Mapping relationships between glutathione and SLC25 transporters in cancers using hybrid machine learning models

Luke Kennedy[1,2], Jagdeep K Sandhu[1,3], Mary-Ellen Harper[1,2] *, and Miroslava Cuperlovic-Culf[1,4] *

*Co-senior authors

Affiliations

[1]Department of Biochemistry, Microbiology and Immunology, Faculty of Medicine University of Ottawa, 451 Smyth Road, Ottawa, ON K1H 8M5, Canada.

[2]Ottawa Institute of Systems Biology, University of Ottawa, 451 Smyth Road, Ottawa, ON K1H 8M5, Canada.

[3]Human Health Therapeutics Research Centre, Bldg M54, National Research Council Canada, 1200 Montreal Road, Ottawa, ON K1A 0R6, Canada.

[4]Digital Technologies Research Centre, Bldg M50, National Research Council Canada, 1200 Montreal Road, Ottawa, ON K1A 0R6, Canada.

Correspondence:

Mary-Ellen Harper **mharper@uottawa.ca** or Miroslava Cuperlovic-Culf **miroslava.cuperlovic-culf@nrc-cnrc.gc.ca**

Abstract

Alterations of metabolism are a well appreciated hallmark of many cancers, including changes in mitochondrial and glutathione (GSH) metabolism. Mitochondrial GSH (mGSH) transport is a poorly characterized aspect of GSH metabolism, which we investigate in the context of cancer. Here, we apply hybrid machine learning (ML) models and leverage cancer cell line multi-omics data and other biological knowledge data as features, to uncover potential candidates involved in mGSH metabolism and membrane transport in cancers. These models predict the known mGSH transporter SLC25A39 but not SLC25A40 as being highly probably related to GSH metabolism in cancers. Additionally, SLC25A24 and SLC25A37 along with the poorly characterized members of the SLC25 family, SLC25A43 and SLC25A50, are predicted to be associated. Similarities in potential substrate binding regions are found between SLC25A39 and the candidates SLC25A24 and A43. These findings have implications for a better understanding of cancer cell metabolism and novel therapeutic targets.

Keywords: Mitochondria, glutathione, machine learning, cancer, protein function annotation, transmembrane transport, SLC25, multi-omics, knowledge-based, Gene Ontology

Introduction

Glutathione (GSH) is a highly abundant tripeptide antioxidant within cells, crucial for many biological processes with its major role in regulation of reactive oxygen species (ROS). Metabolic changes are one of the hallmarks of cancer[1], with well documented alterations in GSH metabolism foremost among them. These alterations of metabolism appear to benefit cancer cells, aiding in tumor proliferation and survival, through not yet fully understood mechanisms and interactions[2]. Direct alterations of GSH metabolism are essential for tumor proliferation in several cancers[3,4] where they mitigate perturbations in redox homeostasis[5,6]. Alternatively, GSH can also modulate ferroptosis[7] and the uptake and metabolism of xenobiotics, including chemotherapeutics[7,8].

Central to the contributions of GSH in the metabolic characteristics of cancer cells is mitochondrial GSH (mGSH), which is found at millimolar concentrations in the organelle, at similar levels to those in the cytoplasm. Mitochondria are the primary source of ROS as a byproduct of oxidative reactions including those in the oxidative phosphorylation system (OXPHOS), and alterations in redox balance and ROS signaling are central to the important role of mitochondria in cancer cell proliferation[9]. Many GSH-utilizing enzymes with altered expression patterns in cancers are found in mitochondria, most notably glutathione peroxidase 4 (GPX4) which is a major regulator of the ferroptosis pathway. Also important is mitochondrial glutathione S-transferase (GST), which conjugates and detoxifies xenobiotics, including chemotherapeutic compounds. Reduction of ROS by GSH with mitochondrial peroxiredoxins (PRDXs) have also been shown to aid cancer cell survival[10].

The processes involved in the uptake of GSH into mitochondria are poorly understood. While high concentrations of GSH are present in both mitochondria and the cytosol, the synthetic enzymes are exclusively within the cytosol[11]. SLC25A10 and SLC25A11, also known as the mitochondrial dicarboxylate and oxoglutarate carriers, respectively, were initially proposed as the proteins responsible for GSH transport into mitochondria in 1998[12]. However, detailed functional experiments in lipid vesicle systems in 2014 convincingly showed that no GSH transport was mediated by these two proteins[13].

Recent evidence from two groups demonstrated mGSH transport roles for SLC25A39 and SLC25A40[14,15]. Specifically, Wang *et al.*[14] identified the transporters through quantitative proteomics of mitochondria from GSH-depleted HeLa cells through depletion of GSH synthetic enzymes or through pharmacological inhibition of GSH synthesis through buthionine sulfoximine. Alternatively, Shi *et al*.[15] leveraged CRISPR screening of gene and environment interactions to demonstrate buffering interactions between SLC25A39 and the mitochondrial iron transporter SLC25A37, revealing SLC25A39 as a candidate. Metabolomics of isolated mitochondria from SLC25A39 KO cell lines identified significantly decreased levels of GSH. Both groups observed decreased mGSH import through transport assays in isolated mitochondria of A39 knockout cells. Additionally, Wang *et al.* report that overexpression of A40 can recover mGSH levels in A39 knockouts[14]. The roles of these transporters in neurodegenerative diseases and cancer have recently been explored[16]. However, it remains unclear if the remaining GSH import in A39 knockouts observed in these studies is facilitated by SLC25A40, which is expressed at roughly one tenth the level of SLC25A39, or if other transport mechanisms exist, such as those for other GSH species[17,18]. Interactions affecting mGSH metabolism and transport, having secondary effects on transport may also be relevant and are even less well understood.

With recent successes of AlphaFold[19] and RosettaFold[20] in the elucidation of protein structure, there is a promising future for computational biology in the related problem of protein function prediction. However, competitions such as Critical Assessment of Functional Annotation (CAFA)[21] for function annotation have not yet found a solution for *de novo* function prediction from sequence. General function annotation models like DeepGO and DeepGOPlus[22,23] represent major advances for the field. Along with sequence-based methods, models which leverage non-sequence features for function annotation exist[24,25] as alternative approaches, yet are less common than sequence-based approaches. In addition to the general function prediction models, there are many sequence-based models designed for specific protein feature annotation such as protein-protein interactions[26] and antibody design[27].

Omics-based methods for function annotation are limited. Recent work by Kunc and Kléma[28] utilize characteristics in co-expression networks constructed from genes to predict shared KEGG pathways between genes. Similarly, Wekesa *et al.*[29] combine differential gene expression information and knowledge of protein-protein interactions through a neighbor-voting algorithm for prediction of shared functions between yeast proteins. Finally, Wang *et al*.[30] predicted gene-gene interactions by combining co-expression features with prior biological knowledge features (e.g., subcellular localization, homology, Reactome similarity).

The approaches of Wekesa *et al.* and Wang *et al*. can be classified as hybrid ML approaches, which aim to capture "the best of both worlds" by combining the predictive powers of data-driven approaches of ML with the interpretability of theory-based models such as mechanistic or knowledge-based models. These types of models have been applied to several domains, most notably in physics-informed ML models[31], and with respect to computational biology have been applied through several frameworks. ML models have been integrated with mechanistic models of metabolism to determine kinetic parameters and predict downstream metabolic effects[32]. While P-net uses a neural network architecture based on biological knowledge of hierarchical gene-pathway-process interactions to predict prostate cancer discovery from gene features such as methylation, and copy number[33]. Alternatively, AlphaFold[19] incorporates evolutionary information through multiple sequence alignments and physical structural constraints to inform its protein structure predictions. These hybrid models retain the predictive power of data-driven methods, often performing better than or comparable to standard data-driven models, while increasing the interpretability of predictions which are based-on, at least in part, understandable biological knowledge.

With this, we sought to develop a ML framework that leverages multi-omics data from human cancer cells and existing knowledge to identify potential genes involved in mGSH transport and potential interacting metabolic processes. Several ML classifier models were developed which utilize cancer cell line encyclopedia (CCLE)[34] transcriptomics features to predict gene ontology (GO) annotations for genes of relevant GO terms. Specifically, three classifiers were developed for annotating to: glutathione

metabolic process (GO:0006749), mitochondrion (GO:0005739), and transmembrane transporter activity (GO:0022857). Additionally, we developed hybrid models for this task which include prior knowledge as features through MitoCarta[35] scores for mitochondrial localization classification, TrSSP[36] scores for transporter classification, or CCLE GSH & glutathione disulfide (GSSG) metabolomics data for glutathione classification. We find that Random Forest (RF) classifiers perform the best from our models tested, with hybrid models outperforming strictly transcriptomics models. The GSH classifier identifies SLC25A37, SLC25A43, and SLC25A50 as possible novel members of the mitochondrial carrier family associated with GSH transport in cancer cells. Furthermore, structural comparisons reveal similarities between the transmembrane amino acid sequences of candidates SLC25A24, SLC25A43 and the known mGSH transporter SLC25A39, which may imply similarities in transported substrates.

Methods

All pre-processing, models, and other computational work were conducted using Python (Python Software Foundation. Python Language Reference, version 3.8. Available at www.python.org). All algorithms and methods used in ML model building and training were implemented via Python's scikit-learn package[37] unless stated otherwise. Figures for evaluations of classifier models were produced via the matplotlib and seaborn libraries[38,39], GO enrichment plots were produced using ShinyGO 0.77[40] and all other images were created with BioRender.com.

Data collection & pre-processing

Metabolomic & transcriptomic CCLE data

The Broad Institute's Cancer Cell Line Encyclopedia (CCLE) provides multi-omics data across over 1,000 human cancer cell lines (CCLs). CCLE transcriptomics, and metabolomics data were used in building out models[34,41]. Raw transcriptomics and metabolomics data were downloaded from broadinstitute.org (available at the time of writing via depmap.org/portal/download/).

Data were cleaned and imputed by removing all CCLs and transcripts/metabolites with >30% missing values or standard deviation values of 0 across samples. Remaining missing values were imputed using the k-nearest neighbors (k-NN) imputation algorithm (using Euclidean distance with k=5 selected to minimize impact of the data structure on the missing data imputation[42]). For metabolomics features, Spearman correlations were generated across common CCLs between GSH, GSSG metabolomics data and transcripts in CCLE transcriptomics data. Non-significant correlations (p-values greater than or equal to 0.05, student's t-test with 2 degrees of freedom) were set to 0. Spearman rho values were normalized by Fisher transformation[43].

