Applying Machine Learning to Predict Adipose Browning Capacity and Mitochondria-Endoplasmic Reticulum Crosstalk

Li Jiang

München 2021

Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

Applying Machine Learning to Predict Adipose Browning Capacity and Mitochondria-Endoplasmic Reticulum Crosstalk

Li Jiang
aus Jining, Shandong, China

2021

Erklärung

Diese Dissertation wurde im Sinne von § 7 der Promotionsordnung vom 28. November 2011 von Frau Dr. Fabiana Perocchi betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, 08.01.2021 .....

Ort, Datum Li Jiang

Dissertation eingereicht am	14.08.2020
1. Gutachterin:	Dr. Fabiana Perocchi
2. Gutachter:	Prof. Dr. Klaus Förstemann
Mündliche Prüfung am	10.12.2020

Acknowledgements

First and foremost I would love to express my sincere thanks to my supervisor Dr. Fabiana Perocchi, who not only offers me with a fantastic opportunity to work in many exciting topics, but also inspires me with her great passion for science, and creative and critical thinking skills.

I feel very grateful to Dr. Mario Halic and Prof. Dr. Klaus Förstemann, who are my thesis advisory committee members of HELENA graduate school. They are never reluctant to provide me with superb advice and precious time.

I acknowledge other members of my examination board (Prof. Dr. em. Ernst-Ludwig Winnacker, Prof. Dr. med. Thomas Misgeld, Prof. Dr. rer. nat. Ming Chen and Prof. Dr. Lucas Jae). I appreciate you for reading my thesis and offering valuable suggestions.

My biggest thanks must go to Yiming, who often discusses projects with me and offers great suggestions. I also feel thankful to Daniela, who collaborated with me scientifically and never felt grudge to share biological knowledge. Thank you, Valerie, for chatting with me - you have brought me with a lot of laugh and optimistic views of the world. In no particular order, I also thank Anja, who helped me in administrative works and offered me a free ride. Many thanks to Jennifer - you saved me a considerable amount of time by sharing with me information for Ph.D. registration. I also thank Hilda, Michael, Natalia and Simona for creating an excellent working atmosphere.

I feel thankful to Zhan - you are a great amiable friend to me. Thank you, Victor, for discussing scientific topics and having funny talks with me.

Last but not least, I owe my deepest thanks to my parents and husband, who are always supporting me emotionally and unconditionally.

Disclaimer

Chapter 2 is copied from the published work below.

Prediction of adipose browning capacity by systematic integration of transcriptional profiles [1].

Yiming Cheng*, **Li Jiang***, Susanne Keipert*, Shuyue Zhang, Andreas Hauser, Elisabeth Graf, Tim Strom, Matthias Tschöp, Martin Jastroch**, and Fabiana Perocchi**. Cell reports 23, no. 10 (2018): 3112-3125.
(* joint first authorship, ** joint corresponding author)

Author contributions: Conceptualization, F.P. and M.J.; Methodology, Y.C., S.K., and L.J.; Software, Y.C., A.H., and L.J.; Formal Analysis, Y.C., A.H., L.J., and T.S.; Investigation, S.K. and E.G.; Resources, F.P., M.J., and M.T.; Data Curation, S.K. and S.Z.; Writing – Original Draft, F.P., M.J., S.K., and Y.C.; Visualization, F.P., L.J., Y.C., and A.H.; Supervision, F.P. and M.J.; Funding Acquisition, F.P.

Summary

The first project aims to predict adipose browning capacity (chapter 2). Activation and recruitment of thermogenic cells in human white adipose tissues ("browning") can counteract obesity and associated metabolic disorders. However, quantifying the effects of therapeutic interventions on browning remains enigmatic. Here, we devise a computational tool, named ProFAT (profiling of fat tissue types), for quantifying the thermogenic potential of heterogeneous fat biopsies based on prediction of white and brown adipocyte content from raw gene expression datasets. ProFAT systematically integrates 103 mouse-fat-derived transcriptomes to identify unbiased and robust gene signatures of brown and white adipocytes. We validate ProFAT on 80 mouse and 97 human transcriptional profiles from 14 independent studies and correctly predict browning capacity upon various physiological and pharmacological stimuli. Our study represents the most exhaustive comparative analysis of public data on adipose biology toward quantification of browning after personalized medical intervention. ProFAT is freely available and should become increasingly powerful with the growing wealth of transcriptomics data.

In virtually all cells mitochondria and ER physically interact forming membrane contact sites termed mitochondria-associated membranes (MAMs). These stable contacts allow the synergistic functioning of the two organelles as well as the concerted regulation of numerous cell biological processes, including Ca^{2+} homeostasis, lipid synthesis and trafficking, protein folding, ROS generation and activity, apoptosis, autophagy, mitochondrial morphology and dynamics. A growing number of studies describing the molecular composition and potential involvement of MAMs in different pathological contexts have highlighted the relevance of ER-mitochondria crosstalk to cell physiology and disease. However, given the dynamic nature of these interactions, protein and lipid complements of MAMs still remain to be fully elucidated. The second project aims to achieve a comprehensive understanding of the molecular players and pathways functionally linking the ER and mitochondria, with a particular focus on inter-organelles Ca^{2+} signaling. To this goal, we have systematically defined a compendium of ER-localized proteins by integrating publicly available genome-wide datasets that provide complementary clues about ER localization using Boosting algorithm (chapter 3). Our compendium of ER proteins, named ERcarta, contains 1023

proteins, which includes 354 novel predictions. To shed light on the functional role of ERcarta genes in the crosstalk between the ER and mitochondria, we have reconstructed a comprehensive Mito-ER regulatory network, which can provide a functional context for interesting candidates after clustering (chapter 4). Given our primary interest in characterizing ER-mitochondria functional associations that regulate Ca^{2+} signaling, we systematically quantified the effect of knocking down ERcarta genes on mitochondrial Ca^{2+} dynamics and identified 294 candidates, which may regulate the mitochondrial Ca^{2+} uptake. Currently, we are prioritizing interesting candidates for follow-up analyses.

List of Abbreviations

ACC	Accuracy
Ago	Argonaute
ARI	Adjusted Rand Index
AUC	Area under Curve
BAs	Brown Adipocytes
BAT	Brown Adipose Tissue
BIND	Biomolecular Interaction Network Database
BioGRID	Biological General Repository for Interaction Datasets
BRH	Best-Reciprocal-Hits
cryo-ET	Electron Cryotomography
CYPs	Cytochrome P450
DFG	German Research Foundation
DHC	Dynamic Hierarchical Clustering
DIP	Database of Interacting Proteins
EM	Electron Microscopy
ER	Endoplasmic Reticulum
ERMES	ER-Mitochondria Encounter Structure
ERRP	ER-Resident Proteins
FDR	False Discovery Rate
FN	False Negative
FP	False Positve
FPR	False Positive Rate
FTP	File Transfer Protocol
GEO	Gene Expression Omnibus
GFP	Green Fluorescent Protein
GO	Gene Ontology
GOBP	Gene Ontology Biological Process
GOCC	Gene Ontology Cellular Component
gWAT	Gonadal WAT
HC	Hierarchical Clustering
HPA	Human Protein Atlas
HPRD	Human Protein Reference Database
IP	Immunoprecipitation

iWAT	Inguinal WAT
KNN	k-Nearest Neighbors
KO	KEGG Orthology
LOOCV	Leave-One-Out Cross-Validation
MAMs	Mitochondria-Associated Membranes
MCSs	Membrane Contact Sites
MCU	Mitochondrial Calcium Uniporter
MFE	Minimum of Free Energy
MINT	The Molecular Interaction Database
miRNAs	microRNAs
mWAT	Mesenteric WAT
NTD	Negative Training Dataset
OBH	One-Way Best-Hits
OMM	Outer-Mitochondrial Membrane
ORFs	Open Reading Frames
PCA	Principle Component Analysis
PDB	Protein Data Bank
PET	Positron Emission Tomography
PID	Pathway Interaction Database
PIR	Protein Information Resource
piRNAs	Piwi-Interacting RNAs
POC	Percentage of Control
PPI	Protein-Protein Interaction
PPIs	Protein-Protein Interactions
PRIDE	The PRoteomics IDEntifications Database
ProFAT	Profiling of Fat Tissue Types
PSC	Pluripotent Stem Cell
PTD	Positive Training Dataset
pvWAT	Perivascular WAT
RF	Random Forest
RG	Rosiglitazone
RIN	RNA Integrity Number
RMA	Robust Multiarray Average
ROC	Receiver Operating Characteristics
ROS	Reactive Oxygen Species
RS	Roscovitine
RyR	Ryanodine Receptor
SC	Index Score
iWAT	Inguinal WAT
SGD	*Saccharomyces* Genome Database
SIB	Institute of Bioinformatics
siRNA	Small Interfering RNA
siRNAs	Small Interfering RNAs
SLNN	Single-Layer Neural Network

SMOTE	Synthetic Minority over-Sampling Technique
SPC	Specificity
SR	Sarcoplasmic Reticulum
sRNAs	Small RNAs
SVFs	Stromal Vascular Fractions
SVMs	Support Vectors Machines
sWAT	Subcutaneous WAT
SYK	Spleen Tyrosine Kinase
TF	True Positive
TN	True Negative
TNR	True Negative Rate
TP	True Positives
TPR	True Positive Rate
TrEMBL	Translated EMBL Nucleotide Sequence Data Library
TSS	Transcription Start Site
UGT	UDP-Glucuronosyltransferase
UniParc	The UniProt Archive
UniProtKB	The UniProt Knowledgebase
UniRef	The UniProt Reference Clusters
UPI	Unique Proteome Identifier
WAs	White Adipocytes
WAT	White Adipose Tissue

Table of contents

List of figures

List of tables

Chapter 1

Introduction

The introduction in section of 1.4.1 of this chapter is from the manuscript "Prediction of adipose browning capacity by systematic integration of transcriptional profiles" by Cheng, Jiang et al. 2018 without modification [1].

1.1 Machine learning

1.1.1 Definitions

As a branch of artificial intelligence, machine learning aims to generate and apply models to do given tasks automatically. It is widely used in many areas including biological sciences. The development of machine learning techniques has enabled scientists to investigate biological problems by analyzing large-scale and complex data sets and data types. Below are three examples of its applications:

(1) Estimation of overall survival of patients with pancreatic adenocarcinoma based on the expression pattern of five microRNAs [2].

(2) Recognition of promoters based on protein-DNA-twist values, *etc.* [3].

(3) Prediction of ten-year risk for the development of type 2 diabetes, based on age, BMI, ethnicity, sex, *etc.* [4].

All of the aforementioned applications include two elements:

- **A set of records**. One record, is composed of

(a) An output, often referred to as a "target" and conventionally denoted by the letter y, represents a value or class to predict (*e.g.*, the survival time of a person, whether a genomic region is a promoter, a 10-year risk for developing type 2 diabetes).

(b) An input, alternatively termed as an "observation" and conventionally represented by the letter $\mathbf{x}$, is usually a set of standalone perceivable knowledge that can aid in the prediction of the output (*e.g.*, the expression pattern of five microRNAs in (1), the protein-DNA-twist values in (2), the age and gender in (3)). One observation can consist of a set of individual "features". For example, in (3), it comprises at least four features (age, BMI, ethnicity, and sex).

Note that the number of records in a study equals to the number of objects included in that study, *e.g.,* in the third example, the number of people involved.

- **A model**. This is a learnt function (f) that maps from an input ($\mathbf{x}$) to an output (y), and is able to make predictions.

Data can then be divided into at least two mutually exclusive types - "numerical" and "categorical". Numerical data are in the form of real or integer numbers. For instance, the survival time of a patient and microRNA expression values, protein-DNA-twist values, and BMI are all real numbers, whereas a person's age is an integer. Instead, categorical data can represent mutually exclusive groups (*e.g.*, in (2) a genomic region can be either a "promoter" or a "non-promoter"; in (3) a person's gender can be either "male" or "female").

1.1.1.1 Supervised learning

Finding out what is the function (f) based on known coupled input-output pairs is called "supervised learning", and the process of finding the function is called "training" or "fitting".

In a supervised learning context, one coupled input-output pair is a "record". The set of records used in the training process is called a "training set". Usually, the final goal of supervised learning is to estimate the values of new targets based on their observations. A popular way to evaluate the model's ability to make reasonable and accurate predictions is by using a "test set". This set contains known observations and corresponding targets, and it should not overlap with any records in the training set. The evaluation process includes two steps, the prediction of targets based on the observations from the test set, and the comparison of predicted targets with known true values of the corresponding targets in the test set.

Supervised learning applied on a numerical target (*e.g.*, survival time) is called "regression", and the model f is called a "regressor" or a "regression model" (see "Linear regression" part in subsection 1.1.2.1). Instead, Supervised learning applied on a categorical target is

called "classification", and the corresponding model a "classifier" or a "classification model" (see subsection 1.1.2.1 except "Linear regression" part).

1.1.1.2 Unsupervised learning

Unsupervised learning is a learning task whose aim is to discover hidden patterns and relationships among data, or to identify a cleaner way to represent data. Due to the lack of a known output (or more formally called "labelled outputs") in the training process, this type of learning is widely used in the field of exploratory data analysis [5].

Usually, data for unsupervised learning lack strict input-output structure. Below are two examples of unsupervised learning:

1. **Hierarchical clustering**: this algorithm is based on the pairwise distance between each of two samples and allows to reveal closeness or similarity among a set of samples. As an example, based on genomic and transcriptomic data, researchers identified ten subgroups of breast cancer, which might help to develop patients-dependent clinic treatments according to diseases' subgroups.

2. **Principle component analysis (PCA)**: this algorithm can be used in dimensionality reduction, so that high dimensional data could be visualized in a two-dimensional plane. For instance, 3,030 samples were visualized in a two-dimensional plot by applying PCA on expression profiles of 20,252 genes [6].

In both examples, the learning tasks are based only on inputs, *e.g.*, transcripts from breast cancer, and gene expression levels across 3,030 samples.

1.1.2 Machine learning models

In order to simplify the description of different machine learning models, a subset of "*Iris* flower data" [7] will be used in some of the subsections below. This dataset contains 100 records, some of which are shown in Table 1.1. Each record contains sepal and petal lengths and widths, together with the species name (two candidate species) of a flower.

1.1.2.1 Supervised machine learning methods

Linear regression

One of the most basic regressors is the "linear regression" model. If it is assumed that the

Sepal length (cm)	Sepal width (cm)	Petal length (cm)	Petal width (cm)	Species
6.9	3.1	4.9	1.5	*versicolor*
5.9	3	4.2	1.5	*versicolor*
6.5	3.2	5.1	2	*virginica*
5.7	2.5	5	2	*virginica*
6	2.2	4	1	*versicolor*
6.1	2.9	4.7	1.4	*versicolor*
5.8	2.8	5.1	2.4	*virginica*
...	...	...	...	...

Table 1.1 *Iris* flower dataset.

target y is petal length and the observation x is sepal length, then the linear regression model can be formulated as

$$y = w_0 + w_1 x \tag{1.1}$$

, where w_0 and w_1 are parameters to be learnt from the data. After being trained on the modified *Iris* dataset, $w_0 = -1.56$ and $w_1 = 1.03$ were found (Fig. 1.1 (A)).

Notably, in the model, there is only one feature included, thus the dimension of the input is 1. As a natural extension of a one-dimensional input, two features x_1 and x_2, can also be taken into account in a linear regression model, *i.e.,*

$$y = w_0 + w_1 x_1 + w_2 x_2 \tag{1.2}$$

For example, x_1 can be sepal length, x_2 petal width, and y petal length, whose corresponding fitted model is visualized in Fig. 1.1 (B).

The input dimension can also be greater than 2, with many features included in the model. However, the visualization of a higher dimensional dataset (*e.g.,* a four-dimension dataset - three features together with one target) is not possible.

Logistic regression

Going back to the modified *Iris* dataset, predicting the species of the flowers (note that species is categorical) is a classification task. The latter can be accomplished for example by applying a cutoff to the linear regression model. If including petal length as a feature and the species of a flower as a target, then the classifier can be defined as

$$z = w_0 + w_1 x \tag{1.3}$$

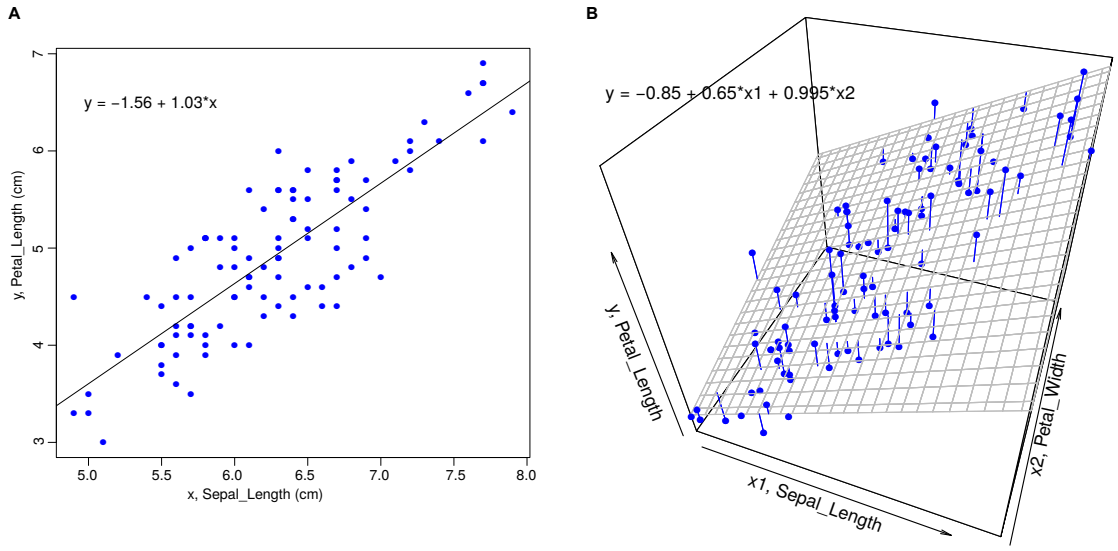


Fig. 1.1 **Linear regression models to predict petal length.** **(A)** Data and learnt model based on one feature. The x-axis represents feature "sepal length" and the y-axis the target "petal length". Each blue dot shows the petal and sepal length of a flower. The fitted model is indicated by the grey line. **(B)** Data and learnt model based on two features x_1 (sepal length) and x_2 (petal width), and the target (y axis) petal length. Each blue dot indicates a flower's sepal and petal length, and its petal width. The learnt model is indicated by the grey plane.

$$y = \begin{cases} \textit{versicolor} & \text{, when } z \leq 0 \\ \textit{virginica} & \text{, when } z > 0 \end{cases} \tag{1.4}$$

Equation (1.3) is almost identical to Eq. (1.1), with the only difference being that, in Eq. (1.3), the variable has been changed into an intermediate variable z. Eq. (1.4) shows the cutoff approach (where cutoff equals 0). In Eq. (1.5) "*versicolor*" is replaced with value 0 and "*virginica*" with value 1:

$$y = \begin{cases} 0 & \text{, when } z \leq 0 \\ 1 & \text{, when } z > 0 \end{cases} \tag{1.5}$$

Eq. (1.5) is called a step function and usually replaced by a sigmoid function Eq. (1.6) for mathematical convenience:

$$y = \frac{1}{1+e^{-z}} \tag{1.6}$$

Values of function (1.6) lie between 0 and 1 and represent the predicted probability of y being "*virginica*". Eq. (1.6) is visualized as the blue curve in Fig. 1.2, which also shows Eq. (1.4) in dark red lines. It can be observed that the shape of the blue curve does approximate the tendency of dark red lines.

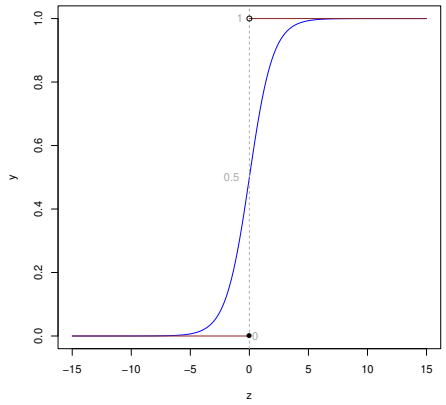


Fig. 1.2 **The sigmoid function and a step function.** The blue curve shows the shape of the sigmoid function defined by Eq. (1.6), while dark red lines a step function defined by Eq. (1.5).

Eq. (1.3) together with Eq. (1.6) form the simplest version of a classification model. For historical reasons, this model, instead of being named as binary logistic classification, got the name of "binary logistic regression". Yet, this is a classifier, instead of a regressor (see section 4.4 of [8] for more details).

K-nearest neighbors (KNN)

K-nearest neighbors (KNN) [9] can be used to solve both regression and classification tasks. Differently from a logistic regression, which learns a set of function parameters, KNN directly estimates a target corresponding to an input based on the input's neighborhood.

Fig. 1.3 demonstrates how a KNN makes target class prediction on the modified *Iris* dataset. The target class (*virginica* or *versicolor*) of a new observation (orange dot in the figure) is predicted on its neighborhood.

In this case, five nearest neighbors (five other observations with known classes) are chosen for the task. Two neighbors are *virginica* and three are *versicolor*, resulting in a final class prediction for this new observation being *versicolor*.

For KNN, k is a critical parameter. In the above example, $k = 5$, and for each prediction, five neighboring samples are chosen for the classification process.

Tree-based models

Tree-based models are used for both classification and regression. For classification tasks, a "decision tree" is the simplest tree-based classifier, which predicts a class based on a series

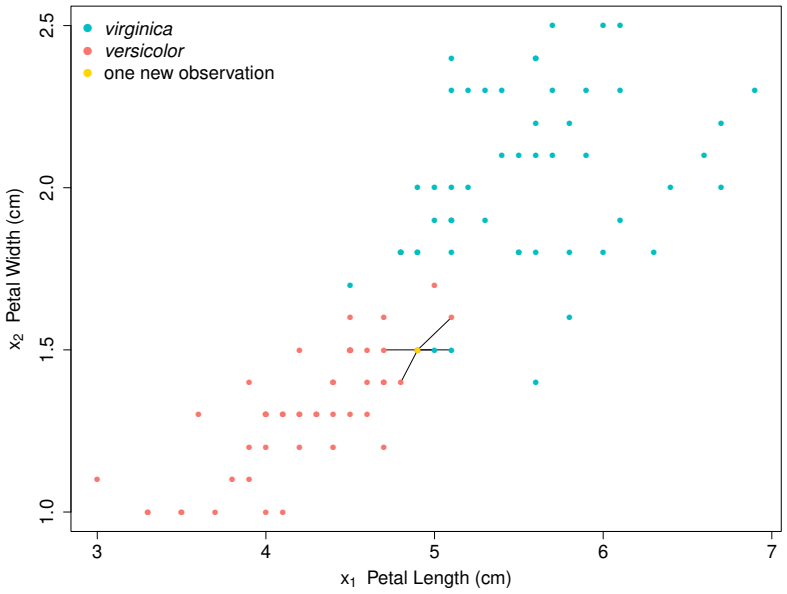


Fig. 1.3 **A KNN classifier.** Each dot in the figure represents a sample, whose observed petal length and width are indicated as x_1 and x_2 coordinate, respectively. The flower classes (*virginica* the cyan dots, and *versicolor* pink)are known for all the samples except one new sample (orange dot). Five neighboring observations, which are connected to the orange dot with black lines, are chosen to vote for the class of the new observation.

of learnt rules. As shown in Fig. 1.4, the tree is trained based on two features - petal length and width - and the target to be predicted is its species (*versicolor* or *virginica*).

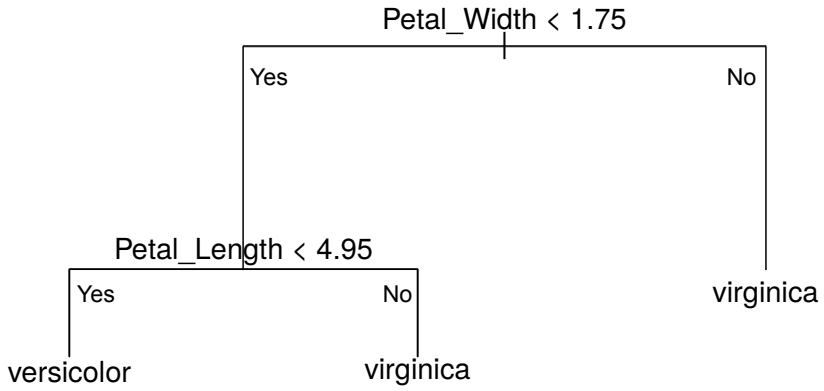


Fig. 1.4 **A fitted decision tree classifier to predict flower species.** The features of the model are petal length and width, and the target is species. The classifier makes a final prediction based on whether petal width is less than 1.75 cm, and whether petal length is less than 4.95cm, sequentially.

A fitted decision tree classifier is highly interpretable. For example, in Fig. 1.4, the model first checks petal width. If the width is not lower than 1.75 cm, the predicted class is *virginica*; otherwise, it continues to judge whether the petal length is less than 4.95 cm. If so, the estimated target is *versicolor*; otherwise, it is *virginica*.

Though a decision tree has the advantage of explainability, it usually lacks enough model complexity for interpretation of biological data. As an improvement, other tree-based "ensemble learning" models have been developed, including bagging, random forest and Boosting. These models make predictions based on a combination of multiple decision trees. However, they differ in many technical details, *i.e.,* choice of sub-training set or sub-features set when generating a single tree, and aggregation style when combining all trees (see [10] for more details).

Additional models for classification are Naive Bayes [8], support vector machines (SVM) [11, 12] and neural network [8], and so on.

1.1.2.2 Unsupervised machine learning methods

Principal components analysis (PCA)

In the modified *Iris* set, there are in total four features including sepal length, sepal width, petal length and petal width, which form a four-dimensional feature space. Thus, the i'th observation can be represented by a dot $\mathbf{x_i} = (x_{i1}, x_{i2}, x_{i3}, x_{i4})$ in a four-dimensional space, where x_{i1} is the sepal length of sample i, x_{i2} the sepal width, x_{i3} the petal length and x_{i4} the

petal width. It is not possible to visualize all these four features by a two-dimensional plot. A naive approach is to select a subset of the two most representative features, *e.g.,* (x_{i1}, x_{i2}) for sample i. A more advanced approach such as principal components analysis (PCA) combines several of the original features to form two new "latent" features that would capture most of the information in the original data.

"Information", in this context, means closenesses (or variances) between each pair of observations. To exemplify, given three samples A, B, and C, if A and B are close to each other (*e.g.*, A-B distance equals 1), and C is further away from dot A (e.g., A-C distance equals 100) and B (B-C distance equals 200), an ideal set of two-dimensional latent features should reflect the original closeness of these observations. To this goal, A is represented by a dot of coordinate $\mathbf{y_A} = (y_{A1}, y_{A2})$, B $\mathbf{y_B}$ and C $\mathbf{y_C}$, so that in the newly found two-dimensional latent feature space, dot A $\mathbf{y_A}$ should be close to dot representing B $\mathbf{y_B}$, while dot C $\mathbf{y_C}$ should be far away from both $\mathbf{y_A}$ and $\mathbf{y_B}$.

As an example, two-dimensional latent feature representation of the modified *Iris* dataset is learnt. Each sample in the dataset is now visualized in a two-dimensional figure Fig. 1.5. The first principal component (PC1) is the first latent feature identified, and the second principal component (PC2) the second. PC1, PC2, in this example, represent 83.72%, 8.61% of the variance of original features, respectively. Though PCA is not a classification learning approach, samples of the same species tend to cluster in this biplot, showing that the latent features were learnt successfully and captures key information for species discrimination.

In principle, more than two latent features can be learnt by PCA, as long as the number is smaller than or equal to the number of features in the original data. Usually, PCA is applied on a normalized dataset, if the dataset has original features in different scales or has extremely different variance across features.

Clustering

Clustering can be used to search subgroups of observations based on a set of features. A widely applied clustering algorithms is “hierarchical clustering (HC)”, which clusters observations hierarchically according to inter-observations similarities. The result of HC-based analysis is usually visualized by a dendrogram, which is a tree-like diagram displaying closeness between observations.

As shown in Fig. 1.6, HC is applied to observations in the modified *Iris* dataset. The four input features are petal length, width, and sepal length and width (note that HC is fully blind of flower class information during training).

HC first regards all observations as one single cluster (n clusters). (i) Then, distances (or dissimilarities) between each pair of observations are calculated, based on the features of each

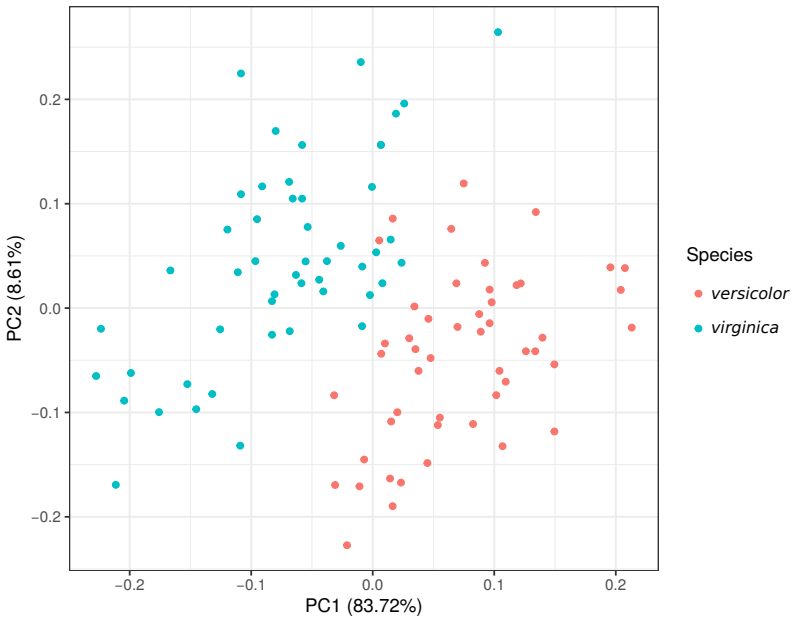


Fig. 1.5 **PCA biplot on the modified *Iris* dataset**. Each dot represents one sample and the color indicates the species (*versicolor*lor in pink and *virginica* in cyan).

observation. Next, each pair with the shortest distance is fused into one cluster. (ii) A set of representative features is then generated for each cluster. The way to generate representative features is linkage process. (iii) The algorithm performs step (i) and (ii) iteratively, based on the representative features until all observations are included into one cluster. To this goal, different distance measurements can be chosen, *e.g.,* Euclidean and Hamming distance as well as correlation-based distances, *e.g.,* Pearson correlation and Spearman correlation.

If one manually defines a place to cut the HC dendrogram and assigns class labels to each cluster, HC then serves as a classifier. The performance of an HC-derived classifier relies on how well the features can predict the interested class labels. In the example, the class label is the species of a sample. If the four features fully capture the species differences, when cutting the HC dendrogram near the root, only two clusters should be obtained - each corresponding to a species. Instead, if the four features are not able to reflect species difference, samples from two species should not cluster but be scattered randomly in the tree. HC of biological data usually results in samples from the same class clustering in smaller branches, as shown in Fig. 1.6.

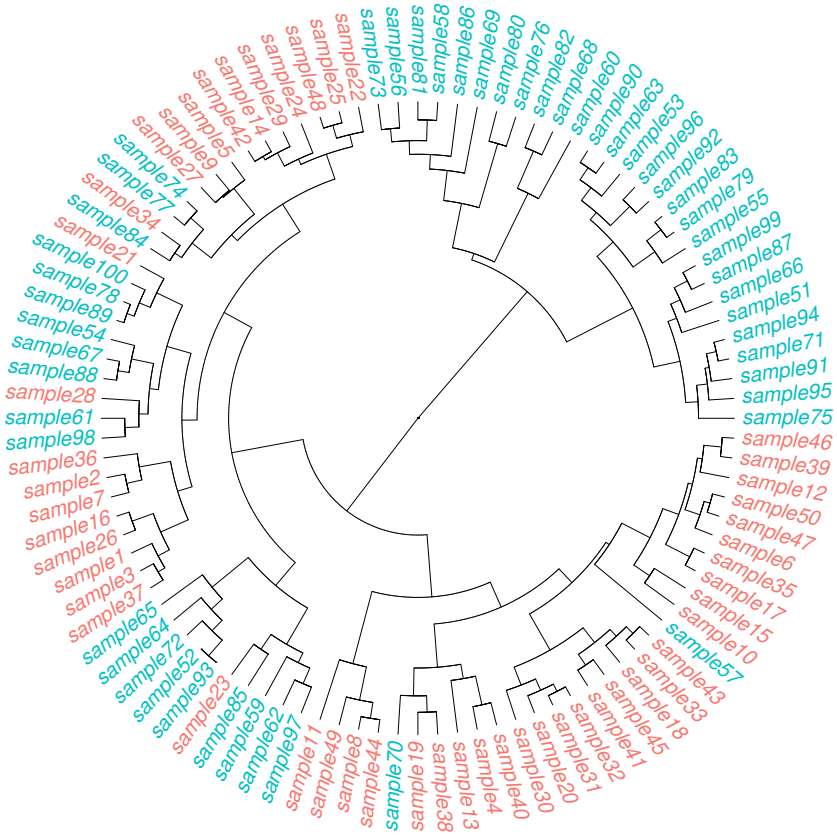


Fig. 1.6 **Hierarchical clustering on *Iris* dataset.** Four features including sepal length and width, petal length and petal width, are used to calculate the distances among observations. To benchmark, samples are color-coded according to their species classes (pink for *versicolor* and cyan for *virginica*).

1.2 Public biological resources

With the advent of ‘omics, there is an increasing volume of publicly available datasets (*e.g.*, transcriptomes from microarray and RNA-Seq; proteomics and protein-protein interactions; metabolomics). Besides, biological knowledge is also increasing (*e.g.*, gene and protein sequences; gene names or gene ID mapping across various resources; functional annotations; domains occurrence; molecular associations from either experiments or computational predictions). Therefore, several public databases have been developed to store biological data and knowledge.

1.2.1 GEO

Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) archives functional genomic data including transcriptomes from microarray and next-generation sequencing, genome-wide methylation, chromatin structure and protein-protein interactions. A GEO sample (coded in the form of "GSMxxx" in GEO) usually includes data acquired from one sample, together with the meta-information on the sample (*e.g.*, condition, treatment, platform; coded in the form of "GPLxxx"). Several related GEO samples form a GEO series (coded in the form of "GSExxx"), which usually corresponds to all samples in a study. Samples can form another type of cohort in GEO - DataSet, which refers to a collection of comparable GEO gene expression samples from the same platform. The GEO DataSet is usually defined by GEO curators. A useful feature related to a DataSet is the GEO profile, which contains the expression of the same gene across all samples in that DataSets. GEO archives both raw and processed data, which are available for query and can be freely downloaded. For example, Entrez Programming Utilities (E-utilities) https://www.ncbi.nlm.nih.gov/home/tools/ can be used to query GEO programmatically. Besides, an interactive webserver GEO2R, mainly based on Biocondutor packages GEOQuery [13] and limma [14], can be used for identification and visualization of differentially expressed genes in GEO series.

1.2.2 ArrayExpress

Hosted by EMBL-EBI, ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) is another widely used database storing functional genomics data. Raw microarray data and experimental details can be directly submitted to ArrayExpress, whereas raw sequencing data generated by high-throughput technologies are submitted to the European Nucleotide Archive (ENA, https://www.ebi.ac.uk/ena/data/view/PRJNA292718) and their corresponding experimental

details and processed data are archived in ArrayExpress. Submitted datasets are saved as "experiments" in ArrayExpress. An experiment refers to a group of assays involved in one study or publication, where one assay indicates either a set of biological replicates for one microarray experiment or the sequencing output of one library for high-throughput sequencing. ArrayExpress also includes experiments from external databases, *e.g.*, GEO. ArrayExpress can be accessed using JSON programmatically and FTP accession is also available for bulk download.

1.2.3 UniProt

The UniProt Consortium (see Fig. 1.7) comprises three research teams - the European Bioinformatics Institute (EMBL-EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR). UniProt integrates cross-species protein sequences and annotations from multiple databases, including GenBank, Ensembl, RefSeq, protein data bank (PDB) *etc.* Originally residing in multiple databases, sequences were sorted to avoid redundancy and then assigned with permanent, immutable unique identifiers in UniProt. UniProt's protein annotations detail such as taxonomy, gene ontology, subcellular localizations, references to external databases and associated human diseases. For each species, unique proteome identifier (UPI) is assigned to indicate individual proteomes. Reference proteomes are proteomes of organisms, which are either well-studied or used for biomedical studies.

1.2.4 PRIDE

The PRoteomics IDEntifications database (PRIDE) was set up in 2004 by EMBL-EBI to store original proteomics data. Apart from proteins or peptides identified by mass spectrometry, PRIDE also collects annotations of proteins' modifications from literature. During the past years, a large number of proteomics data have been submitted to PRIDE, making it the largest proteomics data warehouse [16]. PRIDE allows the use, re-use, reprocess, reanalysis and integration of proteomics data from different cell lines, tissues or species. PRIDE can be mined through web interface or PRIDE Inspector Toolsuite can be used to visualize proteomics data and to assess the quality of ProteomeXchange datasets [17]. The Restful website service (https://www.ebi.ac.uk/pride/ws/archive/) helps to access the archive programmatically, and the file repository (https://asperasoft.com/) allows users to download data via file transfer protocol (FTP) or Aspera.

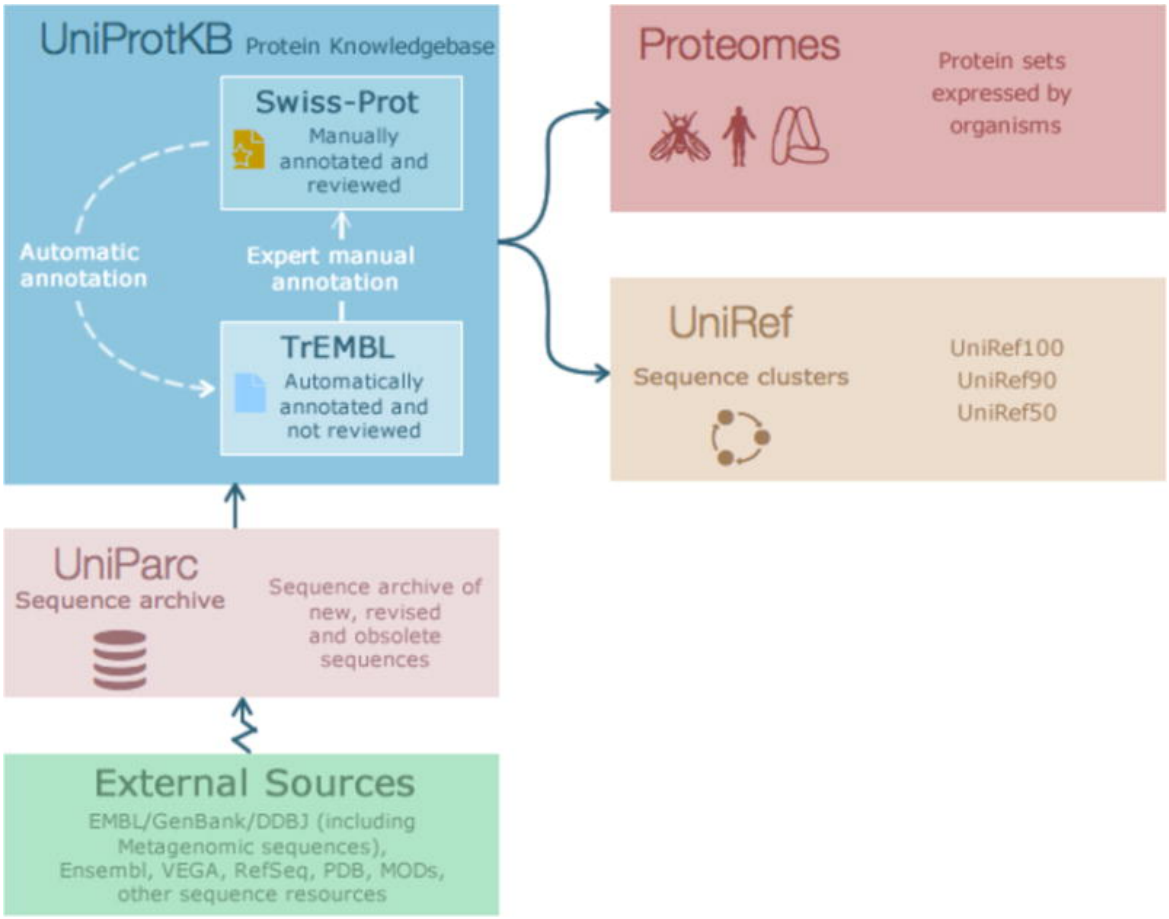


Fig. 1.7 **Overview of four UniProt databases.** Taken from reference [15] (Fig 1). This diagram shows four core datasets hosted by UniProt - the UniProt Archive (UniParc), the UniProt Reference Clusters (UniRef), the UniProt Knowledgebase (UniProtKB) and the Proteomes. UniParc is a hub for storing all publicly accessible protein sequences. UniProtKB includes functional annotations of proteins and is composed of two sub-datasets - reviewed dataset (Swiss-Prot) and unreviewed dataset Translated EMBL Nucleotide Sequence Data Library (TrEMBL). Protein annotation in Swiss-Prot is based on expert curation result while the information in TrEMBL is from computational prediction. UniRef clusters proteins stored in UniProtKB according to sequence similarity at identity levels of 100%, 90% and 50% (UniRef100, UniRef90, UniRef 50). Proteomes contains whole proteomes of several species and is a good resource for functional analysis of species-based proteome.

1.2.5 KEGG

Kyoto Encyclopedia of Genes and Genomes (KEGG, https://www.genome.jp/kegg/pathway.html) project represents a set of widely-used databases, *i.e.,* genomic information (KEGG ORTHOLOGY, KEGG GENOME, KEGG GENES, KEGG SSDB), systems information (KEGG PATHWAY, KEGG BRITE, KEGG MODULE), chemical information and health information (Fig. 1.8). KEGG's biological entities, *e.g.,* genes and proteins, pathways and reactions, are assigned to unique identifiers, which are usually defined in the format of a database-dependent prefix followed by five digits. For instance, map00010 indicates a pathway map from KEGG pathway, and T01001 represents the complete human genome from KEGG GENOME. KEGG PATHWAY stores manually curated pathway maps of molecular interactions and reactions. KEGG MODULE contains manually defined functional units, *e.g.,* pathway modules for metabolic pathway maps, structural complexes forming molecular machinery, other essential functional sets and signature modules, which are used as phenotype markers. KEGG BRITE provides hierarchical relationships among various biological objects, thus helps users to understand cross-talk among genes, proteins, drugs, diseases, *etc*. KEGG ORTHOLOGY (KO) manually defines functional orthologs based on KEGG pathway maps, KEGG Modules as well as hierarchies in KEGG BRITE. Moreover, chemical information and health-related knowledge are collected and saved in multiple KEGG databases, including KEGG COMPOUND, KEGG DISEASE, *etc.* All KEGG databases can be easily accessed by website and KEGG API.

1.2.6 STRING

In living cells, proteins often interact to fulfill dynamic biological processes. Interactions can occur either by direct physical contacts or by indirect functional cooperation. Protein-protein interactions (PPIs) can be identified by experimental methods, *e.g.,* yeast two-hybrid screening [18], affinity purification followed by mass spectrometry [19], protein-fragment complementation assay [20], *etc.* Experimentally identified PPIs are collected and stored in multiple resources, *e.g.,* biomolecular interaction network database (BIND) [21], database of interacting proteins (DIP) [22] and biological general repository for interaction datasets (BioGRID) [23]. Apart from experimental-derived interactions, computationally predicted interactions from integrated analyses of large-scale data, *e.g.,* expression, orthology, protein domains, can also be generated. STRING [24] collects both experiment-verified and computational-based interactions from a variety of databases, *e.g.,* BIND, DIP, BioGRID, Human protein reference database (HPRD) [25], IntAct [26], the Molecular Interaction database (MINT) [27], Pathway Interaction Database (PID) [28], gene ontology (GO) [29], KEGG

Category	Database	Content	Color
Systems information	KEGG PATHWAY	KEGG pathway maps	
	KEGG BRITE	BRITE hierarchies and tables	
	KEGG MODULE	KEGG modules	
Genomic information	KEGG ORTHOLOGY (KO)	Functional orthologs	
	KEGG GENOME	KEGG organisms (complete genomes)	
	KEGG GENES	Genes and proteins	
	KEGG SSDB	GENES sequence similarity	
Chemical information	KEGG COMPOUND	Small molecules	
	KEGG GLYCAN	Glycans	
	KEGG REACTION	Biochemical reactions	
	KEGG RCLASS	Reaction class	
	KEGG ENZYME	Enzyme nomenclature	
Health information	KEGG NETWORK	Disease-related network elements	
	KEGG VARIANT	Human gene variants	
	KEGG DISEASE	Human diseases	
	KEGG DRUG	Drugs	
	KEGG DGROUP	Drug groups	
	KEGG ENVIRON	Health-related substances	

Fig. 1.8 **Categories in KEGG project.** Taken from KEGG website https://www.genome.jp/kegg/kegg1a.html. KEGG consists of 18 databases which can be categorized into systematic, genomic, chemical and health information.

[30] and Reactome [31], *etc*. Moreover, STRING assigns every PPI a confidence score via integrating seven types of interaction evidence, *i.e.,* conserved genomic neighborhood, the co-occurrence of protein pairs in phylogenetic profiles, gene fusion, coexpression of gene pairs within the same or across different species, experimental confirmation, curated information from databases and text mining. When querying STRING with a single gene or a batch of genes, users will obtain a PPI network, which can be easily personalized via changing parameters (*e.g.,* evidence types, number of interactors). The network can be easily downloaded in text or image formats. Besides, function enrichment analysis and clusters prediction with k-means or MCL algorithms are also provided. As an example, the mitochondrial calcium uniporter (MCU) channel network in human is shown in Fig. 1.9. STRING data can be freely accessed via its website, REST API or Cytoscape backend API.

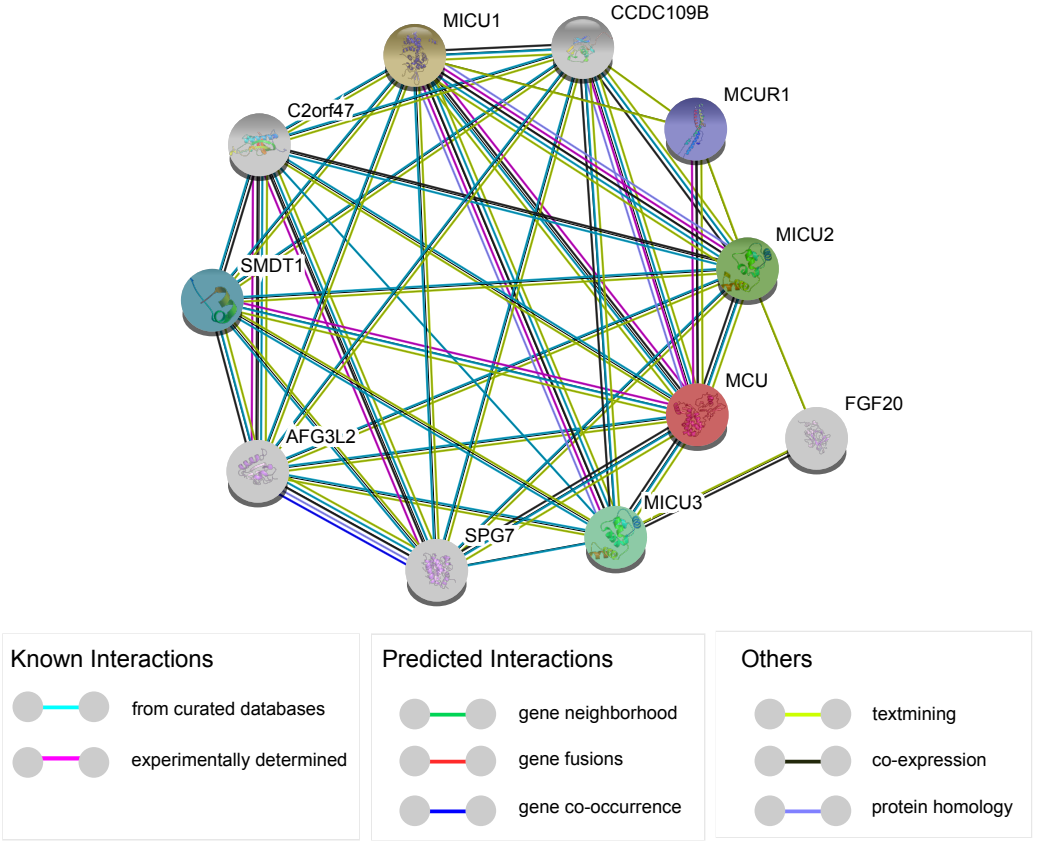
1.3 Application of machine learning in biological research

The availability of machine learning approaches together with the massive amount of biological data stored in various databases, has allowed scientists to integrate dispersive knowledge and generate models to predict biological tasks, such as categorical information of genes or proteins (*e.g.*, subcellular location), relationship between biomolecules (*e.g.*, functional interactions or binding sites) and function-related clusters (*e.g.*, protein complexes). Below are two examples of machine learning-based analysis of biological data.

