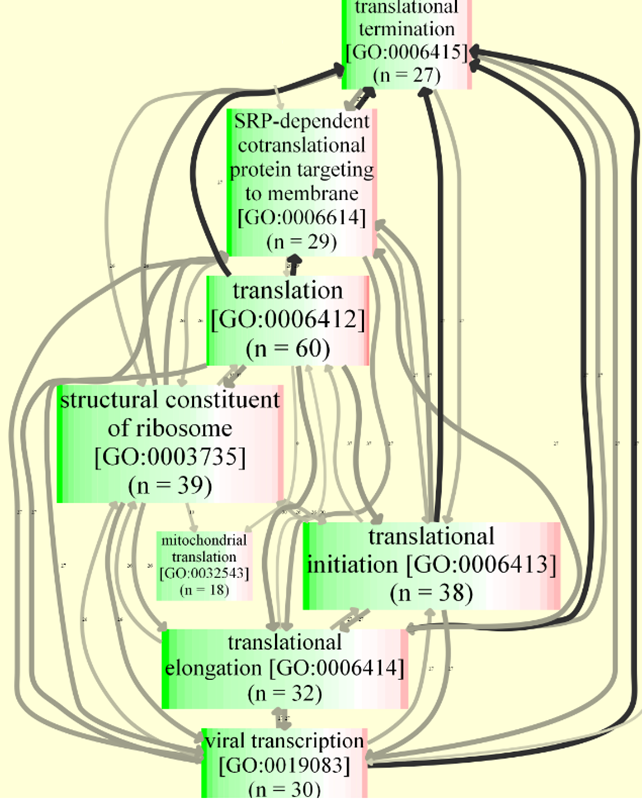
Step (ii) requires in the first place defining the ratio to be calculated via the *Ratios* submenu:

* *Ratio numerator* indicates which folder contains the files required to make up the numerator of the ratio to be calculated (“KO” in this case);
* *Ratio denominator* indicates which folder contains the files required to make up the denominator of the ratio to be calculated (“WT” in this case);
* *Level* specifies whether ratio calculation should be made at the peptide, protein (as is the case here) or the category level;
* *V method* determines the method used to calculate the statistical weight of the newly calculated log2-ratios. In this case we shall be using the maximum between KO and WT;
* *Output* is the name of the folder where the resulting log2-ratio and statistical weight values will be saved (“KO\_vs\_WT” in our example).

Finally, the SBT algorithm will calculate the variance associated to the protein -> category integration, which will be applied to the protein –> proteinall integration. Three more fields in the *SBT* submenu need be completed to that end:

* *Input* is the name of the folder containing the newly calculated, aggregate log2-ratio and statistical weight values (“KO\_vs\_WT” here);
* *Lower level* indicates the starting level for the first SBT integration: peptide, protein (as is the case here) or category;
* *Intermediate level* indicates the ending level for the first SBT integration: peptide, protein or, as in this case, category.

1. **Systems Biology tools: Sanson**

At this point, *iSanXoT* is ready to accurately quantify not only differential protein abundance, but also differential protein function. However, when dealing with functional categories such as the Gene Ontology[[11]](#footnote-11) (GO) terms retrieved from Uniprot by *iSanXoT* (see Section 5), redundancy needs be taken care of. The hierarchical structure of GO vocabularies allows GO terms to have multiple parents, or ascendants[[12]](#footnote-12). As a protein will be annotated by a term and every ancestor of this term, terms at a variety of depths in the hierarchy will appear when assessing differential protein function.

To help you deal with this redundancy issue and assist in the biological interpretation of results, a similarity graph can be generated via the *Sanson* submenu in the *Systems Biology* tab. The similarity graph will show relationships among categories, according to the number of protein components they share, for the set of categories found significantly altered in a given comparison (Fig. 6):

* *Input* specifies the comparison (see previous Section) for which the similarity graph will be created (“KO\_vs\_WT” in our case);

Figure 6. **Zoomed-in portion of a sample similarity graph.** Categories are shown as nodes, while the numbers next to the links indicate the number of protein components shared.

* *Lower level* indicates the type of shared component (“protein” in our example) to be evaluated;
* *Higher level* is the type of entity to be checked for redundancy and included as a node in the similarity graph. Since in this example we have chosen “category”, the graph will include categories that meet the following two conditions in the KO\_vs\_WT comparison:
  1. The category changes in the KO\_vs\_WT comparison based on FDRc < 0.05;
  2. Category size ranges between 10 and 100 protein components.