Feature space reduction while selecting major variances in CCLE transcriptomics was performed using principal component analysis (PCA). These principal components (PCs) were used as features in classifier models (further described in Model development and feature selection process).

Knowledge-based features

Mitochondrial localization scores for genes were based on MitoCarta 3.0[35] (broadinstitute.org/mitocarta). MitoCarta scores are determined through a naive-Bayes model which combines features from several independent domains from features used in our analysis. Localization scores are predicted from features including homology, sequence domain, and tandem mass spectrometry of purified mitochondria to predict mitochondrial localization of genes. CCLE genes were assigned MitoCarta scores according to these data, with 1 indicating mitochondrial localization, 0 indicating non-mitochondrial localization. For genes which are not assigned scores in MitoCarta, a value of -1 is assigned.

Transporter activity was based on TrSSP[36] values, obtained from Zhao lab webpage (www.zhaolab.org/TrSSP). TrSSP utilizes support vector machine models to predict membrane transport proteins and their substrate classes. The models use primary protein sequence features, such as amino acid and biochemical composition, along with position-specific scoring matrices to predict transport function and specificity. Like

MitoCarta scores, TrSSP scores were incorporated into the model as classifications for positive, negative, or no prediction by TrSSP (1/0/-1).

For use as features in our classifiers, these knowledge-based categorical scores were transformed using one-hot encoding. Each categorical feature is converted to multiple separate Boolean features which encode whether a category (e.g., positive TrSSP prediction) is true or false for a given sample (gene).

mGSH transporter classifier models

Training and test gene sets selection

For each classifier, training and test genes were selected based on existing knowledge of gene functions. Gene ontology[44] terms were used to determine genes related to mGSH metabolism: glutathione metabolic process GO:0006749 (65 genes), mitochondrion GO:0005739 (1685 genes), transmembrane transporter activity GO:0022857 (1154 genes). These terms cover all classes of ontology terms: biological process, cellular component, and molecular function, respectively. Annotated genes for each term were identified via AmiGO 2[45] and those genes annotated based on only computational or inferred evidence were removed. This left 40 genes for GO:0006749, 781 genes for GO:0005739, and 700 genes for GO:0022857 annotated based on experimental evidence to create the final list of positive class genes for the classifier models. There is little overlap among these gene sets, primarily, overlapping genes are found between mitochondrial and transmembrane transporter activity genes (72 genes), with 5 genes overlapping the mitochondrial and glutathione metabolic process sets and only one gene found in both transmembrane transporter activity and glutathione metabolic process genes. There are no genes annotated by all three GO terms. AmiGO annotations based on sequence similarities were retained along with the experimental AmiGO annotations used for the other GO annotations to generate the 40 genes used in the final positive gene set for glutathione metabolic process term. These annotations were retained due to the small number of genes annotated by experimental evidence for the GO term, thus creating a larger positive gene set for model training. bioDBnet[46] was used to link gene symbols from AmiGO to Ensembl gene IDs found in the CCLE data.

The negative classification set of genes was generated by randomly undersampling from genes that were not included in the positive set based on GO annotations. By undersampling, we are ensuring that positive and negative genes (samples) for model training and testing are balanced. However, a model trained through this process will be biased by the specific set of randomly selected negative class genes in the training set. Additionally, we are sampling from unlabeled data, thus, it is possible that some genes which are selected for the negative class belong to one of the positive GO terms. To minimize the effects of these limitations and ensure robustness of the models, 100 iterations of random sampling were performed to create different sets of samples for model training. Final predictions and evaluations are the average results over training and testing iterations on the 100 different data sets.

Model development and feature selection process

To determine the best methodology for our task several common ML algorithms and feature sets were tested. Random Forest (RF), Decision Tree (DT), Naive Bayes (NB), and Support Vector Machine (SVM) were applied for classification tasks of mitochondrial, transmembrane transporter activity, and GSH metabolic process GO terms. Models were tested with several sets of features which combine experimental features from CCLE data with knowledge-based features, or feature sets containing only experimental features. Experimental feature sets include the first 5, 14, 30, or 50 CCLE transcriptomics PCs which capture 93%, 95%, 97%, or 98% of the variance, respectively. For classifying for the GSH metabolic process GO term, Spearman correlations values between gene transcriptomics and GSH or GSSG metabolomics were included as features as well. Features from existing biological knowledge were included through MitoCarta scores for mitochondrial localization in the mitochondrial classification task, and TrSSP prediction scores for membrane transport proteins in the transporter classification task.

In contrast, models are also trained using only one source of experimental data as features. Models which classify GO terms using only transcriptomics PCs as features allow for comparison with those models which also include metabolomics features (for

glutathione classification), or existing knowledge features (for mitochondrial and transporter classification). For valid comparisons, either metabolomics or knowledge features were replaced with additional PC features to create the transcriptomics-only models with an equivalent number of features (i.e., the mitochondrial classifier with the first 5 PCs and MitoCarta scores as features instead uses the first 6 PCs as features in the transcriptomics-only model).

Model evaluation

Model performance was analyzed from predicted class probabilities using common evaluation metrics including receiver operating characteristic (ROC) and precision-recall curves (PRC). These methods provide a balanced approach to evaluating classifier performance which considers the effects of true positives (TPs), true negative (TNs), false positives (FPs), and false negatives (FPs). Models were evaluated and selected based on values for the areas under ROC (AUROC) and PRC (AUPRC), and Matthew's correlation coefficient (MCC)[47,48] (reported in **Table** 1). These evaluation metrics were calculated across all test-set predictions from five-fold cross-validations of randomly sampled training/test sets. Reported metrics are calculated as average values across 100 iterations of training/testing using the different gene sets.

Protein structural analysis

SLC25 structures

Protein structures are obtained from the AlphaFold[19] structure database (alphafold.ebi.ac.uk) for use in structural analyses. Additionally, homology models from SWISS-MODEL[49] for both SLC25A39 and A40 are included in our analyses. Homology based structures are modeled on the crystal structure of the bovine ADP/ATP carrier (PDB:OKC1)[50] and cryo-electron microscopy determined structure of human UCP1 (PDB:8HBV)[51] for A39 and A40, respectively.

Protein structure alignments

Comparisons of 3D protein structures were conducted using sequence-independent alignment through the TM-align algorithm[52] and quantified by the corresponding TM-scores which quantifies similarities through similarities in protein structure topology. This method performs comparably to other common alignment methods and does not consider sequence similarity for alignment, which could skew the alignments due to conserved sequences within the SLC25 family. Alignments were implemented using the TMalign module for PyMOL[53] and predicted mGSH transporter structures were compared to known mGSH transporters (SLC25A39, A40).

The CAVER PyMOL plugin version 3.0.3[54] was used to identify tunnel residues for each SLC25 structure. Tunnels were identified by initializing the tunneling at the base or innermost point of the transporters tunnel with maximum starting point distance and desired radius of 3Å for starting point optimization. From this base, the tunneling algorithm probed outward to generate an interior tunnel channel. CAVER default parameters were specified for the tunneling algorithm, using 12 approximating balls with minimum probe radius of 1.5Å, shell depth of 20Å, shell radius of 7Å (or 10Å for structures where a smaller radius fails to fill the tunnel space), and a clustering threshold of 3.5. Default parameters were used for computation memory and speed, and were adjusted as described, if necessary, for finding transporter tunnels. Relevant tunnels were identified from the top-ranked tunnels identified by CAVER as those tunnels which traversed and filled the length of the proteins' tunnel without escaping to the protein surface through a side gap or by "overflowing" the interior tunnel.

In the same fashion as the full structure alignments, analysis of the tunnel regions for the SLC25 family were performed by aligning the tunnel-interacting residues identified from CAVER using TM-align.

Results

Classifier models for mGSH transporters in cancer

To identify top mGSH transporter candidates, three classifier models were developed to classify genes based on our desired candidate characteristics: mitochondrial

localization, association with GSH metabolism, and transmembrane transport function (see **Fig.** 1.a. for overview and **Methods** for details). Each model uses features from CCLE transcriptomics data along with one or more features from other sources; GSH and GSSG metabolomics correlations for GSH metabolic process classification, MitoCarta scores for mitochondrial localization classification, and TrSSP scores for transmembrane transporter activity classification. (**Fig.** 1. b.).

Our classifiers performed well in the tasks of mitochondrial, GSH, and transporter annotation classification across our evaluation metrics, as shown in **Table 1**. Across all models and feature selections, mean five-fold cross-validations shows AUROC is 0.820, and RF classifiers being the best performer with mean AUROC values of 0.900. Classifier performance varied across our three classifier types, with mitochondrial classifiers performing the best and glutathione classifiers performing the worst, on average, which may be attributed to two factors. First, the set of annotated genes related to GSH metabolism is much smaller than the gene sets for the other classifiers (~40 genes vs. ~700 genes), thus providing a smaller training set for these models. Additionally, the GSH classifiers rely solely on CCLE multi-omics data unlike the other two classifiers which also include data from existing knowledge-sources.

Along with testing several ML algorithms in our classifiers, multiple feature sets were explored to determine the best selection for our task. Specifically, variations of the number of top PCs as linear combinations of transcriptomic features (described in Methods) from 5-50, were tested. Increasing the number of these features in our models did not notably improve performance by our metrics (**Supp. Fig.** 1). Based on these results, a small number of the first several PCs are sufficient for effective classification. However, specific gene classification probabilities are variable across models with different numbers of PC features used. Since performance is similar across these models, rather than selecting a specific model we used mean gene classification probabilities with standard error (SE) calculated from models with different numbers of PC features (5, 14, 30, 50) for final classification results.

Comparison of hybrid models and transcriptomics-based models

Consistently across all tested models, classifiers that incorporate both experimental data from CCLE transcriptomics data and data from other knowledge bases as features outperform those models with an equivalent number of experimental data features alone (**Fig**. 1.c.,d., **Table 1**). Although no knowledge component exists for the GSH classifiers, transcriptomics-based models were compared to models which incorporate GSH and GSSG metabolomics data from the CCLE along with the CCLE transcriptomics data for features. Interestingly, the additional data source does not seem to similarly improve performance in the GSH models, with performance of transcriptomics-only GSH classifiers comparable to GSH classifiers using both transcriptomic and metabolomic features (**Fig.** 1.e.).