1.3.1 Application in prediction of mitochondrial proteins

Mitochondria are crucial organelles playing essential roles in cellular metabolic pathways including oxidative phosphorylation which forms ATP, apoptosis and calcium homeostasis [32]. Dysfunction of mitochondrial activity leads to a group of metabolic disorders in human, *e.g.,* type 2 diabetes [33], cancer [34] and a series of neurodegenerative diseases [35]. Thus, to search for effective therapies, deeper understanding of mitochondria and its biology is indispensable. To this end, a comprehensive repository of mitochondrial resident proteins might be beneficial. The first step towards defining a comprehensive mitochondrial proteome parts list was to apply top-down systems level approaches for the systematic and unbiased identification of mitochondrial proteins in the yeast *S. cerevisiae*. In this study, Perocchi *et al.* applied a linear classifier to integrate 24 genome-scale features that provide complementary clues on properties of mitochondrial proteins such as expression profile, subcellular localization, evolution, conservation, and loss-of-function phenotype [36]. As shown in Fig. 1.10, a linear classifier was trained on a reference set of 434 known mitochondrial

A PPI Network of MCU



B Functional Enrichments of Network

Biological Process (GO)			
GO-term	*description*	*count in gene set*	*false discovery rate*
GO:0006851	mitochondrial calcium ion transmembrane transport	10 of 21	7.14e-26
GO:0036444	calcium import into the mitochondrion	8 of 10	2.25e-21
GO:0051560	mitochondrial calcium ion homeostasis	8 of 23	2.70e-19
GO:0006839	mitochondrial transport	9 of 223	6.44e-15
GO:0051561	positive regulation of mitochondrial calcium ion concentrat...	4 of 10	7.17e-10
			(more ...)

PFAM Protein Domains			
domain	*description*	*count in gene set*	*false discovery rate*
PF06480	FtsH Extracellular	2 of 2	1.72e-05
PF04678	Mitochondrial calcium uniporter	2 of 2	1.72e-05
PF01434	Peptidase family M41	2 of 3	1.72e-05
PF13833	EF-hand domain pair	3 of 95	4.87e-05
PF00004	ATPase family associated with various cellular activities (A...	2 of 56	0.00093

Fig. 1.9 **(A) PPI network of MCU.** The PPI network is generated by querying STRING with MCU gene in human, setting interaction confidence score greater than 0.9 and no more than 5 interactors for both first and second shell. Edge colors indicate different types of interaction evidence. **(B) Functional enrichment analysis of genes involved in network (A).** STRING allows functional enrichments analysis in multiple aspects, *e.g.,* biological process (GOBP), Reactome pathways, protein domains.

genes collected from MitoP2 and predicted 346 novel mitochondrial proteins, some of which were confirmed to localize in mitochondria. Next, the authors used STRING to generate a PPI network of the mitochondrial proteome, and HC was used to generate 164 functional modules. The function of 46 modules ($\geq$ 5 proteins) was matched with either KEGG pathways (23 modules) or protein complexes in *Saccharomyces* Genome Database (SGD) (13 modules). More importantly, mutant phenotype profiles and gene expression changes for each predicted module were implemented under non-fermentable and fermentable conditions. Finally, via analyzing protein evolution between human and yeast, they successfully identified candidate genes, which were involved in mitochondrial disorders.

In another study, Calvo *et al.* applied a Naive Bayes model named Maestro to generate a mitochondrial parts list of mammalian mitochondria [37]. The authors compiled a set of targets including 654 and 2817 mitochondrial and non-mitochondrial proteins, respectively, as well as a collection of eight genome-wide features, including: prediction of mitochondrial targeting sequence; presence of protein domains specific for either mitochondrial or non-mitochondrial proteins or shared by both; presence of protein homologs in yeast mitochondria; protein homologs in the bacterial ancestor of mitochondrial *Rickettsia prowazekii*; coexpression with confirmed mitochondrial genes across various tissues; proteomics in isolated mitochondria from four mouse tissues; measurement of upregulation of mRNA after induction of mitochondrial biogenesis. By applying Maestro on the whole human genome, the authors predicted a comprehensive repository of 1080 mitochondrial resident proteins, including 490 novel ones. To assess the accuracy of prediction, the authors applied a ten-fold cross-validation and two types of experimental approaches, including targeted proteomics techniques and fluorescence microscopy with epitope tagging. Moreover, they successfully identified candidates that are involved in mitochondrial disorders . Two years later, the same group reported a more updated compendium known as MitoCarta, which was comprised of three parts [38]. One part of the compendium was generated by applying a naive Bayesian model, which integrated not only newly generated mitochondrial mass spectrometry proteomics across 14 tissues but also six previously mentioned features, such as coexpression and specific protein domain. The second part of MitoCarta was from a large scale of GFP tagging and microscopy analysis, which was performed on mammalian mitochondria, as shown in Fig. 1.11. Moreover, after scanning phylogenetic profiling across 500 fully sequenced species for each of the protein in MitoCarta, the authors successfully detected 19 proteins, which are sharing similar evolutionary profiles with a majority of complex I subunits, and some of them might be involved in complex I deficiency.

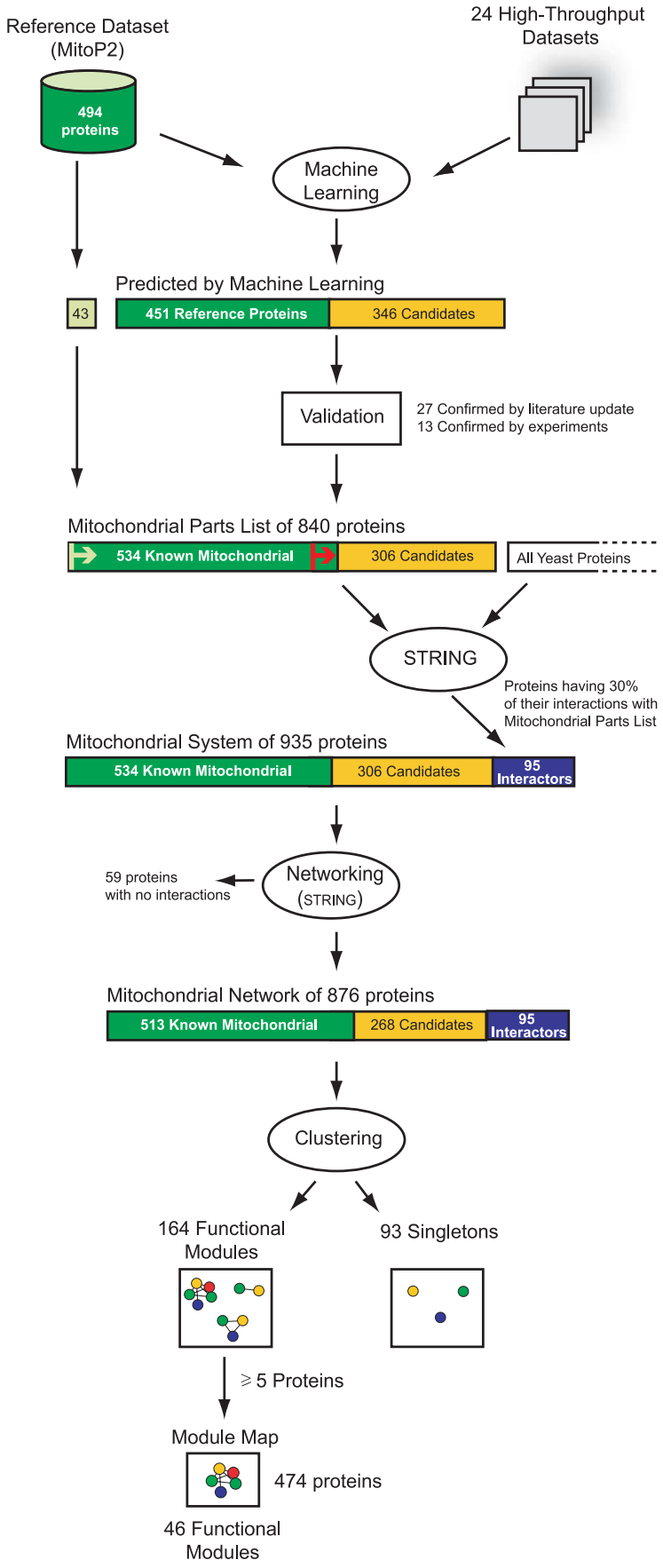


Fig. 1.10 **The integrated approach for yeast mitochondria interaction map.** Taken from paper [36].

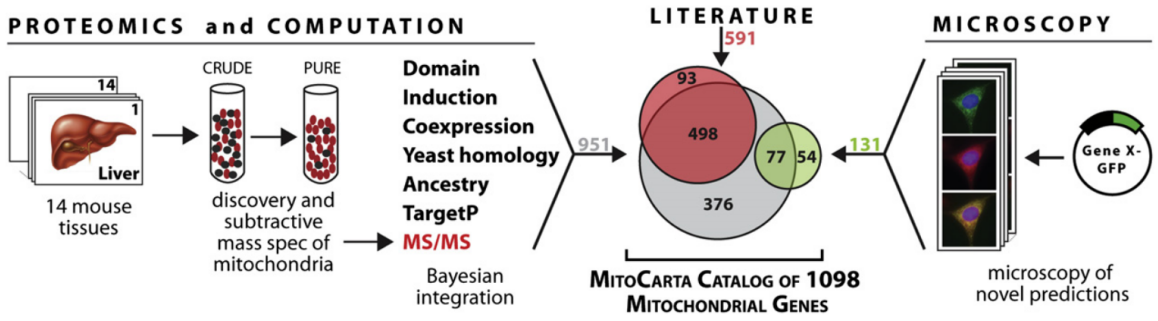


Fig. 1.11 **Building the MitoCarta.** Taken from MitoCarta [38]. MitoCarta comprises in total 1098 mitochondria localized genes which are defined by at least one approach shown in the Venn diagram: (1) gray circle represents prediction of Maestro which integrated seven genome-wide features including the presence of protein domains specific in either mitochondrial or non-mitochondrial proteins or shared by both (Domain), n-fold change of message RNA expression after inducing mitochondrial proliferation (Induction), coexpression with confirmed mitochondrial genes calculated based on message RNA expression across various tissues (Coexpression), presence of mitochondrial homology in yeast (Yeast homology), protein similarity to protein sequences in *Rickettsia prowazekii* (Ancestry), prediction of mitochondrial targeting sequence (TargetP) and mitochondrial mass spectrometry proteomics across 14 tissues (MS/MS), (2) green circles indicates genes from the extensive study of GFP tagging and microscopy, and (3) red circle displays genes collected from literature.

1.3.2 Application in prediction of microRNAs

Machine learning has also been widely used in the study of microRNAs (miRNAs), one type of small RNAs (sRNAs). The biogenesis of miRNAs (usually 22nt in length) takes multiple steps. Firstly, miRNA genes are transcribed into transcripts of primary miRNAs (pri-miRNAs), which are then processed into precursor miRNAs (pre-miRNAs), from where mature miRNAs are generated [39]. Binding to argonaute (Ago) proteins, miRNAs target specific mRNAs via base pairing, which results in repression or cleavage of specific mRNAs [39]. As reviewed by Zhang *et al.* [40], an increasing number of sRNA studies revealed their diverse roles in regulating cellular development or disease, and in diagnosis and therapy. Therefore, a comprehensive understanding of functions of miRNAs renders the necessity of identifying miRNAs from the whole genome, a task which can be achieved by cost-effective and speedy computational algorithms. In 2007, Jiang *et al.* developed MiPred to predict pre-miRNAs by applying a random forest model on sequence-derived features [41]. Their features included: predicted secondary-structure-based minimum of free energy (MFE); composition of 32 types of triplet structure-sequence elements, which were defined by the status of each nucleotide (paired or unpaired) in the secondary structure [42]; a calculated P-value, which represented the MFE difference between the original input sequence and the

List of 125 features used in this work

Feature groups	No. of features	Feature symbol
Sequence-based features	19	Len, %G+C,% A+U, %AA, %AC, %AG, %AU, %CA, %CC, %CG, %CU, %GA, %GC, %GG, %GU, %UA, %UC, %UG, %UU
Secondary structure features	30	MFE, efe, MFEI1, MFEI2, MFEI3, MFEI4, dG, dQ, dD, dF, Prob, zG, zQ, zD, zF, nefe, Freq, diff, dH, dH/L, dS, dS/L, Tm, Tm/L, **MFEI5, MFE/Mean_dG, dH/loop, dS/loop, Tm/Loop, dQ/Loop**
Base pair features	32	dP, zP, div, tot_bp, stem, loop, A-U/L, G-U/L, G-C/L, %A–U/Stem, %G–C/Stem, %G–U/Stem, Probpair1–10, Avg_BP_stem, NonBP_A, NonBP_C, NonBP_G, NonBP_U, Non_BPP, **%A–U/BP, %C–G/BP, %G–U/BP, Avg_BP_Loop**
Triplet sequence structure	32	A(((, A((., A(.., A(.(,A.((,A.(.,A..(, A . . ., C(((, C((., C(.., C(.(, C.((, C.(., C..(, C. . ., G(((, G((., G(.., G(.(, G.((, G.(., G..(, U. . ., U(((, U((., U(.., U(.(, U.((, U.(., U..(, U. . .,
Structural robustness features (SC-derived features)	12	**SC, SC/tot_bp, SC/Len, SC × MFE/Mean_dG, SC × dP, SC × zG, SC/(1 − dP), SC×dP/(1 − dP), SC/NonBP_A, SC/NonBP_C, SC/NonBP_G , SC/NonBP_U**
Total	125	

Fig. 1.12 **List of 125 features used in predicting pre-miRNAs in [44].** Taken from paper [44]. Features are divided into five groups according to their calculation methods. Features in the first group are calculated based on primary sequence, *e.g.,* Len, sequence length; %G+C, GC-content; %A+U, AU-content; frequency of 16 types of di-nucleotide which include AA, AU, AC, AG, *etc*. A number of 30 features listed in the second class are calculated based on thermodynamic stability, *e.g.,* MFE, minimum free energy; efe, ensemble free energy; dG, normalized minimum free energy per length, $dG = MFE/Len$; MFEI1, minimum free energy index1, $MFE1 = dG/\%G+C$; *etc*. In the third group, a number of 32 features are derived from the number of base pairs in the secondary structure, *e.g.,* loop, the number of loops in the secondary structure; stem, the number of stems in the secondary structure; tot_bp, the total number of base pairs in the secondary structure; dP, length-normalized base-pairing propensity, $dP = tot\_bp/Len$; *etc*. In the fourth group, there are 32 types of features from the composition of three neighboring nucleotides. The symbol '(' means paired nucleotide and '.' indicates unpaired, thus "((." indicates that the first and second nucleotides are paired with some other nucleotides in the sequence while the third not. The listed 12 features in the last group measure the structural robustness, *e.g.,* SC, self-containment index score, ranging from 0 to 1 and real pre-miRNAs usually own higher SC scores between 0.85 and 0.98; SC/Len, length normalized SC; SC/tot_bp, tot_bp-normalized SC, *etc*.

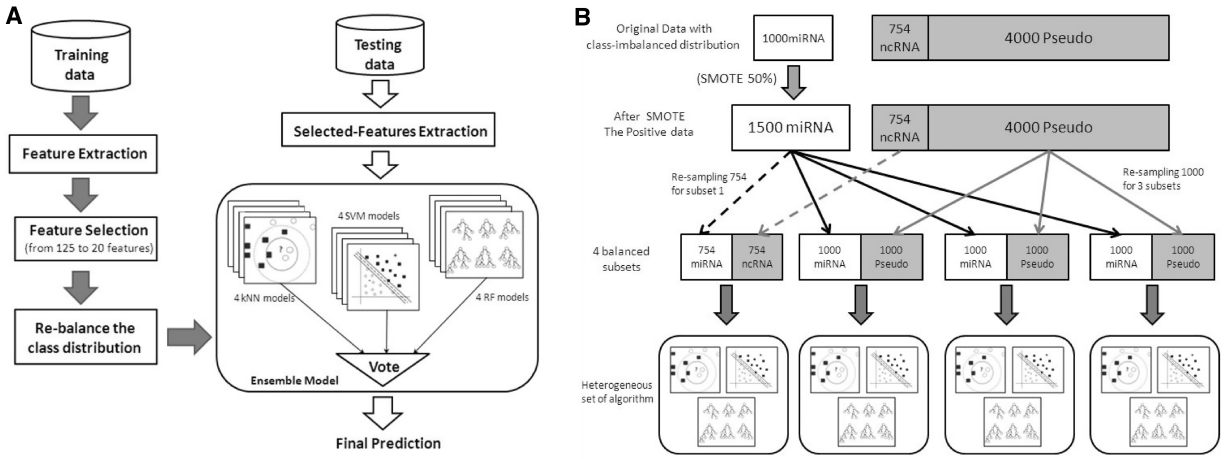


Fig. 1.13 **The pipeline used to predict pre-miRNAs in paper [44].** Taken from paper [44]. **(A)** Overview of the pipeline. Two main steps, *i.e.,* training and test, are taken in the prediction process. A set of targets together with features were extracted and integrated by an ensemble model which makes estimation by combining predictions from 3 types of machine learning techniques, *i.e.,* KNN, SVM and RF. Secondly, a testing dataset is used to test its performance. **(B)** Strategy to balance the training dataset. The training set contains a positive set of 1000 real miRNAs and a negative set comprised of 4000 pseudo miRNAs and 754 non-miRNA ncRNAs. They first enlarge the positive set from 1000 miRNAs to 1500 miRNAs using synthetic minority oversampling technique (SMOTE). Next, via randomly sampling miRNAs from 1500 miRNAs and ncRNAs or pseudo miRNAs from the original negative set, four new training sets were generated and separately trained with 3 different models.

fake sequence, which was generated with di-nucleotide shuffling [43]. Features mentioned above together with a collection of real pre-miRNAs from miRNA registry database and hairpin-contained pseudo miRNA precursors were combined with both a random forest (RF) classifier and a support vector machine (SVM) model, and the RF classifier gave a slightly higher accuracy (1.9%) than SVM. Obtained RF model was tested on data set from *Homo sapiens, Rattus norvegicus, Caenorhabditis elegans, Mus musculus* and *Caenorhabditis briggsae*. Moreover, the authors estimated the importance of features and found that P-value and MFE contributed most in identifying a real pre-miRNA.

In another project [44], Lertampaiporn *et al.* integrated 125 features for miRNA prediction. As shown in figure 1.12, their features comprised of: 19 sequence-based features (*e.g.,* length, GC content); 30 thermodynamics-related features in secondary structure (*e.g.,* MFE, ensemble free energy, normalized entropy by length); 32 base-pair defined features (*e.g.,* the total number of base pairs in the secondary structure, total number of loops or stems); 32 triplet structure-sequence elements; 12 features derived from self-containment index score (SC). Besides, both synthetic minority over-sampling technique (SMOTE) and under-sampling methods were applied to balance the number of targets in the training dataset, after which an ensemble model was generated by integrating predictions from three different machine learning algorithms, *i.e.,* KNN, RF and SVM. This ensemble model was tested on animal and plant data and achieved high accuracy (>93%) in both. In addition to the two introduced models, other tools were also implemented for miRNAs detection, such as miRanalyzer [45] and SMIRP [46].

Both examples mentioned above aim to classify an observation into a specific category, *e.g.,* if a protein is mitochondria-localized or if a sequence is a pre-miRNA.

1.4 Research aims

1.4.1 Predicting adipose browning capacity

Adipose tissue is broadly divided into white and brown, based on key anatomic, structural, molecular and metabolic differences [47]. White adipose tissue (WAT) is specialized to store chemical energy as fat, whereas brown adipose tissue (BAT) can catabolize lipids and glucose for non-shivering thermogenesis, due to the high mitochondrial mass and expression of uncoupling protein 1 (UCP1), a mitochondrial inner membrane protein that dissipates energy from substrate oxidation directly as heat.

Although major WAT and BAT depots are located in anatomically distinct regions, brown-like, UCP1-positive fat cells can be found sporadically and interspersed in various WAT

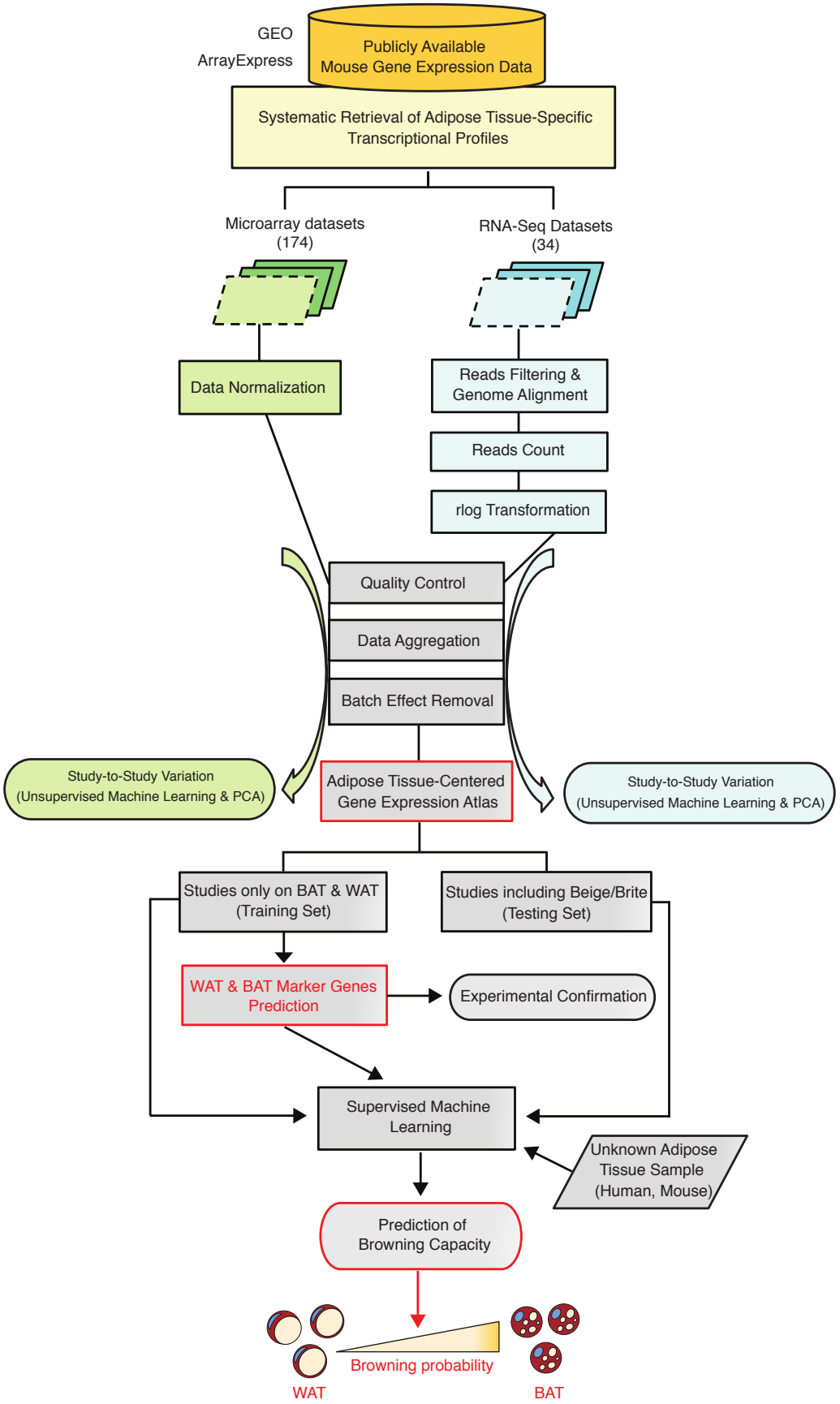


Fig. 1.14 **Pipeline for the Systematic and Unbiased Prediction of Adipose Browning Capacity**

depots in response to cold exposure or β-adrenergic receptor agonists. These cells have been termed beige, brite (brown-in-white), recruitable/inducible brown, or brown-like adipocytes [48], owing to their morphological and metabolic features that are similar to "classical" brown adipocytes and to the expression of thermogenic genes [49]. Several studies have suggested that beige adipocytes can derive from bipotential WAT precursors or mature white adipocytes [50–53]. However, the structural and functional differences that distinguish them from BAT and WAT still remain unclear.

Advance in positron emission tomography (PET)-scanning methods have allowed the discovery that adult humans contain significant deposits of UCP1-positive brown cells in the supraclavicular and neck region [54] as well as in multiple human WAT depots upon exposure to various physiological and pharmacological effectors [55–57]. Promoting the appearance of thermogenic cells in non classical BAT locations can increase energy expenditure and substrate metabolism, improve glucose tolerance, and correct hyperlipidemia, leading to a healthier metabolic phenotype in both rodents [58–60] and humans [61]. Quantifying the browning potential of therapeutic interventions on human BAT activation would therefore accelerate the identification of therapeutic avenues to reduce obesity and its comorbidities. However, this remains challenging, given that human fat contains only a small fraction of brown and brown-like adipocytes.

Lineage-tracing studies for the selective isolation of different adipose cell types have been performed in mice [62] but are not possible in humans. Furthermore, currently available imaging methods have a limited sensitivity and the resulting data are difficult to deconvolute. Besides, there are only a handful of adipose tissue marker genes, which have only been used so far to make a qualitative distinction between human adipocytes and adipose tissue types. Those markers originate from either analyses of whole adipose tissue depots, containing a great proportion of contaminating cells, or *ex vivo* stable and clonally derived adipocytes [55, 63, 64], which are affected by *in vitro* cell culture conditions. Therefore, novel approaches for the unbiased quantification of browning capacity in patients' fat depots are required.

Here, we take advantage of the wealth of data on global transcriptional profiling of fat depots published over the last decade to develop a robust and automated computational pipeline, which we call ProFAT (profiling of fat tissue types), for the systematic prediction of mouse and human adipose browning capacity based on raw gene expression data (Figure 1.14). First, we identify a molecular signature of brown and white adipocytes by integrating 51 and 52 global transcriptional profiles of mouse BAT and WAT from seven independent studies, respectively. Next, we develop a computational model that we train on all 103 datasets and show that it can correctly classify over 80 additional mouse BAT and WAT samples from nine published studies. Importantly, the model can estimate the degree of browning for

WAT-treated samples (beige) independently from biological and technical differences in the anatomical location of the fat depots, experimental models and procedures. We also confirm that our model can be applied to humans and predict the browning capacity of 96 samples derived from heterogeneous tissue biopsies and *ex vivo* immortalized adipocytes. ProFAT is freely available (http://profat.genzentrum.lmu.de) and allows users to automatically perform hierarchical clustering (HC), principal component analysis (PCA) and prediction of browning capacity from raw microarray and RNA-Seq datasets.

1.4.2 Mapping functional crosstalk between mitochondria and ER

Eukaryotic cells are compartmentalized into multiple membrane-bounded organelles, which generate specific cellular micro-environments for certain biological functions [65], such as control of cell's growth and reproduction in nucleus, cellular respiration in mitochondria and protein secretion and Ca^{2+} storage in endoplasmic reticulum (ER). Although organelles are classified into separate cellular structures, more and more studies proved their close communication in performing biological tasks, such as the organization of GPI-anchored proteins between ER and peroxisome [66] and generation of reactive oxygen species (ROS) in mitochondria and peroxisomes [67]. Communication between organelles is implemented through vesicular trafficking [68] or physical membrane contact sites (MCSs) [69], whose sub-components can be identified by biochemical methods, *e.g.,* cellular fractionation [70] and Co-immunoprecipitation (Co-IP) [71], or by image-based approaches, *e.g.,* electron microscopy (EM) [72] and electron cryotomography (cryo-ET) [73]. One of the pivotal MCSs is the physical junction between mitochondria and ER, which is termed as mitochondria-associated membranes (MAMs) and linked to a variety of cellular processes, *e.g.,* ER stress [74], lipid metabolism [75, 73], apoptosis [76] and Ca2+ trafficking [77].

In yeast, the protein complex physically connecting ER to mitochondria is known as ER-mitochondria encounter structure (ERMES), which is composed of protein Mdm12, Mmm1, Mdm10 and Mdm34 [79]. The cytosolic protein Mdm12 acts as the bridge, which links the ER membrane protein Mmm1 and outer-mitochondrial membrane (OMM) protein Mdm10 or Mdm34. Formation of MAMs in mammalian cells is much more complicated considering the nonconservation of core proteins in ERMES and the large number of complexes involved in ER-mitochondrial interaction. Localizing at ER membrane, MFN2 (Mitofusin-2) tethers ER to mitochondria via forming heterotypic or homotypic with mitochondrial membrane protein MFN1 or MFN2 [80]. Another connection of MAMs is Fis1-Bap31 complex (ARCosome), functioning in apoptosis [76]. After the initialization of apoptotic signals by Fis1 in mitochondria, Fis1-Bap31 complex conveys this signal to ER and consequentially activates apoptotic pathway [76]. PTPIP51-VABP complex and IP_3R-Grp75-VDAC complex are two

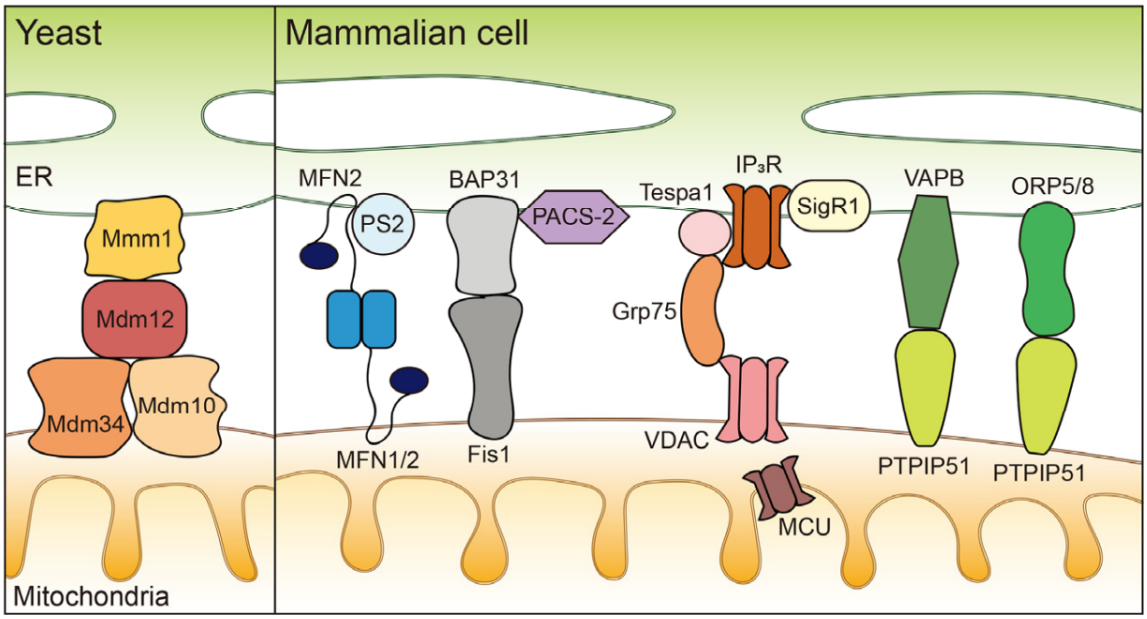


Fig. 1.15 **Complexes forming ER-mitochondria contact sites.** Taken from review [78]. The left part displays the four proteins (Mmm1, Mdm12, Mdm34, Mdm10) forming ERMES in yeast, and the right part shows five protein complexes of MAMs in mammalian cells.

important ER-mitochondria tethers, both of which can regulate Ca^{2+} homeostasis [81, 82]. The variety of roles held by the MAMs has been attributed to their synergistic and dynamic contacts, which are related to their unique protein composition. A growing number of studies describe the relevance of MAMs to cell physiology [83] and diseases [84–87]. However, protein and lipid complement constituting these dynamic interacting structures and how do they interact with each other still remain to be fully elucidated.

To systematically discover ER-mitochondria interactions, it is essential to have a reference proteome for both organelles. Unlike the mitochondrial proteome [88], a comprehensive catalog of ER proteins is still missing. As a membrane-enclosed organelle, ER plays crucial roles in protein synthesis, transport, and folding, as well as in lipid and steroid metabolism and calcium ions (Ca^{2+}) storage [89–91]. ER dysfunction is involved in numerous pathological conditions such as heart failure [92], Alzheimer's disease [93], obesity and type 2 diabetes [94]. The multi-functional and dynamic nature of this organelle requires a large set of proteins as well as functional and physical interconnections with other intracellular organelles and structures. Undoubtedly, the identification and characterization of ER-resident proteins and structures that mediate the crosstalk with other organelles will aid the elucidation of both molecular and mechanistic basis of ER functions, thus shedding lights on potential therapies

for ER-dependent diseases. So a comprehensive and reliable repertoire of ER-resident proteins (ERRP) is important.

A variety of experimental and bioinformatic approaches have been applied to detect or predict ER proteins in several model organisms. Mass spectrometry analyses of ER enriched fractions (microsomes) from mouse [95–97], rat [98, 99] and human [100] tissues have identified several hundreds of associated proteins. However, purified microsomes inevitably contain co-purified contaminants and might miss low expressed proteins. On the other hand, bioinformatically, over a thousand of proteins have been proposed to localize to the ER based on computational predictions by LocTree3 [101] and MultiLoc2 [102], using proteins' sequences as input. But those approaches have limited specificity and sensitivity. These models made their predictions based only on the protein sequence. Moreover, as the training data used by these predictors were directly extracted from the public repositories without manually curating experimental evidence of sub-cellular location, contamination of false positive and false negative proteins in the training data may make the computational tools bear inherent systematic predicting errors. Besides, as these tools are designed for subcellular location prediction, both training data and models are not optimized for ERRP.

There are databases and models designed for ERRPs. Two databases for ERRP were released in 2004 - ER-GolgiDB [103] and Hera database [104]. The former one extracted ERRP from public databases and add new predicted ERRP based on homology information. The latter one manually collected 499 human ER proteins from public database or literature, but only 343 have experimental evidence [104]. To the best of my knowledge, unfortunately, there are no more updated ERRP databases available. In 2017, a machine learning model for ER was optimized and used to predict ERRPs based on protein sequence, using data directly extracted from Swiss-Prot [105]. Yet, this study also suffered from the drawback of lack of reliable training data. Besides, the prediction is also based on the sequence. Considering the recent accumulation of mass spectrometry data and the size and age of the existing ER database, it is necessary to define an updated and reliable list of human ER proteins and integrate more experimental and computational knowledge into a new ER sub-cellular location prediction model.

In this project, we aim to achieve a comprehensive understanding of the molecular players and pathways functionally linking ER and mitochondria, with a focus on Ca^{2+} signaling. To this goal, we have systematically defined a compendium of ERRPs (chapter 3). We first obtained a candidate ERRP set from public databases and literature. Mainly based on this set, we came up with a list of manually confirmed ERRPs, which was used as a positive reference dataset. Meanwhile, a number of 33 helpful features, *e.g.,* ER retention signals, mass spectrometry data and protein-protein interaction, *etc.*, were collected and integrated

by six machine learning models, whose performance were evaluated in a cross-validated fashion. The winning model, Boosting, with an average false discovery rate of 11.5%, was then applied on all human proteins. Our model finally defines a reliable list of 1023 human ER proteins, termed as ERcarta. After searching the OMIM database, we find that 275 genes in ERcarta are involved in 375 diseases. Together with 1307 mitochondrial proteins from MitoCarta2, a comprehensive Mito-ER regulatory network was reconstructed and used to detect small functional modules, which will help shedding light on the functional roles of uncharacterized ER proteins and on the crosstalk between ER and mitochondria (chapter 4). Given our primary interest in characterizing their association in Ca^{2+} signaling, we systematically quantified the effect of knocking down ERcarta genes on mitochondrial Ca^{2+} dynamics by performing a loss-of-function, small interfering RNA (siRNA) screen in HeLa cells that stably express a mitochondria-targeted luminescence Ca^{2+} sensor. Our mt-Ca^{2+} screenings identified a number of 14 potential enhancers and 280 potential inhibitors influencing the level of mt-Ca^{2+}.

Chapter 2

Predicting adipose browning capacity

The results presented in this section are from the manuscript "Prediction of adipose browning capacity by systematic integration of transcriptional profiles" by Cheng, Jiang et al. 2018 without modification [1].

2.1 Results

2.1.1 A Comprehensive Mouse Adipocyte-Centered Gene Expression Atlas

To compile a comprehensive and unbiased gene expression atlas of mouse fat, we systematically retrieved whole-genome transcriptomes from Microarray and RNA-sequencing (RNA-Seq) studies on adipose tissue biopsies and differentiated clonal adipocytes that are publicly available in Gene Expression Omnibus (GEO) and ArrayExpress databases. A total of 16 independent studies on at least two clearly defined adipocytes such as classical brown, white, and inducible brown adipocytes (beige or brite) was selected for downstream computational analyses [106–118, 64, 119, 120] (Table 2.1 and Figure 2.1A), including 174 Microarray and 34 RNA-Seq datasets of high reads quality and correlation between biological replicates (Figures 2.5 and 2.6). This collection includes 83 gene expression datasets on a variety of white fat depots originating from different anatomical locations, such as epididymal, inguinal, gonadal, perivascular, mesenteric and subcutaneous WAT (Figure 2.1B) and mouse models. In addition, it contains 63 gene expression datasets on interscapular BAT and 52 on beige or brite adipocytes originated from different WAT depots in response

to treatments such as cold, PPAR-gamma agonists (rosiglitazone, fexaramine, forskolin, roscovitine), and beta-3 adrenergic receptor agonists (CL316,243; Figures 2.1A and 2B).

2.1.2 Gene expression signatures of brown, white and beige or brite fat

To construct a global adipose tissue-centered gene expression map, we aggregated transcriptional profiles from all Microarray or RNA-Seq-based studies in our atlas (Figure 1.14). First, spurious differences in gene expression between studies, due to technical variation in array platforms and sequencing libraries, were resolved by correcting for batch effects. Next, PCA (Figure Figure 2.1C) and HC (Figures 2.7 and 2.8) were applied to evaluate the relatedness between transcriptional profiles of BAT, WAT and beige or brite-depots-derived datasets from all studies. Both approaches highlighted a strong and robust gene expression signature from BAT and WAT-derived samples, despite their heterogeneous composition. On the whole genome transcriptional level, the variation between WAT depots, due for example to different anatomical regions, proportion of distinct adipocytes, age, food, and gender, had no relevant contribution to the global WAT signature. Furthermore, the gene expression signatures of BAT and WAT were always clearly distinct, independently from the sequencing method (microarray *vs.* RNA-Seq), reflecting robust transcriptional differences in the regulation of their physiology and metabolism. Surprisingly, perivascular WAT (pvWAT) samples from study M9 [108] showed a molecular signature indistinguishable from BAT-derived samples. This result is fully consistent with findings by Fitzgibbons *et al.* that thoracic pvWAT from mice fed either a normal or high-fat diet has virtually identical gene expression profiles to brown adipocytes.

With the exception of samples from Wang *et al.* (study M13), the transcriptional profile of beige or brite adipocytes from other studies was not clearly distinct from either WAT or BAT groups, in both PCA and HC analyses (Figures 2.1C, S3 and S4). For example, gene expression profiles of beige or brite samples from inguinal WAT (iWAT) biopsies of C57BL6 male mice kept in cold for 1-5 weeks (study M8; [119]) were similar to that of BAT samples in the atlas, grouping together in both PCA and HC analyses. On the contrary, beige or brite samples from subcutaneous (sWAT) and mesenteric (mWAT) WAT biopsies of SV129 female mice kept in cold for 10 days (study M10; [113]) showed a gene expression signature similar to WAT samples from the same as well as from other studies. Similarly, beige or brite adipocytes from cold acclimated (study R1; [110]) and fexaramine stimulated (study R2) iWAT and gonadal WAT (gWAT) [107] clustered with WAT samples from other RNA-Seq studies in the atlas, whereas beige or brite adipocytes from iWAT treated with rosiglitazone (study R3; [115]) grouped with BAT samples.

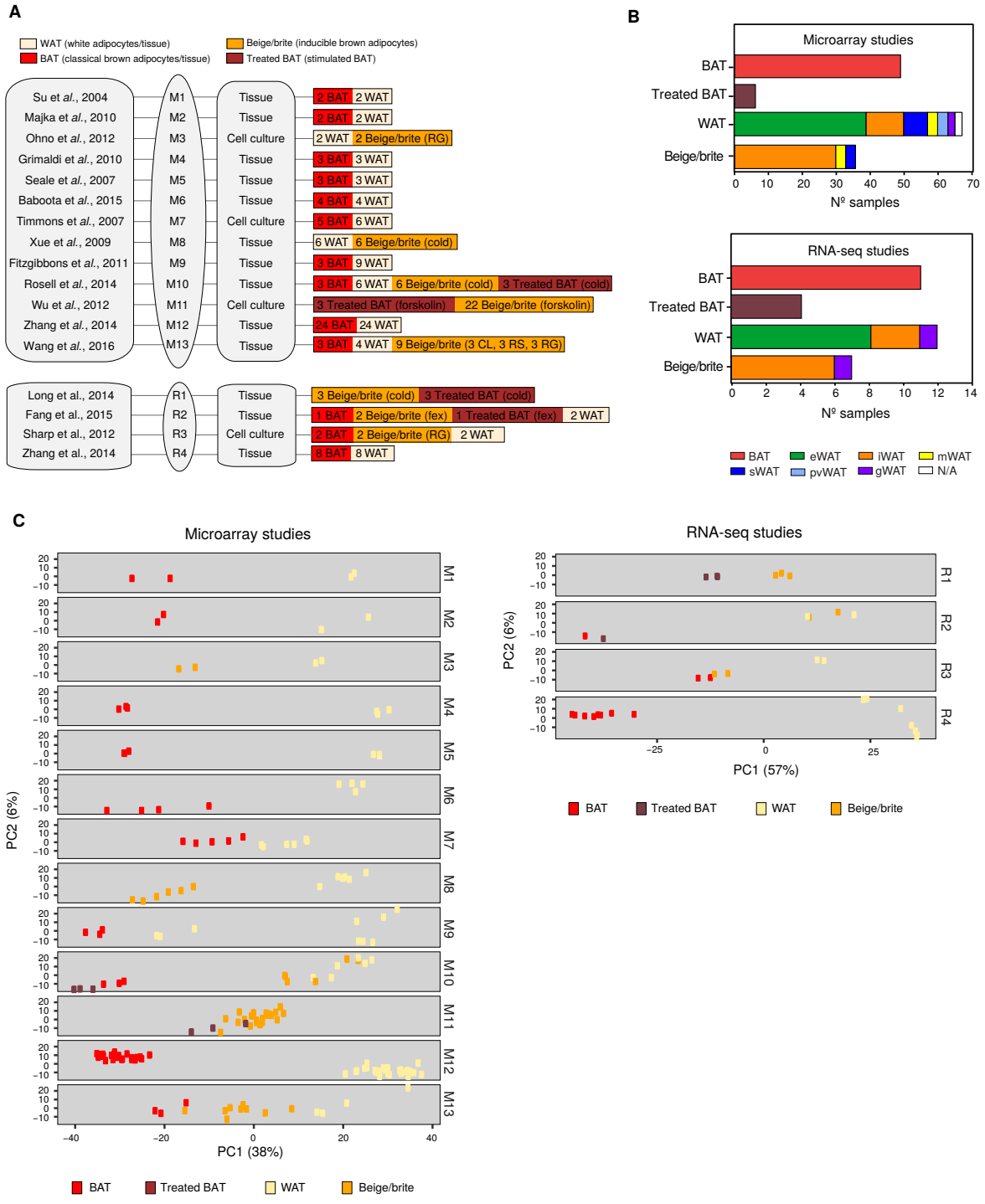


Fig. 2.1 **Mouse-Adipocyte-Centered Gene Expression Atlas.** (A) Summary of microarray (M1-M13) and RNA-seq studies (R1-R4) on fat samples included in the mouse-adipocyte-centered gene expression atlas. The number of samples for each adipose tissue type within a study is indicated. The stimulus applied to induce browning of WAT (beige or brite) is specified in parentheses (CL, CL316,243; fex, fexaramine; RG, rosiglitazone; RS, roscovitine). (B) Sample distribution among different adipose tissue types in all microarray and RNA-seq studies. eWAT, epididymal white adipose tissue; gWAT, gonadal white adipose tissue; iWAT, inguinal white adipose tissue; mWAT, mesenteric white adipose tissue; N/A, not specified; pvWAT, perivascular white adipose tissue; sWAT, subcutaneous white adipose tissue. (C) Study-by-study principle-component analysis (PCA) of normalized gene expression data. See also Table 2.1 and Figures 2.5-2.8.

Taken together, our systematic analysis of transcriptomics data from many published studies highlights robust gene expression differences between BAT and WAT that are independent of experimental procedures, sample purity, origin of fat depots and sequencing methods, and can therefore be used to predict an unbiased molecular signature of BAT and WAT.

2.1.3 Prediction of BAT and WAT molecular signatures

As a first step towards the prediction of brown adipocytes content (browning capacity) in whole adipose tissue depots, we identified marker genes for classical brown and white fat tissue classification (Figure 2.2). To this goal, we integrated 51 BAT and 52 WAT transcriptional profiles from seven out of 16 independent studies in our atlas (M1, M2, M4, M5, M6, M7, M12/R4 in Figure 2.1A). Data normalization and batch effect removal were performed to ensure that differences in gene expression intensities were indeed due to differential expression between BAT and WAT sample groups. Ideally, brown and white fat-specific markers should show an "absolute" difference in expression to allow a clear distinction between BAT and WAT, independently of biological differences in fat depots, sample composition (pure populations *vs.* whole tissue biopsies), and their expression in other cell types. Overall, we found a total of 59 genes (Figure 2.2A) that were consistently and significantly differentially expressed between all BAT and WAT samples (log_2 fold change > 1.5 and P-adj value < 0.01). We identified several known brown fat markers, such as *Ucp1*, *Cidea* (cell death-inducing DFFA-like effector a), *Cox7a1* (cytochrome c oxidase subunit VII a polypeptide 1) and *Zic1* (zinc finger protein of the cerebellum 1), as well as white fat markers (e.g., *Hoxc8*, transcription factor homeobox C8). Due to the high abundance of mitochondria in BAT, brown fat markers included several mitochondrial-targeted proteins that are related to mitochondrial biogenesis and metabolism [88]. Not surprisingly, our marker core set was enriched in biological processes and pathways that are known to be involved in energy production, glucose and lipid metabolism (Figure 2.2B).

