Machine Learning on Protein Expression Data - Predicting Functional Relationships Between Proteins

vorgelegt von
M. Sc.
Piotr Grabowski
ORCID: 0000-0001-9501-6192

von der Fakultät III – Prozesswissenschaften
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
- Dr. rer. nat. -

genehmigte Dissertation

Promotionsausschuss:
Vorsitzende: Prof. Dr. Vera Meyer
Gutachter: Prof. Dr. Juri Rappsilber
Gutachter: Prof. Dr. Matthias Selbach

Tag der wissenschaftlichen Aussprache: 25. Januar 2019

Berlin 2020

Table of contents

Abstract

Integrating gene expression data at transcript and protein level from many experiments helps in understanding functional relationships between genes, transcripts and the proteins they encode. Such approaches, collectively known as co-expression analysis, use various statistical methods to create pairwise association scores between genes or proteins. Co-expression analyses have been traditionally focused on transcript data due to the ever-increasing number of deposited datasets owing to the accessibility of mRNA-based technologies. However, there is growing evidence that protein expression is more closely linked to gene function. In this cumulative dissertation, I present my work on non-functional genomic effects on mRNA co-expression, which are absent on the protein level. These effects are predominantly rooted in genomic features such as 3D genome structure and epigenetic state. Genomic organization seems to have a direct, long-range effect on mRNA co-expression, e.g. through stochastic fluctuations between open and closed chromatin states or DNA replication timing. A considerable proportion of mRNA co-expression of spatially close gene pairs is not functional and buffered on the protein level, possibly through various post-transcriptional mechanisms. I demonstrate this effect in a human lymphoblastoid cell line panel and terminally differentiated mouse tissues by integrating publicly available omics datasets. Moreover, based on the notion of using protein data for co-expression analysis, I show how Random Forests can help in distinguishing patterns of mitochondrial protein localization in high-dimensional interphase chromatin data and even predict potential novel mitochondrial proteins. Finally, I show how machine learning can improve protein co-expression analytics over more classical statistical approaches, such as Pearson correlation. I integrate 294 high-quality SILAC experiments deposited in the PRIDE archive and calculate protein-wise functional links using tree-based unsupervised learning algorithm. The functional links between 5013 proteins resulting from my analysis are becoming part of the widely used STRING tool and thus will benefit biological researchers directly. Additionally, the resulting scores and data were made available via the ProteomeHD web app which I developed (https://www.proteomehd.net). At the methodological level, my work adds to the domain of computational systems biology and has impact on gene and protein function prediction efforts in the field. For example, the analysis of the protein co-expression scores helped to further annotate peroxisomal protein PEX11B and show its dual peroxisomal-mitochondrial function.

Zusammenfassung

Die Integration von Genexpressionsdaten aus Transkript- und Proteinhochdurchsatzmessungen hilft, funktionelle Beziehungen zwischen Genen, Transkripten und Proteinen zu verstehen. Ein bestimmter Ansatz, im Feld auch als Koexpressionsanalyse bezeichnet, nutzt verschiedene statistische Methoden, um paarweise Assoziationsmetriken zwischen Genen und Proteinen zu generieren. Bislang stützen sich Koexpressionsanalysen zumeist auf Transkriptionsdaten, da insbesondere dieser Typ Messdaten generiert und öffentlich verfügbar gemacht wurde. Jüngste Forschungsergebnisse legen jedoch nahe, dass die Expression von Proteinen stärker an die betreffende Genfunktion gebunden sind, als bisher angenommen. Diese kumulative Dissertation behandelt von mir untersuchte nicht-funktionale, genomische Effekte auf die Koexpression von mRNA, welche sich nicht auf die zu regulierenden Proteine auswirken. Diese Effekte beruhen zum überwiegenden Teil auf spezifischen genomischen Eigenschaften, wie der dreidimensionalen Chromatinstruktur und epigenetischer Zustände. Die genomische Architektur scheint direkte, weitreichende Effekte auf die mRNA-Koexpression zu haben, die beispielsweise aus stochastischen Fluktuationen zwischen offenen und geschlossenen Zuständen des Chromatins oder der Replikation von DNA hervorgehen könnte. Ein großer Anteil koexprimierter mRNAs proximal-liegender Gene besitzt keinen funktionalen Zusammenhang und wird auf Proteinebene gepuffert, wahrscheinlich aufgrund verschiedener posttranskriptioneller Mechanismen. Ich zeige diesen Effekt in menschlichen lymphoblastoiden Zelllinien und in differenzierten murinen Geweben durch Integration von öffentlich vorhandenen Omics-Datensätzen. Außerdem lege ich dar, wie ein *Random Forest*-Algorithmus Kovariationsmuster mitochondrialer Proteinen aus hochdimensionalen Interphasen-Chromatin-Daten extrahieren kann, um mögliche neue mitochondriale Proteine vorherzusagen. Schließlich zeige ich wie maschinelles Lernen die Analyse von Proteinkoexpression im Vergleich zu traditionellen statistischen Methoden, wie beispielsweise der Pearson Korrelationsanalyse, verbessern kann. Ich integriere 294 SILAC-Experimente, die im *PRIDE*-Archiv hinterlegt wurden und kalkuliere eine paarweise Protein-Assoziationsmetrik via *Decision Tree*-basiertem maschinellen Lernen. Beispielsweise erbrachte die detaillierte Analyse der Proteinkoexpressionsassoziationsmetrik eine neue Annotation des peroxisomalen Proteins PEX11B und half somit, dessen doppelte peroxisomal-mitochondriale Funktion aufzuklären. Die funktionelle Assoziationsmetrik zwischen den 5013 in meiner Analyse untersuchten Proteinen wird Teil der sehr weit verbreiteten *STRING*-Datenbank und wird die biologische Forschung unterstützen. Zusätzlich wurden die erarbeitete Assoziationsmetrik und Daten über die von mir erstellte ProteomeHD Web App

(https://www.proteomehd.net) verfügbar gemacht. Meine Arbeit fügt ein bedeutendes Werkzeug zur Vorhersage von Gen- und Proteinfunktionen zu bisherigen Mitteln hinzu und trägt somit dazu bei, das Forschungsfeld rechentechnischer Systembiologie weiterzuentwickeln.

Abbreviations

BLAST:	Basic Local Search Alignment Tool
CAFA:	critical assessment of protein function annotation
DBSCAN:	Density-based spatial clustering of applications with noise
GEO:	Gene Expression Omnibus
GO:	Gene Ontology
MIPS:	Munich Information Center for Protein Sequences
miRNAs:	microRNAs
ncRNAs:	non-coding RNAs
PCA:	principal component analysis
PRIDE:	Proteomics Identifications
PSI-BLAST:	Position-specific Iterative Basic Local Alignment Search Tool
PTM:	post-translational modifications
RNAseq:	RNA sequencing
STRING:	Search Tool for the Retrieval of Interacting Genes/Proteins

Introduction

Genes are the basic functional units of genomes. Understanding the functional relationships between genes has long been a major goal of molecular biology. Genes can have regulatory relationships with each other, for example based on their physical proximity on the genome, while their products (RNA and protein) can have functional relationships throughout the cell. Functional relationships of gene products can take many forms, such as direct physical interactions whereby the RNA or proteins they encode form complexes, or indirect “functional” interactions, for example where their encoded RNA or proteins cooperate in a metabolic pathway. Further understanding of these relationships can be applied to improve drug design (Hase et al. 2009), clinical diagnostics (Su, Yoon, and Dougherty 2010; Zhang and Chen 2010) or improve biotechnological processes (D. Li et al. 2016).

Proteins constitute the main functional output of the genome. These amino acid-based machines are responsible for most of the functional aspects of the cell such as creating structural scaffolds, carrying out intra- and intercellular signalling, performing enzymatic reactions, among others. Out of ca. 58.000 known human genes, ~20.000 are protein-coding (Harrow et al. 2012). Studying the function of those genes and their relationship to phenotype has long been a goal of functional genomics, a mature subfield of biology. While traditionally functional genomics has been focused on genomics and transcriptomics technologies, proteomics can help in developing understanding of the relationships between genes by analyzing the functions of the proteins they encode. For this reason, I use here the terms “gene function prediction” and “protein function prediction” interchangeably.

Despite many years of extensive research, the function of many proteins remains elusive. Around 38% of the human proteome is considered “understudied” and without extensive functional annotation (Oprea et al. 2018). The necessary extensive wet-lab characterization is a tedious and and resource-consuming task. To aid this process, systems biology and bioinformatics methods that help infer the function of proteins are being developed to help shed light on the potential function of genes and proteins.

One of such computational methods is gene and protein function prediction. Gene function prediction has a long history, with the main impact coming from the BLAST (Altschul et al. 1990) and PSI-BLAST algorithms (Altschul et al. 1997) which allowed researchers to search (or “blast”) their

sequences of interest against functionally annotated databases and learn something about the possible function of the analyzed gene or protein. With the explosion of genomic data in the last two decades and development of statistical frameworks, sequence-based approaches to function prediction became more sophisticated. Currently, community-based initiatives, like the critical assessment of protein function annotation (CAFA) (Radivojac et al. 2013), document fifty-four such methods for function prediction which include approaches such as Bayesian phylogenomics (Engelhardt et al. 2005), protein function prediction using patterns of native disorder (A. Lobley et al. 2007) or function prediction using sequence-based features (A. E. Lobley et al. 2008).

However, these and similar sequence-based approaches have drawbacks. In most of these methods, gene and protein co-function is typically defined as gene or protein pairs sharing same gene ontology (GO) terms (Ashburner et al. 2000). While such ontologies are useful for functional classification of genes and proteins, they don't always capture their multifunctional character and contain many electronic annotations without manual curation (Khatri and Drăghici 2005). Moreover, it's not clear how much functional information can be extracted from primary amino acid sequence alone, as protein function is mainly conferred by its 3D structure, assemblies of proteins into protein complexes and post-translational modifications (PTMs). Apart from primary amino acid sequence homology analysis, one can use amino acid sequences to predict protein domains (Bateman et al. 2004), protein-protein interactions (Singh et al. 2010; Planas-Iglesias et al. 2013) or subcellular localization signals in their N-terminal amino acid sequence (Dönnes and Höglund 2004). These features are informative on protein functions, but far from offering a complete picture.

Gene and protein expression data has been helpful in further developing our understanding of their functions. Typically, experiments designed to analyze differential gene and protein expression between different conditions, such as gene knock-outs or chemical perturbations, provide hints at their possible role in cellular networks. While the data from such experiments can be informative on its own, additional information can be extracted from integrating many such experiments into one analysis, commonly referred to as co-expression analysis (Serin et al. 2016).

Gene co-expression analysis is a mature subfield with impact on functional annotation of genes and gene-disease relationships (van Dam et al. 2018; Zhao et al. 2010). It offers a way of exploiting the large amounts of data deposited into public repositories such as Gene Expression Omnibus (GEO) (Barrett et al. 2013) and Proteomics Identifications (PRIDE) archive (Vizcaíno et al. 2016). The most prevalent technologies used for generating gene expression data are microarrays and RNA

sequencing (RNAseq). There are over 55.000 microarray and over 21.000 RNAseq experiments ("Series") deposited into GEO alone (state on 09.10.2018). One of the biggest efforts to integrate the gene expression data deposited in GEO is the calculation of co-expression scores between pairs of genes available as part of the STRING database (Szklarczyk et al. 2017).

However, using transcriptomics data for co-expression analysis has pitfalls. The central dogma of biology describes how genes are transcribed into transcripts which, in turn, are being translated into proteins. Transcriptomics technologies monitor transcript levels and do not provide information on their final products, proteins. Transcript levels are only partially informative on the cellular protein levels due to interplay of post-translational mechanisms such as protein synthesis and degradation (Schwanhäusser et al. 2011; Y. Liu, Beyer, and Aebersold 2016). These mechanisms also have strong impact on gene and protein co-expression. For example, in *Saccharomyces cerevisiae* microarray data, gene co-expression of many protein complexes defined by Munich Information Center for Protein Sequences (MIPS) is not much stronger than for random pairs (C.-T. Liu, Yuan, and Li 2009) (excluding large complexes such as ribosomes). Perhaps not surprisingly, monitoring protein levels proved to be more powerful than monitoring transcript levels for co-expression-based gene function prediction (Wang et al. 2017; Grabowski, Kustatscher, and Rappsilber 2018; Kustatscher, Grabowski, and Rappsilber 2017). Protein co-expression is being successfully used also in biomedical context. For example Ryan *et al.* found that genomic mutations in genes encoding one protein complex subunit often lead to downregulation of whole protein complexes (Ryan et al. 2017). Lapek *et al*. integrated protein expression data from 41 breast cancer cell lines to create a map of breast cancer cell line protein-protein associations and looked at affected submodules in the resulting functional networks (Lapek et al. 2017).

Co-expression of genes is affected by their genomic context, which includes genomic distance between genes (Hurst, Pál, and Lercher 2004; Xu, Chen, and Shen 2012; Williams and Bowles 2004; Y.-Y. Li et al. 2006; Purmann et al. 2007), epigenetic signals such as DNA methylation and histone modifications and even higher-order 3D spatial conformations of chromatin (Kustatscher, Grabowski, and Rappsilber 2017; Grabowski, Kustatscher, and Rappsilber 2018; Nora et al. 2012). These genomic effects are largely buffered on the protein level. For example, levels of protein complex subunits do not scale with gene copy number variations in yeast (Dephoure et al. 2014) and are not affected by transcript fluctuations resulting from genetic variation between individuals (Battle et al. 2015). Taken together, this suggests that using protein-level technologies, such as mass spectrometry-based proteomics, is a better choice for functional annotation using co-expression.

However, inferring gene function through protein analysis also has drawbacks. For example, it is not possible to functionally annotate genes which are not protein-coding. This means that such approaches cannot help in understanding the function of the many non-coding RNAs (ncRNAs), including regulatory microRNAs (miRNAs), which account for almost half of the known human genes (Harrow et al. 2012). Moreover, mass spectrometry suffers from sensitivity issues when dealing with complex samples such as the human proteome. Compared to RNA sequencing, which can quantify expression of tens of thousands of genes in one analysis, mass spectrometry is limited to reliably measuring only the more abundant portion of the proteome in one acquisition. Last, but not least, studying the functions of the various protein isoforms encoded by distinct transcripts on a large scale is still challenging for proteomics (Stastna and Van Eyk 2012). This is due to the often limited protein coverage (in terms of distinct peptides detected per protein) and the fact that different isoforms of a protein may differ only by few amino acids.

Analyzing large protein co-expression datasets is not simple. They often contain thousands of proteins and dozens of features (such as measured quantities and changes in experiments), forming massive expression matrices. Calculating co-expression strength between pairs of proteins is typically performed using standard statistics such as Pearson or Spearman correlation coefficients, Biweight midcorrelation, Mutual Information or simple regression models (Song, Langfelder, and Horvath 2012). These long-established methods are used to create N x N protein-protein co-expression matrices (where N is the number of proteins in the input data). Such matrices are then often used by network topology-based algorithms which help define separate co-expressed “modules” (Langfelder and Horvath 2008). Such modules can then be analyzed and screened for novel functional relationships between proteins (and thereby their respective genes).

Machine learning offers an additional mode of exploration of such co-expression datasets. Machine learning is a collective term for computer algorithms that iteratively fit a predictive model to the observed data that are of growing importance in biosciences (Huang, Chaudhary, and Garmire 2017; Camacho et al. 2018; Angermueller et al. 2016). Generally, machine learning approaches are divided into two main classes: supervised and unsupervised algorithms. Supervised algorithms (classifiers and regressors) expect a predefined target variable such as “protein is mitochondrial” vs. “protein is not mitochondrial”. Such algorithms expect training sets of positive and negative examples of the target variable and build a predictive model that can label unseen examples with class probabilities (in case of classifiers) or predicted continuous values (in case of regressors). Examples

of use of machine learning on proteomics data include integration of subcellular fractionation experiments to predict protein subcellular localization (Mulvey et al. 2017; Itzhak et al. 2016; Kustatscher, Grabowski, and Rappsilber 2016) or prediction of peptide chromatographic retention time (Giese, Ishihama, and Rappsilber 2018; Moruz, Tomazela, and Käll 2010). Conversely, unsupervised algorithms do not require a specified target variable. These algorithms are useful for finding structure within data, for example by performing clustering (such as k-means, hierarchical clustering or DBSCAN (Rehman et al. 2014) algorithms). Moreover, this class of algorithms can also calculate pairwise distances between examples (Buttrey and Whitaker 2015) or perform dimensionality reduction (such as the ubiquitous Principal Component Analysis, PCA).

Contributions and Main Findings

In this cumulative dissertation I combine four manuscripts describing primary research to which I contributed significantly (three with first authorships and one with second authorship). Two manuscripts are about the effect of genomic features on gene and protein co-expression. They essentially answer the question “Why is proteomics better than transcriptomics for gene function prediction?”. The third and fourth manuscripts build on the notion of using proteomics for gene function prediction and use machine learning with proteomics datasets for predicting functional relationships between protein-coding genes. Moreover, I add a fifth Opinion manuscript in which I offer an entry point for bench-side biologists to become oriented in the field of computational biology and data analytics. Finally, I list three manuscripts that do not form part of this thesis but to which I contributed significantly during my time as a PhD student.

The first manuscript, entitled “**Pervasive Coexpression of Spatially Proximal Genes Is Buffered at the Protein Level**” (Kustatscher, Grabowski, and Rappsilber 2017), integrated multiple published omics datasets for a lymphoblastoid cell line (LCL) panel to show that mRNA co-expression is strongly affected by genomic features. Protein co-expression, however, was not affected by genomic features and was more closely related to cellular functions than mRNA co-expression. The data ranged from transcriptomics (Pickrell et al. 2010), mass spectrometry (Battle et al. 2015) to epigenomics (ENCODE Project Consortium 2012) and Hi-C 3D genome confirmations (Rao et al. 2014). In this manuscript, I was responsible for Hi-C data analysis and correlating this data with mRNA and protein expression levels.

The second manuscript, entitled “**Epigenetic Variability Confounds Transcriptome but not Proteome Profiling for Coexpression-based Gene Function Prediction**” (Grabowski, Kustatscher, and Rappsilber 2018), in which I am the first author, was a follow-up on the findings in the previous manuscript (Kustatscher, Grabowski, and Rappsilber 2017). Here, I analyzed published mRNA (GEO, ENCODE) and protein expression (Geiger et al. 2013) datasets for mouse tissues which I integrated with epigenomics data available from the ENCODE consortium. In this manuscript, we show that the observations from human cell lines are transferable to mouse tissues. Moreover, we observed that while in the human cell line panel linear proximity between pairs of genes had a stronger impact on mRNA co-expression than epigenetic states, the converse was true for

differentiated mouse tissues. I was responsible for experiment planning, integration and analysis of the data and writing of the manuscript.

In the third manuscript, entitled "**Multiclassifier Combinatorial Proteomics of Organelle Shadows at the Example of Mitochondria in Chromatin Data**" (Kustatscher, Grabowski, and Rappsilber 2016), we looked at the usefulness of integrating published proteomics datasets for subcellular localization prediction. This proof-of-principle work showed that one can train a well-performing classifier with proteins localizing to an organelle that was not enriched in the original proteomics data. We used published interphase chromatin enrichment experiments and trained our machine learning workflow to detect patterns of mitochondrial proteins in these data, which was possible due to the non-random nature of mitochondrial protein contaminants. I was responsible for co-developing the machine learning workflow, analyzing and visualizing the data and writing of the manuscript.

The fourth manuscript, entitled "**The Human Proteome Co-Regulation Map Reveals Functional Relationships Between Proteins**" was submitted to Nature Biotechnology and is currently in revision there. I am a shared first author of this work. This manuscript expands on the idea of protein co-expression and machine learning-based integration of many published datasets. Here, we integrated published SILAC datasets from PRIDE (Vizcaíno et al. 2016) repository, arriving at a data matrix documenting expression of 10.323 proteins over 294 experimental conditions. By using a tree-based unsupervised learning approach (Buttrey and Whitaker 2015), we created a matrix of pairwise functional interaction scores. Furthermore, we improved this matrix by applying a network topology-based algorithm for re-scoring of co-expression data, the Topological Overlap Matrix (Yip and Horvath 2006). This allowed us to find a novel mitochondrion-peroxisome interface protein, PEX11B, formerly annotated as peroxisomal only. Moreover, we show functional relationships of many microproteins for which there is very limited functional annotation, as they generate very few peptides observable by mass spectrometry. Our protein-protein co-expression scores are being currently integrated into the STRING database and will be officially released as part of the STRING version 11. I was responsible for testing and comparing multiple statistical and machine learning approaches and data dimensionality reduction techniques. I created a web application and a web server which constitute the gateway to our resources (available at https://www.proteomehd.net/). Moreover, I was involved in multiple data analysis and visualization stages.

The fifth manuscript, entitled **"A Primer on Data Analytics in Functional Genomics: How to Move from Data to Insight?**" is an Opinion paper published in the peer-reviewed review journal Trends in

Biochemical Sciences. It contains a primer for early stage researchers and students who are predominantly wet-lab oriented and are interested in learning more about analyzing large biological datasets.

Finally, three other manuscripts which I first- or co-authored, but which are not part of this thesis are:

1. **“Proteome Analysis of Human Neutrophil Granulocytes from Patients with Monogenic Disease”,** in which I analyze proteomes from neutrophils of patients with rare monogenic diseases and show that data-independent acquisition (DIA) proteomics can aid genetic medical diagnostics. The manuscript is currently under peer review at Molecular and Cellular Proteomics. I am the first shared author on this manuscript and was the lead driver of the project.
2. “**Machine-Learning Captures Higher-Order Modular Architecture of the Proteome**”, currently in writing. In this manuscript we show that the human proteome is organized into distinct higher-order functional modules, detectable using machine learning and large proteomics datasets. Here, I was mainly responsible for developing a supervised machine-learning workflow running on the computer cluster of the University of Edinburgh.
3. “**The treeClust Algorithm Improves Coexpression Analysis in Large Datasets**”, currently in writing. In this technical manuscript we show how the unsupervised learning algorithm, treeClust, can improve co-expression analysis in large protein expression datasets. We show how it handles outliers, weak and strong correlations, missing values and other properties of large protein expression datasets. We propose treeClust as an attractive alternative to more classical statistical approaches, such as Pearson correlation, for building protein functional association scores.

Manuscript 1. “Pervasive Coexpression of Spatially Proximal Genes Is Buffered at the Protein Level”

Pages 16 - 29

Manuscript available online, DOI: 10.15252/msb.20177548

Published online: August 23, 2017

Article

TRANSPARENT PROCESS OPEN ACCESS

molecular systems biology

Pervasive coexpression of spatially proximal genes is buffered at the protein level

Georg Kustatscher[1,*], Piotr Grabowski[2] & Juri Rappsilber[1,2,**]

Abstract

Genes are not randomly distributed in the genome. In humans, 10% of protein-coding genes are transcribed from bidirectional promoters and many more are organised in larger clusters. Intriguingly, neighbouring genes are frequently coexpressed but rarely functionally related. Here we show that coexpression of bidirectional gene pairs, and closeby genes in general, is buffered at the protein level. Taking into account the 3D architecture of the genome, we find that co-regulation of spatially close, functionally unrelated genes is pervasive at the transcriptome level, but does not extend to the proteome. We present evidence that non-functional mRNA coexpression in human cells arises from stochastic chromatin fluctuations and direct regulatory interference between spatially close genes. Protein-level buffering likely reflects a lack of coordination of post-transcriptional regulation of functionally unrelated genes. Grouping human genes together along the genome sequence, or through long-range chromosome folding, is associated with reduced expression noise. Our results support the hypothesis that the selection for noise reduction is a major driver of the evolution of genome organisation.

Keywords gene expression noise; genome organisation; proteomics; regulatory interference; transcriptomics
Subject Categories Chromatin, Epigenetics, Genomics & Functional Genomics; Genome-Scale & Integrative Biology
DOI 10.15252/msb.20177548 | Received 19 January 2017 | Revised 21 July 2017 | Accepted 24 July 2017
Mol Syst Biol. (2017) 13: 937

Introduction

The position of genes in the human genome is not random (Hurst *et al*, 2004). Genes are often found in pairs or larger clusters that tend to be coexpressed (Caron *et al*, 2001; Lercher *et al*, 2002; Trinklein *et al*, 2004). Some of these coordinate transcription of genes with related functions, for example histone genes and other clusters resulting from gene duplication. However, the majority of closeby, coexpressed human genes appear not to have a higher functional similarity than random gene pairs (Hurst *et al*, 2004; Williams & Bowles, 2004; Li *et al*, 2006; Purmann *et al*, 2007; Michalak, 2008; Xu *et al*, 2012). For example, 35 DNA repair genes are transcribed from bidirectional promoters, but none of their paired genes is involved in DNA repair (Xu *et al*, 2012). This raises intriguing questions: Why are functionally unrelated genes clustered in the genome and how can the cell tolerate their coexpression?

Pioneering work in yeast identified the selection for reduced gene expression noise as a key driver for the evolution of chromosome organisation (Batada & Hurst, 2007; Wang *et al*, 2011). A major cause of gene expression noise is thought to be the random fluctuation of chromatin domains between an active and inactive state, causing mRNAs to be synthesised in short, stochastic bursts (Raj *et al*, 2006). Clusters of active genes may mutually reinforce their open chromatin state, minimising stochastic chromatin remodelling, and thereby reduce expression noise (Batada & Hurst, 2007; Wang *et al*, 2011). Similarly, genes flanking bidirectional promoters have lower expression noise than other genes, even if one of the divergent partners is a noncoding RNA (Wang *et al*, 2011). Noise-sensitive genes, such as those encoding protein complex subunits, are enriched among bidirectional pairs, but neither in yeast nor in human do any of these pairs encode two subunits of the same protein complex (Li *et al*, 2006; Wang *et al*, 2011). Consequently, it has been suggested that bidirectional promoters may drive noise reduction rather than the coexpression of functionally related genes (Wang *et al*, 2011).

The noise reduction model not only provides a potential explanation for the occurrence of clusters of functionally unrelated genes, but also predicts that such genes may be coexpressed (Wang *et al*, 2011). In yeast, chromatin-modifying enzymes are major contributors to gene expression noise (Newman *et al*, 2006) and chromatin remodelling drives the incidental coexpression of neighbouring, functionally unrelated genes (Batada *et al*, 2007). This coexpression may be due to a passive mechanism, whereby random transitions between open and closed chromatin simultaneously expose all genes within a chromatin domain to the transcriptional machinery. Alternatively, for very close genes such as those with bidirectional promoters, the up- or downregulation of one gene may directly affect the transcriptional status of its neighbour (Wang *et al*, 2011). Indeed, such a "ripple effect" of transcriptional activation has been observed in yeast and humans (Ebisuya *et al*, 2008). The noise and

1 Wellcome Trust Centre for Cell Biology, University of Edinburgh, Edinburgh, UK
2 Chair of Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, Berlin, Germany
*Corresponding author. Tel: +44 131 6517057; E-mail: georg.kustatscher@ed.ac.uk
**Corresponding author. Tel: +44 131 6517056; E-mail: juri.rappsilber@ed.ac.uk

expression levels of transgenes also vary with their insertion site, as a result of both domain-wide effects and interference with individual neighbouring genes (Gierman *et al*, 2007; Chen & Zhang, 2016). Transgenes can also affect the mRNA expression levels of endogenous genes located close to the insertion site (Akhtar *et al*, 2013).

If the transcription of noise-reduced, clustered genes is unduly influenced by their neighbours, how can individual genes reach their optimal expression levels? Notably, gene expression is usually measured at the mRNA level. However, protein levels are buffered against certain transcript fluctuations (Liu *et al*, 2016), such as those caused by stochastic transcription initiation (Raj *et al*, 2006; Gandhi *et al*, 2011) and genetic variation between individuals (Battle *et al*, 2015) and species (Khan *et al*, 2013). The abundance of some proteins can also be buffered against gene copy number variations (Geiger *et al*, 2010; Stingele *et al*, 2012; Dephoure *et al*, 2014). We therefore speculated that protein abundances may also be buffered against regulatory interference between genes in close spatial proximity.

Results

Coexpression of bidirectional gene pairs is buffered at the protein level

We investigated the expression of 4,188 genes across 60 different human lymphoblastoid cell lines (LCLs), for which mRNA (Pickrell *et al*, 2010) and protein abundances (Battle *et al*, 2015) have been reported (Fig 1A, Dataset EV1). These genes are highly expressed in all human tissues and their promoters are in active chromatin states (Appendix Fig S1). Although constitutively active, expression levels of these "housekeeping" genes vary between LCLs, as a result of genetic and other differences, including age and growth conditions (Akey *et al*, 2007; Stark *et al*, 2014; Yuan *et al*, 2015). The LCL cell line panel has been instrumental in identifying expression quantitative trait loci, that is DNA sequence variants that specifically influence the expression level of one or more genes (Albert & Kruglyak, 2015). Here, instead of assessing how a gene's expression level depends on the genotype, we analyse how it is influenced by the expression of other, closeby genes. LCLs are a valuable test system as their genome structure and regulatory elements have been mapped at unparalleled resolution (Lieberman-Aiden *et al*, 2009; Ernst *et al*, 2011; ENCODE Project Consortium, 2012; Rao *et al*, 2014).

First, we analysed gene pairs that are transcribed from bidirectional promoters. These are commonly defined as genes that are found in head-to-head orientation with < 1 kb between their transcription start sites (TSSs) (Trinklein *et al*, 2004). Out of 167 such gene pairs in this dataset, the mRNA abundances of 31 (19%) are strongly and significantly co-regulated across LCLs (Pearson's correlation coefficient, PCC > 0.5, BH-adjusted *P*-value < 0.001). However, protein co-regulation is attenuated or buffered for 28 of these (Fig 1B, Appendix Table S1). Literature analysis revealed that the buffered gene pairs generally have unrelated biological functions, in contrast to the three gene pairs whose co-regulation is sustained at the protein level (Appendix Table S1).

We next considered the 929 non-bidirectional gene pairs with up to 50 kb between their TSSs, regardless of their orientation (Dataset EV2). Although these pairs do not share a promoter region, we find that 22% have co-regulated mRNA abundances (PCC > 0.5, BH-adjusted $P < 0.001$). However, only 3% are also co-regulated at the protein level (Fig 1B).

Genes with similar functions have co-regulated mRNA and protein abundances

To confirm that the different impact of gene proximity on mRNA and protein abundances reflects a biological phenomenon, rather than simply a difference in data quality, we assessed the co-regulation of genes with known functional links, irrespective of their genomic position. We analysed subunits of the same protein complex, enzymes catalysing consecutive reactions in metabolic pathways and proteins with identical subcellular localisations. In all cases, we observe strong co-regulation on mRNA and protein levels, but co-regulation of proteins is significantly stronger than that of mRNAs (Fig EV1, $P < 3 \times 10^{-16}$). Therefore, data quality appears not to be limiting. Instead, the observed differences between mRNA and protein co-regulation indicate that post-transcriptional processes eliminate co-regulation of genes which are related spatially, but not functionally.

A fraction of closeby genes is enriched for similar functions

Our observation that only 3% of closeby genes have co-regulated protein abundances appears to contrast with the fact that genes in close genomic proximity are enriched for similar functions (Thévenin *et al*, 2014). However, functional enrichment does not exclude the possibility that the bulk of closeby gene pairs does not share similar functions. For example, we find that co-regulation of transcripts and proteins from closeby genes is more common than for random protein pairs (Fig 1B), and this enrichment is highly significant (3% versus 0.4%, $P < 4 \times 10^{-14}$).

To analyse the relationship between gene distance and function more systematically, we assessed functional associations between our gene pairs using the STRING database (Szklarczyk *et al*, 2017). We considered gene pairs to be functionally associated if their STRING score, that is the likelihood of the association to be biologically meaningful, specific and reproducible, was > 0.7. Using this comprehensive definition, we find that 4.5% of closeby gene pairs, that is those with < 50 kb between their TSSs, are related functionally (Fig EV2A). As observed by Thévenin *et al*, we find this to be a significant enrichment over gene pairs that are farther apart. Likewise, gene pairs from the same chromosome are enriched for similar functions relative to those from different chromosomes. Nevertheless, the extent of mRNA co-regulation (22%) strongly exceeds co-function, and mRNA co-regulation of most closeby gene pairs is not sustained at the protein level (Fig EV2A).

Notably, a similar analysis in yeast has shown that adjacent genes tend to have correlated mRNA expression and are statistically enriched for similar functions (Cohen *et al*, 2000). However, in striking agreement with our observations, only about 2% of these coexpressed neighbouring gene pairs have related functions (Batada *et al*, 2007) and only for these is gene order evolutionarily conserved (Hurst *et al*, 2002). Coexpression of neighbouring genes has also been observed in *Arabidopsis thaliana*, but only a fraction

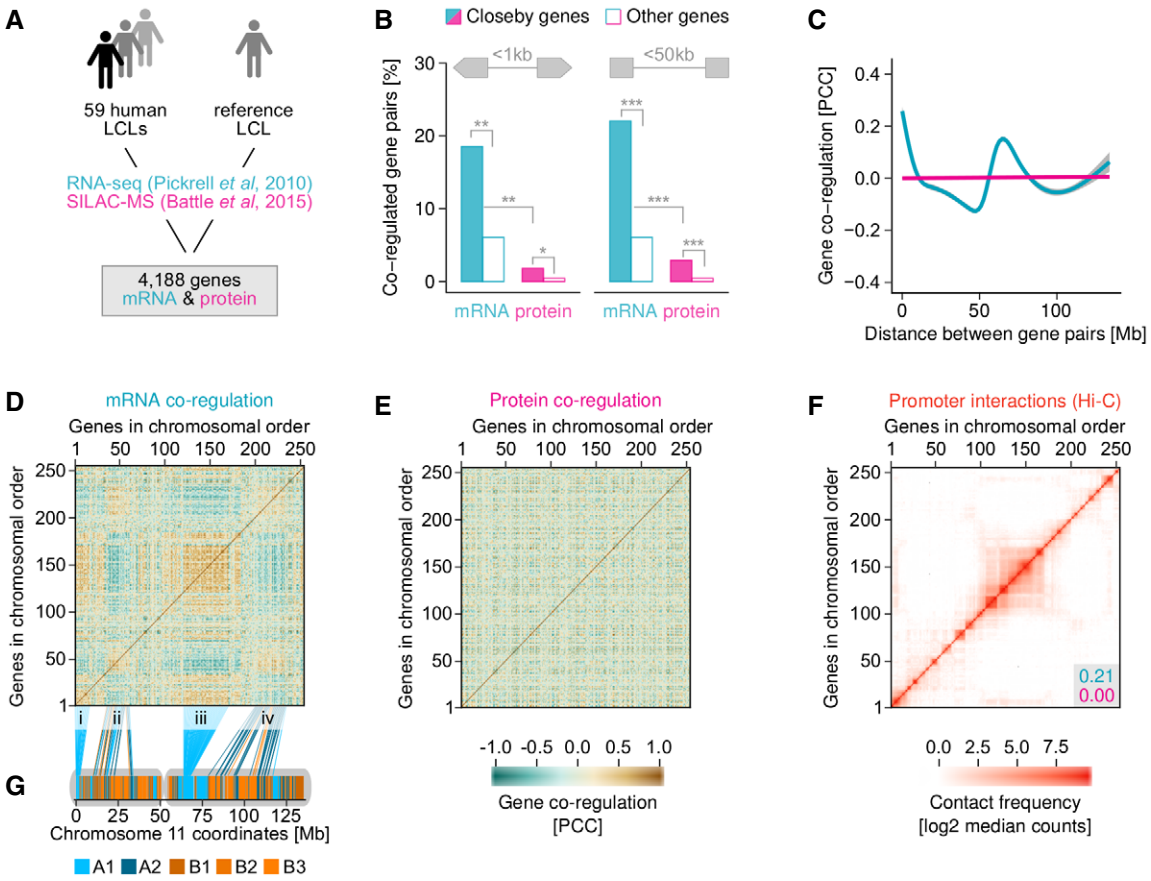


Figure 1. Spatial proximity of genes affects mRNA but not protein regulation.

A We analysed previously reported mRNA and protein abundances in 59 lymphoblastoid cell lines (LCLs), relative to a reference sample.

B Genes transcribed from bidirectional promoters frequently have co-regulated mRNA abundances, but only a fraction of these also have co-regulated protein abundances (left). The same is true for non-bidirectional gene pairs whose transcription start sites (TSS) are < 50 kb apart, irrespective of their orientation (right) (*$P < 0.05$, **$P < 2 \times 10^{-7}$, ***$P < 4 \times 10^{-14}$ based on Fisher's exact test).

C mRNA co-regulation of gene pairs on chromosome 11 decreases with chromosomal distance over many megabases, but not monotonously. Protein co-regulation is unaffected by genomic distance.

D mRNA co-regulation map for chromosome 11 showing large patches of co-regulated (brown) and anti-regulated (blue) gene pairs. Four large, co-regulated patches are highlighted (i–iv).

E No regulation patches exist on the protein level.

F mRNA co-regulation patches partially coincide with physical associations between genes derived from Hi-C data (Rao *et al*, 2014). Numbers in grey box show the Pearson correlation between the Hi-C map and mRNA (blue) or protein (red) co-regulation maps.

G Patches i, iii and ii, iv broadly coincide with genome subcompartments A1 and A2, respectively.

of the observed cases could be explained through a shared function (Williams & Bowles, 2004).

Long-range gene co-regulation leads to coordinated mRNA but not protein expression

The influence of gene distance on co-regulation of transcripts is not limited to genes in close proximity. As seen in the example of chromosome 11, mRNA co-regulation extends over many megabases but does not affect protein abundances (Fig 1C). Although co-regulation generally declines with increasing gene distance, such long-range effects are unlikely to result from transcriptional interference *in cis*. A major co-regulation peak of genes that are more than 50 Mb apart on chromosome 11 suggests that long-range chromosome folding may be involved. In agreement with this, all chromosomes have distinct co-regulation curves (Appendix Fig S2).

The co-regulation map of chromosome 11 shows large patches of genes whose transcripts are coordinately up- and downregulated (Fig 1D). Importantly, no corresponding co-regulation is observed on the protein level (Fig 1E). However, the mRNA co-regulation map shows a striking similarity to physical associations observed for our gene set, as extracted from existing Hi-C data (Rao *et al*, 2014; Fig 1F). The Hi-C contact matrix of chromosome 11 is correlated with the mRNA co-regulation map (PCC 0.21, $P < 2 \times 10^{-318}$), but not the protein map (PCC 0.00, $P = 0.4$). Similar mRNA co-regulation patches can be observed on other chromosomes (Fig EV3) as well as between different chromosomes (Fig EV4). Generally, both intra- and interchromosomal co-regulation patches

correspond to areas with increased Hi-C contacts (Appendix Table S2). Some chromosomes have more prominent patches than others (Fig EV3). Chromosome 19, which is short but exceptionally gene-dense, is unique in forming a single large co-regulation patch (Fig EV3C). Importantly, none of these mRNA co-regulation patches are reflected at the protein level (Figs EV3 and EV4, Appendix Fig S2). This suggests that regulatory interference between genes that are close in 3D could be associated with similar non-functional mRNA co-regulation as observed for neighbouring genes in the genome sequence.

We next sought to determine which structural features of the genome give rise to mRNA co-regulation patches. Four large mRNA co-regulation patches can be observed on chromosome 11 (labelled i–iv in Fig 1D). Co-regulation patches differ widely in size but often span many megabases, likely reflecting broad architectural features. Notably, promoters and enhancers typically interact on a smaller scale, within topologically associated domains (Gibcus & Dekker, 2013). However, co-regulated groups of genes are more reminiscent of genome compartments. Genome compartments were first identified on the basis of long-range interactions mapped by Hi-C, which showed that open and closed chromatin spatially segregate into two genome-wide compartments (Lieberman-Aiden *et al*, 2009). The compartments containing active and repressive chromatin were designated A and B, respectively. A high-resolution Hi-C map of the genome in LCLs subsequently identified that these compartments segregate further into six subcompartments: A1-2 and B1-4 (Rao *et al*, 2014). Genomic loci within each subcompartment tend to be associated with each other more often than with loci from other subcompartments, that is they are in closer spatial proximity. We find that co-regulation patches i and iii of chromosome 11 align with subcompartment A1 and patches ii and iv align with subcompartment A2 (Fig 1G). These are the two subcompartments of the genome formed by transcriptionally active chromatin, which is expected given that we analyse housekeeping genes. Interestingly, genes across patches i and iii are co-regulated, as are genes across patches ii and iv, suggesting that co-localisation in subcompartments may contribute to the existence of these patches.