Comparison of hybrid models to knowledge-based models

Similar to the previous analysis, we sought to evaluate the effects of adding transcriptomics information to the existing predictive models from the literature. Classifications from our mitochondrial and transporter models were compared to those of MitoCarta and TrSSP. The combined transcriptomic and MitoCarta/TrSSP classifiers appear to only show slight differences (**Fig.** 2.a.,b.). The mitochondrial classifier identifies slightly more mitochondrial genes than MitoCarta alone, however there are more false positives (FPs) in this classifier (at a classification threshold of 0.75).

False positive mitochondrial and transporter genes

Interestingly, our classifiers predict more false positives compared to the results of MitoCarta and TrSSP, suggesting a potential for novel functions. To assess this possibility, mitochondrial FP genes were compared to other sources of mitochondrial genes beyond the high confidence GO annotations used for classifier training. Low confidence GO annotations which were originally removed from the training set, Human Protein Atlas subcellular localization annotations from histology images[55], Swiss-Prot annotations[56], and integrated mitochondrial protein index (IMPI) annotations based on MitoMiner[57] were used for other sources of localization evidence. Of the 844 FP genes predicted by the mitochondrial RF classifier, 615 were predicted by both MitoCarta and RF, and 229 were predicted by RF only (**Fig.** 2.c.). 206 of these RF FP genes have no

evidence in the other datasets, while the remaining have evidence in at least one source (**Supp. Data**). Enrichment analysis revealed that the 206 unannotated FPs without existing evidence are found to be significantly enriched in KEGG pathways (**Fig.** 2.e.) and GO terms (**Fig.** 2.f.-h.) associated with degradation pathways, such as the proteasome, RNA stability, degradation and processing pathways, and are furthermore highly associated with one another based on high confidence based evidence through STRING[58] (**Supp. Fig.** 4). The proteasome system is closely linked to mitochondria remodeling in several physiological and diseased states, including cancer[59,60].

Comparisons of FPs in TrSSP and our transporter RF classifier find less overlap between the two (**Fig.** 2.d.). 147 FPs are exclusive to TrSSP, 218 are exclusive to the RF classifier, and 1477 are common to both models. In the same fashion as mitochondrial genes, other evidence sources beyond the high confidence GO term annotations were compared to the FPs predicted by transporter models. Low confidence AMiGO annotations, Transporter Classifier Data Base (TCDB) annotations[61], and the SLCAtlas[62] were used as alternate evidence sources. Swiss-Prot annotations were excluded as these were used in TrSSP model training. For both TrSSP and our RF classifier, only a small number of FP genes have evidence from other sources (2 for TrSSP, and 8 for the RF classifier, **Supp. Data**). Unlike the mitochondrial FPs by our RF classifier, transporter FP genes are not extensively enriched in specific GO terms (**Supp. Fig.** 5), and no enrichment of KEGG pathways is observed. The most apparent enrichment is in biological process terms related to cell-cell interactions, such as cell adhesion, cell junctions, and synapses (**Supp. Fig** 5.c.).

mGSH transporter candidates

The results of classifiers identify several potential mGSH transporters. Within the SLC25 family, the known transporter SLC25A39 is among the most probable GSH-related genes (**Table 2**), and within range of the most probable candidate when considering SE values for probabilities. Surprisingly, SLC25A40, the homolog of SLC25A39, has a very low GSH probability by the RF classifier models and is not predicted to be related to GSH. Amongst the top candidates are several well characterized proteins such as

SLC25A1, SLC25A10, and A11 as well as some orphan members like SLC25A43 and A50. To understand the possible roles of these members in GSH transport, existing evidence was considered and described below.

Validation of metabolite classification method

Several other metabolite classifier models were developed for comparison and validation of the mGSH transporter candidates. Like the GSH classifiers, RF models with 5-50 CCLE transcriptomic PCs and corresponding metabolomic correlations as features were trained for the annotation of genes related to 2-oxoglutarate, glutamate, and carnitine. As with the GSH classifier, relevant genes were identified through high confidence GO annotation based on experimental and sequence similarity evidence: GO:0006103 2-oxoglutarate metabolic process, GO:0006536 glutamate metabolic process, and GO:0009437 carnitine metabolic process. Training/test sets were constructed from these genes and an equal weight of randomly selected remaining genes is included for the negative class. 2-oxoglutarate, glutamate, and carnitine were selected for this evaluation because, out of metabolites found in the CCLE metabolomics set, they have a relatively larger number of annotated genes, making classifier training possible. Additionally, there is evidence for transport of each by one or more SLC25 proteins which allows for some measure of classifier validation. For each of the other metabolite classifiers, performance was similar to the GSH classifier by our evaluation metrics (**Supp. Table** 1).

non-SLC25 mGSH-related proteins

Alongside the SLC25 family, the remaining genes in our dataset were considered for roles in mGSH transporter. Top candidates were identified by their mean GSH probability across RF classifiers. Additionally, the mean mitochondrial and transporter classifier scores were also considered. To identify high confidence candidates, only genes with probabilities greater than 0.80 across all classifications are selected (Genes already annotated with one of our GO terms are said to have probabilities of 1). This thresholding removes 99.95% of the available 49309 genes in the datasets, leaving only 27 genes which meet our criteria (**Supp. Table** 2). Of these, 4 genes are already

annotated with the glutathione metabolic process GO term, 13 are annotated with the mitochondrion GO term, and 5 with the transmembrane transporter activity term.

Structural comparisons

To further examine the SLC25 proteins most probably relevant to GSH transport, we then conducted structural analyses. Experimental determinations of protein structures for the SLC25 family are currently lacking. The first human SLC25 structure was recently determined for UCP1 by cryo-electron microscopy[51]. Beyond this, structures for orthologs in model species have been determined through x-ray crystallography for the bovine[50] and *Thermothelomyces thermophila*[63] ADP/ATP carriers (SLC25A4), and by NMR molecular fragment searching for the yeast UCP2 (SLC25A8)[64]. Predicted protein structures must be used for comparison of family members with unknown structures. Particularly, we were interested in structural comparisons between our candidates with the GSH transporters SLC25A39, and SLC25A40 in order to reveal any similarities in structure and potential function. Structural alignments were determined using TM-align (see Methods) for pairs of predicted 3D structures ("global" alignments) and pairs of tunnel-interacting residues within SLC25 structures ("local" alignments). The residues for local alignments were identified via CAVER[54] (**Supp. Fig.** 6). Rather than considering the entire transporter structure, of more relevance to transporter function and specificity is the interior tunnel of the protein surface, along which substrates interact and move across the IMM[63,65]. By aligning only potentially substrate binding residues, i.e., the interior tunnel residues identified by CAVER, we expected a more relevant analysis of the SLC25 family with respect to substrate binding as these structures are expected to be more variable and better indicators of shared transport activities.

Pairwise global alignments showed high similarities (TM-scores close to 1) across the transporter family (**Supp. Data**). This is expected due to common structural domains across the family, namely the six transmembrane helices. CAVER-identified tunnel-interacting residues of the SLC25 family were aligned via TM-align. Local TM-scores are much lower relative to the global alignment scores (mean scores excluding self-

alignments 0.376 vs. 0.701) (**Supp. Fig.** 7). Furthermore, local alignments are more variable (coefficient of variation values of 0.350 vs. 0.172 for global alignments), indicating the diversity of transport tunnels, and thus transport functions of the SLC25 proteins. **Figure** 3.a. details the TM-scores for local pairwise alignments between top mGSH candidate SLC25s as well as SLC25A39 and SLC25A40.

Amongst these results, relatively high scores between SLC25A24, SLC25A43, and SLC25A39 indicate similarities in transporter tunnels and possible overlapping transport functions of the proteins. Shi *et al*.[15] identify R225, D226, and K329 as important residues for GSH transport in SLC25A39, based on local structural alignments of tunnel regions with SLC25A43 and SLC25A24 (**Fig.** 3.b. & c.), the position of the K329 residue is conserved by K275 in SLC25A43 and by K453 in SLC25A24. However, the R225, D226 motif of SLC25A39 appears less conserved in the aligned structures. In the SLC25A43 alignment, residues G173 and A174 are in closest proximity to the A39 binding residues, and these non-polar residues are not identified by CAVER as tunnel-interacting residues. Interestingly, the sidechains of SLC25A43 residues K226 and R227 appear near R225 and D226 sidechains in SLC25A39, where the positively charged K226 replaces the negative charge of D226 in SLC25A39. Like SLC25A43, the closest aligned residues of SLC25A24 to the GSH binding residues of SLC25A39 are the non-polar G353 and I354, which do not interact with the transporter tunnel. Where K226 and R227 are found in SLC25A43, instead there is R410 and T411 in SLC25A24.

Comparison to DeepGOPlus

For a comparison to the models developed here, genes from our GO terms of interest (GO:0006749 glutathione metabolic process, GO:0005739 mitochondrion, and GO:0022857 transmembrane transporter activity) were annotated by DeepGOPlus. DeepGOPlus[23] is a general function annotation deep learning model based on protein amino acid sequences. The model is constructed as a convolutional neural network (CNN) in which protein sequences are taken as inputs and the predicted GO term annotation and probabilities are returned. Gene ontology terms with predicted annotation probabilities were obtained through the DeepGOWeb API[66] (Version 1.0.13)

using FASTA sequences obtained from canonical Ensembl transcript IDs for the genes under each ontology term. All 39 high confidence GO:0006749 annotations, and 150 randomly selected genes from each other ontology term were annotated using DeepGOWeb. The true positive rates (TPRs) or sensitivity measures of annotations were determined for model comparison (**Fig.** 4, **Supp. Table** 3).

Our classifiers perform better than DeepGOPlus for annotating all GO terms of interest. This is most evident for the set of genes involved in glutathione metabolism, where DeepGOPlus fails to annotate this set of genes with an average TPR of only 0.073 over classification thresholds compared to 0.867 for the RF classifier. DeepGOPlus can reasonably classify mitochondrial localization and transporter activity annotations with mean TPRs of 0.525 and 0.722, respectively. The vast differences in performance are likely due to the size of the gene sets, with the general annotation model performing poorly on the relatively small set of glutathione metabolism genes but much better on the, much larger, mitochondrial and transporter gene sets. This indicates that while general function annotation models are powerful, specific models leveraging additional knowledge are still necessary for any level of confident annotation for underrepresented GO terms.