To further evaluate the predicted marker set, we looked for functional associations between the 59 marker genes (Figure 2.2C). We employed a computational method, called iRegulon, to reverse-engineer the transcriptional regulatory network underlying our set of differentially expressed marker genes. iRegulon searches for *cis*-regulatory regions at 10-20kb around the transcription start site (TSS) of each gene and then it looks for enrichment in any of 10,000 transcription factor (TF) motifs from seven different databases or ChIP-seq peaks that are associated with potential TFs. We identified four key TFs targeting 39 out of the 59 markers, which were also differentially expressed between WAT and BAT samples (log_2 fold-change > 1.5 and P-value < 0.01). Those included two well known

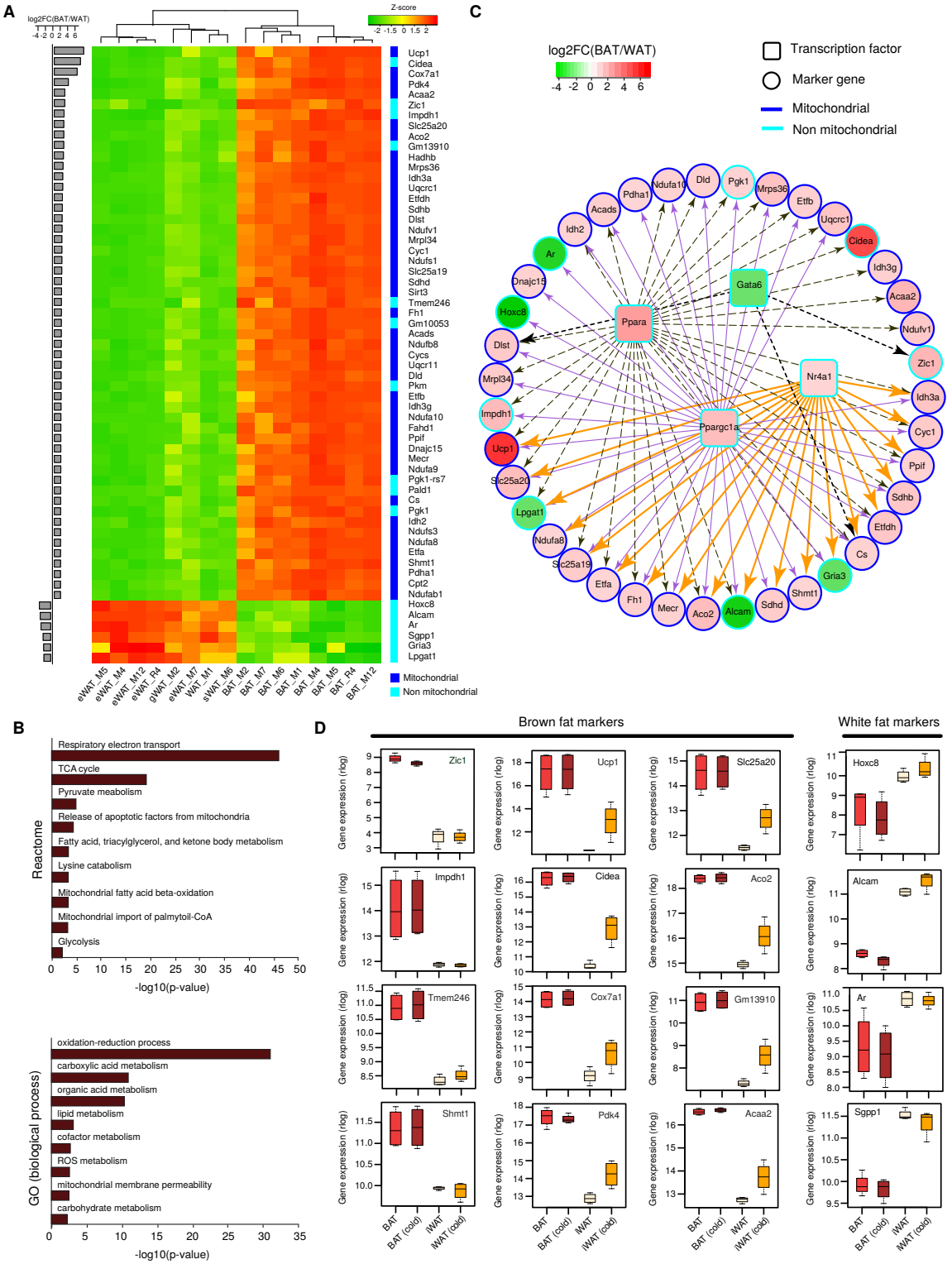


Fig. 2.2 **Prediction and Validation of BAT and WAT Marker Genes.** (A) Relative gene expression changes (Z score) for the predicted marker genes (53 and 6 BAT and WAT markers, respectively). MitoCarta2 is used to predict mitochondrial localization of marker genes. (B) Gene Ontology (GO) and pathway (Reactome) enrichment analysis of marker genes. (C) Transcriptional regulatory network of BAT and WAT marker genes (circles) and predicted targeting transcription factors (squares). Nodes are colored based on the log 2 fold change of the average expression level in BAT and WAT samples used for markers prediction (log_2 FC > 1.5 and p-adj value < 0.01). (D) Experimental validation of BAT and WAT marker genes (n *ge* 4). On each box, central line and edges represent median and 25th and 75th percentiles, respectively, and the whiskers extend to the most extreme data points. See also Figures 2.9 and 2.10.

key adipogenic TFs and co-regulators described in mammals that are part of the subfamily of peroxisome proliferator-activated receptors (*Ppar*α, peroxisome proliferator-activated receptor alpha; *Ppargc1*, peroxisome proliferator-activated receptor gamma coactivator 1-alpha) [121]. Another gene, *Nr4a1* (nuclear receptor subfamily 4, group A, member 1, also known as NUR77), was previously involved in the control of *Ucp1* expression [122]. In addition, we identified *Gata6*, a member of the GATA factors family. Although, those factors are generally considered as negative regulators of adipogenesis, *Gata6* has not yet been implicated in the regulation of adipogenesis in mammals [123]. Next, to validate the predicted BAT and WAT molecular signatures, we quantified the expression of each marker gene in interscapular BAT and iWAT isolated from 16 weeks old female mice kept at either thermoneutrality or cold acclimated for two weeks at 18 °C followed by 4 weeks at 5 °C in order to induce browning (Figure 2.9). We confirmed that all of our brown fat markers were indeed highly expressed in classical BAT from both room temperature and cold exposed mice (Figures 2.2D and S6). The expression of many of those markers, such as for example *Ucp1*, *Cidea*, *Cox7a1*, and *Pdk4*, was also higher in WAT from cold exposed mice, when compared to their expression in untreated WAT, reflecting the induction of browning, whereas others appeared to be brown-specific (e.g., *Zic1*, *Impdh1*, *Tmem246*, *Shmt1*). Similar results were obtained with male mice of the same age and background (data not shown). Notably, several genes have not yet been associated to BAT (*Aco2*, *Gm13910*, *Acaa2*) and WAT (*Alcam*, *Ar*, *Ar*, *Gria3*) and could therefore represent novel BAT and WAT markers.

2.1.4 Automated prediction of mouse adipose tissue browning capacity

To assess the thermogenic potential of fat tissues in response to browning agents, we devised a computational model that can predict brown and white adipocytes content ("BAT probability", probability to be brown-like) independently of sample purity and experimental systems (Figure 2.3A). The model combines into a single-layer neural network (SLNN) the transcriptional profiles of 51 BAT and 52 WAT samples from M1, M2, M4, M5, M6, M7, M12, and R4, which represent our "training set", and the predicted core marker set. Our choice of SLNN was justified by a systematic comparison to the performance of other algorithms, such as random forest, naïve bayes, generalized linear model, recursive partitioning, and support vector machine (Figure 2.11). To this goal, each machine learning algorithm was first trained through a leave-one-out cross-validation (LOOCV) step and the accuracy of different models was then assessed based on the correct classification of BAT and WAT samples from a "testing set" of nine independent studies (M3, M8, M9, M10, M11, M13, R1, R2, R3). As shown in Figure 2.11, SLNN outperformed other algorithms and was therefore implemented for follow-up analyses.

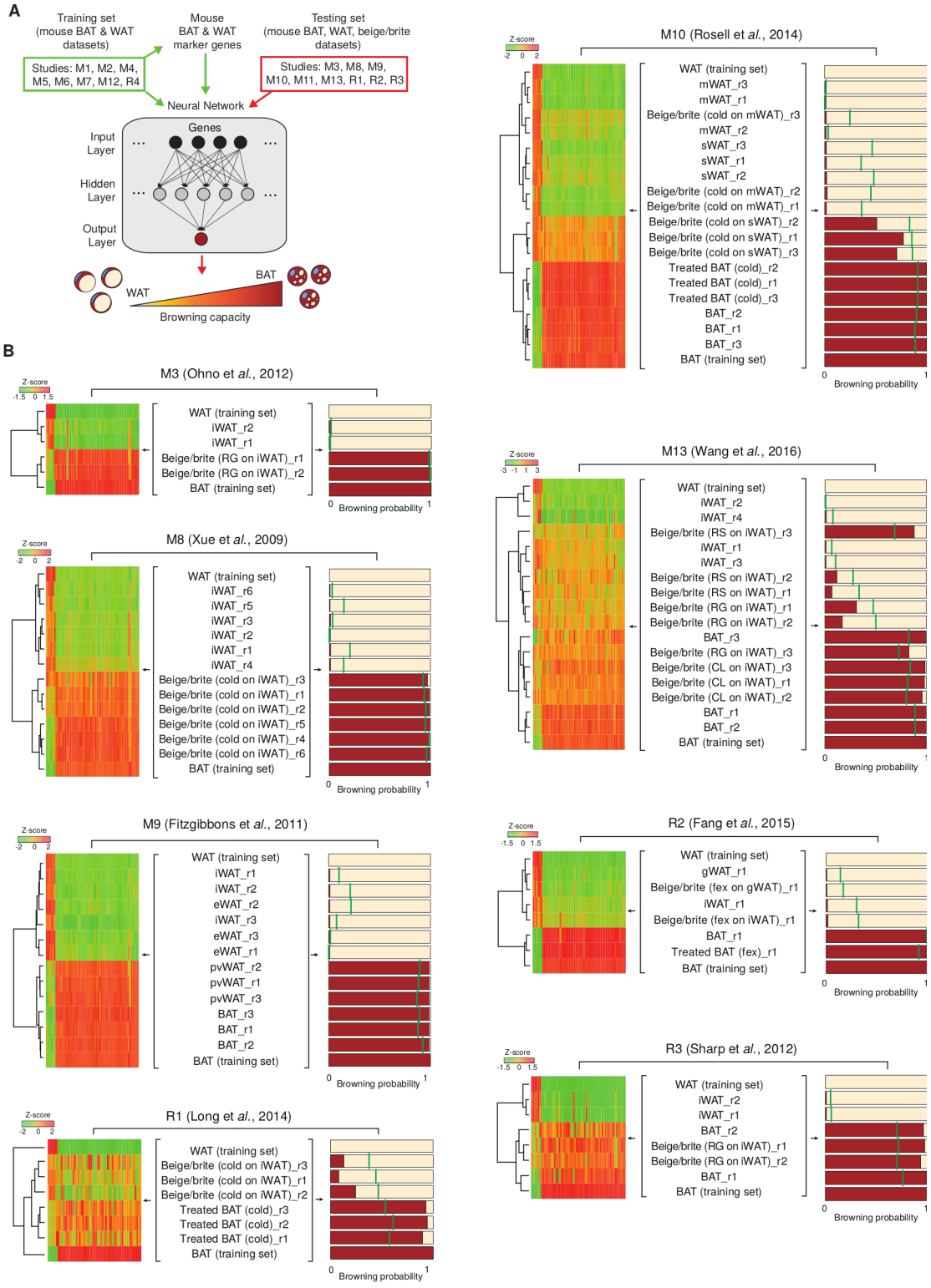


Fig. 2.3 **Prediction of Browning Capacity of Mouse Adipose Tissue Samples from Test Studies by Supervised Machine Learning** (A) Schematic diagram of the supervised machine learning approach. (B) Estimation of browning capacity in samples from each test study (right). HC analysis based on relative gene expression changes (Z score) of marker genes is shown for all samples and biological replicates within each test study (left). The green line on each bar represents the sample's relative *Ucp1* gene expression level calculated as (sample_*Ucp1* - min_*Ucp1*)/(max_*Ucp1* - min_*Ucp1*), where the min_*Ucp1* and max_*Ucp1* indicate the minimum and the maximum value of *Ucp1* gene expression across test and training sets, respectively. BAT (training set), combined BAT samples from all training datasets; r, replicates; WAT (training set), combined WAT samples from all training datasets. See Table 2.1 for detailed description of each sample. See also Figures 2.11 and 2.12.

Next, we tested the predictive power of our model using transcriptomes of white adipocytes from primary cell culture, whole fat tissue biopsies, as well as immortalized clonal lines, in which thermogenesis was activated by either cold, rosiglitazone (RG), roscovitine (RS), CL316,243 (CL), forskolin, or fexaramine (fex) treatment (Figures 2.3B and S8). The model deconvolutes the percentage of brown adipocytes/thermogenic cells and calculates the probability that a specific sample has acquired a brown-like transcriptional signature. A browning probability close to 0% and 100% would indicate a fat sample with WAT-like and BAT-like profiles, respectively. Instead, a browning probability close to 50% would suggest either that the tissue profile is neither BAT nor WAT-like, as expected for example for de-differentiated adipocytes and other tissue types, or that it has features of both fat types (e.g. it consists of an equal mixture of brown and white adipocytes).

As shown in Figure 2.3B, our model always classifies BAT and WAT with almost 100% accuracy and predicts the thermogenic potential of beige or brite samples to be higher than the corresponding untreated WAT samples, a result that is in agreement with the relative UCP1 expression level measured in each sample. A "positive control" in our analysis is represented by study M9. Here, the model "misclassifies" samples from pvWAT as having a high browning probability, thus BAT-like. However, our prediction is fully consistent with findings from the original study of Fitzgibbons *et al.* showing a virtually identical molecular signature between pvWAT and BAT from mice fed either a normal or high-fat diet. Notably, cold-treated sWAT from study M10 showed both, a high UCP1 expression level and browning capacity, whereas the model predicted the same treatment to be ineffective when applied to mesenteric WAT (mWAT). This result is consistent with previous observations that rodents sWAT depots are more sensitive to acquisition of BAT characteristics and have a higher thermogenic potential than visceral depots such as mWAT [124, 125]. When we applied our model on datasets from study M13, we found that samples defined by Wang *et al.* to originate from BAT and iWAT had a browning capacity close to 100% and 0%, respectively. Reassuringly, treatment of iWAT with the browning agent CL was predicted to yield a strong increase in browning capacity, in accordance with functional analyses. Similarly, we found that the thermogenic potential of CL-based iWAT treatment was higher than either RG or RS. Accordingly, measurements of rectal temperature in mice that were exposed to cold after treatment with each browning agent showed that the starting body temperature of CL-treated mice was the highest and CL was the most potent enhancer of glucose tolerance among all three drugs. Moreover, HC analysis also confirmed that at the transcriptional level UCP1-positive adipocytes arising in WAT of mice treated with RG and RS were more similar to each other than to UCP1-positive cells from CL-treated mice, which showed a transcriptome very close to that of BAT. Accordingly, RS and RG-treated cells

expressed several fold lower levels of *Ucp1* than cells from BAT and CL-treated adipocytes. We also obtained consistent results between our predictions and functional characterizations of fex-treated and untreated iWAT and gWAT from study R2. Here, the model predicted that fex treatment would not result in an increased browning activity of WAT. This is in agreement with the low *Ucp1* level measured in those samples and with observations that fex-treated mice show reduction in weight gain and improved metabolic homeostasis upon diet induced obesity, which was largely attributed to enhanced thermogenic activity in BAT rather than browning of iWAT or gWAT. However, the significance of our prediction is difficult to assess for this study, given that only one replicate for each sample is available.

Overall, our predictions are in agreement with HC analyses but while those can only provide a qualitative classification of each sample, our model can also estimate its thermogenic potential in response to a variety of browning stimuli.

2.1.5 Automated prediction of human adipose tissue browning capacity

To evaluate the applicability of our mouse-based model to deconvolute browning capacity of heterogeneous adipocyte populations from human samples we retrieved publicly available transcriptomics analyses of human adipose tissues (Table 2.1, Figures 2.4A and S9). Those included a total of 97 datasets from 3 Microarray and 2 RNA-Seq based studies on a variety of different experimental models: immortalized clonal preadipocyte cell lines derived from stromal vascular fractions (SVFs) of subcutaneous and deep neck of four adult human subjects (study hM1; [126]); primary adipocytes isolated from paired biopsies of deep and subcutaneous neck adipose tissue from six patients undergoing neck surgery (study hM2; [127]); adipose tissue isolated from abdominal subcutaneous fat depots of seven type 2 diabetic (T2D) patients before and after 10 days of cold acclimation (study hM3; [128]); pluripotent stem cell (PSC)-derived white (WAs) and brown (BAs) adipocytes subjected to the Janus kinase 3 (JAK3) and spleen tyrosine kinase (SYK) inhibitors tofacitinib and R406, respectively (study hR1; [129]); immortalized clonal brown and white preadipocytes isolated from SVFs in supraclavicular BAT and sWAT of two adult humans before and after *in vitro* differentiation and in response to forskolin treatment (study hR2; [63]). All of these studies were used as "testing set" in the neural network model (Figure 2.4B), which was instead trained on BAT and WAT samples from mouse-specific studies as previously shown in Figure 2.3A. Each testing dataset was first mapped through orthology to mouse genes. Overall, we observed that the level of *UCP1* expression in the original datasets was not always correlating to the browning capacity predicted by our model. Overall, we found *UCP1*

to be a weak classifier of brown *vs.* white-like depots, particularly when analyzing human tissue biopsies. Our observation is in agreement with previous claims that the thermogenic potential of human adipose tissues does not directly correlate with the simple presence of UCP1-positive cells [130]. Whereas *UCP1* expression can be used as a marker of active brown adipocytes, in heterogeneous populations it would be insufficient to estimate brown adipocyte content. Therefore, we evaluated the predictive value of our model in human samples where *UCP1* level could not be used to quantify browning. As an example, Tew *et al.* (study hM2) looked for functional differences between paired adipose tissue biopsies from deep neck, where human BAT is commonly found, and subcutaneous neck, where WAT is enriched. Accordingly, our model predicted higher browning capacity in samples from deep compared to subcutaneous neck, despite minor changes in *UCP1* expression level measured by microarray analysis. Interestingly, based on our prediction, the deep neck samples of some patients showed stronger browning capacity than others, possibly reflecting biological variations in BAT content or technical differences in the depth of tissue biopsies between individuals. In another study by Hanssen *et al.* (study hM3), chronic cold exposure was employed in seven human patients with T2D as a possible strategy to improve glucose homeostasis. Cold acclimation was previously shown to increase supraclavicular BAT mass and activity and to lead to recruitment of UCP1-positive adipocytes in other adipose tissue depots. Accordingly, all subjects showed an increase in cold-induced glucose uptake rate in the supraclavicular BAT region, although quite different between the individuals. However, BAT activity and mass were unaffected in other fat depots, such as sWAT and visceral WAT and no sign of browning could be detected by microarray-based gene expression analysis of abdominal sWAT biopsies from the same patients before and after cold acclimation. Consistently, we also found that the browning capacity of sWAT from each patient was unaffected by cold acclimation as the difference in browning probability between sWAT samples before and after cold exposure was minor. These results are in agreement with findings from multiple studies showing that cold does not brown all human fat depots equally [131–134]. Findings from our model applied to hR1 datasets were also in agreement with observations in the original study by Moisan *et al.* Here, JAK3 and SYK inhibitors, tofacitinib and R406 respectively, were shown to induce browning of human PSC-WAs. When comparing the browning probability of PSC-WAs samples treated with either DMSO or R406 and tofacitinib, our model correctly predicted a drug-dependent increase in browning. We also predicted a much higher browning capacity for R406 (PSC-WAs SYKi) than for tofacinib (PSC-WAs JAK3)-treated adipocytes, which was consistent with evidence of higher *UCP1* and *FABP4* (fatty acid binding protein 4) expression, small lipid droplet area and mitochondrial content in response to R406. Our data also suggested that both SYK and

JAK3 inhibitors are more potent browning inducers than cell fate conversion methods, as shown by comparing the BAT probability of (PSC)-derived brown adipocytes (PSC-BAs) with PSC-WAs. Finally, when testing samples from study hR2, we found that pre-adipocytes from supraclavicular and subcutaneous fat depots showed very low browning capacity, which increased after differentiation to brown but not to white adipocytes, respectively. As expected, a cAMP stimulus induced by treatment with forskolin increased the browning probability of sWAT derived clonal lines, as also confirmed by the activation of thermogenic markers observed in Shinoda *et al.* Altogether, these results demonstrate that our mouse-based model can be also applied to quantify white and brown adipocytes content in *ex vivo* clonally derived human adipocytes and complex human biopsies, and to reliably predict the thermogenic potential of treatments applied to induce browning of white fat depots.

2.2 Discussion

Integrative data analyses have been extensively shown to outperform the predictive power of individual large-scale studies [37, 135, 38, 36]. Therefore, when combining multiple datasets from different and complementary approaches, we can learn more about the system than what would be gained by analyzing each dataset in isolation. Given the wealthy of transcriptional analyses in the field of adipose biology, obesity and its comorbidities, we found it timely to perform a meta-analysis of published data and to combine into a single framework the knowledge acquired from each and all studies so far. To this goal, we compile the largest adipose-centric gene expression atlas and develop ProFAT, a systematic and automated approach to derive a robust and unbiased molecular signature of mouse BAT and WAT that we use to train a computational model in quantifying the browning capacity of heterogeneous fat tissues in both mouse and humans. We find that BAT and WAT show clearly distinct molecular signatures, irrespective of the anatomical location of the fat depots, their cell types composition, experimental models and procedures employed. Instead, when we apply ProFAT to several transcriptomics data from beige samples we observe that the extent to which beige or brite fat differs from either WAT or BAT greatly depends on study-to-study differences. Indeed, the degree of browning may vary due to samples purity, length and type (cold, PPAR-gamma or beta-3 adrenergic receptor agonists) of browning stimuli, and to whether the fat sample derives from tissue biopsies, primary adipocytes or clonal cell populations. The latter can be affected by *in vitro* adaptations, culture microenvironments and cell-cell-interactions. Unsupervised clustering analyses of gene expression data from pure clonally derived beige adipocytes have suggested that those could be classified as a distinct fat type at the transcriptional level [64]. While our analysis cannot formally rule

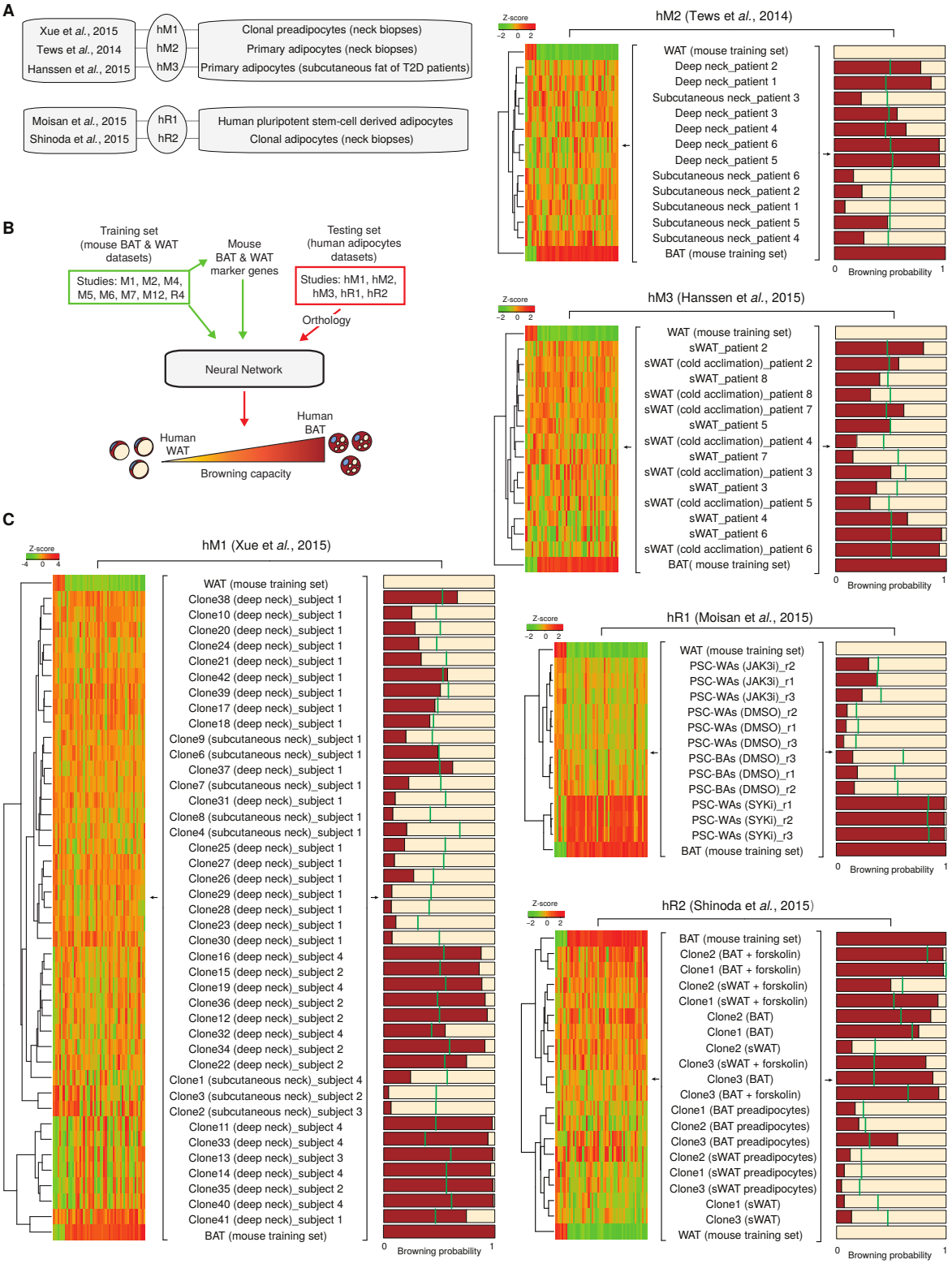


Fig. 2.4 **Prediction of Browning Capacity of Human Adipose Tissue Samples by Supervised Machine Learning** (A) Summary of microarray (hM1-3) and RNA-seq studies (hR1-2) on human fat samples. (B) Schematic diagram of the supervised machine learning approach. (C) Estimation of browning capacity (right) and HC analysis (left) of samples from each human-adipocytes-based study. JAK3i, Janus kinase 3 inhibitor (tofacitinib); PSC-BAs, pluripotent stem cell-derived brown adipocytes; PSC-WAs, pluripotent stem cell derived white adipocytes; SYKi, spleen tyrosine kinase inhibitor (R406). See Table 2.1 for detailed description of each sample. See also Figure 2.13.

out a distinct origin of beige from either brown or white adipocytes, it prompts for caution when defining beige-specific signatures in the context of a few limited dataset and biological models, rather than either systematically across a large and diverse set of data or based on pure populations of UCP1-positive cells [118].

The computational pipeline developed in this study will be especially important when trying to evaluate the thermogenic potential of therapeutic approaches in humans. Human adipose tissue biopsies usually yield limiting amounts of sample to perform an exhaustive functional characterization of browning and classical BAT markers, like *UCP1*, have been shown to be insufficient to predict adipose tissue types. Instead, whole-genome expression analyses typically require little material to be performed and have become a method of choice to infer functional remodeling of white adipose tissues, based on the assumption that the phenotype is reflected in the gene expression signature. Our meta-analysis enables to classify complex tissue samples from distinct fat depots as well as from *in vitro* derived adipocytes of both mouse and humans based on their relative brown and white-like molecular signatures. We envision a scenario in which medical researchers can directly assess the thermogenic potential of the patient's white fat sample, prior and post medical intervention.

Finally, we generate a user-friendly interface where microarray and RNA-Seq-based datasets from mouse and human samples can be directly uploaded and analyzed with both HC and PCA methods and their browning probability can be automatically computed using ProFAT. This resource can be freely accessed and should become increasingly powerful with the growing wealth of transcriptomics data.

2.3 Experimental procedures

2.3.1 Systematic retrieval of adipose tissue-specific transcriptional profiles

NCBI GEO and EBI ArrayExpress databases published before 1st of September 2015 were queried using the following keywords: "adipocyte", "adipose white", "adipose brown", "adipose beige", "fat white", "fat brown", "fat beige", "BAT", and "WAT". Systematic retrieval of whole genome expression profiles for *Mus musculus* and *Homo sapiens* from NCBI GEO and EBI ArrayExpress databases was performed through the Entrez Programming Utilities (E-utilities) and programmatic access, respectively. The GEOquery package from Bioconductor [13] was used to retrieve raw CEL Microarray data. Only Microarray and RNA-Seq datasets generated with Affymetrix and Illumina HiSeq Series sequencing platforms, respectively, were considered for downstream computational analyses.

2.3.2 Data processing

Raw CEL Microarray data were normalized by quantile normalization using the robust multiarray average (RMA) function in affy [136] /oligo [137] R packages. Probe IDs were mapped to Ensembl gene IDs using Biomart [138] based on the following criteria: probes not mapping to any gene ID were excluded; probes mapping to multiple gene IDs were assigned to all genes; for probes mapping to the same gene ID, the mean expression value was considered.

Processing of raw FastQ files from RNA-Seq analyses involved three main steps. First, adapters, barcodes and sequences with a Phred quality scores below 20 were removed using the Trim Galore software. Second, raw reads were mapped against mouse or human reference genomes (Ensembl release 81) using TopHat v2.0.13 [139] with Bowtie index (Bowtie 2.2.0.0) and GTF transcript annotation files. Next, the number of reads mapping to each Ensembl gene ID were counted using the ht-seq count software [140] to obtain raw read counts and quantify gene expression. A gene was defined as expressed if the sum of raw read counts across all datasets within a study was > 1. Last, DESeq2 (regularized logarithm transformation algorithm) [141] was used to perform rlog transformation (conversion of raw read counts in log_2 scale), which minimizes differences between samples and normalize with respect to library size.

For each study, a data matrix was generated, whereby each row and column corresponded to an Ensembl gene ID and sample ID, respectively. Correlation analyses were performed using pheatmap R package based on a pairwise distance matrix generated using Euclidean distance. Biological replicates that did not replicate were considered as outliers and removed from follow-up analyses. Data from all Microarray or RNA-Seq-based studies were aggregated based on Gene IDs and then Combat algorithm [142] was applied to remove the batch effect across multiple batches of microarray and RNA-Seq experiments and to calculate normalized gene expression values. This algorithm is robust to outliers in small sample sizes and performs comparable to existing methods for large samples. Hierarchical clustering was performed using Euclidean distance and complete linkage based on normalized gene expression values. Differential gene expression analysis was performed using the Limma algorithm and significantly differentially expressed genes were defined based on an adjusted P-values < 0.01 and a mean log_2 fold-change threshold > 1.5.

2.3.3 Identification of BAT and WAT marker genes

Microarray and RNA-Seq gene expression data on BAT and WAT samples from the following studies M1, M2, M4, M5, M6, M7, M12, R4 were combined based on Gene IDs. Combat

algorithm [142] was applied to remove batch effects and to calculate normalized gene expression values. Next, the MGFM (Marker Gene Finder in MicroArray) bioinformatics tool [143] was applied to predict genes that allow a robust and specific segregation of samples from BAT and WAT types (http://www.bioconductor.org/packages/release/bioc/html/MGFM.html; default parameters). The subset of 59 genes that were significantly differentially expressed (log_2 fold-change (BAT/WAT) > 1.5 and P-adj value < 0.01) was selected as a core BAT and WAT marker set. The ConcensusPathDB-Mouse (http://cpdb.molgen.mpg.de/MCPDB) was used to identify non-redundant functional categories from Gene Ontology (GO) and Reactome that were enriched within BAT and WAT marker genes (P-value < 0.01). Cytoscape [144] was used to display the predicted regulatory network.

2.3.4 In-house RNA-Seq

Total RNA was extracted from inguinal WAT and interscapular BAT of 16 weeks old female C57BL/6 mice kept either for the whole life at an ambient temperature of 30 °C or for two weeks at 18 °C followed by 4 weeks at 5 °C (n=4; not randomization and blinding applied). Qiazol was used for RNA extraction according to the manufacturer's instructions (Qiazol Lysis Reagent, Qiagen). The quality of the RNA was determined with the Agilent 2100 BioAnalyzer (RNA 6000 Nano Kit, Agilent). All samples had a RNA integrity number (RIN) value greater than 8. For library preparation, 1 µg of total RNA per sample was used. RNA molecules were poly(A) selected, fragmented, and reverse transcribed with the Elute, Prime, Fragment Mix (EPF, Illumina). End repair, A-tailing, adaptor ligation, and library enrichment were performed as described in the Low Throughput protocol of the TruSeq RNA Sample Prep Guide (Illumina). RNA libraries were assessed for quality and quantity with the Agilent 2100 BioAnalyzer and the Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies). RNA libraries were sequenced as 100 bp paired-end runs on an Illumina HiSeq2500 platform. The animal welfare authorities approved animal maintenance and experimental procedures.

2.3.5 Prediction of adipose tissue browning capacity by machine learning

A neural network model was developed with one hidden layer, using caret R package with method set to nnet. Leave-one-out cross validation was used to tune the number of hidden units and weight decay, whereas default values were used for the remaining parameters. Datasets were exclusively assigned to either a test or a training group. The training data only included datasets from study M1, M2, M4, M5, M6, M7, M12 and R4. Instead, the test data included microarray and RNA-Seq datasets from study M3, M8, M9, M10, M11, M13,

R1, R2, and R3. To avoid introducing circularity in the analysis, test data were independent from training data and were never used for training. COMBAT algorithm was applied to normalize any new test dataset against the training set in order to remove batch effects and then the training set together with the core marker set were used in the neural network. Human transcriptional profiles were also used as testing set by mapping human gene IDs to mouse ortholog gene IDs with BioMart (Ensembl release 81) restricted to ortholog_one2one mapping type.

2.3.6 Statistical analysis

Z-score is calculated as (X-μ)/σ, where X is the value of the element, μ is the mean and σ is the standard deviation. A marker gene is defined significant if the adjusted p-value <0.01 and the log_2(fold change) >1.5, where p-value and fold change are calculated with DESeq R package.

2.4 Data and code availability

The data and code can be accessed at: https://github.com/PerocchiLab/ProFAT. Accession codes for publicly available Microarray and RNA-Seq datasets used in this study are also listed in Table 2.1.

2.5 Accession numbers

The accession number for in-house RNA-Seq data is GSE112582.

2.6 Acknowledgements

We acknowledge support from the German Research Foundation (DFG) under the Emmy Noether Programme (PE 2053/1-1) and the Bavarian Ministry of Sciences, Research and the Arts in the framework of the Bavarian Molecular Biosystems Research Network (D2–F5121.2–10c/4822) to F.P., Y.C., and L.J.

2.7 Author contributions

Conceptualization, F.P. and M.J.; Methodology, Y.C., S.K., and L.J.; Software, Y.C., A.H., and L.J.; Formal Analysis, Y.C., A.H., L.J., and T.S.; Investigation, S.K. and E.G.; Resources,

F.P., M.J., and M.T.; Data Curation, S.K. and S.Z.; Writing – Original Draft, F.P., M.J., S.K., and Y.C.; Visualization, F.P., L.J., Y.C., and A.H.; Supervision, F.P. and M.J.; Funding Acquisition, F.P.

2.8 Declaration of interests

The authors declare no competing interests.

2.9 Supplemental information

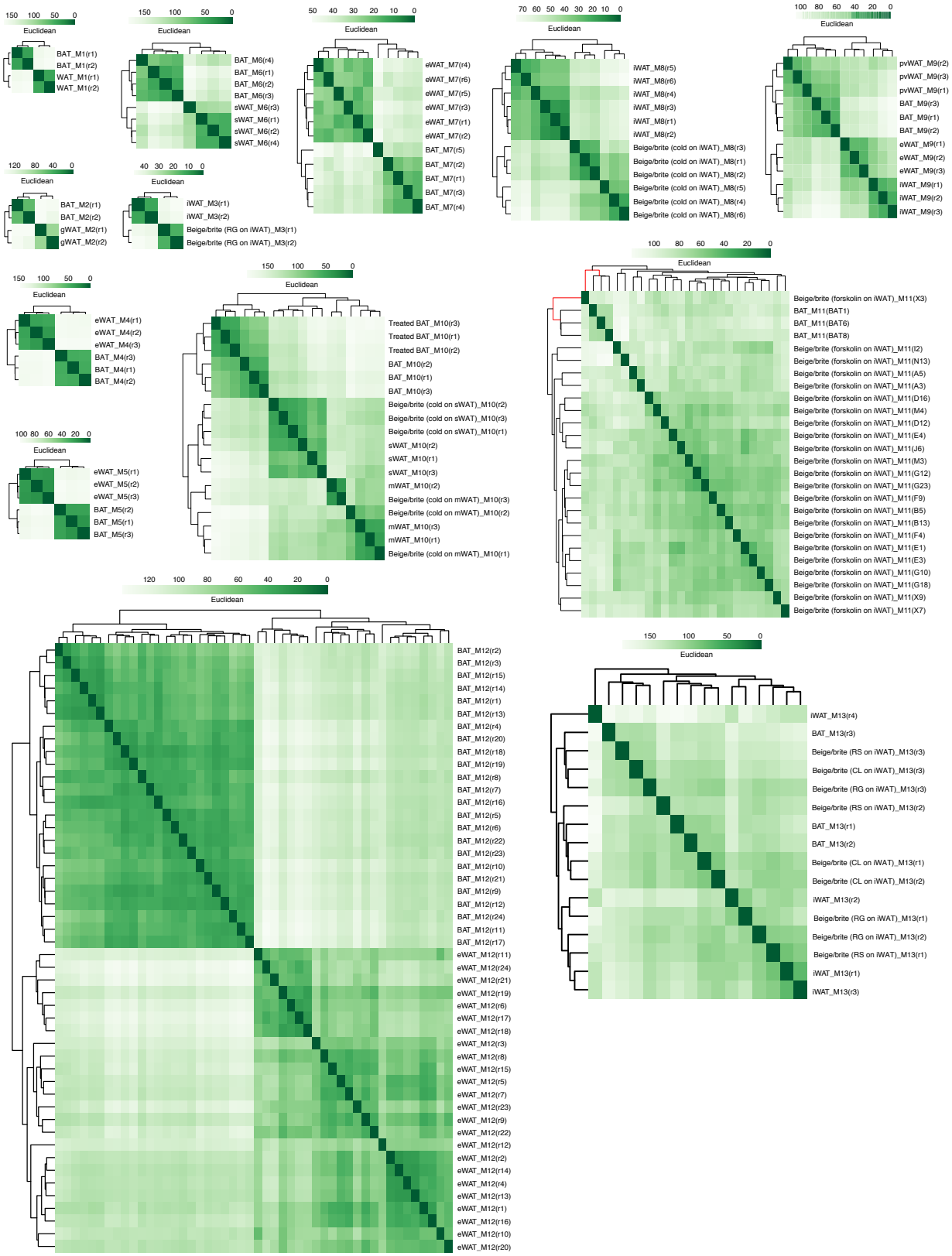


Fig. 2.5 **Hierarchical clustering (HC) of mouse microarray studies.** HC is performed using Euclidian distance and complete linkage. Red lines in the dendrogram indicate biological replicates that are considered as outliers. M, Microarray; r, replicate. The type of treatment applied to induce browning of WAT is indicated in parentheses (RG, rosiglitazone; CL, CL316,243; RS, roscovitine). Related to Figures 1.14 and 2.1.

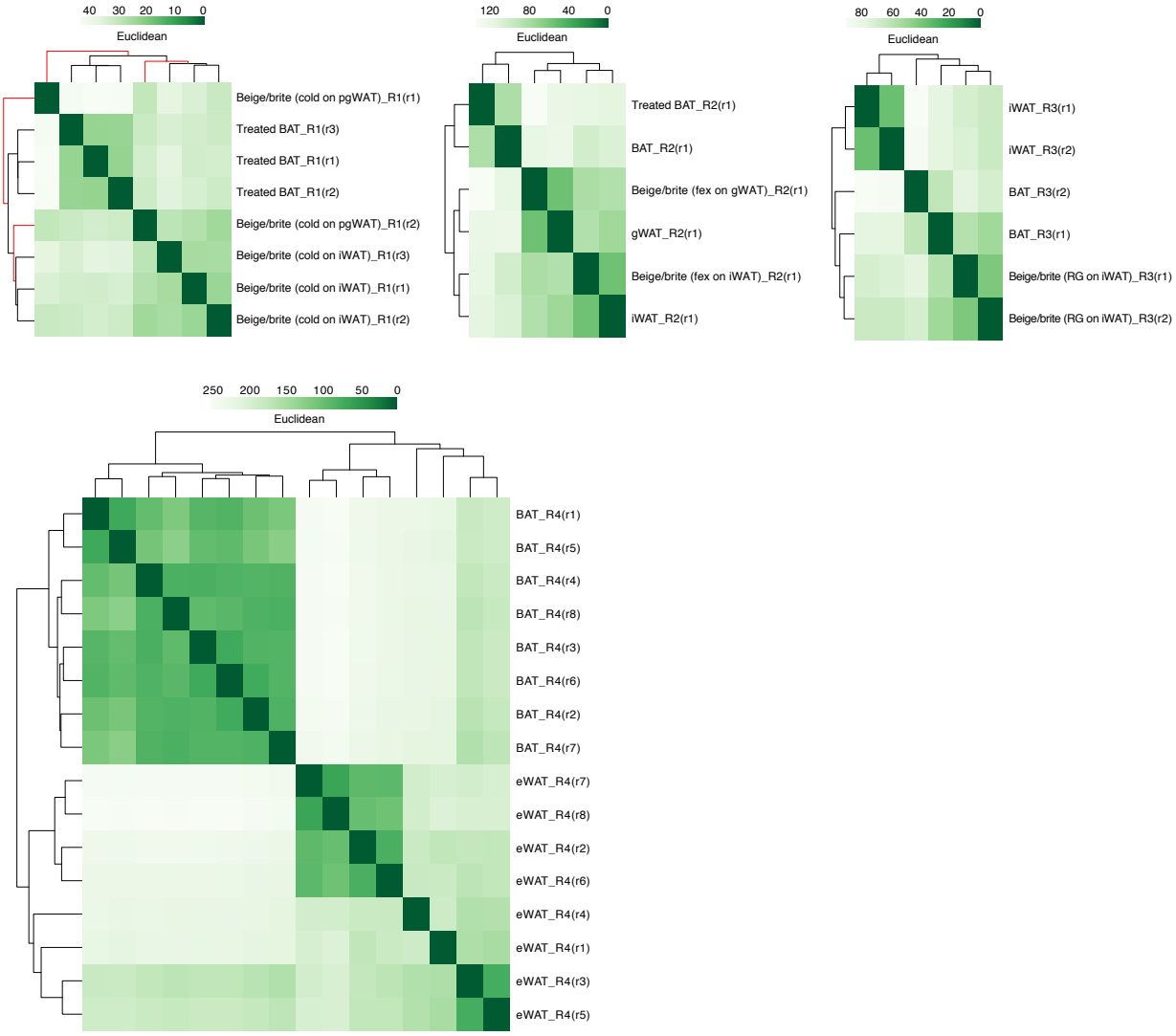


Fig. 2.6 **Hierarchical clustering (HC) of mouse RNA-seq studies.** HC is performed using Euclidian distance and complete linkage. Red lines in the dendrogram indicate biological replicates that are considered as outliers. R, RNA-seq; r, replicate. The type of treatment applied to induce browning of WAT is indicated in parentheses (RG, rosiglitazone; fex, fexaramine). Related to Figures 1.14 and 2.1.

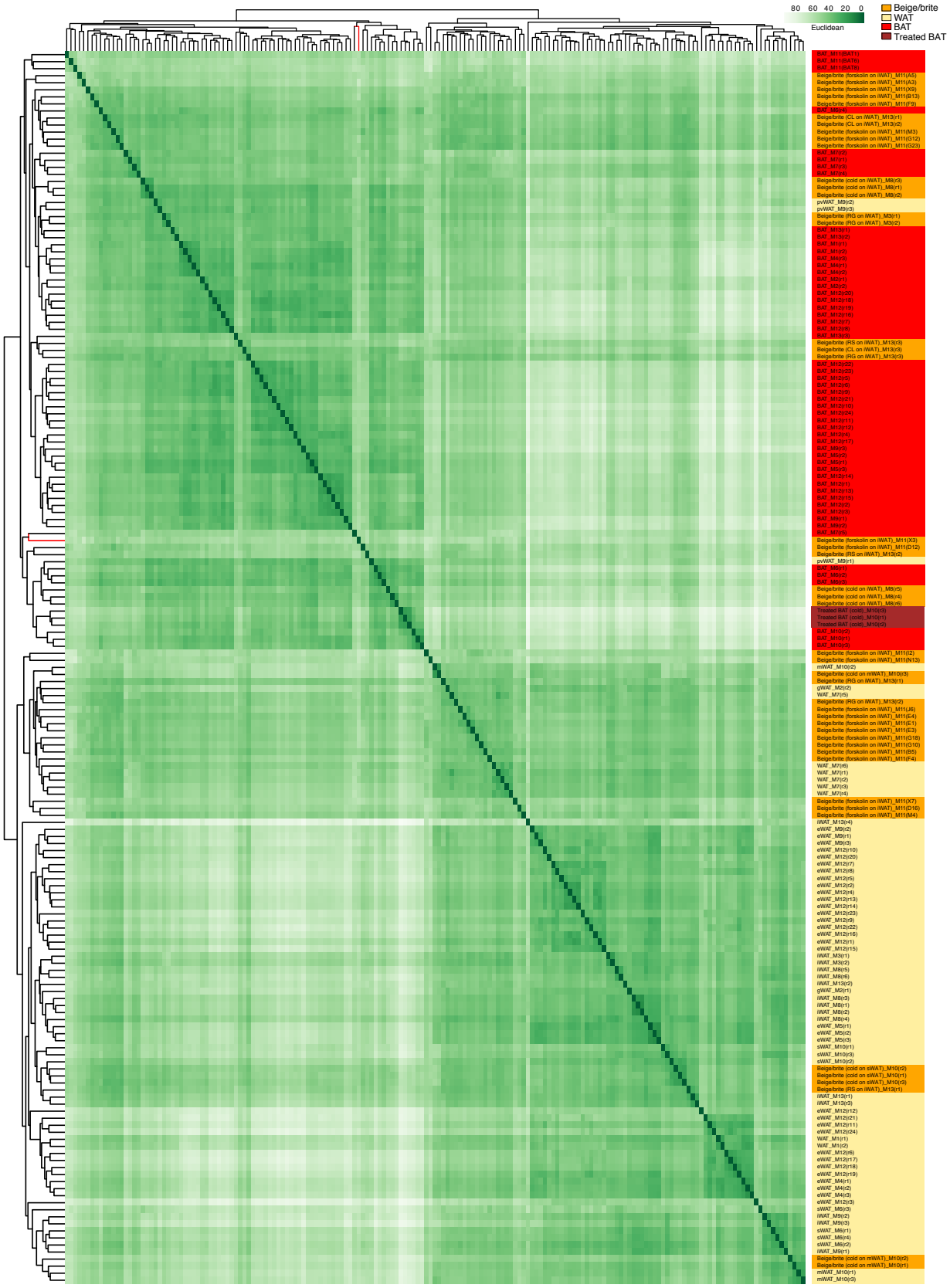


Fig. 2.7 **Hierarchical clustering of mouse samples across all microarray-based studies.** Euclidean distance is used for sample correlation analysis and hierarchical clustering is performed using Euclidian values and complete linkage. See Table 2.1 for detailed description of each sample. Related to Figures 1.14 and 2.1.