1. **Generating reports *à la carte***

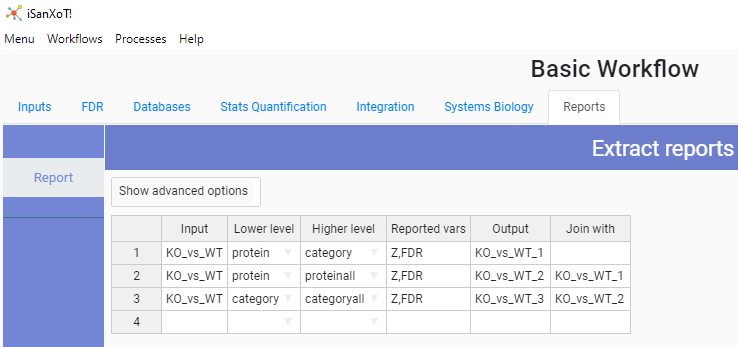
The statistical calculations set on Sections 6 and 7 will generate nine folders containing, among many others, nine files with Zq and FDRq values[[13]](#footnote-13) and nine files with Zc and FDRc values[[14]](#footnote-14) for the *c.a.* 7,000 proteins and 37,000 categories, respectively, that were quantitated in our sample experiment. To save you from the tedious task of browsing through hundreds of files to manually (and painfully) collect the relevant data into a table, the *Reports* menu (Fig. 7) allows the generation of nearly any type of report you may need. In our example, we sought to combine in a single table Z and FDR values for the following three integrations from the KO\_vs\_WT comparison: protein -> proteinall (which quantitates relative protein abundance), category -> categoryall (which quantitates relative category abundance) and protein -> category (which quantitates relative protein abundance within categories and is useful to collect category protein components):

Figure 7. **The Basic Workflow Reports window**.

* *Input* specifies the comparison from which the data will be retrieved (“KO\_vs\_WT” in this case);
* *Lower level* indicates the starting level (peptide, protein, category) for the integration whose data are to be reported;
* *Higher level* indicates the ending level for the integration whose data are to be reported;
* *Reported vars* specifies which statistical variables will be reported (“Z” and “FDR”13,14 in this example);
* *Output* is the report filename;
* *Join with* designates the file whose *Reported vars* will be incorporated into *Output* after intersection with the latter file.

Hence, we start preparing a report (“KO\_vs\_WT\_1”, see row 1 on Fig. 7) that displays category protein components (along with their Zqc and FDRqc values). Then “KO\_vs\_WT\_2” is created by *joining* relative protein abundance data (Zqa and FDRqa) with “KO\_vs\_WT\_1” (row 2); finally, the target report, “KO\_vs\_WT\_3”, is obtained by *joining* relative category abundance data (Zca and FDRca) with “KO\_vs\_WT\_2” (row 3).

1. Ross P.L. *et al*. (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents*. Mol. Cell. Proteomics* **3**, 1154–69. [↑](#footnote-ref-1)
2. Stein D.R. *et al*. (2013) High pH reversed-phase chromatography as a superior fractionation scheme compared to off-gel isoelectric focusing for complex proteome analysis. *Proteomics* **13**, 2956-66. [↑](#footnote-ref-2)
3. https://www.thermofisher.com/order/catalog/product/OPTON-30810#/OPTON-30810. [↑](#footnote-ref-3)
4. https://www.unimod.org/ [↑](#footnote-ref-4)
5. Eng J.K. *et al.* (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom*. **5**, 976–89. [↑](#footnote-ref-5)
6. Martínez-Bartolomé S. *et al*. (2008) Properties of average score distributions of SEQUEST: the probability ratio method. *Mol. Cell. Proteomics* **7**, 1135-45. [↑](#footnote-ref-6)
7. Bonzon-Kulichenko E. *et al*. (2015) Revisiting Peptide Identification by High-Accuracy Mass Spectrometry: Problems Associated with the Use of Narrow Mass Precursor Windows. *J. Proteome Res.* **14**, 700–10. [↑](#footnote-ref-7)
8. García-Marqués F. *et al*. (2016) A Novel Systems-Biology Algorithm for the Analysis of Coordinated Protein Responses Using Quantitative Proteomics. *Mol. Cell. Proteomics* **15**, 1740-60. [↑](#footnote-ref-8)
9. https://www.uniprot.org/. [↑](#footnote-ref-9)
10. Trevisan-Herraz M. *et al*. (2019) SanXoT: a modular and versatile package for the quantitative analysis of high-throughput proteomics experiments. *Bioinformatics*. **35**, 1594-96. [↑](#footnote-ref-10)
11. Gene Ontology Consortium (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res*. **49**, 325-34. [↑](#footnote-ref-11)
12. Ashburner M. *et al*. (2001) Creating the gene ontology resource: Design and implementation. Genome Res. 11, 1425-33. [↑](#footnote-ref-12)
13. Zq values are mean-corrected protein log2-ratios expressed in units of standard deviation. FDRq determines which Zq values stand out as outliers in the protein quantification data. [↑](#footnote-ref-13)
14. Zc values are mean-corrected category log2-ratios expressed in units of standard deviation. FDRc determines which Zc values stand out as outliers in the category quantification data. [↑](#footnote-ref-14)