Discussion

The importance of mitochondria and GSH in cancers is widely appreciated. Mitochondria have a major role in metabolic shifts, one of the hallmarks of cancer, while mGSH is important in metabolism and ROS processing, as well as post-translation proteins modifications and tumor drug resistance. An improved understanding of mitochondrial transport mechanisms in cancer cells, including GSH transport mechanisms is important for advancing the cancer biology field. To this end we developed a hybrid ML framework that utilizes multi-omics data and existing knowledge for annotating genes with GO terms of interest. We applied this model to the problem of mitochondrial GSH metabolism, specifically focusing on the aspect of transport. We propose several potential candidates that are predicted to be related to, and possibly

modulated by, mGSH transport identified by this model to guide future experimentation in this area of research.

Random forest classifiers were developed here for the annotation of mGSH transporter characteristics: mitochondrial localization, relation to GSH metabolism, and transporter function. Models combine CCLE transcriptomics and metabolomics data for annotation of genes related to GSH metabolism, and knowledge from existing predictive models for mitochondrial localization and transporter activity annotation. This framework resulted in models with high performance (mean RF AUROC of 0.900, **Table** 1) across each classification task.

For comparison, we attempted a similar annotation classification using the recent state-of-the-art sequence-based DeepGOPlus[23,66] for annotation of our terms of interest. The general GO annotation model performed worse, as measured by sensitivity in correctly classifying annotated genes, than our RF classifier (**Fig.** 4.). This is most evident in the GSH metabolic process term classification, possibly due to the relatively small set of annotated genes for this term which is underrepresented in the general models training. Furthermore, as a sequence-based approach, DeepGOPlus cannot be applied to disease contexts as we do with our proposed method, in this case using cancer cell transcriptomics and metabolomics data to focus analysis on the cancer context. This highlights the benefits of leveraging high-throughput experimental datasets and the additional value of incorporating a diverse set of knowledge bases in function annotation models, which notably improves model performance (mean RF AUROC of 0.838 in transcriptomics-only models, **Table** 1).

As a further validation of our methodology, several other classifiers using our framework were developed and tested for annotating other GO terms. Identical RF models to the GSH classifier were developed which instead use genes annotated with GO terms for glutamate, 2-oxoglutarate, and carnitine metabolic process for training. These models show similar performance to the GSH classifier (**Supp. Table** 1) despite a small number of annotated genes for training. Furthermore, with respect to our interest in transport,

these classifiers correctly identify many previously known SLC25 transporters as their most probable candidates.

Looking at the family of mitochondrial carriers, the SLC25 proteins, we find several predicted to be related to GSH metabolic processes. Most notably, SLC25A39, the recently identified GSH transporter is amongst our top hits (ranked 4th with a GSH metabolic process probability of 0.772 ± 0.023). Furthermore, several SLC25 which are known to be relevant to mGSH metabolism are amongst the top candidates: SLC25A1[67,68] (ranked 1st, probability of 0.789 ± 0.017), SLC25A10[69,70] (ranked 2nd, probability of 0.780 ± 0.026), SLC25A13[71] (ranked 3rd, probability of 0.774 ± 0.016), and the iron transporter SLC25A37[15] (ranked 8th, probability of 0.749 ± 0.026). Surprisingly, the other novel GSH transporter, SLC25A40, is not predicted to be relevant, with a GSH metabolic process probability of only 0.470 ± 0.015 (ranked 37th of the 53 SLC25 family proteins). This may be due to the very low expression of SLC25A40 across the CCLE cell lines (**Supp. Fig.** 2), suggesting that SLC25A40 is less relevant to mGSH transport in cancer cells. Other potential mitochondrial carriers of interest include SLC25A43 (ranked 6th), SLC25A24 (ranked 7th), and the OMM transporters SLC25A50 (ranked 5th) and SLC25AA49 (ranked 12th). SLC25A43 is an orphan transporter which has recently been found to affect redox homeostasis[72], but the other candidates have not been, to our knowledge, associated with GSH metabolism.

Structural similarities in the candidates, specifically within the tunnel region of the proteins, can be used to aid in uncovering the potential roles of these transporters in mGSH metabolism. We investigated this possibility using predicted structures for the SLC25 proteins and the structure-independent alignment method, TMalign. Through this, we find that the known GSH transporter, SLC25A39, shows structural similarities to our candidates SLC25A43, and SLC25A24. Important residues for GSH binding in SLC25A39 are somewhat conserved in our candidates, specifically, SLC25A43 shows a shift in residue position within the transporter tunnel and a substitution of an aspartate to lysine residues, whereas SLC25A24 substitutes a threonine. These findings indicate that despite overall similarities in transporter tunnels between the GSH transporter and

our candidates SLC25A43 and SLC25A24, substrate binding and transport mechanism is likely somewhat different.

Expanding our search beyond the SLC25 family, 27 genes were classified with high probability (>0.80) for all classification tasks (**Supp. Table** 5). Of these, some of the most probable candidates are discussed here for their possible roles in mGSH metabolism and transport.

Pyruvate Carboxylase (PC) is a mitochondrial protein which catalyzes the conversion of pyruvate to oxaloacetate. Agreeing with the high probability annotation for the GSH metabolic process term by our models, PC has been suggested to link glucose and GSH synthesis, along with protecting against oxidative stress and maintaining redox metabolism[73]. Recently, pyruvate metabolism has been closely linked to GSH metabolism and ferroptosis in lung cancer cells through the plasma membrane cysteine transporter SLC7A11[74] (ranked 3rd by GSH metabolic process GO term annotation probability in RF classifiers (**Supp. Data**)). PC is reported to localize to the mitochondrial matrix with little evidence supporting it being membrane-bound. It may be that PC interacts with the IMM-bound transporters to facilitate and modulate mGSH transport.

Mitochondrial Calcium Uniporter (MCU) is also a top hit in our models. Associated with its role in Ca^{2+} homeostasis are iron and redox homeostasis, and cell death pathways[75]. With respect to GSH, MCU is regulated via s-glutathionylation[75] and downregulation is associated with enriched GSH metabolism in melanoma cells[76]. It is possible that MCU plays a greater role in mitochondrial iron and GSH metabolism and transport, in the context of cancer metabolism, than has been reported thus far.

ATP-binding cassette family B6 (ABCB6) is a transporter protein with evidence for both plasma membrane and mitochondrial localization[77]. This protein transports the heme precursor porphyrin[78], however the mechanism of transport has been shown to be GSH-dependent, with significantly increased activity in the presence of GSH[79]. These suggest GSH-mediated transport of substrates by ABCB6 but it is unclear if GSH merely

modulates transport, or if it is transported as well, likely through GSH-conjugate species. Similarly, Neudesin Neurotrophic Factor (NENF) is a mitochondrial localized protein primarily involved in the differentiation/development of neuronal cells[80]. NENF activity is modulated by the binding of heme[81], which presents a connection to GSH that, to the best of our knowledge, has not yet been explored. Some evidence suggests NENF is membrane-associated[80], however, due to its small size any transmembrane transport activity would be surprising.

The results of our models find several genes which may play roles in mGSH metabolism and transport, thus attempting to guide future experimental analyses. Our findings are limited by our relatively small training dataset, where more data would certainly improve model performance and possibly confidence of predictions. Furthermore, our findings are generalized across cancer cell data from a range of tissue and cancer types. Future work, leveraging larger datasets, or datasets of specific cancer types or stages could possibly uncover interactions of more relevance to the specific diseases. Another limitation of this work is the use of gene ontology terms for identifying relevant genes. Specific and manually curated databases for these functions, similar to MitoCarta, likely provide a more accurate set of relevant genes; however, these databases are limited. Using a database like gene ontology provides a resource for the annotation of many functions, processes, and cell localizations which cannot be replicated by specific databases like MitoCarta.

Finally, the classifications from the individual models provide novel candidates for areas much beyond our scope in this work. For example, the classifiers for other metabolites here were used for validation of the glutathione classifier, but we did not investigate any predicted relevant genes for those metabolites. Nevertheless, we anticipate that our findings here can be instrumental in the identification of potentially novel proteins involved in mitochondrial glutathione metabolism and transport in cancer cells. Results here should also be useful in hypothesis generation and in guiding future experimentation for knowledge discovery in cancer settings.

Acknowledgements

This research was funded by the National Research Council Canada's New Beginnings Initiative-Ideation Fund (MH, MCC), Canadian Institutes of Health Research (MH FDN 143278) and the NSERC-CREATE Metabolomics Advanced Training and International Exchange Program (MATRIX).

Author contributions

L.K. designed and trained models and performed all other analyses. L.K. M.-E.H., and M.C.-C. interpreted the results and wrote the manuscript. M.-E.H., and M.C.-C., jointly supervised the study and acquired funding. L.K. M.-E.H., and M.C.-C. envisioned the study. All authors provided feedback and suggestions to develop the research and manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Data availability

All data used and generated here are publicly available at https://github.com/lkenn012/mGSH_cancerClassifiers[82].

Code availability

All code used to generate the results and figures is publicly available at https://github.com/lkenn012/mGSH_cancerClassifiers[82].

Tables

	Metric		
Algorithm	AUROC	AUPRC	MCC
Decision tree	0.7663	0.8270	0.5371
	0.6896†	0.7692†	0.3840†
Naive-Bayes	0.8553	0.8422	0.3517*
	0.7433†	0.7334†	0.2028† *
Random forest	**0.9003**	**0.8941**	**0.6349**
	0.8380†	**0.8156†**	**0.5437†**
Support vector machine	0.7614	0.7558	0.3487
	0.7130†	0.7250†	0.3125†

† scores from models with transcriptomics-only features.
* low MCC values relative to other evaluation metrics due to low positive class accuracy.

Table 1. **Average performance of classifiers in mGSH transporter classification**. Scores are mean values across each classifier type and all feature sets which include transcriptomic and other data. Individual scores are calculated as averages across bootstrap iterations. The highest scoring model for each evaluation metric is bolded.

Rank	Gene Symbol	GSH probability
1	SLC25A1	0.7891 (0.0173)
2	SLC25A10	0.7800 (0.0256)
3	SLC25A13	0.7737 (0.0156)
4	**SLC25A39**	**0.7716 (0.0230)**
5	SLC25A50	0.7620 (0.0222)
6	SLC25A43	0.7603 (0.0318)
7	SLC25A24	0.7577 (0.0345)
8	SLC25A37	0.7494 (0.0260)
9	SLC25A3	0.7410 (0.0489)
10	SLC25A11	0.7322 (0.0074)
37	**SLC25A40**	**0.4702 (0.0152)**

Table 2: Most probable GSH transporting SLC25 proteins by mean GSH RF classifier probability. GSH probabilities are mean probabilities with standard error for RF classifiers over different feature sets. The known GSH transporters SLC25A39 and SLC25A40 are in bold.