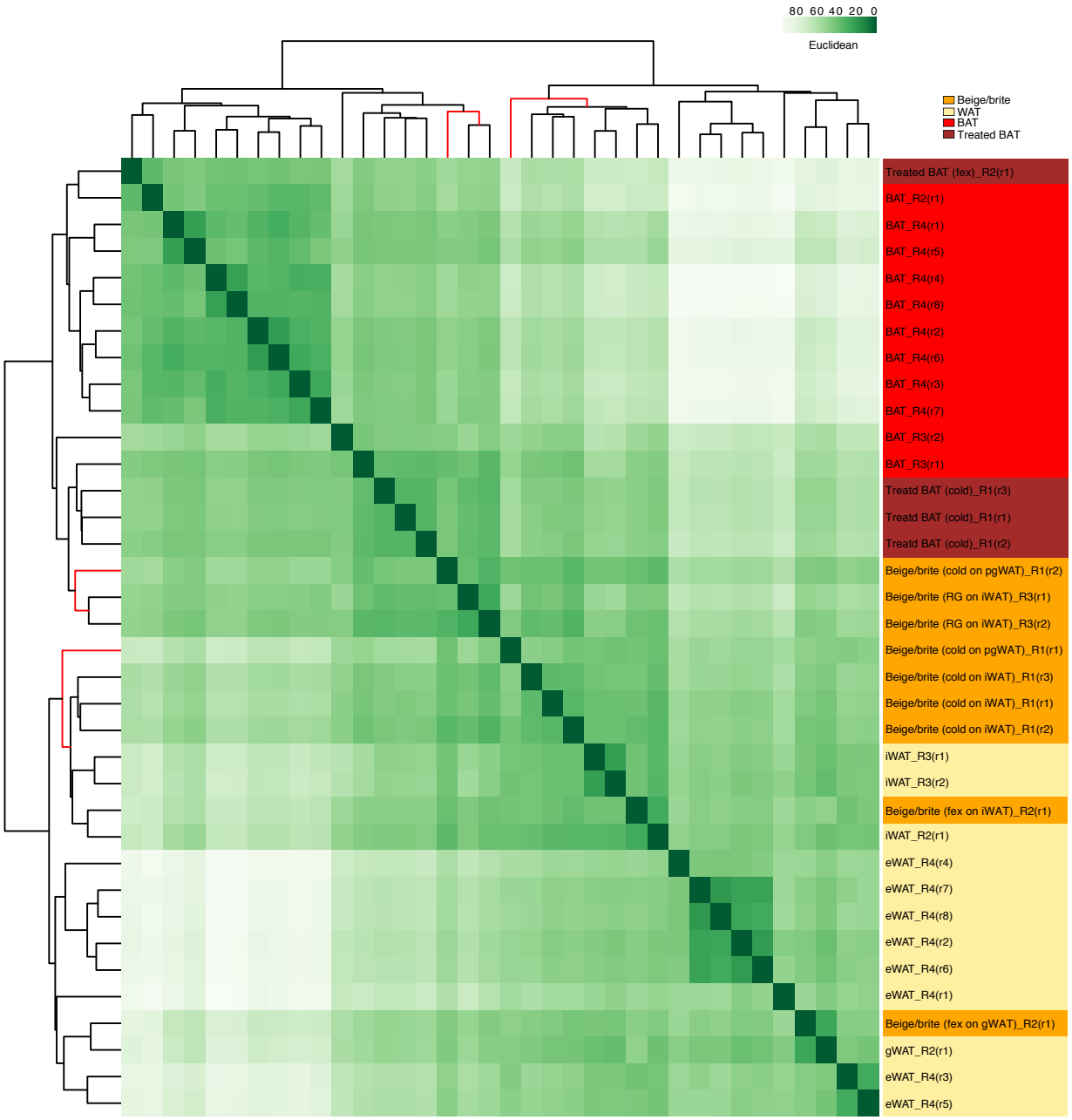


Fig. 2.8 **Hierarchical clustering of mouse samples across all RNA-seq-based studies.** Euclidean distance is used for sample correlation analysis and hierarchical clustering is performed using Euclidian values and complete linkage. See Table 2.1 for detailed description of each sample. Related to Figures 1.14 and 2.1.

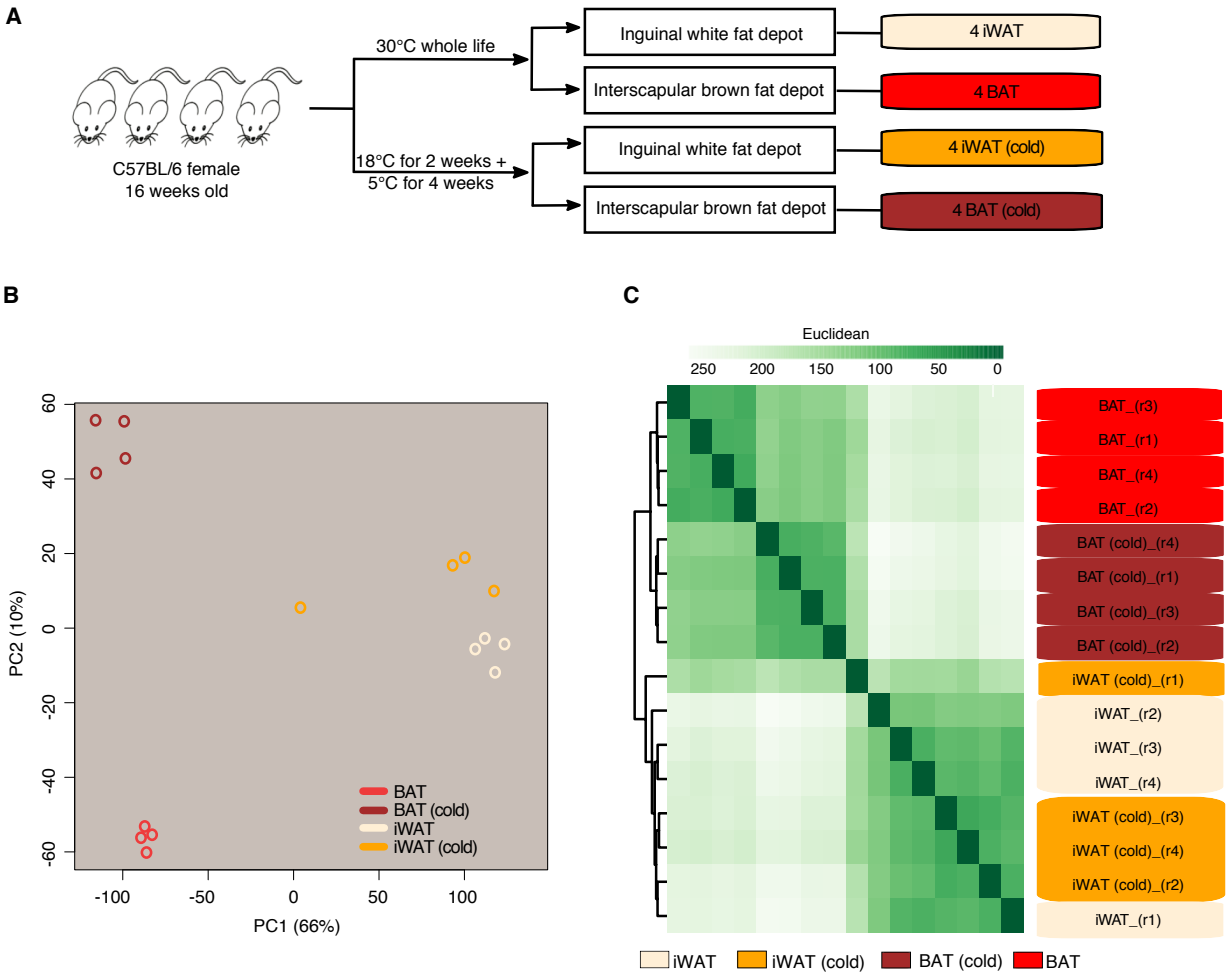


Fig. 2.9 **In-house transcriptome analysis of BAT and WAT from wild-type and cold-exposed mice.** (A) Experimental design. BAT and BAT (cold), interscapular brown adipose tissue from mice kept at either 30 °C or in cold (n≥4), respectively; iWAT and iWAT (cold), inguinal white adipose tissue from mice kept at either 30 °C or in cold (n≥4), respectively. (B) Principle component analysis of normalized gene expression data for all biological replicates. (C) Hierarchical clustering of all samples based on gene expression data. Euclidean distance is used for pairwise distance matrix and hierarchical clustering is performed with Euclidean distance and complete linkage. Related to Figure 2.2.

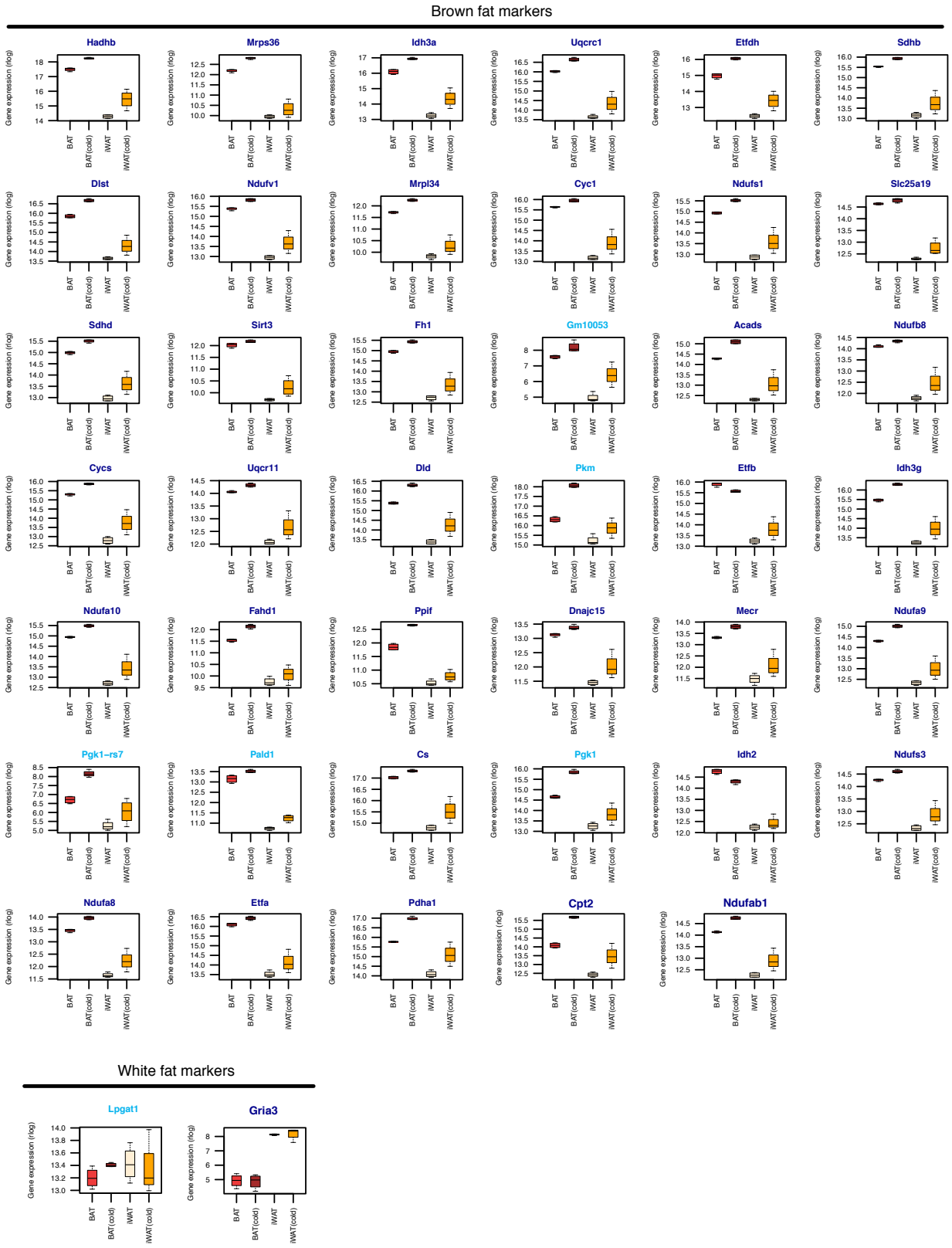


Fig. 2.10 **Experimental validation of BAT and WAT marker genes.** Gene names in blue and magenta indicate mitochondria and non-mito- chondria localized markers, respectively. BAT and BAT (cold), interscapular brown adipose tissue from mice kept at either 30 °C or in cold (n≥4), respectively; iWAT and iWAT (cold), inguinal white adipose tissue from mice kept at either 30 °C or in cold (n≥4), respectively. Related to Figure 2.2.

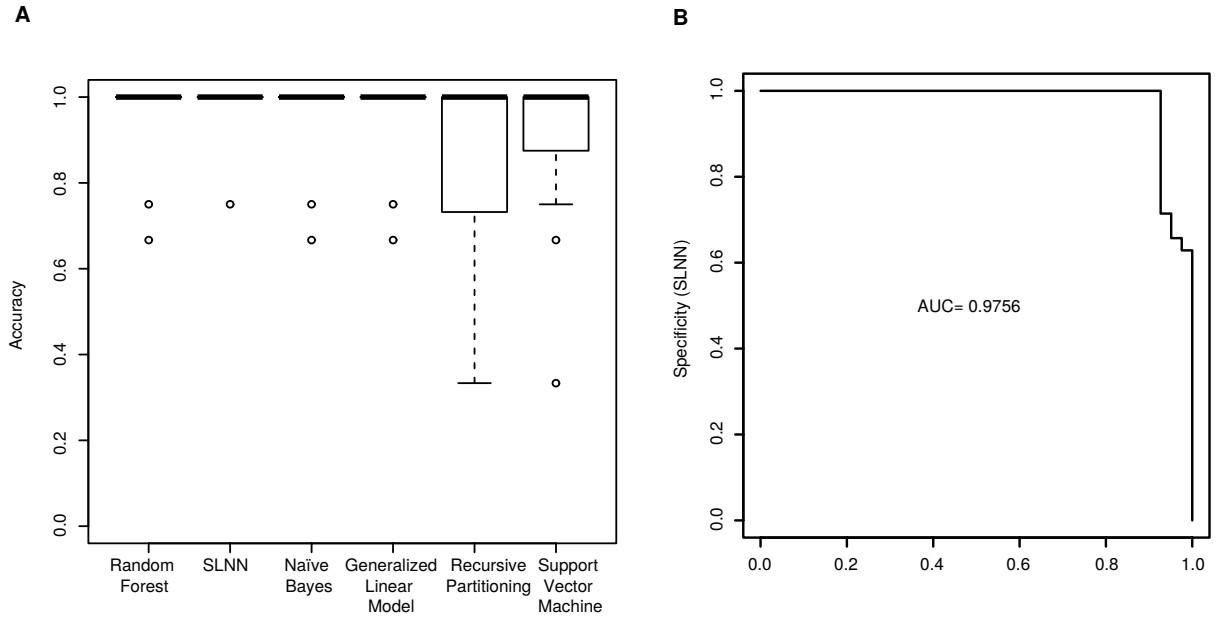


Fig. 2.11 **Systematic comparison of the performance of machine learning algorithms for the prediction of browning capacity.** (A) Accuracy of six machine learning algorithms (random forest, SLNN, naive bayes, generalized linear model, recursive partitioning, support vector machine), trained using BAT and WAT-specific microarray and RNA-seq datasets from studies M1, M2, M4, M5, M6, M7, M12 and R4, in classifying BAT and WAT samples from an indipendent set of studies (M3, M8, M9, M10, M11, M13, R1, R2, R3). (B) Classification accuracy of SLNN. The accuracy is calculated as: (TP+TN)/total samples. A sample is a true positive if BAT was predicted with a probability > 0.5. Related to Figure 2.3.

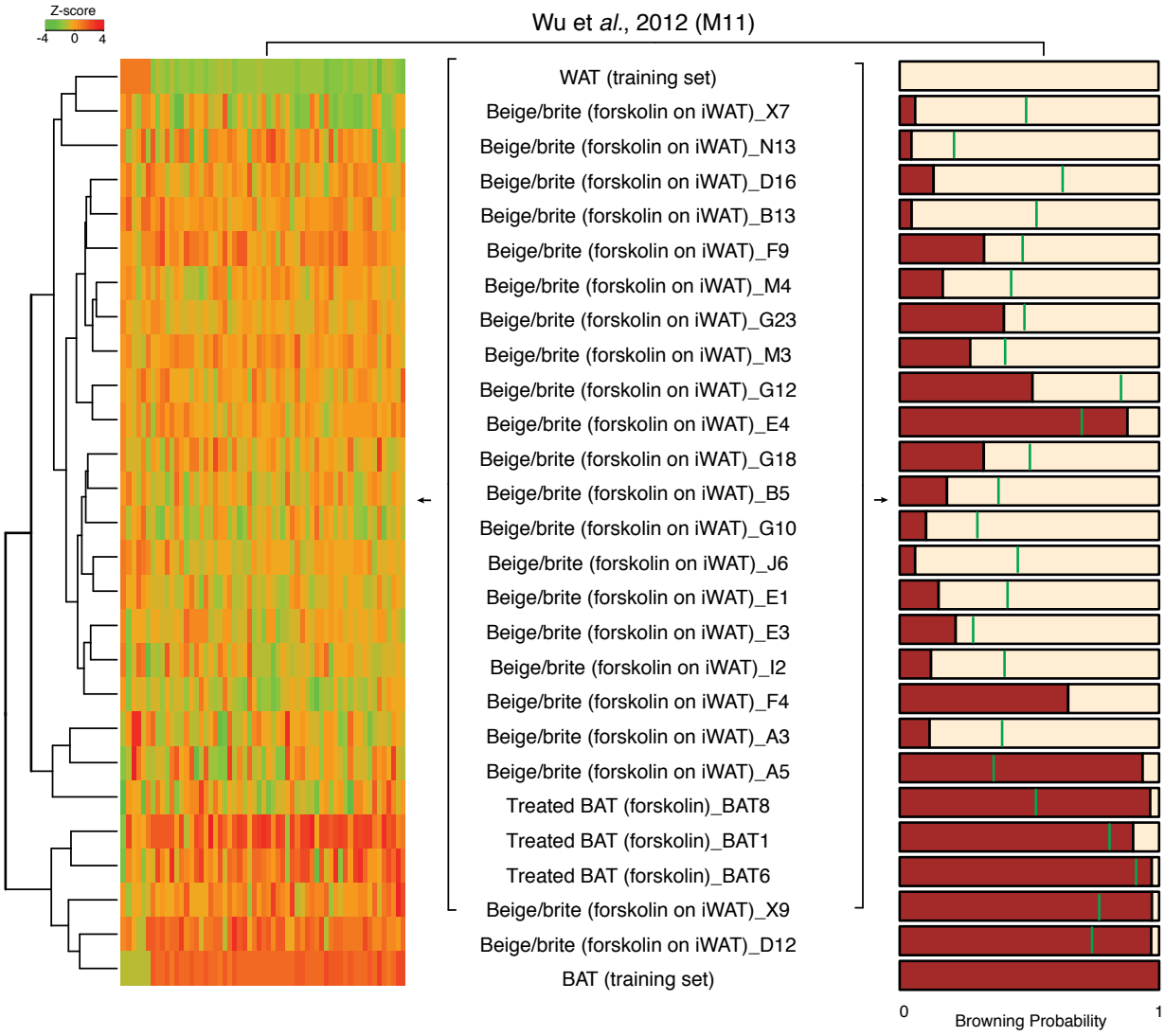


Fig. 2.12 **Prediction of browning capacity for study M11.** Related to Figure 2.3.

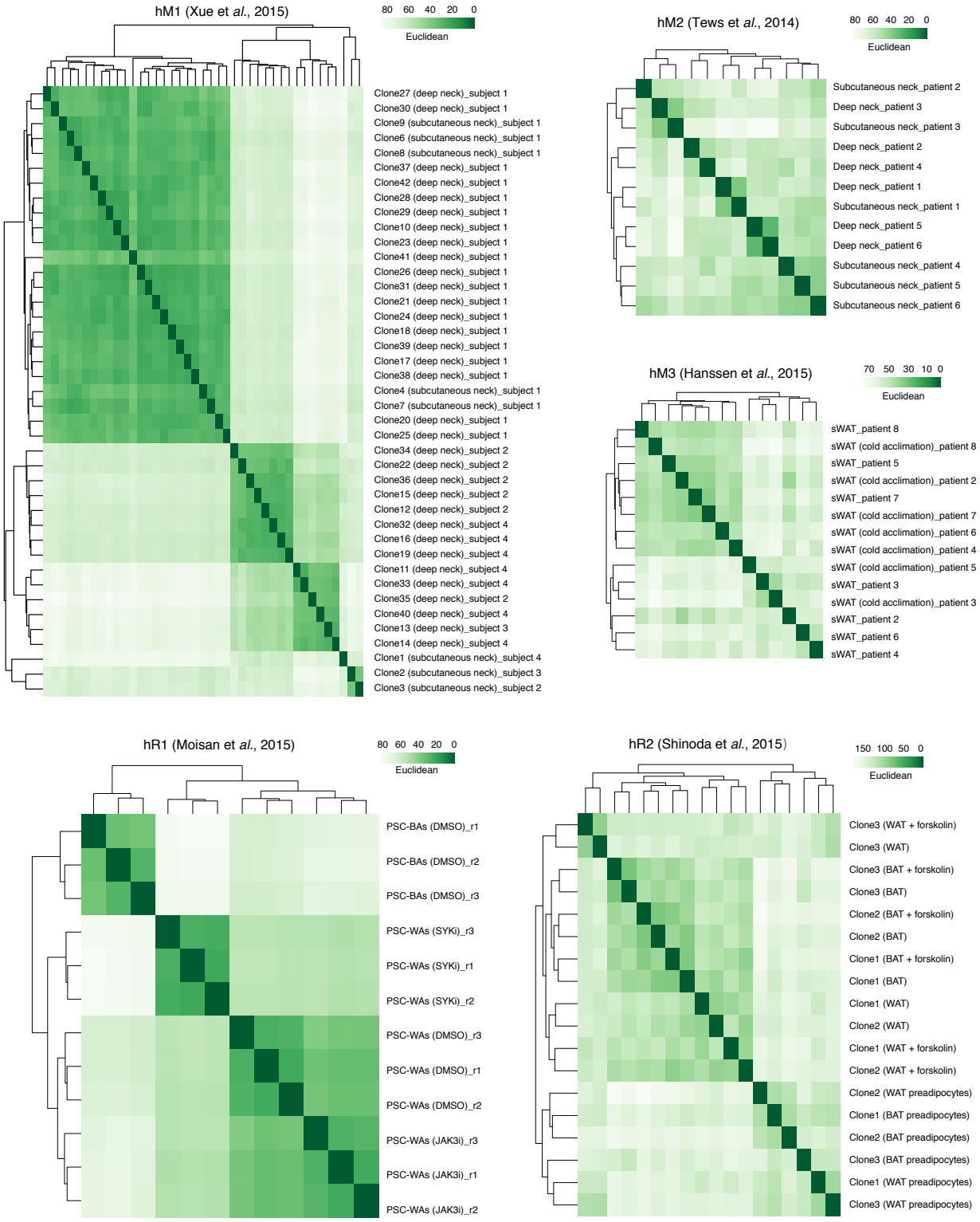


Fig. 2.13 **Hierarchical clustering (HC) of samples within each human microarray and RNA-seq study.** HC is performed using Euclidian distance and complete linkage. hM, microarray on human samples; hR, RNA-seq on human samples; r, replicate. Related to Figure 2.4

Table 2.1 **Mouse and human adipose tissue-centered gene expression atlas.** WAT, white adipose tissue; iWAT, inguinal white adipose tissue; gWAT, gonadal white adipose tissue; eWAT, epididymal white adipose tissue; sWAT, subcutaneous white adipose tissue; mWAT, mesenteric white adipose tissue; pvWAT, perivascular white adipose tissue; BAT, brown adipose tissue; (RG), rosiglitazone; (RS), roscovitine; fex, fexaramine; CL, CL316,243; DIO, diet-induced obesity; N/A, not available. M, microarrays; R, RNA-seq; r, biological replicate; hM, human Microarray; hR, human RNA-Seq; SVF, stromal vascular fraction; PSC, pluripotent stem-cells; BAs, brown adipocytes; WAs, white adipocytes; MPCs, mesenchymal progenitor cells. Related to Figures 1.14-2.4.

Study ID	Dataset ID	Pubmed ID	Database ID	Sample ID	Sample type	Age	Sample origin	Tissue type	Treatment	Annotation
M1	BAT_M1(r1)	15075390	GSE10246	GSM258611	Tissue	8-10 weeks	C57BL6, male	N/A	None	BAT
M1	BAT_M1(r2)	15075390	GSE10246	GSM258612	Tissue	8-10 weeks	C57BL6, male	N/A	None	BAT
M1	WAT_M1(r1)	15075390	GSE10246	GSM258613	Tissue	8-10 weeks	C57BL6, male	N/A	None	WAT
M1	WAT_M1(r2)	15075390	GSE10246	GSM258614	Tissue	8-10 weeks	C57BL6, male	N/A	None	WAT
M2	BAT_M2(r1)	20679227	GSE19757	GSM493479	Tissue	N/A	N/A	Interscapular brown adipose tissue	None	BAT
M2	BAT_M2(r2)	20679227	GSE19757	GSM493480	Tissue	N/A	N/A	Interscapular brown adipose tissue	None	BAT
M2	gWAT_M2(r1)	20679227	GSE19757	GSM493477	Tissue	N/A	N/A	Gonadal white adipose tissue	None	WAT
M2	gWAT_M2(r2)	20679227	GSE19757	GSM493478	Tissue	N/A	N/A	Gonadal white adipose tissue	None	WAT
M3	iWAT_M3(r1)	22405074	GSE35011	GSM860620	Primary adipocytes	N/A	C57BL6, male	Inguinal white adipose tissue	None	WAT
M3	iWAT_M3(r2)	22405074	GSE35011	GSM860621	Primary adipocytes	N/A	C57BL6, male	Inguinal white adipose tissue	None	WAT
M3	Beige/brite (RG on iWAT)_M3(r1)	22405074	GSE35011	GSM860622	Primary adipocytes	N/A	C57BL6, male	Inguinal white adipose tissue	Rosiglitazone (1μM)	Beige/brite
M3	Beige/brite (RG on iWAT)_M3(r2)	22405074	GSE35011	GSM860623	Primary adipocytes	N/A	C57BL6, male	Inguinal white adipose tissue	Rosiglitazone (1μM)	Beige/brite
M4	BAT_M4(r1)	21035761	GSE20165	GSM506030	Tissue	20 weeks	SV129, male	Interscapular brown adipose tissue	None	BAT
M4	BAT_M4(r2)	21035761	GSE20165	GSM506031	Tissue	20 weeks	SV129, male	Interscapular brown adipose tissue	None	BAT
M4	BAT_M4(r3)	21035761	GSE20165	GSM506032	Tissue	20 weeks	SV129, male	Interscapular brown adipose tissue	None	BAT
M4	eWAT_M4(r1)	21035761	GSE20165	GSM506024	Tissue	20 weeks	SV129, male	Epididymal white adipose tissue	None	WAT
M4	eWAT_M4(r2)	21035761	GSE20165	GSM506025	Tissue	20 weeks	SV129, male	Epididymal white adipose tissue	None	WAT
M4	eWAT_M4(r3)	21035761	GSE20165	GSM506026	Tissue	20 weeks	SV129, male	Epididymal white adipose tissue	None	WAT
M5	BAT_M5(r1)	17618855	GSE8044	GSM198456	Tissue	10-12 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M5	BAT_M5(r2)	17618855	GSE8044	GSM198457	Tissue	10-12 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M5	BAT_M5(r3)	17618855	GSE8044	GSM198458	Tissue	10-12 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M5	eWAT_M5(r1)	17618855	GSE8044	GSM198496	Tissue	10-12 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M5	eWAT_M5(r2)	17618855	GSE8044	GSM198523	Tissue	10-12 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M5	eWAT_M5(r3)	17618855	GSE8044	GSM198545	Tissue	10-12 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M6	BAT_M6(r1)	26010905	GSE67389	GSM1646139	Tissue	N/A	LACA, male	Interscapular brown adipose tissue	None	BAT
M6	BAT_M6(r2)	26010905	GSE67389	GSM1646141	Tissue	N/A	LACA, male	Interscapular brown adipose tissue	None	BAT
M6	BAT_M6(r3)	26010905	GSE67389	GSM1646143	Tissue	N/A	LACA, male	Interscapular brown adipose tissue	None	BAT
M6	BAT_M6(r4)	26010905	GSE67389	GSM1646145	Tissue	N/A	LACA, male	Interscapular brown adipose tissue	None	BAT
M6	sWAT_M6(r1)	26010905	GSE67389	GSM1646131	Tissue	N/A	LACA, male	Subcutaneous white adipose tissue	None	WAT
M6	sWAT_M6(r2)	26010905	GSE67389	GSM1646133	Tissue	N/A	LACA, male	Subcutaneous white adipose tissue	None	WAT
M6	sWAT_M6(r3)	26010905	GSE67389	GSM1646135	Tissue	N/A	LACA, male	Subcutaneous white adipose tissue	None	WAT
M6	sWAT_M6(r4)	26010905	GSE67389	GSM1646137	Tissue	N/A	LACA, male	Subcutaneous white adipose tissue	None	WAT
M7	BAT_M7(r1)	17360536	GSE7032	GSM162537	Primary adipocytes	3-4 weeks	NMRI	Interscapular brown adipose tissue	None	BAT
M7	BAT_M7(r2)	17360536	GSE7032	GSM162538	Primary adipocytes	3-4 weeks	NMRI	Interscapular brown adipose tissue	None	BAT
M7	BAT_M7(r3)	17360536	GSE7032	GSM162539	Primary adipocytes	3-4 weeks	NMRI	Interscapular brown adipose tissue	None	BAT
M7	BAT_M7(r4)	17360536	GSE7032	GSM162540	Primary adipocytes	3-4 weeks	NMRI	Interscapular brown adipose tissue	None	BAT
M7	BAT_M7(r5)	17360536	GSE7032	GSM162541	Primary adipocytes	3-4 weeks	NMRI	Interscapular brown adipose tissue	None	BAT
M7	eWAT_M7(r1)	17360536	GSE7032	GSM162550	Primary adipocytes	3-4 weeks	NMRI	Epididymal white adipose tissue	None	WAT
M7	eWAT_M7(r2)	17360536	GSE7032	GSM162551	Primary adipocytes	3-4 weeks	NMRI	Epididymal white adipose tissue	None	WAT
M7	eWAT_M7(r3)	17360536	GSE7032	GSM162552	Primary adipocytes	3-4 weeks	NMRI	Epididymal white adipose tissue	None	WAT
M7	eWAT_M7(r4)	17360536	GSE7032	GSM162553	Primary adipocytes	3-4 weeks	NMRI	Epididymal white adipose tissue	None	WAT
M7	eWAT_M7(r5)	17360536	GSE7032	GSM162554	Primary adipocytes	3-4 weeks	NMRI	Epididymal white adipose tissue	None	WAT
M7	eWAT_M7(r6)	17360536	GSE7032	GSM162555	Primary adipocytes	3-4 weeks	NMRI	Epididymal white adipose tissue	None	WAT
M8	iWAT_M8(r1)	19117550	GSE13432	GSM338989	Tissue	5 weeks	C57BL6, male	Inguinal white adipose tissue	30 °C (5 weeks)	WAT
M8	iWAT_M8(r2)	19117550	GSE13432	GSM338990	Tissue	5 weeks	C57BL6, male	Inguinal white adipose tissue	30 °C (5 weeks)	WAT
M8	iWAT_M8(r3)	19117550	GSE13432	GSM338991	Tissue	5 weeks	C57BL6, male	Inguinal white adipose tissue	30 °C (5 weeks)	WAT
M8	iWAT_M8(r4)	19117550	GSE13432	GSM338983	Tissue	5 weeks	C57BL6, male	Inguinal white adipose tissue	30 °C (1 week)	WAT
M8	iWAT_M8(r5)	19117550	GSE13432	GSM338984	Tissue	5 weeks	C57BL6, male	Inguinal white adipose tissue	30 °C (1 week)	WAT
M8	iWAT_M8(r6)	19117550	GSE13432	GSM338985	Tissue	5 weeks	C57BL6, male	Inguinal white adipose tissue	30 °C (1 week)	WAT
M8	Beige/brite (cold on iWAT)_M8(r1)	19117550	GSE13432	GSM338992	Tissue	5 weeks	C57BL6, male	Inguinal white adipose tissue	4 °C (5 weeks)	Beige/brite
M8	Beige/brite (cold on iWAT)_M8(r2)	19117550	GSE13432	GSM338993	Tissue	5 weeks	C57BL6, male	Inguinal white adipose tissue	4 °C (5 weeks)	Beige/brite
M8	Beige/brite (cold on iWAT)_M8(r3)	19117550	GSE13432	GSM338994	Tissue	5 weeks	C57BL6, male	Inguinal white adipose tissue	4 °C (5 weeks)	Beige/brite
M8	Beige/brite (cold on iWAT)_M8(r4)	19117550	GSE13432	GSM338986	Tissue	5 weeks	C57BL6, male	Inguinal white adipose tissue	4 °C (1 week)	Beige/brite
M8	Beige/brite (cold on iWAT)_M8(r5)	19117550	GSE13432	GSM338987	Tissue	5 weeks	C57BL6, male	Inguinal white adipose tissue	4 °C (1 week)	Beige/brite
M8	Beige/brite (cold on iWAT)_M8(r6)	19117550	GSE13432	GSM338988	Tissue	5 weeks	C57BL6, male	Inguinal white adipose tissue	4 °C (1 week)	Beige/brite
M9	BAT_M9(r1)	21765057	GSE28440	GSM703077	Tissue	20 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M9	BAT_M9(r2)	21765057	GSE28440	GSM703078	Tissue	20 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M9	BAT_M9(r3)	21765057	GSE28440	GSM703079	Tissue	20 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M9	eWAT_M9(r1)	21765057	GSE28440	GSM703086	Tissue	20 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M9	eWAT_M9(r2)	21765057	GSE28440	GSM703087	Tissue	20 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M9	eWAT_M9(r3)	21765057	GSE28440	GSM703088	Tissue	20 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M9	pvWAT_M9(r1)	21765057	GSE28440	GSM703080	Tissue	20 weeks	C57BL6, male	Perivascular white adipose tissue	None	WAT
M9	pvWAT_M9(r2)	21765057	GSE28440	GSM703081	Tissue	20 weeks	C57BL6, male	Perivascular white adipose tissue	None	WAT
M9	pvWAT_M9(r3)	21765057	GSE28440	GSM703082	Tissue	20 weeks	C57BL6, male	Perivascular white adipose tissue	None	WAT
M9	iWAT_M9(r1)	21765057	GSE28440	GSM703083	Tissue	20 weeks	C57BL6, male	inguinal white adipose tissue	None	WAT
M9	iWAT_M9(r2)	21765057	GSE28440	GSM703084	Tissue	20 weeks	C57BL6, male	inguinal white adipose tissue	None	WAT
M9	iWAT_M9(r3)	21765057	GSE28440	GSM703085	Tissue	20 weeks	C57BL6, male	inguinal white adipose tissue	None	WAT
M10	BAT_M10(r1)	24549398	GSE51080	GSM1237806	Tissue	10 weeks	SV129, female	Interscapular brown adipose tissue	28 °C (10 days)	BAT
M10	BAT_M10(r2)	24549398	GSE51080	GSM1237794	Tissue	10 weeks	SV129, female	Interscapular brown adipose tissue	28 °C (10 days)	BAT
M10	BAT_M10(r3)	24549398	GSE51080	GSM1237795	Tissue	10 weeks	SV129, female	Interscapular brown adipose tissue	28 °C (10 days)	BAT
M10	Treated BAT (cold)_M10(r1)	24549398	GSE51080	GSM1237792	Tissue	10 weeks	SV129, female	Interscapular brown adipose tissue	6 °C (10 days)	Treated BAT
M10	Treated BAT (cold)_M10(r2)	24549398	GSE51080	GSM1237800	Tissue	10 weeks	SV129, female	Interscapular brown adipose tissue	6 °C (10 days)	Treated BAT
M10	Treated BAT (cold)_M10(r3)	24549398	GSE51080	GSM1237803	Tissue	10 weeks	SV129, female	Interscapular brown adipose tissue	6 °C (10 days)	Treated BAT
M10	sWAT_M10(r1)	24549398	GSE51080	GSM1237797	Tissue	10 weeks	SV129, female	Subcutaneous white adipose tissue	28 °C (10 days)	WAT
M10	sWAT_M10(r2)	24549398	GSE51080	GSM1237798	Tissue	10 weeks	SV129, female	Subcutaneous white adipose tissue	28 °C (10 days)	WAT
M10	sWAT_M10(r3)	24549398	GSE51080	GSM1237793	Tissue	10 weeks	SV129, female	Subcutaneous white adipose tissue	28 °C (10 days)	WAT
M10	Beige/brite (cold on sWAT)_M10(r1)	24549398	GSE51080	GSM1237790	Tissue	10 weeks	SV129, female	Subcutaneous white adipose tissue	6 °C (10 days)	Beige/brite
M10	Beige/brite (cold on sWAT)_M10(r2)	24549398	GSE51080	GSM1237796	Tissue	10 weeks	SV129, female	Subcutaneous white adipose tissue	6 °C (10 days)	Beige/brite
M10	Beige/brite (cold on sWAT)_M10(r3)	24549398	GSE51080	GSM1237799	Tissue	10 weeks	SV129, female	Subcutaneous white adipose tissue	6 °C (10 days)	Beige/brite
M10	mWAT_M10(r1)	24549398	GSE51080	GSM1237801	Tissue	10 weeks	SV129, female	Mesenteric white adipose tissue	28 °C (10 days)	WAT
M10	mWAT_M10(r2)	24549398	GSE51080	GSM1237802	Tissue	10 weeks	SV129, female	Mesenteric white adipose tissue	28 °C (10 days)	WAT
M10	mWAT_M10(r3)	24549398	GSE51080	GSM1237791	Tissue	10 weeks	SV129, female	Mesenteric white adipose tissue	28 °C (10 days)	WAT
M10	Beige/brite (cold on mWAT)_M10(r1)	24549398	GSE51080	GSM1237804	Tissue	10 weeks	SV129, female	Mesenteric white adipose tissue	6 °C (10 days)	Beige/brite
M10	Beige/brite (cold on mWAT)_M10(r2)	24549398	GSE51080	GSM1237805	Tissue	10 weeks	SV129, female	Mesenteric white adipose tissue	6 °C (10 days)	Beige/brite
M10	Beige/brite (cold on mWAT)_M10(r3)	24549398	GSE51080	GSM1237807	Tissue	10 weeks	SV129, female	Mesenteric white adipose tissue	6 °C (10 days)	Beige/brite

M11	Treated BAT (forskolin)_M11(BAT1)	22796012	GSE39562	GSM971717	Primary adipocytes	N/A	N/A	Interscapular brown adipose tissue	Forskolin (10μM, 4 hrs)	Treated BAT
M11	Treated BAT (forskolin)_M11(BAT6)	22796012	GSE39562	GSM971718	Primary adipocytes	N/A	N/A	Interscapular brown adipose tissue	Forskolin (10μM, 4 hrs)	Treated BAT
M11	Treated BAT (forskolin)_M11(BAT8)	22796012	GSE39562	GSM971719	Primary adipocytes	N/A	N/A	Interscapular brown adipose tissue	Forskolin (10μM, 4 hrs)	Treated BAT
M11	Beige/brite (forskolin on iWAT)_M11(X9)	22796012	GSE39562	GSM971716	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	Forskolin (10μM, 4 hrs)	Beige/brite
M11	Beige/brite (forskolin on iWAT)_M11(D12)	22796012	GSE39562	GSM971698	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	Forskolin (10μM, 4 hrs)	Beige/brite
M11	Beige/brite (forskolin on iWAT)_M11(G10)	22796012	GSE39562	GSM971705	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	Forskolin (10μM, 4 hrs)	Beige/brite
M11	Beige/brite (forskolin on iWAT)_M11(I2)	22796012	GSE39562	GSM971709	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	Forskolin (10μM, 4 hrs)	Beige/brite
M11	Beige/brite (forskolin on iWAT)_M11(A5)	22796012	GSE39562	GSM971695	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	Forskolin (10μM, 4 hrs)	Beige/brite
M11	Beige/brite (forskolin on iWAT)_M11(A3)	22796012	GSE39562	GSM971694	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	Forskolin (10μM, 4 hrs)	Beige/brite
M11	Beige/brite (forskolin on iWAT)_M11(B5)	22796012	GSE39562	GSM971696	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	Forskolin (10μM, 4 hrs)	Beige/brite
M11	Beige/brite (forskolin on iWAT)_M11(B13)	22796012	GSE39562	GSM971697	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	Forskolin (10μM, 4 hrs)	Beige/brite
M11	Beige/brite (forskolin on iWAT)_M11(D16)	22796012	GSE39562	GSM971699	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	Forskolin (10μM, 4 hrs)	Beige/brite
M11	Beige/brite (forskolin on iWAT)_M11(E1)	22796012	GSE39562	GSM971700	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	Forskolin (10μM, 4 hrs)	Beige/brite
M11	Beige/brite (forskolin on iWAT)_M11(E3)	22796012	GSE39562	GSM971701	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	Forskolin (10μM, 4 hrs)	Beige/brite
M11	Beige/brite (forskolin on iWAT)_M11(E4)	22796012	GSE39562	GSM971702	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	Forskolin (10μM, 4 hrs)	Beige/brite
M11	Beige/brite (forskolin on iWAT)_M11(F4)	22796012	GSE39562	GSM971703	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	Forskolin (10μM, 4 hrs)	Beige/brite
M11	Beige/brite (forskolin on iWAT)_M11(F9)	22796012	GSE39562	GSM971704	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	Forskolin (10μM, 4 hrs)	Beige/brite
M11	Beige/brite (forskolin on iWAT)_M11(G12)	22796012	GSE39562	GSM971706	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	Forskolin (10μM, 4 hrs)	Beige/brite
M11	Beige/brite (forskolin on iWAT)_M11(G18)	22796012	GSE39562	GSM971707	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	Forskolin (10μM, 4 hrs)	Beige/brite
M11	Beige/brite (forskolin on iWAT)_M11(G23)	22796012	GSE39562	GSM971708	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	Forskolin (10μM, 4 hrs)	Beige/brite
M11	Beige/brite (forskolin on iWAT)_M11(J6)	22796012	GSE39562	GSM971710	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	Forskolin (10μM, 4 hrs)	Beige/brite
M11	Beige/brite (forskolin on iWAT)_M11(M3)	22796012	GSE39562	GSM971711	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	Forskolin (10μM, 4 hrs)	Beige/brite
M11	Beige/brite (forskolin on iWAT)_M11(M4)	22796012	GSE39562	GSM971712	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	Forskolin (10μM, 4 hrs)	Beige/brite
M11	Beige/brite (forskolin on iWAT)_M11(N13)	22796012	GSE39562	GSM971713	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	Forskolin (10μM, 4 hrs)	Beige/brite
M11	Beige/brite (forskolin on iWAT)_M11(X7)	22796012	GSE39562	GSM971715	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	Forskolin (10μM, 4 hrs)	Beige/brite
M12	BAT_M12(r1)	25349387	GSE54650	GSM1321062	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M12	BAT_M12(r2)	25349387	GSE54650	GSM1321063	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M12	BAT_M12(r3)	25349387	GSE54650	GSM1321064	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M12	BAT_M12(r4)	25349387	GSE54650	GSM1321065	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M12	BAT_M12(r5)	25349387	GSE54650	GSM1321066	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M12	BAT_M12(r6)	25349387	GSE54650	GSM1321067	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M12	BAT_M12(r7)	25349387	GSE54650	GSM1321068	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M12	BAT_M12(r8)	25349387	GSE54650	GSM1321069	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M12	BAT_M12(r9)	25349387	GSE54650	GSM1321070	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M12	BAT_M12(r10)	25349387	GSE54650	GSM1321071	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M12	BAT_M12(r11)	25349387	GSE54650	GSM1321072	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M12	BAT_M12(r12)	25349387	GSE54650	GSM1321073	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M12	BAT_M12(r13)	25349387	GSE54650	GSM1321074	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M12	BAT_M12(r14)	25349387	GSE54650	GSM1321075	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M12	BAT_M12(r15)	25349387	GSE54650	GSM1321076	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M12	BAT_M12(r16)	25349387	GSE54650	GSM1321077	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M12	BAT_M12(r17)	25349387	GSE54650	GSM1321078	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M12	BAT_M12(r18)	25349387	GSE54650	GSM1321079	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M12	BAT_M12(r19)	25349387	GSE54650	GSM1321080	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M12	BAT_M12(r20)	25349387	GSE54650	GSM1321081	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M12	BAT_M12(r21)	25349387	GSE54650	GSM1321082	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M12	BAT_M12(r22)	25349387	GSE54650	GSM1321083	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M12	BAT_M12(r23)	25349387	GSE54650	GSM1321084	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M12	BAT_M12(r24)	25349387	GSE54650	GSM1321085	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
M12	eWAT_M12(r1)	25349387	GSE54650	GSM1321254	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M12	eWAT_M12(r2)	25349387	GSE54650	GSM1321255	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M12	eWAT_M12(r3)	25349387	GSE54650	GSM1321256	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M12	eWAT_M12(r4)	25349387	GSE54650	GSM1321257	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M12	eWAT_M12(r5)	25349387	GSE54650	GSM1321258	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M12	eWAT_M12(r6)	25349387	GSE54650	GSM1321259	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M12	eWAT_M12(r7)	25349387	GSE54650	GSM1321260	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M12	eWAT_M12(r8)	25349387	GSE54650	GSM1321261	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M12	eWAT_M12(r9)	25349387	GSE54650	GSM1321262	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M12	eWAT_M12(r10)	25349387	GSE54650	GSM1321263	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M12	eWAT_M12(r11)	25349387	GSE54650	GSM1321264	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M12	eWAT_M12(r12)	25349387	GSE54650	GSM1321265	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M12	eWAT_M12(r13)	25349387	GSE54650	GSM1321266	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M12	eWAT_M12(r14)	25349387	GSE54650	GSM1321267	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M12	eWAT_M12(r15)	25349387	GSE54650	GSM1321268	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M12	eWAT_M12(r16)	25349387	GSE54650	GSM1321269	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M12	eWAT_M12(r17)	25349387	GSE54650	GSM1321270	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M12	eWAT_M12(r18)	25349387	GSE54650	GSM1321271	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M12	eWAT_M12(r19)	25349387	GSE54650	GSM1321272	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M12	eWAT_M12(r20)	25349387	GSE54650	GSM1321273	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M12	eWAT_M12(r21)	25349387	GSE54650	GSM1321274	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M12	eWAT_M12(r22)	25349387	GSE54650	GSM1321275	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M12	eWAT_M12(r23)	25349387	GSE54650	GSM1321276	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M12	eWAT_M12(r24)	25349387	GSE54650	GSM1321277	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
M13	iWAT_M13(r1)	27974179	GSE87191	GSM2324311	Tissue	8-10 weeks	UCP1/Tomato mice	Inguinal white adipose tissue	None	WAT
M13	iWAT_M13(r2)	27974179	GSE87191	GSM2324312	Tissue	8-10 weeks	UCP1/Tomato mice	Inguinal white adipose tissue	None	WAT
M13	iWAT_M13(r3)	27974179	GSE87191	GSM2324313	Tissue	8-10 weeks	UCP1/Tomato mice	Inguinal white adipose tissue	None	WAT
M13	iWAT_M13(r4)	27974179	GSE87191	GSM2324314	Tissue	8-10 weeks	UCP1/Tomato mice	Inguinal white adipose tissue	None	WAT
M13	BAT_M13(r1)	27974179	GSE87191	GSM2324315	Tissue	8-10 weeks	UCP1/Tomato mice	Brown adipose tissue	None	BAT
M13	BAT_M13(r2)	27974179	GSE87191	GSM2324316	Tissue	8-10 weeks	UCP1/Tomato mice	Brown adipose tissue	None	BAT
M13	BAT_M13(r3)	27974179	GSE87191	GSM2324317	Tissue	8-10 weeks	UCP1/Tomato mice	Brown adipose tissue	None	BAT
M13	Beige/brite (CL on iWAT)_M13(r1)	27974179	GSE87191	GSM2324318	Tissue	8-10 weeks	UCP1/Tomato mice	Inguinal white adipose tissue	CL316,243L	Beige/brite
M13	Beige/brite (CL on iWAT)_M13(r2)	27974179	GSE87191	GSM2324319	Tissue	8-10 weeks	UCP1/Tomato mice	Inguinal white adipose tissue	CL316,243L	Beige/brite
M13	Beige/brite (CL on iWAT)_M13(r3)	27974179	GSE87191	GSM2324320	Tissue	8-10 weeks	UCP1/Tomato mice	Inguinal white adipose tissue	CL316,243L	Beige/brite
M13	Beige/brite (RS on iWAT)_M13(r1)	27974179	GSE87191	GSM2324321	Tissue	8-10 weeks	UCP1/Tomato mice	Inguinal white adipose tissue	Roscovitine	Beige/brite
M13	Beige/brite (RS on iWAT)_M13(r2)	27974179	GSE87191	GSM2324322	Tissue	8-10 weeks	UCP1/Tomato mice	Inguinal white adipose tissue	Roscovitine	Beige/brite
M13	Beige/brite (RS on iWAT)_M13(r3)	27974179	GSE87191	GSM2324323	Tissue	8-10 weeks	UCP1/Tomato mice	Inguinal white adipose tissue	Roscovitine	Beige/brite
M13	Beige/brite (RG on iWAT)_M13(r1)	27974179	GSE87191	GSM2324324	Tissue	8-10 weeks	UCP1/Tomato mice	Inguinal white adipose tissue	Rosiglitazone	Beige/brite
M13	Beige/brite (RG on iWAT)_M13(r2)	27974179	GSE87191	GSM2324325	Tissue	8-10 weeks	UCP1/Tomato mice	Inguinal white adipose tissue	Rosiglitazone	Beige/brite
M13	Beige/brite (RG on iWAT)_M13(r3)	27974179	GSE87191	GSM2324326	Tissue	8-10 weeks	UCP1/Tomato mice	Inguinal white adipose tissue	Rosiglitazone	Beige/brite
R1	Treated BAT (cold)_R1(r1)	24709624	GSE56248	GSM1357681	Tissue	6 weeks	C57BL6 UCP1-TRAP, female	Interscapular brown adipose tissue	4 °C (2 weeks)	Treated BAT
R1	Treated BAT (cold)_R1(r2)	24709624	GSE56248	GSM1357682	Tissue	6 weeks	C57BL6 UCP1-TRAP, female	Interscapular brown adipose tissue	4 °C (2 weeks)	Treated BAT
R1	Treated BAT (cold)_R1(r3)	24709624	GSE56248	GSM1357683	Tissue	6 weeks	C57BL6 UCP1-TRAP, female	Interscapular brown adipose tissue	4 °C (2 weeks)	Treated BAT
R1	Beige/brite (cold on iWAT)_R1(r1)	24709624	GSE56248	GSM1357676	Tissue	6 weeks	C57BL6 UCP1-TRAP, female	Inguinal white adipose tissue	4 °C (2 weeks)	Beige/brite
R1	Beige/brite (cold on iWAT)_R1(r2)	24709624	GSE56248	GSM1357677	Tissue	6 weeks	C57BL6 UCP1-TRAP, female	Inguinal white adipose tissue	4 °C (2 weeks)	Beige/brite
R1	Beige/brite (cold on iWAT)_R1(r3)	24709624	GSE56248	GSM1357678	Tissue	6 weeks	C57BL6 UCP1-TRAP, female	Inguinal white adipose tissue	4 °C (2 weeks)	Beige/brite