Genes with co-regulated mRNAs co-localise in genome subcompartments

To assess systematically the overlap of co-regulated gene groups with genome compartments, we clustered genes by co-regulation. We found four transcriptome regulation groups T1-4 (Fig 2A and Dataset EV3), explaining more than 50% of the total variance (Appendix Fig S3). Transcripts within each group are co-regulated (Fig 2A and B). Genes from T1 and T2 are strongly enriched for subcompartments A2 and A1, respectively (Fig 2C). Curiously, they are anti-correlated, that is when T1 genes are upregulated, T2 genes tend to be downregulated, and vice versa (Fig 2B). Co-regulated genes of the T3 and T4 groups are also enriched for A1 and A2 subcompartments, respectively. However, they are independent of T1 and T2, that is there is neither a positive nor a negative correlation between T1/T2 and T3/T4 (Fig 2B). Therefore, while subcompartments A1 and A2 are strongly related to transcriptome regulation groups, they are not sufficient to explain them.

Genome compartments and subcompartments were defined solely based on their physical interaction patterns, but also have

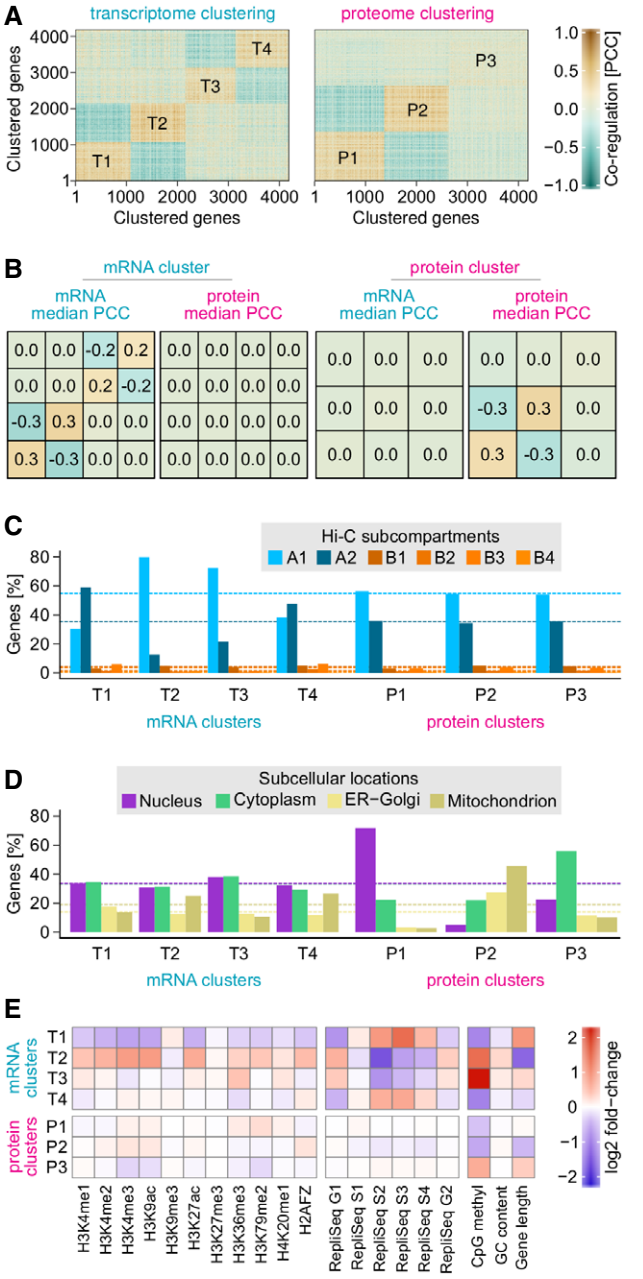


Figure 2. Transcriptome and proteome regulation are driven by different factors.

A *k*-means clustering of genes based on their mRNA or protein abundance changes across LCLs.

B Median Pearson's correlation coefficients (PCCs) for each transcriptome and proteome *k*-means cluster. Genes assigned to different *k*-means clusters can either be anti-regulated (e.g. T1 and T2) or not correlated (e.g. T1 and T3). *k*-means clusters formed by genes that are co-regulated at the mRNA level are not generally co-regulated at the protein level, and vice versa.

C Transcriptome clusters are strongly enriched for subcompartment A1 or A2. Dashed lines indicate the percentage of genes expected if subcompartments were evenly distributed across clusters.

D Proteome clusters are mainly composed of proteins from distinct subcellular locations. Dashed lines indicate the percentage of genes expected if subcellular locations were evenly distributed across clusters.

E Genomic and epigenomic features enriched in each cluster relative to the whole dataset.

different genomic and epigenomic characteristics. A1 and A2 subcompartments are both enriched for features associated with transcriptionally active chromatin, but to different extents (Rao *et al*, 2014). Interestingly, we also found clear differences in histone modifications and DNA methylation associated with transcriptome regulation groups (Fig 2E). For example, in comparison with T2, T1 gene bodies are enriched for H3K9me3, depleted in activating marks such as H3K4me3 and H3K27ac, are longer, replicate later and have a lower GC content. These differences mirror those observed between A2 and A1 subcompartments (Rao *et al*, 2014). In contrast, T3 and T4 do not show these features despite preferentially localising to A1 and A2 subcompartments. Instead, T3 genes display heavy CpG methylation, which is almost an order of magnitude stronger than for T4 genes. Consequently, T3 and T4 define their own epigenetic subpopulation within A-type compartments.

Genes with co-regulated protein abundances are related functionally, not spatially

Clustering analysis of protein expression profiles led to three proteome regulation groups P1-3 (Fig 2A and Dataset EV3), explaining more than 50% of the total variance (Appendix Fig S3). Neither genome compartments nor epigenomic signatures appear to be associated with proteome regulation groups (Fig 2C and E). In contrast, proteome regulation groups broadly correspond to subcellular locations: nucleus (P1), mitochondria, ER and Golgi (P2) and cytoplasm (P3) (Fig 2D). They are also enriched for biological processes taking place in these subcellular locations (Appendix Fig S4). In contrast, T1-4 only weakly coincide with subcellular locations or biological processes.

Intriguingly, T1-4 and P1-3 are independent of each other, that is genes that are clustered based on their transcript expression signature are generally not co-regulated on the protein level, and vice versa (Fig 2B). This suggests that much of the mRNA coexpression of genes from the same subcompartment may be non-functional. Note that as for sequence proximity (see above), this appears to contrast with a previous report that genes which are close in 3D nuclear space often have similar functions (Thévenin *et al*, 2014). However, we also find significant enrichment of functional associations between genes from the same subcompartment (Fig EV2B). Nevertheless, in quantitative terms, the extent of mRNA co-regulation strongly exceeds co-function as well as protein co-regulation. For example, while 11% of gene pairs in the same (intrachromosomal) subcompartment have co-regulated mRNAs, < 1% have similar functions according to STRING and are co-regulated at the protein level (Fig EV2B).

Gene clustering within but not between chromosomes associates with reduced expression noise

In yeast, clustering of genes in the genome sequence is associated with reduced expression noise (Batada & Hurst, 2007; Wang *et al*, 2011). However, the situation is more complex when considering the 3D structure of the genome. Highly transcribed gene clusters tend to form fewer contacts with other chromosomes, and genomic loci with more interchromosomal contacts tend to have higher expression noise (McCullagh *et al*, 2010; Sandhu, 2012).

We tested whether gene clustering has a similar effect in human cells. For each gene in our dataset, we calculated a clustering degree, defined as the average distance to its three nearest neighbouring genes along the DNA sequence. We then compared the expression noise of the 5% most and least clustered genes, respectively. As observed in yeast, we find that gene expression noise in LCLs is significantly reduced for genes in gene-dense areas (Fig 3A). The noise-reducing effect is much more significant on the mRNA than the protein level.

In a second step, we investigated whether gene clustering in nuclear space has a similar noise-reducing effect. In principle, gene-dense regions may interact with each other in 3D to benefit from further noise reduction by forming "super-clusters". The three human histone gene clusters on chromosome 6, for example, converge in 3D to form such a super-cluster (Sandhu *et al*, 2012). Therefore, we calculated a second clustering degree for each gene, defined as the average distance to its three nearest neighbours in 3D, using Hi-C contacts. To capture long-range interactions resulting from chromosome folding, we only considered neighbouring genes that were on the same chromosome, but at least 500 kb up- or downstream in terms of DNA sequence. There is a positive correlation between the clustering degree in 1D and 3D (PCC 0.32, $P < 6 \times 10^{-97}$), suggesting that genes clustered along the sequence are also more densely packed in the 3D structure of a chromosome. Moreover, this gene clustering due to chromosome folding is also associated with a significant reduction of gene expression noise, albeit not as strongly as sequence-based clusters (Fig 3A).

Next, we investigated clusters that genes from different chromosomes may form in nuclear space, calculating a third clustering degree based on interchromosomal Hi-C contacts. As shown in yeast (McCullagh *et al*, 2010; Sandhu, 2012), we find a negative correlation between sequence-based and interchromosomal clustering (PCC -0.1, $P < 5 \times 10^{-11}$). This suggests that gene-dense regions, while forming long-range, noise-reducing interactions within the same chromosome, are less likely to interact with gene clusters on a different chromosome. Moreover, genes forming interchromosomal clusters are associated with higher expression noise than those with fewer interactions (Fig 3A). This difference is not statistically significant but is in agreement with earlier findings in yeast (McCullagh *et al*, 2010; Sandhu, 2012).

Coexpression of closeby genes is driven by stochastic epigenetic fluctuations and regulatory interference

How can gene proximity lead to mRNA coexpression? Many incidents of coexpressed genes that are close in sequence have been linked to stochastic alternation between an active and inactive chromatin state (Batada *et al*, 2007). Such chromatin fluctuations can lead to coordinated transcriptional bursts of all genes within a chromatin domain (Raj *et al*, 2006). We first compared the chromatin environment of genes that are co-regulated with their sequence neighbours with genes that show no such co-regulation ("neighbours" being defined as genes whose TSSs are < 50 kb away). We find that genes which are coexpressed with their neighbours are more often flanked by heterochromatin, upstream of their transcription start site (Fig 3B). This is consistent with mRNA coexpression driven by stochastic spreading of the adjacent heterochromatin domain into the active locus, silencing all genes therein. This is

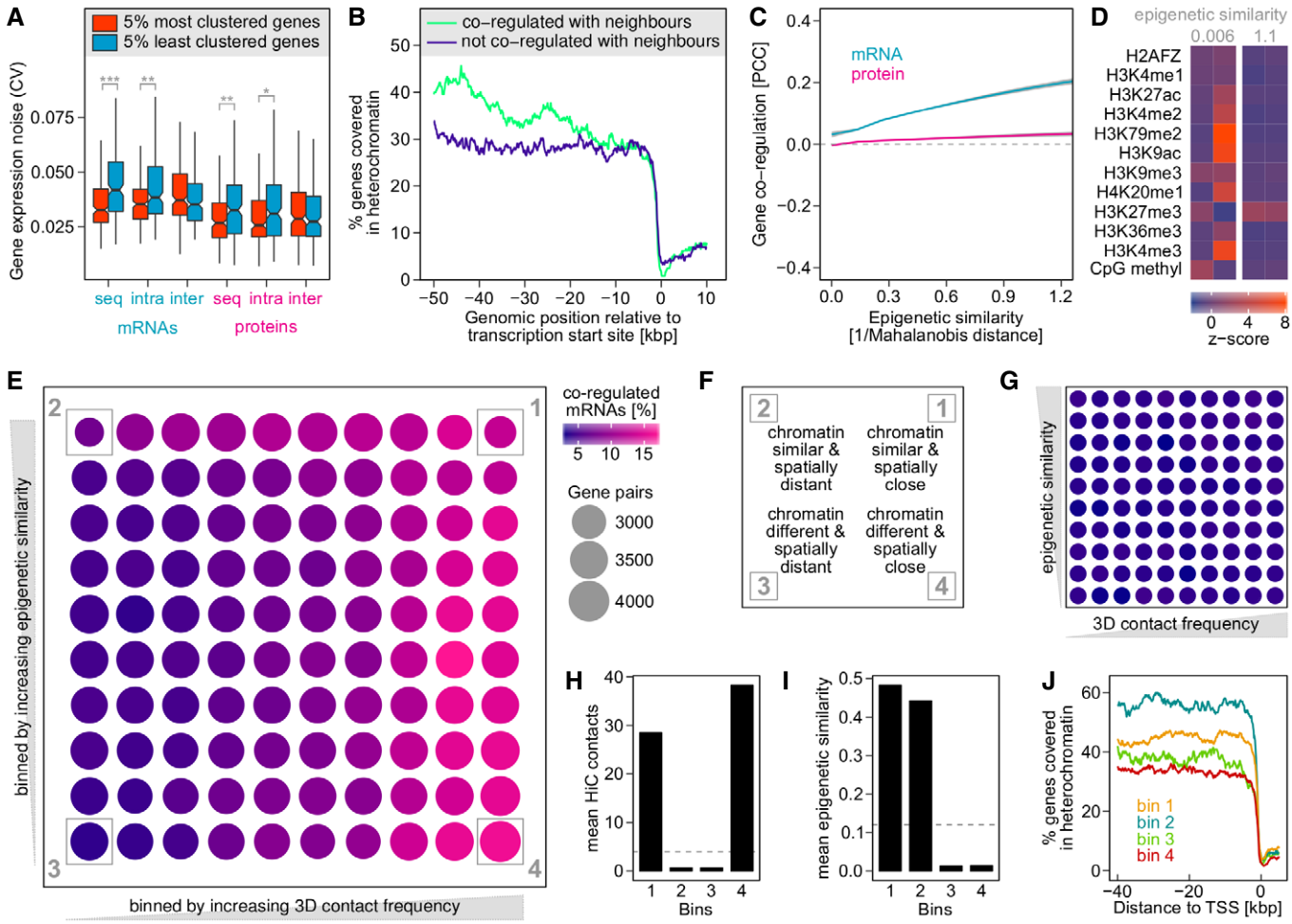


Figure 3. mRNA coexpression of neighbouring genes is driven by chromatin fluctuations and regulatory interference.

A Intrachromosomal gene clustering reduces gene expression noise. We determined the expression noise (coefficient of variation, CV) of the most and least densely clustered genes, considering three different types of clustering: in terms of sequence proximity (seq), using long-range Hi-C contacts (> 500 kb) within the same chromosome (intra) and using interchromosomal Hi-C contacts (inter). Expression noise is reduced for clustered genes, except for genes forming more interchromosomal contacts (*$P < 0.01$, **$P < 0.002$, ***$P < 5 \times 10^{-6}$ based on Kolmogorov–Smirnov test). Boxplot drawn in the style of Tukey, that is box limits indicate the first and third quartiles, central lines the median, whiskers extend 1.5 times the interquartile range from the box limits. Notches indicate the 95% confidence interval for comparing medians.

B The upstream region of genes that are co-regulated with their neighbours, that is other genes within 50 kb, is more likely to be occupied by heterochromatin than that of genes showing no such co-regulation. Heterochromatin regions in LCLs have been reported previously (Ernst *et al*, 2011).

C Epigenetic similarity calculated on the basis of histone marks and CpG methylation is a strong general predictor of mRNA co-regulation. Curves are fitted to all intrachromosomal gene pairs irrespective of their genomic distance.

D Two randomly picked gene pairs exemplifying low and high epigenetic similarity, respectively. Each column represents a gene and each row an epigenetic feature. Colours show the standardised, average abundance of each mark across the gene body.

E mRNA co-regulation requires epigenetic similarity or spatial proximity, but not both. Intrachromosomal gene pairs were binned by epigenetic similarity and spatial proximity (Hi-C contacts), and the percentage of co-regulated mRNAs is shown in colour. Note bins 2 and 4 are both enriched for co-regulated mRNAs despite containing gene pairs that are spatially distant and epigenetically different, respectively.

F Description of bins highlighted in panel (E).

G Gene pairs binned as in (E) but colour showing percentage of co-regulated proteins. Protein co-regulation does not depend on epigenetic similarity or spatial proximity.

H On average, gene pairs in bins 1 and 4 have many more Hi-C contacts than those in bins 2 and 3, that is they are spatially closer. Dashed line shows average Hi-C contacts between genes in the dataset.

I On average, gene pairs in bins 1 and 2 are epigenetically much more similar than those in bins 3 and 4. Dashed line shows average epigenetic similarity between genes in the dataset.

J Heterochromatin profile for genes in bins 1–4.

reminiscent of subtelomeric regions in yeast, which are hot spots for expression noise (Batada & Hurst, 2007) due to transient spreading of telomeric heterochromatin (Anderson *et al*, 2014).

Notably, chromatin fluctuations may lead to mRNA coexpression that is not restricted to genes in close spatial proximity. Chromatin factors play a key role in creating gene expression noise (Newman

et al, 2006). Fluctuating expression levels of, for example, a histone-modifying enzyme may simultaneously affect all its target chromatin domains in the genome. To test for such a global chromatin-mediated co-regulation effect, we determined the epigenetic similarity between all genes in our dataset. We defined "epigenetic similarity" based on the abundance of various histone marks within gene bodies. We used the Mahalanobis distance to measure similarity, as this takes into account that some histone marks are strongly co-dependent, for example H3K9ac and H3K4me3. Genes with similar epigenetic profiles are targeted by a similar set of chromatin-modifying factors, and are therefore expected to respond similarly to stochastic fluctuations of these factors. Indeed, we find that the epigenetic similarity is a strong predictor of non-functional mRNA co-regulation (Fig 3C and D).

This chromatin fluctuation scenario is a passive mechanism where genes simply respond to changes in their chromatin domain. However, on a local scale, transcriptional changes of one gene may directly affect the transcription of its neighbours, if chromatin remodelling or transcription factors spill over to adjacent genomic regions (Ebisuya *et al*, 2008; Wang *et al*, 2011). This "regulatory interference" model crucially depends on spatial proximity, but does not require co-regulated genes to be part of the same chromatin domain. To compare the impact of chromatin and gene distance on non-functional mRNA coexpression, we grouped gene pairs based on epigenetic similarity as well as based on Hi-C contact frequency. We then observed which groups contain co-regulated mRNAs (Fig 3E). This shows that gene pairs which are far apart both spatially and epigenetically are rarely co-regulated (bin 3 in Fig 3E and F). Gene pairs with similar histone marks tend to be co-regulated, even if they are spatially distant (Fig 3E and H). Co-regulation of such genes is consistent with the passive chromatin fluctuation model, but not the transcriptional interference model. Importantly, spatially close gene pairs can be co-regulated even if their histone marks show no similarity (bin 4 in Fig 3E and I). This type of coexpression is not consistent with the passive chromatin fluctuation model, since the epigenetic differences between the gene pairs suggest that, in steady state, they occupy distinct chromatin domains. These genes are also the least likely to be flanked by heterochromatin (Fig 3J). However, the behaviour of gene pairs in bin 4 is consistent with the regulatory interference model, where fluctuations in one gene affect the chromatin and transcriptional state of its neighbours, in sequence and 3D. Note that this effect is buffered at the protein level (Fig 3G), which is in agreement with this type of coexpression being not functional.

Buffering of non-functional mRNA coexpression tends to be a non-selective process

Finally, we asked which post-transcriptional mechanisms might buffer the coexpression of genes that are spatially close, but functionally unrelated. In principle, this could be a selective process that specifically targets closeby genes and disentangles their expression patterns. Alternatively, buffering could be a neutral process, where the lack of coordination between post-transcriptional mechanisms prevents the mRNA coexpression to be propagated to the protein level. In this case, a selective process would need to exist to ensure that functionally related genes do in fact have co-regulated protein abundances. To distinguish between these two possibilities, we analysed five measures of post-transcriptional gene expression control (Fig 4).

First, we tested whether gene pairs with sustained protein co-regulation are more likely to have similar mRNA half-lives in LCLs (Duan *et al*, 2013), relative to co-regulated gene pairs with buffered protein abundances. Indeed, we find this to be the case, even though the difference is modest (Fig 4A). Next, we analysed which co-regulated gene pairs are more likely to be targeted by the same miRNA (Helwak *et al*, 2013). Again, gene pairs that are also co-regulated on the protein level are enriched for pairs sharing at least one miRNA. Third, as an indication for translation-related effects, we took into account ribosome profiling data for the LCL cell line panel (Battle *et al*, 2015), which reflect both the abundance of mRNAs and the extent to which they are occupied by ribosomes (Ingolia, 2014). Gene pairs with coexpressed proteins are almost three times as likely to have correlated ribosome profiles than pairs which only have co-regulated mRNA abundances. Then, we looked at the impact of protein degradation, by considering the occurrence of non-exponentially degraded proteins (NEDs) (McShane *et al*, 2016). These are proteins that are rapidly degraded after synthesis, for example because they are protein complex subunits produced in super-stoichiometric amounts. Again, we find that NEDs are enriched among gene pairs with co-regulated proteins rather than those with buffered protein levels. Finally, we show that the protein sequence length, which strongly correlates with the extent of post-transcriptional control (Vogel *et al*, 2010), is more similar for co-regulated than buffered proteins. Proximity in the genome seemed to have no impact on the similarity of gene pairs in any of the five measures of post-transcriptional gene expression control investigated here (Fig 4B). Taken together, these results suggest that buffering of co-regulated closeby genes may occur via a neutral mechanism, with buffered gene pairs consistently lacking the extent of shared post-transcriptional processing observed for functionally related gene pairs. If mRNA coexpression is functionally relevant, multiple layers of post-transcriptional control appear to work together to ensure that this is propagated to the protein level.

Discussion

Genes are not randomly distributed across the sequence and structure of the genome, forming clusters that tend to be coexpressed but do not generally have a shared function. Gene expression noise is detrimental to cell fitness, especially for housekeeping genes (Fraser *et al*, 2004). Clusters of actively transcribed genes have low expression noise, which may drive the evolution of non-random gene order (Batada & Hurst, 2007). The coexpression of functionally unrelated neighbouring genes may then be a side effect of the selection for noise reduction. However, such coexpression is not necessarily deleterious. As we show here, non-functional co-regulation is frequently observed at the mRNA level, but is largely buffered at the protein level. Consequently, non-functional coexpression is unlikely to offset the benefit of noise reduction.

The expression profiles of genes in a cluster co-evolve, such that the evolutionary change in expression of one gene on average predicts changes in its neighbours (Ghanbarian & Hurst, 2015).

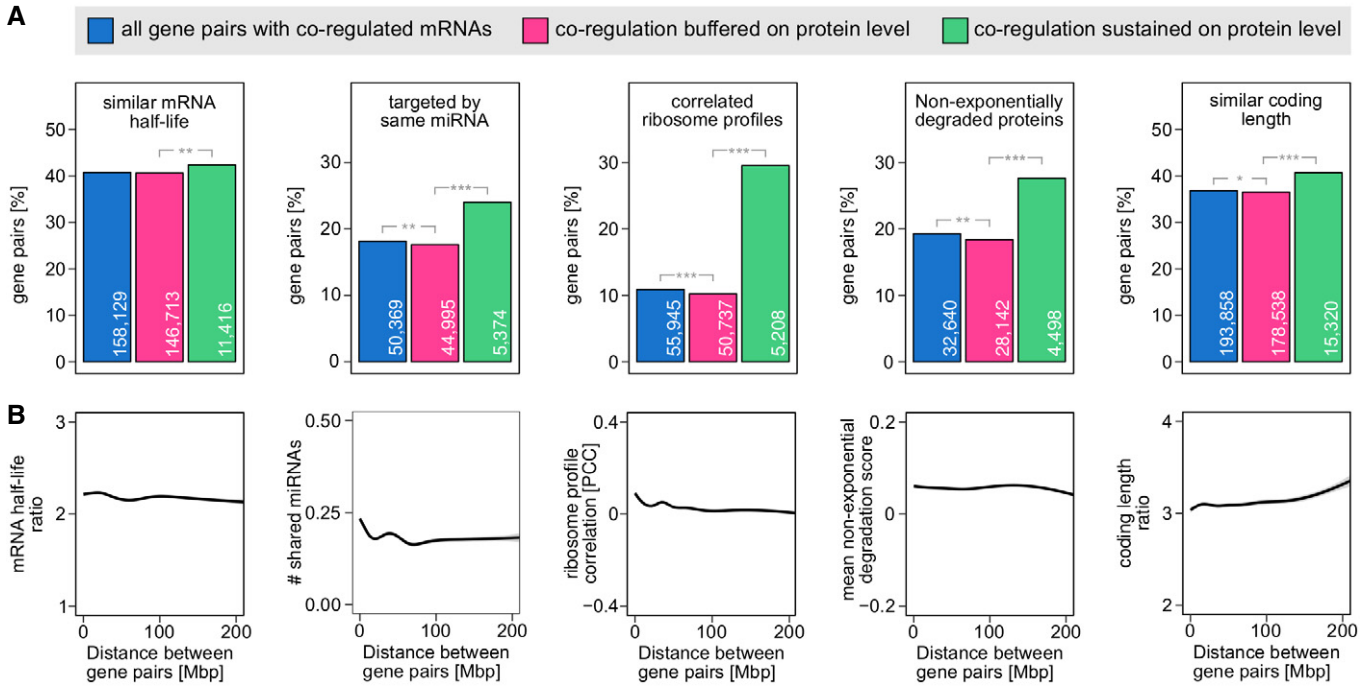


Figure 4. Buffering of non-functional mRNA co-regulation likely is a passive process.

A Percentage of gene pairs with coordinated post-transcriptional regulation, irrespective of genomic distance. Gene pairs with sustained protein co-regulation consistently stand out as more likely to share similar aspects of post-transcriptional control. Genes were considered to have a similar mRNA half-life if the half-life ratio between the more and less stable gene was < 1.5. For miRNAs, all gene pairs targeted by at least one shared miRNA were considered. Gene pairs were said to have correlated ribosome profiles if their ribosome occupancy correlated with PCC > 0.5 (BH adj. $P < 0.001$) across LCLs. For the non-exponentially degraded proteins (NEDs) barchart, gene pairs containing at least one NED were counted. Coding length was considered similar if the longer protein was < 1.5-fold longer than the shorter protein. Numbers of gene pairs are shown inside the bars. Statistical significance was calculated using Fisher's exact test (*$P < 0.01$, **$P < 1 \times 10^{-6}$, ***$P < 3 \times 10^{-27}$).

B No striking relationship between gene distance and the extent to which gene pairs show similar post-transcriptional regulation. Note that the small increase of similar ribosome occupancy towards closeby genes may be explained by the fact that ribosome profiles partially reflect mRNA abundance.

Nevertheless, it is still unclear whether expression clusters are the result of natural selection. In yeast, only the most highly coexpressed neighbours are conserved as a pair, but these also tend to be functionally related (Hurst *et al*, 2002). Neighbouring gene pairs that separate tend to show interchromosomal co-localisation (Dai *et al*, 2014). In *Drosophila*, highly coexpressed neighbouring gene pairs are less likely to be conserved than expected (Weber & Hurst, 2011). In mammals, although some coexpression clusters are evolutionarily maintained (Sémon & Duret, 2006), natural selection generally tends to separate gene pairs that show a strong position-related coexpression effect (Liao & Zhang, 2008) or that involve tissue-specific expression (Lercher *et al*, 2002). This indicates that non-functional coexpression can affect cell fitness under some circumstances, possibly if it becomes so strong that it persists through the uncoordinated post-transcriptional processes.

The existence of coexpression clusters may also reflect the way new genes originate. For example, highly transcribed chromatin regions are more susceptible to retroposition (Hurst *et al*, 2004). Recently, it has been proposed that the large number of human gene pairs in head-to-head orientation may arise from divergent transcription of single genes, when initially noncoding, antisense transcripts evolve into new protein-coding genes (Wu & Sharp, 2013). In both of these cases, new genes would have no sequence homology with their neighbours, and would therefore be unlikely to share their function. However, some of the most well-known coexpression clusters, such as histone gene clusters, arose by gene duplication. Gene duplicates could potentially explain why some gene clusters are functionally related. There are 30 gene pairs in our dataset that are located within 50 kb from each other and are coexpressed on both the mRNA and the protein level. Of these, 10 (33%) are classified as paralogues by Ensembl, a strong enrichment considering that paralogues account for only 1.5% of these closeby gene pairs overall. However, 20 (66%) of the clustered gene pairs with co-regulated protein abundances show no evidence for paralogy, suggesting that functionally relevant clusters need not necessarily arise by gene duplication.

Our analysis focussed on housekeeping genes, because comparable data for tissue- or condition-specific genes were not available. Housekeeping genes constitute about half of all human genes (Uhlén *et al*, 2015). They have a higher tendency to cluster than other genes (Lercher *et al*, 2002), presumably because they are more sensitive to gene expression noise (Fraser *et al*, 2004). Interestingly, post-transcriptional expression control is particularly important for housekeeping genes (Gandhi *et al*, 2011; Jovanovic *et al*, 2015). Notably, transcriptional activation of induced genes can also lead to co-activation of functionally unrelated neighbouring genes (Spitz *et al*, 2003; Ebisuya *et al*, 2008). However, it remains to be seen if such co-activation is also buffered at the protein level.

In conclusion, non-functional mRNA coexpression, due to chromatin fluctuations and regulatory interference, is far more common than previously thought. Generally, this does not hamper cell fitness as post-transcriptional regulatory mechanisms enforce functional coexpression while dampening non-functional coexpression. Our observations suggest that evolution of human genome organisation is driven by noise reduction, which is a hypothesis initially made in yeast (Batada & Hurst, 2007). The large presence of non-functional coexpression of genes at the transcript but not protein level has implications for the fields of transcriptomics and proteomics when screening for functional links between genes.

Materials and Methods

mRNA abundances in human lymphoblastoid cell lines

RNA-sequencing data for human lymphoblastoid cell lines (LCLs) have been reported (Pickrell *et al*, 2010). Counts per mapped reads were downloaded from http://eqtl.uchicago.edu and converted to log2 "reads per kilobase transcript per million mapped reads" (RPKMs). Genes expressed in < 30 LCLs were removed. In order to make mRNA measurements comparable to proteomics data, expression levels needed to be analysed relative to the same reference LCL. To do so, log2 RPKMs values from the reference cell line GM19238 were subtracted from all other LCLs.

Protein abundances in human lymphoblastoid cell lines

Protein abundances in LCLs have also been reported (Battle *et al*, 2015). They have been measured by mass spectrometry and quantified relative to the reference cell line GM19238, using stable isotope labelling by amino acids in cell culture (SILAC) (Ong *et al*, 2002). Mass spectrometry raw files were downloaded from the PRIDE repository (Vizcaíno *et al*, 2016) (project identifier PXD001406) and re-processed using MaxQuant 1.5.2.8 (Cox & Mann, 2008). Raw files tagged as "run2" were omitted. Mass spectra were searched against human Swiss-Prot sequences downloaded from Uniprot (UniProt Consortium, 2015). To facilitate combining mRNA and protein datasets, no protein isoforms were considered. We used non-normalised SILAC ratios obtained by MaxQuant with at least two ratio counts. Because the internal standard had been used as heavy SILAC sample, heavy/light (H/L) SILAC ratios were inverted to obtain L/H ratios (i.e. test LCLs / reference LCL). Proteins that could not be unambiguously mapped to a single gene were removed, as were proteins detected in 30 LCLs or less. SILAC ratios were also log2-transformed.

Combining mRNA and protein expression data

To combine mRNA and protein data, ENSEMBL gene IDs from RNA sequencing were mapped to Uniprot IDs using Uniprot's webtool (UniProt Consortium, 2015). Genes with ambiguous mappings were removed. We also only considered LCLs for which both mRNA and protein data were available. The resulting file contains mRNA and protein abundances for 4,188 human genes in 59 LCLs, relative to the GM19238 reference sample (Dataset EV1). It contains 0.1 and 6.7% missing values for mRNA and protein measurements, respectively.

Defining positions of genes in the genome

Genomic coordinates of human genes (dataset version GRCh38.p5) were downloaded from ENSEMBL (Yates *et al*, 2016). As we are considering genes but not specific transcript or protein isoforms, transcription start sites (TSSs) were defined as the start site of the outermost transcript of a gene.

Testing gene pairs for co-regulation

Coordinated up- and downregulation of gene expression was measured using Pearson's correlation coefficient (PCC). The gene expression datasets for LCLs (Dataset EV1) were used as input. The median log2 fold change of each LCL was set to zero, in order to prevent correlations reflecting irrelevant data features such as uneven mixing of light and heavy SILAC samples. Gene pairs were considered to be co-regulated at PCC > 0.5, but only if the correlation was significant (Benjamini and Hochberg-adjusted *P*-values < 0.001).

Characterisation of genes as housekeeping genes

To demonstrate that the 4,188 genes in the LCL dataset belong to the constitutively expressed core proteome, we performed a number of tests:

Chromatin states of gene promoters

Chromatin states of the genome of the GM12878 lymphoblastoid cell line were determined previously (Ernst *et al*, 2011). They were downloaded as hg19 genome coordinates from the USCS genome browser (Rosenbloom *et al*, 2015) and converted to GRCh38 coordinates using the liftOver command line tool (available at https://genome-store.ucsc.edu/). Genomic regions with conflicting chromatin state annotations, resulting from the genome coordinates update, were removed. For each gene in our dataset, the chromatin state mapping to its transcription start site was determined.

GO term enrichment

A statistical overrepresentation test was performed using the PANTHER classification system (Mi *et al*, 2016) according to the reported protocol (Mi *et al*, 2013). Overrepresentation of Gene Ontology Biological Process (slim) terms was assessed for our 4,188 genes compared to the entire human genome. Only significantly enriched terms (more than twofold; $P < 0.05$ after Bonferroni correction) were considered.

mRNA tissue expression data

mRNA expression levels in different human tissues have been assessed using RNA sequencing (Uhlén *et al*, 2015). Transcripts detected with FPKM ≥ 1 were considered to be expressed.

Protein tissue expression data

Protein expression levels in different human tissues have been assessed using mass spectrometry (Wilhelm *et al*, 2014) (available at www.proteomicsdb.org). To avoid bias due to the incomplete nature of current proteome maps, only tissues with expression values for more than 6,000 proteins were considered.

Defining pairs of genes with related functions (focussed on accuracy)

To test whether genes with related functions are co-regulated across LCLs, we defined three sets of functionally linked gene pairs. Functional associations in these test sets are as accurate—not as comprehensive—as possible.

Gene pairs from same protein complexes
Human protein–protein interaction pairs based on Reactome pathways (Fabregat *et al*, 2016) were downloaded from www.reactome.org (homo_sapiens.interactions.txt file; March 2016). They were filtered for physical interactions of the "direct_complex" category. Gene pairs belonging to more than one complex and homodimeric interactions were removed.

Gene pairs encoding enzymes from consecutive metabolic reactions
As for protein complexes, human protein–protein interaction pairs based on Reactome pathways (Fabregat *et al*, 2016) were downloaded from www.reactome.org (homo_sapiens.interactions.txt file; March 2016). They were filtered for interactions of the "neighbouring_reactions" category. These are interactions where one gene/protein produces the input or catalyst for the second reaction. Any gene pairs known to interact also physically, that is belonging to the "direct_complex" or "indirect_complex" categories, were removed. In addition, gene pairs were filtered for those involved in *metabolic* pathways, as opposed to, for example, the cell cycle pathway which would contain irrelevant reactions such as "Mis18 complex binds the centromere". To do so, we first inferred all pathways mapping to the metabolism root pathway, using the pathway hierarchy relationship file (ReactomePathwaysRelation.txt, available on www.reactome.org). Enzymatic reactions belonging to each metabolic pathway were then identified using another interaction file available from Reactome (homo_sapiens.mitab.interactions.txt). Finally, to avoid "trivial" consecutive reactions such as those involving ubiquitous metabolites like NAD^+, we removed metabolic reactions with more than ten neighbouring reactions.

Gene pairs from identical subcellular locations
Subcellular localisations of human proteins were downloaded from Uniprot (UniProt Consortium, 2015). Proteins localising to more than one subcellular location were removed. To avoid trivial localisations such as "cytoplasm", only subcellular compartments with 200 or less known protein components were considered.

Defining pairs of genes with related functions (focussed on completeness)

To estimate an upper limit for how many coexpressed neighbouring genes may be functionally related, we defined a separate test set based on the STRING database (Szklarczyk *et al*, 2017). Functional associations in this test set are as comprehensive as possible. Protein network data for *Homo sapiens* were downloaded from http://string-db.org. We considered all functional associations with a combined STRING score > 0.7. This score integrates various types of evidence and indicates the likelihood of the association to be biologically meaningful, specific and reproducible.

Testing functionally related gene pairs for co-regulation

Correlation coefficients were obtained for every gene pair in our three test sets (protein complexes, consecutive metabolic reactions, subcellular locations) and their distribution was displayed in histograms. As a control, gene pairs were randomly shuffled to break the link between the pairs. For example, gene pairs encoding subunits of the same protein complexes were shuffled such that the same genes were paired randomly, in which case most gene pairs encode subunits of different protein complexes. The Kolmogorov–Smirnov test was used to assess whether PCC distributions of relevant gene pairs were significantly different from those obtained with randomised pairs.

Chromosome co-regulation mapping

PCCs were calculated for all relevant gene combinations, as described for histograms above. For chromosome co-regulation curves, PCCs were plotted against the genomic distance between transcription start sites, with curves fitted by a generalised additive model. For chromosome co-regulation maps, genes were plotted in their chromosomal order and PCCs between all gene combinations were represented by a colour scale.

Hi-C interactions for our gene set

Hi-C contact matrices for a lymphoblastoid cell line (Rao *et al*, 2014) were downloaded from NCBI GEO database (accession GSE63525). An unpublished script from Liz Ing-Simmons (available at https://github.com/liz-is/readhic) was adapted (available at https://github.com/Rappsilber-Laboratory/readhic) and then used to import the Hi-C contact matrices into R, using 10-kb resolution and "KRnorm" normalisation for intrachromosomal pairs and 50-kb resolution and "INTERKRnorm" normalisation for interchromosomal pairs. All reads used passed the MAPQ>0 filter. Hi-C data are based on GRCh37 genome coordinates. GRCh37 transcription start sites for all genes were obtained using the biomaRt R package (Durinck *et al*, 2009), considering only the TSS of the outermost transcript of each gene. The GenomicInteractions R package (Harmston *et al*, 2015) was used to determine the contact frequency between the genes in our dataset, considering the median read count of all Hi-C pixels in a range $\pm$ 40 kb around the TSS of each gene.

Analysis of genome subcompartments

Nuclear subcompartments A1, A2, B1, B2, B3 and B4 have been defined previously (Rao *et al*, 2014). A genome-wide mapping of subcompartments in a lymphoblastoid cell line is available via the NCBI GEO database (accession GSE63525). Subcompartment annotations were lifted from hg19/b37 to GRCh38 genome coordinates using the UCSC genome browser service (Rosenbloom *et al*, 2015).

k-means clustering of transcript and protein expression changes

k-means clustering was performed using the default algorithm and settings in R (R Core Team, 2016), with $k = 4$ (mRNAs) or $k = 3$ (proteins) and five random start sets. Values of k were chosen such that the clusters explain at least 50% of the total variance.

Analysis of cluster features

Subcellular locations

To get a broad understanding of subcellular locations enriched in *k*-means clusters, we downloaded all Uniprot entries mapping to the locations Nucleus (Uniprot subcellular location ID: SL-0191), Endoplasmic reticulum (SL-0095), Golgi apparatus (SL-0132), Mitochondrion (SL-0173) and Cytoplasm (SL-0086) (UniProt Consortium, 2015). Proteins localising to the Endoplasmic reticulum and/or the Golgi apparatus were combined as "ER-Golgi". Proteins mapping to more than one organelle were removed.

GO term enrichment

A statistical overrepresentation test was performed using the PANTHER classification system (Mi *et al*, 2016) according to the reported protocol (Mi *et al*, 2013). Overrepresentation of Gene Ontology Biological Process (complete) terms in each cluster, relative to other clusters, was assessed. Using PANTHER's GO hierarchy annotation, we reported only the most specific GO terms and omitted any co-enriched parent terms for clarity. All reported GO terms were significantly enriched ($P < 0.05$ after Bonferroni correction).