Rank	Gene Symbol	Glutamate probability	Rank	Gene Symbol	2-oxoglutarate probability	Rank	Gene Symbol	Carnitine probability
1	SLC25A39	0.7998 (0.0409)	1	SLC25A39	0.8408 (0.0166)	1	SLC25A30	0.8204 (0.0361)
2	SLC25A1	0.7889 (0.0357)	**2**	**SLC25A11**	**0.8304 (0.0228)**	2	SLC25A29	0.8116 (0.0196)
3	SLC25A50	0.7784 (0.0322)	3	SLC25A1	0.8287 (0.0351)	3	SLC25A36	0.7682 (0.0131)
4	SLC25A5	0.7706 (0.0148)	4	SLC25A28	0.8060 (0.0145)	4	SLC25A25	0.7632 (0.0249)
5	SLC25A10	0.7617 (0.0248)	5	SLC25A10	0.8059 (0.0292)	5	SLC25A38	0.7618 (0.0211)
11	**SLC25A13**	**0.7378 (0.0305)**				**28**	**SLC25A20**	**0.6415 (0.0345)**
23	**SLC25A22**	**0.6847 (0.0304)**						
36	**SLC25A12**	**0.6139 (0.0469)**						
42	**SLC25A18**	**0.4663 (0.0178)**						

Table 3: **Top SLC25 candidates for non-GSH metabolites by RF classifiers**. Candidates ranked by mean RF classifier probabilities (5-50 PC features) with standard error. Known metabolite transporters are bolded.

Figure legends

Figure 1: Overview of mGSH transporter classifier model design and evaluation. (**a**) Schematic of the workflow for ML model development (inspired by designs by Fellows Yates *et al.*[83]). In this diagram, lines represent the flow of information from the initially collected raw data through several transformation steps to the processed model features, and then through the ML training, evaluation, and selection steps to obtain the final classification of our unknown genes. (**b**) Example feature vectors assigned to genes in this framework, where transcriptomic features are PCs and model-specific features are MitoCarta scores, for example. ROC curves are presented for mitochondrial (**c**), transporter (**d**) and glutathione (**e**) classifiers with various feature sets. Models which incorporate transcriptomics and other feature data are colored while transcriptomics-only models are gray. Differently shaded curves represent models with various numbers of transcriptomics features (range: 5-50).

Figure 2: Confusion matrices of existing mitochondrial and transporter models (left) and models produced here (right). (**a**) MitoCarta compared to our mitochondrial RF classifier. (**b**) TrSSP compared to our transporter RF classifier. Matrices are colored based on the number of samples contained within each quadrant. A classification threshold of 0.75 was used for the mitochondrial and transporter classifiers as it provided the best balance of maximizing true positives and minimizing false positives. (**c**) Venn diagram of false positive mitochondrial genes predicted by MitoCarta 3.0 and by the mitochondrial RF classifier produced here. (**d**) Venn diagram of false positive transporter genes predicted by TrSSP and by the transporter RF classifier produced here. (**e**-**h**) GO enrichment analysis of genes predicted to localize to mitochondria by RF classifiers with no existing evidence in localization databases. Enrichment plots of false positive mitochondrial genes for KEGG pathways (**e**) cellular component (**f**), molecular function (**g**), and biological process (**h**) GO terms.

Figure 3: Structural alignment by TM-align of CAVER-identified transporter tunnels in SLC25 family members predicted to be associated with glutathione metabolic processes. (**a**) Heatmap of TM-scores for pairwise alignments of tunnel-interacting regions identified by CAVER for candidate SLC25 members. Values are TM-scores normalized to the sequence length of the column protein. Highest scoring alignments for each column are bolded. (**b**) & (**c**) Pymol visualization of transporter tunnel alignments for SLC25A39 (dark purple) and most similar SLC25 candidates by TM-score, SLC25A43 (green) and SLC25A24 (orange). Indicated residues are those identified as relevant for GSH transport by SLC25A39 and corresponding residues in SLC25A43 and SLC25A24. Protein tunnels visualized from the intermembrane space with tunnel-interacting residues identified by CAVER highlighted and non-tunnel residues faded.

Figure 4: Evaluation of gene ontology annotation predictions by DeepGOPlus (dashed) and RF classifier models (solid). Sensitivity, or TPR, are calculated at various classification thresholds.

References

1. Pavlova, N. N., Zhu, J. & Thompson, C. B. The hallmarks of cancer metabolism: Still emerging. Cell Metab. 34, 355–377 (2022).

2. Kennedy, L., Sandhu, J. K., Harper, M.-E. & Cuperlovic-Culf, M. Role of Glutathione in Cancer: From Mechanisms to Therapies. Biomolecules 10, (2020).

3. Lien, E. C. et al. Glutathione biosynthesis is a metabolic vulnerability in PI(3)K/Akt-driven breast cancer. Nat. Cell Biol. 18, 572–578 (2016).

4. Harris, I. S. et al. Glutathione and thioredoxin antioxidant pathways synergize to drive cancer initiation and progression. Cancer Cell 27, 211–222 (2015).

5. Kong, H. & Chandel, N. S. Regulation of redox balance in cancer and T cells. J. Biol. Chem. 293, 7499–7507 (2018).

6. Piskounova, E. et al. Oxidative stress inhibits distant metastasis by human melanoma cells. Nature 527, 186–191 (2015).

7. Yang, W. S. et al. Regulation of Ferroptotic Cancer Cell Death by GPX4. Cell 156, 317–331 (2014).

8. Ballatori, N., Krance, S. M., Marchan, R. & Hammond, C. L. Plasma membrane glutathione transporters and their roles in cell physiology and pathophysiology. Mol. Aspects Med. 30, 13–28 (2009).

9. Wallace, D. C. Mitochondria and cancer. Nat. Rev. Cancer 12, 685–698 (2012).

10. Whitaker, H. C. et al. Peroxiredoxin-3 is overexpressed in prostate cancer and promotes cancer cell survival by protecting cells from oxidative stress. Br. J. Cancer 109, 983–993 (2013).

11. Oestreicher, J. & Morgan, B. Glutathione: subcellular distribution and membrane transport1. Biochem. Cell Biol. (2018) doi:10.1139/bcb-2018-0189.

12. Chen, Z. & Lash, L. H. Evidence for mitochondrial uptake of glutathione by dicarboxylate and 2-oxoglutarate carriers. J. Pharmacol. Exp. Ther. 285, 608–618 (1998).

13. Booty, L. M. et al. The mitochondrial dicarboxylate and 2-oxoglutarate carriers do not transport glutathione. FEBS Lett. 589, 621–628 (2015).

14. Wang, Y. et al. SLC25A39 is necessary for mitochondrial glutathione import in mammalian cells. Nature 599, 136–140 (2021).

15. Shi, X. et al. Combinatorial GxGxE CRISPR screen identifies SLC25A39 in mitochondrial glutathione transport linking iron homeostasis to OXPHOS. Nat. Commun. 13, 2483 (2022).

16. von Bohlen und Halbach, O. Controlling glutathione entry into mitochondria: potential roles for SLC25A39 in health and (treatment of) disease. Signal Transduction and Targeted Therapy 7, 1–2 (2022).

17. Cooper, A. J. L. & Hanigan, M. H. Metabolism of glutathione S-conjugates: Multiple pathways. in Comprehensive Toxicology 363–406 (Elsevier, 2018).

18. Ramsay, E. E. & Dilda, P. J. Glutathione S-conjugates as prodrugs to target drug-resistant tumors. Front. Pharmacol. 5, 181 (2014).

19. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).

20. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science 373, 871–876 (2021).

21. Zhou, N. et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. Genome Biol. 20, 1–23 (2019).

22. Kulmanov, M., Khan, M. A. & Hoehndorf, R. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. Bioinformatics 34, 660–668 (2017).

23. Kulmanov, M. & Hoehndorf, R. DeepGOPlus: improved protein function prediction from sequence. Bioinformatics 36, 422–429 (2019).

24. Zhou, J., Xiong, W., Wang, Y. & Guan, J. Protein Function Prediction Based on PPI Networks: Network Reconstruction vs Edge Enrichment. Front. Genet. 12, 758131 (2021).

25. Hu, S. et al. Protein function annotation based on heterogeneous biological networks. BMC Bioinformatics 23, 1–16 (2022).

26. Lei, Y. et al. A deep-learning framework for multi-level peptide–protein interaction prediction. Nat. Commun. 12, 1–10 (2021).

27. Shanehsazzadeh, A. et al. Unlocking de novo antibody design with generative artificial intelligence. bioRxiv 2023.01.08.523187 (2023) doi:10.1101/2023.01.08.523187.

28. Kunc, V. & Klema, J. On functional annotation with gene co-expression networks. in 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (IEEE, 2022). doi:10.1109/bibm55620.2022.9995542.

29. Wekesa, J. S., Luan, Y. & Meng, J. Predicting Protein Functions Based on Differential Co-expression and Neighborhood Analysis. J. Comput. Biol. 28, 1 (2021).

30. Wang, Y. et al. Using Machine Learning to Measure Relatedness Between Genes: A Multi-Features Model. Sci. Rep. 9, 1–15 (2019).

31. Karpatne, A., Kannan, R. & Kumar, V. Knowledge Guided Machine Learning: Accelerating Discovery using Scientific Knowledge and Data. (CRC Press, 2022).

32. Liu, F., Heiner, M. & Gilbert, D. Hybrid modelling of biological systems: current progress and future prospects. Brief. Bioinform. 23, (2022).

33. Elmarakeby, H. A. et al. Biologically informed deep neural network for prostate cancer discovery. Nature 598, 348–352 (2021).

34. Yu, K. et al. Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. Nat. Commun. 10, 1–11 (2019).

35. Rath, S. et al. MitoCarta3.0: an updated mitochondrial proteome now with sub-organelle localization and pathway annotations. Nucleic Acids Res. 49, D1541–D1547 (2021).

36. Mishra, N. K., Chang, J. & Zhao, P. X. Prediction of membrane transport proteins and their substrate specificities using primary sequence information. PLoS One 9, e100278 (2014).

37. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. arXiv (2012) doi:10.48550/arXiv.1201.0490.

38. Hunter, J. D. Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. 9, 90–95 (2007).

39. Waskom, M. et al. mwaskom/seaborn: v0.8.1 (September 2017). (2017) doi:10.5281/zenodo.883859.

40. Ge, S. X., Jung, D. & Yao, R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics 36, 2628–2629 (2020).

41. Li, H. et al. The landscape of cancer cell line metabolism. Nat. Med. 25, 850–860 (2019).

42. Beretta, L. & Santaniello, A. Nearest neighbor imputation algorithms: a critical evaluation. BMC Med. Inform. Decis. Mak. 16, 197–208 (2016).