R2	BAT_R2(r1)	25559344	SRA188644	SRR1602555	Tissue	14 weeks	C57BL6 DIO	Interscapular brown adipose tissue	Vehicle	BAT
R2	Treated BAT (fex)_R2(r1)	25559344	SRA188644	SRR1602556	Tissue	14 weeks	C57BL6 DIO	Interscapular brown adipose tissue	Fex	Treated BAT
R2	iWAT_R2(r1)	25559344	SRA188644	SRR1597458	Tissue	14 weeks	C57BL6 DIO	Inguinal white adipose tissue	Vehicle	WAT
R2	Beige/brite (fex on iWAT)_R2(r1)	25559344	SRA188644	SRR1597454	Tissue	14 weeks	C57BL6 DIO	Inguinal white adipose tissue	Fex	Beige/brite
R2	gWAT_R2(r1)	25559344	SRA188644	SRR1596699	Tissue	14 weeks	C57BL6 DIO	Gonadal white adipose tissue	Vehicle	WAT
R2	Beige/brite (fex on gWAT)_R2(r1)	25559344	SRA188644	SRR1597452	Tissue	14 weeks	C57BL6 DIO	Gonadal white adipose tissue	Fex	Beige/brite
R3	BAT_R3(r1)	23166672	E-MTAB-2624	ERS472803	Primary adipocytes	N/A	N/A	Interscapular brown adipose tissue	None	BAT
R3	BAT_R3(r2)	23166672	E-MTAB-2624	ERS472808	Primary adipocytes	N/A	N/A	Interscapular brown adipose tissue	None	BAT
R3	iWAT_R3(r1)	23166672	E-MTAB-2624	ERS472804	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	None	WAT
R3	iWAT_R3(r2)	23166672	E-MTAB-2624	ERS472806	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	None	WAT
R3	Beige/brite (RG on iWAT)_R3(r1)	23166672	E-MTAB-2624	ERS472807	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	Rosiglitazone (1μM)	Beige/brite
R3	Beige/brite (RG on iWAT)_R3(r2)	23166672	E-MTAB-2624	ERS472805	Primary adipocytes	N/A	N/A	Inguinal white adipose tissue	Rosiglitazone (1μM)	Beige/brite
R4	BAT_R4(r1)	25349387	GSE54651	GSM1321302	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
R4	BAT_R4(r2)	25349387	GSE54651	GSM1321303	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
R4	BAT_R4(r3)	25349387	GSE54651	GSM1321304	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
R4	BAT_R4(r4)	25349387	GSE54651	GSM1321305	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
R4	BAT_R4(r5)	25349387	GSE54651	GSM1321306	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
R4	BAT_R4(r6)	25349387	GSE54651	GSM1321307	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
R4	BAT_R4(r7)	25349387	GSE54651	GSM1321308	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
R4	BAT_R4(r8)	25349387	GSE54651	GSM1321309	Tissue	6 weeks	C57BL6, male	Interscapular brown adipose tissue	None	BAT
R4	eWAT_R4(r1)	25349387	GSE54651	GSM1321366	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
R4	eWAT_R4(r2)	25349387	GSE54651	GSM1321367	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
R4	eWAT_R4(r3)	25349387	GSE54651	GSM1321368	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
R4	eWAT_R4(r4)	25349387	GSE54651	GSM1321369	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
R4	eWAT_R4(r5)	25349387	GSE54651	GSM1321370	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
R4	eWAT_R4(r6)	25349387	GSE54651	GSM1321371	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
R4	eWAT_R4(r7)	25349387	GSE54651	GSM1321372	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
R4	eWAT_R4(r8)	25349387	GSE54651	GSM1321373	Tissue	6 weeks	C57BL6, male	Epididymal white adipose tissue	None	WAT
hM1	Clone1 (subcutaneous neck)_subject 4	26076036	GSE68544	GSM1674867	SVF cells from subcutaneous and subplatysmal neck	N/A	Subject 4	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone2 (subcutaneous neck)_subject 3	26076036	GSE68544	GSM1674868	SVF cells from subcutaneous and subplatysmal neck	N/A	Subject 3	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone3 (subcutaneous neck)_subject 2	26076036	GSE68544	GSM1674869	SVF cells from subcutaneous and subplatysmal neck	N/A	Subject 2	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone4 (subcutaneous neck)_subject 1	26076036	GSE68544	GSM1674870	SVF cells from subcutaneous and subplatysmal neck	N/A	Subject 1	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone6 (subcutaneous neck)_subject 1	26076036	GSE68544	GSM1674871	SVF cells from subcutaneous and subplatysmal neck	N/A	Subject 1	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone7 (subcutaneous neck)_subject 1	26076036	GSE68544	GSM1674872	SVF cells from subcutaneous and subplatysmal neck	N/A	Subject 1	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone8 (subcutaneous neck)_subject 1	26076036	GSE68544	GSM1674873	SVF cells from subcutaneous and subplatysmal neck	N/A	Subject 1	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone9 (subcutaneous neck)_subject 1	26076036	GSE68544	GSM1674874	SVF cells from subcutaneous and subplatysmal neck	N/A	Subject 1	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone10 (deep neck)_subject 1	26076036	GSE68544	GSM1674875	SVF cells from deep neck	N/A	Subject 1	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone11 (deep neck)_subject 4	26076036	GSE68544	GSM1674876	SVF cells from deep neck	N/A	Subject 4	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone12 (deep neck)_subject 2	26076036	GSE68544	GSM1674877	SVF cells from deep neck	N/A	Subject 2	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone13 (deep neck)_subject 3	26076036	GSE68544	GSM1674878	SVF cells from deep neck	N/A	Subject 3	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone14 (deep neck)_subject 4	26076036	GSE68544	GSM1674879	SVF cells from deep neck	N/A	Subject 4	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone15 (deep neck)_subject 2	26076036	GSE68544	GSM1674880	SVF cells from deep neck	N/A	Subject 2	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone16 (deep neck)_subject 4	26076036	GSE68544	GSM1674881	SVF cells from deep neck	N/A	Subject 4	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone17 (deep neck)_subject 1	26076036	GSE68544	GSM1674882	SVF cells from deep neck	N/A	Subject 1	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone18 (deep neck)_subject 1	26076036	GSE68544	GSM1674883	SVF cells from deep neck	N/A	Subject 1	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone19 (deep neck)_subject 4	26076036	GSE68544	GSM1674884	SVF cells from deep neck	N/A	Subject 4	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone20 (deep neck)_subject 1	26076036	GSE68544	GSM1674885	SVF cells from deep neck	N/A	Subject 1	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone21 (deep neck)_subject 1	26076036	GSE68544	GSM1674886	SVF cells from deep neck	N/A	Subject 1	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone22 (deep neck)_subject 2	26076036	GSE68544	GSM1674887	SVF cells from deep neck	N/A	Subject 2	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone23 (deep neck)_subject 1	26076036	GSE68544	GSM1674888	SVF cells from deep neck	N/A	Subject 1	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone24 (deep neck)_subject 1	26076036	GSE68544	GSM1674889	SVF cells from deep neck	N/A	Subject 1	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone25 (deep neck)_subject 1	26076036	GSE68544	GSM1674890	SVF cells from deep neck	N/A	Subject 1	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone26 (deep neck)_subject 1	26076036	GSE68544	GSM1674891	SVF cells from deep neck	N/A	Subject 1	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone27 (deep neck)_subject 1	26076036	GSE68544	GSM1674892	SVF cells from deep neck	N/A	Subject 1	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone28 (deep neck)_subject 1	26076036	GSE68544	GSM1674893	SVF cells from deep neck	N/A	Subject 1	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone29 (deep neck)_subject 1	26076036	GSE68544	GSM1674894	SVF cells from deep neck	N/A	Subject 1	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone30 (deep neck)_subject 1	26076036	GSE68544	GSM1674895	SVF cells from deep neck	N/A	Subject 1	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone31 (deep neck)_subject 1	26076036	GSE68544	GSM1674896	SVF cells from deep neck	N/A	Subject 1	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone32 (deep neck)_subject 4	26076036	GSE68544	GSM1674897	SVF cells from deep neck	N/A	Subject 4	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone33 (deep neck)_subject 4	26076036	GSE68544	GSM1674898	SVF cells from deep neck	N/A	Subject 4	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone34 (deep neck)_subject 2	26076036	GSE68544	GSM1674899	SVF cells from deep neck	N/A	Subject 2	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone35 (deep neck)_subject 2	26076036	GSE68544	GSM1674900	SVF cells from deep neck	N/A	Subject 2	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone36 (deep neck)_subject 2	26076036	GSE68544	GSM1674901	SVF cells from deep neck	N/A	Subject 2	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone37 (deep neck)_subject 1	26076036	GSE68544	GSM1674902	SVF cells from deep neck	N/A	Subject 1	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone38 (deep neck)_subject 1	26076036	GSE68544	GSM1674903	SVF cells from deep neck	N/A	Subject 1	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone39 (deep neck)_subject 1	26076036	GSE68544	GSM1674904	SVF cells from deep neck	N/A	Subject 1	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone40 (deep neck)_subject 4	26076036	GSE68544	GSM1674905	SVF cells from deep neck	N/A	Subject 4	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone41 (deep neck)_subject 1	26076036	GSE68544	GSM1674906	SVF cells from deep neck	N/A	Subject 1	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM1	Clone42 (deep neck)_subject 1	26076036	GSE68544	GSM1674907	SVF cells from deep neck	N/A	Subject 1	Immortalized human clonal preadipocytes	Immortalization, culture, differentiation	N/A
hM2	Deep neck_patient 1	25102227	GSE54280	GSM1311783	Deep neck	N/A	Patient 1	Primary human adipocytes	None	N/A
hM2	Deep neck_patient 2	25102227	GSE54280	GSM1311785	Deep neck	N/A	Patient 2	Primary human adipocytes	None	N/A
hM2	Deep neck_patient 3	25102227	GSE54280	GSM1311787	Deep neck	N/A	Patient 3	Primary human adipocytes	None	N/A
hM2	Deep neck_patient 4	25102227	GSE54280	GSM1311789	Deep neck	N/A	Patient 4	Primary human adipocytes	None	N/A
hM2	Deep neck_patient 5	25102227	GSE54280	GSM1311791	Deep neck	N/A	Patient 5	Primary human adipocytes	None	N/A
hM2	Deep neck_patient 6	25102227	GSE54280	GSM1311793	Deep neck	N/A	Patient 6	Primary human adipocytes	None	N/A
hM2	Subcutaneous neck_patient 1	25102227	GSE54280	GSM1311784	Subcutaneous neck	N/A	Patient 1	Primary human adipocytes	None	N/A
hM2	Subcutaneous neck_patient 2	25102227	GSE54280	GSM1311786	Subcutaneous neck	N/A	Patient 2	Primary human adipocytes	None	N/A
hM2	Subcutaneous neck_patient 3	25102227	GSE54280	GSM1311788	Subcutaneous neck	N/A	Patient 3	Primary human adipocytes	None	N/A
hM2	Subcutaneous neck_patient 4	25102227	GSE54280	GSM1311790	Subcutaneous neck	N/A	Patient 4	Primary human adipocytes	None	N/A
hM2	Subcutaneous neck_patient 5	25102227	GSE54280	GSM1311792	Subcutaneous neck	N/A	Patient 5	Primary human adipocytes	None	N/A
hM2	Subcutaneous neck_patient 6	25102227	GSE54280	GSM1311794	Subcutaneous neck	N/A	Patient 6	Primary human adipocytes	None	N/A
hM3	sWAT (cold acclimation)_patient 2	26147760	GSE67297	GSM1644008	Abdominal subcutaneous fat	N/A	Patient 2	Primary adipocytes from human type 2 diabetic patients	After 10 days of cold acclimation	N/A

hM3	sWAT (cold acclimation)_patient 3	26147760	GSE67297	GSM1644012	Abdominal subcutaneous fat	N/A	Patient 3	Primary adipocytes from human type 2 diabetic patients	After 10 days of cold acclimation	N/A
hM3	sWAT (cold acclimation)_patient 4	26147760	GSE67297	GSM1644016	Abdominal subcutaneous fat	N/A	Patient 4	Primary adipocytes from human type 2 diabetic patients	After 10 days of cold acclimation	N/A
hM3	sWAT (cold acclimation)_patient 5	26147760	GSE67297	GSM1644005	Abdominal subcutaneous fat	N/A	Patient 5	Primary adipocytes from human type 2 diabetic patients	After 10 days of cold acclimation	N/A
hM3	sWAT (cold acclimation)_patient 6	26147760	GSE67297	GSM1644009	Abdominal subcutaneous fat	N/A	Patient 6	Primary adipocytes from human type 2 diabetic patients	After 10 days of cold acclimation	N/A
hM3	sWAT (cold acclimation)_patient 7	26147760	GSE67297	GSM1644013	Abdominal subcutaneous fat	N/A	Patient 7	Primary adipocytes from human type 2 diabetic patients	After 10 days of cold acclimation	N/A
hM3	sWAT (cold acclimation)_patient 8	26147760	GSE67297	GSM1644017	Abdominal subcutaneous fat	N/A	Patient 8	Primary adipocytes from human type 2 diabetic patients	After 10 days of cold acclimation	N/A
hM3	sWAT_patient 2	26147760	GSE67297	GSM1644006	Abdominal subcutaneous fat	N/A	Patient 2	Primary adipocytes from human type 2 diabetic patients	Before cold acclimation	N/A
hM3	sWAT_patient 3	26147760	GSE67297	GSM1644010	Abdominal subcutaneous fat	N/A	Patient 3	Primary adipocytes from human type 2 diabetic patients	Before cold acclimation	N/A
hM3	sWAT_patient 4	26147760	GSE67297	GSM1644014	Abdominal subcutaneous fat	N/A	Patient 4	Primary adipocytes from human type 2 diabetic patients	Before cold acclimation	N/A
hM3	sWAT_patient 5	26147760	GSE67297	GSM1644004	Abdominal subcutaneous fat	N/A	Patient 5	Primary adipocytes from human type 2 diabetic patients	Before cold acclimation	N/A
hM3	sWAT_patient 6	26147760	GSE67297	GSM1644007	Abdominal subcutaneous fat	N/A	Patient 6	Primary adipocytes from human type 2 diabetic patients	Before cold acclimation	N/A
hM3	sWAT_patient 7	26147760	GSE67297	GSM1644011	Abdominal subcutaneous fat	N/A	Patient 7	Primary adipocytes from human type 2 diabetic patients	Before cold acclimation	N/A
hM3	sWAT_patient 8	26147760	GSE67297	GSM1644015	Abdominal subcutaneous fat	N/A	Patient 8	Primary adipocytes from human type 2 diabetic patients	Before cold acclimation	N/A
hR1	PSC-BAs (DMSO)_r1	25487280	SRP042186	GSM1396748	rentiated from PPARG2-CEBPB	N/A	Replicate 1	Human pluripotent stem-cell derived brown adipose cells	DMSO (after 7 days of induction/differentiation)	N/A
hR1	PSC-BAs (DMSO)_r2	25487280	SRP042186	GSM1396749	rentiated from PPARG2-CEBPB	N/A	Replicate 2	Human pluripotent stem-cell derived brown adipose cells	DMSO (after 7 days of induction/differentiation)	N/A
hR1	PSC-BAs (DMSO)_r3	25487280	SRP042186	GSM1396750	rentiated from PPARG2-CEBPB	N/A	Replicate 3	Human pluripotent stem-cell derived brown adipose cells	DMSO (after 7 days of induction/differentiation)	N/A
hR1	PSC-WAs (DMSO)_r1	25487280	SRP042186	GSM1396736	and differentiated from PPARG	N/A	Replicate 1	Human pluripotent stem-cell derived white adipose cells	DMSO (after 7 days of induction/differentiation)	N/A
hR1	PSC-WAs (DMSO)_r2	25487280	SRP042186	GSM1396737	and differentiated from PPARG	N/A	Replicate 2	Human pluripotent stem-cell derived white adipose cells	DMSO (after 7 days of induction/differentiation)	N/A
hR1	PSC-WAs (DMSO)_r3	25487280	SRP042186	GSM1396738	and differentiated from PPARG	N/A	Replicate 3	Human pluripotent stem-cell derived white adipose cells	DMSO (after 7 days of induction/differentiation)	N/A
hR1	PSC-WAs (JAK3i)_r1	25487280	SRP042186	GSM1396741	and differentiated from PPARG	N/A	Replicate 1	Human pluripotent stem-cell derived white adipose cells	JAK3 inhibitor tofacitinib (2 μM), (after 7 days of induction/differentiation)	N/A
hR1	PSC-WAs (JAK3i)_r2	25487280	SRP042186	GSM1396739	and differentiated from PPARG	N/A	Replicate 2	Human pluripotent stem-cell derived white adipose cells	JAK3 inhibitor tofacitinib (2 μM), (after 7 days of induction/differentiation)	N/A
hR1	PSC-WAs (JAK3i)_r3	25487280	SRP042186	GSM1396740	and differentiated from PPARG	N/A	Replicate 3	Human pluripotent stem-cell derived white adipose cells	JAK3 inhibitor tofacitinib (2 μM), (after 7 days of induction/differentiation)	N/A
hR1	PSC-WAs (SYKi)_r1	25487280	SRP042186	GSM1396742	and differentiated from PPARG	N/A	Replicate 1	Human pluripotent stem-cell derived white adipose cells	SYK inhibitor R406 (1 μM), (after 7 days of induction/differentiation)	N/A
hR1	PSC-WAs (SYKi)_r2	25487280	SRP042186	GSM1396743	and differentiated from PPARG	N/A	Replicate 2	Human pluripotent stem-cell derived white adipose cells	SYK inhibitor R406 (1 μM), (after 7 days of induction/differentiation)	N/A
hR1	PSC-WAs (SYKi)_r3	25487280	SRP042186	GSM1396744	and differentiated from PPARG	N/A	Replicate 3	Human pluripotent stem-cell derived white adipose cells	SYK inhibitor R406 (1 μM), (after 7 days of induction/differentiation)	N/A
hR2	Clone1 (BAT preadipocytes)	25774848	E-MTAB-2602	ERR522178	SVF cells from supraclavicular BAT	N/A	Replicate 1	Immortalized human clonal brown preadipocytes	None	N/A
hR2	Clone2 (BAT preadipocytes)	25774848	E-MTAB-2602	ERR522180	SVF cells from supraclavicular BAT	N/A	Replicate 2	Immortalized human clonal brown preadipocytes	None	N/A
hR2	Clone3 (BAT preadipocytes)	25774848	E-MTAB-2602	ERR522169	SVF cells from supraclavicular BAT	N/A	Replicate 3	Immortalized human clonal brown preadipocytes	None	N/A
hR2	Clone1 (BAT)	25774848	E-MTAB-2602	ERR522176	SVF cells from supraclavicular BAT	N/A	Replicate 1	Immortalized and differentiated human clonal brown adipocytes	None	N/A
hR2	Clone2 (BAT)	25774848	E-MTAB-2602	ERR522186	SVF cells from supraclavicular BAT	N/A	Replicate 2	Immortalized and differentiated human clonal brown adipocytes	None	N/A
hR2	Clone3 (BAT)	25774848	E-MTAB-2602	ERR522185	SVF cells from supraclavicular BAT	N/A	Replicate 3	Immortalized and differentiated human clonal brown adipocytes	None	N/A
hR2	Clone1 (BAT + forskolin)	25774848	E-MTAB-2602	ERR522172	SVF cells from supraclavicular BAT	N/A	Replicate 1	Immortalized and differentiated human clonal brown adipocytes	Forskolin (10 μM, 4 hours)	N/A
hR2	Clone2 (BAT + forskolin)	25774848	E-MTAB-2602	ERR522184	SVF cells from supraclavicular BAT	N/A	Replicate 2	Immortalized and differentiated human clonal brown adipocytes	Forskolin (10 μM, 4 hours)	N/A
hR2	Clone3 (BAT + forskolin)	25774848	E-MTAB-2602	ERR522173	SVF cells from supraclavicular BAT	N/A	Replicate 3	Immortalized and differentiated human clonal brown adipocytes	Forskolin (10 μM, 4 hours)	N/A
hR2	Clone1 (sWAT preadipocytes)	25774848	E-MTAB-2602	ERR522170	SVF cells from subcutaneous WAT	N/A	Replicate 1	Immortalized human clonal white preadipocytes	None	N/A
hR2	Clone2 (sWAT preadipocytes)	25774848	E-MTAB-2602	ERR522181	SVF cells from subcutaneous WAT	N/A	Replicate 2	Immortalized human clonal white preadipocytes	None	N/A
hR2	Clone3 (sWAT preadipocytes)	25774848	E-MTAB-2602	ERR522175	SVF cells from subcutaneous WAT	N/A	Replicate 3	Immortalized human clonal white preadipocytes	None	N/A
hR2	Clone1 (sWAT)	25774848	E-MTAB-2602	ERR522174	SVF cells from subcutaneous WAT	N/A	Replicate 1	Immortalized and differentiated human clonal white adipocytes	None	N/A
hR2	Clone2 (sWAT)	25774848	E-MTAB-2602	ERR522182	SVF cells from subcutaneous WAT	N/A	Replicate 2	Immortalized and differentiated human clonal white adipocytes	None	N/A
hR2	Clone3 (sWAT)	25774848	E-MTAB-2602	ERR522179	SVF cells from subcutaneous WAT	N/A	Replicate 3	Immortalized and differentiated human clonal white adipocytes	None	N/A
hR2	Clone1 (sWAT + forskolin)	25774848	E-MTAB-2602	ERR522177	SVF cells from subcutaneous WAT	N/A	Replicate 1	Immortalized and differentiated human clonal white adipocytes	Forskolin (10 μM, 4 hours)	N/A
hR2	Clone2 (sWAT + forskolin)	25774848	E-MTAB-2602	ERR522183	SVF cells from subcutaneous WAT	N/A	Replicate 2	Immortalized and differentiated human clonal white adipocytes	Forskolin (10 μM, 4 hours)	N/A
hR2	Clone3 (sWAT + forskolin)	25774848	E-MTAB-2602	ERR522171	SVF cells from subcutaneous WAT	N/A	Replicate 3	Immortalized and differentiated human clonal white adipocytes	Forskolin (10 μM, 4 hours)	N/A

Chapter 3

ERcarta: an inventory of human endoplasmic reticulum resident proteins

3.1 Results

3.1.1 Training sets

A list of 1017 human ERRP candidates (Table 3.1) was retrieved from literature mining and publicly available databases. As shown in Fig. 3.1, though those ERRPs are annotated as ER-localized in each queried database, only 1 out of 1017 ERRPs is shared by all six databases. Moreover, the number of ERRPs in a single database is limited (58 in CORUM and 118 in LIFEdb), indicating the necessity of curating a gold standard reference set of true localized ER proteins. After manually checking ER-localization evidence from biochemical or image-based technologies, *e.g.,* immunoprecipitation (IP) and immunofluorescence (IF), protein candidates, which were not verified by experiments, were excluded, yielding a list of 630 confirmed ERRPs (Table 3.1). In addition to the 21 ERRPs from literature, a curated list of 651 ERRPs was obtained and used as positive training dataset (PTD) for ERRP prediction. After combining non-ER proteins from Swiss-Prot and the primary ERRPs list, a number of 5282 none-ER-localized proteins were collected and used as negative training dataset (NTD). As shown in Table 3.1, though collected ERRPs are annotated as ER-localized in databases, some of them are not supported by literature. The percentages of true ERRPs ranging from 36.4% to 84.2%, highlighting the importance and necessity of manual curation. HPAD yields the highest percentage of ERRP (84,2 %), while the LIFEdb the lowest (36.4%).

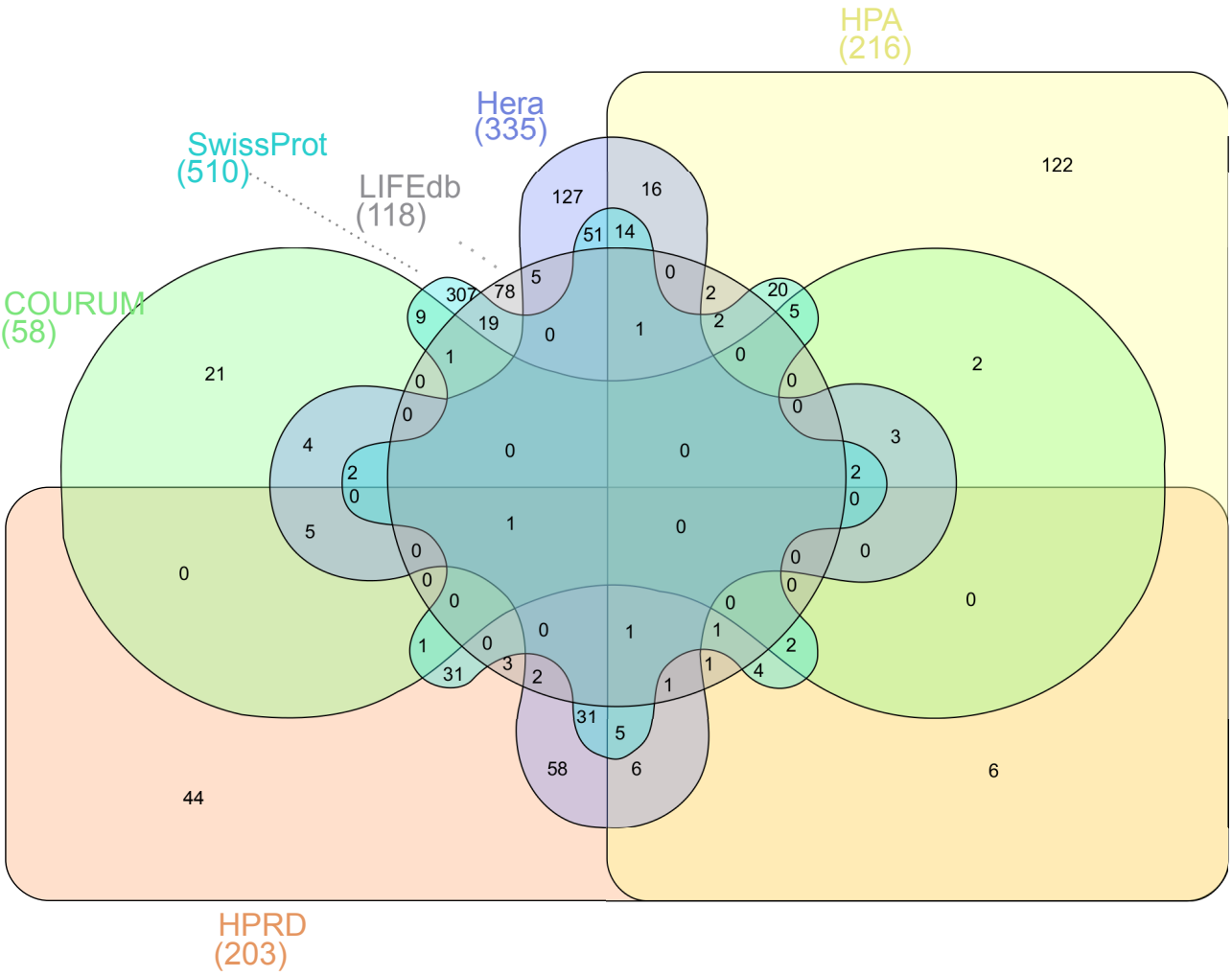


Fig. 3.1 **Distribution of confirmed ERRPs collected from databases.** HPA, Human protein atlas; HPRD, Human protein reference database.

Database	Before curation	After curation	ERRPs (%)
Swiss-Prot	510	369	72.4
Hera	335	279	83.3
HPA	216	116	53.7
HPRD	203	171	84.2
LIFEdb	118	43	36.4
CORUM	58	39	67.2
Literature	-	21	-
In total	1017	651	-

Table 3.1 Collection of ERRPs from databases and literature

	Features	Description	Predicted proteins	FDR (%)
Positive predictors	Pfam domains	Protein domains that are specific for ER proteins	75	12.96
	Mouse prot8	ER proteome from liver [145]	113	15.19
	ER targeting signals	KDEL-like motifs for ER lumen proteins [104]	58	16.28
	WoLF PSORT	WoLF PSORT predicted ER proteins in human	98	19.57
	LocTree3	LocTree3 predicted ER proteins in human	1013	19.76
	Human PPI	Protein interaction among human proteins [146]	768	31.09
	MultiLoc2	MultiLoc2 predicted ER proteins in human	248	31.41
	Yeast ortholog	Orthologs of yeast ER proteins	744	34.52
	Mouse prot5	ER proteome from rER pancreatic beta cells [147]	674	39.82
	Mouse prot2	ER proteome from liver [96]	701	51.51
	ER membrane motifs	KKxx, KxKxx motifs for ER membrane proteins [148–150]	452	52.60
	Yeast PPI	Protein interaction among yeast orthologs [146]	1129	55.95
	Mouse prot4	ER proteome from liver [97]	524	60.11
	Rat prot2	ER proteome from pancreas [99]	1384	61.85
	Mouse prot1	ER proteome from liver [95]	2855	70.67
	Rat prot1	ER proteome from live [98]	3797	72.79
	Human prot2	ER proteome from liver [100]	4563	73.52
	Mouse prot7	ER proteome from liver and kidney microsomes [100]	1932	73.53
	PSORT II	PRSOT II predicted ER proteins in human	7611	75.72
	Mouse prot6	ER proteome from liver [151]	5576	76.14
	Human prot1	ER proteome from kidney [100]	5200	78.53
	Mouse prot3	ER proteome from liver [97]	755	78.82
	STAT	Promoter motif (TCCMAGAA) in human [96]	14425	88.77
	MyoD	Promoter motif (CNGNRNCAGGTGNNGNA) [96]	18044	88.98
	C-Myc/Max2	Promoter motif (SCRCRTGGC) in human [96]	15566	89.18
	NRF-2	Promoter motif (ACCGGAAGNG) in human [96]	13554	89.44
	C-Myc/Max1	Promoter motif (ACCACGTGGT) in human [96]	8047	90.03
	N-Myc	Promoter motif (CCACGTG) in human [96]	8819	90.08
	AhR	Promoter motif (CACGCNA) in human [96]	8646	90.09
Negative predictors	LocTree3	LocTree3 predicted secretory proteins in human	3544	88.66
	Secretory set	A list of collected human secretory proteins	283	88.96
	SignalP	Presence of signal peptide for human proteins	3600	91.05
	TargetP	Presence of signal peptide for human proteins	5416	93.49
	ERcarta		**1023**	**11.5**

Table 3.2 Features used to predict ERRP

3.1.2 ERcarta: a repository of human ER proteome

A total of 33 genome-wide features, which are derived from both sequence-based computational prediction and experimental identified ER proteome, were collected (Table 3.2). Due to ER's function in protein secretion, many proteins may be wrongly annotated as ER-localized, though they do not reside in the organelle. To Minimize the inclusion of secreted proteins in our ER parts-list, four features were selected as negative predictors of ER localization (subsubsection 3.3.3.8). Though each of these features can be used as a weak predictor in defining a human ER proteome, their performance was limited. A majority of individual features give either high false discovery rate (FDR) or extremely small numbers of human ER proteins (Table 3.2). Seven *cis*-regulatory motifs yield a minimum of 88% FDR. Even seven of published ER proteomic datasets, which were generated from experimentally purifies microsomes, give FDRs higher than 70%. One exception is the mouse set8, an ER proteome extracted from mouse liver, which achieves 15.19% FDR but with a sacrifice in the number of detected ERRPs, predicting only 113 ERRPs in total, and the similar performance occurs for Pfam domains and ER targeting signals. Detailed performance of each feature (sensitivity and specificity) is shown in Fig. 3.2. Among all features, LocTree3 gives the best performance in ERRP identification, yielding 61.75% sensitivity and 98.12% specificity with 19.76% FDR.

To take advantage of all features, and thus improve the predictive power, all features were integrated by six machine learning algorithms. Ten-fold cross-validation was used during training to avoid overfitting and to calculate the performance indexes, *e.g.,* accuracy, sensitivity, specificity and FDR. Except for naive Bayes, five models outperformed the predictive power of all individual features (Fig. 3.2). Boosting model performed slightly better than other classifiers, yielding 71.6% sensitivity, 98.9% specificity and FDR of 11.5%, and was selected to predict human ERRPs. Using a probability threshold of 0.5 to define a predicted ERRP, boosting predicted a list of 1025 ERRPs, termed as ERcarta. Evidence of ER-localization of newly predicted ERRPs was manually checked, based on which a number of 18 ERRPs and two non-ERRP were confirmed and added to the PTD and NTD, respectively. Using the updated gold standard reference set, containing 669 ERRP in PTD and 5284 non-ERRP in NTD, boosting was trained for an additional round, ending up with the final version of ERcarta, which stores 1023 ERRPs. Not surprisingly, ERcarta accurately predicts 511 (76.4%) ERRPs out of 669 in the PTD. More importantly, a number of 354 novel ERRPs are identified. Annotation from cellular component of gene ontology (GOCC) of newly predicted ERRPs displays high confidence of ER-localization. Nearly half of the novel ERRPs (48.3%) are annotated as endoplasmic reticulum (GO:0005783), endoplasmic reticulum lumen (GO:0005788) or endoplasmic reticulum membrane (GO:0005789).

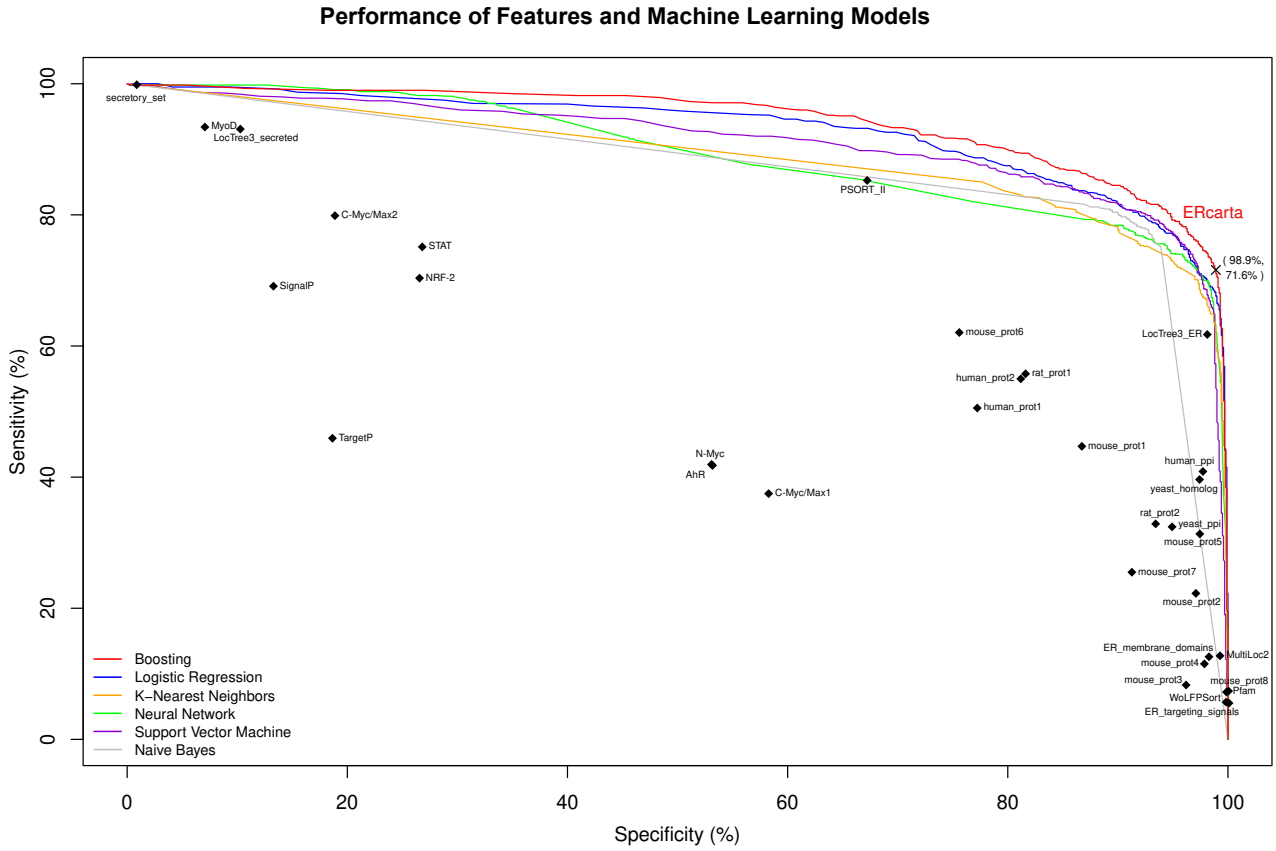


Fig. 3.2 **Sensitivity and specificity of features and integrated models.** Performance of individual features (black diamonds) and integrated machine learning models (solid lines) are displayed. Boosting model (solid red line) performs slightly better than other models and thus is selected for human ERRP prediction. Using a cutoff of 0.5 in defining an ERRP, boosting yields a 71.6% sensitivity and a 98.9%. ERcarta (the black cross over the red line) outperforms the predictive power of all individual features.

3.1.3 Function annotation of ERcarta

Function enrichment analysis on Pfam domains, KEGG pathway and biological process of GO (GOBP) were performed. The first enriched Pfam domain was Cytochrome P450 (CYPs), a group of heme proteins, which are usually bound to either ER or inner mitochondrial membrane in human [152]. CYPs oxidize fatty acids and xenobiotics, and also play essential roles in drug metabolism [153]. Consistently, two KEGG pathways, *i.e.,* "Drug metabolism - cytochrome P450" and "Metabolism of xenobiotics by cytochrome P450", and a biological process named "xenobiotic metabolic process" were also detected (Fig. 3.3B and 3.3C). Moreover, newly predicted ERRPs were found involved in ER-related activities such as "protein processing in endoplasmic reticulum", "protein transport", "post-translational protein modification".

To further explore biological roles of ERRPs, information which sheds light on potential functions, *e.g.,* the existence of transmembrane domains, ATP-binding sites and EF-hand for calcium-binding or the involvement in human diseases, was collected. As shown in Fig. 3.4A, more than 200 novel ERRPs, containing transmembrane domains, might represent potential ER transporters, channels or protein complexes; seven proteins, containing ATP-binding site, could mediate energy-consuming processes. A total number of 275 ERcarta genes were associated with 375 human diseases from the OMIM database, and 60 novel ERRPs were linked to 76 OMIM diseases. Abnormality of corresponding ER genes was observed in patients suffering cancers, liver or kidney anomalies [154, 155], *etc.*. The association was also supported by the GAD disease class analysis (Fig. 3.4B), which displays disease class of renal, metabolic or chemical dependency.

Finally, the evolutionary route of ERcarta proteins across 2048 species from metazoa, other eukaryotes, bacteria and archaea, was analyzed using ProtPhylo [156] database. As shown in Fig. 3.5, proteins sharing similar evolutionary trace were grouped together; some proteins were conserved in all four categories (blue-colored in all four categories), while others were only conserved in metazoa (blue-colored in metazoa only). Besides, ERcarta-associated diseases, *i.e.,* Alzheimer's disease, Parkinson's disease, cancer and liver diseases, were extracted from GAD, summarized and displayed in the side-heatmap (colored in green-red-purple) in Fig. 3.5. Phylogenetic profiles have been applied in previous studies to pinpoint potential function or pathogenicity of proteins [157, 158], based on the idea that functionally-related genes tend to be present or absent together during evolution. Two small clusters are shown as examples (on the right side of Fig. 3.5). The first cluster contains EPHX1 and proteins from the UDP-glucuronosyltransferase (UGT) family, which play crucial roles in human drug metabolism via catalyzing glucuronic acid [159], and the majority of UGT members are linked to cancer and liver diseases. Interestingly, EPHX1 is also known

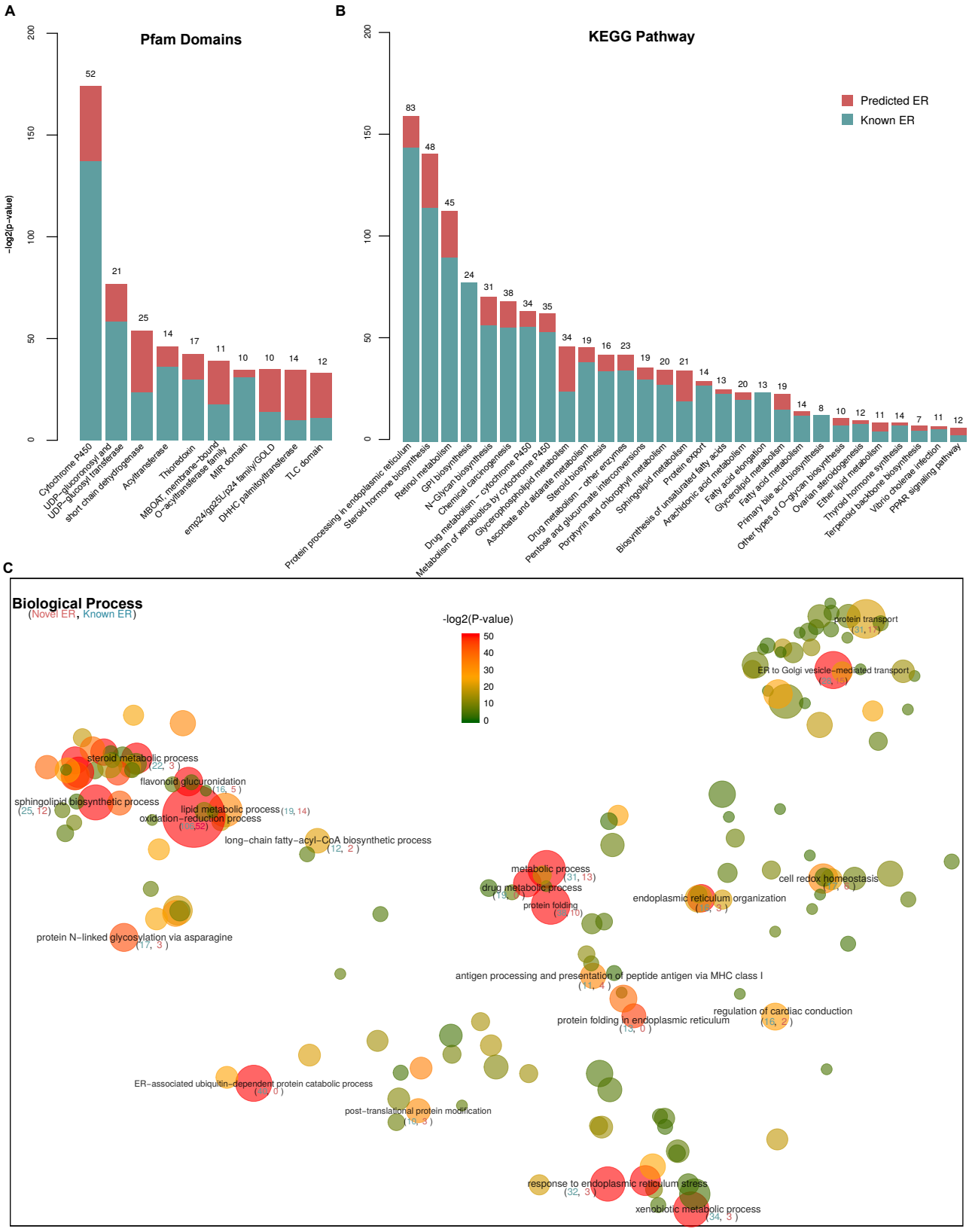


Fig. 3.3 **Function enrichment of ERcarta.** Enriched Pfam domains (**A**), KEGG pathways (**B**) and biological processes of gene ontology (**C**) for ERcarta. The black integers on the top of bars in **A** and **B** indicate the total number of ERRPs, which are involved in the corresponding Pfam domains and KEGG pathways. Blue- and red-colored areas of each bar in **A** and **B** indicate the percentage of known and newly predicted ERRPs, respectively.