Genomic and epigenomic features

Raw signals of ChIP-seq experiments for lymphoblastoid cells were downloaded from ENCODE (ENCODE Project Consortium, 2012) in hg19 genomic coordinates. ENCODE accessions were ENCFF000ARW (H2AZ), ENCFF000ARZ (H3K4me1), ENCFF000ATL (H3K4me2), ENCFF001EXX (H3K4me3), ENCFF000ASJ (H3K27ac), ENCFF000ATX (H3K79me2), ENCFF000AUF (H3K9ac), ENCFF000 AUL (H3K9me3), ENCFF000AUS (H4K20me1), ENCFF001EXC (H3K 27me3), ENCFF001EXP (H3K36me3), ENCFF001GNK (RepliSeq G1b), ENCFF001GNN (RepliSeq G2), ENCFF001GNR (RepliSeq S1), ENCFF001GNT (RepliSeq S2), ENCFF001GNX (RepliSeq S3) and ENCFF001GOA (RepliSeq S4). These bigWig files were converted to bedGraph files, lifted over to GRCh38 coordinates, cleared of any resulting overlaps and converted back to bigWig files using command line tools from the UCSC genome browser (Rosenbloom *et al*, 2015) (tools available at https://genome-store.ucsc.edu/). GC percentage over 5-bp windows was downloaded from the UCSC genome browser (Rosenbloom *et al*, 2015). Average signals over gene bodies were calculated with the UCSC bigWigAverageOverBed command line utility, using the coordinates of our genes as bed files. CpG methylation from reduced representation bisulphite sequencing of a lymphoblastoid cell line was also available from ENCODE (ENCODE Project Consortium, 2012) (experiment ENCSR000DFT; file accession ENCFF001TLQ). After lifting the hg19 bedMethyl file over to GRCh38 genomic coordinates, the mean percentage of CpG methylation in gene bodies was calculated using an R script. For each epigenomic or genomic feature, the median enrichment for genes in each *k*-means cluster, compared to all genes in our dataset, was calculated and plotted as log2 ratio in a heatmap.

Calculation of gene expression noise

Gene expression noise at the mRNA and protein levels was calculated as the coefficient of variation (CV; standard deviation divided by the mean) of log2-transformed RPKM and SILAC ratios, respectively. To avoid dividing by zero (for unchanged genes with a log2 ratio of zero), a constant value of 10 was added to all mRNA and protein log2 ratios before calculating the noise.

Calculating the clustering degree

To define local gene density in a manner that can be applied to both the sequence and the 3D structure of the genome, we determined the average distance of a gene to its three nearest neighbouring genes. We calculated three such "clustering degrees" for each gene in our dataset. For the sequence-based clustering degree, the distance to neighbouring genes was calculated in base pairs. For intrachromosomal clustering in 3D, gene distance was calculated based on Hi-C counts. However, we only considered "nearest" neighbours which were at least 500 kb away in terms of DNA sequence, to catch long-range interactions and avoid replicating the sequence-based clustering degree. For interchromosomal clustering, we considered the three nearest neighbours on other chromosomes, based on interchromosomal Hi-C contacts.

Heterochromatin profiles of upstream regions

Chromatin states throughout the LCL genome were previously described (Ernst *et al*, 2011). To simplify the analysis, we combined the five inactive chromatin states defined by Ernst *et al* ("Heterochromatin", "Repressed", "Repetitive", "Poised Promoter" and "Insulator") into one "heterochromatin" state. We then scanned the promoter region of test genes for the presence of heterochromatin, moving in 100-bp intervals from −50,000 bp to +10,000 bp relative to their transcription start site.

Calculating epigenetic similarity

Epigenetic similarity was calculated on the basis of the histone mark abundance within gene bodies (see section "Analysis of cluster features" for processing of ChIP-seq data). For this analysis, we considered H2AFZ, H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K79me2, H3K9ac, H3K9me3, H4K20me1, H3K27me3, H3K36me3 and CpG methylation, but not GC content, gene length and replication timing. For every pair of genes, we then determined how similar or dissimilar they are regarding the abundance of these epigenetic features. This was calculated using the Mahalanobis distance measure, which takes into account that some histone marks strongly covary.

Analysis of post-transcriptional mechanisms

mRNA half-lives in seven different LCLs were previously reported (Duan *et al*, 2013). We first calculated the average half-life of each mRNA in these LCLs. We considered two mRNAs to have a similar stability if the half-life of the more stable one was < 1.5-fold longer than the less stable one. mRNA targets of human miRNAs were also described previously (Helwak *et al*, 2013). Ribosome occupancy profiles for the LCL cell line panel were recently published (Battle *et al*, 2015). We considered ribosome profiles for 57 LCLs and 4,033 genes for which we had matching mRNA and protein measurements. We calculated Pearson correlation coefficients (PCCs) for ribosome profiles between all gene pairs. Two genes were said to

have correlated ribosome profiles at PCC > 0.5 (BH-adjusted *P*-value < 0.001). Proteins subjected to non-exponential degradation in human RPE-1 cells were also described recently (McShane *et al*, 2016). Finally, protein sequence lengths were downloaded from Uniprot (UniProt Consortium, 2015).

Human paralogous genes

Human gene duplicates were downloaded from ENSEMBL (Yates *et al*, 2016). We only considered paralogues with at least 25% sequence identity.

General data processing and plotting

Data processing was performed in R (R Core Team, 2016), unless indicated otherwise. Plots were created using the ggplot2 package (Wickham, 2009).

Expanded View for this article is available online.

Acknowledgements
This work was supported by the Wellcome Trust through a Senior Research Fellowship to JR (grant number 103139). The Wellcome Trust Centre for Cell Biology is supported by core funding from the Wellcome Trust (grant number 203149).

Author contributions
PG analysed Hi-C contact frequencies between the genes in our dataset. GK and JR designed the study, analysed the data and wrote the paper.

Conflict of interest
The authors declare that they have no conflict of interest.

References

Akey JM, Biswas S, Leek JT, Storey JD (2007) On the design and analysis of gene expression studies in human populations. *Nat Genet* 39: 807 – 808; author reply 808–9

Akhtar W, de Jong J, Pindyurin AV, Pagie L, Meuleman W, de Ridder J, Berns A, Wessels LFA, van Lohuizen M, van Steensel B (2013) Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell* 154: 914 – 927

Albert FW, Kruglyak L (2015) The role of regulatory variation in complex traits and disease. *Nat Rev Genet* 16: 197 – 212

Anderson MZ, Gerstein AC, Wigen L, Baller JA, Berman J (2014) Silencing is noisy: population and cell level noise in telomere-adjacent genes is dependent on telomere position and sir2. *PLoS Genet* 10: e1004436

Batada NN, Hurst LD (2007) Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nat Genet* 39: 945 – 949

Batada NN, Urrutia AO, Hurst LD (2007) Chromatin remodelling is a major source of coexpression of linked genes in yeast. *Trends Genet* 23: 480 – 484

Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, Pritchard JK, Gilad Y (2015) Genomic variation. Impact of regulatory variation from RNA to protein. *Science* 347: 664 – 667

Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voûte PA, Heisterkamp S, van Kampen A, Versteeg R (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* 291: 1289 – 1292

Chen X, Zhang J (2016) The genomic landscape of position effects on protein expression level and noise in yeast. *Cell Syst* 2: 347 – 354

Cohen BA, Mitra RD, Hughes JD, Church GM (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet* 26: 183 – 186

Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 26: 1367 – 1372

Dai Z, Xiong Y, Dai X (2014) Neighboring genes show interchromosomal colocalization after their separation. *Mol Biol Evol* 31: 1166 – 1172

Dephoure N, Hwang S, O'Sullivan C, Dodgson SE, Gygi SP, Amon A, Torres EM (2014) Quantitative proteomic analysis reveals posttranslational responses to aneuploidy in yeast. *Elife* 3: e03023

Duan J, Shi J, Ge X, Dölken L, Moy W, He D, Shi S, Sanders AR, Ross J, Gejman PV (2013) Genome-wide survey of interindividual differences of RNA stability in human lymphoblastoid cell lines. *Sci Rep* 3: 1318

Durinck S, Spellman PT, Birney E, Huber W (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* 4: 1184 – 1191

Ebisuya M, Yamamoto T, Nakajima M, Nishida E (2008) Ripples from neighbouring transcription. *Nat Cell Biol* 10: 1106 – 1113

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57 – 74

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473: 43 – 49

Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S, Matthews L, May B, Milacic M, Rothfels K, Shamovsky V, Webber M, Weiser J, Williams M, Wu G, Stein L *et al* (2016) The reactome pathway knowledgebase. *Nucleic Acids Res* 44: D481 – D487

Fraser HB, Hirsh AE, Giaever G, Kumm J, Eisen MB (2004) Noise minimization in eukaryotic gene expression. *PLoS Biol* 2: e137

Gandhi SJ, Zenklusen D, Lionnet T, Singer RH (2011) Transcription of functionally related constitutive genes is not coordinated. *Nat Struct Mol Biol* 18: 27 – 34

Geiger T, Cox J, Mann M (2010) Proteomic changes resulting from gene copy number variations in cancer cells. *PLoS Genet* 6: e1001090

Ghanbarian AT, Hurst LD (2015) Neighboring genes show correlated evolution in gene expression. *Mol Biol Evol* 32: 1748 – 1766

Gibcus JH, Dekker J (2013) The hierarchy of the 3D genome. *Mol Cell* 49: 773 – 782

Gierman HJ, Indemans MHG, Koster J, Goetze S, Seppen J, Geerts D, van Driel R, Versteeg R (2007) Domain-wide regulation of gene expression in the human genome. *Genome Res* 17: 1286 – 1295

Harmston N, Ing-Simmons E, Perry M, Barešić A, Lenhard B (2015) GenomicInteractions: an R/Bioconductor package for manipulating and investigating chromatin interaction data. *BMC Genom* 16: 963

Helwak A, Kudla G, Dudnakova T, Tollervey D (2013) Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* 153: 654 – 665

Hurst LD, Williams EJB, Pál C (2002) Natural selection promotes the conservation of linkage of co-expressed genes. *Trends Genet* 18: 604 – 606

Hurst LD, Pál C, Lercher MJ (2004) The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet* 5: 299 – 310

Ingolia NT (2014) Ribosome profiling: new views of translation, from single codons to genome scale. *Nat Rev Genet* 15: 205 – 213

Jovanovic M, Rooney MS, Mertins P, Przybylski D, Chevrier N, Satija R, Rodriguez EH, Fields AP, Schwartz S, Raychowdhury R, Mumbach MR, Eisenhaure T, Rabani M, Gennert D, Lu D, Delorey T, Weissman JS, Carr SA, Hacohen N, Regev A (2015) Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. *Science* 347: 1259038

Khan Z, Ford MJ, Cusanovich DA, Mitrano A, Pritchard JK, Gilad Y (2013) Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science* 342: 1100 – 1104

Lercher MJ, Urrutia AO, Hurst LD (2002) Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat Genet* 31: 180 – 183

Li Y-Y, Yu H, Guo Z-M, Guo T-Q, Tu K, Li Y-X (2006) Systematic analysis of head-to-head gene organization: evolutionary conservation and potential biological relevance. *PLoS Comput Biol* 2: e74

Liao B-Y, Zhang J (2008) Coexpression of linked genes in Mammalian genomes is generally disadvantageous. *Mol Biol Evol* 25: 1555 – 1565

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326: 289 – 293

Liu Y, Beyer A, Aebersold R (2016) On the dependency of cellular protein levels on mRNA abundance. *Cell* 165: 535 – 550

McCullagh E, Seshan A, El-Samad H, Madhani HD (2010) Coordinate control of gene expression noise and interchromosomal interactions in a MAP kinase pathway. *Nat Cell Biol* 12: 954 – 962

McShane E, Sin C, Zauber H, Wells JN, Donnelly N, Wang X, Hou J, Chen W, Storchova Z, Marsh JA, Valleriani A, Selbach M (2016) Kinetic analysis of protein stability reveals age-dependent degradation. *Cell* 167: 803 – 815.e21

Mi H, Muruganujan A, Casagrande JT, Thomas PD (2013) Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc* 8: 1551 – 1566

Mi H, Poudel S, Muruganujan A, Casagrande JT, Thomas PD (2016) PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res* 44: D336 – D342

Michalak P (2008) Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics* 91: 243 – 248

Newman JRS, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, DeRisi JL, Weissman JS (2006) Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* 441: 840 – 846

Ong S-E, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* 1: 376 – 386

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard JK (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464: 768 – 772

Purmann A, Toedling J, Schueler M, Carninci P, Lehrach H, Hayashizaki Y, Huber W, Sperling S (2007) Genomic organization of transcriptomes in mammals: coregulation and cofunctionality. *Genomics* 89: 580 – 587

R Core Team (2016) *R: a language and environment for statistical computing*. Vienna, Austria: R Core Team. Available at: https://www.R-project.org/

Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S (2006) Stochastic mRNA synthesis in mammalian cells. *PLoS Biol* 4: e309

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159: 1665 – 1680

Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, Harte RA, Heitner S, Hickey G, Hinrichs AS, Hubley R, Karolchik D, Learned K, Lee BT, Li CH, Miga KH *et al* (2015) The UCSC genome browser database: 2015 update. *Nucleic Acids Res* 43: D670 – D681

Sandhu KS (2012) Did the modulation of expression noise shape the evolution of three dimensional genome organizations in eukaryotes? *Nucleus* 3: 286 – 289

Sandhu KS, Li G, Poh HM, Quek YLK, Sia YY, Peh SQ, Mulawadi FH, Lim J, Sikic M, Menghi F, Thalamuthu A, Sung WK, Ruan X, Fullwood MJ, Liu E, Csermely P, Ruan Y (2012) Large-scale functional organization of long-range chromatin interaction networks. *Cell Rep* 2: 1207 – 1219

Sémon M, Duret L (2006) Evolutionary origin and maintenance of coexpressed gene clusters in mammals. *Mol Biol Evol* 23: 1715 – 1723

Spitz F, Gonzalez F, Duboule D (2003) A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell* 113: 405 – 417

Stark AL, Hause RJ Jr, Gorsic LK, Antao NN, Wong SS, Chung SH, Gill DF, Im HK, Myers JL, White KP, Jones RB, Dolan ME (2014) Protein quantitative trait loci identify novel candidates modulating cellular response to chemotherapy. *PLoS Genet* 10: e1004192

Stingele S, Stoehr G, Peplowska K, Cox J, Mann M, Storchova Z (2012) Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. *Mol Syst Biol* 8: 608

Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, Jensen LJ, von Mering C (2017) The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res* 45: D362 – D368

Thévenin A, Ein-Dor L, Ozery-Flato M, Shamir R (2014) Functional gene groups are concentrated within chromosomes, among chromosomes and in the nuclear space of the human genome. *Nucleic Acids Res* 42: 9854 – 9861

Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM (2004) An abundance of bidirectional promoters in the human genome. *Genome Res* 14: 62 – 66

Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CA-K, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T *et al* (2015) Proteomics. Tissue-based map of the human proteome. *Science* 347: 1260419

UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res* 43: D204 – D212

Vizcaíno JA, Csordas A, del-Toro N, Dianes JA, Griss J, Lavidas I, Mayer G, Perez-Riverol Y, Reisinger F, Ternent T, Xu Q-W, Wang R, Hermjakob H (2016) 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res* 44: D447 – D456

Vogel C, Abreu Rde S, Ko D, Le S-Y, Shapiro BA, Burns SC, Sandhu D, Boutz DR, Marcotte EM, Penalva LO (2010) Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol Syst Biol* 6: 400

Wang G-Z, Lercher MJ, Hurst LD (2011) Transcriptional coupling of neighboring genes and gene expression noise: evidence that gene orientation and noncoding transcripts are modulators of noise. *Genome Biol Evol* 3: 320 – 331

Weber CC, Hurst LD (2011) Support for multiple classes of local expression clusters in *Drosophila melanogaster*, but no evidence for gene order conservation. *Genome Biol* 12: R23

Wickham H (2009) *Ggplot2 elegant graphics for data analysis*. New York: Springer

Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, Mathieson T, Lemeer S, Schnatbaum K, Reimer U, Wenschuh H, Mollenhauer M, Slotta-Huspenina J, Boese J-H, Bantscheff M, Gerstmair A *et al* (2014) Mass-spectrometry-based draft of the human proteome. *Nature* 509: 582 – 587

Williams EJB, Bowles DJ (2004) Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. *Genome Res* 14: 1060 – 1067

Wu X, Sharp PA (2013) Divergent transcription: a driving force for new gene origination? *Cell* 155: 990 – 996

Xu C, Chen J, Shen B (2012) The preservation of bidirectional promoter architecture in eukaryotes: what is the driving force? *BMC Syst Biol* 6 (Suppl 1): S21

Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, Girón CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Keenan S, Lavidas I, Martin FJ *et al* (2016) Ensembl 2016. *Nucleic Acids Res* 44: D710 – D716

Yuan Y, Tian L, Lu D, Xu S (2015) Analysis of genome-wide RNA-sequencing data suggests age of the CEPH/Utah (CEU) lymphoblastoid cell lines systematically biases gene expression profiles. *Sci Rep* 5: 7960

Manuscript 2. “Epigenetic Variability Confounds Transcriptome but not Proteome Profiling for Coexpression-based Gene Function Prediction”

Pages 31 - 39

Manuscript available online, DOI: 10.1074/mcp.RA118.000935

⌘ Author's Choice

Epigenetic Variability Confounds Transcriptome but Not Proteome Profiling for Coexpression-based Gene Function Prediction*

Piotr Grabowski‡, Georg Kustatscher§, and Juri Rappsilber‡§¶

Genes are often coexpressed with their genomic neighbors, even if these are functionally unrelated. For small expression changes driven by genetic variation within the same cell type, non-functional mRNA coexpression is not propagated to the protein level. However, it is unclear if protein levels are also buffered against any non-functional mRNA coexpression accompanying large, regulated changes in the gene expression program, such as those occurring during cell differentiation. Here, we address this question by analyzing mRNA and protein expression changes for housekeeping genes across 20 mouse tissues. We find that a large proportion of mRNA coexpression is indeed non-functional and does not lead to coexpressed proteins. Chromosomal proximity of genes explains a proportion of this nonfunctional mRNA coexpression. However, the main driver of non-functional mRNA coexpression across mouse tissues is epigenetic similarity. Both factors together provide an explanation for why monitoring protein coexpression outperforms mRNA coexpression data in gene function prediction. Furthermore, this suggests that housekeeping genes translocating during evolution within genomic subcompartments might maintain their broad expression pattern. ***Molecular & Cellular Proteomics 17: 2082–2090, 2018. DOI: 10.1074/mcp.RA118.000935.***

Genes are not arranged randomly but tend to be clustered in the genome into coexpressed domains (1). Such clustering can be a regulatory strategy of both prokaryotic and eukaryotic genomes. Interestingly, this does not mean that genes that are coexpressed are necessarily also linked functionally. There exist gene clusters that tend to be coexpressed, yet lack evident cofunctionality (1, 2). This is especially visible for bidirectional gene pairs which are coexpressed because of shared regulatory context, but commonly seem to lack a functional relationship (3). This has an impact on gene coexpression studies which infer functional associations between genes based on similar gene activity. Coexpression of spatially close genes can be driven by stochastic transcriptional bursting (4) or transcriptional interference between neighboring genes (5). The existence of coexpressed gene clusters that lack a functional connection is intriguing given that nonspecific gene expression should have a negative impact on cell fitness. Interestingly, Hurst and colleagues have shown that clustered genes mutually reinforce their active state and are less likely to be accidentally silenced, for example by stochastic fluctuations of chromatin states (6). Therefore, clustered genes show lower expression noise, a benefit that may offset the negative impact of their coincidental coexpression. In agreement with this, we have recently demonstrated that coexpression of proximal genes, both in terms of sequence and 3D genomic proximity, is pervasive in the human genome. Importantly, however, coexpression of spatially close, functionally unrelated genes is restricted to their mRNA abundances and is not propagated to the protein level (7). This protein-level buffering of non-functional mRNA coexpression supports the idea that reduction of expression noise is a key driver of the evolution of genome organization. Consequently, function prediction is based better on protein coexpression than mRNA coexpression data (8, 9).

Our previous analysis was based on a panel of human lymphoblastoid cell lines (LCLs)[1] for which the expression changes had a prominent noise component owing to the little variability between the cell lines. A related analysis of human cancer panels also found mRNA—but not protein—coexpression to reflect chromosomal gene colocalization (8). However, it remains to be seen if a similar uncoupling of transcriptome and proteome exists also for strong, regulated and biologically important expression changes. For example, different cell types have different metabolic needs, morphology, organelle numbers and sizes. Even for ubiquitously expressed

From the ‡Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, 13355 Berlin, Germany; §Wellcome Centre for Cell Biology, University of Edinburgh, Edinburgh EH9 3BF, UK

Received June 22, 2018, and in revised form, July 9, 2018
Published, MCP Papers in Press, July 24, 2018, DOI 10.1074/mcp.RA118.000935

housekeeping genes, this can amount to large quantitative differences in expression levels. Here, we investigate the impact of genome organization and epigenetic states on mRNA and protein coexpression across different mouse tissues by integrating multiple published omics data sets. We show that the observations made on cell lines regarding factors governing mRNA and protein coexpression also hold in tissues, with changes in the relative weights of the contributions from genome position *versus* epigenetic state. We point at possible biases in expression profiling for functional genomics that researchers should consider.

EXPERIMENTAL PROCEDURES

Mouse Tissue mRNA and Protein Expression Data Set Assembly—SILAC mouse tissue proteomes were downloaded from (10), normalized SILAC H/L ratios for each tissue extracted and log2-transformed. SILAC kidney values were obtained by averaging expression values for kidney cortex and medulla.

Transcriptomics profiling data of tissues were obtained from (11–15) (links in supplemental Table S1). Data downloaded from ENCODE were in Gencode M4-aligned bam format with the only exception of the skeletal muscle data which were downloaded in fastq format and aligned using TopHat v2.0.9 and Gencode M4 annotation. The TopHat settings were set to default apart from using "bowtie1" parameter and library type set to "fr-secondstrand." The bam files were subsequently processed using Cufflinks 2.2.1 with default settings to obtain gene expression (fragments per kilobase of exon model per million mapped reads, FPKM) values. The three tissues downloaded from GEO were in normalized FPKM or RPKM format. All the mRNA expression data were transformed into a common transcripts per million (TPM) unit. To make the RNAseq data set comparable with the proteomics data, each mRNA expression value was divided by a median expression value for a given gene in all 20 tissues (analogously to the Super-SILAC approach (16) used in the proteomics data set). Finally, the normalized TPM ratios were log2-transformed.

The resulting mRNA and protein expression data set contains 3391 genes with expression values in at least 8 tissues on both mRNA and protein levels. The proteomics data and mRNA data contain 15.5% and 6.7% missing values, respectively.

The processed data set is available as supplemental File S1.

Epigenetics Data Processing—ChIPseq data for 9 mouse tissues (marks: H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K79me2) were obtained from ENCODE in bigWig format (fold change *versus* control). The data for H3K9ac was available only for two tissues. To extract mean ChIPseq signal per gene body for all tissues, a UCSC bigWigAverageOverBed command line tool was used in conjunction with a custom-made bed file based on Gencode M4 mouse gene annotation. The processed ChIPseq data set is available as supplemental File S2.

Gene Expression Correlation Analysis—To obtain the between-gene correlation values the data were centered at 0 for each experiment and a Pearson correlation coefficient was calculated using R function "corr.test" from the psych package with the "use" parameter set to "pairwise.complete.obs." For improved statistical power, correlations were calculated only for genes which had data in at least 8 overlapping tissues (both on protein and mRNA levels). Gene pairs were considered correlated if their PCC value was > 0.5. For subsequent analyses, only correlations with Benjamini-Hochberg adjusted p values < 0.05 were considered.

Genomic Positions of Genes and Intergenic Distances—Mouse gene positions on mm10 genome were obtained from Ensembl Biomart (17, 18) (state on 29.06.2017). For gene distance calculation, first base pair of each gene's outermost transcription start site (TSS) was used and distances between those positions calculated for each gene pair.

Statistical Significance Analysis of Close-by and Other Coregulated Genes—Two Pearson Chi-squared tests were performed on two 2×2 contingency tables (for mRNA and protein levels). The first contingency table (mRNA-level) divided gene pairs by two variables. The first variable considered genomic distances between the gene pairs (close-by/other) and the second variable divided the gene pairs according to their mRNA coexpression (gene pairs with mRNA Pearson correlation coefficient > 0.5 and BH-adjusted p value < 0.05 were considered correlated and all other pairs were considered uncorrelated). Similarly, for the protein-level analysis, the first variable was genomic proximity. In the second variable, pairs were correlated if they both had mRNA and protein PCC > 0.5 and the BH-adjusted p value < 0.05.

Analysis of Post-transcriptional Mechanisms—The miRNA/gene mapping data for mouse brain were obtained from (19). The CDS lengths of coexpressed genes were obtained from Biomart using Ensembl Genes 92 database and the GRCm38.p6 data set. The genes were considered to have similar CDS length if the ratio of the length of the longer CDS to the shorter CDS was below 1.5. The liver time-series ribosome profiling data was obtained from (20). Ribosome profiling matrices were scaled using the accompanying mRNA expression data and the resulting ratios were log2-transformed. Finally, Pearson correlation coefficients between genes were calculated using R function "corr.test" from the "psych" package (21). Gene pairs with Pearson correlation coefficient > 0.5 and the Holm-adjusted p value < 0.001 were considered as correlated. Protein translation rates were obtained from (22). For each gene pair, a ratio of their translation rates was calculated, log2-transformed and the absolute values taken. Gene pairs were considered to have similar translation rates if this absolute log2 ratio was lower or equal to 1. The protein degradation profiles were obtained from (23) and gene pairs coding at least one nonexponentially degraded protein were counted.

K-means Clustering of mRNA and Protein Expression Data—The Pearson correlation coefficients for all gene pairs were used to cluster the mRNA and protein data separately. An R clustering function "kmeans" was used for this purpose. The first k value that explained 50% of the variance in the data was selected. The percentage of variance explained was defined as the ratio of the between sum of squares to the total sum of squares for every given k. The parameter "nstart" was set to 3 and "max.iter" set to 20.

Subcellular Localization Enrichment—Subcellular localization annotation was obtained from Uniprot (24). Proteins localized to more than one subcellular compartment were removed. Endoplasmic reticulum was joined with Golgi as "ER/Golgi" to balance the group sizes. Only "nucleus," "mitochondrion," and joined "ER/Golgi" groups were considered for subcellular localization enrichments. The expected value for each cluster was defined as the percentage of proteins with the given subcellular localization annotation in the data. The observed value was calculated as a percentage of those proteins in the given cluster. Finally, log2 observed/expected values were calculated for each of the cluster and subcellular localization.

[1] The abbreviations used are: LCL, lymphoblastoid cell lines; CDS, coding sequence; CV, coefficient of variation; FDR, false discovery rate; FPKM, fragments per kilobase million; GAM, generalized additive model; GEO, gene expression omnibus; GO, gene ontology; PCC, Pearson correlation coefficient; RPKM, reads per kilobase million; SILAC, stable isotope labeling with amino acids in cell culture; TPM, transcripts per million; TSS, transcription start site; UTR, untranslated region.

GO Enrichment Analysis—Gene Ontology enrichments were performed using DAVID online service (25). All Uniprot Accession numbers belonging to each of the clusters were used as a query and the whole mouse genome used as background for statistical analysis. The top 5 significantly enriched terms were reported for each cluster (FDR < 0.01).

Tissue-specific Epigenetic Cluster Profiling—The median log2 fold-change values used in Fig. 2*E* were calculated as follows: the median of the epigenetic signal of genes over all clusters in each tissue served as the expected value. The observed value was the median epigenetic signal in a given combination of cluster and tissue. Finally, a log2 observed/expected value was obtained showing the relative enrichment of the epigenetic signal between clusters for each tissue.

Calculating Epigenetic Similarity—Inverted Mahalanobis distance (1/Mahalanobis distance) was used to calculate the similarity between epigenetic profiles of genes. The "mahalanobis" R function was used with a user-specified covariance matrix.

Calculation of Gene Positional Clustering—Distances between all possible pairs of genes located on same chromosomes were calculated. For each gene, the mean distance to its 5 nearest neighbors was calculated. The list of genes was sorted by increasing mean distance to their 5 nearest neighbors. Finally, the genes at the top and bottom 5% of the list were labeled as most and least positionally clustered, respectively.

Calculation of Gene Expression Variability—Gene expression variability at the mRNA and protein levels was calculated as the coefficient of variation (CV; standard deviation divided by the mean) of log2-transformed TPM and SILAC ratios. To avoid dividing by zero (for unchanged genes with a log2 ratio of zero), a constant value of 10 was added to all mRNA and protein log2 ratios before calculating the variability.

Data Processing and Plotting—All data processing was performed in R (26) and the plots made using the ggplot2 package (27). The R scripts used to analyze data and generate most of the figures can be found on our GitHub (https://github.com/Rappsilber-Laboratory/tissue_mRNA_protein_scripts_MCP).

RESULTS AND DISCUSSION

Coexpression of Nearby Gene Pairs Is Buffered at the Protein Level in Mouse Tissues—We assembled a mouse tissue expression data set comprising 3391 genes in 20 different tissues by combining proteomics and transcriptomics from different sources. Protein abundance data were derived from a quantitative proteomics data set based on metabolic isotope labeling of mice (10). Transcriptomics data were obtained from the ENCODE Consortium (11) and Gene Expression Omnibus (GEO) repository (12) (Fig. 1*A*). The tissue collection comprises few main broad functional categories such as the nervous system (cerebellum, brain cortex), digestive system (stomach, intestine, pancreas), immune system (thymus, spleen) and multifunctional organs such as the liver and kidney. To compare the gene expression between multiple tissues with enough statistical power, we used only genes expressed ubiquitously in all tissues as opposed to using tissue-specific genes. These so-called housekeeping genes account for about half of the genome in human (28) and presumably also in mouse. They are involved in basic cellular functions such as energy metabolism (including mitochondrial proteins), genome integrity maintenance, gene expression, protein trafficking, and cell structural functions.

To generate a coexpression matrix for all observed gene pairs on both mRNA and protein level, we calculated their Pearson correlation coefficients (PCCs) across the 20 tissues (exemplified in supplemental Fig. S1). Importantly, compared with a previous study on lymphoblastoid cell lines (LCLs) (7), the expression changes observed between tissues and consequently many different cell types were substantially larger (fold-change increased by a mean of ~75% for both mRNA and proteins, Fig. 1*B*). We then assessed the quality and information content of the integrated data set by plotting the mRNA- and protein-level correlations for functionally related gene pairs. As expected, functional gene pairs have much higher correlation coefficients than randomly shuffled gene pairs (supplemental Fig. S2). This effect is more pronounced on protein than mRNA level (Fig. 1*C*). Subunits of the same complex correlated to a median of 0.59 at protein level and 0.35 at mRNA level. For comparison, in lymphoblastoid cell lines we observed 0.61 and 0.27, respectively. As one would expect, mRNA coexpression appears to be closer linked to function across tissues than closely related cell lines. Nevertheless, protein coexpression remains more indicative of shared function.

Next, we wondered about the impact of gene proximity on their correlated expression. We took gene pairs separated by less than 50 Kb between their transcription start sites ("close-by genes") and looked at their mRNA correlation compared with gene pairs further apart (Fig. 1*D*). We observe 13% of close-by genes to have coregulated mRNAs. However, only a quarter of these (3.3%) are also coregulated on the protein level. This suggests that only a fraction of those coregulated mRNA pairs is functionally related. It is worth noting that even though our mRNA and protein data have similar numbers of data points per gene, the protein data is slightly sparser (15.5% and 6.7% missing values, respectively). Despite the numerical disadvantage of the protein data set, protein-level correlations are still more informative on the function than mRNA (Fig. 1*C*, supplemental Fig. S2). The data also differs in their measurement-based variation as they were acquired by different technologies. However, we are limiting our comparisons in most cases to within-mRNA and within-protein, avoiding direct mRNA-protein comparison.

As a second line of inquiry into the impact of gene proximity on their correlated expression, we grouped the gene pairs by chromosomes, arranged them in their genomic order and plotted their correlation values as a coregulation map (Fig. 1*E*). Patches of coregulated mRNAs are clearly visible on chromosome 17 that are not reflected on the protein level. The patches are seen along the diagonal, suggesting that neighboring genes tend to be cotranscribed. Patches are also found away from the diagonal. These patches likely reflect large-scale 3D architecture as we have shown in human (7). Fitting a generalized additive model (GAM) to the linear correlation data further highlights the observed coregulation patches which might be indicative of the chromosome folding

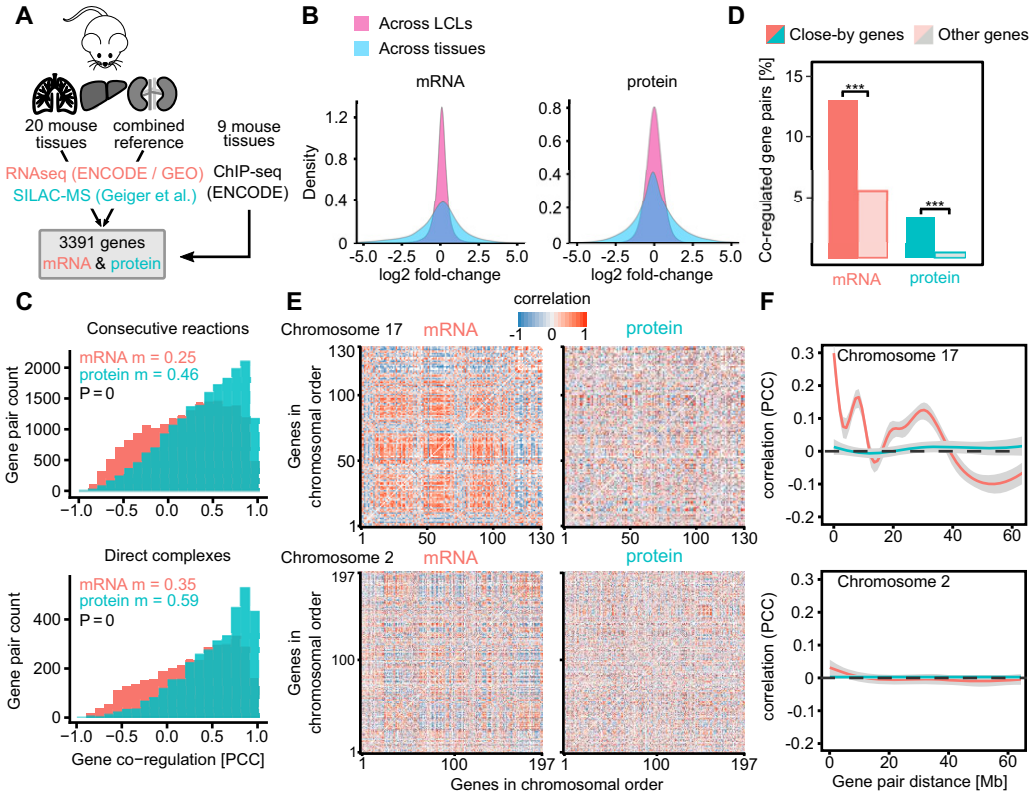


FIG. 1. **Genomic distance between gene pairs affects their coexpression stronger on the mRNA than on the protein level.** *A*, We analyzed mRNA and protein expression changes between 20 different mouse tissues. Additionally we analyzed epigenetic profiles of genes by using ENCODE data for 9 different tissues. *B*, The global log2-fold changes in the mouse tissue data set are larger on both mRNA and protein levels compared with the LCL data set as used in (7). *C*, The coregulation of enzymes catalyzing consecutive metabolic reactions and protein complexes is significantly stronger on protein level compared with mRNA level (Mann-Whitney test p value < 0.0001 in both cases, m = median). *D*, The fraction of close-by genes (< 50 kilobases separation) coregulated on mRNA level is four times as large as on protein level which suggests that only about a quarter of the proximal mRNA coregulation is functional. Statistical significance was assessed using a Pearson's Chi-squared test (***p value < 0.0001). *E*, Chromosomal gene coregulation patterns are visible on mRNA level but disappear on protein level on chromosome 17. However, this effect seems not to be as strong for chromosome 2. *F*, The mRNA coregulation decreases with the linear gene separation albeit not monotonously, reflecting the observed chromosomal coregulation patches on chromosome 17. This effect is not observed on protein level. No long-range effects can be observed for chromosome 2. The gray area around the lines signifies 95% confidence intervals.

(Fig. 1*F*, chromosome 17). The patches are not equally pronounced in all chromosomes, for example see chromosome 2 (Fig. 1*E*, 1*F*).

Gene Pairs with Sustained Coexpression Have Similar Post-transcriptional Regulation—For many gene pairs, protein coexpression correlates with mRNA coexpression, while for other gene pairs mRNA and protein coexpression are not correlated. To identify possible mechanisms leading to buffered or sustained gene coexpression we conducted an analysis of post-transcriptional mechanisms using five published data sets (Fig. 2*A*). First, we looked at how many miRNAs are shared between gene pairs. miRNAs have been implicated in post-transcriptional gene expression control by binding to transcripts and regulating mRNA degradation and protein translation (29). Using miRNA-gene interaction data generated using the CLEAR-CLIP protocol (19), we found that gene pairs with sustained coexpression tend to share significantly more miRNAs than pairs with buffered coexpression (Mann Whitney U test p value = 0.002). We then looked at protein coding sequence (CDS) lengths which are a general indicator of the extent of post-transcriptional control (30). Gene pairs with sustained coexpression had significantly (Chi-squared Test p value < 0.0001) more similar CDS lengths than gene pairs with buffered coexpression patterns. Subsequently, we looked at levels of ribosome occupancy using ribosome profiling data from mouse liver (20) and protein translation rates determined using mass spectrometry (22). In both cases, gene pairs with sustained coexpression tend to have similar translation levels (Chi-squared Test p values < 0.0001 in both cases). Finally, we looked at protein degradation profiles by considering gene pairs having at least one nonexponentially degraded protein (NEDs) (23). We found that gene pairs with sustained coexpression are significantly enriched in NEDs (Chi-squared Test p value < 0.0001). Together, this suggests that various post-transcriptional mechanisms are involved in propagating functional gene coexpression to the protein level.

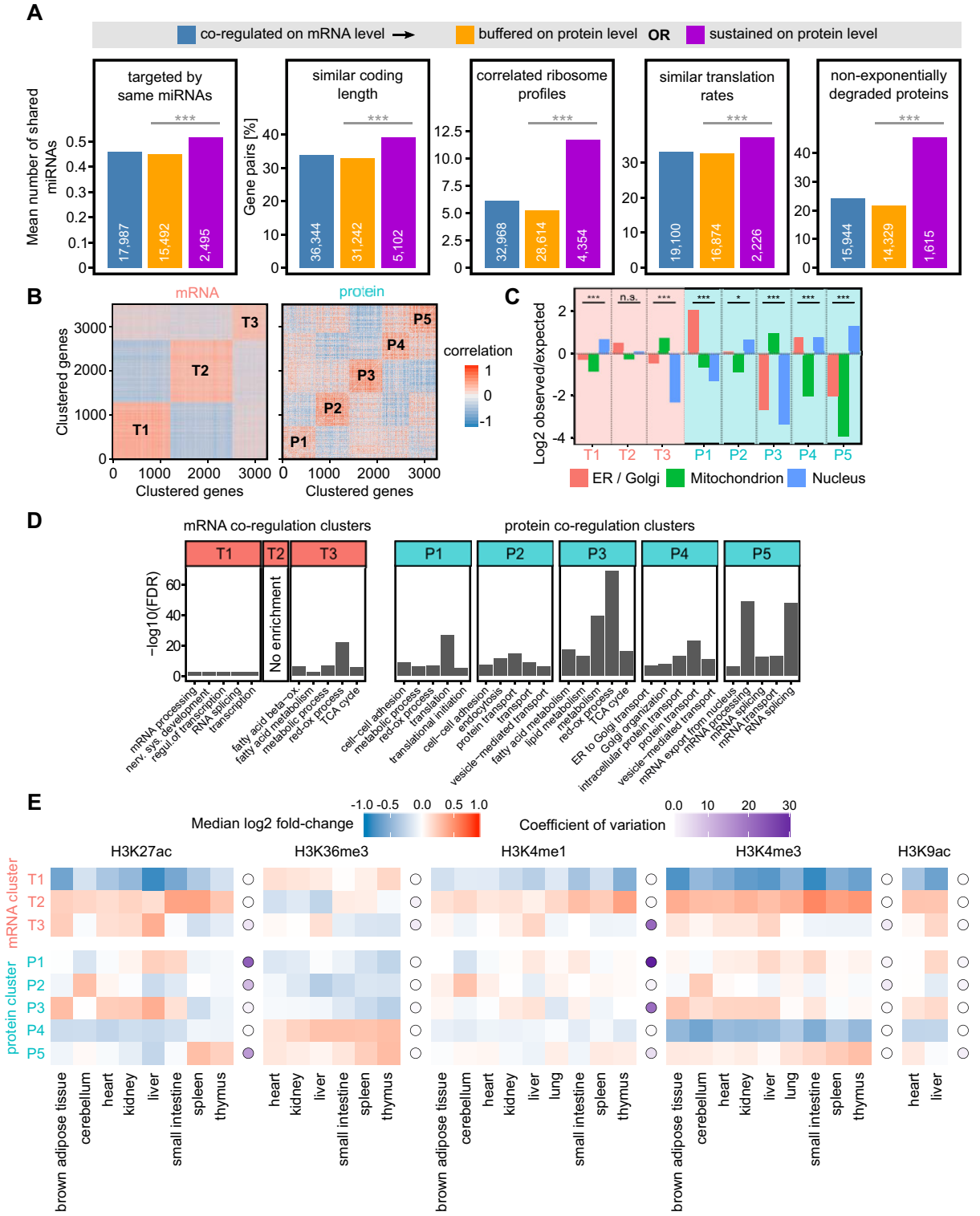


FIG. 2. **mRNA and protein coregulation clusters are functionally distinct and display different epigenetic signatures.** *A*, Analysis of post-transcriptional regulation of gene pairs coexpressed on mRNA level. Gene pairs with sustained coexpression on the protein level share on average more miRNA targeting than pairs with buffered coexpression on the protein level (Mann Whitney p value < 0.0001). Gene pairs were considered to have similar CDS length if the ratio of the longer sequence to the shorter was < 1.5. Gene pairs were considered to have correlated ribosome profiles if their ribosome occupancy profiles had Pearson correlation coefficient > 0.5 (Holm adj. p value < 0.001). Gene pairs were considered to have similar translation rates if the absolute log2 ratio of their translation rates was lower or equal 1. For the non-exponentially degraded proteins (NEDs) bar chart, gene pairs containing at least one NED were counted. *B*, K-means clustering of the

Protein Coregulation Clusters Are More Functional Than mRNA Coregulation Clusters—To group genes with similar coexpression patterns we used k-means clustering (Fig. 2*B*). This expands our analysis of coregulation from gene pairs to gene groups. This revealed specific coregulation patterns in which each cluster tends to be coregulated or antiregulated with other clusters (supplemental Fig. S3). Of the three transcript-based gene clusters, cluster T1 and T2 are anticorrelated. A similar anticorrelation was observed in human, which could be traced there to chromosome subcompartments A1 and A2 (7). Briefly, compartments are regions of the genome defined by 3D analysis of chromosome structure (31). Compartment A is characterized by active gene expression whereas compartment B mostly by suppressed gene expression. It was later discovered that both A and B compartments are divided further into subcompartments A1, A2 and B1 to B4, each with distinct epigenetic marks and spatial interaction patterns (32).