43. Fisher, R. A. Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. Biometrika 10, 507–521 (1915).

44. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29 (2000).

45. Carbon, S. et al. AmiGO: online access to ontology and annotation data. Bioinformatics 25, 288–289 (2009).

46. Mudunuri, U., Che, A., Yi, M. & Stephens, R. M. bioDBnet: the biological database network. Bioinformatics 25, 555 (2009).

47. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure 405, 442–451 (1975).

48. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21, 1–13 (2020).

49. Waterhouse, A. et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 46, W296–W303 (2018).

50. Pebay-Peyroula, E. et al. Structure of mitochondrial ADP/ATP carrier in complex with carboxyatractyloside. Nature 426, 39–44 (2003).

51. Kang, Y. & Chen, L. Structural basis for the binding of DNP and purine nucleotides onto UCP1. Nature 620, 226–231 (2023).

52. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 33, 2302–2309 (2005).

53. The PyMOL Molecular Graphics System, Version 2.4.1 Schrödinger, LLC.

54. Chovancova, E. et al. CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. PLoS Comput. Biol. 8, e1002708 (2012).

55. Thul, P. J. & Lindskog, C. The human protein atlas: A spatial map of the human proteome. Protein Sci. 27, 233–244 (2018).

56. The UniProt Consortium et al. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 51, D523–D531 (2022).

57. Smith, A. C. & Robinson, A. J. MitoMiner v4.0: an updated database of mitochondrial localization evidence, phenotypes and diseases. Nucleic Acids Res. 47, D1225–D1228 (2019).

58. Szklarczyk, D. et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 47, D607–D613 (2019).

59. Yue, S. et al. Highly conserved shifts in ubiquitin-proteasome system (UPS) activity drive mitochondrial remodeling during quiescence. Nat. Commun. 13, 1–14 (2022).

60. Sharma, A. et al. Therapeutic implications of mitochondrial stress-induced proteasome inhibitor resistance in multiple myeloma. Sci Adv 8, eabq5575 (2022).

61. Saier, M. H. et al. The Transporter Classification Database (TCDB): 2021 update. Nucleic Acids Res. 49, D461–D467 (2020).

62. Gyimesi, G. & Hediger, M. A. Systematic in silico discovery of novel solute carrier-like proteins from proteomes. PLoS One 17, e0271062 (2022).

63. Ruprecht, J. J. et al. The Molecular Mechanism of Transport by the Mitochondrial ADP/ATP Carrier. Cell 176, 435–447.e15 (2019).

64. Berardi, M. J., Shih, W. M., Harrison, S. C. & Chou, J. J. Mitochondrial uncoupling protein 2 structure determined by NMR molecular fragment searching. Nature 476, 109–113 (2011).

65. Ruprecht, J. J. & Kunji, E. R. S. The SLC25 Mitochondrial Carrier Family: Structure and Mechanism. Trends Biochem. Sci. 45, 244–258 (2020).

66. Kulmanov, M., Zhapa-Camacho, F. & Hoehndorf, R. DeepGOWeb: fast and accurate protein function prediction on the (Semantic) Web. Nucleic Acids Res. 49, W140–W146 (2021).

67. Fernandez, H. R. et al. The mitochondrial citrate carrier, SLC25A1, drives stemness and therapy resistance in non-small cell lung cancer. Cell Death Differ. 25, 1239–1258 (2018).

68. Wadey, A. L., Muyderman, H., Kwek, P. T. & Sims, N. R. Mitochondrial glutathione uptake: characterization in isolated brain mitochondria and astrocytes in culture. J. Neurochem. 109 Suppl 1, 101–108 (2009).

69. Punzi, G. et al. SLC25A10 biallelic mutations in intractable epileptic encephalopathy with complex I deficiency. Hum. Mol. Genet. 27, 499–504 (2017).

70. Zhang, D. et al. Mitochondrial outer membrane protein FUNDC2 promotes ferroptosis and contributes to doxorubicin-induced cardiomyopathy. Proceedings of the National Academy of Sciences 119, (2022).

71. Palmieri, L. et al. Citrin and aralar1 are Ca(2+)-stimulated aspartate/glutamate transporters in mitochondria. EMBO J. 20, 5060–5069 (2001).

72. Zhang, J. et al. Genome-wide screening in the haploid system reveals Slc25a43 as a target gene of oxidative toxicity. Cell Death Dis. 13, 1–11 (2022).

73. Fu, A. et al. Glucose metabolism and pyruvate carboxylase enhance glutathione synthesis and restrict oxidative stress in pancreatic islets. Cell Rep. 37, 110037 (2021).

74. Zhao, L. et al. A non-canonical function of LDHB promotes SLC7A11-mediated glutathione metabolism and protects against glutaminolysis-dependent ferroptosis in KRAS-driven lung cancer. bioRxiv 2023.02.12.525859 (2023) doi:10.1101/2023.02.12.525859.

75. Dong, Z. et al. Mitochondrial Ca2+ Uniporter is a Mitochondrial Luminal Redox Sensor That Augments MCU Channel Activity. Mol. Cell 65, 1014 (2017).

76. Stejerean-Todoran, I. et al. MCU controls melanoma progression through a redox-controlled phenotype switch. EMBO Rep. 23, e54746 (2022).

77. Song, G. et al. Molecular insights into the human ABCB6 transporter. Cell Discovery 7, 1–11 (2021).

78. Xie, L.-H., Fefelova, N., Pamarthi, S. H. & Gwathmey, J. K. Molecular Mechanisms of Ferroptosis and Relevance to Cardiovascular Disease. Cells 11, 2726 (2022).

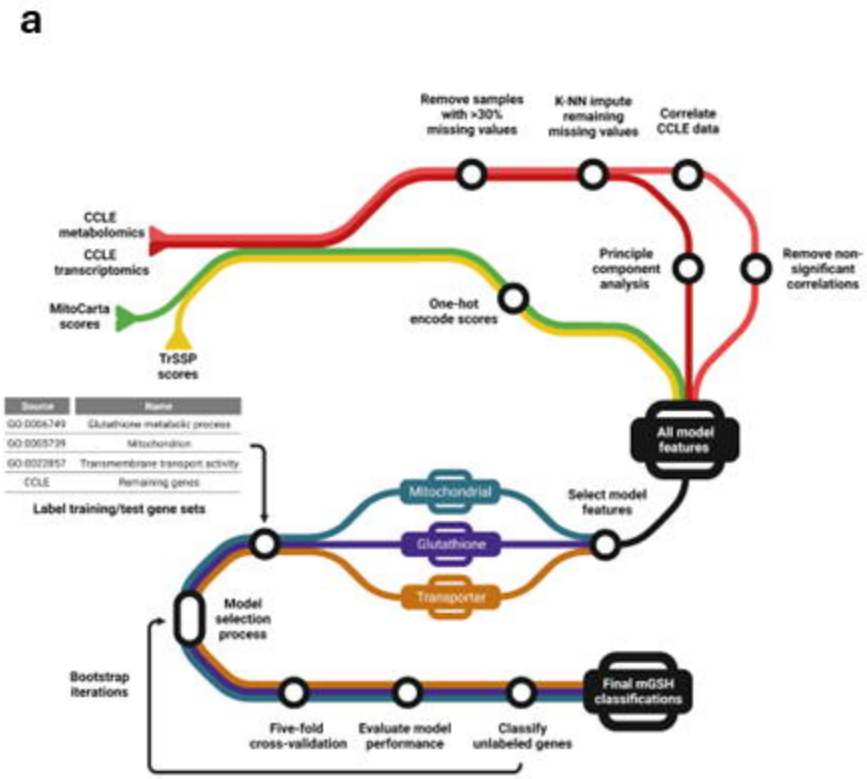
79. Wang, C. et al. Cryo-electron microscopy structure of human ABCB6 transporter. Protein Sci. 29, 2363 (2020).

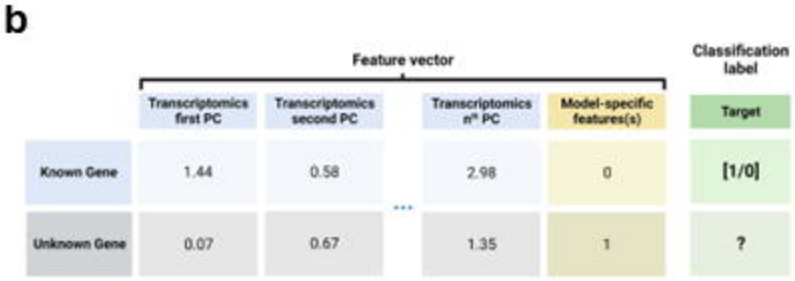
80. Kimura, I., Nakayama, Y., Zhao, Y., Konishi, M. & Itoh, N. Neurotrophic effects of neudesin in the central nervous system. Front. Neurosci. 7, 54475 (2013).

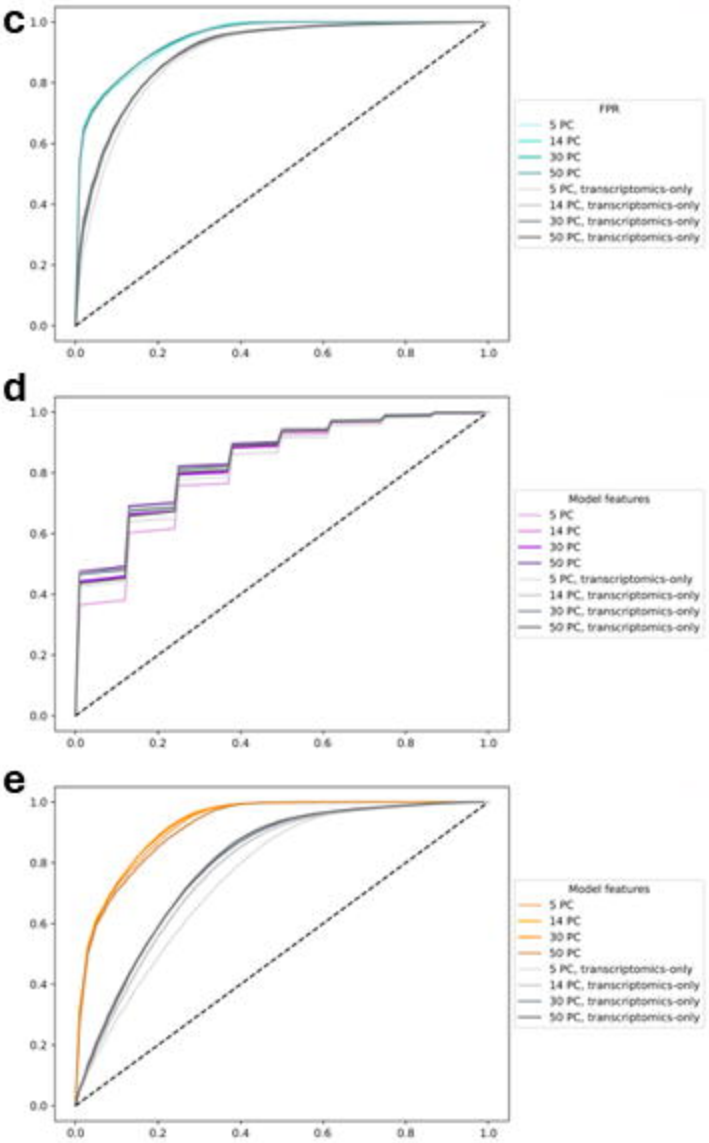
81. Moutaoufik, M. T. et al. Rewiring of the Human Mitochondrial Interactome during Neuronal Reprogramming Reveals Regulators of the Respirasome and Neurogenesis. iScience 19, 1114–1132 (2019).