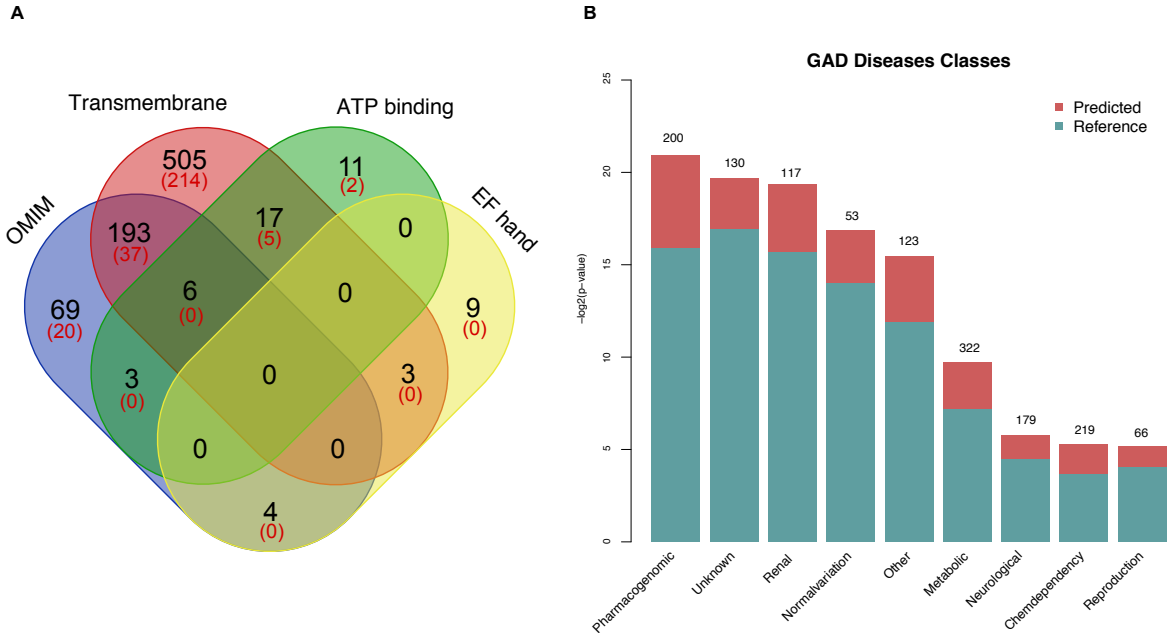


Fig. 3.4 **Function annotation on protein domains and diseases.** **(A)** Overlap among ERcarta proteins, which have transmembrane domains, ATP binding domains and EF-hand or involve in diseases. Black integers indicate the total number of genes involved while the red shows the number of newly predicted ERRPs. **(B)** Enriched disease categories for ERcarta. The integers on top of the bars indicate the number of ER genes that are involved in corresponding disease category, blue- and red-colored areas in the bar present the percentage of know and newly predicted ER proteins in each category, respectively.

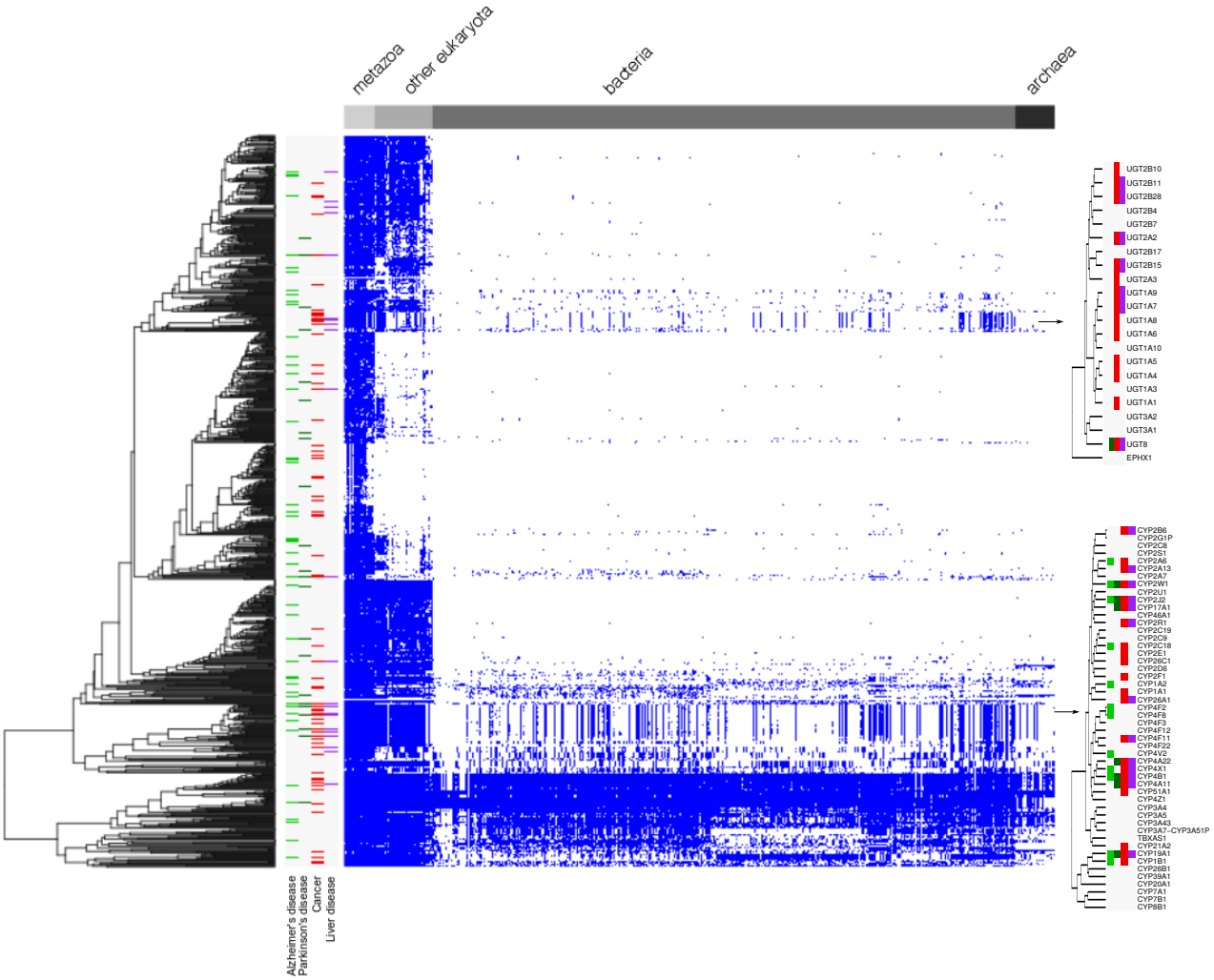


Fig. 3.5 **Phylogenetic profiling of ERcarta across 2048 organisms.** Blue-colored cell indicates the existence of a protein (row) in corresponding species (column). Left-side heatmap shows proteins' involvement in GAD diseases, *i.e.,* Alzheimer's disease (light green), Parkinson's disease (dark green), cancer (red) and liver diseases (purple). The black dendrogram shows protein groups, which are clustered by the similarity of evolutionary conservation. Right-side heatmaps are two examples of protein sub-groups.

as an enzyme transforming epoxides to diols [160], and contribute to pregnancy-induced hypertension, which might cause impaired liver function in severe cases [161]. GO annotation of EPHX1 is related to drug metabolism according to GeneCards, and HPA considers it as a potential drug target. The second group displays the members of the Cytochromes P450 (CYPs) family, a group of prevalent drug-metabolizing enzymes [162, 163].

3.2 Conclusion

Our manual confirmation of ERRPs indicates the dispersal of ERRPs among repositories (Fig. 3.1). Thus, it is difficult to choose a reliable and complete ER reference proteome when trying to elucidate the function of ER organelle systematically. To solve this, based on collected ERRPs in the PTD, non-ERRP in the NTD and features derived from protein sequence, protein interactions and experiments, we systematically extended the catalog of human ERRPs to 1023 proteins using boosting, which included 354 newly identified ERRPs.

Taking advantage of multiple resources, ERcarta is the most updated and comprehensive ER proteome in human and has the potential to be a valuable resource for elucidating the molecular basis of ER functions.

3.3 Methods

3.3.1 Human reference proteome

A complete, non-redundant set of 20,996 human proteins was obtained from EBI (ftp://ftp.ebi.ac.uk/pub/databases/reference_proteomes/QfO/Eukaryota/) based on UniProt release 2018_04, Ensembl release 91, and Ensembl Genome release 38. Within this set, 20,790 proteins that are at least 50 amino acids long were considered as human reference proteome for further analyses. For each protein in the reference set, its corresponding Ensembl gene id, Ensembl transcript ID and Ensembl peptide ID, gene name, NCBI entrez gene ID, RefSeq mRNA and peptide ID were assigned by R package biomaRt. If an ID was missing with biomaRt, corresponding information was then extracted from the human ID mapping file (UP000005640_9606.idmapping.gz from ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes/Eukaryota/). Biological annotation of proteins in the proteome was retrieved from UniProt release 2018_05.

3.3.2 Training sets

A reference dataset of ERRPs was curated based on literature survey and six public databases, including Hera [104], Human LIFEdb [164], CORUM [165], Swiss-Prot [166], The Human Protein Atlas (HPA) [167, 168] and Human Protein Reference Database (HPRD) [169]. Hera database provides the first list of manually curated ERRPs in human. LIFEdb stores human ERRPs, which are identified with green fluorescent protein (GFP). Based on a collection of open reading frames (ORFs) from Molecular Genome Analysis and the ORFeome resource, LIFEdb developed tools to tag ORFs with GFP and further determined subcellular location of ORFs via tracking fusion proteins [170]. CORUM serves as the most comprehensive resource of experimentally confirmed mammalian protein complexes, which are manually collected by checking published literature. Swiss-Prot contains a large number of human ERRPs, whose localizations are assigned by experiments and computational prediction. HPA aims to generate proteins atlas of human cells, tissues and organs, and its subcellular location annotation of proteins is based on immunofluorescently stained cells. Manually collecting biological knowledge for the human proteome from scientific papers, HPRD integrates various annotations, *e.g.,* protein domains, interactions, subcellular location, *etc.* Proteins annotated as ER-localizing were retrieved from each database, manually curated, and mapped to the human reference proteome. Only proteins with strong evidence of ER-localization from complementary biochemical and imaging-based assays (*e.g.,* immunoprecipitation, immunofluorescence, western blot, electron microscopy, mass spectrometry, *etc.*) were included in the PTD.

A NTD of human proteins was compiled including secretory proteins supported by literature survey, and proteins, which had no evidence of ER localization in Swiss-Prot database (release 2018_05). Briefly, all human proteins, excluding these in the PTD, were firstly filtered by limiting subcellular locations evidence to 'any experimental assertion'. Secondly, proteins, which contain words 'probable', 'putative', 'by similarity', 'possible' or 'endoplasmic reticulum' in their description of the subcellular location, were excluded. Thirdly, proteins, annotated as 'endoplasmic reticulum' in cellular component of gene ontology, were removed.

3.3.3 Features

Our collection of features includes ER retention signals, ER-specific protein domains, ER orthologs in yeast, ER-enriched *cis*-regulatory motifs, published mass spectrometry data, protein-protein interaction, predicted localization from available software and N-terminal

signal for secretory proteins (please refer to Table 3.2). The detailed method of each feature is introduced below.

3.3.3.1 ER retention signals

Motif KDEL> (PS00014 in PROSITE) at the C-terminal is a classical signal for ER lumen proteins [148]. Motifs KKXX> or KXKXX> (where X may be any kind of amino acid) in the cytosolic C terminus are the retention signal for type I ER membrane proteins [148–150]. Besides, two PROSITE motifs (PS00951, PS00952) are also signals for ER lumen proteins. Perl script of PROSITE (modified on 23.04.2018) was downloaded from ftp://ftp.expasy.org/databases/prosite/ps_scan/ps_scan.pl and used to detect the existence of these ER-retention signals for the whole human reference proteome with default parameters [171]. For each human protein, a score of 1 was assigned if as least one ER retrieval signal was found; otherwise, 0 was assigned.

3.3.3.2 Pfam domains

Protein domains are conserved protein sequences, which form specific functional units. Thus the existence of the same domains usually indicates the involvement in the same biological activities, which usually perform in the same organelle. A list of 116 ER-specific cellular component GO terms were extracted from AmiGO 2 [29, 172]. Pfam-A database (version 31.0) and associated GO annotation were downloaded from ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam31.0/. A Pfam domain is considered as an ER-specific domain if its associated GO terms are included in the 116 ER-specific GO terms. Hmmscan from HMMER v3.1b1 (http://hmmer.org/) was applied to identify Pfam domains for each human protein (cutoff 0.05). For each human protein, the total number of detected ER-specific domains was assigned.

3.3.3.3 Yeast ortholog

In order to get all ER-localized proteins in yeast, *Saccharomyces* genome database (SGD) (version 04/04/2018) was searched by limiting GO ID to GO:0005783, and three tables listing ER genes, whose localization was annotated by 'Manually Curated', 'High-throughput' and 'Computational', were returned (https://www.yeastgenome.org/go/GO:0005783) [173]. After merging, we got a list of 601 ER proteins in yeast. Next, to identify potential homologs, an all-versus-all proteome comparison between human and yeast was performed using blastp (version 2.2.29+) with expected value < 1E-5. For each human protein, if its yeast homolog localized at ER, a score of 1 was assigned; otherwise, 0 was assigned.

3.3.3.4 *Cis*-regulatory motifs in promoters

Enriched *cis*-regulatory motifs vary among genes localizing at different organelles. Seven motifs, *i.e.,* AhR (CACGCNA), c-Myc/Max(ACCACGTGGT and SCRCRTGGC), N-Myc (CCACGTG), NRF-2(ACCGGAAGNG), STAT (TCCMAGAA), MyoD (CNGNRNCAGGT-GNNGNA), were found enriched in upstream of ER-localized genes [96]. To detect the existence of these motifs, a region of 4kb around transcription starting site (TSS) (2kb from upstream and 2kb from downstream) were extracted based on annotation from http://genome.ucsc.edu/cgi-bin/hgTables, and searched against the seven motifs by FIMO using default parameters (p-value < 1E-4) [174]. For each protein, if one motif was detected, a score of 1 was assigned to this motif; otherwise, 0 was assigned.

3.3.3.5 Published ER datasets

A number of 12 published ER datasets were collected from literature and proteomics databases *e.g.*, PRIDE Archive [175] and ProteomeXchange [176]. These datasets comprise 6 sets from mouse liver [95–97, 151, 145], 1 from mouse rER pancreatic beta cells[147], 1 from mouse liver and kidney microsomes [100], 2 from rat liver [98] and pancreas [99] and 2 from human kidney and liver [100]. Briefly, Foster et al. generated an ER proteome from mouse liver homogenates, treated with gradient centrifugation followed by liquid chromatography/tandem mass spectrometry (LC/MS/MS) [96]. Similarly, Song et al. separated ER fraction from mouse liver homogenates using centrifugation, extracted peptides of proteins in the fraction via a gel-based approach and finally identified them by Nano-LC MS/MS [95]. In another project, separating proteins with both two-dimensional gel electrophoresis (2DE) and one-dimensional gel electrophoresis (1DE), Peng et al. obtained a proteomic profile of mouse liver microsomes [97]. Combining 1D SDS-PAGE and HPLC-MS/MS, Lee et al. performed a proteomic analysis of the ER pellets, which were isolated from MIN6 cells, a widely used mouse pancreatic beta cell line [147]. In Aumailley's paper, ER fractions from mouse liver were obtained using an ER enrichment assay kit (Novus Biologicals, Burlington, ON, Canada), followed with proteins digestion and identification with NanoLC-MS/MS [151]. Albertolle et al. generated multiple lists of sulfenylated microsomal proteins from mouse liver and kidney, human liver and kidney [100]. Gilchrist et al. isolated rough ER microsomes and smooth ER microsomes from rat liver, and quantified peptides with tandem mass spectrometry [98]. Chen et al. isolated rough ER of rat pancreas with centrifugation and quantified associated peptides with 2D LC-MALDI-MS/MS [99]. For each dataset, sequences of identified ER proteins were extracted, based on given ids such as Uniprot entry, gi number .*etc*, and used to create blast databases. Independent all-verse-all blasts (Evalue

1e-05) between extracted ER proteins and the human reference proteome were performed for homology detection. For each human protein in the reference proteome, a score of 1 was assigned if it has a mouse/rat/human ER homolog; otherwise, 0 was assigned.

3.3.3.6 Protein-protein interaction

Protein-protein interaction (PPI) was used to improve prediction of subcellular localization, especially for proteins locating at ER, plasma membrane and cytosol [146]. Human and yeast PPI were downloaded from STRING database (version 10.5) [177], and only physical links ('binding' actions) with combined score >= 800 were kept for further analysis. File '9606.protein.aliases.v10.5.txt' from STRING was used to map human proteins from reference proteome (with UniProt entry) to STRING (with ENSEMBLE IDs), and if a UniProt entry was missing in the aliases file, an all-verse-all blast between human reference proteome and all STRING human proteins was then performed, and the best match was kept as mapped STRING protein. For each human protein, their interaction partners were extracted, and the percentage of ER partners was calculated and assigned to the protein. Regarding the yeast PPI, if a human protein does not have a yeast homolog, a score of 0 was assigned. Otherwise, the percentage of yeast ER partners was calculated and assigned to the protein. An ER partner indicates an ER-localized protein from human RFER (see subsection 3.3.2) or the list of yeast ER proteins mentioned in subsubsection 3.3.3.3.

3.3.3.7 Subcellular localization prediction

Localization predictors, *i.e.,* LocTree3 [101], MultiLoc2 [102], PSORT II [178] and WoLF PSORT [179], were applied separately to predict subcellular localization of proteins in human reference proteome. Combining LocTree2, an SVM-based localization classifier, and homology information, LocTree3 can predict 18 subcellular localizations for Eukaryota. Integrating phylogenetic profiling and GO terms with SVMs, MultiLoc2 predicts 11 main subcellular localizations for Eukaryota. PSORT II can predict 11 subcellular classes based on protein sorting signals. Extending from PSORT II, WoLF PSORT predicts localizations via applying k-nearest neighbors on sequence-based features, *e.g.,* amino acid composition, *etc*. According to the output of each tool, original scores, indicating the probability of being ER-localized, were assigned to human proteins. If probability was not provided, a score of 1 or 0 was assigned to indicate being ER-localized or not.

3.3.3.8 Secretory proteins (negative predictors)

The signal sequence is a short peptide, present at the N-terminal of secretory proteins. As the starting point of secretion, rough ERs unavoidably contain many secretory proteins, which are considered as false positive ERRPs. SignalP-4.1 [180] and TargetP v1.1 [181] were used to predict the existence or absence of signal peptide and returned scores from both tools were assigned to corresponding human proteins. Moreover, LocTree3 [101] was also applied to detect secreted proteins, and the output scores, indicating the probability of secretion, were directly assigned to feature 'LocTree3_secreted' for all proteins. Besides the prediction, a list of human secreted proteins was extracted from http://www.bci.mcgill.ca/~hera/PSLT/datasets/, and corresponding UniProt entries of secretory proteins were extracted using biomaRt [138, 182]. For each human protein, if it was in the collected list, a score of 1 was assigned; otherwise, 0 was assigned.

3.3.4 Machine learning-based data integration

In total, six machine learning models, *i.e.,* Boosting, logistic regression, naive Bayes, support vector machine, neural network and k-nearest neighbors, were applied to integrate all features. Ten-fold cross-validation was used to train models and measure performance. Average sensitivity (TPR), specificity (SPC), false discovery rate (FDR) and accuracy (ACC) among ten folds were calculated and used for model selection. Boosting, giving the best performance, was selected for the final prediction of ERRPs (Fig. 3.2). R package "gbm" was used to implement the boosting algorithm. TPR, SPC, FDR and ACC are defined as (3.1)

$$\begin{aligned} TPR &= \frac{TP}{TP+FN} \\ SPC &= \frac{TN}{TN+FP} \\ FDR &= \frac{FP}{TP+FP} \\ ACC &= \frac{TP+TN}{TP+TN+FP+FN} \\ FPR &= \frac{FP}{FP+TN} \\ &= 1-SPC \end{aligned} \tag{3.1}$$

, where TP and FP represent true positive and false positive ERRPs, TN and FN are true negative and false negative ERRPs.

Chapter 4

Mito-ER crosstalk

4.1 Results

4.1.1 Selection of protein interaction links and clustering methods

Multiple types of protein interaction links and clustering approaches were tested for clusters prediction. The selection of protein associations included nine types of phylogenetic links obtained from ProtPhylo database [156] and the functional interaction from STRING database [177]. The four clustering techniques are ClusterONE, MCL, HC and DHC. As shown in appendix Fig. 4.3, regardless of the selection of clustering methods or the referenced database (KEGG or CORUM), ARI and F1 scores, obtained from STRING-based clusters, are much higher than those from pure phylogenetic links. Thus, PPI from STRING database was chosen for further analysis. Considering that STRING uses KEGG pathways to assign the confidence scores of PPI, only CORUM complexes were used as reference clusters for unbiased evaluation. As shown in appendix Fig. 4.4, the performance varies algorithms, depending on the total number of genes included in the clusters (cluster size $>= 3$). Aiming to keep as many genes as possible, DHC, yielding higher and more stable ARI and F1 scores, was selected for final clustering prediction.

4.1.2 Mito-ER regulatory network and functional modules

A Mito-ER regulatory map was reconstructed and shown in Fig. 4.1**A**. This network was then divided into small functional modules by using DHC method with various combination of parameters. The best prediction, yielding an ARI of 0.30 and an F1 score of 0.31, was selected to define functional clusters, which included in total 1954 proteins including 1095 from MitoCarta (94.56% MitoCarta) and 895 from ERcarta (87.49% ERcarta). These proteins

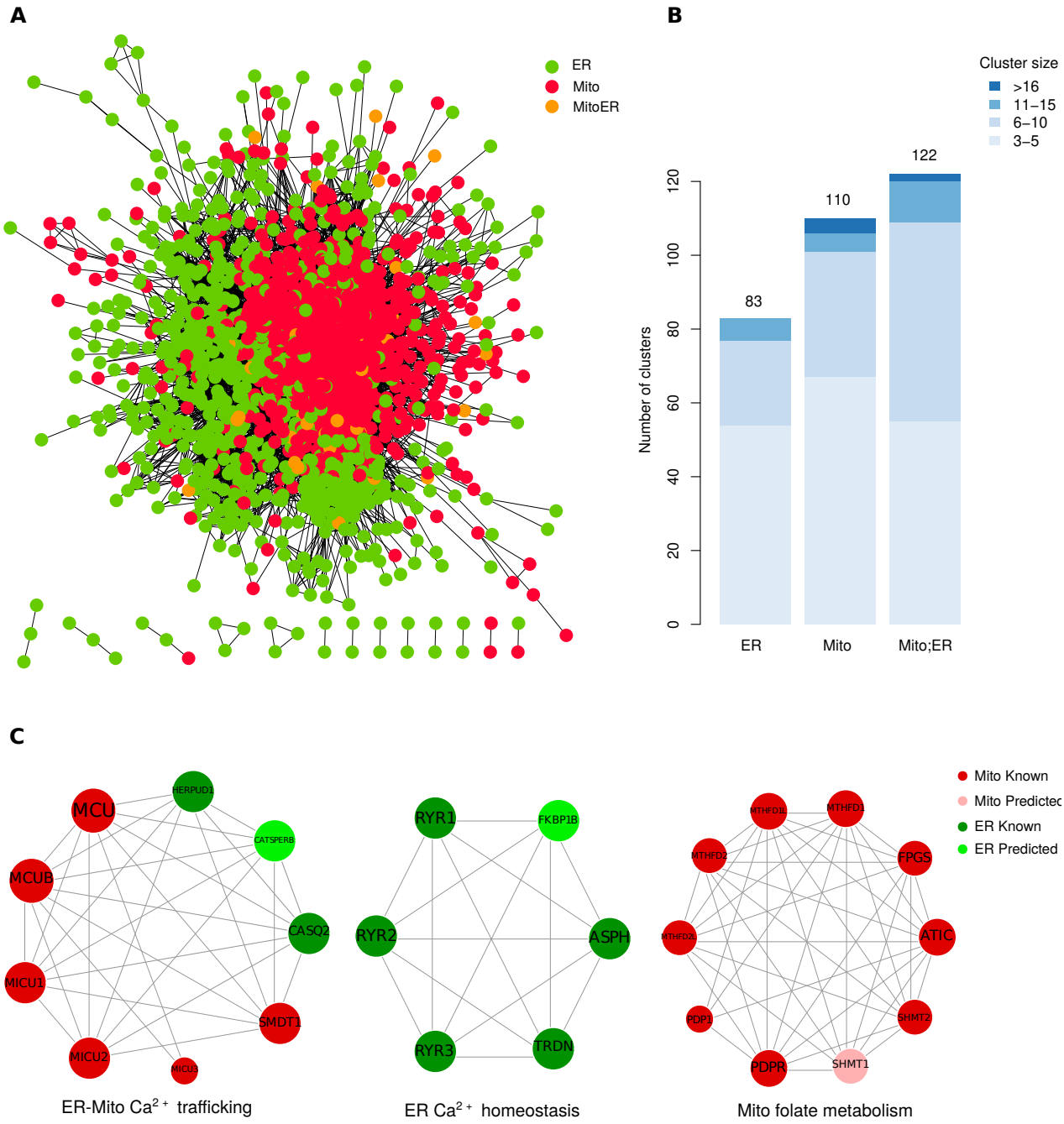


Fig. 4.1 **The comprehensive Mito-ER regulatory network and predicted functional clusters. (A)** an overview of the comprehensive Mito-ER regulatory map. Interaction associations among genes are from STRING database. Each node represents a gene, which is from ER-carta (green), MitoCarta (red) or both (orange). **(B)** Type and size of predicted clusters. Heights of stacked bar plots indicate the number of functional modules of three types of cluster, and the shadow display percentage of clusters with different size within each type. **(C)** Examples of three predicted functional modules. From left to right, they are a Mito-ER cluster functioning in Mito-ER Ca^{2+} trafficking, an ER cluster in Ca^{2+} homeostasis and a mitochondrial cluster in folate metabolism.

were divided into 321 clusters, each of which consists of at least three proteins. According to the localization of involved proteins, these functional modules are classified into pure mitochondrial clusters, pure ER clusters and Mito-ER clusters. In total, we obtained 83 pure ER clusters, 110 pure mitochondrial clusters and 122 Mito-ER clusters. Full list of predicted functional models is shown in supplementary table 4.2. Regardless of the cluster type, the majority of clusters are small-size modules, which are comprised of 3-10 proteins (Fig. 4.1**B**)). Each cluster is representing a certain functional module, and potential function of proteins can be inferred from proteins of known function in the same cluster. Three predicted modules, which are involved in ER-Mito Ca^{2+} trafficking, ER Ca^{2+} homeostasis and mitochondrial folate metabolism, are shown in Fig. 4.1**C**.

The first Mito-ER Ca^{2+} trafficking module contains nine members, including 6 mitochondrial proteins and 3 ER-localized proteins. All 6 mitochondrial proteins, *i.e.,* MCU, MCUB, MICU1, MICU2, MICU3 and SMDT1, are known for their roles in mitochondrial Ca^{2+} uptake. CASQ2, a calcium-binding protein, can influence calcium ion level inside sarcoplasmic reticulum (SR) by interacting with RYR2 Ca^{2+} channel [183]. HERPUD1, also known as HERP, can activate the protective responses to ER stress by preventing overload of Ca^{2+} level inside ER, stabilize mitochondrial function and thus promotes the survival of cells [184]. Moreover, predicted ER proteins CATSPERB is also found in this cluster. Despite the lack of a reliable description of function, its sequence similarity to mouse CATSPERB suggests a role in fertilization [185]. Considering its appearance in this calcium ion trafficking cluster, the involvement of CATSPERB in calcium transportation is plausible. Similarly, the second cluster represents the ryanodine receptor channel (RyR), which is responsible for the release of Ca^{2+} from ER/SR lumen to cytosol. RyRs have three mammalian isoforms (RYR1, RYR2, RYR3), which are abundant in different tissues. RYR1 is the predominant isoform in skeletal muscle. RYR2 is usually abundant in the myocardium (heart muscle), and RYR3 expresses in a wide range of tissues such as smooth muscle and brain widespread [186, 187]. The interaction between FKBP1B (FKBP12.6) and RyRs has already been proved in previous studies, suggesting that the dissociation between RyR channel and FKBP may activate the release of calcium ions in ER/SR [188, 189]. Besides, another two ER proteins (ASPH and TRDN) are also included. ASPH (Junctate) is a membrane-bounded Ca^{2+} sensitive protein which has proved function in calcium ion binding [190, 191]. TRDN (Triadin) regulates Ca^{2+} release probably via interacting to CASQ2. Briefly, the RYR2 channel inhibitor CASQ2 may stabilize SR Ca^{2+} release by inhibiting RYR2 channel via its combination to TRDN [192]. The third cluster containing 10 proteins acting in folate metabolism. According to the GOBP, nine proteins are directly involved in the folic acid metabolic process while two others indirectly influence this pathway via regulating the biosynthesis of acetyl-CoA, which

can regulate 4'-phosphopantetheinylation, an important post-transcriptional modification for 10-formyltetrahydrofolate dehydrogenase [193].

4.1.3 Screen for ERcarta dependent regulation of mt-Ca^{2+}

To explore the roles of ERcarta genes in regulating Ca^{2+} signaling, a library of 989 siRNAs targeting individual ERcarta genes was used to transfect HeLa cells expressing a stable mt-targeted aequorin as calcium sensor. After 72 hours, mt-Ca^{2+} dynamics were recorded upon histamine stimulation. At the same time cell viability was also measured to account for the potential toxic effect of siRNA or gene knock-down. A number of 48 plates (two replicates $24 * 2$) were screened. Each plate includes siRNAs from negative control (nt), siMCU and siMCUb. The latter are used as positive controls for low mt-Ca^{2+} uptake or high mt-Ca^{2+}, respectively, compared to negative controls. An example plate is given in Fig. 4.2**A**. Cubic spline function and exponential decay models were fitted to smoothen Ca^{2+}-dependent signals inside mitochondria and ER, respectively. Viabilities, peaks, left slopes and decay rates were normalized to POC based on corresponding negative controls before comparison among plates.

The effect of siRNA on cell viability was firstly evaluated. Viability measurements replicate very well with only a few outliers (Fig. 4.2**B**). A cut-off of 0.75 for the viability screens was set to filter genes whose averaged viability was not greater than 0.75, and 42 genes were removed. Only three of excluded genes were found as essential, indicating the possibility of a compensatory mechanism in place. Both peaks (Fig. 4.2**B**) and left slope (Fig. 4.5) of remaining genes show good reproducibility, and due to their high correlation (Fig. 4.2**C**), averaged peak scores were chosen for hits selection for simplicity. Another 10 genes were considered as outliers according to the 99.99% confidence interval of two replications and thus removed. Finally, a number of 937 genes were kept for hits selection. Distribution of averaged peaks among four groups is shown in Fig. 4.2**D**. Peaks in the negative control group (blue box) range between 0.63 (lower whisker) and 1.35 (upper whisker). MCU functions as an enhancer in calcium uptake while MCUb is a calcium channel inhibitor. Expectedly, comparing to the negative control group (nt1), knockdown of MCU (siMCU, green box) and MCUb (siMCUb, red box) display decrease and increase in calcium uptake peaks, respectively. Promisingly, comparing to the peaks in the nt1 group, loss function of CATSPERB and HERPUD1, members of predicted Mito-ER Ca^{2+} trafficking cluster (Fig. 4.1**C**), leads to a peak increase in the mt-Ca^{2+} uptake (Table 4.1). Besides, knocking down of genes in predicted ER Ca^{2+} homeostasis also influence the peaks, for instance, loss function of TRDN decreases the peak while RYR3, FKBP1B and ASPH can increase the

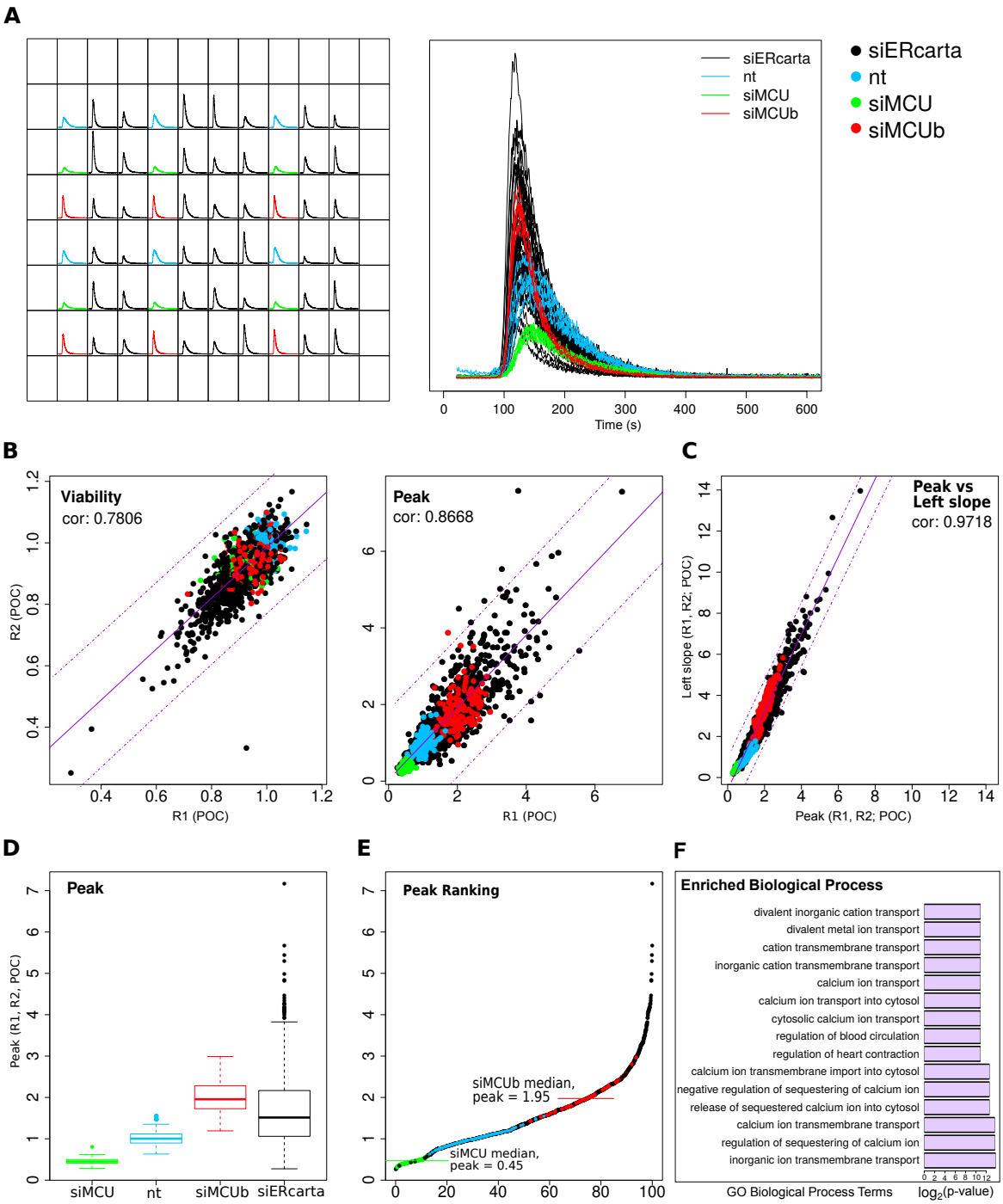


Fig. 4.2 **Mitochondrial Ca^{2+} screens. (A)** displays raw data of Ca^{2+}-dependent luminescence signal on a 96-well plate. Edge wells are designed as empty to avoid edge effect. None-empty wells contain cells, which are treated by specific siRNA libraries from four groups (nt: negative control, siRNAs sequence targeting no genes; siMCU: siRNAs targeting MCU; siMCUb: siRNAs targeting MCUb; siERcarta: siRNAs targeting ERcarta genes). For each well, Ca^{2+}-dependent signals are smoothened by cubic spline function to capture peak and left slope of calcium uptake. **(B)** Reproducibility of the calcium screens on cell viability and peak. Linear regression (solid violet lines) are fitted with normalized viability or peak of each siRNA (dot). Dots beyond the 99.99% confidence interval (dotted violet lines) are considered as outliers. **(C)** Correlation between peak and left slope. **(D)** Distribution of averaged peaks among four groups. **(E)** Ranking of siERcarta based on averaged peaks. siERcarta with peaks greater than the median of siMCUb group (1.95, indicated by the red line) are considered as potential inhibitors, and smaller than the median of siMCU group (0.45, green line) are potential enhancers. **(F)** Enriched biological processes for hits.

peaks (Table 4.1). The result displays a high correlation between these genes and calcium signaling, indicating high confidence in our detection of function modules.

Types	Genes	Mean of Peak
ER Ca^{2+} homeostasis	TRDN	0.33
Negative control (nt1)	lower whisker	0.63
Mito-ER Ca^{2+} trafficking	CASQ2	1.19
ER Ca^{2+} homeostasis	RYR2	1.19
Negative control (nt1)	upper whisker	1.35
ER Ca^{2+} homeostasis	RYR1	1.56
Mito-ER Ca^{2+} trafficking	CATSPERB	1.64
Mito-ER Ca^{2+} trafficking	HERPUD1	1.72
ER Ca^{2+} homeostasis	RYR3	2
ER Ca^{2+} homeostasis	FKBP1B	2.34
ER Ca^{2+} homeostasis	ASPH	3.47

Table 4.1 The peak of mt-Ca^{2+} after knocking down of nine genes.

Based on the rank of peaks (Fig. 4.2**E**), we identified 14 enhancers and 280 inhibitors by using peak cut-off 0.45 (median of siMCU group) and 1.95 (median of siMCUb group), respectively. However, the detailed mechanisms of how these hits influence the mt-Ca^{2+} uptake are still unknown. Served as a Ca^{2+} store in cells [194], abnormality of ER function such as ER stress might affect the calcium mobilization [195]. It is possible that knockdown of ER-resident genes might change the structure, amount or micro-environment of ER, which initializes ER stress, and thus sequentially induce the release of Ca^{2+} from ER lumen first to the cytosol and then mitochondria [196], or directly to mitochondrial matrix through ER-mitochondria contact sites [195]. Apart from the indirect change of mitochondrial calcium uptake by influencing ER function, *e.g.,* ER stress, our hits might include potential members of calcium transfer channels, which are directly involved in the process of Ca^{2+} shuttling. Expectedly, analysis of enriched biological processes on hits shows several calcium-related terms (Fig. 4.2**F**), *e.g.,* calcium ion transmembrane transport, release of sequestered calcium ion into cytosol and negative regulation of sequestering of calcium ion, *etc.*

Further secondary screens will be necessary to prioritize interesting hits for follow up studies. Those will be done in collaboration with Dr. Marta Giacomello from university of Padova and include screening of the effect of the 294 hits on cytosolic calcium, mitochondrial membrane potential, Mito-ER contacts formation, ER calcium, mitochondrial mass and mitochondrial bioenergetics.

4.2 Conclusion

In cells, ER is physically and functionally connected to mitochondria. These crosstalks allow the synergistic functioning of the two organelles. Our systematic prediction of functionally-related sub-modules not only provides functional context for newly predicted ERRPs (identified in chapter 3), but also highlights candidates in Mito-ER contacts. Moreover, the mt-Ca^{2+} screen highlights potential mechanisms for ER-dependent regulation of mitochondrial Ca^{2+} uptake.

4.3 Methods

4.3.1 Data collection

Human mitochondrial genes were downloaded from MitoCarta2 [88] and ER-resident genes were from ERcarta, which was predicted in chapter 3. Links of protein-protein interaction (PPI) were downloaded from STRING (version 10.5) [177]. Phylogenetic profiles of genes obtained by five orthology detection methods (*i.e.,* One-way Best-Hits (OBH), Best-Reciprocal-Hits (BRH) [197], OrthoMCL [198] and eggNOG version 4 [199]), were extracted from ProtPhylo [156]. The attribute 'external_gene_name' of R package biomaRt [138, 182] was used to assign consistent gene symbols to genes from MitoCarta2, ERcarta, STRING and ProtPhylo. Human protein complexes were downloaded from the CORUM database (release 3.0) [165]. Pathways used in the evaluation were extracted from KEGG PATHWAY Database (release 87.1) [30], excluded were three large pathways (*i.e.,* 'Environmental Information Processing', 'Organismal Systems' and 'Human Diseases') and genes involved in more than three pathways.

4.3.2 Functional clusters prediction

Four different methods, *i.e.,* hierarchical clustering (HC), dynamic hierarchical clustering (DHC) [200], ClusterONE [201] and MCL[202], were applied to detect functional clusters. HC and DHC were implemented with hclust function using five linkage algorithms, namely, ward.D linkage, ward.D2 linkage, average linkage, single linkage and complete linkage. Dendrograms generated by hclust were cutting into sub-groups by using various heights in HC and by using different combination of 'deepSplit' and 'minClusterSize' in DHC, respectively. With DHC method, big clusters were cut recursively until the number of genes in all predicted clusters was smaller than the given 'maxClusterSize'. ClusterONE1.1 was

applied with setting cluster size to 3 and density ranging from 0.3 to 0.9. MCL was used by setting inflation ranging from 1.2 to 10 and edge weights from 0.3 to 0.9.

4.3.3 Selection of predicted clusters

To compare the performance of predicted clusters, which were predicted by using different PPI links and clustering methods, and thus select the best prediction, indexes (RI, ARI and F1 score) were calculated with R package clusterCrit and flexclust by using KEGG pathways and CORUM complexes as referenced clusters. Visualization of performance was performed using R package ggplot2. Definitions of indexes are shown below.

Rand index (RI, shown in Eq. (4.1)) [203] seeks to calculate the percentage of correctly clustered pairs among all possible pairs, adjusted rand index (ARI) is a corrected version of rand index [204], which considers the chance grouping. F1 score (F_1, shown in Eq. (4.2)) is another popular index based on precision (also known as positive predictive value, PPV) and recall (also known as sensitivity or true positive rate, TPR).

$$RI = \frac{TP+TN}{TP+TN+FP+FN} \tag{4.1}$$

$$\begin{aligned} PPV &= \frac{TP}{TP+FP} \\ TPR &= \frac{TP}{TP+FN} \\ F_1 &= \frac{2.PPV.TPR}{PPV+TPR} \end{aligned} \tag{4.2}$$

where TP, TN, FP and FN were defined based on a pair of genes

TP: if they are in the same cluster in both predicted and referenced clusters

TN: if they are in different clusters in both predicted and referenced clusters

FP: if they are in the same cluster by prediction but in different clusters according to the reference

FN: if they are in different clusters by prediction but in the same cluster according to the reference

4.3.4 Calcium screening analysis

Mitochondrial calcium screens were performed using the strategy described previously [205]. Briefly, for each plate, dynamics of Ca^{2+}-dependent luminescence in the mitochondria were

smoothened with cubic spline function to extract peaks and left slopes of mitochondrial calcium kinetics. Fitted peaks and left slopes were then normalized to percentage of control (POC), which is defined in Eq. (4.3).

$$POC = \frac{x_i}{\bar{c}} \tag{4.3}$$

where x_i is the fitted peak or left slope for i^{th} sample and $\bar{c}$ is the averaged peak or left slope of samples of the negative control group on the same plate.

4.3.5 Other analysis

Enrichment of biological process terms for selected hits was performed using GOrilla webserver [206]. Venn diagrams were generated with the website (http://bioinformatics.psb.ugent.be/webtools/Venn/)

4.4 Appendix figures and tables

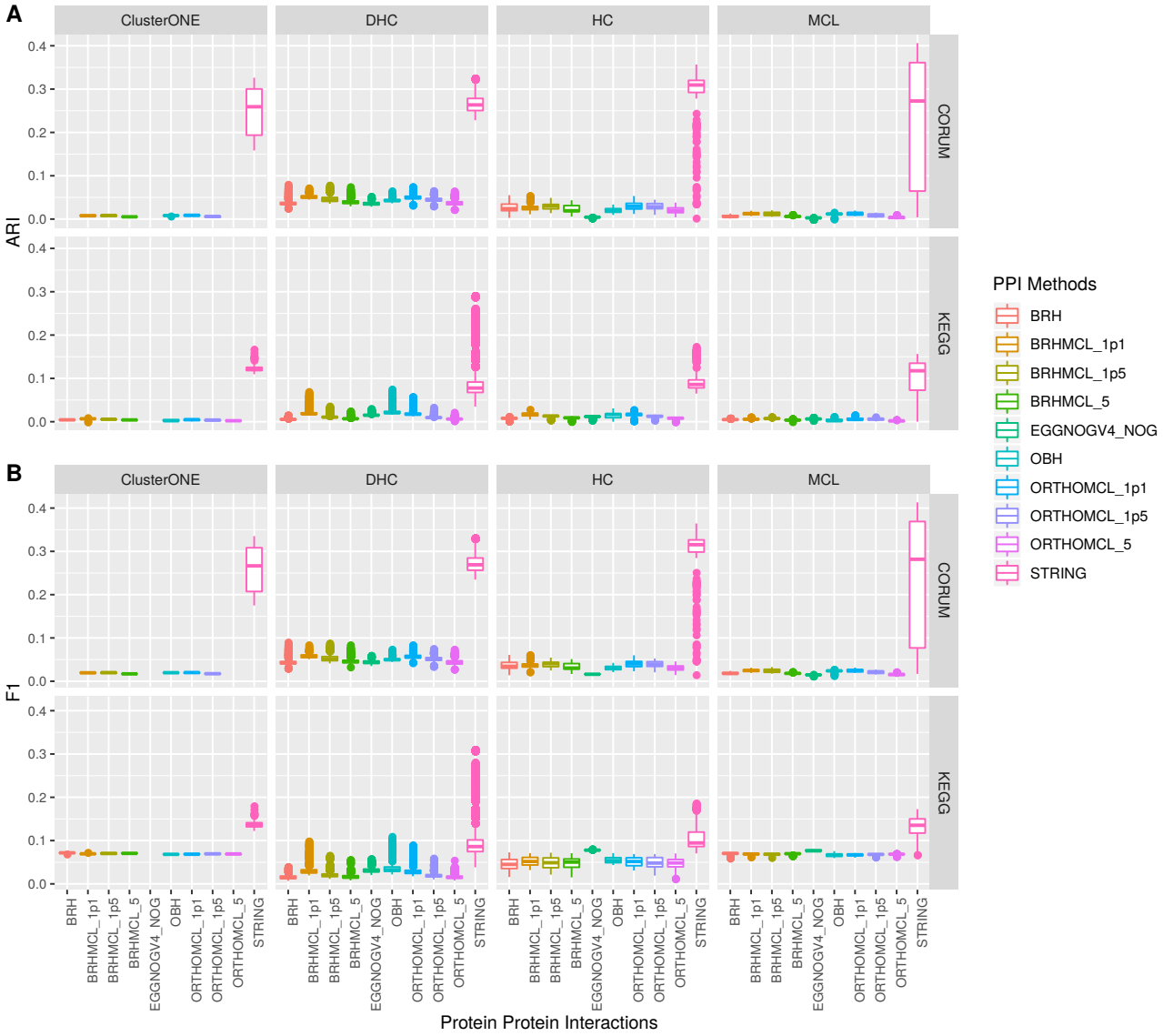


Fig. 4.3 **Performance of predicted clusters in selecting the type of protein links.** KEGG pathways and CORUM complexes are used as a reference to calculate ARI and F1 score. STRING-based clusters yield better performance than those predicted based on pure phylogenetic links.