In the absence of equivalent high-resolution HiC data for mouse tissues we tested for epigenetic similarity within these clusters as epigenetic signatures closely link to chromatin subcompartments (32). Indeed, the epigenetic signatures of T1 and T2 clusters resemble those found in chromatin subcompartments A1 and A2 (see next paragraph). Notably, neither in mouse nor in human do the transcript-based gene clusters inform on protein coexpression. The marked exception is given by cluster T3 which displays coexpression behavior also at the protein level. Looking at the function of genes present in each of the clusters by performing subcellular localization (Fig. 2*C*) and Gene Ontology (33) term enrichment (Fig. 2*D*) reveals that cluster T3 is enriched for mitochondrial functions. This indicates large differences in the energetic needs of different tissues, which may require gene regulation at both the transcriptional and protein level. The five protein-based gene clusters correlate with each other to various degrees, with the anti-correlations of P2 *versus* P4 and P3 *versus* P5 being most pronounced. These likely reflect commitments of cell types to different large cellular processes (Fig. 2*D*). Interestingly, we observed a large overlap between the clusters T3 and P3. They had 734 and 686 members, respectively, and around half of the members were shared between them (365 genes). Similarly, to cluster T3, the protein cluster P3 was enriched in mitochondrial functions (Fig. 2*C*, 2*D*). This suggests that the coordination of mitochondrial protein coexpression could be tightly controlled already on the mRNA level.

Except for P3, the protein-based gene clusters are not reflected in transcript coexpression (supplemental Fig. S3). In summary, one of the three transcript-based gene clusters show some functional enrichment. However, all five protein coregulation clusters show well-defined subcellular localization patterns and functional GO term enrichments. As observed in other systems, protein coexpression links closer to function than transcript coexpression (7, 8).

We added a regulatory dimension to the expression data set by leveraging the ENCODE ChIP-seq data resources for nine different mouse tissues. This allowed us to estimate epigenetic variability of the gene pairs in the data. We calculated ChIP-seq signal enrichment for gene bodies belonging to the mRNA and protein coregulation clusters (Fig. 2*E*). Transcript clusters T1 and T2, which cover about 80% of the genes, maintain their epigenetic profile across all tissues with T2 being more enriched in activating marks compared with T1. While these two groups are defined through their chromatin state, they do not experience tissue specific regulation through epigenetic processes. This might be linked to chromatin subcompartments. Indeed, the epigenetic patterns of mouse clusters T1 and T2 closely resemble human chromatin subcompartments A2 and A1, respectively (7). This suggests a similar chromatin subcompartmentalization in mouse as is found in human. In contrast, transcript cluster T3 and most protein clusters display epigenetic variation across tissues indicating the action of an epigenetic program which is in line with epigenetic processes being involved in cell differentiation (34). It may initially surprise that protein clusters have epigenetic tissue-specific changes while transcript clusters T1 and T2 lack these (for example see H3K27ac or H3K4me1). This is consistent with subcompartments dominating the epigenetic signature that is associated with mRNA coexpression. It is worth keeping in mind that we analyze housekeeping genes, for which one would expect adjustments in expression rather than on/off changes and consequently only weak epigenetic influences. Interestingly, a strong between-cluster difference can be seen for the H3K36me3 mark which displays almost no variability between tissues for protein clusters. The H3K36me3 mark has been shown to be implicated in gene expression noise control through a mechanism of transcriptional burst frequency modulation (35) and to be enriched among noise-sensitive, highly expressed genes (36, 37). In full agreement with this, the mRNA cluster enriched in the H3K36me3 mark (T1) has significantly lower expression variability compared with other clusters (supplemental Fig. S4*A*).

mRNA and protein coexpression data. Three distinct mRNA clusters and five distinct proteins clusters explained ~50% of the variance in the respective data. *C*, mRNA coregulations clusters (T1–T3) have lower protein subcellular localization enrichments than protein coregulation clusters (P1–P5). The significance of enrichments/depletions in each cluster was tested using Pearson's Chi-squared test. ***p value < 0.0001, *p value < 0.05, n.s. = not significant. *D*, GO enrichment analysis of the genes in the mRNA and protein coregulation clusters. More GO terms are enriched in protein than in mRNA clusters. *E*, mRNA-based clusters T1 and T2 have uniform epigenetic signal distributions displaying little between-tissue variability as opposed to protein clusters which show large between-tissue and between-cluster variability. Epigenetic signal enrichment in tissue (squares), coefficient of variation for each histone mark (circle), color code as shown.

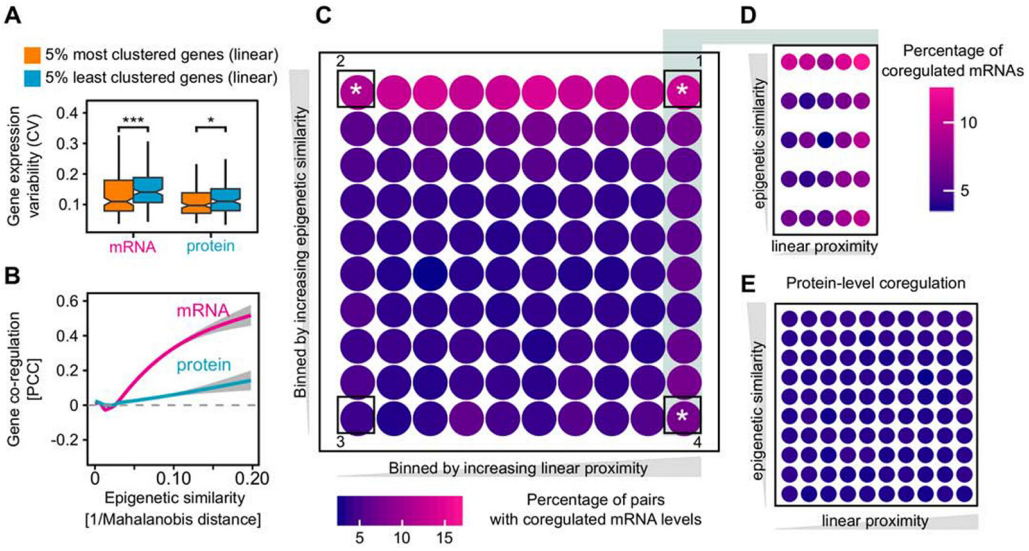


FIG. 3. **The impact of gene proximity and epigenetic similarity on mRNA- and protein-level coregulation.** *A*, Positional gene clustering reduces the expression variability on mRNA level. We calculated the expression variability (coefficient of variation, CV) of the 5% most and 5% least positionally clustered genes on the genome (*i.e.* considering their sequence proximity). The difference is significant (using Mann-Whitney test) on both mRNA level (***p value = 0.00029) and protein level (*p value = 0.019). When using 10 and 1% most and least clustered genes, we obtain the same statistical results as with 5% (data not shown). Boxplot drawn in the style of Tukey, *i.e.* box limits indicate the first and third quartiles, central lines the median, whiskers extend 1.5 times the interquartile range from the box limits. Notches indicate the 95% confidence interval for comparing medians. *B*, Gene coregulation increases with epigenetic similarity at the mRNA level, whereas it remains largely independent from epigenetic similarity at the protein level. *C*, Epigenetic similarity is the major driver of the mRNA coregulation. Gene pairs were considered coregulated if their mRNA level correlation was > 0.5 and the BH-adjusted p values < 0.05. The bins were created by dividing gene pair distances and epigenetic similarity (1/Mahalanobis distance) into 10 roughly equal sets. This yielded 100 unique bin combinations. The color signifies the percentage of coregulated mRNA in each bin. The mean gene pair distance in the left-most column is 115 Mb and 2 Mb in the right-most column. White stars (*) mark corner sectors which have significantly higher mRNA coexpression compared with an equal-sized random background sample as judged by Kolmogorov-Smirnov test. The procedure was repeated 1000 times. The mean p values for sectors 1, 2, 3 and 4 were 0, 10^{-13}, 0.039 and $6*10^{-9}$, respectively. p value of 0 is reported by the KS test for extremely low values. *D*, Effects in linear distance are confined to very close proximity. The 10 bins constituting the right-most column in Fig. 3*C* were extracted and magnified. The mean gene pair distance for the left-most column is 4 Mb and 240 Kb for the right-most column. *E*, Protein-level coregulation of housekeeping genes is not generally affected by epigenetic similarity or linear distance.

Curiously, we also observed a strong expression variability difference for protein clusters P4 and P5 which are enriched for H3K36me3 compared with other three protein clusters. However, it is not clear if the differences in H3K36me3 signal in mRNA and protein clusters are a cause of different expression variability or an effect of differences in the ongoing transcription.

Gene Clustering Reduces mRNA Expression Variability in Mouse Tissues—We determined the gene expression variability (coefficient of variation, CV) of the most and least densely clustered genes, considering sequence proximity (Fig. 3*A*). Transcript expression variability is reduced significantly for genes clustered in the genome sequence while the effect is less pronounced for protein expression variability. Importantly, although gene expression variability generally covariates with expression level, no difference in expression levels was observed here for the top and lowest 5% positionally clustered genes (53,000 and 56,000 mean TPM, respectively). As observed previously for yeast (38) and human (7) gene clustering may safeguard against accidental silencing and the resulting expression noise. However, gene expression variability is not exactly the same as bona fide gene expression noise. It is interesting therefore that our observations using global between-tissue variability of expression reflect the observations based on expression noise in its classical sense in other systems. As a further link of expression variability between tissues to noise, we noted a strong dependence of both mRNA and protein expression variability on H3K36me3 signal in gene bodies. Genes lacking H3K36me3 signal are the most variably expressed between the tissues whereas the opposite is true for genes with strong H3K36me3 signal (supplemental Fig. S4*B*). This resembles the role of this mark in expression noise control (36, 37).

Epigenetic Similarity Is the Main Driver of Nonfunctional mRNA Coexpression—Coexpression of close-by, unrelated genes can be driven by at least two distinct mechanisms. First, stochastic fluctuations between the on and off state of a chromatin domain can affect multiple genes simultaneously and lead to their coexpression (4, 39). In addition, coexpression can reflect a transcriptional "ripple effect," where the activation of one gene leads to the upregulation of other genes in its immediate neighborhood (5). We investigated

which of these factors drives non-functional mRNA coexpression across mouse tissues. To estimate which genes may be affected by the same chromatin fluctuations, we first determined the epigenetic profile of each gene, based on 7 histone marks in 9 different tissues reported by ENCODE. We then calculated the epigenetic similarity between all gene pairs using the Mahalanobis distance, which considers that some histone marks are codependent (exemplified in supplemental Fig. S5). As one might expect, we observed that correlation of mRNA abundances increases dramatically with increasing epigenetic similarity of their respective genes. Interestingly, the effect is largely buffered on the protein level (Fig. 3*B*). This suggests that many mRNA pairs are coactivated as a side-effect of their genes being in the same genomic neighborhood which in turn confers a specific epigenetic profile. To place the epigenetic similarity and coregulation into gene position context, we plotted the coregulation values as a function of both epigenetic similarity and a linear genomic separation of the gene pairs (Fig. 3*C*). Strikingly, epigenetic similarity drives mRNA coexpression irrespective of whether genes are far apart (Fig. 3*C*, sector 2) or close-by in the genome (sector 1). For the gene pairs that are on average within 2 Mb to each other, those that have very different epigenetic profiles are much less likely to be coexpressed than those with similar chromatin features (Fig. 3*C*, sector 4 *versus* 1). This is most likely an effect of global fluctuations of chromatin factors shown previously in yeast (40). Gene proximity only starts to be a driving factor for genes less than 240 Kb apart (Fig. 3*D*, right-most column) which agrees with previous observations of a local transcriptional ripple effect (5). Notably, most of this mRNA coexpression is non-functional, because the same group of genes show, on average, no coexpression at the protein level (Fig. 3*E*).

CONCLUSIONS

In an LCL cell line panel and in cancer samples, at homeostatic conditions much of mRNA coexpression is non-functional, *i.e.* does not affect protein coexpression and instead can be traced back to genome organization. We wondered how much coexpression of mRNA and proteins would be linked when comparing very different cellular states given by multiple fully differentiated tissues. mRNA coexpression is indeed more closely linked to function in mouse tissues than in homeostatic conditions, although protein coexpression is significantly more indicative of function. The epigenetic profiling of coexpression clusters revealed that mRNA coexpression is affected by two distinct epigenetic states, most likely reflecting the different genomic subcompartments in which they reside. As observed in homeostatic conditions, this broad positioning effect on mRNA coexpression is then buffered on the protein level. However, in mouse tissues the non-functional mRNA coexpression is linked more closely to epigenetic states than to linear gene proximity. Epigenetic differences between the tissues dwarf the linear proximity effect on coexpression. Notably, we chose to use housekeeping genes only as they conferred enough data points to be usable in this correlation-based study. It is not clear to what extent do the observations on housekeeping genes generalize to the rest of the genome. Taken together, our observations lend support to the notion of monitoring protein coexpression for functional genomics. However, to fully understand the impact of epigenetics on mRNA and protein coexpression and the underlying mechanisms, more in-depth experimental studies are needed.

Acknowledgements—We thank Laurence Hurst for critically reading the manuscript.

* This work was supported by the Wellcome Trust through a Senior Research Fellowship to JR (grant number 103139). The Wellcome Centre for Cell Biology is supported by core funding from the Wellcome Trust (grant number 203149).

[S] This article contains supplemental Figures and Tables.

¶ To whom correspondence should be addressed: Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, 13355 Berlin, Germany. Tel.: +49 30 314-72374; E-mail: juri.rappsilber@ed.ac.uk.

Other contacts: Piotr Grabowski, Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, 13355 Berlin, Germany; E-mail: grabowski@tu-berlin.de. Georg Kustatscher, Wellcome Centre for Cell Biology, University of Edinburgh, Edinburgh EH9 3BF, UK; E-mail: georg.kustatscher@ed.ac.uk.

Author contributions: P.G., G.K., and J.R. designed research; P.G., G.K., and J.R. performed research; P.G., G.K., and J.R. analyzed data; P.G., G.K., and J.R. wrote the paper.

REFERENCES

1. Hurst, L. D., Pál, C., and Lercher, M. J. (2004) The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* **5,** 299–310
2. Williams, E. J. B., and Bowles, D. J. (2004) Coexpression of neighboring genes in the genome of Arabidopsis thaliana. *Genome Res.* **14,** 1060–1067
3. Xu, C., Chen, J., and Shen, B. (2012) The preservation of bidirectional promoter architecture in eukaryotes: what is the driving force? *BMC Syst. Biol.* **6,** S21
4. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y., and Tyagi, S. (2006) Stochastic mRNA synthesis in mammalian cells. *PLos Biol.* **4,** e309
5. Ebisuya, M., Yamamoto, T., Nakajima, M., and Nishida, E. (2008) Ripples from neighbouring transcription. *Nat. Cell Biol.* **10,** 1106–1113
6. Batada, N. N., and Hurst, L. D. (2007) Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nat. Genet.* **39,** 945–949
7. Kustatscher, G., Grabowski, P., and Rappsilber, J. (2017) Pervasive coexpression of spatially proximal genes is buffered at the protein level. *Mol. Syst. Biol.* **13,** 937
8. Wang, J., Ma, Z., Carr, S. A., Mertins, P., Zhang, H., Zhang, Z., Chan, D. W., Ellis, M. J. C., Townsend, R. R., Smith, R. D., McDermott, J. E., Chen, X., Paulovich, A. G., Boja, E. S., Mesri, M., Kinsinger, C. R., Rodriguez, H., Rodland, K. D., Liebler, D. C., and Zhang, B. (2017) Proteome Profiling Outperforms Transcriptome Profiling for Coexpression Based Gene Function Prediction. *Mol. Cell Proteomics* **16,** 121–134
9. Lapek, J. D., Jr, Greninger, P., Morris, R., Amzallag, A., Pruteanu-Malinici, I., Benes, C. H., and Haas, W. (2017) Detection of dysregulated protein-association networks by high-throughput proteomics predicts cancer vulnerabilities. *Nat. Biotechnol.* **35,** 983–989
10. Geiger, T., Velic, A., Macek, B., Lundberg, E., Kampf, C., Nagaraj, N., Uhlen, M., Cox, J., and Mann, M. (2013) Initial quantitative proteomic map of 28 mouse tissues using the SILAC mouse. *Mol. Cell Proteomics* **12,** 1709–1722
11. ENCODEProject Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74

12. Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., and Soboleva, A. (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* **41,** D991–D995
13. Brosens, J. J., Salker, M. S., Teklenburg, G., Nautiyal, J., Salter, S., Lucas, E. S., Steel, J. H., Christian, M., Chan, Y.-W., Boomsma, C. M., Moore, J. D., Hartshorne, G. M., Sućurović, S., Mulac-Jericevic, B., Heijnen, C. J., Quenby, S., Koerkamp, M. J. G., Holstege, F. C. P., Shmygol, A., and Macklon, N. S. (2014) Uterine selection of human embryos at implantation. *Sci. Rep.* **4,** 3894
14. Kim, H., Toyofuku, Y., Lynn, F. C., Chak, E., Uchida, T., Mizukami, H., Fujitani, Y., Kawamori, R., Miyatsuka, T., Kosaka, Y., Yang, K., Honig, G., van der Hart, M., Kishimoto, N., Wang, J., Yagihashi, S., Tecott, L. H., Watada, H., and German, M. S. (2010) Serotonin regulates pancreatic beta cell mass during pregnancy. *Nat. Med.* **16,** 804–808
15. Mustafi, D., Kevany, B. M., Genoud, C., Okano, K., Cideciyan, A. V., Sumaroka, A., Roman, A. J., Jacobson, S. G., Engel, A., Adams, M. D., and Palczewski, K. (2011) Defective photoreceptor phagocytosis in a mouse model of enhanced S-cone syndrome causes progressive retinal degeneration. *FASEB J.* **25,** 3157–3176
16. Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J. R., and Mann, M. (2010) Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat. Methods* **7,** 383–385
17. Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Keenan, S., Lavidas, I., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Nuhn, M., Parker, A., Patricio, M., Pignatelli, M., Rahtz, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Birney, E., Harrow, J., Muffato, M., Perry, E., Ruffier, M., Spudich, G., Trevanion, S. J., Cunningham, F., Aken, B. L., Zerbino, D. R., and Flicek, P. (2016) Ensembl 2016. *Nucleic Acids Res.* **44,** D710–D716
18. Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4,** 1184–1191
19. Moore, M. J., Scheel, T. K. H., Luna, J. M., Park, C. Y., Fak, J. J., Nishiuchi, E., Rice, C. M., and Darnell, R. B. (2015) miRNA–target chimeras reveal miRNA 3′-end pairing as a major determinant of Argonaute target specificity. *Nat. Commun.* **6,** 8864
20. Janich, P., Arpat, A. B., Castelo-Szekely, V., Lopes, M., and Gatfield, D. (2015) Ribosome profiling reveals the rhythmic liver translatome and circadian clock regulation by upstream open reading frames. *Genome Res.* **25,** 1848–1859
21. Revelle, W. R. (2017) psych: Procedures for personality and psychological research.
22. Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011) Global quantification of mammalian gene expression control. *Nature* **473,** 337–342
23. McShane, E., Sin, C., Zauber, H., Wells, J. N., Donnelly, N., Wang, X., Hou, J., Chen, W., Storchova, Z., Marsh, J. A., Valleriani, A., and Selbach, M. (2016) Kinetic Analysis of Protein Stability Reveals Age-Dependent Degradation. *Cell* 167, 803–815.e21
24. The UniProt Consortium. (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45,** D158–D169
25. Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4,** 44–57
26. RCore Team. (2017) R: A Language and Environment for Statistical Computing.
27. Wickham, H. (2009) *ggplot2: Elegant graphics for data analysis,* Springer, New York
28. Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson Å Kampf. C, Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., and Pontén, F. (2015) Proteomics. Tissue-based map of the human proteome. *Science* **347,** 1260419
29. Wilczynska, A., and Bushell, M. (2015) The complexity of miRNA-mediated repression. *Cell Death Differ.* **22,** 22–33
30. Vogel, C., de Sousa Abreu, R., Ko, D., Le, S., Shapiro, B. A., Burns, S. C., Sandhu, D., Boutz, D. R., Marcotte, E. M., and Penalva, L. O. (2010) Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol. Syst. Biol.* **6,** 400
31. Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293
32. Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., and Aiden, E. L. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159,** 1665–1680
33. Gene Ontology Consortium. (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **43,** D1049–D1056
34. Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., and Bernstein, B. E. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49
35. Kim, J. K., and Marioni, J. C. (2013) Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol.* **14,** R7
36. Wu, S., Li, K., Li, Y., Zhao, T., Li, T., Yang, Y.-F., and Qian, W. (2017) Independent regulation of gene expression level and noise by histone modifications. *PLoS Comput. Biol.* **13,** e1005585
37. Faure, A. J., Schmiedel, J. M., and Lehner, B. (2017) Systematic Analysis of the Determinants of Gene Expression Noise in Embryonic Stem Cells. *Cell Syst.* **5,** 471–484.e4
38. Hurst, L. D., Williams, E. J. B., and Pál, C. (2002) Natural selection promotes the conservation of linkage of coexpressed genes. *Trends Genet.* **18,** 604–606
39. Batada, N. N., Urrutia, A. O., and Hurst, L. D. (2007) Chromatin remodelling is a major source of coexpression of linked genes in yeast. *Trends Genet.* **23,** 480–484
40. Newman, J. R. S., Ghaemmaghami, S., Ihmels, J., Breslow, D. K., Noble, M., DeRisi, J. L., and Weissman, J. S. (2006) Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* **441,** 840–846

Manuscript 3. “Multiclassifier Combinatorial Proteomics of Organelle Shadows at the Example of Mitochondria in Chromatin Data”

Pages 41 - 49

Manuscript available online, DOI: 10.1002/pmic.201500267

RESEARCH ARTICLE

Multiclassifier combinatorial proteomics of organelle shadows at the example of mitochondria in chromatin data

Georg Kustatscher[1]*, *Piotr Grabowski*[1,2]* *and Juri Rappsilber*[1,2]

[1] Wellcome Trust Centre for Cell Biology, University of Edinburgh, UK
[2] Department of Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, Berlin, Germany

Received: July 2, 2015
Revised: September 3, 2015
Accepted: October 15, 2015

Subcellular localization is an important aspect of protein function, but the protein composition of many intracellular compartments is poorly characterized. For example, many nuclear bodies are challenging to isolate biochemically and thus remain inaccessible to proteomics. Here, we explore covariation in proteomics data as an alternative route to subcellular proteomes. Rather than targeting a structure of interest biochemically, we target it by machine learning. This becomes possible by taking data obtained for one organelle and searching it for traces of another organelle. As an extreme example and proof-of-concept we predict mitochondrial proteins based on their covariation in published interphase chromatin data. We detect about ⅓ of the known mitochondrial proteins in our chromatin data, presumably most as contaminants. However, these proteins are not present at random. We show covariation of mitochondrial proteins in chromatin proteomics data. We then exploit this covariation by multiclassifier combinatorial proteomics to define a list of mitochondrial proteins. This list agrees well with different databases on mitochondrial composition. This benchmark test raises the possibility that, in principle, covariation proteomics may also be applicable to structures for which no biochemical isolation procedures are available.

Keywords:
Chromatin / Machine learning / Mitochondria / Organelle / Systems biology

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

Eukaryotic cells contain organelles and other specialized compartments, whose protein composition can be analyzed by proteomics to provide important clues regarding their biological function [1, 2]. Organelle proteomics approaches traditionally depend on the biochemical isolation of the analyzed structure, which can be relatively straightforward for membrane-enclosed organelles such as mitochondria [3]. However, the majority of spatial compartments cannot be adequately enriched for conclusive analysis, as their isolates may be contaminated with too many functionally unrelated proteins that copurify. Alternative approaches have therefore been developed to infer the composition of organelles that cannot be purified to homogeneity. For example, subtractive [4] and quantitative [5] proteomics approaches have been employed to distinguish between genuine components and contaminants in biochemical isolates of nuclear envelopes and lipid rafts, respectively. Partial enrichment combined with quantitative proteomic analysis was used to broadly categorize the cell into cytoplasm, nucleus, and nucleolus [6]. Protein correlation profiling was developed to study the composition of the centrosome [7] and later provided a mammalian

Correspondence: Professor Juri Rappsilber, Wellcome Trust Centre for Cell Biology, Institute of Cell Biology, University of Edinburgh, Michael Swann Building 4.18a, Max Born Crescent, Edinburgh EH9 3BF, UK
E-mail: Juri.Rappsilber@ed.ac.uk

Abbreviations: **ChEP**, chromatin enrichment for proteomics; **MCCP**, multiclassifier combinatorial proteomics

*These authors contributed equally to this work.

Significance of the study

This study introduces a new concept for organelle proteomics. Until now, specific biochemical enrichment was paramount to study biological structures by proteomics. However, many compartments in the cell simply cannot be isolated or even partially separated from the rest of the cell. Examples for this include chromatin, which is highly charged and invariably "absorbs" functionally unrelated proteins, and nuclear bodies that are not surrounded by a membrane and most likely disintegrate upon cell lysis. We present here a method that may overcome such challenges in the future. The basic idea is that machine-learning can identify organelle-specific patterns across many comparative proteomics studies, even if the organelle was just present as contamination in the original experiment. As a proof-of-principle we identified mitochondrial proteins from chromatin proteomics data. While we do not have enough data at the moment to define the entire mitochondrial proteome in this way, our experiment shows that enriching an organelle through biochemical fractionation is no longer a strict requirement to analyze its composition. We envisage that this method may be useful to study a multitude of nonpurifiable biological structures in the future.

organelle map [8]. Using a related method, localization of organelle proteins by isotope tagging, proteins were assigned to the various compartments of the endomembrane system, which cannot be efficiently distinguished biochemically [9].

When analyzing mitotic chromosomes we also encountered an abundant presence of background proteins [10]. Importantly, mitotic chromosomes are large and highly charged, attracting many functionally unrelated proteins, and thus are physically contaminated themselves. This made it difficult to identify contaminants using the existing fractionation-based procedures. We therefore proposed a machine learning approach, multiclassifier combinatorial proteomics (MCCP), as a solution. Taking the outcome of multiple proteomic analyses of mitotic chromosomes that were done under biochemically or genetically distinct conditions, and integrating those by Random Forest analysis provided a ranked list of protein components of mitotic chromosomes. Interphase chromatin is another example of a specialized functional compartment whose biochemical isolates remain highly impure [11]. Working with partially purified material, we used MCCP to infer the protein composition of interphase chromatin from biological covariation. For this we analyzed chromatin-enriched samples from a wide variety of biological conditions and showed that proteins with well-known chromatin functions tend to respond in a similar way to various perturbations, such as drug treatments. We subsequently used a machine learning algorithm to capture the covariation pattern corresponding to chromatin factors. The resulting model allowed us to predict hundreds of new potential interphase chromatin proteins simply based on their covariation with already known chromatin proteins.

Some compartments may be inherently unstable in vitro. For example, it has been proposed that many intracellular bodies represent liquid droplets that form by phase transition from the surrounding cyto- or nucleoplasm [12]. Such compartments may be very difficult or even impossible to purify biochemically, and presumably would start to disintegrate after cell lysis. Therefore, new approaches may be required to determine their protein composition. Possibly, also here a solution could come from machine learning.

One conclusion from our analysis of interphase chromatin was that covariation with reference proteins was more accurate than biochemical enrichment in identifying chromatin components. This raised the intriguing possibility that biochemical enrichment may not be an essential element of determining the composition of cellular structures by proteomics. To push this hypothesis to its extreme we wondered if an entirely untargeted organelle could be defined through its changing coappearance in the analysis of chromatin. This would offer a way to study the composition of nonpurifiable compartments, especially that of many elusive nuclear bodies that may stick to chromatin when it is isolated but that cannot be isolated on their own.

To test the hypothesis that covariation in proteomic datasets can be the central element of studying the composition of cellular structures by proteomics, we attempted to define the composition of mitochondria on the basis of our chromatin proteomics dataset. Our intention was not to present an alternative or even superior way of analyzing mitochondria but simply to use mitochondria as a test system for other organelles or structures that challenge current analysis approaches. Mitochondria are large, well defined, and not functionally linked to chromatin in any obvious way, but are frequently part of the background of our chromatin enrichment procedure. We defined a high-quality reference set of mitochondrial proteins and used this to train a machine-learning algorithm to spot other mitochondrial proteins in our chromatin dataset. The results agreed well with the current consensus of which proteins are in mitochondria. We could not expect to obtain a comprehensive mitochondrial protein inventory, because only ⅓ of the known mitochondrial proteins were detected in our chromatin samples. However, this proof-of-principle experiment demonstrates the possibility that targeted biochemical enrichment may be optional and not essential for defining organelles. Subcellular localization may be predicted through covariation, thus allowing targeting a structure during data analysis rather than experimentally.

2 Materials and methods

2.1 Chromatin proteomics data

Proteomic analyses of interphase chromatin were described previously [11]. For this project we only considered 45 SILAC ratios comparing chromatin under different biological conditions. Only those 4565 proteins with values in at least ten out of all 45 chromatin proteomics experiments were considered (Supporting Information Table 1). In brief, these experiments consisted of human cell lines grown in SILAC medium and subjected to various perturbations, such as treatment with drugs, growth factors, or irradiation. They also include SILAC-based comparisons of different cell types and cell-cycle phases. In order to preferentially detect chromatin-bound proteins, all samples were subjected to the chromatin enrichment for proteomics (ChEP) procedure [13]. Tryptic digests were analyzed by LC-MS/MS on an LTQ-Orbitrap or LTQ-Orbitrap Velos (Thermo Fisher Scientific). These samples are described as "biological classifier" experiments in Table 1 of Kustatscher et al. [11] in more detail. Raw data have been deposited in the PRIDE [14] repository (www.ebi.ac.uk/pride) as part of the dataset PXD000493 (for this study we only used a subset of these data, namely experiments 3–7 and 18–35).

2.2 High-confidence mitochondrial reference protein set

We compiled a set of well-studied, high-confidence mitochondrial reference proteins. As a starting point, we downloaded all 1065 human proteins that mapped to "mitochondrion" in Uniprot's [15] subcellular localization database (www.uniprot.org/locations) and that were part of Swiss-Prot. We kept only proteins with an annotation score of at least four out of five. To remove proteins with ambiguous localization we filtered out proteins whose localization annotation matched the following keywords: nucleus, reticulum, Golgi, secreted, cytosol, peroxisome, and cell projection. This shortlisted 653 proteins, for which we manually evaluated Uniprot and GO [16] annotations and, where necessary, searched the available literature to extract a final list of 486 bona fide mitochondrial proteins with no reported functions elsewhere in the cell. Of these 486 mitochondrial reference proteins, 172 (35%) were detected in the chromatin proteomics dataset (Supporting Information Table 2).

2.3 Random Forest prediction of mitochondrial proteins

For supervised machine learning we used the Weka 3.7 [17] implementation of the Random Forest algorithm [18], executed through an in-house workflow built on the KNIME data analytics platform [19]. This implementation of Random Forest does not impute missing values. The Random Forest was constructed using 500 trees, six random features at each split and an unlimited maximum tree depth. The high-confidence mitochondrial reference protein set was used as positive training data. Negative training data were randomly selected from all nonmitochondrial proteins in our chromatin proteomics dataset (for this purpose, nonmitochondrial was defined as having no such annotation in GO or Uniprot). To avoid using unbalanced training data, only 172 negative training instances were selected, i.e. the same number as positive training instances. However, rather than constructing just one Random Forest, the workflow was executed ten times with different randomly drawn negative training data. The average Random Forest scores and their standard deviation were collected. Prediction accuracy was assessed in two different ways. The out-of-bag error, an unbiased estimate of the test set error inbuilt to the algorithm, was collected. In addition, the training dataset was cross-validated 100-fold, and the cross-validated data were used to judge performance based on the area under a ROC curve. Random Forest scores, including the cross-validated scores for the mitochondrial training dataset, are reported in the Supporting Information Table 2.

2.4 Comparison with other mitochondrial datasets

We compared our mitochondrial predictions to five different sources of mitochondrial annotation. The human version of MitoCarta [20] was downloaded on May 1, 2015 from www.broadinstitute.org/pubs/MitoCarta. GO annotations [16] were downloaded from QuickGO [21] using the identifier "mitochondrion" (GO:0005739), restricted to the qualifiers "contributes to," "colocalizes with" and "none". Only annotations with evidence level "manual experimental" were considered. The third external mitochondrial protein set consisted of proteins annotated as mitochondrial in Uniprot and was downloaded as described for the high-confidence mitochondrial reference protein set, without filtering against multiple localizations. An immunofluorescence-based list of proteins with mitochondrial localization was retrieved from the Human Protein Atlas [22], omitting proteins with "uncertain" reliability status. The fifth reference set consisted of mitochondrial matrix proteins identified via spatially restricted enzymatic tagging and MS [23].

2.5 Further data processing and visualization

Data were processed using R version 3.2 [24]. ID conversions, where necessary, were done using Bioconductor biomaRt package [25]. ROC curves were generated and visualized using ROCR R package [26]. Data analysis plots were prepared using ggplot2 R package [27].

3 Results and discussion

Covariation proved to be a successful concept in defining core proteins of mitotic and interphase chromatin when starting from multiple but impure proteomic lists of these structures [10, 11]. To test if this approach could be expanded to structures that have not been the target of experimental data collection we attempted here to define mitochondria through their coappearance in chromatin analyses. We chose mitochondria for this proof-of-principle experiment, because this organelle has been well-defined through other studies and thus allows us to evaluate the success of our approach. Mitochondria are membrane-enclosed and thus presumably clearly defined, and their composition has been studied for decades with many different experimental approaches, including proteomics. This makes them a good reference point to assess the performance of novel organelle proteomics approaches. Moreover, mitochondria contain more than thousand proteins [20], several hundred of which are detected in our chromatin samples, providing a reasonably sized test set for our setup. It should be noted that some mitochondrial proteins, such as prohibitins, have genuine additional functions as nuclear transcription factors and so would be expected to be found in chromatin [28]. However, the majority of mitochondrial proteins in our assay likely become associated with chromatin in an artificial way at some point during chromatin enrichment, i.e. they are likely contaminants in our chromatin analysis.

3.1 Biological perturbations affect the abundance of mitochondrial proteins in chromatin samples

We observed that the presence of mitochondria in chromatin samples tends to change—very gently—in response to biological perturbations (Fig. 1). This is initially counterintuitive as one would expect from a contaminating protein that its presence would be largely unaffected by biological changes in chromatin. Surprisingly, mitochondrial proteins become mildly but significantly depleted ($p = 1.13 \times 10^{-10}$) in chromatin samples after treating cells with TNFα (Fig. 1A), they are more abundant ($p = 7.26 \times 10^{-22}$) in chromatin samples from HepG2 than HEK293 cells (Fig. 1B) and they are depleted ($p = 7.95 \times 10^{-30}$) from chromatin following 4-hydroxytamoxifen treatment (Fig. 1C). Indeed, in most comparative chromatin proteomics experiments, we find that mitochondria are slightly enriched or depleted in one condition compared to the other (Supporting Information Table 3). These changes are likely due to the primary or secondary effects of a perturbation on the cell, although we can only speculate about the precise mechanisms involved. For example, alterations in chromatin structure may affect the association of background proteins, leading to increased or decreased copurification of mitochondria with chromatin under different conditions. In addition, the number of mitochondria per cell may also be altered in some experiments, e.g. when comparing different cell types. While it is difficult to pinpoint the exact reasons for mitochondrial abundance variation in chromatin samples, we set out to test whether these changes can be exploited to study mitochondrial proteins.

3.2 Mitochondria are not major contaminants in chromatin samples

To ensure that mitochondria are a valid initial test system for our method, we first confirmed that mitochondria are not preferentially coenriched with chromatin. First, we noted that mitochondrial proteins are nearly an order of magnitude less abundant than chromatin proteins in these samples (Fig. 1D). To further confirm their status as contaminants, we turned to our chromatin proteome study, in which we assigned probabilities for genuine chromatin-based functions to human proteins. As expected, the vast majority of mitochondrial proteins (94%) are not predicted to have a functional association with chromatin (Supporting Information Fig. 1A). Finally, we tested how mitochondrial abundance in chromatin samples compares to that of various other organelles and common contaminants, such as ribosomes, the cytoskeleton and the Golgi apparatus. In fact, mitochondria are the least abundant of the tested chromatin contaminants (Supporting Information Fig. 1B).

3.3 Covariation in chromatin samples can predict mitochondrial proteins

We previously observed coordinated bulk behavior for chromatin proteins versus background proteins across various biological situations [11]. This covariation of chromatin factors allowed us to construct a comprehensive inventory of interphase chromatin. We defined a reference set of known chromatin factors and then used a Random Forest machine learning algorithm to identify proteins with similar behavior across various chromatin proteomics experiments. Now, we tested whether this approach could also capture a mitochondria-specific pattern across the same set of chromatin proteomics experiments.