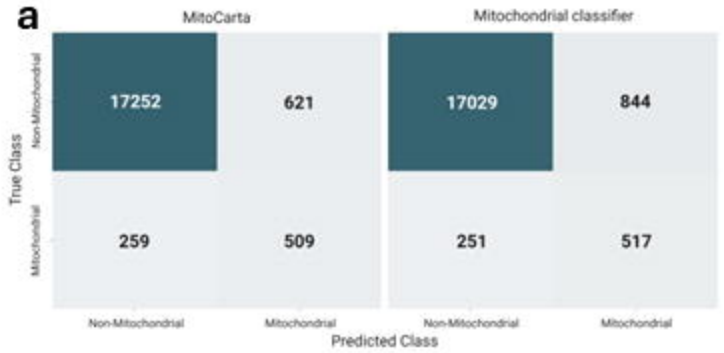
82. GitHub - lkenn012/mGSH_cancerClassifiers: Code, data, and other information for mGSH transporter models. GitHub https://github.com/lkenn012/mGSH_cancerClassifiers.

83. Fellows Yates, J. A. et al. Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager. PeerJ 9, e10947 (2021).

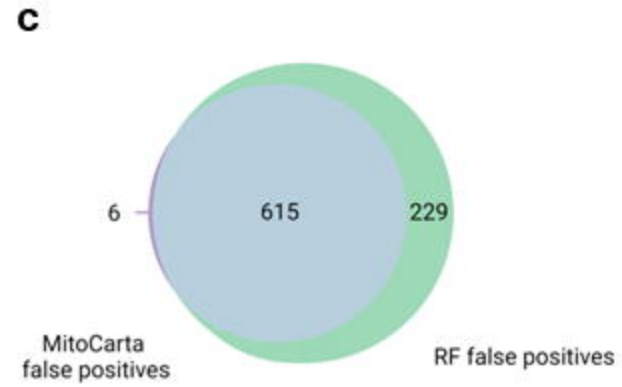

a
Remove samples with >30% missing values
K-NN impute remaining missing values
Correlate CCLE data
CCLE metabolomics
CCLE transcriptomics
MitoCarta scores
TrSSP scores
One-hot encode scores
Principle component analysis
Remove non-significant correlations
All model features
Source
Name
GO:0006749
Glutathione metabolic process
GO:0005739
Mitochondrion
GO:0022857
Transmembrane transport activity
CCLE
Remaining genes
Label training/test gene sets
Mitochondrial
Glutathione
Transporter
Select model features
Model selection process
Bootstrap iterations
Final mGSH classifications
Five-fold cross-validation
Evaluate model performance
Classify unlabeled genes

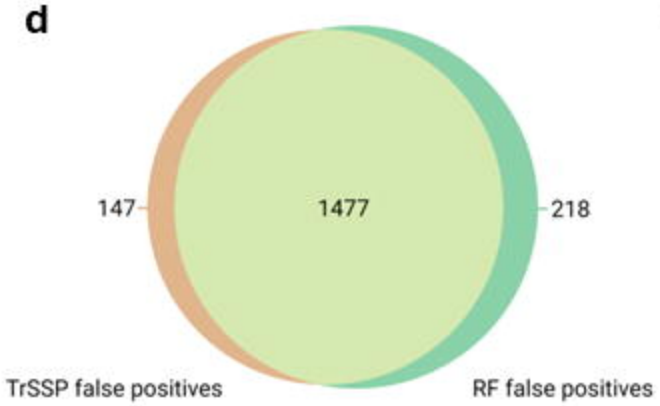

b
Feature vector
Classification label
Transcriptomics first PC
Transcriptomics second PC
Transcriptomics nth PC
Model-specific features(s)
Target
Known Gene
1.44
0.58
2.98
0
[1/0]
Unknown Gene
0.07
0.67
1.35
1
?

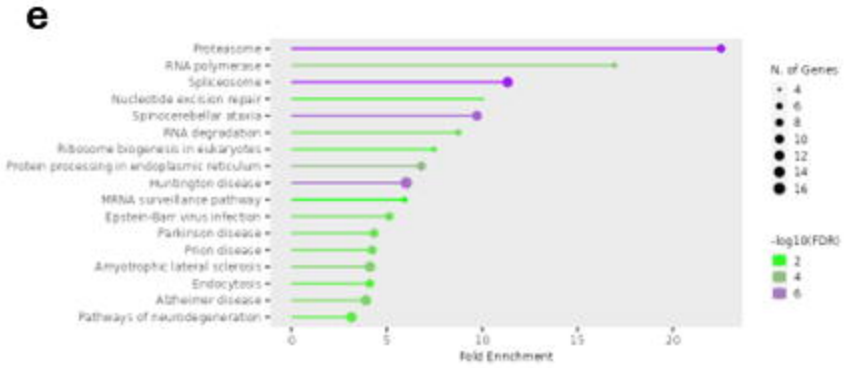

c
FPR
5 PC
14 PC
30 PC
50 PC
5 PC, transcriptomics-only
14 PC, transcriptomics-only
30 PC, transcriptomics-only
50 PC, transcriptomics-only
d
Model features
5 PC
14 PC
30 PC
50 PC
5 PC, transcriptomics-only
14 PC, transcriptomics-only
30 PC, transcriptomics-only
50 PC, transcriptomics-only
e
Model features
5 PC
14 PC
30 PC
50 PC
5 PC, transcriptomics-only
14 PC, transcriptomics-only
30 PC, transcriptomics-only
50 PC, transcriptomics-only

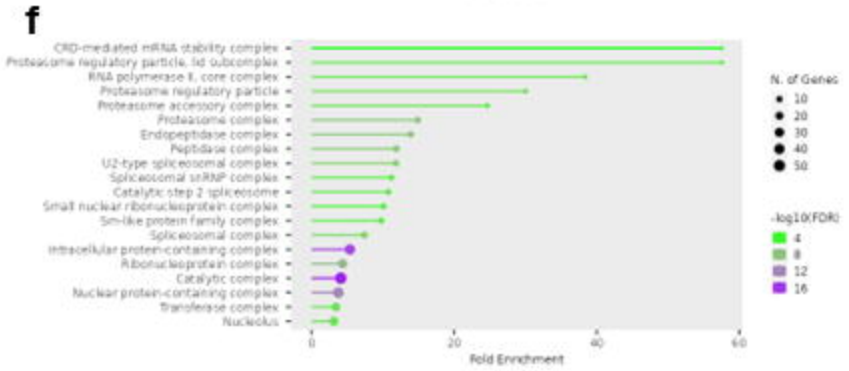

a
MitoCarta
Mitochondrial classifier
True Class
Non-Mitochondrial
Mitochondrial
17252
621
259
509
17029
844
251
517
Non-Mitochondrial
Mitochondrial
Non-Mitochondrial
Mitochondrial
Predicted Class

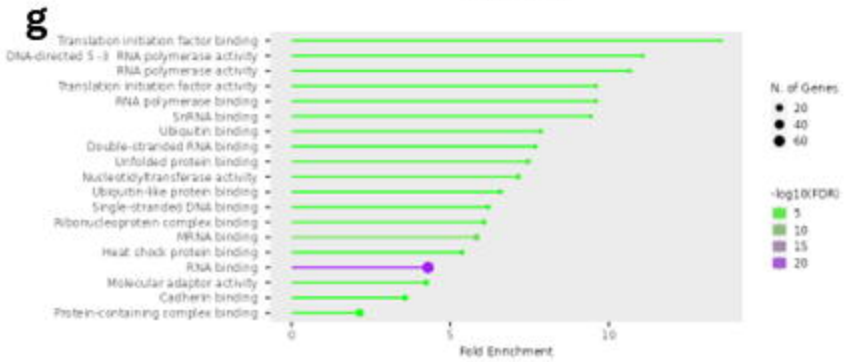
b
TrSSP
Transporter classifier
True Class
Non-Transporter
Transporter
16115
1624
280
416
15930
1809
286
410
Non-Transporter
Transporter
Non-Transporter
Transporter
Predicted Class

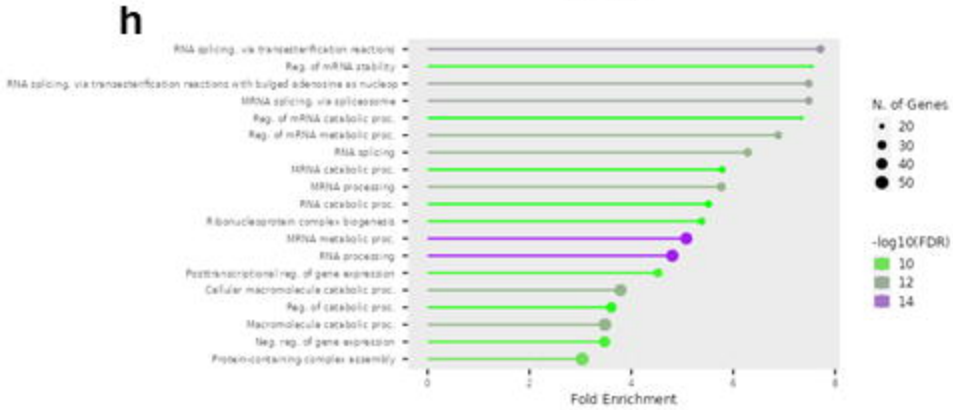

c
6
615
229
MitoCarta
false positives
RF false positives


d
147
1477
218
TrSSP false positives
RF false positives


e
Proteasome
RNA polymerase
Spliceosome
Nucleotide excision repair
Spinocerebellar ataxia
RNA degradation
Ribosome biogenesis in eukaryotes
Protein processing in endoplasmic reticulum
Huntington disease
MRNA surveillance pathway
Epstein-Barr virus infection
Parkinson disease
Prion disease
Amyotrophic lateral sclerosis
Endocytosis
Alzheimer disease
Pathways of neurodegeneration
0
5
10
15
20
Fold Enrichment
N. of Genes
4
6
8
10
12
14
16
-log10(FDR)
2
4
6