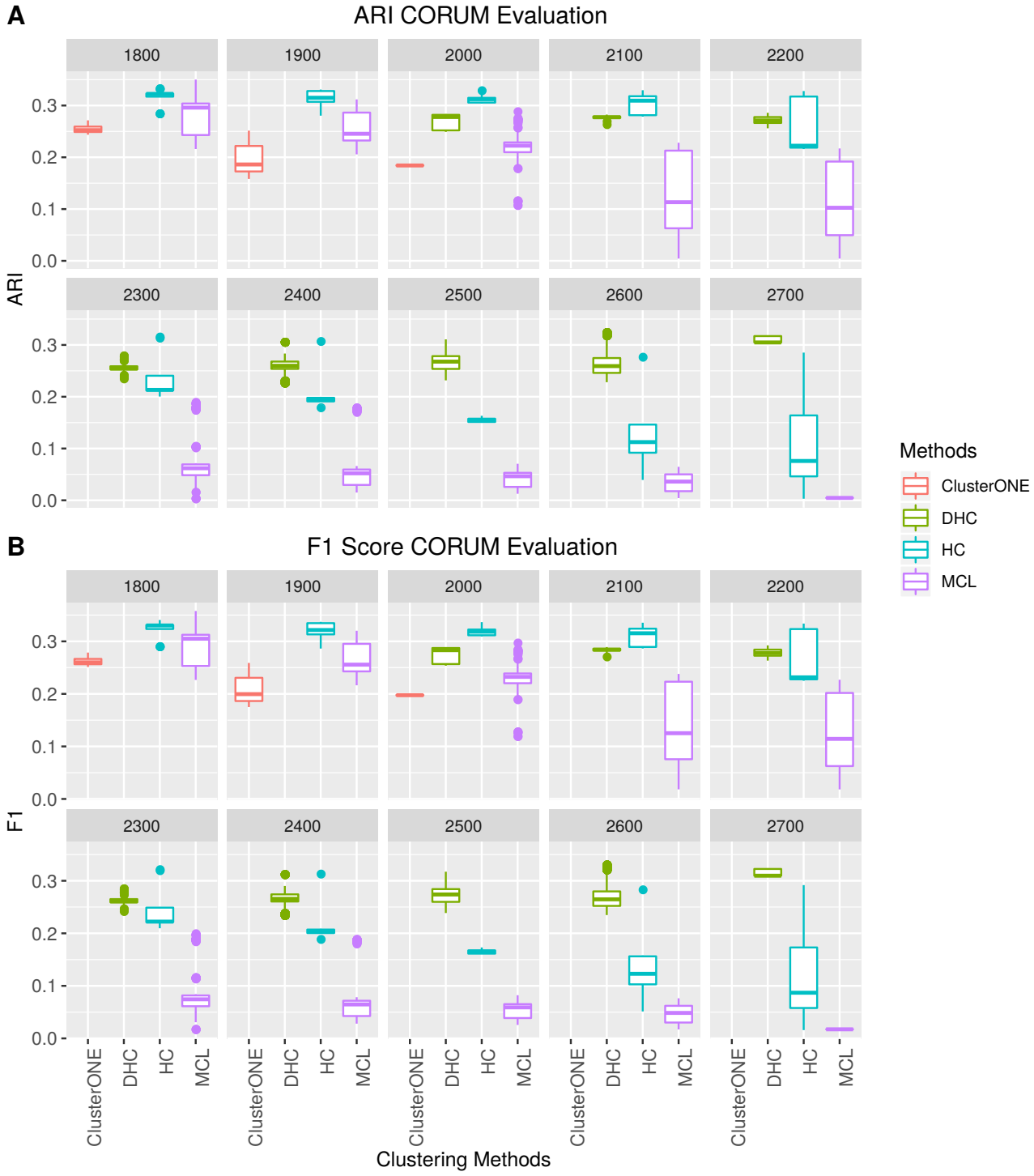


Fig. 4.4 **Performance of predicted clusters in selecting clustering approaches.** CORUM complexes are used as the reference to calculate ARI and F1 score.

Table 4.2 Full list of predicted MitoER clusters

cluster1_Mito_size48: FOXRED1;MT-ND1;MT-ND2;MT-ND3;MT-ND4;MT-ND4L;MT-ND5;MT-ND6;NDUFA1;NDUFA10;NDUFA11;NDUFA12;NDUFA13;NDUFA2;NDUFA3;NDUFA4;NDUFA5;NDUFA6;NDUFA7;NDUFA8;NDUFA9;NDUFAB1;NDUFAF2;NDUFB1;NDUFB10;NDUFB11;NDUFB2;NDUFB3;NDUFB4;NDUFB5;NDUFB6;NDUFB7;NDUFB8;NDUFB9;NDUFC1;NDUFC2;NDUFS1;NDUFS2;NDUFS3;NDUFS4;NDUFS5;NDUFS6;NDUFS7;NDUFS8;NDUFV1;NDUFV2;NDUFV3;NUBPL
cluster2_Mito_size13: COX4I1;COX5A;COX5B;COX6A1;COX6B1;COX6C;COX7A2L;COX7B;COX7C;COX8A;MT-CO1;MT-CO2;MT-CO3
cluster3_MitoER;Mito;ER_size6: ATXN2;ERP44;PRDX3;PRDX4;TXN2;TXNRD2
cluster4_Mito;ER_size6: CDIPT;CDS1;CDS2;CRLS1;LPGAT1;PGS1
cluster5_Mito;ER_size6: ACAA2;AUH;HMGCL;HMGCR;HMGCS1;HMGCS2
cluster6_Mito_size6: BCAT2;LDHAL6B;ME1;ME3;PDHA2;SDSL
cluster7_ER_size6: ASPH;FKBP1B;RYR1;RYR2;RYR3;TRDN
cluster8_Mito_size6: CCDC90B;HIGD1A;MCUR1;METTL5;MMACHC;MMADHC
cluster9_Mito_size6: ATP23;ATPIF1;DNA2;MLH1;PIF1;TOP3A
cluster10_ER_size6: ATF6B;CREB3;CREB3L1;CREB3L3;MALL;WFS1
cluster11_ER_size6: AKR1C3;CYP19A1;HSD17B2;SRD5A1;UGT2B15;UGT2B17
cluster12_Mito;ER_size6: ICMT;NLN;OMA1;RCE1;ROMO1;ZMPSTE24
cluster13_ER_size13: ALG10;ALG10B;DAD1;DDOST;KRTCAP2;MAGT1;MOGS;OSTC;RPN1;RPN2;STT3A;STT3B;TUSC3
cluster14_ER_size6: CYP2C19;CYP2J2;CYP2U1;CYP4F2;CYP4F3;SLC27A6
cluster15_ER_size6: INPP5K;MARCH2;MARCH3;MPPE1;SACM1L;TPTE2
cluster16_Mito;ER_size6: LAP3;P4HA1;P4HA2;P4HA3;PRODH;PYCR1
cluster17_Mito;ER_size6: FABP1;FADS1;PLIN2;SLC27A1;SLC27A4;UBXN4
cluster18_Mito;ER_size6: LPCAT1;LPCAT2;LPCAT4;LYPLA1;PNPLA6;PNPLA7
cluster19_Mito;ER_size6: GAA;GANC;HK1;HK2;MGAM;SI
cluster20_MitoER;ER_size6: ATP2A1;ATP2A2;ATP2A3;CASQ1;DMPK;PLN
cluster21_Mito;ER_size6: CMC2;COX7A2;DBI;HINT1;SLIRP;TMEM258
cluster22_ER_size6: EDEM1;EDEM2;EDEM3;FOXRED2;TXNDC11;TXNDC16
cluster23_MitoER;Mito_size6: MECR;MMAB;RDH13;RDH14;RTN4IP1;ZADH2
cluster24_Mito_size13: AADAT;ACCS;ACO1;ACO2;IDH1;IDH2;IDH3A;IDH3B;IDH3G;NIT2;OGDHL;PARK7;TACO1
cluster25_Mito;ER_size6: CST3;FRAS1;GRIP1;MIPEP;PDZD2;PICK1
cluster26_Mito_size6: C6orf203;NRDC;OPA3;RCC1L;STARD7;TEFM
cluster27_Mito_size6: ATPAF1;ATPAF2;FMC1;NUDT19;OXLD1;SLC25A51
cluster28_MitoER;Mito_size6: C15orf48;LETM1;RDH11;SLC25A1;TAMM41;TMEM70
cluster29_Mito;ER_size6: OXR1;TMBIM4;TMED5;TMEM123;TMX1;TMX2
cluster30_Mito;ER_size6: BNIP3;BNIP3L;CLN8;SELENOK;STEAP3;TMEM11
cluster31_Mito_size6: MACROD1;MTPAP;SIRT3;SIRT4;SIRT5;TOP1MT
cluster32_Mito;ER_size6: AL049779.1;DHRS13;DHRSX;HSDL2;HTATIP2;PCBD2
cluster33_Mito_size6: DNLZ;SLC25A18;SLC25A22;SLC25A23;SLC25A25;SLC25A41
cluster34_Mito;ER_size6: AFG1L;DDAH1;ETFRF1;NT5DC2;RETREG1;XPNPEP3
cluster35_Mito_size13: DNAJC19;GRPEL2;HSCB;HSPA9;PAM16;PMPCA;PMPCB;TIMM17A;TIMM17B;TIMM21;TIMM23;TIMM44;TIMM50
cluster36_Mito;ER_size6: ATL1;MPV17L;PXMP2;RTN1;RTN3;SPAST
cluster37_Mito;ER_size6: CLCC1;CLIC4;RAB24;RNF128;TMEM189;WLS
cluster38_Mito;ER_size6: AKAP10;KCNRG;MARCH1;SCAPER;SUCO;TMCC1
cluster39_Mito;ER_size6: ANGEL2;ANKRD13C;CHERP;FAM185A;MTFR1;PPWD1
cluster40_Mito_size5: AFG3L2;ATAD3A;OXA1L;PHB;PHB2
cluster41_Mito;ER_size5: G6PC;MCEE;MUT;PCCA;PCCB
cluster42_ER_size5: CANX;GANAB;MLEC;PDIA3;PRKCSH
cluster43_Mito_size5: MINOS1;MTX1;MTX2;RMDN1;SAMM50
cluster44_Mito_size5: FDX1;FDX2;FDXR;MGARP;STAR
cluster45_ER_size5: ORAI1;ORAI2;ORAI3;STIM1;STIM2
cluster46_ER_size13: CYP2A7;CYP2W1;CYP3A43;CYP4B1;CYP4F11;CYP4F22;CYP4V2;CYP4X1;CYP4Z1;CYP51A1;LSS;SDR42E1;SQLE
cluster47_Mito_size5: CLPP;CLPX;HTRA2;LONP1;OPA1
cluster48_Mito_size5: SLC25A31;TSPO;VDAC1;VDAC2;VDAC3
cluster49_Mito;ER_size5: CDC25C;DUSP26;PKMYT1;PTPN5;PTPRR
cluster50_ER_size5: DERL3;ERLIN1;ERLIN2;RNF185;RNF5
cluster51_Mito_size5: AGMAT;ARG2;CPS1;OAT;OTC
cluster52_ER_size14: ALG12;ALG6;ALG8;ALG9;PIGZ;PIGV;PIGX;DPM1;DPM2;DPM3;ALG3;PIGB;PIGM;PIGN
cluster53_Mito;ER_size5: MRM1;NAXE;NUDT8;RTN2;NAXD
cluster54_Mito_size5: ECHDC1;ACSS1;ACSS3;ALDH2;ALDH9A1
cluster55_Mito_size5: GLRX5;ABHD11;BOLA1;BOLA3;DHRS4
cluster56_Mito_size5: GLYCTK;GRHPR;HAO2;AGXT;HOGA1
cluster57_ER_size13: CNIH1;CNIH2;CNIH3;LMAN2;LMAN2L;PORCN;SAR1B;SEC16A;SEC22A;SEC22C;SEC23IP;SHISA7;TRAPPC2L
cluster58_Mito;ER_size5: THEM4;HSD17B10;RRBP1;SCP2;TRMT10C
cluster59_ER_size9: PIGG;ELOVL3;ELOVL6;HSD17B12;ELOVL1;ELOVL4;ELOVL7;ELOVL5;ELOVL2
cluster60_ER_size5: INSIG2;SCAP;INSIG1;SREBF1;SREBF2
cluster61_Mito;ER_size5: MPC2;MPC1;VMP1;REEP1;REEP3
cluster62_Mito;ER_size5: BCAP31;TMBIM6;RER1;BFAR;CASP8
cluster63_Mito;ER_size5: TMEM173;MAVS;NLRX1;RNF26;FASTKD5
cluster64_ER_size5: MOGAT1;MOGAT2;DGAT1;DGAT2;MOGAT3
cluster65_Mito;ER_size5: POR;CYP24A1;CYP27B1;CYP2R1;CYP4F12
cluster66_ER_size5: ALG5;AWAT1;DOLK;DOLPP1;DPAGT1
cluster67_Mito;ER_size13: DNAJA3;DNAJB14;DNAJB2;DNAJB9;DNAJC1;DNAJC11;DNAJC15;DNAJC25;HINT2;HINT3;HSPA13;TMTC3;YBEY
cluster68_Mito_size5: ACADSB;ACADS;ACOX3;ACSM1;GCDH
cluster69_ER_size5: HSD17B6;CYP1B1;HSD17B1;UGT1A4;UGT1A6
cluster70_MitoER;Mito_size5: RHOT1;FKBP8;IDE;MUL1;PTRH2
cluster71_Mito;ER_size5: BPHL;TTC19;CD320;LAMC1;PAK5
cluster72_MitoER;Mito;ER_size5: CYP2F1;CYP2A13;EPHX1;MGST1;GSTK1
cluster73_Mito;ER_size5: SCD;SCD5;RNFT1;ACOT2;ACOT7
cluster74_Mito_size5: UCP3;SLC25A14;SLC25A27;UCP1;UCP2
cluster75_Mito;ER_size5: AC135050.2;GGCX;PGAP3;TBRG4;VKORC1L1
cluster76_ER_size5: ENTPD6;CANT1;ENTPD5;PDE3A;PDE9A
cluster77_ER_size5: MAN1B1;AGPAT2;AGPAT3;MBOAT4;PLPPR4
cluster78_Mito;ER_size12: ARF5;KDELR1;KDELR2;KDELR3;KIF1B;RAB1A;SURF4;TMED10;TMED2;TMED3;TMED7;TMED9
cluster79_ER_size5: ERO1A;PDIA4;ERO1B;ERP29;GPX7
cluster80_Mito_size5: AASS;MAOB;ACYP2;ALDH7A1;MAOA
cluster81_Mito;ER_size5: PDIA5;GCAT;PDILT;PRDX2;PRDX5
cluster82_Mito;ER_size5: OXNAD1;TMX3;CYB5R1;CYB5R2;DHODH
cluster83_Mito;ER_size5: YARS2;DDRGK1;LYRM7;SPATA20;TRUB2
cluster84_Mito_size5: CARS2;IARS2;LARS2;PARS2;VARS2
cluster85_Mito_size5: PMAIP1;ERAL1;NIF3L1;SFXN4;SLC25A19
cluster86_MitoER;Mito;ER_size5: HIGD2A;COA7;FKBP10;P3H4;SDHAF2
cluster87_ER_size5: ARL6IP5;REEP6;LNPK;PRAF2;REEP5
cluster88_ER_size5: ATL2;RNF13;PGRMC1;ATL3;GLYATL2
cluster89_ER_size12: AGMO;CYP39A1;CYP8B1;ERG28;FAXDC2;HSD17B7;HSD3B7;MSMO1;NSDHL;SDR42E2;SIGMAR1;TM7SF2
cluster90_Mito;ER_size5: LTK;FAM213A;FAM162A;PTPN1;RMDN3
cluster91_ER_size5: YIF1A;YIPF7;IER3IP1;RABAC1;VKORC1
cluster92_Mito;ER_size5: CKMT2;QTRT1;HHATL;HRC;HSPB7
cluster93_Mito;ER_size5: FKBP11;FKBP14;FKBP9;PI4KA;RCN3
cluster94_Mito;ER_size5: RAB10;FKBP2;GLOD4;RAB32;TBC1D20
cluster95_Mito_size9: COX16;COX17;CMC1;COA5;PET100;CHCHD7;CHCHD4;COX19;GFER
cluster96_Mito_size5: SLC25A36;SUPV3L1;C15orf40;LRPPRC;METTL17
cluster97_Mito;ER_size5: MARCH8;SLC25A37;TMEM56;FAM210B;RHBDD1
cluster98_Mito;ER_size5: REEP2;TM4SF20;P4HTM;JPH3;TMEM186
cluster99_Mito_size4: LYRM4;NFS1;ISCU;FXN
cluster100_Mito;ER_size12: ARSA;ARSB;ARSD;ARSF;ARSI;ARSJ;MESD;NEU4;STS;SUMF1;SUMF2;TMLHE
cluster101_ER_size4: STX18;RINT1;ZW10;USE1
cluster102_Mito_size4: PRELID3B;PRELID3A;PRELID1;TRIAP1
cluster103_Mito_size4: ABCD2;ABCD1;SLC25A17;ABCD3
cluster104_ER_size4: ESYT3;ESYT1;ESYT2;YIF1B
cluster105_Mito;ER_size4: SOD2;SOD1;CAT;GPX8
cluster106_Mito_size4: TMEM177;NADK2;NMNAT3;NNT
cluster107_Mito;ER_size4: PXMP4;CIDEC;LIPE;PLIN1
cluster108_Mito_size4: GLS;ALDH18A1;GLUD1;L2HGDH
cluster109_Mito_size4: PNPT1;DGUOK;GUK1;NME3
cluster110_Mito_size31: AURKAIP1;CHCHD1;DAP3;MRPS10;MRPS11;MRPS12;MRPS15;MRPS16;MRPS17;MRPS18B;MRPS18C;MRPS2;MRPS21;MRPS22;MRPS23;MRPS24;MRPS25;MRPS26;MRPS27;MRPS28;MRPS31;MRPS33;MRPS34;MRPS35;MRPS5;MRPS6;MRPS7;MRPS9;PTCD3;TSFM;TUFM
cluster111_Mito;ER_size12: ACAD10;ACAD11;ACADL;ACOX1;ACSL1;ACSL4;ACSL5;ACSL6;ACSM3;CPT1A;CPT1B;FADS2
cluster112_Mito_size4: LDHB;HCCS;ME2;PC
cluster113_ER_size4: CYP2E1;CYP2B6;CYP2C8;UGT2B4
cluster114_Mito_size4: CPT1C;CPT2;CRAT;SLC25A20
cluster115_ER_size4: CES3;CES1;CES2;CES5A
cluster116_MitoER;ER_size4: PCYT1B;CEPT1;CHPT1;PAFAH2
cluster117_Mito;ER_size4: AGPAT1;GNPAT;GPD2;PLPP3
cluster118_ER_size4: SDF2L1;POMT1;POMT2;SDF2
cluster119_ER_size4: EXT1;EXT2;SLC35D1;UGT3A1
cluster120_ER_size4: UGT2A3;CYP26A1;CYP2C18;CYP3A5
cluster121_MitoER;Mito;ER_size12: AHCYL1;AHCYL2;COMTD1;CPXM2;GNG5;HEBP1;ITPR1;ITPR2;ITPR3;LRMP;MRVI1;PPP1R15A
cluster122_Mito;ER_size4: ALG11;ABCC6;ALG2;METTL8
cluster123_Mito_size4: CYCS;PAICS;SLC25A12;SLC25A13
cluster124_Mito_size4: GOT2;GPT2;MPST;NAGS
cluster125_Mito_size4: SUOX;ETHE1;SQOR;TST
cluster126_Mito;ER_size4: CEP89;TMEM231;TMEM67;TRPV1
cluster127_ER_size4: DIO1;DIO2;UGT1A8;UGT2B11
cluster128_Mito;ER_size4: TYSND1;MTDH;SND1;TMEM214
cluster129_ER_size4: AWAT2;RDH10;RDH12;RDH5
cluster130_Mito_size4: ABCB6;ABCB8;ABCF2;RAB11FIP5
cluster131_Mito;ER_size12: ERN1;ERN2;FGFR3;LRRC8C;PDP2;PPM1K;PTPN2;PTPN4;PTPRE;SLMAP;TMEM184A;TMEM184B
cluster132_ER_size4: NOMO2;SELENOF;UGGT1;UGGT2
cluster133_Mito;ER_size4: SFXN2;SLC25A3;SPG7;TMCO1
cluster134_Mito_size4: C1QBP;MAIP1;OCIAD1;STOML2
cluster135_Mito_size4: ACP6;SDHAF1;SDHAF3;SDHAF4
cluster136_Mito;ER_size4: FAM210A;REEP4;TMEM33;USP19
cluster137_Mito_size4: IMMP1L;IMMP2L;PLPBP;RSAD1
cluster138_Mito_size4: D2HGDH;HAGH;LACTB2;LDHD
cluster139_ER_size4: JSRP1;SRL;TMEM38A;TMEM38B
cluster140_Mito;ER_size4: CALU;NUCB2;PSTK;SARS2
cluster141_ER_size4: CLPTM1L;NCLN;NOMO1;NOMO3
cluster142_Mito_size11: ACAA1;ACADM;ACADVL;ACAT1;ECHS1;EHHADH;HADH;HADHA;HADHB;HIBCH;HSD17B4

cluster143_Mito_size4: AIFM2;SFXN5;SLC25A34;SLC25A35
cluster144_Mito;ER_size4: CDKAL1;DNAJB12;DNAJC4;FHIT
cluster145_MitoER;Mito;ER_size4: ABCB10;ABCB9;AL669918.1;SLC35G1
cluster146_Mito_size4: MARS2;MTFMT;PDF;PUS1
cluster147_Mito_size4: ACOT9;C12orf10;ELAC2;EXOG
cluster148_Mito;ER_size4: CCDC51;HAX1;PIEZO2;PKD2
cluster149_MitoER;Mito;ER_size7: MTERF1;PNPO;SLC30A9;UNG;TMED6;SLC30A6;SLC39A7
cluster150_Mito_size4: MARC1;COX18;MARC2;SLC25A46
cluster151_Mito;ER_size4: DHRS7C;HSD11B1L;DHRS7;HSDL1
cluster152_MitoER;ER_size4: HHAT;ARL6IP1;GOLPH3;PGAP2
cluster153_Mito;ER_size11:
EMC1;EMC10;EMC2;EMC3;EMC4;EMC6;EMC7;MMGT1;OCIAD2;TMEM43;ZFAND2B
cluster154_ER_size4: ATP13A1;AL136295.3;ASNA1;WRB
cluster155_Mito_size7: COA4;COX14;UQCC2;COX20;UQCC1;COA3;COA1
cluster156_Mito;ER_size4: PITPNM1;SEPT4;CCDC88B;TUBB3
cluster157_Mito_size3: ETFB;ETFDH;ETFA
cluster158_Mito_size3: TIMM8B;TIMM10B;TIMM13
cluster159_MitoER;Mito;ER_size3: STX3;STX17;SNAP29
cluster160_Mito;ER_size3: PGAM1;TPI1;GAPDH
cluster161_Mito_size3: MCCC2;IVD;MCCC1
cluster162_MitoER;Mito_size3: CBR4;CRYZ;HSD17B8
cluster163_ER_size3: HSP90B1;ATF6;CALR
cluster164_ER_size11:
SEC11A;SEC11C;SEC61A2;SPCS1;SPCS2;SPCS3;SSR1;SSR2;SSR3;SSR4;TRAM1
cluster165_Mito_size3: RARS2;RARS;KARS
cluster166_Mito_size3: LIAS;LIPT1;LIPT2
cluster167_Mito;ER_size3: RFK;TYW1B;FLAD1
cluster168_Mito_size3: SLC25A4;PPIF;SLC25A5
cluster169_MitoER;ER_size3: SC5D;DHCR24;DHCR7
cluster170_Mito_size3: ALDH4A1;PRODH2;ACSM2A
cluster171_Mito_size3: TFB2M;POLRMT;TFAM
cluster172_Mito_size3: GTPBP3;MTO1;TRMU
cluster173_MitoER;Mito;ER_size3: CYP4A22;ACSF3;ALDH3A2
cluster174_Mito_size3: QDPR;SPR;PTS
cluster175_Mito;ER_size11: COQ2;DHDDS;EBP;EBPL;FDFT1;FDPS;IDI1;NUS1;PDSS1;PDSS2;VWA8
cluster176_Mito_size3: GLS2;KMO;KYAT3
cluster177_ER_size3: ANKLE2;MAL2;VRK1
cluster178_ER_size3: KCNQ1;KCNE2;KCNE1
cluster179_ER_size3: LMF2;LMF1;CLPTM1
cluster180_ER_size3: UGT2B10;UGT2B7;HSD11B1
cluster181_Mito_size3: ECH1;ECI1;ECI2
cluster182_MitoER;Mito;ER_size3: AGK;AGPAT5;GPAT2
cluster183_ER_size3: UGT1A10;SRD5A3;HSD17B3
cluster184_Mito;ER_size3: SLC27A2;PHYH;AMACR
cluster185_Mito_size3: EEFSEC;MRPS14;SECISBP2
cluster186_MitoER;Mito;ER_size10: AIFM1;BAD;BAK1;BAX;BCL2;BCL2L1;BCL2L2;BID;BIK;DIABLO
cluster187_ER_size3: SAMD8;SMPD4;CERS1
cluster188_Mito;ER_size3: SDR16C5;BDH1;OXCT1
cluster189_Mito;ER_size3: G6PC3;DCXR;AKR1B10
cluster190_Mito_size3: ALAS2;GATM;ALAS1
cluster191_ER_size3: LRAT;DHRS3;DHRS9
cluster192_ER_size3: MAN1A1;MAN1A2;MAN1C1
cluster193_Mito;ER_size3: RHOT2;OBSCN;RASGRF2
cluster194_ER_size3: EXTL3;EXTL2;EXTL1
cluster195_MitoER;Mito_size3: MSRA;MSRB3;SELENOO
cluster196_Mito;ER_size3: DNAJB8;DNAJC16;CLPB
cluster197_Mito_size10: ATIC;FPGS;MTHFD1;MTHFD1L;MTHFD2;MTHFD2L;PDP1;PDPR;SHMT1;SHMT2
cluster198_Mito_size3: MPV17L2;MPV17;SLC25A24
cluster199_MitoER;ER_size3: SLC19A3;SERAC1;SLC19A2
cluster200_MitoER;Mito_size3: SFXN1;TMEM14C;RCN2
cluster201_Mito;ER_size3: PRDX6;ERP27;PITRM1
cluster202_Mito_size3: ACOT13;TK2;TRNT1
cluster203_Mito;ER_size3: RPUSD4;IKBIP;MRM3
cluster204_Mito;ER_size3: SLC25A10;TMEM65;RTN4
cluster205_Mito_size3: ACAD8;ACSF2;ECHDC2
cluster206_Mito_size3: C12orf65;COASY;DCAKD
cluster207_ER_size3: SLC33A1;PDIA2;CLGN
cluster208_ER_size10: AMFR;DERL1;DERL2;FAF2;HM13;RNF139;SELENOS;SVIP;TMEM129;UBAC2
cluster209_Mito_size3: PTPMT1;MTCH2;C20orf24
cluster210_ER_size3: MGAM2;MYORG;GNPTG
cluster211_Mito_size3: GSTZ1;FAHD1;FAHD2A
cluster212_ER_size3: PPIC;TMED7-TICAM2;TMED4
cluster213_ER_size3: CTSE;DNAJC14;CPVL
cluster214_Mito;ER_size3: HSD17B11;ISOC2;DECR1
cluster215_Mito_size3: ISCA1;ISCA2;NFU1
cluster216_Mito_size3: DGLUCY;HDHD5;MSRB2
cluster217_Mito_size3: SLC25A33;PACSIN2;THG1L
cluster218_MitoER;Mito_size3: SUGCT;SLC16A11;CLYBL
cluster219_Mito;ER_size26:
ATP5A1;ATP5B;ATP5C1;ATP5D;ATP5E;ATP5F1;ATP5G1;ATP5G2;ATP5G3;ATP5H;ATP5I;ATP5J;ATP5J2;A
TP5J2-PTCD1;ATP5L;ATP5O;ATP5S;ATP6AP1;C14orf2;EIF2AK3;MT-ATP6;MT-
ATP8;PPA2;PTCD1;TCIRG1;USMG5
cluster220_Mito_size10:
ACAD9;ECSIT;NDUFAF1;NDUFAF3;NDUFAF4;NDUFAF5;NDUFAF6;NDUFAF7;TIMMDC1;TMEM126B
cluster221_ER_size3: KSR1;SLC27A3;VRK2
cluster222_ER_size3: EPM2AIP1;STBD1;EPM2A
cluster223_ER_size3: OSBPL8;COL4A3BP;OSBPL5
cluster224_Mito_size3: FASTKD2;FASTKD3;NGRN
cluster225_Mito;ER_size3: ATP10D;ANO5;FKRP
cluster226_ER_size3: FA2H;CYB5R4;CYP20A1
cluster227_ER_size3: SPPL3;TRAM1L1;TRAM2
cluster228_Mito;ER_size3: PDK4;PTPRN2;TLR9
cluster229_ER_size10:
CYP1A2;CYP26B1;CYP2A6;CYP2C9;CYP3A4;CYP3A7;CYP4A11;UGT1A1;UGT1A9;UGT2A2
cluster230_ER_size3: MYRF;TFPI;TMPRSS3
cluster231_ER_size3: MTTP;TM6SF2;XDH
cluster232_Mito_size3: LETM2;MRS2;SLC25A28
cluster233_ER_size3: ATP8B3;JAGN1;ZDHHC21
cluster234_ER_size3: FITM1;FITM2;ZDHHC6
cluster235_Mito;ER_size3: GLYAT;TMEM174;TMEM72
cluster236_Mito;ER_size3: AKR7A2;ARL2;BSCL2
cluster237_Mito_size3: NIPSNAP3A;SLC25A32;SLC25A44
cluster238_ER_size3: CNPY3;SELENOM;SELENON
cluster239_Mito;ER_size3: MPDU1;SLC25A11;TMEM143
cluster240_Mito;ER_size10:
OSBP2;OSBPL1A;OSBPL3;OSBPL6;OSBPL7;PPM1L;TOMM34;VAPA;VAPB;ZFYVE27
cluster241_Mito_size3: PANK2;SLC22A4;SLC25A16
cluster242_Mito_size60:
GADD45GIP1;GFM1;GFM2;HEMK1;MRPL1;MRPL10;MRPL11;MRPL12;MRPL13;MRPL14;MRPL15;MRPL1
6;MRPL17;MRPL18;MRPL19;MRPL2;MRPL20;MRPL21;MRPL22;MRPL23;MRPL24;MRPL27;MRPL28;MRP
L3;MRPL30;MRPL32;MRPL33;MRPL34;MRPL35;MRPL36;MRPL37;MRPL38;MRPL39;MRPL4;MRPL40;MR
PL41;MRPL42;MRPL43;MRPL44;MRPL45;MRPL46;MRPL47;MRPL48;MRPL49;MRPL50;MRPL51;MRPL52
;MRPL53;MRPL54;MRPL55;MRPL57;MRPL58;MRPL9;MRPS18A;MRPS30;MRPS36;MRRF;MTIF2;MTIF3;
MTRF1L
cluster243_Mito;ER_size10: ARV1;PIGA;PIGC;PIGH;PIGL;PIGP;PIGQ;PIGW;PIGY;PYURF
cluster244_Mito;ER_size10: CHDH;MBOAT1;MBOAT2;PEMT;PISD;PLD3;PLD4;PTDSS1;PTDSS2;SELENOI
cluster245_Mito_size10: AARS2;FARS2;GARS;HARS2;POLG;POLG2;TARS;TARS2;THNSL1;WARS2
cluster246_Mito;ER_size10:
PABPC5;ZDHHC1;ZDHHC11;ZDHHC11B;ZDHHC12;ZDHHC14;ZDHHC24;ZDHHC4;ZDHHC5;ZDHHC8
cluster247_ER_size9: GPAA1;LGMN;PGAP1;PIGF;PIGK;PIGO;PIGS;PIGT;PIGU
cluster248_MitoER;ER_size9:
ORMDL1;ORMDL2;ORMDL3;SPTLC1;SPTLC2;SPTLC3;SPTSSA;SPTSSB;ZDHHC9
cluster249_Mito_size9: PINK1;TOMM20;TOMM22;TOMM40;TOMM40L;TOMM5;TOMM6;TOMM7;TOMM70
cluster250_Mito;ER_size18:
DDX28;FKBP7;LSG1;METAP1D;MRM2;MTERF4;MTG1;NOA1;NSUN3;NSUN4;PPTC7;PUSL1;RPUSD3;TF
B1M;TRMT1;TRMT11;TRMT2B;TRMT61B
cluster251_Mito;ER_size9: CASQ2;CATSPERB;HERPUD1;MCU;MCUB;MICU1;MICU2;MICU3;SMDT1
cluster252_MitoER;ER_size9: AADAC;AADACL3;AADACL4;CYP2D6;FMO1;FMO3;FMO4;FMO5;NCEH1
cluster253_Mito_size9: DUS2;GTPBP10;GUF1;MALSU1;METTL15;MTG2;MTRF1;OSGEPL1;YME1L1
cluster254_Mito;ER_size9:
ALOX5AP;CCDC58;HMOX1;HMOX2;LTC4S;REXO2;SSBP1;TMEM126A;TMEM205
cluster255_Mito;ER_size9: CD164;DPY19L1;DPY19L3;MR1;PNPLA8;RIC3;SARAF;SERINC1;TMEM30A
cluster256_Mito;ER_size8: RPL10A;RPL34;RPL35A;RPS14;RPS15A;RPS18;SRPRA;SRPRB
cluster257_Mito;ER_size8: DARS2;EARS2;GATB;GATC;NARS2;PTRH1;QRSL1;RCN1
cluster258_Mito;ER_size8: G6PD;GPI;H6PD;MINPP1;PGLS;PKLR;RPIA;TKT
cluster259_Mito;ER_size8: COQ10A;COQ10B;COQ4;COQ5;COQ7;COQ8B;COQ9;SLC6A4
cluster260_ER_size8: CNPY2;DNAJB11;DNAJC10;DNAJC3;HSPA5;HYOU1;MANF;PDIA6
cluster261_Mito_size16:
BCKDHA;BCKDHB;BCKDK;DBT;DHTKD1;DLAT;DLD;DLST;OGDH;PDHA1;PDHB;PDHX;PDK1;PDK2;PDK3;
TXNRD1
cluster262_Mito;ER_size8: HLA-E;HLA-F;HLA-G;IFI27;RSAD2;TAP1;TAP2;TAPBP
cluster263_ER_size8: CERCAM;COL22A1;COL28A1;COLGALT1;COLGALT2;PLOD1;PLOD2;PLOD3
cluster264_Mito;ER_size8: CYP11A1;CYP11B2;CYP17A1;CYP21A2;CYP7B1;HSD11B2;HSD3B1;HSD3B2
cluster265_ER_size8: ACER3;CERS2;CERS5;CERS6;KDSR;SGPL1;SGPP1;SGPP2
cluster266_MitoER;Mito;ER_size8: CBR3;COMT;CYP1A1;CYP26C1;CYP2S1;UGT1A3;UGT1A5;UGT1A7
cluster267_Mito;ER_size8: BTRC;CRY1;FBXL4;KLHL41;RNF19B;UBE2J1;UBE2J2;UFL1
cluster268_Mito_size8: ALKBH1;ALKBH3;ALKBH7;APEX2;MUTYH;NTHL1;OGG1;PDE12
cluster269_Mito;ER_size8: ABCB7;ALG13;CA5B;CLCN4;FUNDC1;FUNDC2;SLC25A43;TMEM164
cluster270_Mito;ER_size8: EFHD1;KIAA0100;KRT5;MTFR1L;MYH15;SCARF1;SDF4;TCHP
cluster271_Mito_size7: SDHA;SDHB;SDHC;SDHD;SUCLA2;SUCLG1;SUCLG2
cluster272_Mito_size15: COA6;COX4I2;COX6A2;COX6B2;COX7A1;COX8C;CYC1;MT-
CYB;UQCR10;UQCR11;UQCRB;UQCRC1;UQCRC2;UQCRFS1;UQCRQ
cluster273_Mito;ER_size7: CCT7;DNAJC25-GNG10;DNAJC30;HSPD1;HSPE1;LONP2;TRAP1
cluster274_Mito;ER_size7: CNIH4;KTN1;MTCH1;PSMA6;PSMD7;PSMF1;RNF170
cluster275_ER_size7: AD000671.1;APH1A;APH1B;NCSTN;PSEN1;PSEN2;PSENEN
cluster276_Mito_size7: COX10;COX11;COX15;RIDA;SCO1;SCO2;SURF1
cluster277_MitoER;ER_size7: CH25H;CYP27A1;CYP46A1;CYP7A1;SLC27A5;SOAT1;SOAT2
cluster278_Mito_size7: AK2;DTYMK;DUT;MOCS1;NME4;NME6;NUDT2
cluster279_ER_size7: CTAGE5;LMAN1;LMAN1L;MCFD2;MIA3;PREB;TMED1
cluster280_MitoER;Mito_size7: AKAP1;MARCH5;MFN1;MFN2;MTERF3;SLC25A26;SLC25A38
cluster281_ER_size7: ASPHD2;KLHL14;TOR1A;TOR1AIP2;TOR1B;TOR2A;TOR3A
cluster282_Mito;ER_size7: ENDOG;PTGES;PTGES2;PTGIS;PTGS1;PTGS2;TBXAS1
cluster283_Mito;ER_size15:
ABHD1;ADIG;CALR3;CTAGE1;CTAGE9;FATE1;FTMT;GSG1;LRIT1;MARCH10;RNF148;RNF183;SPATA19;
ZDHHC19;ZNRF4
cluster284_MitoER;ER_size7: AGPAT4;GPAM;GPAT3;GPAT4;LCLAT1;LPCAT3;MBOAT7
cluster285_Mito;ER_size7: NBEAL2;NLGN2;NLGN3;NLGN4X;TPRA1;UVRAG;WDR81
cluster286_Mito;ER_size7: ACER1;ASAH2;CERS3;CERS4;DEGS1;DEGS2;UGT8
cluster287_MitoER;Mito;ER_size7: GPX1;GPX4;GSR;GSTO1;MGST2;MGST3;TXNDC12
cluster288_Mito;ER_size7: AK3;AK4;CMPK2;NT5C;NT5M;PDE3B;PDE4D
cluster289_Mito;ER_size7: ADAMTS9;B3GLCT;CTSW;POFUT1;POFUT2;SCCPDH;THBS1
cluster290_Mito;ER_size7: BLOC1S1;CPD;FTH1;KDELC1;KDELC2;NCOA4;TXNDC5
cluster291_MitoER;Mito;ER_size7: AIFM3;CCDC47;CISD1;CISD2;CYB5A;CYB5B;CYB5R3
cluster292_Mito_size7: C21orf33;HDHD3;NIT1;NUDT13;NUDT5;NUDT9;PYCR2
cluster293_MitoER;Mito;ER_size7: ACSM5;ADHFE1;ALDH1L1;ALDH1L2;DHRS1;DHRS7B;HSD17B13
cluster294_Mito;ER_size14:
AUP1;ERLEC1;FAM8A1;KIAA0141;MARCH6;OS9;RNF103;SEL1L;SEL1L2;SEL1L3;SHH;SYVN1;TMEM41B
;TRIM13
cluster295_Mito_size7: CHCHD10;CHCHD2;GHITM;NDUFAF8;RNASEH1;SLC16A1;SPRYD4
cluster296_Mito_size7: ABHD10;CISD3;DHX30;FAM136A;GRSF1;NIPSNAP1;NIPSNAP2
cluster297_MitoER;ER_size7: ATAD1;MARCH4;NT5DC3;RNF103-CHMP3;RNF150;UBQLN1;UCHL1
cluster298_Mito_size7: DNM1L;FIS1;GDAP1;MFF;MIEF1;PEX11B;PGAM5
cluster299_Mito;ER_size7: ATAD3B;GTPBP6;NPHS2;SFXN3;STOM;STOML1;TRIT1
cluster300_MitoER;ER_size7: ABCA13;ABCA9;DGAT2L6;DUOXA1;DUOXA2;SLC38A10;ZDHHC16
cluster301_Mito;ER_size7: BCAP29;BCL2L13;CKMT1B;FICD;LRPAP1;NBR1;RMND1
cluster302_Mito_size7: SLC25A21;SLC25A29;SLC25A39;SLC25A40;SLC25A42;SLC25A45;SLC25A48
cluster303_Mito;ER_size7: ALG14;DMAC2;LRRC59;SLC35B1;TMEM109;TMEM147;TMEM208
cluster304_Mito_size6: ACACA;ACACB;FASN;MCAT;MLYCD;OXSM
cluster305_Mito;ER_size14:
ATG9A;ERAP1;ERAP2;ERMP1;GLRX2;LCTL;MTHFS;NAALAD2;PARL;RHBDF1;RHBDF2;TSTD1;UGT2B2
8;UGT3A2
cluster306_Mito_size6: BCS1L;SLC25A6;TIMM10;TIMM22;TIMM8A;TIMM9
cluster307_Mito_size6: AMT;DMGDH;GCSH;GLDC;IBA57;SARDH
cluster308_ER_size6: SEC61A1;SEC61B;SEC61G;SEC62;SEC63;SERP1
cluster309_Mito_size6: ACLY;CS;FH;MDH1;MDH2;MGME1
cluster310_Mito_size6: ADCK1;ADCK2;ADCK5;COQ3;COQ6;COQ8A
cluster311_Mito_size6: CPOX;FECH;HMBS;LYPLAL1;PPOX;SDR39U1
cluster312_Mito_size6: ABAT;AGXT2;ALDH1B1;ALDH5A1;ALDH6A1;HIBADH
cluster313_ER_size6: HACD1;HACD2;HACD3;HACD4;TECR;TECRL
cluster314_MitoER;Mito_size6: APOO;APOOL;C19orf70;CHCHD3;CHCHD6;IMMT
cluster315_MitoER;ER_size6: CRTAP;MZB1;P3H1;P4HB;PPIB;SERPINH1

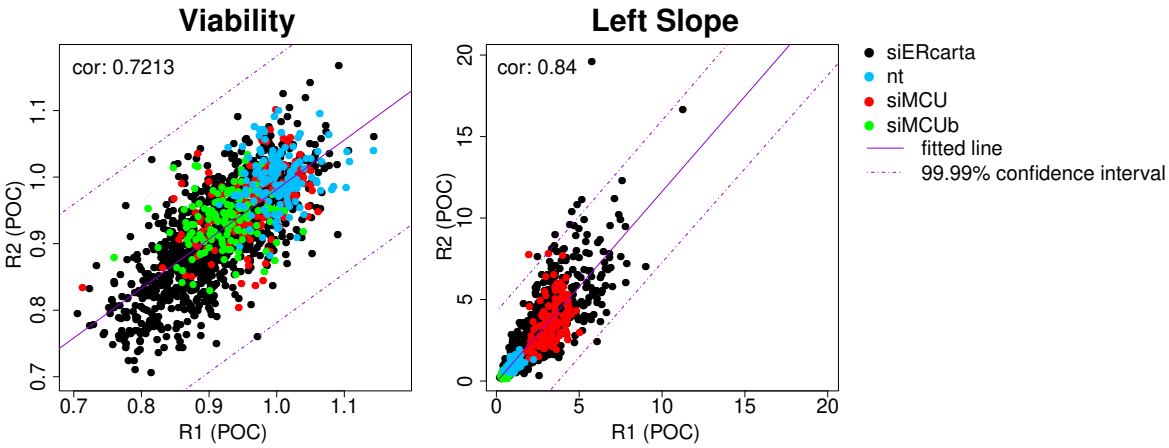


Fig. 4.5 **Reproducibility of viability for mt-Ca^{2+} screen.** Dots from different groups are indicated by colors (nt: negative control, siRNAs targeting no genes; siMCU: siRNAs targeting MCU; siMCUb: siRNAs targeting MCUb; siERcarta: siRNAs targeting ERcarta genes).

References

[1] Yiming Cheng, Li Jiang, Susanne Keipert, Shuyue Zhang, Andreas Hauser, Elisabeth Graf, Tim Strom, Matthias Tschöp, Martin Jastroch, and Fabiana Perocchi. Prediction of adipose browning capacity by systematic integration of transcriptional profiles. *Cell reports*, 23(10):3112–3125, 2018.

[2] Xiu-Hui Shi, Xu Li, Hang Zhang, Rui-Zhi He, Yan Zhao, Min Zhou, Shu-Tao Pan, Chun-Le Zhao, Ye-Chen Feng, Min Wang, et al. A five-microrna signature for survival prognosis in pancreatic adenocarcinoma based on tcga data. *Scientific reports*, 8(1): 1–10, 2018.

[3] Yanglan Gan, Jihong Guan, and Shuigeng Zhou. A comparison study on feature selection of dna structural properties for promoter prediction. *BMC bioinformatics*, 13 (1):4, 2012.

[4] Julia Hippisley-Cox, Carol Coupland, John Robson, Aziz Sheikh, and Peter Brindle. Predicting risk of type 2 diabetes in england and wales: prospective derivation and validation of qdscore. *Bmj*, 338:b880, 2009.

[5] Diego Kuonen. Data mining and statistics: What is the connection? *The Data Administration Newsletter*, 30, 2004.

[6] Patrick R Schmid, Nathan P Palmer, Isaac S Kohane, and Bonnie Berger. Making sense out of massive data by going beyond differential expression. *Proceedings of the National Academy of Sciences*, 109(15):5594–5599, 2012.

[7] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[8] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[9] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[10] Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC press, 2012.

[11] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20 (3):273–297, 1995.

[12] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[13] Sean Davis and Paul S Meltzer. Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics*, 23(14):1846–1847, 2007.

[14] Belinda Phipson, Stanley Lee, Ian J Majewski, Warren S Alexander, and Gordon K Smyth. Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *The annals of applied statistics*, 10 (2):946, 2016.

[15] Sangya Pundir, Maria J Martin, and Claire O'Donovan. Uniprot protein knowledgebase. In *Protein Bioinformatics*, pages 41–55. Springer, 2017.

[16] Juan Antonio Vizcaíno, Attila Csordas, Noemi Del-Toro, José A Dianes, Johannes Griss, Ilias Lavidas, Gerhard Mayer, Yasset Perez-Riverol, Florian Reisinger, Tobias Ternent, et al. 2016 update of the pride database and its related tools. *Nucleic acids research*, 44(D1):D447–D456, 2015.

[17] Yasset Perez-Riverol, Qing-Wei Xu, Rui Wang, Julian Uszkoreit, Johannes Griss, Aniel Sanchez, Florian Reisinger, Attila Csordas, Tobias Ternent, Noemi del Toro, et al. Pride inspector toolsuite: moving toward a universal visualization tool for proteomics data standard formats and quality assessment of proteomexchange datasets. *Molecular & Cellular Proteomics*, 15(1):305–317, 2016.