We first assembled a high-confidence set of mitochondrial proteins. We started from a list of proteins annotated as mitochondrial in Uniprot and removed all entries with potentially ambiguous subcellular localization, such as mitochondrial proteins with additional reported functions in the endoplasmic reticulum or elsewhere in the cell. For this we considered information from Uniprot, GO, and the primary literature. We further removed proteins which were generally not well characterized, and could therefore not be considered bona fide mitochondrial proteins. Of the remaining 486 proteins we observed 172 (35%) in our data. We also sought to define a high-confidence set of nonmitochondrial proteins without introducing a bias. Such a bias could result from simply

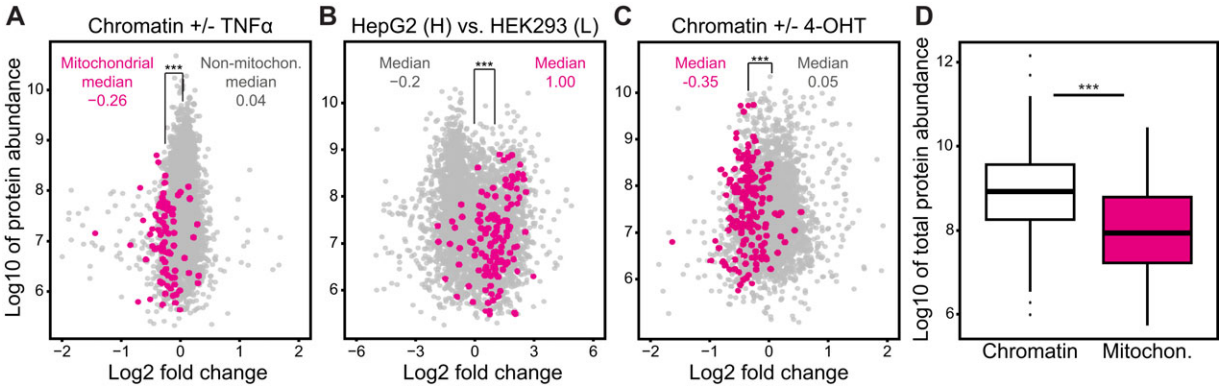


Figure 1. Mitochondrial proteins in interphase chromatin samples. (A–C) Mitochondrial proteins (magenta) are present in chromatin proteomics data, and are up- or downregulated in response to biological perturbations. For example, they are downregulated after treating HeLa cells for 10 min with TNFα compared to untreated controls (A). They are upregulated in chromatin samples purified from HepG2 as opposed to HEK293 cells (B). Mitochondria are also depleted from chromatin samples after treating estradiol-treated MCF7 cells with 4-hydroxytamoxifen (4-OHT) (C). The fold change is the SILAC ratio and protein abundance is the sum of measured peptide intensities. The significance of mitochondrial fold-changes was evaluated by *t*-test and yielded *p*-values < 0.001 in all three cases, as illustrated by the triple asterisks. (D) Boxplot of protein abundances showing that chromatin proteins are nearly an order of magnitude more abundant than mitochondrial (Mitochon.) proteins, supporting the contaminant status of the latter. The sum of protein intensities measured across all experiments is plotted. The intensity difference is highly significant according to a *t*-test (*p*-value $= 5.4 \times 10^{-32}$).

selecting nuclear proteins, for example. We solved this by drawing nonmitochondrial proteins randomly from all proteins in our dataset, except from proteins that had mitochondrial annotations in either Uniprot or GO.

We then conducted a supervised machine learning experiment based on the Random Forest algorithm [18] to distinguish mitochondrial from nonmitochondrial proteins using a publically available chromatin proteomics dataset (ProteomeXchange PXD000493) [11]. The dataset was obtained by analyzing chromatin-enriched samples from human cell lines grown in SILAC medium and subjected to various perturbations, such as treatment with drugs, growth factors, or irradiation. They also include SILAC-based comparisons of different cell types and cell-cycle phases. In order to preferentially detect chromatin-bound proteins, all samples were subjected to the ChEP procedure [13]. The chromatin dataset comprised 23 double- and triple-SILAC experiments with 45 SILAC ratios in total (Supporting Information Table 1). The Random Forest was trained using the reference sets of mitochondrial and nonmitochondrial proteins. For the nonmitochondrial training set to be representative of most nonmitochondrial cellular compartments we would have had to use significantly more than 172 proteins, as we used for the mitochondrial training set. However, using unbalanced training data skews the resulting scores. We therefore opted to train ten Random Forests, each time with the same 172 mitochondrial proteins but a different set of 172 randomly chosen nonmitochondrial training proteins. We collected the average Random Forest scores for each protein. This approach has the advantage of using a balanced training set and still sample a large fraction of all nonmitochondrial proteins to minimize prediction bias. In addition, the standard deviation of the score across the ten different Random Forest models reveals the impact of the choice of nonmitochondrial training proteins. The resulting set of Random Forests could distinguish between the known mitochondrial and nonmitochondrial training proteins very well, as indicated by the mean out-of-bag error of 0.1 ± 0.008. This shows that we can identify mitochondrial proteins only based on their SILAC ratios across many chromatin proteomics experiments.

We next performed 100-fold cross-validation to determine reliable prediction scores for our high-confidence mitochondrial proteins. This means we constructed 100 Random Forests and in each left out a different 1% of the reference data, using the model generated with the remaining 99% to obtain unbiased prediction scores for these proteins. This allowed us to use a ROC curve to estimate the model's performance, in addition to the inbuilt out-of-bag error estimate of the Random Forest algorithm. The mean area under the ROC curve we obtained was 0.96 (Fig. 2A). This confirms the high accuracy of our prediction already indicated by the low out-of-bag error.

In addition to our reference mitochondrial proteins, many other proteins with known mitochondrial functions received high Random Forest scores (Fig. 2B). To evaluate our prediction accuracy in a systematic way, we searched for false positive predictions among our top hits. For this we manually assessed the available literature and labeled proteins as false positives if they were well-characterized but lacked evidence for mitochondrial localization. At a Random Forest score cutoff of 0.69 we had 169 proteins of which 18 were clearly not mitochondrial (~10% false positives). Of the remaining 151 proteins (Fig. 2B), 94 are part of our high-confidence mitochondrial protein set and an additional 51 proteins are known to be mitochondrial. Six proteins were poorly or ambiguously annotated. For example, the prolyl hydroxylase LEPRE1 has

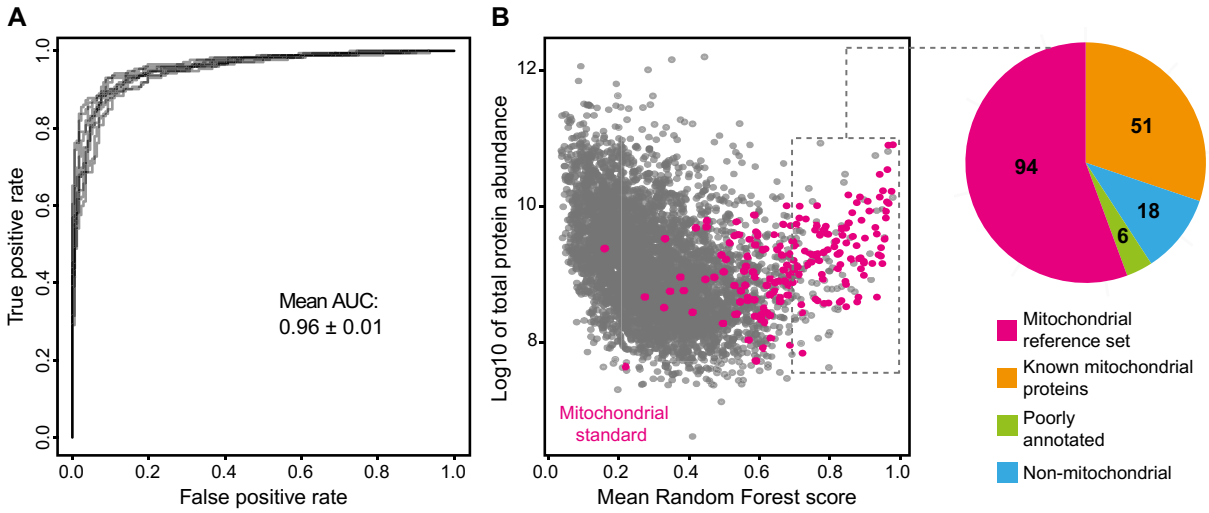


Figure 2. A Random Forest model can predict mitochondrial proteins based on their covariation in chromatin proteomics data. (A) High accuracy of mitochondrial prediction is shown by ROC curves derived from the 100-fold cross-validated mitochondrial and nonmitochondrial reference set. The ten curves correspond to ten Random Forests generated with different negative training data, highlighting the robustness of the Random Forest model. AUC: area under the curve. (B) Random Forest scores for the 4565 proteins (gray) in our analysis. High-confidence mitochondrial reference proteins (magenta) are heavily enriched toward higher scores. The pie-chart shows the manual annotation of proteins within the dashed rectangle, corresponding to a score cut-off of 0.69. Most proteins are either part of our high-confidence mitochondrial reference set or other known mitochondrial proteins. Six proteins were poorly annotated. 18 proteins were classified as nonmitochondrial, i.e. they are well-annotated but no evidence for mitochondrial function exists. This group was used to estimate that we have about 10% false positives at this score cut-off.

one isoform that is thought to be secreted [29] and another one that may reside in mitochondria [30], but our data do not allow us to distinguish between the two. The other five proteins are candidates for novel mitochondrial proteins, warranting further study into their biological function.

3.4 Mitochondria predictions agree well with established mitochondrial protein inventories

To determine the specificity and sensitivity of our approach in more detail, we compared its predictions to existing mitochondrial annotation data (Fig. 3). The most comprehensive inventory of mitochondrial proteins yet, MitoCarta, combined proteomic analysis of isolated mitochondria with GFP tagging and microscopy, and included additional genome-scale datasets such as the occurrence of mitochondrial targeting sequences [20]. The vast majority of proteins that receive high mitochondrial scores in our study are indeed found in MitoCarta, confirming the high specificity of our predictions (Fig. 3A). There is also strong enrichment of MitoCarta proteins toward high Random Forest scores. However, a number of MitoCarta proteins do not score high in our approach, raising the possibility that "prediction by covariation" may lack sensitivity. Alternatively, low-scoring proteins in our model may have been falsely assigned to mitochondria by classical proteomic approaches, for example due to an artificial copurification with mitochondria-enriched biochemical fractions. To test this possibility, we followed three separate lines of evidence.

First, we compared MitoCarta's confidence measure, the Maestro score, to our Random Forest score. MitoCarta entries which scored low in our analysis also tend to have been assigned to MitoCarta with lower confidence (Supporting Information Fig. 2). Next, we compared our predictions to a second, independent proteomic dataset that targeted proteins of the mitochondrial matrix rather than the entire mitochondrion [23]. In this approach, a genetically modified peroxidase enzyme is fused to a localization signal that specifically targets it to the mitochondrial matrix, where it biotinylates proteins in close physical proximity. This method results in very high specificity, because the inner mitochondrial membrane acts as a barrier confining the biotin label to matrix proteins. Interestingly, when compared to our Random Forest predictions, there are far fewer low-scoring proteins among mitochondrial factors identified in this way (Fig. 3B). This is also exemplified by a shift of median Random Forest score from 0.60 for MitoCarta proteins to 0.76 for mitochondrial proteins listed by Rhee et al. [23].

For a third specificity test, we compiled a consensus list of mitochondrial proteins by integrating four subcellular localization databases: MitoCarta, Uniprot, GO, and the Human Protein Atlas [15, 16, 20, 22]. There was complete agreement among the four databases on 245 proteins. One hundred forty-three of these we observed in our study. Similar to the matrix annotations from Rhee et al. [23], we find that the vast majority of these 143 consensus proteins rank very

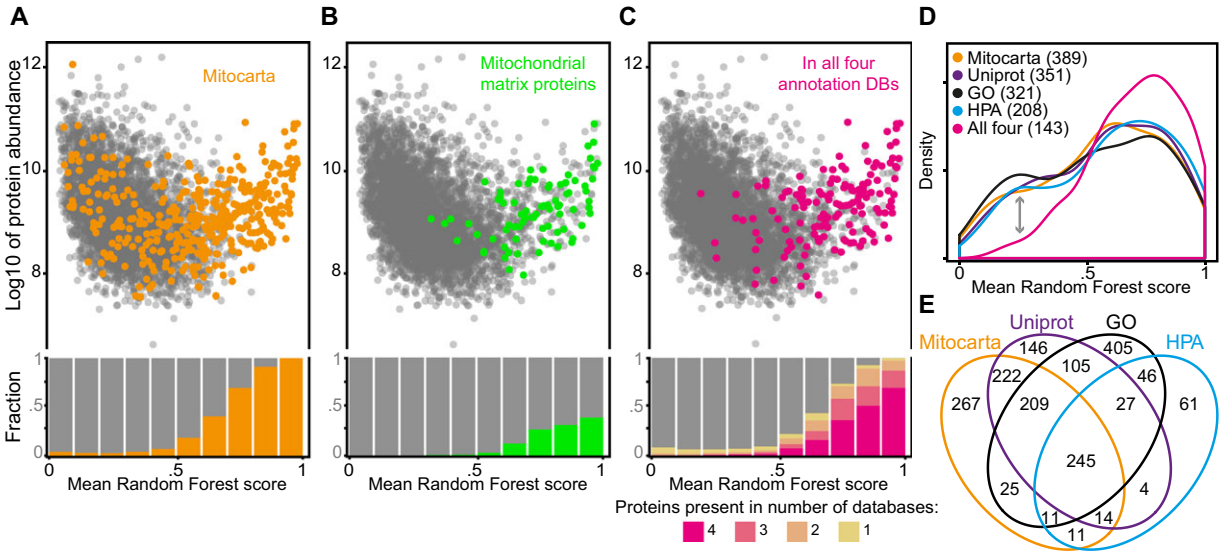


Figure 3. Covariation-based prediction evaluated by comparison to existing mitochondrial protein inventories. (A) Mitochondrial prediction for all 4565 proteins is shown as their Random Forest machine learning score. Proteins that are present in MitoCarta [20] are highlighted in orange. There is a strong enrichment of MitoCarta proteins toward high Random Forest scores (see bar chart). (B) Same plot but highlighting proteins in green that were specifically assigned to the mitochondrial matrix by Rhee et al. [23]. Note that very few of these annotations receive low Random Forest scores. Many high scoring proteins are mitochondrial but not in Rhee et al.'s [23] matrix proteome. (C) Proteins in magenta are annotated as mitochondrial in four different databases: MitoCarta, Uniprot, GO, and the Human Protein Atlas (HPA). These overlapping, high-confidence annotations include fewer low-scoring predictions than proteins found in only 1, 2, or 3 of these databases (see bar chart). (D) Distribution of mean Random Forest scores for proteins annotated as mitochondrial in either of the four indicated databases. Individually, all databases show a bimodal distribution. Restricting the analysis to the overlapping annotations shows a marked reduction in low-scoring annotations (see arrow). This indicates that such proteins are not just missed in our prediction due to lack of sensitivity, but are also not supported by other databases. (E) Venn diagram depicting the overlap of mitochondrial annotations in the four databases, including proteins not detected in our dataset.

high in our predictions (median Random Forest score 0.74) (Fig. 3C). Interestingly, any individual database contains a number of mitochondrial annotations that receive low scores in our assay (Fig. 3D). Increasing the number of databases that must agree on a protein to be mitochondrial decreases the number of low scoring annotations and improves the median Random Forest score (any database: 0.27, any two databases 0.46, any three databases: 0.63, all four databases: 0.74).

These three points suggest that our Random Forest analysis succeeds in recognizing bona fide components of mitochondria. Scoring low in our analysis indicates that a protein is less likely to be a genuine component of mitochondria. A conclusive evaluation of false negatives in our analysis is complicated by an absence of large consensus on mitochondrial proteins. A total of 1798 proteins are labeled "mitochondrial" by at least one database while the four databases agree on only 245 (Fig. 3E). However, two reasons could account for genuine mitochondrial proteins scoring low in our assay. First, the accuracy of the Random Forest classification depends on the number of experiments available to it, so increasing the number of input experiments will increase performance further. Also, we cannot expect to identify "conditional" mitochondrial factors, i.e. proteins that only localize to mitochondria under certain biological conditions. This is because such proteins may have a predominant function elsewhere in the cell and therefore not covary with mitochondrial reference proteins.

Due to the low coverage of mitochondrial proteins in the chromatin dataset, we are unable to make predictions on the majority of the estimated 1129 mitochondrial proteins [20]. For example, we detected 389 (38%) of the 1013 proteins in human MitoCarta. Therefore, we cannot carry out a comprehensive analysis of the entire organelle and cannot match existing resources like MitoCarta in terms of completeness. Most published proteomics data now become available through repositories such as PRIDE, so we expect that in the future it will be possible to mine much larger datasets for mitochondrial proteins in this way. While we only show here the example of mitochondria in chromatin samples, we would expect that, in principle, any comparative proteomics experiment could be used as input dataset, as long as some components of the target structure have been detected and accurately quantified in it. It should be noted that with this method no individual experiment needs to strongly separate the target structure from the rest of the cell, but separation is achieved by integrating many small, apparently insignificant differences into one machine learning score.

3.5 Feature count influences prediction accuracy

One important parameter affecting the accuracy of mitochondrial predictions by machine learning is the "feature count," i.e. the number of different experiments in which a protein was quantified. The more feature counts (SILAC ratios) are available to establish the "covariation pattern" of a protein, the better a protein can be assigned to a certain functional group. For example, some of the 143 mitochondrial consensus proteins, on which all annotation databases agree, remain below our mitochondrial prediction cut-off. These mitochondrial proteins have been quantified in a median of 16 ± 7 SILAC experiments. By contrast, the consensus proteins that score above cut-off and are thus successfully predicted to be mitochondrial, have a median of 22 ± 9 features, and this difference is statistically significant (p-value < 0.001).

Our mitochondrial protein predictions are based on SILAC data, i.e. relative rather than absolute protein abundances. This implies that protein abundance itself should not have a direct impact on prediction accuracy, but there is arguably an indirect effect of protein abundance on performance. For example, abundant proteins will generally be observed more often, resulting in higher feature counts. SILAC quantitation itself also tends to be more accurate for abundant proteins.

3.6 Implications for the design of SILAC studies

The observation that background in SILAC experiments changes systematically has implications for the design of comparative proteomics studies. For example, studies that test the effect of a drug on chromatin proteins would typically compare chromatin fractions from treated samples with a mock control and may conclude that all measured changes relate to the drug's effect on chromatin composition. However, our observations suggest that care should be taken when drawing such conclusions. Changes among the purification background, either through direct or indirect effects of the perturbation, are in fact widespread. This is illustrated by the fact that mitochondria can be up- and downregulated significantly in chromatin samples between experiment and control (Fig. 1A–C). We made a similar observation in a screen for novel DNA replication factors, where we isolated nascent chromatin using an unrelated biochemical procedure [31]. Upon comparing nascent and mature chromatin we observed many differences that were clearly unrelated to chromatin replication. These rather reflected alterations in chromatin association of background proteins. To obtain high-confidence DNA replication factors we filtered the data using interphase chromatin probabilities [11].

4 Concluding remarks

We provide proof-of-principle data to show that background in comparative proteomics experiments is not static or random, but exhibits fluctuations that are possibly biologically meaningful and can, in fact, be exploited. Background proteins with similar functions, such as mitochondrial factors, are coordinately up- or downregulated in chromatin analyses. We demonstrate that this makes it possible to predict components of mitochondria based solely on their behavior in chromatin samples, by quantifying their presence across a diverse range of conditions and using machine learning to compare it to reference proteins of known function. In principle, we would expect our approach to work for any organelle or compartment that has been detected in quantitative proteomics data although this remains to be demonstrated. With specific significance to nuclei, a large number of nuclear bodies have been difficult to purify on their own and may well be seen as "shadows" in our chromatin data. Future work will have to show if shadow proteomics provides a path to mapping these and other elusive structures in cells.

We would like to thank Jimi-Carlo Bukowski-Wills for his support with getting started in KNIME and Lutz Fischer for his continuous bioinformatics support. This work was supported by a Wellcome Trust Senior Research Fellowship to J.R. [103139] and instrument grant [091020]. The Wellcome Trust Centre for Cell Biology is supported by core funding from the Wellcome Trust [092076]. G.K. was supported by a FEBS long-term fellowship.

The authors have declared no conflict of interest.

5 References

[1] Gatto, L., Vizcaíno, J. A., Hermjakob, H., Huber, W., Lilley, K. S., Organelle proteomics experimental designs and analysis. *Proteomics* 2010, *10*, 3957–3969.

[2] Drissi, R., Dubois, M. L., Boisvert, F. M., Proteomics methods for subcellular proteome analysis. *FEBS J.* 2013, *280*, 5626–5634.

[3] Mootha, V. K., Bunkenborg, J., Olsen, J. V., Hjerrild, M. et al., Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell* 2003, *115*, 629–640.

[4] Schirmer, E. C., Florens, L., Guan, T., Yates, J. R., 3rd, Gerace, L., Nuclear membrane proteins with potential disease links found by subtractive proteomics. *Science* 2003, *301*, 1380–1382.

[5] Foster, L. J., De Hoog, C. L., Mann, M., Unbiased quantitative proteomics of lipid rafts reveals high specificity for signaling factors. *Proc. Natl. Acad. Sci. USA* 2003, *100*, 5813–5818.

[6] Boisvert, F. M., Lam, Y. W., Lamont, D., Lamond, A. I., A quantitative proteomics analysis of subcellular proteome localization and changes induced by DNA damage. *Mol. Cell. Proteomics* 2010, *9*, 457–470.

[7] Andersen, J. S., Wilkinson, C. J., Mayor, T., Mortensen, P. et al., Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* 2003, *426*, 570–574.

[8] Foster, L. J., de Hoog, C. L., Zhang, Y., Xie, X. et al., A mammalian organelle map by protein correlation profiling. *Cell* 2006, *125*, 187–199.

[9] Dunkley, T. P., Watson, R., Griffin, J. L., Dupree, P., Lilley, K. S., Localization of organelle proteins by isotope tagging (LOPIT). *Mol. Cell. Proteomics* 2004, *3*, 1128–1134.

[10] Ohta, S., Bukowski-Wills, J. C., Sanchez-Pulido, L., Alves, F. e. L. et al., The protein composition of mitotic chromosomes determined using multiclassifier combinatorial proteomics. *Cell* 2010, *142*, 810–821.

[11] Kustatscher, G., Hégarat, N., Wills, K. L., Furlan, C. et al., Proteomics of a fuzzy organelle: interphase chromatin. *EMBO J.* 2014, *33*, 648–664.

[12] Hyman, A. A., Simons, K., Cell biology. Beyond oil and water–phase transitions in cells. *Science* 2012, *337*, 1047–1049.

[13] Kustatscher, G., Wills, K. L., Furlan, C., Rappsilber, J., Chromatin enrichment for proteomics. *Nat. Protoc.* 2014, *9*, 2090–2099.

[14] Vizcaíno, J. A., Côté, R. G., Csordas, A., Dianes, J. A. et al., The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* 2013, *41*, D1063–D1069.

[15] UniProt Consortium, UniProt: a hub for protein information. *Nucleic Acids Res.* 2015, *43*, D204–D212.

[16] Gene Ontology Consortium, Gene Ontology Consortium: going forward. *Nucleic Acids Res.* 2015, *43*, D1049–D1056.

[17] Hall, M., Frank, E., Holmes, G., Pfahringer, B. et al., The WEKA data mining software: an update. *SIGKDD Explorations* 2009, *11*, 10–18.

[18] Breiman, L., Random forests. *Machine Learn.* 2001, *45*, 5–32.

[19] Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R. et al., KNIME: the Konstanz information miner. *Data Anal. Machine Learn. Appl.* 2008, 319–326.

[20] Pagliarini, D. J., Calvo, S. E., Chang, B., Sheth, S. A. et al., A mitochondrial protein compendium elucidates complex I disease biology. *Cell* 2008, *134*, 112–123.

[21] Binns, D., Dimmer, E., Huntley, R., Barrell, D. et al., QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics* 2009, *25*, 3045–3046.

[22] Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C. et al., Proteomics. Tissue-based map of the human proteome. *Science* 2015, *347*, 1260419.

[23] Rhee, H. W., Zou, P., Udeshi, N. D., Martell, J. D. et al., Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science* 2013, *339*, 1328–1331.

[24] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 2015.

[25] Durinck, S., Spellman, P. T., Birney, E., Huber, W., Mapping identifiers for the integration of genomic datasets with the R/bioconductor package biomaRt. *Nat. Protoc.* 2009, *4*, 1184–1191.

[26] Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., ROCR: visualizing classifier performance in R. *Bioinformatics* 2005, *21*, 3940–3941.

[27] Wickham, H., *ggplot2: Elegant Graphics for Data Analysis*, Springer, New York 2009.

[28] Wang, S., Fusaro, G., Padmanabhan, J., Chellappan, S. P., Prohibitin co-localizes with Rb in the nucleus and recruits N-CoR and HDAC1 for transcriptional repression. *Oncogene* 2002, *21*, 8388–8396.

[29] Willaert, A., Malfait, F., Symoens, S., Gevaert, K. et al., Recessive osteogenesis imperfecta caused by LEPRE1 mutations: clinical documentation and identification of the splice form responsible for prolyl 3-hydroxylation. *J. Med. Genet.* 2009, *46*, 233–241.

[30] Kazak, L., Reyes, A., Duncan, A. L., Rorbach, J. et al., Alternative translation initiation augments the human mitochondrial proteome. *Nucleic Acids Res.* 2013, *41*, 2354–2369.

[31] Alabert, C., Bukowski-Wills, J. C., Lee, S. B., Kustatscher, G. et al., Nascent chromatin capture proteomics determines chromatin dynamics during DNA replication and identifies unknown fork components. *Nat. Cell Biol.* 2014, *16*, 281–293.

Manuscript 4. “The Human Proteome Co-Regulation Map Reveals Functional Relationships Between Proteins”

Pages 51 - 80

Pre-print available online, DOI: https://doi.org/10.1101/582247

The human proteome co-regulation map reveals functional relationships between proteins

Georg Kustatscher[1]*, Piotr Grabowski[2]*, Tina A. Schrader[3], Josiah B. Passmore[3], Michael Schrader[3], Juri Rappsilber[1,2#]

[1] Wellcome Trust Centre for Cell Biology, University of Edinburgh, Edinburgh EH9 3BF, UK
[2] Chair of Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, 13355 Berlin, Germany
[3] Biosciences, University of Exeter, Exeter EX4 4QD, UK

* Equal contribution
# Communicating author: juri.rappsilber@ed.ac.uk

Submission as *Resource* article.

The annotation of protein function is a longstanding challenge of cell biology that suffers from the sheer magnitude of the task. Here we present ProteomeHD, which documents the response of 10,323 human proteins to 294 biological perturbations, measured by isotope-labelling mass spectrometry. Using this data matrix and robust machine learning we create a co-regulation map of the cell that reflects functional associations between human proteins. The map identifies a functional context for many uncharacterized proteins, including microproteins that are difficult to study with traditional methods. Co-regulation also captures relationships between proteins which do not physically interact or co-localize. For example, co-regulation of the peroxisomal membrane protein PEX11β with mitochondrial respiration factors led us to discover a novel organelle interface between peroxisomes and mitochondria in mammalian cells. The co-regulation map can be explored at www.proteomeHD.net.

Functional genomics approaches often use a “guilt-by-association” strategy to determine the biological function of genes and proteins on a system-wide scale. For example, high-throughput measurement of protein-protein interactions[1–4] and subcellular localization[5–8] has delivered invaluable insights into proteome organisation. A limitation of these techniques is that extensive biochemical sample processing and non-specific antibodies may introduce artifacts. Moreover, not all proteins that function in the same biological process also interact physically or co-localize. Such functional relationships may be uncovered by assays with phenotypic readouts, including genetic interactions[9] and metabolic profiles[10], but these have yet to be applied on a genomic scale in humans. One of the oldest functional genomics methods is gene expression profiling[11]. Genes with correlated activity often participate in similar cellular functions, which can be exploited to infer the function of uncharacterized genes based on their coexpression with known genes[12–16].

However, predicting gene function from coexpression alone often leads to inaccurate results[17,18]. One possible reason for this is that gene activity is generally measured at the mRNA level, neglecting the contribution of protein synthesis and degradation to gene

expression control. The precise extent to which protein levels depend on mRNA abundances is still debated, and likely differs between genes and test systems[19–21]. However, some fundamental differences between mRNA and protein expression control have recently emerged. For example, many genes have coexpressed mRNAs due to their chromosomal proximity rather than any functional similarity[17,22–24]. Such non-functional mRNA coexpression results from stochastic transitions between active and inactive chromatin that affect wide genomic loci[22,23,25], and transcriptional interference between closeby genes[23,26]. Importantly, coexpression of spatially close, but functionally unrelated genes is buffered at the protein level[17,23]. Protein abundances are also less affected than mRNA levels by genetic variation[27,28], including variations in gene copy numbers[29–31]. Consequently, protein expression profiling outperforms mRNA expression profiling with regard to gene function prediction[17,18]. Protein-based profiling not only allows for a more accurate measurement of gene activity, but can determine additional aspects of a cell's response to a perturbation, such as changes in protein localization and modification state. At the proteome level, expression profiling can therefore be extended to a more comprehensive protein covariation analysis.

Proof-of-principle studies by us and others have shown that protein covariation can be used to infer, for example, the composition of protein complexes and organelles[32–40]. However, these studies have focussed on relatively small sets of proteins or biological conditions, or used samples tailored to the analysis of specific cellular structures. In addition to the limited amount of data, coexpression analyses may be held back by the statistical tools used to pinpoint genes with similar activity. Coexpressed genes are commonly identified using Pearson's correlation, which is restricted to linear correlations and susceptible to outliers. Machine-learning may offer an increase in sensitivity and specificity.

Despite the success of functional genomics, many human proteins remain uncharacterized, especially small proteins that are difficult to study by biochemical methods. The emergence of big proteomics data and new computational approaches could provide an opportunity to look at these proteins from a different angle. We wondered if protein covariation would assign functions to previously uncharacterized proteins or novel roles to characterized ones. The resulting resource is available at www.proteomeHD.net to generate hypotheses on the cellular functions of proteins of interest in a straightforward manner.

RESULTS

ProteomeHD is a data matrix for functional proteomics

To turn protein covariation analysis into a system-wide, generally applicable method, we created ProteomeHD. In contrast to previous drafts of the human proteome[7,8,20,41,42], ProteomeHD does not catalogue the proteome of specific tissues or subcellular compartments. Instead, ProteomeHD catalogues the transitions between different proteome states, i.e. changes in protein abundance or localization resulting from cellular perturbations. HD, or high-definition, refers to two aspects of the dataset. First, all experiments are quantified using SILAC (stable isotope labelling by amino acids in cell culture)[43]. SILAC essentially eliminates sample processing artifacts and is especially accurate when quantifying small fold-changes. This is crucial to detect subtle, system-wide effects of a

perturbation on the protein network. Second, HD refers to the number of observations (pixels) available for each protein. As more perturbations are analysed, regulatory patterns become more refined and can be compared more accurately.

To assemble ProteomeHD we processed the raw data from 5,288 individual mass-spectrometry runs into one coherent data matrix, which covers 10,323 proteins (from 9,987 genes) and 294 biological conditions (Supplementary Table 1). About 20% of the experiments were performed in our laboratory and the remaining data were collected from the Proteomics Identifications (PRIDE)[44] repository (Fig. 1a). The data cover a wide array of quantitative proteomics experiments, such as perturbations with drugs and growth factors, genetic perturbations, cell differentiation studies and comparisons of cancer cell lines (Supplementary Table 2). All experiments are comparative studies using SILAC[43], i.e. they do not report absolute protein concentrations but highly accurate fold-changes in response to perturbation (Fig. 1b). About 60% of the included experiments analysed whole-cell samples. The remaining measurements were performed on samples that had been fractionated after perturbation, e.g. to enrich for chromatin-based or secreted proteins. This allows for the detection of low-abundance proteins that may not be detected in whole-cell lysates.

Machine-learning captures functional protein associations

Proteins that are functioning together have similar patterns of up- and downregulation across the many conditions and samples in ProteomeHD (Fig. 1c). Proteins with unrelated function can be clearly distinguished, even though most expression changes are well below 2-fold. Therefore, we reasoned that it should be possible to reveal functional links between proteins on the basis of such regulatory patterns, and reveal the function of unknown proteins by associating them with well-characterized ones. To increase the accuracy of pattern recognition we focussed on the 5,013 proteins that were quantified in at least 95 of the 294 perturbation experiments. We used the treeClust[45] machine-learning algorithm to calculate how similar any two proteins behave across ProteomeHD, resulting in a “co-regulation score”. We define proteins with a co-regulation score above 0.5 as “co-regulated”. In this way, we find 56,587 protein co-regulation pairs, or 0.5% of the 1.2×10^7 possible pairs (Fig. 1d).

We then tested whether co-regulation indicates co-function. Indeed, we find that co-regulated protein pairs are heavily enriched for subunits of the same protein complex, enzymes catalysing consecutive metabolic reactions and proteins occupying the same subcellular compartments (Fig. 1e). Most proteins are co-regulated with at least one other protein, and nearly half have more than ten co-regulation partners (Fig. 1f). For 97% of the tested proteins, the group of their co-regulation partners is enriched in at least one Gene Ontology[46] biological process (Fig. 1h).

The extent of coexpression between two genes is often determined using Pearson’s correlation coefficient (PCC). We calculated correlation coefficients for protein pairs across ProteomeHD. Surprisingly, they do not agree well with the co-regulation scores obtained by treeClust machine-learning (Supplementary Fig. 1a). To assess which metric identifies known protein associations more accurately we performed a precision-recall analysis. At a recall of 1,000 functionally related protein pairs, the precision is 0.99, 0.67 and 0.10 for treeClust learning, PCC and a random classifier, respectively (Supplementary Fig. 1b). This

indicates that treeClust strongly outperforms PCC at predicting functional relationships from the same dataset. To understand the reason for this we inspected a number of protein pairs in detail. We find that treeClust's robustness towards data outliers allows it to detect more true-positive and fewer false-positive protein associations (Supplementary Fig. 1c,d).

A co-regulation map of the human proteome

As a result of treeClust learning we know for each protein how strongly - or weakly - it is co-regulated with any other protein. To visualize this complex dataset in a human-readable form we applied t-Distributed Stochastic Neighbor Embedding (t-SNE)[47,48]. This produces a two-dimensional proteome co-regulation map in which the distance between proteins indicates how similar they responded to the various perturbations in ProteomeHD (Fig. 1g). The map shows that protein co-regulation is closely related to co-function. From a global perspective, the map reflects the subcellular organization of the cell (Fig. 1i). For example, it separates the nucleolus from the nucleus, sets apart most secreted proteins and broadly distinguishes between small and large subunits of the ribosome. A closer look into three sections of the map reveals that it captures more detailed functional relationships, too. For example, the five protein complexes of the respiratory chain are almost resolved (Fig. 1i, section i). The section also contains the phosphate and ADP carriers that transport the substrates for ATP synthesis through the inner mitochondrial membrane. Similarly, a section containing various RNA-related processes shows most subunits of the exosome complex grouped together, next to other nucleolar rRNA processing factors and the nuclear pore complex (Fig. 1i, section ii). In a third example section, cytoskeleton proteins such as actins and myosins are next to their regulators, including Rho GTPases and the Arp2/3 complex (Fig. 1i, section iii). Notably, these annotations are only used to illustrate that the co-regulation map reflects functional similarity; the map itself is generated without any curated information, solely on the basis of protein abundance changes in ProteomeHD. Therefore, the co-regulation map provides a data-driven overview of the proteome, connecting proteins into functionally related groups.

Co-regulation complements existing functional genomics methods

We next asked if protein co-regulation can predict associations that are not detected by other methods. We first compared co-regulation to physical protein-protein interactions determined by various methods, as catalogued in BioGrid[49] (Fig. 2a). In total, only about 10% of co-regulated protein pairs showed evidence of physical interaction. These were mainly derived from co-fractionation experiments, which tend to capture indirect interactions, rather than methods that detect direct interactions, such as two-hybrid screens. In addition, we assessed how co-regulation compares to functional associations mapped by STRING[50], based on methods such as text mining, evolutionary co-occurrence or mRNA coexpression. We find that the combined STRING evidence captures about 26% of all co-regulation pairs, showing that co-regulation analysis confirms existing links, but also provides many additional ones.

We then compared the co-regulation approach to an individual functional genomics experiment. BioPlex 2.0 is the most comprehensive affinity purification–mass spectrometry (AP-MS) study reported to date[4]. BioPlex reports 11,229 physical interactions between the

proteins used in our study, of which 12% are also co-regulated (Fig. 2b). An additional 54,064 potential links between these proteins are identified uniquely by co-regulation. These are strongly enriched for functional protein associations found in STRING, compared to a random set of protein pairs (Fig. 2b). Importantly, co-regulation links complement physical interactions not only in numbers, but also qualitatively. For example, FAM45A is a protein of unknown function that BioPlex reports to interact with two protein complexes involved endosomal cargo sorting, CCC and retriever[51]. FAM45A is also co-regulated with several CCC subunits, various other endosomal trafficking proteins and three regulatory factors of NF-κB signaling, suggesting that FAM45A may be an additional link between this key signaling pathway and endocytic trafficking[52] (Fig. 2c).

Co-regulation provides functional annotation for uncharacterized proteins

The co-regulation map contains 339 uncharacterized proteins, which we define as proteins with a Uniprot[53] annotation score of 3 or less (Fig. 2d). Of these, 80% are co-regulated with at least one fully characterized protein, i.e. a protein with an annotation score of 4 or 5 (Fig. 2e). A median of 9 well-studied proteins are co-regulated with each uncharacterized protein, making it possible to predict the potential function of hundreds of uncharacterized proteins on the basis of their co-regulation partners. We observe a similar connectivity for the cancer gene census[54], i.e. genes that cause cancer when mutated, and for DisGeNET[55] genes, which are genes implicated in a broad range of human diseases (Fig. 2e). Therefore, protein co-regulation may be helpful for functional analysis of human disease genes. To facilitate such functional annotation efforts we created the website www.ProteomeHD.net, which allows proteins to be queried regarding their position in the co-regulation map and their co-regulation partners.

A common property of uncharacterized proteins is their small size. For example, proteins smaller than 15 kDa constitute 16% of the uncharacterized proteins in the human proteome, but only 5% of the characterized ones. Among the least well understood fraction of the proteome, i.e. proteins with an annotation score of 1, 38% are smaller than 15 kDa (Fig. 2f). This discrepancy is set to increase further, since hundreds or thousands such microproteins have so far been overlooked by genome annotation efforts[56,57]. Microproteins can regulate fundamental biological processes[58], but their small size makes it difficult to identify interaction partners[56,59] or to target them in mutagenesis screens[56]. Microprotein sequences also tend to be less conserved than those of longer protein-coding genes[60]. We reasoned that protein covariation may help to reduce the annotation gap for small proteins, because simply quantifying proteins in cell extracts should introduce less bias against small proteins than methods which require extensive genetic or biochemical sample processing. Indeed, we find that 12% of the uncharacterized proteins in the co-regulation map are smaller than 15 kDa. While this is less than the 16% in the proteome overall, the bias is considerably smaller than that of physical protein-protein-interaction maps. For example, BioPlex contains 6% uncharacterized microproteins (Fig. 2g).

Protein co-regulation can predict the potential function of uncharacterized microproteins that are absent from BioPlex 2.0, and in some cases these predictions are supported by additional evidence from small-scale studies. For example, the mitochondrial proteolipid MP68 is co-regulated with subunits of the ATP synthase complex, suggesting a

function in ATP production (Fig. 2h). Intriguingly, MP68 co-purifies with the ATP synthase complex, but only in buffers containing specific phospholipids[61,62], and knockdown of MP68 decreases ATP synthesis in HeLa cells[63]. In addition, several membrane proteins of the endoplasmic reticulum are co-regulated with MP68, suggesting an additional function in that membrane. Second, the conserved C11orf98 microprotein is located in the nucleolar area of the co-regulation map (Fig. 2i). C11orf98 was also identified as a nucleolar protein by *in situ* proximity tagging, another approach proposed to reduce non-specific interactions obtained by affinity-purification of microproteins[59].

A new function for PEX11β in peroxisome-mitochondria interplay

Some well-characterized proteins have unexpected co-regulation partners. For example, PEX11β is a key regulator of peroxisomal membrane dynamics and division[64]. However, only one of PEX11β's co-regulation partners is a peroxisomal protein. Instead, it is most strongly co-regulated with subunits of the mitochondrial ATP synthase and other components of the electron transport chain (Fig. 2j). These proteins are located to the inner mitochondrial membrane, making a physical interaction with PEX11β unlikely. However, peroxisomes and mitochondria in mammals are intimately linked cooperating in fatty acid β-oxidation and ROS homeostasis[65]. How these organelles communicate or mediate metabolite flux has been elusive. Live cell imaging revealed that expression of PEX11β-EGFP in mammalian cells induced the formation of peroxisomal membrane protrusions, which interact with mitochondria (Fig. 3, Supplementary movies 1-3). Interactions of elongated peroxisomes with mitochondria were more frequent than those of spherical organelles, and long-lasting excluding random events (Fig. 3n,o). Miro1 (RHOT1), a membrane adaptor for the microtubule-dependent motors kinesin and dynein[66], is also co-regulated with PEX11β (Fig. 2j). We recently showed that Miro1 distributes to mitochondria and peroxisomes[67] indicating that it coordinates mitochondrial and peroxisomal dynamics with local energy turnover. Peroxisome-targeted Miro1 (Myc-Miro-PO) can be used as a tool to exert pulling forces at peroxisomal membranes, which results in the formation of membrane protrusions in certain cell types (Supplementary Fig. 2) (I Castro, DM Richards, J Metz, JL Costello, JB Passmore, TAS, A Gouveia, D Ribeiro, MS, submitted). We show here that silencing of PEX11β inhibits membrane elongation by Myc-Miro-PO, confirming that PEX11β is required for the formation of peroxisomal membrane protrusions (Supplementary Fig. 2). These findings are in agreement with studies in plants, where *At*PEX11a has been reported to mediate the formation of peroxisomal membrane extensions in response to ROS[68]. In yeast, peroxisome-mitochondria contact sites are established by *Sc*Pex11 and *Sc*Mdm34, a component of the ERMES complex[69]. We conclude that PEX11β and Miro1 contribute to peroxisome membrane protrusions, which present a new mechanism of interaction between peroxisomes and mitochondria in mammals. They likely function in the metabolic cooperation and crosstalk between both organelles, and may facilitate transfer of metabolites such as acetyl-CoA and/or ROS homeostasis during mitochondrial ATP production. These findings now enable future studies on the precise functions of peroxisome membrane protrusions in mammalian cells and the role of PEX11β.