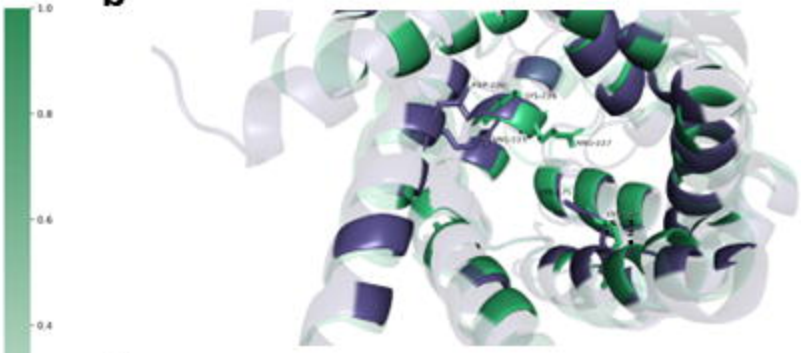

f
CRD-mediated mRNA stability complex
Proteasome regulatory particle, lid subcomplex
RNA polymerase II, core complex
Proteasome regulatory particle
Proteasome accessory complex
Proteasome complex
Endopeptidase complex
Peptidase complex
U2-type spliceosomal complex
Spliceosomal snRNP complex
Catalytic step 2 spliceosome
Small nuclear ribonucleoprotein complex
Sm-like protein family complex
Spliceosomal complex
Intracellular protein-containing complex
Ribonucleoprotein complex
Catalytic complex
Nuclear protein-containing complex
Transferase complex
Nucleolus
0
20
40
60
Fold Enrichment
N. of Genes
10
20
30
40
50
-log10(FDR)
4
8
12
16


g
Translation initiation factor binding
DNA-directed 5′-3′ RNA polymerase activity
RNA polymerase activity
Translation initiation factor activity
RNA polymerase binding
SnRNA binding
Ubiquitin binding
Double-stranded RNA binding
Unfolded protein binding
Nucleotidyltransferase activity
Ubiquitin-like protein binding
Single-stranded DNA binding
Ribonucleoprotein complex binding
MRNA binding
Heat shock protein binding
RNA binding
Molecular adaptor activity
Cadherin binding
Protein-containing complex binding
0
5
10
Fold Enrichment
N. of Genes
20
40
60
-log10(FDR)
5
10
15
20


h
RNA splicing, via transesterification reactions
Reg. of mRNA stability
RNA splicing, via transesterification reactions with bulged adenosine as nucleop
MRNA splicing, via spliceosome
Reg. of mRNA catabolic proc.
Reg. of mRNA metabolic proc.
RNA splicing
MRNA catabolic proc.
MRNA processing
RNA catabolic proc.
Ribonucleoprotein complex biogenesis
MRNA metabolic proc.
RNA processing
Posttranscriptional reg. of gene expression
Cellular macromolecule catabolic proc.
Reg. of catabolic proc.
Macromolecule catabolic proc.
Neg. reg. of gene expression
Protein-containing complex assembly
0
2
4
6
8
Fold Enrichment
N. of Genes
20
30
40
50
-log10(FDR)
10
12
14

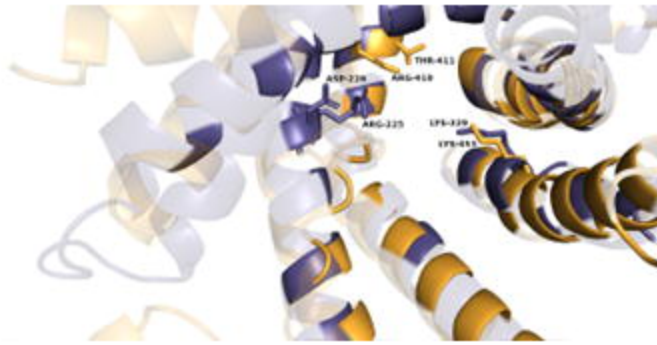
a

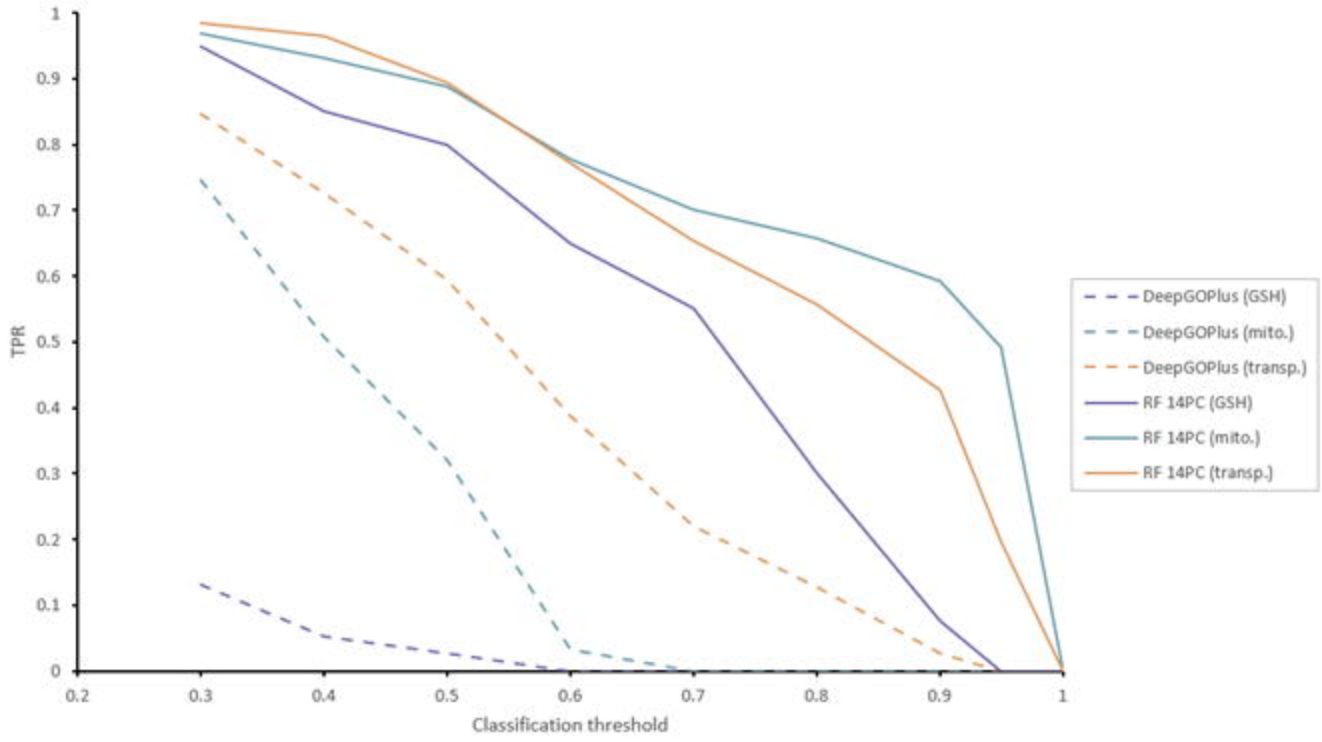
	SLC25A1_AF	SLC25A3_AF	SLC25A10_AF	SLC25A11_AF	SLC25A13_AF	SLC25A24_AF	SLC25A37_AF	SLC25A43_AF	SLC25A50_AF	SLC25A39_AF	SLC25A40_AF	SLC25A39_homology_1oxc	SLC25A40_homology_8h0v
SLC25A1_AF	1	0.21	0.24	0.21	0.31	0.38	0.35	0.31	0.39	0.31	0.28	0.39	0.31
SLC25A3_AF	0.18	1	0.26	0.22	0.17	0.16	0.21	0.19	0.22	0.29	0.23	0.2	0.26
SLC25A10_AF	0.22	0.26	1	0.34	0.32	0.25	0.28	0.26	0.15	0.2	0.2	0.23	0.34
SLC25A11_AF	0.19	0.22	0.3	1	0.46	0.35	0.31	0.33	0.17	0.25	0.18	0.39	0.48
SLC25A13_AF	0.28	0.17	0.3	0.48	1	0.23	0.37	0.24	0.22	0.28	0.19	0.38	0.41
SLC25A24_AF	0.39	0.19	0.3	0.51	0.28	1	0.41	0.52	0.3	0.24	0.17	0.55	0.52
SLC25A37_AF	0.37	0.25	0.32	0.44	0.49	0.42	1	0.34	0.21	0.21	0.22	0.37	0.53
SLC25A43_AF	0.3	0.19	0.28	0.4	0.28	0.45	0.3	1	0.34	0.24	0.29	0.49	0.38
SLC25A50_AF	0.34	0.21	0.16	0.17	0.22	0.26	0.19	0.32	1	0.23	0.15	0.28	0.37
SLC25A39_AF	0.32	0.34	0.21	0.31	0.33	0.24	0.2	0.25	0.29	1	0.72	0.32	0.33
SLC25A40_AF	0.29	0.25	0.21	0.18	0.25	0.17	0.21	0.3	0.16	0.74	1	0.32	0.39
SLC25A39_homology_1oxc	0.37	0.2	0.24	0.46	0.42	0.5	0.34	0.5	0.33	0.31	0.31	1	0.55
SLC25A40_homology_8h0v	0.3	0.27	0.35	0.6	0.46	0.46	0.45	0.38	0.4	0.31	0.35	0.54	1



b

c




TPR
Classification threshold
DeepGOPlus (GSH)
DeepGOPlus (mito.)
DeepGOPlus (transp.)
RF 14PC (GSH)
RF 14PC (mito.)
RF 14PC (transp.)