[18] AA Terentiev, NT Moldogazieva, and KV Shaitan. Dynamic proteomics in modeling of the living cell. protein-protein interactions. *Biochemistry (Moscow)*, 74(13):1586–1607, 2009.

[19] Anne-Claude Gavin, Patrick Aloy, Paola Grandi, Roland Krause, Markus Boesche, Martina Marzioch, Christina Rau, Lars Juhl Jensen, Sonja Bastuck, Birgit Dümpelfeld, et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440 (7084):631, 2006.

[20] Ingrid Remy, Galia Ghaddar, and Stephen W Michnick. Using the β-lactamase protein-fragment complementation assay to probe dynamic protein–protein interactions. *Nature protocols*, 2(9):2302, 2007.

[21] Gary D Bader, Doron Betel, and Christopher WV Hogue. Bind: the biomolecular interaction network database. *Nucleic acids research*, 31(1):248–250, 2003.

[22] Ioannis Xenarios, Danny W Rice, Lukasz Salwinski, Marisa K Baron, Edward M Marcotte, and David Eisenberg. Dip: the database of interacting proteins. *Nucleic acids research*, 28(1):289–291, 2000.

[23] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl_1):D535–D539, 2006.

[24] Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, et al. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*, page gkw937, 2016.

[25] Suraj Peri, J Daniel Navarro, Ramars Amanchy, Troels Z Kristiansen, Chandra Kiran Jonnalagadda, Vineeth Surendranath, Vidya Niranjan, Babylakshmi Muthusamy, TKB Gandhi, Mads Gronborg, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome research*, 13 (10):2363–2371, 2003.

[26] Sandra Orchard, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H Campbell, Gayatri Chavali, Carol Chen, Noemi Del-Toro, et al. The mintact project—intact as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, 42(D1):D358–D363, 2013.

[27] Andrew Chatr-Aryamontri, Arnaud Ceol, Luisa Montecchi Palazzi, Giuliano Nardelli, Maria Victoria Schneider, Luisa Castagnoli, and Gianni Cesareni. Mint: the molecular interaction database. *Nucleic acids research*, 35(suppl_1):D572–D574, 2006.

[28] Carl F Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H Buetow. Pid: the pathway interaction database. *Nucleic acids research*, 37(suppl_1):D674–D679, 2008.

[29] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25, 2000.

[30] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45(D1):D353–D361, 2016.

[31] David Croft, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, Phani Garapati, Marc Gillespie, Maulik R Kamdar, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 42(D1):D472–D477, 2013.

[32] Gráinne S Gorman, Patrick F Chinnery, Salvatore DiMauro, Michio Hirano, Yasutoshi Koga, Robert McFarland, Anu Suomalainen, David R Thorburn, Massimo Zeviani, and Douglass M Turnbull. Mitochondrial diseases. *Nature reviews Disease primers*, 2 (1):1–22, 2016.

[33] Bradford B Lowell and Gerald I Shulman. Mitochondrial dysfunction and type 2 diabetes. *Science*, 307(5708):384–387, 2005.

[34] Douglas C Wallace. A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. *Annu. Rev. Genet.*, 39:359–407, 2005.

[35] Michael R Duchen. Mitochondria in health and disease: perspectives on a new mitochondrial biology. *Molecular aspects of medicine*, 25(4):365–451, 2004.

[36] Fabiana Perocchi, Lars J Jensen, Julien Gagneur, Uwe Ahting, Christian Von Mering, Peer Bork, Holger Prokisch, and Lars M Steinmetz. Assessing systems properties of yeast mitochondria through an interaction map of the organelle. *PLoS genetics*, 2(10), 2006.

[37] Sarah Calvo, Mohit Jain, Xiaohui Xie, Sunil A Sheth, Betty Chang, Olga A Goldberger, Antonella Spinazzola, Massimo Zeviani, Steven A Carr, and Vamsi K Mootha. Systematic identification of human mitochondrial disease genes through integrative genomics. *Nature genetics*, 38(5):576, 2006.

[38] David J Pagliarini, Sarah E Calvo, Betty Chang, Sunil A Sheth, Scott B Vafai, Shao-En Ong, Geoffrey A Walford, Canny Sugiana, Avihu Boneh, William K Chen, et al. A mitochondrial protein compendium elucidates complex i disease biology. *Cell*, 134 (1):112–123, 2008.

[39] Jacob O'Brien, Heyam Hayder, Yara Zayed, and Chun Peng. Overview of microrna biogenesis, mechanisms of actions, and circulation. *Frontiers in endocrinology*, 9:402, 2018.

[40] Chunxiang Zhang. Novel functions for small rna molecules. *Current opinion in molecular therapeutics*, 11(6):641, 2009.

[41] Peng Jiang, Haonan Wu, Wenkai Wang, Wei Ma, Xiao Sun, and Zuhong Lu. Mipred: classification of real and pseudo microrna precursors using random forest prediction model with combined features. *Nucleic acids research*, 35(suppl_2):W339–W344, 2007.

[42] Chenghai Xue, Fei Li, Tao He, Guo-Ping Liu, Yanda Li, and Xuegong Zhang. Classification of real and pseudo microrna precursors using local structure-sequence features and support vector machine. *BMC bioinformatics*, 6(1):310, 2005.

[43] Christopher Workman and Anders Krogh. No evidence that mrnas have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic acids research*, 27(24):4816–4822, 1999.

[44] Supatcha Lertampaiporn, Chinae Thammarongtham, Chakarida Nukoolkit, Boonserm Kaewkamnerdpong, and Marasri Ruengjitchatchawalya. Heterogeneous ensemble approach with discriminative features and modified-smotebagging for pre-mirna classification. *Nucleic acids research*, 41(1):e21–e21, 2013.

[45] Michael Hackenberg, Naiara Rodríguez-Ezpeleta, and Ana M Aransay. miranalyzer: an update on the detection and analysis of micrornas in high-throughput sequencing experiments. *Nucleic acids research*, 39(suppl_2):W132–W138, 2011.

[46] Robert J Peace, Kyle K Biggar, Kenneth B Storey, and James R Green. A framework for improving microrna prediction in non-human genomes. *Nucleic acids research*, 43(20):e138–e138, 2015.

[47] Andrea Frontini and Saverio Cinti. Distribution and development of brown adipocytes in the murine and human adipose organ. *Cell metabolism*, 11(4):253–256, 2010.

[48] Jeff Ishibashi and Patrick Seale. Beige can be slimming. *Science*, 328(5982):1113–1114, 2010.

[49] Irina G Shabalina, Natasa Petrovic, Jasper MA de Jong, Anastasia V Kalinovich, Barbara Cannon, and Jan Nedergaard. Ucp1 in brite/beige adipose tissue mitochondria is functionally thermogenic. *Cell reports*, 5(5):1196–1203, 2013.

[50] G Barbatelli, I Murano, Lise Madsen, Q Hao, M Jimenez, Karsten Kristiansen, JP Giacobino, Rita De Matteis, and S Cinti. The emergence of cold-induced brown adipocytes in mouse white fat depots is determined predominantly by white to brown adipocyte transdifferentiation. *American Journal of Physiology-Endocrinology and Metabolism*, 298(6):E1244–E1253, 2010.

[51] J Himms-Hagen, A Melnyk, MC Zingaretti, E Ceresi, G Barbatelli, and S Cinti. Multilocular fat cells in wat of cl-316243-treated rats derive directly from white adipocytes. *American Journal of Physiology-Cell Physiology*, 279(3):C670–C681, 2000.

[52] Tim J Schulz, Tian Lian Huang, Thien T Tran, Hongbin Zhang, Kristy L Townsend, Jennifer L Shadrach, Massimiliano Cerletti, Lindsay E McDougall, Nino Giorgadze, Tamara Tchkonia, et al. Identification of inducible brown adipocyte progenitors residing in skeletal muscle and white fat. *Proceedings of the National Academy of Sciences*, 108(1):143–148, 2011.

[53] Qiong A Wang, Caroline Tao, Rana K Gupta, and Philipp E Scherer. Tracking adipogenesis during white adipose tissue development, expansion and regeneration. *Nature medicine*, 19(10):1338, 2013.

[54] Stephen R Farmer. Obesity: Be cool, lose weight. *Nature*, 458(7240):839, 2009.

[55] Aaron M Cypess, Andrew P White, Cecile Vernochet, Tim J Schulz, Ruidan Xue, Christina A Sass, Tian Liang Huang, Carla Roberts-Toler, Lauren S Weiner, Cathy Sze, et al. Anatomical localization, gene expression profiling and functional characterization of adult human neck brown fat. *Nature medicine*, 19(5):635, 2013.

[56] Naja Zenius Jespersen, Therese Juhlin Larsen, Lone Peijs, Søren Daugaard, Preben Homøe, Annika Loft, Jasper de Jong, Neha Mathur, Barbara Cannon, Jan Nedergaard, et al. A classical brown adipose tissue mrna signature partly overlaps with brite in the supraclavicular region of adult humans. *Cell metabolism*, 17(5):798–805, 2013.

[57] Martin E Lidell, Matthias J Betz, Olof Dahlqvist Leinhard, Mikael Heglind, Louise Elander, Marc Slawik, Thomas Mussack, Daniel Nilsson, Thobias Romu, Pirjo Nuutila, et al. Evidence for two types of brown adipose tissue in humans. *Nature medicine*, 19 (5):631, 2013.

[58] Alexander Bartelt, Oliver T Bruns, Rudolph Reimer, Heinz Hohenberg, Harald Ittrich, Kersten Peldschus, Michael G Kaul, Ulrich I Tromsdorf, Horst Weller, Christian Waurisch, et al. Brown adipose tissue activity controls triglyceride clearance. *Nature medicine*, 17(2):200, 2011.

[59] So Yun Min, Jamie Kady, Minwoo Nam, Raziel Rojas-Rodriguez, Aaron Berkenwald, Jong Hun Kim, Hye-Lim Noh, Jason K Kim, Marcus P Cooper, Timothy Fitzgibbons, et al. Human'brite/beige'adipocytes develop from capillary networks, and their implantation improves metabolic homeostasis in mice. *Nature medicine*, 22(3):312, 2016.

[60] Kristin I Stanford, Roeland JW Middelbeek, Kristy L Townsend, Ding An, Eva B Nygaard, Kristen M Hitchcox, Kathleen R Markan, Kazuhiro Nakano, Michael F Hirshman, Yu-Hua Tseng, et al. Brown adipose tissue regulates glucose homeostasis and insulin sensitivity. *The Journal of clinical investigation*, 123(1), 2012.

[61] Masayuki Saito, Yuko Okamatsu-Ogura, Mami Matsushita, Kumiko Watanabe, Takeshi Yoneshiro, Junko Nio-Kobayashi, Toshihiko Iwanaga, Masao Miyagawa, Toshimitsu Kameya, Kunihiro Nakada, et al. High incidence of metabolically active brown adipose tissue in healthy adult humans: effects of cold exposure and adiposity. *Diabetes*, 58(7):1526–1531, 2009.

[62] Alexander Bartelt and Joerg Heeren. Adipose tissue browning and metabolic health. *Nature Reviews Endocrinology*, 10(1):24, 2014.

[63] Kosaku Shinoda, Ineke HN Luijten, Yutaka Hasegawa, Haemin Hong, Si B Sonne, Miae Kim, Ruidan Xue, Maria Chondronikola, Aaron M Cypess, Yu-Hua Tseng, et al. Genetic and functional characterization of clonally derived adult human brown adipocytes. *Nature medicine*, 21(4):389, 2015.

[64] Jun Wu, Pontus Boström, Lauren M Sparks, Li Ye, Jang Hyun Choi, An-Hoa Giang, Melin Khandekar, Kirsi A Virtanen, Pirjo Nuutila, Gert Schaart, et al. Beige adipocytes are a distinct type of thermogenic fat cell in mouse and human. *Cell*, 150(2):366–376, 2012.

[65] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. The compartmentalization of cells. In *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.

[66] Michael Schrader, Sandra Grille, H Dariush Fahimi, and Markus Islinger. Peroxisome interactions and cross-talk with other subcellular compartments in animal cells. In *Peroxisomes and their Key Role in Cellular Signaling and Metabolism*, pages 1–22. Springer, 2013.

[67] Jean Demarquoy and Françoise Le Borgne. Crosstalk between mitochondria and peroxisomes. *World journal of biological chemistry*, 6(4):301, 2015.

[68] Jun-ya Shoji, Takashi Kikuma, and Katsuhiko Kitamoto. Vesicle trafficking, organelle functions, and unconventional secretion in fungal physiology and pathogenicity. *Current opinion in microbiology*, 20:1–9, 2014.

[69] Sarah Cohen, Alex M Valm, and Jennifer Lippincott-Schwartz. Interacting organelles. *Current opinion in cell biology*, 53:84–91, 2018.

[70] Mariusz R Wieckowski, Carlotta Giorgi, Magdalena Lebiedzinska, Jerzy Duszynski, and Paolo Pinton. Isolation of mitochondria-associated membranes and mitochondria from animal tissues and cells. *Nature protocols*, 4(11):1582, 2009.

[71] Emily R Eden, Ian J White, Anna Tsapara, and Clare E Futter. Membrane contacts between endosomes and er provide sites for ptp1b–epidermal growth factor receptor interaction. *Nature cell biology*, 12(3):267, 2010.

[72] Dan H Moore and Helmut Ruska. Electron microscope study of mammalian cardiac muscle cells. *The Journal of Cell Biology*, 3(2):261–268, 1957.

[73] Rubén Fernández-Busnadiego, Yasunori Saheki, and Pietro De Camilli. Three-dimensional architecture of extended synaptotagmin-mediated endoplasmic reticulum–plasma membrane contact sites. *Proceedings of the National Academy of Sciences*, 112(16):E2004–E2013, 2015.

[74] Mounia Chami, Bénédicte Oulès, György Szabadkai, Rachida Tacine, Rosario Rizzuto, and Patrizia Paterlini-Bréchot. Role of serca1 truncated isoform in the proapoptotic calcium transfer from er to mitochondria during er stress. *Molecular cell*, 32(5): 641–651, 2008.

[75] Scot J Stone and Jean E Vance. Phosphatidylserine synthase-1 and-2 are localized to mitochondria-associated membranes. *Journal of Biological Chemistry*, 275(44): 34534–34540, 2000.

[76] Ryota Iwasawa, Anne-Laure Mahul-Mellier, Christoph Datler, Evangelos Pazarentzos, and Stefan Grimm. Fis1 and bap31 bridge the mitochondria–er interface to establish a platform for apoptosis induction. *The EMBO journal*, 30(3):556–568, 2011.

[77] Rosario Rizzuto, Paolo Pinton, Walter Carrington, Frederic S Fay, Kevin E Fogarty, Lawrence M Lifshitz, Richard A Tuft, and Tullio Pozzan. Close contacts with the endoplasmic reticulum as determinants of mitochondrial ca2+ responses. *Science*, 280 (5370):1763–1766, 1998.

[78] Soyeon Lee and Kyung-Tai Min. The interface between er and mitochondria: Molecular compositions and functions. *Molecules and cells*, 41(12):1000, 2018.

[79] Benoît Kornmann, Erin Currie, Sean R Collins, Maya Schuldiner, Jodi Nunnari, Jonathan S Weissman, and Peter Walter. An er-mitochondria tethering complex revealed by a synthetic biology screen. *science*, 325(5939):477–481, 2009.

[80] Olga Martins De Brito and Luca Scorrano. Mitofusin 2 tethers endoplasmic reticulum to mitochondria. *Nature*, 456(7222):605, 2008.

[81] Kurt J De Vos, Gabor M Morotz, Radu Stoica, Elizabeth L Tudor, Kwok-Fai Lau, Steven Ackerley, Alice Warley, Christopher E Shaw, and Christopher CJ Miller. Vapb interacts with the mitochondrial protein ptpip51 to regulate calcium homeostasis. *Human molecular genetics*, 21(6):1299–1311, 2011.

[82] György Szabadkai, Katiuscia Bianchi, Péter Várnai, Diego De Stefani, Mariusz R Wieckowski, Dario Cavagna, Anikó I Nagy, Tamás Balla, and Rosario Rizzuto. Chaperone-mediated coupling of endoplasmic reticulum and mitochondrial ca2+ channels. *J Cell Biol*, 175(6):901–911, 2006.

[83] Charles Betz, Daniele Stracka, Cristina Prescianotto-Baschong, Maud Frieden, Nicolas Demaurex, and Michael N Hall. mtor complex 2-akt signaling at mitochondria-associated endoplasmic reticulum membranes (mam) regulates mitochondrial physiology. *Proceedings of the National Academy of Sciences*, 110(31):12526–12534, 2013.

[84] Estela Area-Gomez, Ad de Groof, Eduardo Bonilla, Jorge Montesinos, Kurenai Tanji, Istvan Boldogh, Liza Pon, and Eric A Schon. A key role for mam in mediating mitochondrial dysfunction in alzheimer disease. *Cell death & disease*, 9(3):1–10, 2018.

[85] Ornella Moltedo, Paolo Remondelli, and Giuseppina Amodio. The mitochondria-endoplasmic reticulum contacts and their critical role in ageing and age-associated diseases. *Frontiers in cell and developmental biology*, 7:172, 2019.

[86] Saverio Marchi and Paolo Pinton. Alterations of calcium homeostasis in cancer cells. *Current opinion in pharmacology*, 29:1–6, 2016.

[87] Chloe N Poston, Srinivasan C Krishnan, and Carthene R Bazemore-Walker. In-depth proteomic analysis of mammalian mitochondria-associated membranes (mam). *Journal of proteomics*, 79:219–230, 2013.

[88] Sarah E Calvo, Karl R Clauser, and Vamsi K Mootha. Mitocarta2. 0: an updated inventory of mammalian mitochondrial proteins. *Nucleic acids research*, 44(D1): D1251–D1257, 2015.

[89] Daniel Sorger and Günther Daum. Triacylglycerol biosynthesis in yeast. *Applied microbiology and biotechnology*, 61(4):289–299, 2003.

[90] David K Breslow. Sphingolipid homeostasis in the endoplasmic reticulum and beyond. *Cold Spring Harbor perspectives in biology*, 5(4):a013326, 2013.

[91] Eva Sammels, Jan B Parys, Ludwig Missiaen, Humbert De Smedt, and Geert Bultynck. Intracellular ca 2+ storage in health and disease: a dynamic equilibrium. *Cell calcium*, 47(4):297–314, 2010.

[92] Tetsuo Minamino and Masafumi Kitakaze. Er stress in cardiovascular disease. *Journal of molecular and cellular cardiology*, 48(6):1105–1110, 2010.

[93] Benoit D Roussel, Antonina J Kruppa, Elena Miranda, Damian C Crowther, David A Lomas, and Stefan J Marciniak. Endoplasmic reticulum dysfunction in neurological disease. *The Lancet Neurology*, 12(1):105–118, 2013.

[94] Miriam Cnop, Fabienne Foufelle, and Licio A Velloso. Endoplasmic reticulum stress, obesity and diabetes. *Trends in molecular medicine*, 18(1):59–68, 2012.

[95] Yanping Song, Ying Jiang, Wantao Ying, Yan Gong, Yujuan Yan, Dong Yang, Jie Ma, Xiaofang Xue, Fan Zhong, Songfeng Wu, et al. Quantitative proteomic survey of endoplasmic reticulum in mouse liver. *Journal of proteome research*, 9(3):1195–1202, 2010.

[96] Leonard J Foster, Carmen L de Hoog, Yanling Zhang, Yong Zhang, Xiaohui Xie, Vamsi K Mootha, and Matthias Mann. A mammalian organelle map by protein correlation profiling. *Cell*, 125(1):187–199, 2006.

[97] Fang Peng, Xianquan Zhan, Mao-Yu Li, Fan Fang, Guoqing Li, Cui Li, Peng-Fei Zhang, and Zhuchu Chen. Proteomic and bioinformatics analyses of mouse liver microsomes. *International journal of proteomics*, 2012, 2012.

[98] Annalyn Gilchrist, Catherine E Au, Johan Hiding, Alexander W Bell, Julia Fernandez-Rodriguez, Souad Lesimple, Hisao Nagaya, Line Roy, Sara JC Gosline, Michael Hallett, et al. Quantitative proteomics analysis of the secretory pathway. *Cell*, 127(6): 1265–1281, 2006.

[99] Xuequn Chen, Maria Dolors Sans, John R Strahler, Alla Karnovsky, Stephen A Ernst, George Michailidis, Philip C Andrews, and John A Williams. Quantitative organellar proteomics analysis of rough endoplasmic reticulum from normal and acute pancreatitis rat pancreas. *Journal of proteome research*, 9(2):885–896, 2009.

[100] Matt Albertolle, Thanh TN Phan, Ambra Pozzi, and F Peter Guengerich. Sulfenylation of human liver and kidney microsomal cytochromes p450 and other drug metabolizing enzymes as a response to redox alteration. *Molecular & Cellular Proteomics*, pages mcp–RA117, 2018.

[101] Tatyana Goldberg, Maximilian Hecht, Tobias Hamp, Timothy Karl, Guy Yachdav, Nadeem Ahmed, Uwe Altermann, Philipp Angerer, Sonja Ansorge, Kinga Balasz, et al. Loctree3 prediction of localization. *Nucleic acids research*, 42(W1):W350–W355, 2014.

[102] Torsten Blum, Sebastian Briesemeister, and Oliver Kohlbacher. Multiloc2: integrating phylogeny and gene ontology terms improves subcellular protein localization prediction. *BMC bioinformatics*, 10(1):1, 2009.

[103] KO Wrzeszczynski and B Rost. Annotating proteins from endoplasmic reticulum and golgi apparatus in eukaryotic proteomes. *Cellular and Molecular Life Sciences CMLS*, 61(11):1341–1353, 2004.

[104] M Scott, Guoqing Lu, M Hallett, and David Y Thomas. The hera database and its use in the characterization of endoplasmic reticulum proteins. *Bioinformatics*, 20(6): 937–944, 2004.

[105] Ravindra Kumar, Bandana Kumari, and Manish Kumar. Prediction of endoplasmic reticulum resident proteins using fragmented amino acid composition and support vector machine. *PeerJ*, 5:e3561, 2017.

[106] Ritesh K Baboota, Siddhartha M Sarma, Ravneet K Boparai, Kanthi Kiran Kondepudi, Shrikant Mantri, and Mahendra Bishnoi. Microarray based gene expression analysis of murine brown and subcutaneous adipose tissue: significance with human. *PloS one*, 10(5):e0127701, 2015.

[107] Sungsoon Fang, Jae Myoung Suh, Shannon M Reilly, Elizabeth Yu, Olivia Osborn, Denise Lackey, Eiji Yoshihara, Alessia Perino, Sandra Jacinto, Yelizaveta Lukasheva, et al. Intestinal fxr agonism promotes adipose tissue browning and reduces obesity and insulin resistance. *Nature medicine*, 21(2):159, 2015.

[108] Timothy P Fitzgibbons, Sophia Kogan, Myriam Aouadi, Greg M Hendricks, Juerg Straubhaar, and Michael P Czech. Similarity of mouse perivascular and brown adipose tissues and their resistance to diet-induced inflammation. *American Journal of Physiology-Heart and Circulatory Physiology*, 301(4):H1425–H1437, 2011.

[109] Benedetto Grimaldi, Marina Maria Bellet, Sayako Katada, Giuseppe Astarita, Jun Hirayama, Rajesh H Amin, James G Granneman, Daniele Piomelli, Todd Leff, and Paolo Sassone-Corsi. Per2 controls lipid metabolism by direct regulation of pparγ. *Cell metabolism*, 12(5):509–520, 2010.

[110] Jonathan Z Long, Katrin J Svensson, Linus Tsai, Xing Zeng, Hyun C Roh, Xingxing Kong, Rajesh R Rao, Jesse Lou, Isha Lokurkar, Wendy Baur, et al. A smooth muscle-like origin for beige adipocytes. *Cell metabolism*, 19(5):810–820, 2014.

[111] Susan M Majka, Keith E Fox, John C Psilas, Karen M Helm, Christine R Childs, Alistaire S Acosta, Rachel C Janssen, Jacob E Friedman, Brian T Woessner, Theodore R Shade, et al. De novo generation of white adipocytes from the myeloid lineage via mesenchymal intermediates is age, adipose depot, and gender specific. *Proceedings of the National Academy of Sciences*, 107(33):14781–14786, 2010.

[112] Haruya Ohno, Kosaku Shinoda, Bruce M Spiegelman, and Shingo Kajimura. Pparγ agonists induce a white-to-brown fat conversion through stabilization of prdm16 protein. *Cell metabolism*, 15(3):395–404, 2012.

[113] Meritxell Rosell, Myrsini Kaforou, Andrea Frontini, Anthony Okolo, Yi-Wah Chan, Evanthia Nikolopoulou, Steven Millership, Matthew E Fenech, David MacIntyre, Jeremy O Turner, et al. Brown and white adipose tissues: intrinsic differences in gene expression and response to cold exposure in mice. *American Journal of Physiology-Endocrinology and Metabolism*, 306(8):E945–E964, 2014.

[114] Patrick Seale, Shingo Kajimura, Wenli Yang, Sherry Chin, Lindsay M Rohas, Marc Uldry, Geneviève Tavernier, Dominique Langin, and Bruce M Spiegelman. Transcriptional control of brown fat determination by prdm16. *Cell metabolism*, 6(1):38–54, 2007.

[115] Louis Z Sharp, Kosaku Shinoda, Haruya Ohno, David W Scheel, Emi Tomoda, Lauren Ruiz, Houchun Hu, Larry Wang, Zdena Pavlova, Vicente Gilsanz, et al. Human bat possesses molecular signatures that resemble beige/brite cells. *PloS one*, 7(11):e49452, 2012.

[116] Andrew I Su, Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith A Ching, David Block, Jie Zhang, Richard Soden, Mimi Hayakawa, Gabriel Kreiman, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6062–6067, 2004.

[117] James A Timmons, Kristian Wennmalm, Ola Larsson, Tomas B Walden, Timo Lassmann, Natasa Petrovic, D Lee Hamilton, Ruth E Gimeno, Claes Wahlestedt, Keith Baar, et al. Myogenic gene expression signature establishes that brown and white adipocytes originate from distinct cell lineages. *Proceedings of the National Academy of Sciences*, 104(11):4401–4406, 2007.

[118] Hong Wang, Libin Liu, Jean Z Lin, Tamar R Aprahamian, and Stephen R Farmer. Browning of white adipose tissue with roscovitine induces a distinct population of ucp1+ adipocytes. *Cell metabolism*, 24(6):835–847, 2016.

[119] Yuan Xue, Natasa Petrovic, Renhai Cao, Ola Larsson, Sharon Lim, Shaohua Chen, Helena M Feldmann, Zicai Liang, Zhenping Zhu, Jan Nedergaard, et al. Hypoxia-independent angiogenesis in adipose tissues during cold acclimation. *Cell metabolism*, 9(1):99–109, 2009.

[120] Ray Zhang, Nicholas F Lahens, Heather I Ballance, Michael E Hughes, and John B Hogenesch. A circadian gene expression atlas in mammals: implications for biology and medicine. *Proceedings of the National Academy of Sciences*, 111(45):16219–16224, 2014.

[121] Juan R Alvarez-Dominguez, Zhiqiang Bai, Dan Xu, Bingbing Yuan, Kinyui Alice Lo, Myeong Jin Yoon, Yen Ching Lim, Marko Knoll, Nikolai Slavov, Shuai Chen, et al. De novo reconstruction of adipose tissue transcriptomes reveals long non-coding rna regulators of brown adipocyte development. *Cell metabolism*, 21(5):764–776, 2015.

[122] Timo Kanzleiter, Tatjana Schneider, Isabel Walter, Florian Bolze, Christoph Eickhorst, Gerhard Heldmaier, Susanne Klaus, and Martin Klingenspor. Evidence for nr4a1 as a cold-induced effector of brown fat thermogenesis. *Physiological genomics*, 24(1): 37–44, 2006.

[123] Marta Bou, Jerôme Montfort, Aurélie Le Cam, Cécile Rallière, Véronique Lebret, Jean-Charles Gabillard, Claudine Weil, Joaquim Gutiérrez, Pierre-Yves Rescan, Encarnación Capilla, et al. Gene expression profile during proliferation and differentiation of rainbow trout adipocyte precursor cells. *BMC genomics*, 18(1):347, 2017.

[124] Patrick Seale, Heather M Conroe, Jennifer Estall, Shingo Kajimura, Andrea Frontini, Jeff Ishibashi, Paul Cohen, Saverio Cinti, and Bruce M Spiegelman. Prdm16 determines the thermogenic program of subcutaneous white adipose tissue in mice. *The Journal of clinical investigation*, 121(1):96–105, 2011.

[125] Claire Tiraby and Dominique Langin. Conversion from white to brown adipocytes: a strategy for the control of fat mass? *Trends in Endocrinology & Metabolism*, 14(10): 439–441, 2003.

[126] Ruidan Xue, Matthew D Lynes, Jonathan M Dreyfuss, Farnaz Shamsi, Tim J Schulz, Hongbin Zhang, Tian Lian Huang, Kristy L Townsend, Yiming Li, Hirokazu Takahashi, et al. Clonal analyses and gene profiling identify genetic biomarkers of the thermogenic potential of human brown and white preadipocytes. *Nature medicine*, 21(7):760, 2015.

[127] D Tews, V Schwar, M Scheithauer, T Weber, T Fromme, M Klingenspor, TF Barth, P Möller, K Holzmann, KM Debatin, et al. Comparative gene array analysis of progenitor cells from human paired deep neck and subcutaneous adipose tissue. *Molecular and cellular endocrinology*, 395(1-2):41–50, 2014.

[128] Mark JW Hanssen, Joris Hoeks, Boudewijn Brans, Anouk AJJ Van Der Lans, Gert Schaart, José J Van Den Driessche, Johanna A Jörgensen, Mark V Boekschoten, Matthijs KC Hesselink, Bas Havekes, et al. Short-term cold acclimation improves insulin sensitivity in patients with type 2 diabetes mellitus. *Nature medicine*, 21(8): 863, 2015.

[129] Annie Moisan, Youn-Kyoung Lee, Jitao David Zhang, Carolyn S Hudak, Claas A Meyer, Michael Prummer, Sannah Zoffmann, Hoa Hue Truong, Martin Ebeling, Anna Kiialainen, et al. White-to-brown metabolic conversion of human adipocytes by jak inhibition. *Nature cell biology*, 17(1):57, 2015.

[130] Matthias Rosenwald, Aliki Perdikari, Thomas Rülicke, and Christian Wolfrum. Bidirectional interconversion of brite and white adipocytes. *Nature cell biology*, 15(6): 659, 2013.

[131] TJ Conere, SW Church, and WS Lowry. The radioprotective role of common anaesthetic and other narcotic agents. *Radiotherapy and Oncology*, 5(4):347–348, 1986.

[132] Brooks P Leitner, Shan Huang, Robert J Brychta, Courtney J Duckworth, Alison S Baskin, Suzanne McGehee, Ilan Tal, William Dieckmann, Garima Gupta, Gerald M Kolodny, et al. Mapping of human brown adipose tissue in lean and obese young men. *Proceedings of the National Academy of Sciences*, page 201705287, 2017.

[133] Thobias Romu, Camilla Vavruch, Olof Dahlqvist-Leinhard, Joakim Tallberg, Nils Dahlström, Anders Persson, Mikael Heglind, Martin E Lidell, Sven Enerbäck, Magnus Borga, et al. A randomized trial of cold-exposure on energy expenditure and supraclavicular brown adipose tissue volume in humans. *Metabolism-Clinical and Experimental*, 65(6):926–934, 2016.

[134] Maarten J Vosselman, Guy HEJ Vijgen, Boris RM Kingma, Boudewijn Brans, and Wouter D van Marken Lichtenbelt. Frequent extreme cold exposure and brown fat and cold-induced thermogenesis: a study in a monozygotic twin. *PloS one*, 9(7):e101653, 2014.

[135] Edison T Liu. Systems biology, integrative biology, predictive biology. *Cell*, 121(4): 505–506, 2005.

[136] Laurent Gautier, Leslie Cope, Benjamin M Bolstad, and Rafael A Irizarry. affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3): 307–315, 2004.

[137] Benilton S Carvalho and Rafael A Irizarry. A framework for oligonucleotide microarray preprocessing. *Bioinformatics*, 26(19):2363–2367, 2010.

[138] Steffen Durinck, Paul T Spellman, Ewan Birney, and Wolfgang Huber. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature protocols*, 4(8):1184, 2009.

[139] Cole Trapnell, Lior Pachter, and Steven L Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[140] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. Htseq—a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, 2015.

[141] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

[142] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[143] Khadija El Amrani, Harald Stachelscheid, Fritz Lekschas, Andreas Kurtz, and Miguel A Andrade-Navarro. Mgfm: a novel tool for detection of tissue and cell specific marker genes from microarray gene expression data. *BMC genomics*, 16(1): 645, 2015.

[144] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[145] Oh Kwang Kwon, JuHee Sim, Sun Ju Kim, Eunji Sung, Jin Young Kim, Tae Cheon Jeong, and Sangkyu Lee. Comprehensive analysis of in vivo phosphoproteome of mouse liver microsomes. *Journal of proteome research*, 14(12):5215–5224, 2015.

[146] Cornelia M Hooper, Sandra K Tanz, Ian R Castleden, Michael A Vacher, Ian D Small, and A Harvey Millar. Subacon: a consensus algorithm for unifying the subcellular localization data of the arabidopsis proteome. *Bioinformatics*, 30(23):3356–3364, 2014.

[147] Jin-sook Lee, Yanning Wu, Patricia Schnepp, Jingye Fang, Xuebao Zhang, Alla Karnovsky, James Woods, Paul M Stemmer, Ming Liu, Kezhong Zhang, et al. Proteomics analysis of rough endoplasmic reticulum in pancreatic beta cells. *Proteomics*, 15(9):1508–1511, 2015.

[148] Mariano Stornaiuolo, Lavinia V Lotti, Nica Borgese, Maria-Rosaria Torrisi, Giovanna Mottola, Gianluca Martire, and Stefano Bonatti. Kdel and kkxx retrieval signals appended to the same reporter protein determine different trafficking between endoplasmic reticulum, intermediate compartment, and golgi complex. *Molecular biology of the cell*, 14(3):889–902, 2003.

[149] Martin J Vincent, Annelet S Martin, and Richard W Compans. Function of the kkxx motif in endoplasmic reticulum retrieval of a transmembrane protein depends on the length and structure of the cytoplasmic domain. *Journal of Biological Chemistry*, 273 (2):950–956, 1998.

[150] Roland Kabuß, Angel Ashikov, Stefan Oelmann, Rita Gerardy-Schahn, and Hans Bakker. Endoplasmic reticulum retention of the large splice variant of the udp-galactose transporter is caused by a dilysine motif. *Glycobiology*, 15(10):905–911, 2005.

[151] Lucie Aumailley, Florence Roux-Dalvai, Isabelle Kelly, Arnaud Droit, and Michel Lebel. Vitamin c alters the amount of specific endoplasmic reticulum associated proteins involved in lipid metabolism in the liver of mice synthesizing a nonfunctional werner syndrome (wrn) mutant protein. *PloS one*, 13(3):e0193170, 2018.

[152] Danièle Werck-Reichhart and René Feyereisen. Cytochromes p450: a success story. *Genome biology*, 1(6):reviews3003–1, 2000.

[153] F Peter Guengerich. Cytochrome p450 and chemical toxicology. *Chemical research in toxicology*, 21(1):70–83, 2007.

[154] Mustafa Erman, Mustafa Benekli, Mert Basaran, Sevil Bavbek, Suleyman Buyukberber, Ugur Coskun, Gokhan Demir, Bulent Karabulut, Berna Oksuzoglu, Metin Ozkan, et al. Renal cell cancer: overview of the current therapeutic landscape. *Expert review of anticancer therapy*, 16(9):955–968, 2016.

[155] Lale Ozcan and Ira Tabas. Role of endoplasmic reticulum stress in metabolic disease and other disorders. *Annual review of medicine*, 63:317–328, 2012.

[156] Yiming Cheng and Fabiana Perocchi. Protphylo: identification of protein–phenotype and protein–protein functional associations via phylogenetic profiling. *Nucleic acids research*, 43(W1):W160–W168, 2015.

[157] Yang Li, Sarah E Calvo, Roee Gutman, Jun S Liu, and Vamsi K Mootha. Expansion of biological pathways based on evolutionary inference. *Cell*, 158(1):213–225, 2014.

[158] Philip R Kensche, Vera van Noort, Bas E Dutilh, and Martijn A Huynen. Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. *Journal of the Royal Society Interface*, 5(19):151–170, 2008.

[159] Heidi L Liston, John S Markowitz, and C Lindsay DeVane. Drug glucuronidation in clinical psychopharmacology. *Journal of clinical psychopharmacology*, 21(5): 500–515, 2001.

[160] F Oesch, N Kaubisch, DM Jerina, and JW Daly. Hepatic epoxide hydrase. structure-activity relations for substrates and inhibitors. *Biochemistry*, 10(26):4858–4866, 1971.

[161] Jaana Laasanen, Eeva-Liisa Romppanen, Mikko Hiltunen, Seppo Helisalmi, Arto Mannermaa, Kari Punnonen, and Seppo Heinonen. Two exonic single nucleotide polymorphisms in the microsomal epoxide hydrolase gene are jointly associated with preeclampsia. *European Journal of Human Genetics*, 10(9):569, 2002.

[162] Ana M Gomes, Stefan Winter, Kathrin Klein, Miia Turpeinen, Elke Schaeffeler, Matthias Schwab, and Ulrich M Zanger. Pharmacogenomics of human liver cytochrome p450 oxidoreductase: multifactorial analysis and impact on microsomal drug oxidation. 2009.

[163] Steven N Hart, Shuang Wang, Kaori Nakamoto, Christopher Wesselman, Ye Li, and Xiao-bo Zhong. Genetic polymorphisms in cytochrome p450 oxidoreductase influence microsomal p450-catalyzed drug metabolism. *Pharmacogenetics and genomics*, 18 (1):11–24, 2008.

[164] Detlev Bannasch, Alexander Mehrle, Karl-Heinz Glatting, Rainer Pepperkok, Annemarie Poustka, and Stefan Wiemann. Lifedb: a database for functional genomics experiments integrating information from external sources, and serving as a sample tracking system. *Nucleic Acids Research*, 32(suppl 1):D505–D508, 2004.

[165] Andreas Ruepp, Brigitte Waegele, Martin Lechner, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and H-Werner Mewes. Corum: the comprehensive resource of mammalian protein complexes—2009. *Nucleic acids research*, 38(suppl 1):D497–D501, 2010.

[166] UniProt Consortium et al. Uniprot: a hub for protein information. *Nucleic acids research*, page gku989, 2014.

[167] Mathias Uhlén, Linn Fagerberg, Björn M Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, et al. Tissue-based map of the human proteome. *Science*, 347(6220):1260419, 2015.

[168] Mathias Uhlen, Per Oksvold, Linn Fagerberg, Emma Lundberg, Kalle Jonasson, Mattias Forsberg, Martin Zwahlen, Caroline Kampf, Kenneth Wester, Sophia Hober, et al. Towards a knowledge-based human protein atlas. *Nature biotechnology*, 28(12): 1248–1250, 2010.

[169] TS Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, et al. Human protein reference database—2009 update. *Nucleic acids research*, 37(suppl 1):D767–D772, 2009.

[170] Jeremy C Simpson, Ruth Wellenreuther, Annemarie Poustka, Rainer Pepperkok, and Stefan Wiemann. Systematic subcellular localization of novel proteins identified by large-scale cdna sequencing. *EMBO reports*, 1(3):287–292, 2000.

[171] Edouard De Castro, Christian JA Sigrist, Alexandre Gattiker, Virginie Bulliard, Petra S Langendijk-Genevaux, Elisabeth Gasteiger, Amos Bairoch, and Nicolas Hulo. Scanprosite: detection of prosite signature matches and prorule-associated functional and structural residues in proteins. *Nucleic acids research*, 34(suppl 2):W362–W365, 2006.

[172] Gene Ontology Consortium et al. Gene ontology consortium: going forward. *Nucleic acids research*, 43(D1):D1049–D1056, 2015.

[173] J Michael Cherry, Eurie L Hong, Craig Amundsen, Rama Balakrishnan, Gail Binkley, Esther T Chan, Karen R Christie, Maria C Costanzo, Selina S Dwight, Stacia R Engel, et al. Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic acids research*, page gkr1029, 2011.

[174] Charles E Grant, Timothy L Bailey, and William Stafford Noble. Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

[175] Juan Antonio Vizcaíno, Attila Csordas, Noemi del Toro, José A Dianes, Johannes Griss, Ilias Lavidas, Gerhard Mayer, Yasset Perez-Riverol, Florian Reisinger, Tobias Ternent, et al. 2016 update of the pride database and its related tools. *Nucleic acids research*, 44(D1):D447–D456, 2016.

[176] Juan A Vizcaíno, Eric W Deutsch, Rui Wang, Attila Csordas, Florian Reisinger, Daniel Rios, Jose A Dianes, Zhi Sun, Terry Farrah, Nuno Bandeira, et al. Proteomexchange provides globally coordinated proteomics data submission and dissemination. *Nature biotechnology*, 32(3):223–226, 2014.

[177] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, et al. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, page gku1003, 2014.

[178] Kenta Nakai and Paul Horton. Psort: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends in biochemical sciences*, 24(1): 34–35, 1999.

[179] Paul Horton, Keun-Joon Park, Takeshi Obayashi, Naoya Fujita, Hajime Harada, CJ Adams-Collier, and Kenta Nakai. Wolf psort: protein localization predictor. *Nucleic acids research*, 35(suppl 2):W585–W587, 2007.

[180] Thomas Nordahl Petersen, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. Signalp 4.0: discriminating signal peptides from transmembrane regions. *Nature methods*, 8(10):785, 2011.

[181] Olof Emanuelsson, Henrik Nielsen, Søren Brunak, and Gunnar Von Heijne. Predicting subcellular localization of proteins based on their n-terminal amino acid sequence. *Journal of molecular biology*, 300(4):1005–1016, 2000.

[182] Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber. Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–3440, 2005.

[183] Dmitry Terentyev, Alessandra Nori, Massimo Santoro, Serge Viatchenko-Karpinski, Zuzana Kubalova, Inna Gyorke, Radmila Terentyeva, Srikanth Vedamoorthyrao, Nico A Blom, Giorgia Valle, et al. Abnormal interactions of calsequestrin with the ryanodine receptor calcium release channel complex linked to exercise-induced sudden cardiac death. *Circulation research*, 98(9):1151–1158, 2006.

[184] Sic L Chan, Weiming Fu, Peisu Zhang, Aiwu Cheng, Jaewon Lee, Koichi Kokame, and Mark P Mattson. Herp stabilizes neuronal ca2+ homeostasis and mitochondrial function during endoplasmic reticulum stress. *Journal of Biological Chemistry*, 279 (27):28733–28743, 2004.

[185] Jin Liu, Jingsheng Xia, Kwang-Hyun Cho, David E Clapham, and Dejian Ren. Catsperβ, a novel transmembrane protein in the catsper channel complex. *Journal of Biological Chemistry*, 282(26):18945–18952, 2007.

[186] Hiroshi Takeshima, Seiichiro Nishimura, Takeshi Matsumoto, Hiroyuki Ishida, Kenji Kangawa, Naoto Minamino, Hisayuki Matsuo, Masamichi Ueda, Masao Hanaoka, Tadaaki Hirose, et al. Primary structure and expression from complementary dna of skeletal muscle ryanodine receptor. *Nature*, 339(6224):439, 1989.

[187] Grace E Stutzmann and Mark P Mattson. Endoplasmic reticulum ca2+ handling in excitable cells in health and disease. *Pharmacological reviews*, 63(3):700–727, 2011.

[188] Ying Qi, Eunice M Ogunbunmi, Eileen A Freund, Anthony P Timerman, and Sidney Fleischer. Fk-binding protein is associated with the ryanodine receptor of skeletal muscle in vertebrate animals. *Journal of Biological Chemistry*, 273(52):34813–34819, 1998.

[189] T. Ozawa. Modulation of ryanodine receptor Ca2+ channels (Review). *Mol Med Rep*, 3(2):199–204, 2010.

[190] Susan Treves, Giordana Feriotto, Luca Moccagatta, Roberto Gambari, and Francesco Zorzato. Molecular cloning, expression, functional characterization, chromosomal localization, and gene structure of junctate, a novel integral calcium binding protein of sarco (endo) plasmic reticulum membrane. *Journal of Biological Chemistry*, 275 (50):39555–39568, 2000.

[191] Sonal Srikanth, Marcus Jew, Kyun-Do Kim, Ma-Khin Yee, Jeff Abramson, and Yousang Gwack. Junctate is a ca2+-sensing structural component of orai1 and stromal interaction molecule 1 (stim1). *Proceedings of the National Academy of Sciences*, 109 (22):8682–8687, 2012.

[192] Dmitry Terentyev, Serge Viatchenko-Karpinski, Srikanth Vedamoorthyrao, Sridhar Oduru, Inna Györke, Simon C Williams, and Sandor Györke. Protein–protein interactions between triadin and calsequestrin are involved in modulation of sarcoplasmic reticulum calcium release in cardiac myocytes. *The Journal of physiology*, 583(1): 71–80, 2007.

[193] Joris Beld, Eva C Sonnenschein, Christopher R Vickery, Joseph P Noel, and Michael D Burkart. The phosphopantetheinyl transferases: catalysis of a post-translational modification crucial for life. *Natural product reports*, 31(1):61–108, 2014.

[194] Rosario Rizzuto and Tullio Pozzan. Microdomains of intracellular ca2+: molecular determinants and functional consequences. *Physiological reviews*, 86(1):369–408, 2006.

[195] Aurelien Deniaud, E Maillier, D Poncet, G Kroemer, C Lemaire, C Brenner, et al. Endoplasmic reticulum stress induces calcium-dependent permeability transition, mitochondrial outer membrane permeabilization and apoptosis. *Oncogene*, 27(3): 285–299, 2008.

[196] Wolfgang F Graier, Maud Frieden, and Roland Malli. Mitochondria and ca 2+ signaling: old guests, new functions. *Pflügers Archiv-European Journal of Physiology*, 455(3):375–396, 2007.

[197] Martijn A Huynen and Peer Bork. Measuring genome evolution. *Proceedings of the National Academy of Sciences*, 95(11):5849–5856, 1998.

[198] Li Li, Christian J Stoeckert, and David S Roos. Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome research*, 13(9):2178–2189, 2003.

[199] Sean Powell, Kristoffer Forslund, Damian Szklarczyk, Kalliopi Trachana, Alexander Roth, Jaime Huerta-Cepas, Toni Gabaldon, Thomas Rattei, Chris Creevey, Michael Kuhn, et al. eggnog v4. 0: nested orthology inference across 3686 organisms. *Nucleic acids research*, 42(D1):D231–D239, 2014.

[200] Peter Langfelder, Bin Zhang, and Steve Horvath. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. *Bioinformatics*, 24(5):719–720, 2007.

[201] Tamás Nepusz, Haiyuan Yu, and Alberto Paccanaro. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature methods*, 9(5):471, 2012.

[202] Anton J Enright, Stijn Van Dongen, and Christos A Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, 30(7):1575–1584, 2002.

[203] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[204] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[205] Daniela M Arduino, Jennifer Wettmarshausen, Horia Vais, Paloma Navas-Navarro, Yiming Cheng, Anja Leimpek, Zhongming Ma, Alba Delrio-Lorenzo, Andrea Giordano, Cecilia Garcia-Perez, et al. Systematic identification of mcu modulators by orthogonal interspecies chemical screening. *Molecular cell*, 67(4):711–723, 2017.

[206] Eran Eden, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC bioinformatics*, 10(1):48, 2009.