Proteomics enables higher accuracy but lower coverage than transcriptomics

To compare the impact of mRNA and protein abundances on expression profiling we first focussed on 59 SILAC ratios in ProteomeHD that measured abundance changes across a panel of lymphoblastoid cell lines[28]. For these samples, corresponding mRNA abundance changes have been determined using RNA-sequencing[70]. Repeating treeClust learning on the basis of these data, we observed that protein coexpression predicts functional associations with far higher precision than mRNA coexpression (Fig. 4a). Similar results have recently been reported for a panel of human cancer samples[17].

Such analyses show that in a direct gene-by-gene, sample-by-sample comparison, protein expression levels are better indicators for gene function than mRNA expression. However, the amount of transcriptomics data published to date vastly exceeds that of proteomics studies. For example, the NCBI GEO repository currently holds mRNA expression profiling data from more than one million human samples[71]. This raises the possibility that the sheer quantity of available transcriptomics data could overcome their reduced reflection of functional links and, in combined form, perform better than protein-based measurements. To test this we compared the ProteomeHD co-regulation score with Pearson correlation coefficients obtained by STRING, which leverages the vast amount of mRNA expression experiments deposited in GEO[50,72]. Remarkably, precision-recall analysis shows that the protein co-regulation score still outperforms mRNA coexpression, despite being based on only 294 SILAC ratios (Fig. 4b). Much of this improvement is due to the robustness of treeClust machine-learning, as Pearson's correlation coefficients derived from the same ProteomeHD data work only partially better than mRNA correlation (Fig. 4b). While only gene pairs with both mRNA and protein expression measurements were considered for the precision-recall analysis, the transcriptomics and proteomics datasets individually covered 17,436 and 4,976 genes, respectively (Fig. 4b). Therefore, mRNA profiling outperforms protein profiling in terms of gene coverage.

DISCUSSION

ProteomeHD in conjunction with machine learning provides an entry point for "big-data"-type protein co-regulation analysis into the functional genomics methods repertoire. It is possible that accuracy and coverage could be increased further by adding additional proteomics data. To test this we randomly removed 5%, 10% or 15% of the data points in ProteomeHD. This decreases performance reproducibly and proportionally to the amount of removed data (Supplementary Fig. 3), suggesting that ProteomeHD has not reached saturation and expanding it will further enhance its performance. One possibility would be to incorporate other types of proteomics experiments, such as affinity-purifications or indeed the entire PRIDE[44] repository. The latter approach is for instance taken by the Tabloid Proteome, which infers protein associations based on detecting them in the same subset of many different proteomics experiments[39]. However, there is a benefit of restricting ProteomeHD on perturbation experiments. It supports a biological interpretation of protein associations derived from it: two co-regulated proteins are part of the same cellular response to changing biological conditions, even though the precise molecular nature of the connection remains

unknown. In this way, protein co-regulation analysis is analogous to genetic interaction screening. This also sets protein co-regulation apart from indiscriminate protein covariation or co-occurrence analyses, which find protein links in a mix of proteomics data and therefore give no insight into the possible biological connection.

A key difference between our approach and previous gene coexpression studies is our application of two machine-learning algorithms, treeClust[45] and tSNE[47,48]. Inferring protein associations through treeClust learning is both more robust and sensitive than a traditional correlation-based approach, providing a leap in the accuracy with which functionally relevant interactions can be identified from the same dataset. For example, a recent study reported a protein co-regulation network across 41 cancer cell lines and subsequently identified dysregulated protein associations that predict drug sensitivities of these cell lines[18]. High-quality proteomics data allowed Lapek *et al*[18] to detect protein-protein associations with an accuracy that was tenfold higher than that based on matching mRNA coexpression data. When applying treeClust, strikingly, we can further improve this performance (Supplementary Fig. 4a). This suggests that treeClust may be helpful for the detection of "dysregulation biomarkers" in the future. The second machine-learning tool we apply here, tSNE, visualizes treeClust-learned protein associations as a 2D map. Correlation networks are typically built from a limited number of the strongest pairwise interactions, whereas tSNE takes into account the similarity - or dissimilarity - between all possible pairwise protein combinations. It creates the map that best reflects both direct and indirect relationships between all proteins. In this way, also proteins that are not directly linked to the core network can be placed into a functional context. For example, a tSNE co-regulation map obtained for Lapek *et al*'s cancer proteomics dataset contains the complete set of ~6,800 proteins, rather than the 3,024 proteins that are directly correlated with another protein (Supplementary Fig. 4b). Moreover, protein-protein associations visualized by tSNE can be explored in a hierarchical manner, with larger distances indicating weaker co-regulation. This may be useful for studying connections between related protein complexes (Fig. 1i) or to reveal broad functional clues for uncharacterized proteins for which no detailed predictions are available, such as the C11orf98 protein assigned to the nucleolar area of the co-regulation map (Fig. 2j). Our web application at www.proteomeHD.net is designed to support researchers in exploring co-regulation data at multiple scales, to validate existing hypotheses or create new ones.

Protein coexpression analysis identifies functional connections between proteins with an accuracy and sensitivity that is substantially higher than traditional mRNA coexpression analysis. This may be particularly important for constitutively active genes, which constitute about half of human genes[42] and are primarily controlled at the protein level[73,74]. With an ever increasing amount of protein expression data making their way into the public domain, and the simplicity of exploiting the analysis results by the scientific community, protein coexpression analysis has a large potential in gene function annotation. Only 300 quantitative proteomics measurements sufficed in conjunction with machine learning to establish functional connections between many human genes, which may be of considerable interest for proteome annotation in less studied or difficult to study organisms.

ACKNOWLEDGEMENTS

We are grateful to Damian Szklarczyk for providing the mRNA Pearson correlation data used by String. We also thank Karen Wills, Kyosuke Nakamura, Constance Alabert and Anja Groth for contributing chromatin enrichment experiments, and Afsoon S. Azadi for support with live-cell-imaging. This work was supported by the Wellcome Trust through a Senior Research Fellowship to JR (grant number 103139) and by the Biotechnology and Biological Sciences Research Council (BB/K006231/1, BB/N01541X/1) to MS. The Wellcome Trust Centre for Cell Biology is supported by core funding from the Wellcome Trust (grant number 203149).

AUTHOR CONTRIBUTIONS

G. K. and J. R. conceived the project. G. K. conducted the data analysis. P. G. created the web application. T. A. S., J. B. P. and M. S. conducted the Pex11β analysis. All authors contributed to writing the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

FIGURE 1

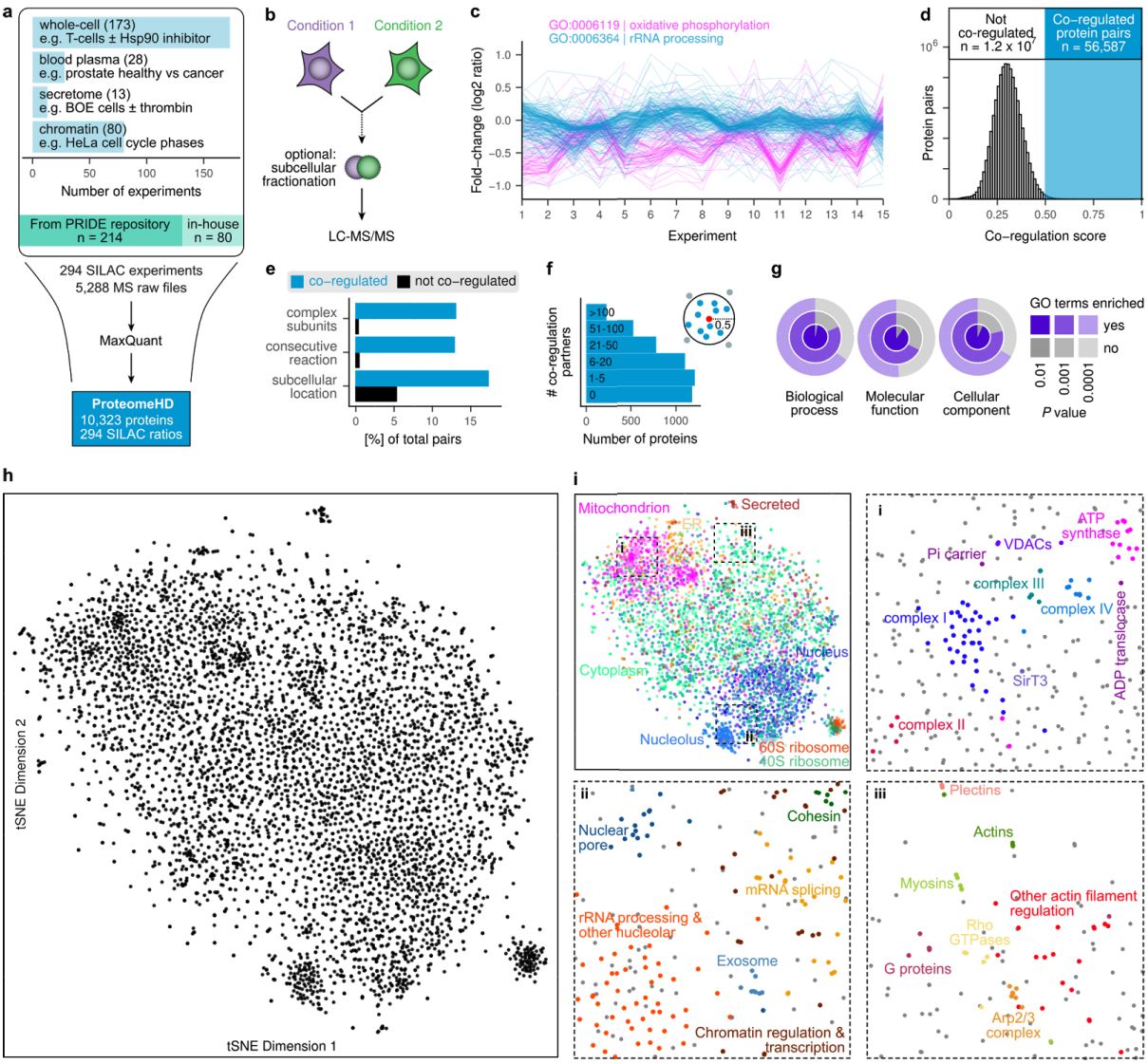


Figure 1. The co-regulation map shows functional associations between human proteins.

(**a**) Assembly of ProteomeHD, which quantifies the protein response to 294 perturbations using SILAC[43]. Most measurements document protein abundance changes in whole-cell samples, but in some cases subcellular fractions were enriched to detect low-abundance proteins. Data were collected from PRIDE[44] and produced in-house. (**b**) All data are comparative, i.e. SILAC-labelled samples are quantified relative to each other. (**c**) Example experiments showing that groups of proteins with related functions, e.g. Gene Ontology[46] (GO) biological processes, display similar expression changes. Note that the fold-changes are often very small. (**d**) Unsupervised machine learning using the treeClust[45] algorithm produces a co-regulation score, indicating the extent of covariation between two proteins. A small fraction of protein pairs exceeds the 0.5 cut-off and is defined to be “co-regulated”. (**e**) Co-regulated protein pairs are strongly enriched for subunits of the same protein complex,

363 enzymes catalysing consecutive metabolic reactions and proteins with identical subcellular
364 localization. (**f**) Most proteins are co-regulated with 1 to 5 other proteins, but many have
365 more co-regulated partners. (**g**) Considering proteins that are co-regulated with ≥10 proteins,
366 these groups of co-regulated proteins are almost always enriched in one or more GO terms.
367 (**h**) The global co-regulation map of ProteomeHD created using t-Distributed Stochastic
368 Neighbor Embedding (t-SNE)[47,48]. Distances between proteins indicate how similar their
369 expression patterns are. See www.proteomeHD.net for an interactive version of the map. (**i**)
370 The co-regulation map broadly corresponds to subcellular compartments, and more detailed
371 functional associations can be observed at higher resolution, as exemplified in i-iii.

FIGURE 2

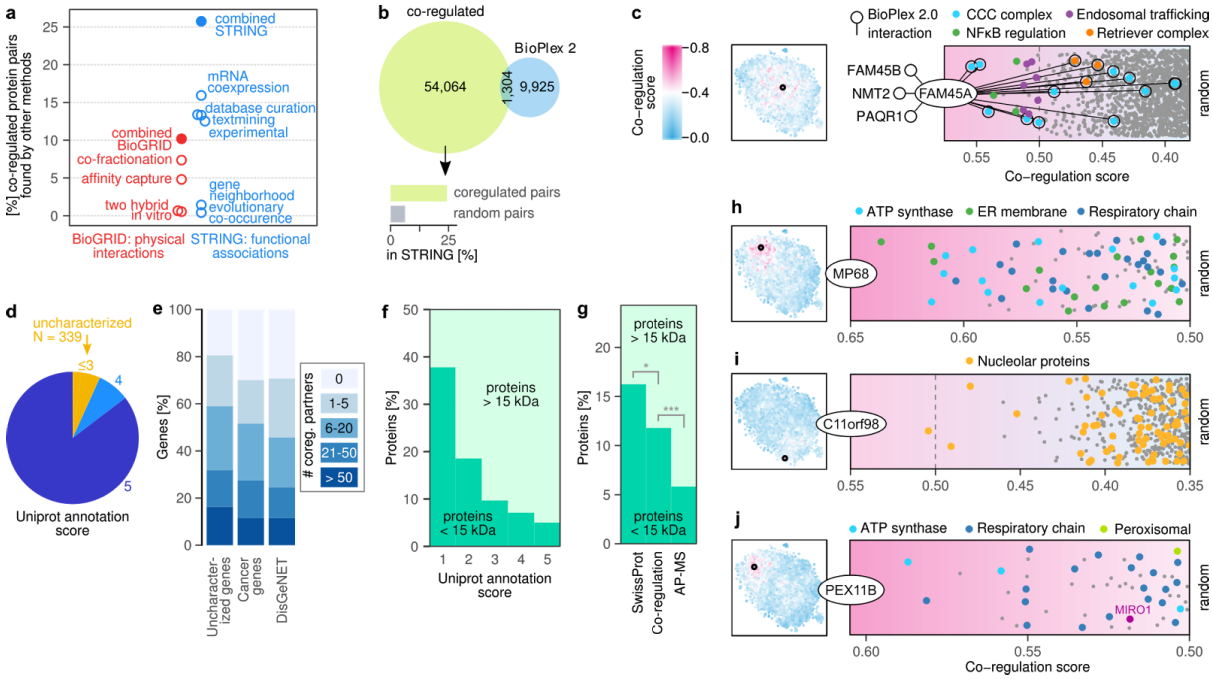


Figure 2. Protein co-regulation complements existing methods and predicts functions of unknown proteins.

(**a**) Percentage of co-regulated protein pairs that were previously linked physically (BioGrid[49]) or functionally (STRING[50]) by a range of functional genomics methods. BioGrid and STRING integrate data from many small and large-scale studies. (**b**) Number of co-regulation links compared to links found for the same set of genes by BioPlex 2.0[4], the largest protein-protein-interaction (PPI) dataset reported to date by a single study. Associations unique to co-regulation are also enriched for links in STRING, compared to random protein pairs. (**c**) Co-regulation and BioPlex complement each other to predict an endosomal trafficking function for uncharacterized protein FAM45A, possibly related to NFκB signaling. Inset shows the position of FAM45A (black circle) in the co-regulation map, other proteins are color-coded by their co-regulation score with FAM45A. (**d**) Proteins in the co-regulation map are defined as uncharacterized if their Uniprot annotation score is ≤3. (**e**) Connectivity of uncharacterized and disease genes to well-characterized genes (annotation score ≥4). 80% of uncharacterized proteins have at least one co-regulation partner, 60% have more than five. (**f**) Microproteins are heavily enriched among the uncharacterized human proteins in SwissProt. (**g**) The co-regulation map contains fewer microproteins (12%) than SwissProt overall (16%), but this bias is smaller than that of a state-of-the-art affinity-purification mass-spectrometry (AP-MS) experiment, represented by BioPlex (6%). *P*-values are from one-sided Fisher's Exact test (* $P < 0.05$, *** $P < 0.001$). (**h, i**) Protein co-regulation reveals potential functions of two uncharacterized microproteins that are absent from BioPlex. (**j**) Unexpected behavior of peroxisomal PEX11β, which is co-regulated with mitochondrial respiration factors.

FIGURE 3

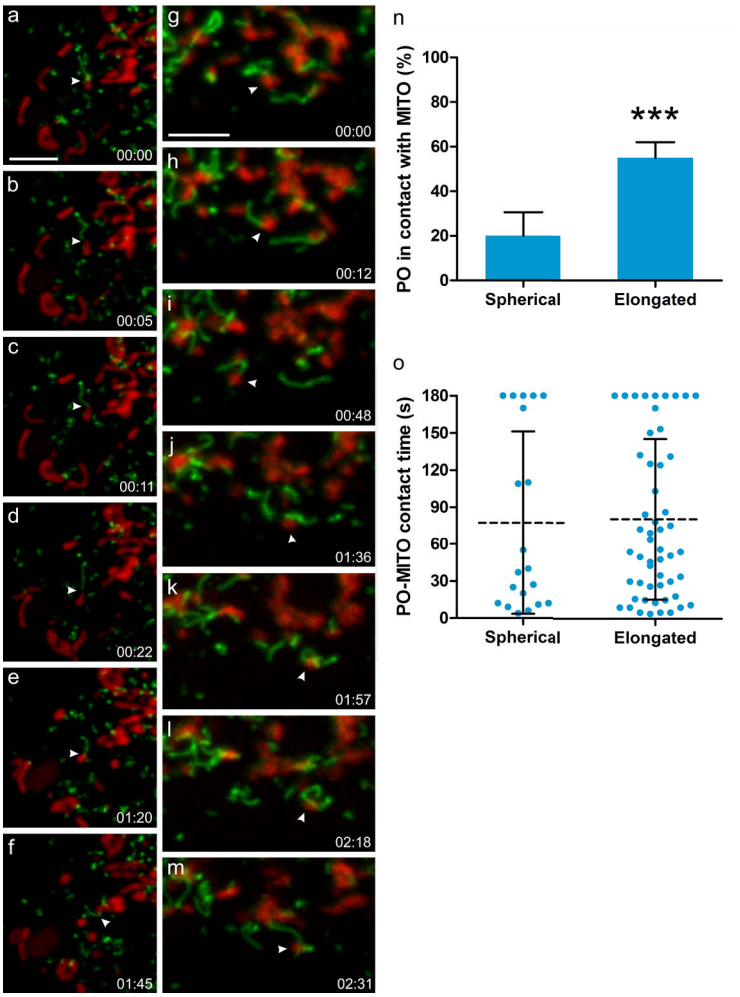


Figure 3. PEX11β mediates the formation of peroxisomal membrane protrusions which interact with mitochondria in mammalian cells.

(**a-m**) COS-7 cells were transfected with PEX11β-EGFP, mitochondria were stained with Mitotracker (red) and cells observed live using a spinning disc microscope. PEX11β, a membrane shaping protein, induces the formation of tubular membrane protrusions from globular peroxisomes. We show here that those membrane protrusions can interact with mitochondria. (**a-f**) shows a peroxisome which interacts with a mitochondrion via its membrane protrusion (arrowhead), and follows it, occasionally detaching and re-establishing contact before interacting with another mitochondrion (*see* Supplementary Movie 1). (**g-m**) shows a mitochondrion (arrowhead) which interacts with a peroxisome via a peroxisomal membrane protrusion. It then detaches and moves away to interact with another peroxisome, which wraps its protrusion around it, before interacting with another mitochondrion (*see* Supplementary Movie 2). (**n**) Quantification of interactions between spherical or elongated peroxisomes (PO) with mitochondria (MITO). The average result of 3 independent experiments is shown, error bars indicate standard deviation. (**o**) Quantification of contact time. Note that elongated PO interact more frequently with MITO than spherical PO, but for similar time periods. PO-MITO interactions are generally long-lasting and not random (*see* Supplementary Movie 3) (n=200 peroxisomes from 5 different cells). Dotted line indicates the mean, error bars indicate standard deviation. *** $P < 0.001$ from a two-tailed unpaired *t* test; Time (min:sec). Scale bars, 5 μm.

FIGURE 4

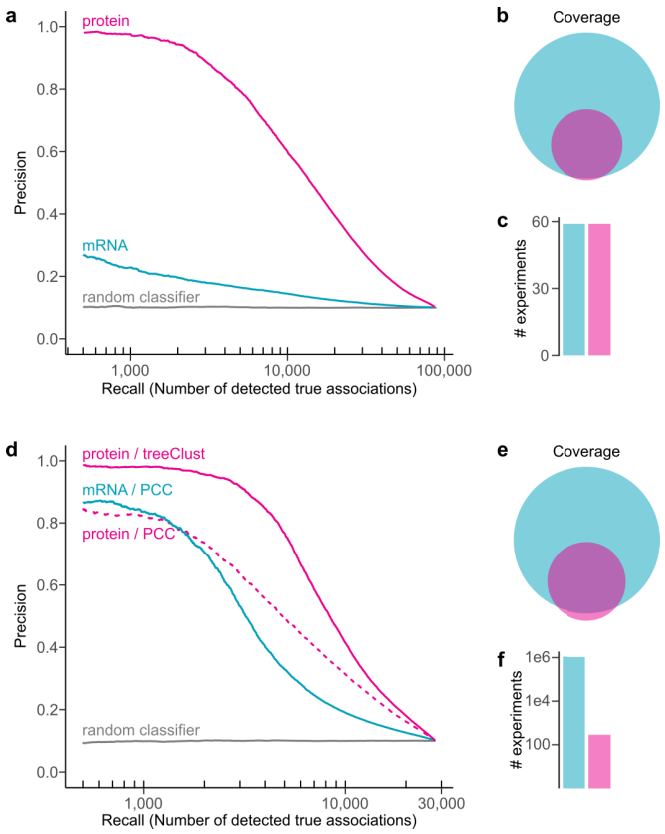
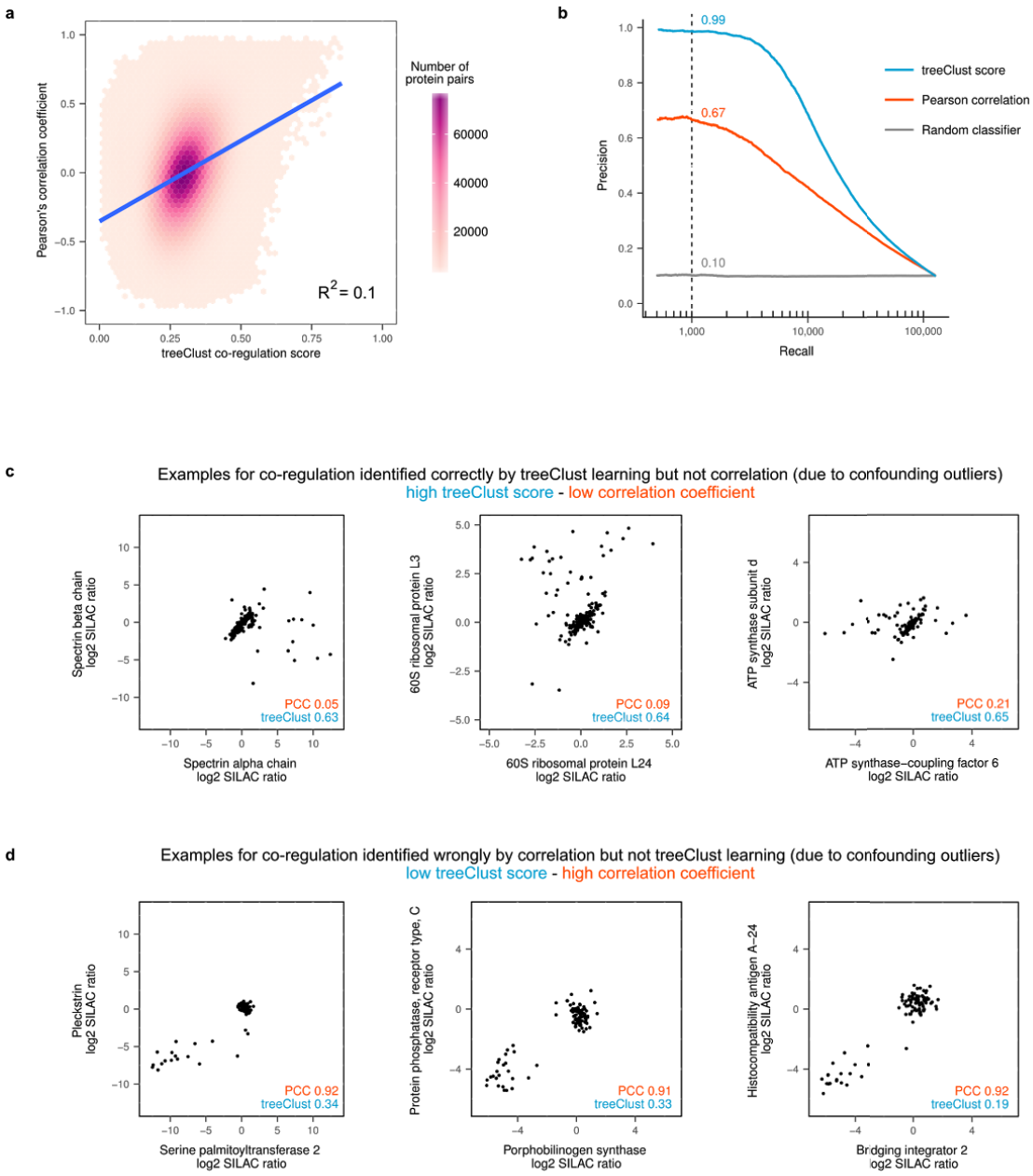
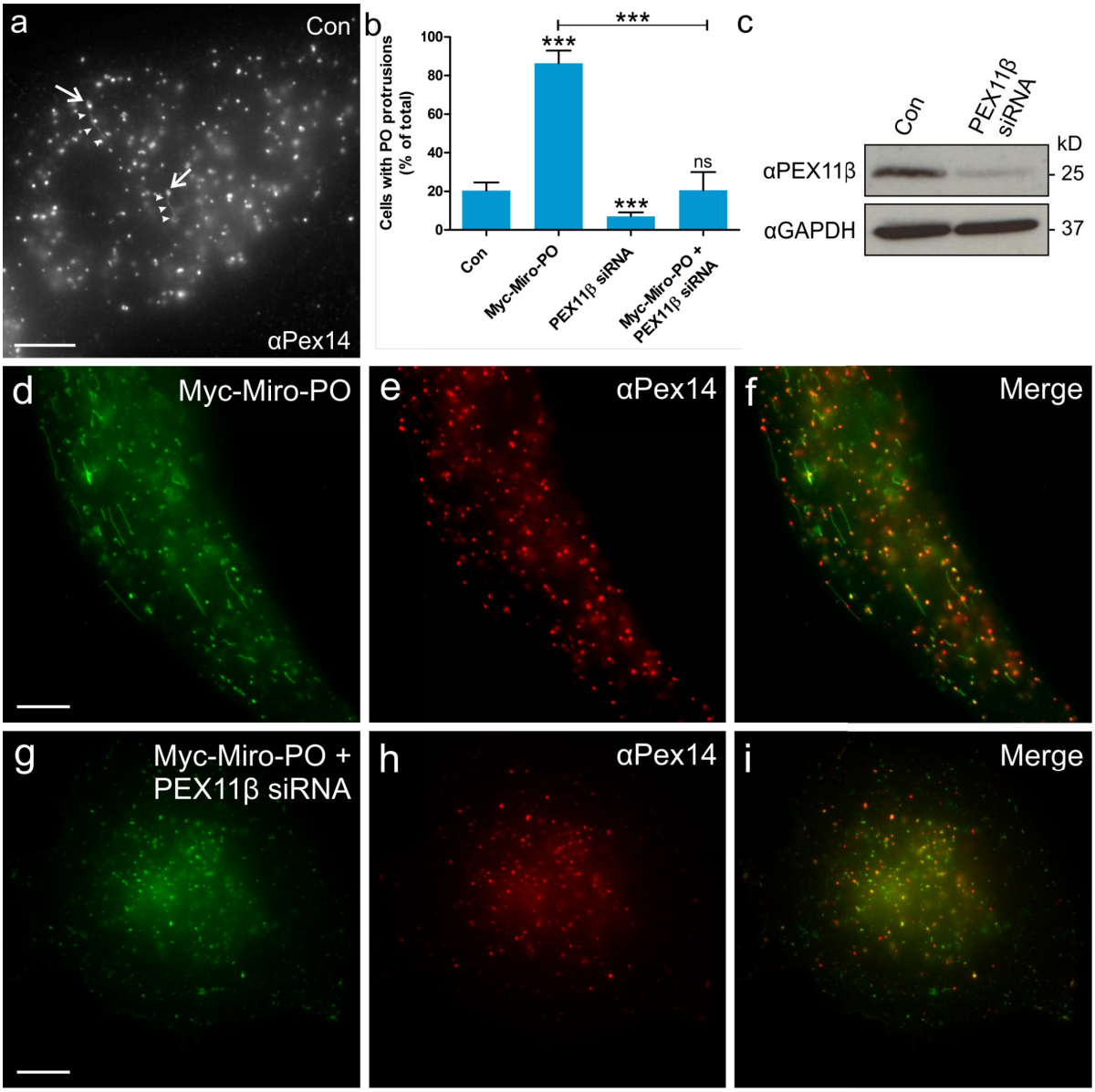


Figure 4. Protein co-regulation enables higher precision from less data, but lower coverage than classic mRNA coexpression. (**a**) Precision-recall analysis of treeClust machine-learning on a subset of ProteomeHD, that is 59 samples for which matching RNA-seq data were available from a separate study[70]. Reactome pathways were used as gold standard for true functional associations (proteins found in same pathway) and false associations (never found in same pathway). (**b**) Venn diagram showing number of genes covered by each analysis (only genes covered by both were considered for precision-recall curves). (**c**) Barchart showing number of experiments the curves are based on. (**d**) Similar precision-recall analysis of treeClust machine-learning on the full ProteomeHD database ("protein / treeClust"), in comparison to Pearson correlation obtained by STRING[50] on the basis of one million human mRNA profiling samples deposited in the NCBI Gene Expression Omnibus[71] ("mRNA / PCC"). Protein co-regulation outperforms mRNA correlation despite being based on orders-of-magnitude less data. This is partially due to the use of machine-learning, as predicting associations from ProteomeHD using PCC decreases performance markably ("protein / PCC"). (**e, f**) same as (b, c).



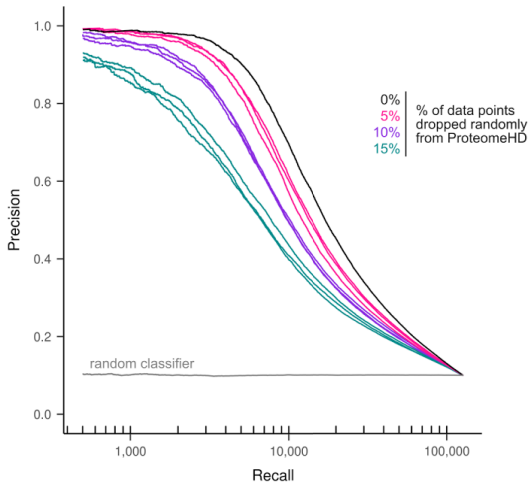
Supplementary Figure S1. treeClust machine learning is more robust than Pearson's correlation coefficient (PCC).

(**a**) Comparison of treeClust score and PCC for the same protein pairs. The two measures for expression similarity are correlated, but only weakly (R squared 0.1). Hexagons are bins of protein pairs falling in each range of the plot. A linear regression is shown in blue. (**b**) Precision - recall plot showing that treeClust learning is more precise than PCC. Pairs of proteins mapping to the same Reactome pathway were used as bona fide "functionally related proteins" (true positives). Protein pairs that do not map to any common Reactome pathway were used as bona fide "functionally unrelated proteins" (false positives). At a recall of 1,000 functionally related protein pairs, the precision is 0.99, 0.67 and 0.10 for treeClust learning, PCC and a random prediction, respectively. (**c**) Three examples for coordinated expression changes of functionally related protein pairs, which are identified by treeClust but not PCC. Note that treeClust is not confounded by the outlier measurements. (**d**) Three examples of proteins with unrelated expression changes and functions, which are mistakenly identified as co-regulated by PCC, but not treeClust. Most measurements for these proteins do not fall on a diagonal line, but a few broadly correlated outlier measurements drive the high PCC.



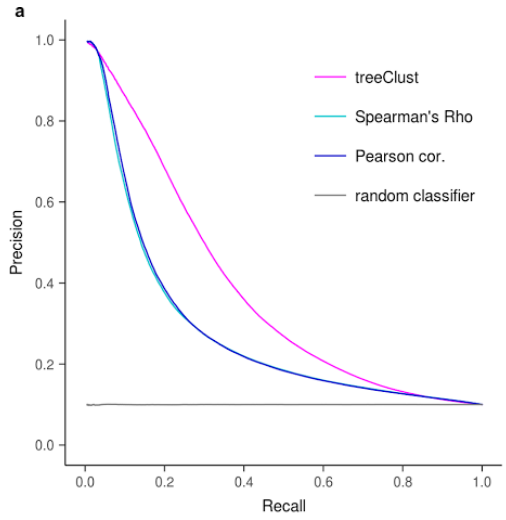
Supplementary Figure 2. MIRO1-induced peroxisomal membrane protrusions depend on PEX11β.

(**a-i**) PEX5-deficient human skin fibroblasts were mock-treated (control), or transfected with Myc-Miro-PO, a peroxisome-targeted Miro1 variant, in the presence of control- or PEX11β-specific siRNA. Cells were processed for immunofluorescence using anti-Myc and anti-PEX14 antibodies (peroxisomal marker). (**b**) Quantification of cells with peroxisomal protrusions. The average result of 3 independent experiments is shown, error bars indicate standard deviation. (**c**) Immunoblots of cell lysates showing efficient silencing of Pex11β. Loading control: GAPDH. (**a, b**) Control cells occasionally contain peroxisomes with membrane protrusions (< 5 per cell; up to 5 µm in length). (**d-f, b**) Myc-Miro-PO induces the formation of peroxisomal membrane protrusions (> 5 per cell; > 5 µm in length). (**g-i, b**) Silencing of PEX11β by siRNA significantly reduces the number of cells with peroxisomal membrane protrusions in controls and Myc-Miro-PO expressing cells. Globular peroxisomes (arrows) with membrane protrusions (arrowheads) in (a) are highlighted. *** $P < 0.001$; ** $P < 0.01$ from a two-tailed unpaired t test; ns, not significant. Scale bars, 10 µm.

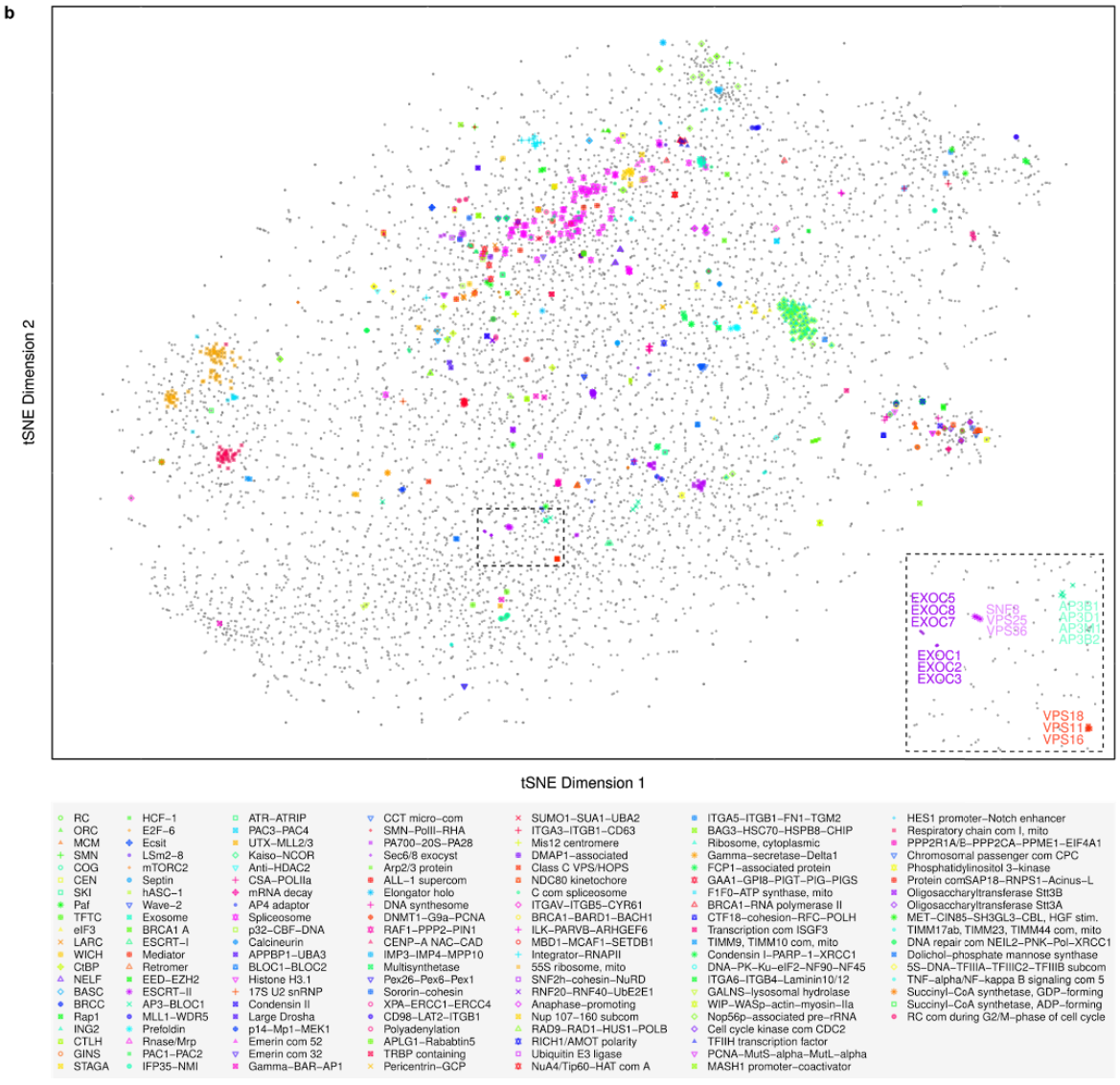


Supplementary Figure 3. Information content of ProteomeHD has not reached saturation yet

We randomly removed 5%, 10% and 15% of the data points across the ProteomeHD matrix, in triplicate, and repeated treeClust learning to predict protein associations. The Precision-Recall analysis shows that removing data points decreases performance proportionally to the amount of removed data, suggesting that adding additional data would likely enhance performance further.



Supplementary Figure 4. Validation of treeClust and tSNE performance on an independent proteomics dataset
(**a**) treeClust was applied to the cancer proteomics dataset from Lapek *et al*[18]. It outperforms both Pearson and Spearman correlation, as shown by a Precision-Recall analysis using Reactome annotations as the gold standard. Note that treeClust builds only one decision tree per condition, i.e. 41 trees on this dataset, too few for a standard analysis. Therefore, treeClust was performed iteratively, obtaining the mean co-regulation score of 100 treeClust forests, each generated from 10 random experiments. (**b**) Co-regulation map for the Lapek *et al* dataset, made by tSNE from treeClust scores. As in the correlation network of the original report (Fig. 2 in ref. 18), CORUM protein complexes are colored. In contrast to a network there is not a limited number of arbitrarily arranged, pairwise links, but the position of each protein reflects its similarity or dissimilarity with all other proteins in the map. This makes it possible to place all proteins in a functional context, not just those that are directly linked to members of the core network. It also allows for a hierarchical analysis of protein associations, with increasing distances indicating weaker co-regulation. For example, the subunits of the protein complexes in the enlarged map area (inset) are clustered together, and the distances between the complexes are larger. However, all four complexes have roles in vesicular trafficking.



SUPPLEMENTARY MOVIE LEGENDS

Supplementary Movie 1. Interaction of peroxisomal membrane protrusions with mitochondria in COS-7 cells. *See* Fig. 4a-f.

COS-7 cells were transfected with PEX11β-EGFP, mitochondria were stained with Mitotracker (red), and analysed by live-cell imaging using an IX81 microscope (Olympus) equipped with a CSUX1 spinning disk head (Yokogawa). A peroxisome interacts with a mitochondrion via its membrane protrusion, and follows it, occasionally detaching and re-establishing contact. 200 stacks of 9 planes (0.5 µm thickness, 100 ms exposure) were taken in a continuous stream. 118 frames, 14× speed. Scale bar, 5 µm.

Supplementary Movie 2. Interaction of peroxisomal membrane protrusions with mitochondria in COS-7 cells. *See* Fig. 4g-m and legend Movie 1.

Note a peroxisome at the bottom, which interacts with a mitochondrion via its membrane protrusion and then wraps around it, possibly to increase the membrane contact area. 200 stacks of 9 planes (0.5 µm thickness, 100 ms exposure) were taken in a continuous stream. 200 frames, 14× speed. Scale bar, 5 µm.

Supplementary Movie 3. Interaction of peroxisomal membrane protrusions with mitochondria in COS-7 cells. *See* legend Movie 1.

A mitochondrion, which moves to the left, is dragging a peroxisome with a membrane protrusion with it, indicating that the organelles are tightly tethered to each other. 200 stacks of 9 planes (0.5 µm thickness, 100 ms exposure) were taken in a continuous stream. 100 frames, 14× speed. Scale bar, 5 µm.

ONLINE METHODS

Data selection for ProteomeHD

MS raw data were produced in-house or downloaded from the PRIDE repository[44]. Only experiments fulfilling the following inclusion criteria were considered:

(1) Comparative proteomics experiments, i.e. relative protein quantitations of two or more biological states. For example, cells treated with an inhibitor *vs.* mock control. (2) Biological - not biochemical - comparisons, i.e. fold-changes must have been brought about *in vivo*, not by differential biochemical purification. For example, SILAC-labelled cells were treated with inhibitor or mock control, harvested and combined, and chromatin was enriched on the combined sample. In such cases any observed fold-change reflects the response to the inhibitor in the living cell, for example a protein re-localising from cytoplasm onto chromatin. We did not consider experiments that compared, for example, a whole-cell lysate with a chromatin-enriched fraction, as this would measure the impact of the biochemical enrichment rather than a biological event. (3) Quantitation by “stable isotope labeling by amino acids in cell culture” (SILAC)[43]. (4) Samples of human origin.

In addition to these conceptual considerations, the following restrictions were imposed by the data processing pipeline: (5) The SILAC mass shift introduced by heavy arginine must be distinct from heavy lysine. (6) Raw data acquired on an Orbitrap mass spectrometer. (7) Samples alkylated with iodoacetamide, resulting in carbamidomethylation of cysteines.

In total, we considered 294 experiments (SILAC ratios) from 31 projects. A full list of these is provided in Supplementary Table 2.

In-house data collection

80 experiments were performed in-house and analyzed chromatin-enriched samples. Of these, 65 measured the effect of growth factors, radiation and other perturbations on interphase chromatin, which was prepared using Chromatin Enrichment for Proteomics (ChEP)[75]. About half of these experiments had previously been published[34]. Another 15 experiments documented perturbations specifically on freshly replicated chromatin, which was prepared using Nascent Chromatin Capture (NCC)[76].

MS raw data processing

The 5,288 MS raw files were processed using MaxQuant 1.5.2.8[77] on a Dell PowerEdge R920 server. Default MaxQuant search parameters were used with the following exceptions: In group-specific parameters, match type was set to “No matching”. In global parameters, “Re-quantify” was enabled, minimum ratio count was set to 1 and “Discard unmodified counterpart peptide” was disabled. Also in global parameters, writing of large tables was disabled. SILAC labels were set as group-specific parameters as indicated in Supplementary Table 2. Canonical and isoform protein sequences were downloaded from Uniprot[53] on 28th May 2015, considering only reviewed SwissProt entries that were part of the human proteome.

Protein fold-changes were then extracted from the proteinGroups file returned by MaxQuant. Non-normalized SILAC ratios were considered for downstream analysis and log2

transformed. From triple labelling experiments, the heavy/light and medium/light ratios - but not the heavy/medium ratios - were considered. Proteins detected in less than 4 experiments were discarded, as were proteins labeled as contaminants, reverse hits and those only identified by a modification site. We named the resulting data matrix ProteomeHD. It covers 10,323 proteins and protein isoforms, mapping to 9,987 genes, and 294 SILAC ratios. On average, each protein has 112 SILAC measurements. Each experiment covers, on average, 3,928 proteins. ProteomeHD can be downloaded as Supplementary Table 1.

Protein co-regulation analysis using unsupervised machine-learning

We used the R[78] package for the treeClust[45] algorithm to learn expression dissimilarities between proteins in ProteomeHD. For improved accuracy, we only considered 5,013 proteins that were detected in ≥ 95 experiments. TreeClust is an unsupervised machine-learning algorithm based on decision trees that can handle missing values. Note that treeClust was designed not only to measure inter-point dissimilarities but also to perform clustering[45]. However, in this study we use it only to calculate dissimilarities, via the treeClust.dist function. The dissimilarity specifier was set to d.num = 2, so that dissimilarities are weighted according to tree quality. The protein co-regulation score between two proteins was defined as 1 - treeClust dissimilarity. While the co-regulation score is continuous, some analyses benefitted from a simplified categorical approach. In these cases, an arbitrary cut-off was chosen to define “co-regulated protein pairs” (> 0.5) and “not co-regulated pairs” (≤ 0.5).

To visualize ProteomeHD as a 2D co-regulation map, treeClust dissimilarities were subjected to t-Distributed Stochastic Neighbor Embedding (t-SNE)[47,48] using the Rtsne package for R. The theta parameter was set to zero, perplexity to 50 and 1,500 iterations were performed.

Functional annotation of co-regulated proteins

To test if protein co-regulation reflects co-function (Fig. 1e) we defined three sets of “functionally related” protein pairs (subunits of the same protein complexes, enzymes catalyzing consecutive metabolic reactions and proteins with identical subcellular localization) as previously described[23].

To test larger groups (not pairs) of co-regulated proteins for functional enrichment, we analyzed enrichment of Gene Ontology terms using the topGO[79] R package (Fig. 1g). For each protein we tested the group of its co-regulation partners for GO term enrichment. Because some proteins are co-regulated with no or very few other proteins, we restricted the analysis to the 2,139 proteins that are co-regulated with at least 10 proteins. The three aspects (Biological process, Molecular function, Cellular component) of GO were downloaded from QuickGO[80] with taxon set to human and qualifier to null. Rather than the whole proteome, only proteins that were included in the treeClust analysis and had GO annotations were used as the gene “universe” or background for the topGO analysis. Enrichment of GO terms among protein co-regulation groups was tested considering GO graph structure and using a Fisher’s exact test.

Annotation of the co-regulation map

Proteins localizing to specific subcellular compartments were downloaded from Uniprot[53] using the following tags: Nucleus (SL-0191), Nucleolus (SL-0188), Endoplasmic reticulum (SL-0095), Mitochondrion (SL-0173), Cytoplasm (SL-0086), Secreted (SL-0243). Proteins and protein complexes were annotated individually based on the available literature (Fig. 1h).

Creating the www.proteomeHD.net framework

The ProteomeHD online application was written in Python Flask web framework. The interactive plots are generated using Bokeh visualization library for Python (https://github.com/bokeh/bokeh). The Gene Ontology and KEGG enrichment statistics are obtained from a STRING[50] server using an API call with maximally top 100 proteins co-regulated with the query. Only significantly enriched terms (Bonferroni adjusted *P* value < 0.1) are displayed.

Comparison to orthogonal methods

Physical protein-protein-interactions detected by a comprehensive range of small- and large-scale methods were assessed using BioGRID[49], version 3.4.152. BioPlex 2.0[4] served as an example for physical interactions mapped by a single project. Functional protein associations mapped by a large range of methods and publications were inferred from STRING[50], version 10.5.

Annotation of uncharacterized and disease genes

Proteins were defined as “uncharacterized” on the basis of having an annotation score ≤ 3 in Uniprot[53]. The Cancer Gene Census, i.e. genes that can cause cancer when mutated, was curated by COSMIC (Catalogue Of Somatic Mutations In Cancer, version 81)[54]. DisGeNET was used as a comprehensive, curated list of human gene - disease associations[55].

Precision - Recall analyses

A gold standard set of reference proteins was defined using Reactome[81]. Bona fide functionally associated protein pairs (true positives) were defined as protein pairs found in the same “detailed” Reactome pathway. This was inferred from the file UniProt2Reactome.txt (available at https://reactome.org/download-data), where each protein is annotated to the lowest level subset of Reactome pathways. To make sure that only closely related protein pairs were assigned the “true positive” label, we excluded two pathways that were composed of > 200 proteins. We defined protein pairs that are not functionally associated (false positives) as proteins that are never in the same Reactome pathway, at any annotation level. This was inferred from UniProt2Reactome_All_Levels.txt (also available at https://reactome.org/download-data), a file that maps proteins to all levels of the Reactome pathway hierarchy.

On the basis of these true and false positive protein pairs, precision - recall analyses were carried out using the ROCR package[82] for R. The number of false positive protein pairs was randomly downsampled to 90%, so that a random classifier used as reference would have a consistent precision of 0.1.

Comparison of treeClust and Pearson correlation

Pearson's correlation coefficients (PCCs) for proteins across ProteomeHD were obtained using R. As for treeClust learning, only proteins quantified in ≥ 95 experiments were considered.

Comparison of mRNA and protein expression profiling

For the comparison of matched samples and proteins we considered mRNA and protein expression changes across 59 lymphoblastoid cell lines (Fig. 4a). The protein fold-changes are part of ProteomeHD and were originally published by Battle and colleagues[28]. RNA-sequencing data for the same cell lines and proteins were also previously reported[70]. We used the RNA-sequencing data to calculate mRNA fold-changes relative to a 60th cell line, which was the same cell line used as a SILAC reference for the protein expression data. The combined mRNA and protein dataset has been described in more detail elsewhere[23].

For a more comprehensive comparison we considered protein associations predicted using treeClust learning or PCC on the basis of all 294 SILAC ratios in ProteomeHD (Fig. 4b). This was compared to mRNA associations inferred by PCC on the basis of all human mRNA expression data processed by STRING. STRING's state-of-the-art mRNA coexpression analysis pipeline considers all microarray and RNA-sequencing data deposited in the GEO repository[71], resulting in one of the largest mRNA coexpression analyses available to date[50,72]. Note that for this comparison we did not use the STRING coexpression score, which is calibrated against the KEGG database, but the original uncalibrated Pearson's correlations, which were kindly provided by Damian Szklarczyk. STRING PCCs are calculated separately for one- and two-channel microarrays and RNA-sequencing experiments. We used the average of these for the precision - recall analysis, which performed better than any individual experiment type.

Validation of treeClust and tSNE on the cancer proteomics dataset

Lapek *et al* measured the abundances for 6,911 proteins in 41 different breast cancer cell lines[18]. These data are available as Supplementary Table 2 (tab 3) of their report. As described by Lapek *et al*, we converted the protein intensities into log2 fold-changes over the median intensity measured for each protein across all cell lines. We then calculated Pearson's and Spearman's rank correlations for all possible protein pairs using R's base function. The Spearman's correlation coefficients obtained in this way are identical to the ones obtained by Lapek *et al* using the cor.prob function (Supplementary Table 6 in their report[18]). We also determined treeClust co-regulation scores for all protein pairs. However, treeClust can only grow one decision tree per input variable, i.e. 41 in this dataset, which would be too few for it to perform properly. To circumvent this, we forced treeClust to generate 1,000 decision trees by applying it iteratively. We created 100 treeClust forests, each generated with a random subset of 10 of the 41 variables, and used the average co-regulation score for downstream analysis. Precision-recall analysis using a Reactome gold standard and tSNE visualization were performed as described above. The CORUM protein complexes displayed in Lapek *et al*'s Figure 2, reported in their Supplementary Table 7[18], were color-coded in the co-regulation map.

General data processing

Unless specified otherwise, all data processing was performed in R[78], where possible using the data.table package[83]. All plots were created using the ggplot2 package[84].

Plasmids, siRNA, and antibodies

For cloning of peroxisome-targeted Miro1, the C-terminal TMD and tail of Myc-Miro1 (kindly provided by P. Aspenström, Karolinska Institute, Sweden) was exchanged by a PEX26/ALDP fragment previously shown to target proteins to the peroxisome membrane (I Castro, DM Richards, J Metz, JL Costello, JB Passmore, TAS, A Gouveia, D Ribeiro, MS, submitted). PEX11β-EGFP was kindly provided by G. Dodt (Univ. of Tuebingen, Germany). PEX11β siRNA (AUU AGG GUG AGA AUA GAC AGG AUGG) (Eurofins) was previously verified[85]. Control siRNA (si-GENOME nontargeting siRNA pool #2) was obtained from GE Healthcare (D-001206-14-05). Antibodies used were as follows: rabbit polyclonal antibody against PEX14 (1:1400, kindly provided by D. Crane, Griffith University, Australia); mouse monoclonal antibody 9E10 against the Myc epitope (1:200, Santa Cruz Biotechnology, Inc., sc-40), rabbit monoclonal antibody against PEX11β (1:1000, Abcam, ab181066); rabbit polyclonal antibody against GAPDH (1:2000, ProSci3783). Secondary anti-IgG antibodies against rabbit (Alexa 594, 1:1000, Molec. Probes/Life Technol. A21207) and mouse (Alexa 488, 1:400, Molec. Probes/Life Technol. A21202) were obtained from ThermoFisher Scientific. HRP-coupled donkey polyclonal antibody against rabbit IgG (1:5000) was obtained from Biorad (172-1013).

Cell culture and transfection

COS-7 cells (African green monkey kidney cells; ATCC CRL-1651), and PEX5 deficient fibroblasts (kindly provided by H. Waterham, AMC, University of Amsterdam, NL) were cultured in DMEM (high glucose, 4.5 g/L) supplemented with 10% FBS, 100 U/ml penicillin and 100 µg/ml streptomycin at 37°C (5% CO_2, 95% humidity) (HERACell 240i CO_2 incubator). COS-7 cells were transfected using diethylaminoethyl-dextran (Sigma-Aldrich). dPEX5 fibroblasts have enlarged peroxisomes, which facilitates the visualization of membrane extensions. For transfection of dPEX5 fibroblasts, the Neon® Transfection System (Thermo Fisher Scientific) was used following the manufacturer's protocol. Briefly, cells (seeded 24h before transfection) were washed once with PBS and trypsinized using TrypLE Express. Trypsinized cells were resuspended in complete medium, pelleted by centrifugation, and washed with PBS. The cells were once again centrifuged and carefully resuspended in 110 µl buffer R. For each condition, 4×10^5 cells were mixed with the DNA construct (5 µg) or with 100 nM siRNA. Cells were microporated using a 100 µl Neon tip with the following settings: 1400 V, 20 ms, one pulse. Microporated cells were immediately seeded into plates with prewarmed complete medium (without antibiotics) and incubated at 37°C with 5% CO_2 and 95% humidity. The efficiency of silencing was monitored by immunoblotting of cell lysates and confirmed as previously reported[85].

Immunofluorescence and microscopy

Cells grown on glass coverslips were processed for immunofluorescence 24h after transfection. Cells were fixed for 20 min with 4% paraformaldehyde in PBS (pH 7.4),

permeabilized with 0.2% Triton X-100, and blocked with 1% BSA, each for 10 min. Incubation with primary and secondary antibodies took place for 1h each in a humid chamber. Coverslips were washed with ddH_2O to remove PBS and mounted with Mowiol medium on glass slides. All immunofluorescence steps were performed at room temperature and cells were washed three times with PBS between each individual step. Cell imaging was performed using an IX81 microscope (Olympus) equipped with an UPlanSApo 100×/1.40 oil objective (Olympus). Digital images were taken with a CoolSNAP HQ2 CCD camera and adjusted for contrast and brightness using the Olympus Soft Imaging Viewer software and MetaMorph 7 (Molecular Devices). For live-cell imaging, COS-7 cells were plated in 3.5 cm diameter glass bottom dishes (Cellvis). MitoTracker Red CMXRos (Life Technologies) at 100 nM was used for visualisation of mitochondria. Live-cell imaging data was collected using an Olympus IX81 microscope equipped with a Yokogawa CSUX1 spinning disk head, CoolSNAP HQ2 CCD camera, 60 x/1.35 oil objective. Digital images were taken and processed using VisiView software (Visitron Systems, Germany). Prior to image acquisition, a controlled temperature chamber was set-up on the microscope stage at 37ºC, as well as an objective warmer. During image acquisition, cells were kept at 37ºC and in CO_2–independent medium (HEPES buffered). 200 stacks of 9 planes (0.5 µm thickness, 100 ms exposure) were taken in a continuous stream. All conditions and laser intensities were kept between experiments.

Quantification and statistical analysis of peroxisome morphology and interaction

Analysis of statistical significance was performed using GraphPad Prism 5 software. A two-tailed unpaired *t* test was used to determine statistical difference against the indicated group. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$. For analysis of peroxisome morphology, a minimum of 150 cells were examined per condition, and organelle parameters (e.g. membrane protrusions) were microscopically assessed in at least three independent experiments. The analysis was made blind and in different areas of the coverslip. Organelle interaction and contact time were analysed manually from live-cell imaging data using MetaMorph 7 (Molecular Devices). A region of interest (ROI) was drawn in different areas of the cell. Spherical and elongated peroxisomes within the ROI were tracked over the whole time course, and the frequency and duration of contacts monitored. Multiple interactions of the same peroxisome with mitochondria were treated as separate events. Data are presented as mean ± SD.

Data availability

The data that support the findings of this study are available as Supplementary Tables 1 and 2, and on www.proteomeHD.net. All data analysis has been performed with publically available and documented R packages that are referred to in the online methods section.

REFERENCES

1. Gavin, A.-C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440,** 631–636 (2006).
2. Havugimana, P. C. *et al.* A census of human soluble protein complexes. *Cell* **150,** 1068–1081 (2012).
3. Hein, M. Y. *et al.* A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* **163,** 712–723 (2015).
4. Huttlin, E. L. *et al.* Architecture of the human interactome defines protein communities and disease networks. *Nature* **545,** 505–509 (2017).
5. Dunkley, T. P. J., Watson, R., Griffin, J. L., Dupree, P. & Lilley, K. S. Localization of organelle proteins by isotope tagging (LOPIT). *Mol. Cell. Proteomics* **3,** 1128–1134 (2004).
6. Foster, L. J. *et al.* A mammalian organelle map by protein correlation profiling. *Cell* **125,** 187–199 (2006).
7. Christoforou, A. *et al.* A draft map of the mouse pluripotent stem cell spatial proteome. *Nat. Commun.* **7,** 8992 (2016).
8. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science* **356,** (2017).
9. Costanzo, M. *et al.* A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353,** (2016).
10. Mülleder, M. *et al.* Functional Metabolomics Describes the Yeast Biosynthetic Regulome. *Cell* **167,** 553–565.e12 (2016).
11. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270,** 467–470 (1995).
12. DeRisi, J. L., Iyer, V. R. & Brown, P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278,** 680–686 (1997).
13. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 14863–14868 (1998).
14. Kim, S. K. *et al.* A gene expression map for Caenorhabditis elegans. *Science* **293,** 2087–2092 (2001).
15. Hughes, T. R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102,** 109–126 (2000).
16. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302,** 249–255 (2003).
17. Wang, J. *et al.* Proteome Profiling Outperforms Transcriptome Profiling for Coexpression Based Gene Function Prediction. *Mol. Cell. Proteomics* **16,** 121–134 (2017).
18. Lapek, J. D., Jr *et al.* Detection of dysregulated protein-association networks by high-throughput proteomics predicts cancer vulnerabilities. *Nat. Biotechnol.* **35,** 983–989 (2017).
19. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165,** 535–550 (2016).
20. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509,**

582–587 (2014).
21. Fortelny, N., Overall, C. M., Pavlidis, P. & Freue, G. V. C. Can we predict protein from mRNA levels? *Nature* **547,** E19–E20 (2017).
22. Batada, N. N., Urrutia, A. O. & Hurst, L. D. Chromatin remodelling is a major source of coexpression of linked genes in yeast. *Trends Genet.* **23,** 480–484 (2007).
23. Kustatscher, G., Grabowski, P. & Rappsilber, J. Pervasive coexpression of spatially proximal genes is buffered at the protein level. *Mol. Syst. Biol.* **13,** 937 (2017).
24. Hurst, L. D. It's easier to get along with the quiet neighbours. *Mol. Syst. Biol.* **13,** 943 (2017).
25. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* **4,** e309 (2006).
26. Ebisuya, M., Yamamoto, T., Nakajima, M. & Nishida, E. Ripples from neighbouring transcription. *Nat. Cell Biol.* **10,** 1106–1113 (2008).
27. Khan, Z. *et al.* Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science* **342,** 1100–1104 (2013).
28. Battle, A. *et al.* Genomic variation. Impact of regulatory variation from RNA to protein. *Science* **347,** 664–667 (2015).
29. Geiger, T., Cox, J. & Mann, M. Proteomic changes resulting from gene copy number variations in cancer cells. *PLoS Genet.* **6,** e1001090 (2010).
30. Stingele, S. *et al.* Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. *Mol. Syst. Biol.* **8,** 608 (2012).
31. Dephoure, N. *et al.* Quantitative proteomic analysis reveals posttranslational responses to aneuploidy in yeast. *Elife* **3,** e03023 (2014).
32. Ohta, S. *et al.* The protein composition of mitotic chromosomes determined using multiclassifier combinatorial proteomics. *Cell* **142,** 810–821 (2010).
33. Wu, L. *et al.* Variation and genetic control of protein abundance in humans. *Nature* **499,** 79–82 (2013).
34. Kustatscher, G. *et al.* Proteomics of a fuzzy organelle: interphase chromatin. *EMBO J.* **33,** 648–664 (2014).
35. Wu, Y. *et al.* Multilayered genetic and omics dissection of mitochondrial activity in a mouse reference population. *Cell* **158,** 1415–1430 (2014).
36. Kustatscher, G., Grabowski, P. & Rappsilber, J. Multiclassifier combinatorial proteomics of organelle shadows at the example of mitochondria in chromatin data. *Proteomics* **16,** 393–401 (2016).
37. Okada, H., Ebhardt, H. A., Vonesch, S. C., Aebersold, R. & Hafen, E. Proteome-wide association studies identify biochemical modules associated with a wing-size phenotype in Drosophila melanogaster. *Nat. Commun.* **7,** 12649 (2016).
38. Williams, E. G. *et al.* Systems proteomics of liver mitochondria function. *Science* **352,** aad0189 (2016).
39. Gupta, S., Turan, D., Tavernier, J. & Martens, L. The online Tabloid Proteome: an annotated database of protein associations. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkx930
40. Rieckmann, J. C. *et al.* Social network architecture of human immune cells unveiled by quantitative proteomics. *Nat. Immunol.* **18,** 583–593 (2017).

41. Kim, M.-S. *et al.* A draft map of the human proteome. *Nature* **509,** 575–581 (2014).
42. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347,** 1260419 (2015).
43. Ong, S.-E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1,** 376–386 (2002).
44. Vizcaíno, J. A. *et al.* 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44,** D447–56 (2016).
45. Buttrey, S. E. & Whitaker, L. R. treeClust: an R package for tree-based clustering dissimilarities. (2015).
46. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **45,** D331–D338 (2017).
47. Van Der Maaten, L. & Hinton, G. Visualizing High-Dimensional Data Using t-SNE. *J. Mach. Learn. Res.* **9,** 26 (2008).
48. Van Der Maaten, L. Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **15,** 3221–3245 (2014).
49. Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34,** D535–9 (2006).
50. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **45,** D362–D368 (2017).
51. McNally, K. E. *et al.* Retriever is a multiprotein complex for retromer-independent endosomal cargo recycling. *Nat. Cell Biol.* (2017). doi:10.1038/ncb3610
52. Mamińska, A. *et al.* ESCRT proteins restrict constitutive NF-κB signaling by trafficking cytokine receptors. *Sci. Signal.* **9,** ra8 (2016).
53. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45,** D158–D169 (2017).
54. Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45,** D777–D783 (2017).
55. Piñero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* **45,** D833–D839 (2017).
56. Andrews, S. J. & Rothnagel, J. A. Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.* **15,** 193–204 (2014).
57. Aspden, J. L. *et al.* Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. *Elife* **3,** e03528 (2014).
58. D'Lima, N. G. *et al.* A human microprotein that interacts with the mRNA decapping complex. *Nat. Chem. Biol.* **13,** 174–180 (2017).
59. Chu, Q. *et al.* Identification of Microprotein-Protein Interactions via APEX Tagging. *Biochemistry* (2017). doi:10.1021/acs.biochem.7b00265
60. Slavoff, S. A. *et al.* Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.* **9,** 59–64 (2013).
61. Meyer, B., Wittig, I., Trifilieff, E., Karas, M. & Schägger, H. Identification of two proteins associated with mammalian ATP synthase. *Mol. Cell. Proteomics* **6,** 1690–1699 (2007).
62. Chen, R., Runswick, M. J., Carroll, J., Fearnley, I. M. & Walker, J. E. Association of two

proteolipids of unknown function with ATP synthase from bovine heart mitochondria. *FEBS Lett.* **581,** 3145–3148 (2007).
63. Fujikawa, M., Ohsakaya, S., Sugawara, K. & Yoshida, M. Population of ATP synthase molecules in mitochondria is limited by available 6.8-kDa proteolipid protein (MLQ). *Genes Cells* **19,** 153–160 (2014).
64. Schrader, M., Costello, J. L., Godinho, L. F., Azadi, A. S. & Islinger, M. Proliferation and fission of peroxisomes - An update. *Biochim. Biophys. Acta* **1863,** 971–983 (2016).
65. Schrader, M., Costello, J., Godinho, L. F. & Islinger, M. Peroxisome-mitochondria interplay and disease. *J. Inherit. Metab. Dis.* **38,** 681–702 (2015).
66. Devine, M. J., Birsa, N. & Kittler, J. T. Miro sculpts mitochondrial dynamics in neuronal health and disease. *Neurobiol. Dis.* **90,** 27–34 (2016).
67. Costello, J. L. *et al.* Predicting the targeting of tail-anchored proteins to subcellular compartments in mammalian cells. *J. Cell Sci.* **130,** 1675–1687 (2017).
68. Rodríguez-Serrano, M., Romero-Puertas, M. C., Sanz-Fernández, M., Hu, J. & Sandalio, L. M. Peroxisomes Extend Peroxules in a Fast Response to Stress via a Reactive Oxygen Species-Mediated Induction of the Peroxin PEX11a. *Plant Physiol.* **171,** 1665–1674 (2016).
69. Mattiazzi Ušaj, M. *et al.* Genome-Wide Localization Study of Yeast Pex11 Identifies Peroxisome-Mitochondria Interactions through the ERMES Complex. *J. Mol. Biol.* **427,** 2072–2087 (2015).
70. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464,** 768–772 (2010).
71. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.* **41,** D991–5 (2013).
72. Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43,** D447–52 (2015).
73. Gandhi, S. J., Zenklusen, D., Lionnet, T. & Singer, R. H. Transcription of functionally related constitutive genes is not coordinated. *Nat. Struct. Mol. Biol.* **18,** 27–34 (2011).
74. Jovanovic, M. *et al.* Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. *Science* **347,** 1259038 (2015).
75. Kustatscher, G., Wills, K. L. H., Furlan, C. & Rappsilber, J. Chromatin enrichment for proteomics. *Nat. Protoc.* **9,** 2090–2099 (2014).
76. Alabert, C. *et al.* Nascent chromatin capture proteomics determines chromatin dynamics during DNA replication and identifies unknown fork components. *Nat. Cell Biol.* **16,** 281–293 (2014).
77. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372 (2008).
78. R Core Team. R: A Language and Environment for Statistical Computing. (2017).
79. Alexa, A. & Rahnenfuhrer, J. topGO: enrichment analysis for gene ontology. *R package version 2.30.0* (2016).
80. Binns, D. *et al.* QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics* **25,** 3045–3046 (2009).
81. Fabregat, A. *et al.* The Reactome pathway Knowledgebase. *Nucleic Acids Res.* **44,**

D481–7 (2016).
82. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **21,** 3940–3941 (2005).
83. Dowle, M. & Srinivasan, A. data.table: Extension of `data.frame`. (2017).
84. Wickham, H. *ggplot2: Elegant graphics for data analysis*. (Springer, 2009).
85. Costello, J. L. *et al.* ACBD5 and VAPB mediate membrane associations between peroxisomes and the ER. *J. Cell Biol.* **216,** 331–342 (2017).
85. Costello, J. L. *et al.* ACBD5 and VAPB mediate membrane associations between peroxisomes and the ER. *J. Cell Biol.* **216,** 331–342 (2017).

Manuscript 5. “A Primer on Data Analytics in Functional Genomics: How to Move from Data to Insight?”

Pages 82 - 96

Manuscript available online, DOI: https://doi.org/10.1016/j.tibs.2018.10.010

A primer on data analytics in functional genomics: how to move from data to insight?

Piotr Grabowski[1], Juri Rappsilber[1,2#]

[1] Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, 13355 Berlin, Germany
[2] Wellcome Centre for Cell Biology, University of Edinburgh, Edinburgh EH9 3BF, UK
[#] Correspondence: juri.rappsilber@ed.ac.uk

Keywords: functional genomics, systems biology, machine learning, data science, data integration

Abstract

High-throughput methodologies and machine learning have been central in developing systems-level perspectives in molecular biology. Unfortunately, performing such integrative analyses has traditionally been reserved for bioinformaticians. This is now changing with the appearance of resources to help bench-side biologists become skilled at computational data analysis and handling large omics datasets. Here, we show an entry route into the field of omics data analytics. We provide information about easily accessible data sources and suggest some first steps for aspiring computational data analysts. Moreover, we highlight how machine learning is transforming the field and how it can help make sense of biological data. Finally, we suggest good starting points for self-learning and hope to convince readers that computational data analysis and programming is not intimidating.

Can a “traditional” biologist handle big data?

Biologists are facing an exciting yet challenging time with the increasing availability of high-throughput datasets that need to be analyzed and understood. These omics datasets can be either integrated with self-generated data or re-analyzed independently. In the former case, the extra dimension provided by the new data can help generate additional hypotheses on biological systems or support hypothesis validation. In the second case, one can consider published data from a different perspective than that intended in the original study, integrating additional data sources, to make new discoveries without having to invest the time and funds in acquiring new data. Re-analysis and re-purposing of published data is a growing trend [1]. The field of biological sciences is expecting a rise in specialists in data integration and interpretation.

Integrative multi-omics is a rapidly growing field, as reviewed by [2,3]. Additionally, one of the exciting fields with increasing amounts of impact and deposited data are the single cell technologies which encompass genomics, transcriptomics and epigenomics [4,5]. These technologies can be especially powerful when combined with other types of data [6].

The term “multi-omics” refers to the process of integrating data from different high-throughput technologies. Examples of such combinations are:

1. Genomics + transcriptomics, often used in expression quantitative trait loci (eQTL) studies, which can elucidate genomic variants that are important for cellular functions and disease

2. Transcriptomics + proteomics, relating how the transcriptome is shaping the proteome to the possible post-transcriptional and post-translational mechanisms governing this process , as reviewed in [7]
3. Proteomics + metabolomics, correlating differences in protein levels with the metabolites they regulate, synthesize or degrade [8,9]
4. Epigenetics + transcriptomics + proteomics, particularly how the regulatory state of the genome influences gene expression [10] or to obtain a holistic view of stem cell differentiation [11]
5. Phenomics + genomics + transcriptomics, relating external phenotypic traits to genetic sequences and gene expression, which can be helpful in plant biotechnology, for example [12]

Analyzing and making sense of such large datasets can be challenging. A natural ally for this task is machine learning, which is becoming the go-to method for developing analytical workflows for multivariate omics data. It can be used to build models for data classification (for example, to separate healthy and sick patients or protein members of different subcellular components), to cluster data into separate groups, reduce the dimensionality of the dataset for visualization and perform missing value estimation. However, using machine learning requires more knowledge and experience than performing basic statistical hypothesis testing in Excel-like spreadsheet environments. One has to understand the basic concepts in order to avoid producing nonsensical results.

Moreover, data processing, integration and modeling require some degree of programming skills. For this reason, analyzing such data and using machine learning has traditionally been delegated to computer-savvy experts. This often prohibits any hands-on contact from the domain specialists with their data, especially in heavily wet lab-oriented fields. Programming languages such as R and Python offer unlimited power for analysis but require some level of fluency in writing instructions and knowing relevant functions and packages. Knowing at least one analytics platform is paramount to performing any integrative omics study.

This manuscript is a conceptual primer aimed mainly at graduate students, PhD students and post-doctoral researchers who want to start their journey into computational data analysis, but are not sure about the overall breadth of the field, which are the important first steps to take, and what resources are available. We propose a meta-level workflow consisting of four elements: 1) obtaining processed data from public repositories, which can be used alone or in conjunction with self-generated data, 2) hands-on manipulation and processing methods for large datasets, 3) using statistics and machine learning to find significant differences and/or relationships, 4) accessing knowledge and annotation databases to help extract novel insights (Figure 1). Finally, we give some tips on learning resources that might be helpful to start one's journey into integrative data analytics and machine learning.

Where to find publicly available data?

The volume of biological and biomedical data deposited into public repositories and databases is vast and growing every week. This offers a valuable resource to those who are able to navigate it. The data is free and instantly available. This can allow for rapid testing of one's ideas without delays associated with experiment planning and data acquisition. Some of these repositories are listed in Table 1.

The NCBI's Gene Expression Omnibus (GEO) is an example of such a repository which, as of May 2018, contains nearly 4500 curated datasets on gene expression, epigenetics and genome variation profiling. A useful web resource for GEO-deposited data is the ARCHS4 (https://amp.pharm.mssm.edu/archs4/) from the Ma'ayan lab [13], which provides access to processed gene expression tables from the raw data deposited in GEO and Sequence Read Archive (SRA). Of note, the main difference between GEO and SRA is that GEO contains processed data while the raw data (such as FASTQ files from a sequencing run) are deposited into the SRA. This means that if one is looking for "ready-to-use" gene expression tables, one should search the GEO.

The Encyclopedia of DNA Elements (ENCODE, [14]) consortium provides a high quality multi-omics data resource for human, mouse, worm and fruit fly models. It contains data on gene expression, epigenetics and 3D genome conformations that are generated through a variety of technologies. Additionally, the ENCODE consortium provides computational annotation such as predicted DNA regulatory elements.

ProteomeXchange [15] stores published proteomics datasets from over 9000 projects, covering a multitude of species. The datasets tagged 'biological/biomedical' pertain to the general research audience, or can be tagged 'technical', if they are more relevant to the specialized proteomics community. Sometimes, the deposited data is in the so-called "raw" format only, which would require a preliminary processing step using proteomics software before it can be interpreted. However, one can typically find the processed protein or peptide quantification tables in the accompanying manuscript.

The European Genome-Phenome Archive (https://ega-archive.org [16]) offers a large collection of biomedical omics data from multiple studies. However, as is often the case with medical databases containing sensitive patient information, one has to apply to gain access via official channels.

The GTEx Consortium Portal [17] (www.gtexportal.org) stores omics data from a panel of 53 human tissues from densely genotyped donors. The combination of gene expression data with genomic variants and patient information greatly facilitates eQTL studies.

dbGaP [18] (https://www.ncbi.nlm.nih.gov/gap) is a database archiving data about interactions of human genotype and phenotype. The data types encompass DNA variation, SNP assays, DNA methylation, copy number variation and gene expression profiling using technologies such as RNAseq and microarrays. Those are linked to phenotype data such as disease-related clinical status.

Single Cell Expression Atlas (https://www.ebi.ac.uk/gxa/sc/) and SCPortalen (http://single-cell.clst.riken.jp/) are repositories for data acquired using single cell technologies, such as single cell RNAseq.

Aside from technology- and domain-specific resources, initiatives now exist for the global integration of omics datasets according to the FAIR principles ("findable, accessible, interoperable and reusable"). The biggest such initiative is the Omics Discovery Index [19] (https://www.omicsdi.org/), which provides an open-source platform for discovery, access and dissemination of published omics data, and currently integrates 11 repositories. An interesting feature available on this platform is the "similar dataset" section, which can be used to search for other datasets that are conceptually related, similarly to recommended products in online stores.

Repository	**Data type**	**Link**
Gene Expression Omnibus	Gene expression, non-coding RNA profiling, epigenetics, genome variation profiling	www.ncbi.nlm.nih.gov/geo/
ENCODE	Epigenetics, gene expression, computational predictions	www.encodeproject.org
ArrayExpress	DNA sequencing, gene and protein expression, epigenetics	www.ebi.ac.uk/arrayexpress/
European Genome-Phenome Archive*	Various omics with phenotype data (biomedical studies)	https://ega-archive.org
PRIDE, ProteomeXchange	Proteomics, protein expression, post-translational modifications	www.ebi.ac.uk/pride/archive/ http://www.proteomexchange.org/
1000 Genomes	Genome sequences, sequence variants	www.internationalgenome.org
MetaboLights	Metabolomics	www.ebi.ac.uk/metabolights/

GTEx*	Gene expression (microarrays and RNAseq), genome sequences	www.gtexportal.org
NIH/NCI Genomic Data Commons	Gene expression, epigenetics, miRNA-seq (focus on cancer)	https://portal.gdc.cancer.gov
NIH dbGaP*	Genotypes, gene expression, epigenetics, phenotypes	https://www.ncbi.nlm.nih.gov/gap
cBioPortal	Focused on cancer, contains data on gene copy numbers, gene and protein expression, DNA methylation and clinical data	http://www.cbioportal.org
Single Cell Expression Atlas	Single cell gene expression (RNAseq)	https://www.ebi.ac.uk/gxa/sc/
RIKEN SCPortalen	Single cell gene expression (RNAseq)	http://single-cell.clst.riken.jp/

Table 1. Summary of large data repositories for omics analytics.
* needs granted access for individual-level data

How to analyze big datasets?

After downloading the dataset, the next step is to carry out an integrative analysis. Initially, this process involves a series of data quality checks (such as looking at data distributions and ranges or looking for any missing values) and joining of datasets based on common ID systems (usually requires downloading ID translation tables). Subsequently, one can then perform the desired statistical analyses or run machine learning workflows and/or annotate the data using external knowledge bases. All of these steps require appropriate software.

Next-generation sequencing data often needs processing before it can be represented in e.g. expression table. To help with these steps, the Galaxy platform [20] offers powerful solutions. It was developed with user-friendliness and simplicity in mind to allow non-specialists to handle genome and transcriptome data using a simple web-based user interface. Importantly, the user doesn't have to worry about providing enough computational resources as these are provided by many Galaxy-hosting institutions. Alternatively, a Galaxy server can be quickly set-up on a local server.

KNIME [21] is an accessible entry point for time-constrained biologists or for those daunted by programming. It is a graphical user interface (GUI) analytics environment that offers a 'point and click' alternative to classical programming. One can create node-based workflows in which each node is a function that takes in a certain object (for example, gene and protein expression tables), processes it and outputs the results (for example, combined expression data as one matrix). This modular approach offers flexibility and allows one to be creative while keeping the entire workflow easy to follow and reproducible. The "Node Guide" section of the KNIME web page is a great starting point with many examples and downloadable workflows (https://www.knime.com/nodeguide). Moreover, a hub for bioinformatics problems was recently developed to share KNIME workflows for biological data processing and analysis (https://cibi.uni-konstanz.de/hub). More information on using KNIME in the life sciences can be found here [22].

Choosing between GUI-based analytical platforms such as KNIME or "classical" programming languages is a personal matter. KNIME offers a lot of ready-to-use functionalities to combine using a graphical user interface. While this allows for a quicker start, it also has limitations (for example, the user is limited only to the implemented nodes). Programming languages such as R and Python offer much more flexibility for data analytics and are considered the standard tools of trade in research and industry. The choice between R and Python is mostly related to personal preferences. However, it might be more productive to start with a language that is more commonly used in one's professional environment as this enables code sharing and hands-on help from colleagues. Both R and Python offer very versatile and powerful analytical environments. Until recently, R was a more popular choice among biologists as it had more mature libraries for biological data (including the popular Bioconductor package repository). This is changing now, as the statistical and biological analytics suite for Python is being constantly expanded. Both languages have a syntax that is relatively easy to learn and there are no major speed differences between the two when it comes to typical data operations. One advice is to simply try both for a short period of time and see which language is a better fit.

An important aspect of productive analytical programming is selecting the integrated development environment (IDE). IDEs are programs that help programmers to write code by providing access to coding tools, an interactive programming console, plotting areas and variable inspectors. Analyzing data using R and Python without an IDE is more challenging and we highly recommend using one such as RStudio for R and PyCharm or Spyder for Python.

One has to be cautious when integrating data from many sources such as multiple technologies and even laboratories. Most quantification technologies require proper data normalization procedures, for example using a control sample that can take into account measurement noise related to a given platform. It is advisable to work using normalized values or to calculate them, if both the sample of interest and a control are available in the repository. In the worst case, the observed signal in the data might be simply technical noise and not genuine biological change, due to lack of proper normalization. Furthermore, it is important to understand how a given unit is being used in the field. For example, RNAseq expression values FPKM and RPKM are typically

used for visualization and ranking. However, one should avoid using those widely used units for differential gene expression analysis [23]. Good practices for other types of data, such as ChIP-seq, can be found elsewhere [24]. We strongly recommend familiarizing oneself with the way analyses are carried out in respective fields prior to downloading and integrating omics datasets.

How can machine learning help you with your data?

Dealing with big datasets is not easy. To address this, one of the tools that has become very popular in the life sciences is machine learning. In brief, machine learning is a collective term for computer algorithms that iteratively fit a predictive model to the observed data. This model can then be generally applied to predict properties of yet unencountered data, as long as they can be described by the same features. The breadth and depth of this dynamic field has been extensively reviewed [21–23]. Here, we will focus on the practical basics regarding the usefulness of machine learning in biology and provide an example of a machine learning workflow design in Box 1.

Generally, machine learning approaches are divided into two main classes: supervised and unsupervised algorithms. Supervised learning algorithms build a mathematical description (a model) of how a combination of features, such as a gene expression values, relates to some target variable, such as “is important in cancer progression”. These models can then be used to predict the target variable (classes) for data that the model has not yet encountered. An example of this is predicting subcellular localization of proteins [24–26]. Here, one has to first feed the algorithm a dataset together with high-quality annotation, such as proteins assigned to known subcellular compartments (a training set), on which to train the model. After this process, the trained classifier can be used to assign subcellular localizations of other proteins in the dataset. Similarly to the classification task, a supervised machine learning algorithm can be trained to predict continuous values instead of classes (i.e. perform regression), such as chromatographic retention times of peptides [25] or predicting gene expression levels using data on epigenetics and genomic features [26].

Unsupervised approaches, as opposed to supervised approaches, don’t require a pre-specified target variable of interest. Instead, this broad group of algorithms can help find (and exploit) structure in the data. An example of such approach widely used in biology is data clustering which allows to group observations according to their properties. One can imagine a panel of samples which are not clearly distinguished by some binary classification (like cancer/healthy), but rather having various genomic mutations. Having obtained protein expression profiles for each of the samples, one can use an unsupervised approach to see which of those mutations behave similarly to one another. A typical algorithm used in such situation is hierarchical clustering which generates a dendrogram in the process. Cutting this dendrogram at a selected height, results in formation of distinct clusters. These clusters can be then analyzed for functional enrichments (described in more detail in the next section). Unsupervised approaches have been useful for finding groups of co-regulated proteins in cancer [30], finding co-behaving mRNA and miRNA modules in time-series data [31] or finding co-expressed genes in many samples [7,32].

Yet another type of unsupervised algorithms allows for dealing with high-dimensional data, for example when one is interested in visualizing it or detecting outliers, by performing dimensionality reduction. One of the most popular algorithms for this task is principal component analysis (PCA). A good description of how it works and its applications in biology can be found elsewhere [27].

Another interesting application of machine learning is identification of novel predictive features for an observed phenotype (known collectively as "feature importance analysis" or "feature selection"). Here, machine learning is first used for a classification or regression task as described above. However, during this process, many algorithms can inform the user about which of the used features were the most important for a given task. Subsequently, one can look at how well the selected features correlate with the target variable. An example of such approach is expanding the model of nonsense-mediated mRNA decay (NMD) [28]. Here, Lindeboom et al. looked at levels of NMD in human cancers and developed additional descriptors based on genomic features (such as length of an exon harboring mutation). By using Random Forest-based regression they could identify which of those new features are important for predicting NMD efficiency, thereby expanding the current model. A short review of such approaches in biology can be found in [29].

Machine learning pipelines can be built using R, Python and KNIME (among many other languages and platforms). While KNIME offers a great selection of machine learning nodes, including WEKA [30] and H2O (http://docs.h2o.ai/) implementations, it offers less flexibility for pipeline development compared to programming languages such as R and Python. We found that starting with machine learning in KNIME and switching to "classical" programming languages worked best for many of our students. This allowed them to first learn the absolute basics of analytics and subsequently give them more creative freedom.

One of the best places to start using machine learning in R is the "caret" package which offers functions for data processing, classification and regression algorithms, feature selection and model evaluation tools. Similarly to R, Python offers a powerful machine learning environment: "scitkit-learn" [31]. Moreover, a good place to start one's journey with machine learning is downloading the Iris dataset and following one of the many tutorials for a respective machine learning environment (for example, http://scikit-learn.org/stable/auto_examples/datasets/plot_iris_dataset.html).

How to annotate results and generate hypotheses?

Biological data that has accumulated over the last decades is collated in databases using systems of annotations and ontologies. One can use these external databases to help explain functional relationships between genes or proteins of interest in new datasets. For example, using information about pathways can indicate if observed expression changes are modulating particular cellular functions.

A popular knowledge base is Uniprot (www.uniprot.org), which is a protein-centric resource, annotating the proteomes of many studied species. Swiss-Prot is the manually curated part of the

database, offering high-quality annotation. It should be preferred over the electronically generated TrEMBL annotation for functional genomics analyses. The “Retrieve/ID mapping” tool by Uniprot allows mapping of both protein/gene identifiers between different systems (such as RefSeq to Uniprot Accession numbers) and query lists of proteins in order to annotate them with biological properties such as protein sequences, domains, subcellular localization, etc.

BioMart is another widely-used database (with a helpful R package biomaRt [32] and a Python library [33]), found at https://www.ensembl.org/biomart/martview/. BioMart offers biological annotation such as genomic coordinates, transcripts and proteins associated with a given gene; sequences, GC-content, genetic variants or protein domains. The genes of interest (“the query”) are configured in the “Filters” section of the database while the relevant biological information that one may wish to download is configured in the “Attributes” section. The resulting annotated data table can be then saved to disk as a .csv file and integrated into the analytical workflow by matching the gene, transcript or protein IDs.

For genome-level annotation, NCBI offers the Genome Data Viewer (https://www.ncbi.nlm.nih.gov/genome/gdv/), a tool for exploring eukaryotic genomes. This tool can be used to find positions of genes and annotate the genome track with various types of external data. Another similar tool for genome-level analysis is the UCSC Genome Browser (https://genome.ucsc.edu/), which focuses predominantly on human and mouse genomes and offers vast amounts of functional data integrated in “tracks” that are aligned to a given genome. UCSC Genome Browser can seem overwhelming at first, but the steep learning curve for this tool is worth enduring.

Some online resources offer even more “distilled” levels of biological information. STRING [34] offers a database (https://www.string-db.org) on functional connectivity between genes/protein. Users can search for interaction networks between genes/proteins of interest or download the entire database. STRING collates an array of biological sources such as biochemical experiments, text mining and co-expression studies and produces an integrated score. It offers a very simple and fast way to check if a group of genes/proteins are functionally related. Apart from the integrated score, STRING also performs simple GO and KEGG enrichment analysis, further aiding hypotheses development. An alternative resource to STRING is the BioGRID [35], which hosts a variety of interaction data for multiple species. Other easy-to-use tools for functional enrichment and pathway analysis are the Gene Ontology-centered DAVID [36], which can help discover biologically important modules after performing differential expression analysis or data clustering. Enrichr [37], available at http://amp.pharm.mssm.edu/Enrichr/enrich, is another tool that takes a list of genes and calculates enrichments in many functional categories such as pathways, ontologies or transcription-factor binding. Finally, Gene Set Enrichment Analysis (GSEA, [38]) can help analyze whether an *a priori* defined group of genes is significantly affected in given biological states.

In addition to knowledge bases that contain annotation for multiple species, there are specialized resources curated by communities that are focused on specific organisms or groups of organisms. Examples include the Saccharomyces Genome Database (https://www.yeastgenome.org/),

WormBase aimed at nematodes (https://wormbase.org), FlyBase aimed at *Drosophila* (http://flybase.org) or SubtiWiki focused on *Bacillus subtilis* biology (http://subtiwiki.uni-goettingen.de/).

Annotation database / Tool name	**Description**	**Link**
Uniprot	Comprehensive proteomics knowledge base (functions, pathways, sequences, modifications, literature references, ID conversion)	https://www.uniprot.org
Biomart	Gene-centric database with ID conversion, genomic features (such as exons, introns, UTRs), sequences, positions of genes in the genome	https://www.ensembl.org/biomart/martview/
NCBI Genome Data Viewer	A web tool for exploration and analysis of eukaryotic genome assemblies	https://www.ncbi.nlm.nih.gov/genome/gdv/
UCSC Genome Browser	A collection of tools for analysis of genomes with a plethora of available data “tracks” such as epigenetic signals and genomic features	https://genome.ucsc.edu/
StringDB	A database of known and predicted protein-protein interactions. Integrates functional relationship data from various sources.	https://www.string-db.org/
BioGRID	Curated database of physical and genetic interactions based on various experimental sources	https://thebiogrid.org/
DAVID	Gene Ontology and pathway analysis web tool for calculation of functional enrichments in lists of genes or proteins	https://david.ncifcrf.gov/
Enrichr	Web tool for calculating various functional enrichments in lists of genes or proteins	https://amp.pharm.mssm.edu/Enrichr/

G:Profiler	Web tools for functional profiling of groups of genes or proteins. Contains useful ID conversion and orthology mapping tools.	https://biit.cs.ut.ee/gprofiler/

Table 2. Summary of annotation databases and post-analysis tools helpful in making sense of results in computational analytics.

Where to find help?

Learning to handle large-scale data analysis has become increasingly accessible thanks to numerous resources available on the Internet. Acquiring these specialized skills is no longer limited to hands-on training organized at institutes, but can be done from the comfort of one's office or home and with exceptional time flexibility. One of the largest providers of such resources are Coursera and edX.org. These commercial platforms offer dozens of courses on programming, statistics, machine learning and even genomics.

One of the most popular courses on machine learning is the course offered by Andrew Ng (simply called "Machine Learning", found at https://www.coursera.org/learn/machine-learning). This course is a good place to start one's adventure with machine learning, as the concepts are explained in a very intuitive and math-light way. Another important skill (especially for people interested in using Python and R) is understanding the basic concepts of programming and computer science. One of these courses is the "Introduction to Computer Science and Programming Using Python" offered by MIT on the edX.org platform. Even though the course is Python-based, the concepts learned are transferable to other programming languages, such as R. An advantage of this course is that it is free. In addition to this, a good read is "Ten simple rules for biologists learning to program" [39].

Another worthwhile resource is Coursera's "Statistics with R Specialization", which is a bundle of courses that teaches statistics and R simultaneously and can be of benefit to anyone who is interested in functional genomic analyses (which are inherently statistics-heavy). Aside from commercial providers, there are high quality online courses from EMBL-EBI, which can be found at www.ebi.ac.uk/training/online/. Here, the spectrum of skills is more concentrated on applied biological problems and specific platforms, such as analyzing RNAseq data. Moreover, the genomics and biostatistics courses from Rafael Irizarry (found at https://rafalab.github.io/pages/teaching.html) are another high-quality and free learning resource on biological data analytics. For a more general selection of courses on R, Python and Data Science, one can refer to DataCamp (https://www.datacamp.com/). It offers high quality courses with a free (albeit limited) membership plan.

Furthermore, specialists in the omics field can be accessed through various forums with specific questions. We strongly encourage referring to those forums when analyzing data. This improves one's understanding of the data peculiarities and various analytical approaches needed to extract knowledge from the datasets and allows developing all the required skills much faster, while avoiding potential beginner's mistakes. Biostars (https://www.biostars.org/) and SEQanswers

(http://seqanswers.com/) are forums with very active bioinformatics communities and good places to seek help.

Concluding remarks

The curricula of most bioscience programs already contain elements of computational data analytics. However, there is a need for increased focus on this subject, to encourage students to complete their degrees with a working knowledge of at least one programming language and statistics. Luckily for those who have already finished their formal education, many learning resources are available that are well-structured and contain high-quality material, while forums offer expert advice to overcome any challenges. The only prerequisite is that one has to be prepared to battle through the initial confusion and understand that the time investment will pay off in the near future. Just do it!

Acknowledgements

We would like to thank Francis O'Reilly and Sven Giese for critically reading the manuscript and their helpful suggestions.

References

1 Piwowar, H.A. and Vision, T.J. (2013) Data reuse and the open data citation advantage. *PeerJ* 1, e175
2 Hasin, Y. *et al.* (2017) Multi-omics approaches to disease. *Genome Biol.* 18, 83
3 Kim, M. and Tagkopoulos, I. (2018) Data integration and predictive modeling methods for multi-omics datasets. *Mol Omics* 14, 8–25
4 Liberali, P. *et al.* (2015) Single-cell and multivariate approaches in genetic perturbation screens. *Nat. Rev. Genet.* 16, 18–32
5 Gawad, C. *et al.* (2016) Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* 17, 175–188
6 Karemaker, I.D. and Vermeulen, M. (2018) Single-Cell DNA Methylation Profiling: Technologies and Biological Applications. *Trends Biotechnol.* 36, 952–965
7 Liu, Y. *et al.* (2016) On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* 165, 535–550
8 Boccio, P. *et al.* (2016) Integration of metabolomics and proteomics in multiple sclerosis: From biomarkers discovery to personalized medicine. *PROTEOMICS-Clinical Applications* 10, 470–484
9 Bahado-Singh, R. *et al.* (2017) Integrated Proteomic and Metabolomic prediction of Term Preeclampsia. *Sci. Rep.* 7, 16189
10 Kustatscher, G. *et al.* (2017) Pervasive coexpression of spatially proximal genes is buffered at the protein level. *Mol. Syst. Biol.* 13, 937
11 Lindeboom, R.G.H. *et al.* (2018) Integrative multi-omics analysis of intestinal organoid differentiation. *Mol. Syst. Biol.* 14, e8227
12 Li, D. *et al.* (2016) Integrated analysis of phenome, genome, and transcriptome of hybrid rice uncovered multiple heterosis-related loci for yield increase. *Proceedings of the National Academy of Sciences* 113, E6026–E6035
13 Lachmann, A. *et al.* (2018) Massive mining of publicly available RNA-seq data from human

and mouse. *Nat. Commun.* 9, 1366
14 ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74
15 Vizcaíno, J.A. *et al.* (2016) 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* 44, 11033
16 Lappalainen, I. *et al.* (2015) The European Genome-phenome Archive of human data consented for biomedical research. *Nat. Genet.* 47, 692–695
17 Carithers, L.J. *et al.* (2015) A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreserv. Biobank.* 13, 311–319
18 Mailman, M.D. *et al.* (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* 39, 1181–1186
19 Perez-Riverol, Y. *et al.* (2017) Discovering and linking public omics data sets using the Omics Discovery Index. *Nat. Biotechnol.* 35, 406–409
20 Afgan, E. *et al.* (2016) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* 44, W3–W10
21 Berthold, M.R. *et al.* (2009) KNIME - the Konstanz Information Miner: Version 2.0 and Beyond. *SIGKDD Explor. Newsl.* 11, 26–31
22 Fillbrunn, A. *et al.* (2017) KNIME for reproducible cross-domain analysis of life science data. *J. Biotechnol.* 261, 149–156
23 Conesa, A. *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol.* 17, 13
24 Bailey, T. *et al.* (2013) Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput. Biol.* 9, e1003326
25 Moruz, L. *et al.* (2010) Training, selection, and robust calibration of retention time models for targeted proteomics. *J. Proteome Res.* 9, 5209–5216
26 Li, Y. *et al.* (2014) Regression analysis of combined gene expression regulation in acute myeloid leukemia. *PLoS Comput. Biol.* 10, e1003908
27 Ringnér, M. (2008) What is principal component analysis? *Nat. Biotechnol.* 26, 303–304
28 Lindeboom, R.G.H. *et al.* (2016) The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nat. Genet.* 48, 1112–1118
29 He, Z. and Yu, W. (2010) Stable feature selection for biomarker discovery. *Comput. Biol. Chem.* 34, 215–225
30 Hall, M. *et al.* (2009) The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter* 11, 10–18
31 Pedregosa, F. *et al.* (2011) Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830
32 Durinck, S. *et al.* (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* 4, 1184–1191
33 Cokelaer, T. *et al.* (2013) BioServices: a common Python package to access biological Web Services programmatically. *Bioinformatics* 29, 3241–3242
34 Szklarczyk, D. *et al.* (2017) The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D368
35 Chatr-Aryamontri, A. *et al.* (2017) The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* 45, D369–D379
36 Huang, D.W. *et al.* (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57
37 Kuleshov, M.V. *et al.* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90–7
38 Subramanian, A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102, 15545–15550

39 Carey, M.A. and Papin, J.A. (2018) Ten simple rules for biologists learning to program. *PLoS Comput. Biol.* 14, e1005871

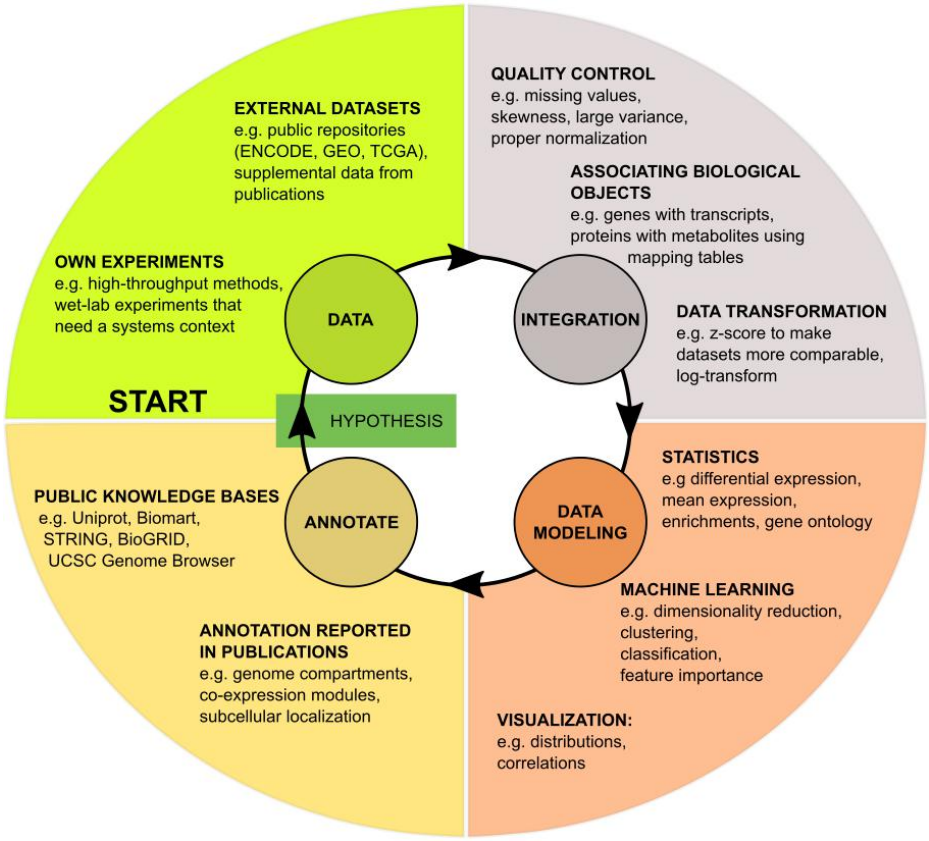


Figure 1. Basic high-level flow of omics data analytics in the life sciences.

BOX 1. Machine learning in biology. How to approach a machine learning analysis for biological questions?

Planning a machine learning analysis can be an overwhelming task for a researcher lacking computational experience. In Figure 2, we divided an example classification workflow (mitochondrial protein prediction) into separate stages while emphasizing important questions that one should consider at each stage.

First, one has to define the target variable of interest and think about what can represent the positive and negative examples of the target.

Secondly, one has to carefully assemble a training set (for supervised methods). Selecting only confident positive and negative examples is essential for the quality of the final analysis. One can perform a manual literature search or take examples (such as proteins) on which there is largest agreement between the different databases.

Thirdly, the input data used by the algorithm should contain enough positive and negative training examples. Importantly, machine learning should not be treated as remedy for low quality data. The classical statistical rule (crap-in, crap-out) applies to machine learning as well.

Subsequently, one should select a proper algorithm for the task. This step depends on the target type (classification vs. regression), number of available training data and technicalities such as presence of missing values.

Finally, the resulting class probabilities (or predicted continuous target values in regression) should be manually evaluated. At this stage one can check if there are any over- or underfitting problems and evaluate the workflow's performance using statistics such as mean accuracy (for classification) or mean squared error (for regression) using left-out ("test") data. Such statistically evaluated ranking can then be used with external annotation databases such as STRING or Enrichr and further validated in the wet-lab or used to build new hypothesis for more computational exploration.

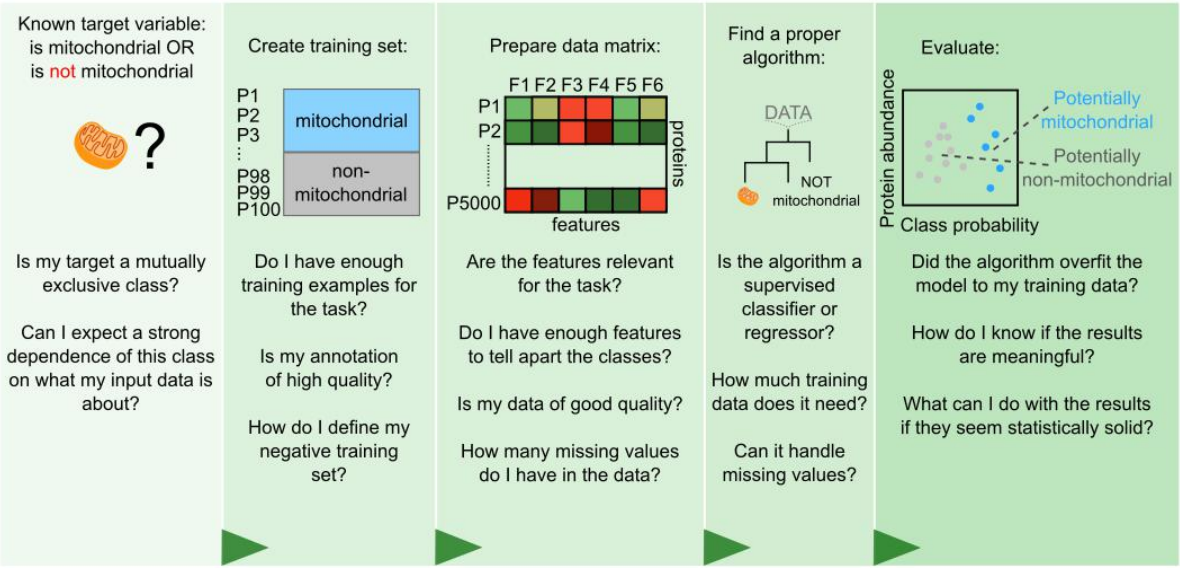


Figure 2. Planning a machine learning-based analysis requires careful consideration at each stage of the analysis. We listed the most general elements of designing such workflow using mitochondrial protein classification task as an example. However, same thinking patterns apply to regression tasks or for feature importance analysis.

Outlook

The combination of advanced analytics such as machine learning and high quality proteomics datasets holds a great promise for developing predictive pipelines. Such pipelines can be used to rank thousands of proteins and millions of possible functional interactions by their strength (or statistical plausibility) and allow the typically expensive and time-consuming follow-up work to focus on the most promising hits. We expect these methods to flourish with the increase of numbers of well-annotated published proteomics datasets and increasing accessibility of data analytics tools for domain specialists. Currently, most of the published protein data and analytical efforts are focused on human, but understanding functional relationships between proteins in other species is also of high importance. Species such as the house mouse, fruit fly or zebrafish are important model organisms used in molecular biology and further expanding our knowledge about those species is paramount for all facets of biology. Furthermore, I expect that protein co-expression analytics will become an important tool in characterizing unknown or understudied microorganisms. By combining RNA sequencing and proteomics, one could measure mRNA and protein expression in many different growth conditions and create a draft functional map of the proteome within a matter of few months. Such organisms could be then more quickly utilized for drug discovery, enzyme production or even help with environmental processes, such as degrading plastic pollution or cleaning up air in the cities. Last, but not least, there is an emerging field of tissue-specific gene co-expression analytics that can shed light on functional interactions between proteins in more natural states, instead of looking only at cultivated cell lines. Such analyses will help speed-up drug discovery and drug toxicity assessment (which are inherently very tissue-specific), by providing more accurate protein functional relationship maps for the many human tissues and cell types making up our bodies. Finally, I strongly believe that with the ever-increasing availability of machine learning resources (both regarding learning how to use it and how to apply it to data), more people will be diving deep into big datasets and make use of their domain knowledge without having to delegate this task to bioinformaticians, who are still in short supply. I believe, the XXI century will be truly the golden age of biology.

Acknowledgments

I thank my beloved parents for giving me the best possible upbringing and supporting me in all possible ways. You encouraged me all my life to see what works best for me and gave me all the resources I needed to move forward, without ever pushing me in the direction you thought is best. I guess this PhD is also a doctorate for you in a way.

I thank my two older brothers, Paweł and Bartek, for being the best role-models in my life and always being there when I needed them. Being the youngest kid in the family has its perks!

Moltes gràcies to my fiancé, Mercè, you've been the most supportive person to have around. You helped me focus when I was down and made the whole process a lot more bearable. Molts petons!

I also thank my close friend, Marcin, who had a huge impact on my life and made me interested in science when I was younger. It all started with physics, but hey, we can't all be physicists! Thanks for cheering for me throughout the whole journey. Dzida!

Completing a PhD doesn't happen in vacuum but in an academic environment. I have to admit that mine was great. I thank my supervisor, Prof. Dr. Juri Rappsilber, for giving me an opportunity to show that I am able to deliver good quality science when I was looking for a new home. You provided me with the necessary resources and enough intellectual freedom, so that I could develop myself into a critically-thinking scientist.

I thank my close collaborator, Dr. Georg Kustatscher, for working with me throughout the last four years and simply being an inspiring researcher with really good ideas. I hope we will co-author yet another manuscript some day.

Finally I thank all the Rappsilber lab members, both in Edinburgh and Berlin, for creating a friendly and motivated workplace. Francis and Ludwig, it's sad to leave and not be able to throw more "Polish vs. Irish vs. Germans" jokes at lunch. Having said that, I am happy to leave all the future mass spec problems to you guys! I'll miss you all.

References

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3): 403–10.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." *Nucleic Acids Research* 25 (17): 3389–3402.

Angermueller, Christof, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. 2016. "Deep Learning for Computational Biology." *Molecular Systems Biology* 12 (7): 878.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, et al. 2000. “Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium.” *Nature Genetics* 25 (1): 25–29.

Barrett, Tanya, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, et al. 2013. “NCBI GEO: Archive for Functional Genomics Data Sets—update.” *Nucleic Acids Research* 41 (D1): D991–95.

Bateman, Alex, Lachlan Coin, Richard Durbin, Robert D. Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, et al. 2004. “The Pfam Protein Families Database.” *Nucleic Acids Research* 32 (Database issue): D138–41.

Battle, Alexis, Zia Khan, Sidney H. Wang, Amy Mitrano, Michael J. Ford, Jonathan K. Pritchard, and Yoav Gilad. 2015. “Genomic Variation. Impact of Regulatory Variation from RNA to Protein.” *Science* 347 (6222): 664–67.

Buttrey, Samuel E., and Lyn R. Whitaker. 2015. “treeClust: An R Package for Tree-Based Clustering Dissimilarities.” https://calhoun.nps.edu/handle/10945/48179.

Camacho, Diogo M., Katherine M. Collins, Rani K. Powers, James C. Costello, and James J. Collins. 2018. “Next-Generation Machine Learning for Biological Networks.” *Cell*, May. https://doi.org/10.1016/j.cell.2018.05.015.

Dam, Sipko van, Urmo Võsa, Adriaan van der Graaf, Lude Franke, and João Pedro de Magalhães. 2018. “Gene Co-Expression Analysis for Functional Classification and Gene–disease Predictions.” *Briefings in Bioinformatics* 19 (4): 575–92.

Dephoure, Noah, Sunyoung Hwang, Ciara O’Sullivan, Stacie E. Dodgson, Steven P. Gygi, Angelika Amon, and Eduardo M. Torres. 2014. “Quantitative Proteomic Analysis Reveals Posttranslational Responses to Aneuploidy in Yeast.” *eLife* 3 (July): e03023.

Dönnes, Pierre, and Annette Höglund. 2004. “Predicting Protein Subcellular Localization: Past, Present, and Future.” *Genomics, Proteomics & Bioinformatics* 2 (4): 209–15.

ENCODE Project Consortium. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” *Nature* 489 (7414): 57–74.

Engelhardt, Barbara E., Michael I. Jordan, Kathryn E. Muratore, and Steven E. Brenner. 2005. “Protein Molecular Function Prediction by Bayesian Phylogenomics.” *PLoS Computational Biology* 1 (5): e45.

Geiger, Tamar, Ana Velic, Boris Macek, Emma Lundberg, Caroline Kampf, Nagarjuna Nagaraj, Mathias Uhlen, Juergen Cox, and Matthias Mann. 2013. “Initial Quantitative Proteomic Map of 28 Mouse Tissues Using the SILAC Mouse.” *Molecular & Cellular Proteomics: MCP* 12 (6): 1709–22.

Giese, Sven H., Yasushi Ishihama, and Juri Rappsilber. 2018. “Peptide Retention in Hydrophilic Strong Anion Exchange Chromatography Is Driven by Charged and Aromatic Residues.” *Analytical Chemistry* 90 (7): 4635–40.

Grabowski, Piotr, Georg Kustatscher, and Juri Rappsilber. 2018. “Epigenetic Variability Confounds Transcriptome but Not Proteome Profiling for Coexpression-Based Gene Function Prediction.” *Molecular & Cellular Proteomics: MCP*, July. https://doi.org/10.1074/mcp.RA118.000935.

Harrow, Jennifer, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, et al. 2012. “GENCODE: The Reference Human Genome Annotation for The ENCODE Project.” *Genome Research* 22 (9): 1760–74.

Hase, Takeshi, Hiroshi Tanaka, Yasuhiro Suzuki, So Nakagawa, and Hiroaki Kitano. 2009. “Structure of Protein Interaction Networks and Their Implications on Drug Design.” *PLoS Computational Biology* 5 (10): e1000550.

Huang, Sijia, Kumardeep Chaudhary, and Lana X. Garmire. 2017. “More Is Better: Recent Progress in Multi-Omics Data Integration Methods.” *Frontiers in Genetics* 8 (June): 84.

Hurst, Laurence D., Csaba Pál, and Martin J. Lercher. 2004. “The Evolutionary Dynamics of Eukaryotic Gene Order.” *Nature Reviews. Genetics* 5 (4): 299–310.

Itzhak, Daniel N., Stefka Tyanova, Jürgen Cox, and Georg Hh Borner. 2016. “Global, Quantitative and Dynamic Mapping of Protein Subcellular Localization.” *eLife* 5 (June). https://doi.org/10.7554/eLife.16950.

Khatri, Purvesh, and Sorin Drăghici. 2005. “Ontological Analysis of Gene Expression Data: Current Tools, Limitations, and Open Problems.” *Bioinformatics* 21 (18): 3587–95.

Kustatscher, Georg, Piotr Grabowski, and Juri Rappsilber. 2016. “Multiclassifier Combinatorial Proteomics

of Organelle Shadows at the Example of Mitochondria in Chromatin Data." *Proteomics* 16 (3): 393–401.
———. 2017. "Pervasive Coexpression of Spatially Proximal Genes Is Buffered at the Protein Level." *Molecular Systems Biology* 13 (8): 937.
Langfelder, Peter, and Steve Horvath. 2008. "WGCNA: An R Package for Weighted Correlation Network Analysis." *BMC Bioinformatics* 9 (December): 559.
Lapek, John D., Jr, Patricia Greninger, Robert Morris, Arnaud Amzallag, Iulian Pruteanu-Malinici, Cyril H. Benes, and Wilhelm Haas. 2017. "Detection of Dysregulated Protein-Association Networks by High-Throughput Proteomics Predicts Cancer Vulnerabilities." *Nature Biotechnology* 35 (10): 983–89.
Li, Dayong, Zhiyuan Huang, Shuhui Song, Yeyun Xin, Donghai Mao, Qiming Lv, Ming Zhou, et al. 2016. "Integrated Analysis of Phenome, Genome, and Transcriptome of Hybrid Rice Uncovered Multiple Heterosis-Related Loci for Yield Increase." *Proceedings of the National Academy of Sciences* 113 (41): E6026–35.
Liu, Ching-Ti, Shinsheng Yuan, and Ker-Chau Li. 2009. "Patterns of Co-Expression for Protein Complexes by Size in Saccharomyces Cerevisiae." *Nucleic Acids Research* 37 (2): 526–32.
Liu, Yansheng, Andreas Beyer, and Ruedi Aebersold. 2016. "On the Dependency of Cellular Protein Levels on mRNA Abundance." *Cell* 165 (3): 535–50.
Li, Yuan-Yuan, Hui Yu, Zong-Ming Guo, Ting-Qing Guo, Kang Tu, and Yi-Xue Li. 2006. "Systematic Analysis of Head-to-Head Gene Organization: Evolutionary Conservation and Potential Biological Relevance." *PLoS Computational Biology* 2 (7): e74.
Lobley, A. E., T. Nugent, C. A. Orengo, and D. T. Jones. 2008. "FFPred: An Integrated Feature-Based Function Prediction Server for Vertebrate Proteomes." *Nucleic Acids Research* 36 (Web Server issue): W297–302.
Lobley, Anna, Mark B. Swindells, Christine A. Orengo, and David T. Jones. 2007. "Inferring Function Using Patterns of Native Disorder in Proteins." *PLoS Computational Biology* 3 (8): e162.
Moruz, Luminita, Daniela Tomazela, and Lukas Käll. 2010. "Training, Selection, and Robust Calibration of Retention Time Models for Targeted Proteomics." *Journal of Proteome Research* 9 (10): 5209–16.
Mulvey, Claire M., Lisa M. Breckels, Aikaterini Geladaki, Nina Kočevar Britovšek, Daniel J. H. Nightingale, Andy Christoforou, Mohamed Elzek, Michael J. Deery, Laurent Gatto, and Kathryn S. Lilley. 2017. "Using hyperLOPIT to Perform High-Resolution Mapping of the Spatial Proteome." *Nature Protocols* 12 (6): 1110–35.
Nora, Elphège P., Bryan R. Lajoie, Edda G. Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, et al. 2012. "Spatial Partitioning of the Regulatory Landscape of the X-Inactivation Centre." *Nature* 485 (7398): 381–85.
Oprea, Tudor I., Cristian G. Bologa, Søren Brunak, Allen Campbell, Gregory N. Gan, Anna Gaulton, Shawn M. Gomez, et al. 2018. "Unexplored Therapeutic Opportunities in the Human Genome." *Nature Reviews. Drug Discovery* 17 (5): 317–32.
Pickrell, Joseph K., John C. Marioni, Athma A. Pai, Jacob F. Degner, Barbara E. Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K. Pritchard. 2010. "Understanding Mechanisms Underlying Human Gene Expression Variation with RNA Sequencing." *Nature* 464 (7289): 768–72.
Planas-Iglesias, Joan, Manuel A. Marin-Lopez, Jaume Bonet, Javier Garcia-Garcia, and Baldo Oliva. 2013. "iLoops: A Protein–protein Interaction Prediction Server Based on Structural Features." *Bioinformatics* 29 (18): 2360–62.
Purmann, Antje, Joern Toedling, Markus Schueler, Piero Carninci, Hans Lehrach, Yoshihide Hayashizaki, Wolfgang Huber, and Silke Sperling. 2007. "Genomic Organization of Transcriptomes in Mammals: Coregulation and Cofunctionality." *Genomics* 89 (5): 580–87.
Radivojac, Predrag, Wyatt T. Clark, Tal Ronnen Oron, Alexandra M. Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, et al. 2013. "A Large-Scale Evaluation of Computational Protein Function Prediction." *Nature Methods* 10 (3): 221–27.
Rao, Suhas S. P., Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, et al. 2014. "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping." *Cell* 159 (7): 1665–80.
Rehman, S. U., S. Asghar, S. Fong, and S. Sarasvady. 2014. "DBSCAN: Past, Present and Future." In

The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014), 232–38.
Ryan, Colm J., Susan Kennedy, Ilirjana Bajrami, David Matallanas, and Christopher J. Lord. 2017. “A Compendium of Co-Regulated Protein Complexes in Breast Cancer Reveals Collateral Loss Events.” *Cell Systems* 5 (4): 399–409.e5.
Schwanhäusser, Björn, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. 2011. “Global Quantification of Mammalian Gene Expression Control.” *Nature* 473 (7347): 337–42.
Serin, Elise A. R., Harm Nijveen, Henk W. M. Hilhorst, and Wilco Ligterink. 2016. “Learning from Co-Expression Networks: Possibilities and Challenges.” *Frontiers in Plant Science* 7 (April): 444.
Singh, Rohit, Daniel Park, Jinbo Xu, Raghavendra Hosur, and Bonnie Berger. 2010. “Struct2Net: A Web Service to Predict Protein–protein Interactions Using a Structure-Based Approach.” *Nucleic Acids Research* 38 (suppl_2): W508–15.
Song, Lin, Peter Langfelder, and Steve Horvath. 2012. “Comparison of Co-Expression Measures: Mutual Information, Correlation, and Model Based Indices.” *BMC Bioinformatics* 13 (December): 328.
Stastna, Miroslava, and Jennifer E. Van Eyk. 2012. “Analysis of Protein Isoforms: Can We Do It Better?” *Proteomics* 12 (19-20): 2937–48.
Su, Junjie, Byung-Jun Yoon, and Edward R. Dougherty. 2010. “Identification of Diagnostic Subnetwork Markers for Cancer in Human Protein-Protein Interaction Network.” *BMC Bioinformatics* 11 Suppl 6 (October): S8.
Szklarczyk, Damian, John H. Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, et al. 2017. “The STRING Database in 2017: Quality-Controlled Protein–protein Association Networks, Made Broadly Accessible.” *Nucleic Acids Research* 45 (D1): D362–68.
Vizcaíno, Juan Antonio, Attila Csordas, Noemi Del-Toro, José A. Dianes, Johannes Griss, Ilias Lavidas, Gerhard Mayer, et al. 2016. “2016 Update of the PRIDE Database and Its Related Tools.” *Nucleic Acids Research* 44 (22): 11033.
Wang, Jing, Zihao Ma, Steven A. Carr, Philipp Mertins, Hui Zhang, Zhen Zhang, Daniel W. Chan, et al. 2017. “Proteome Profiling Outperforms Transcriptome Profiling for Coexpression Based Gene Function Prediction.” *Molecular & Cellular Proteomics: MCP* 16 (1): 121–34.
Williams, Elizabeth J. B., and Dianna J. Bowles. 2004. “Coexpression of Neighboring Genes in the Genome of Arabidopsis Thaliana.” *Genome Research* 14 (6): 1060–67.
Xu, Chao, Jiajia Chen, and Bairong Shen. 2012. “The Preservation of Bidirectional Promoter Architecture in Eukaryotes: What Is the Driving Force?” *BMC Systems Biology* 6 Suppl 1 (July): S21.
Yip, Andy M., and Steve Horvath. 2006. “The Generalized Topological Overlap Matrix for Detecting Modules in Gene Networks.” In *BIOCOMP*, 451–57.
Zhang, Fan, and Jake Y. Chen. 2010. “Discovery of Pathway Biomarkers from Coupled Proteomics and Systems Biology Methods.” *BMC Genomics* 11 Suppl 2 (November): S12.
Zhao, Wei, Peter Langfelder, Tova Fuller, Jun Dong, Ai Li, and Steve Hovarth. 2010. “Weighted Gene Coexpression Network Analysis: State of the Art.” *Journal of Biopharmaceutical Statistics* 20 (2): 281–300.

Zhao, Wei, Peter Langfelder, Tova Fuller, Jun Dong, Ai Li, and Steve Hovarth. 2010. “Weighted Gene Coexpression Network Analysis: State of the Art.” *Journal of Biopharmaceutical Statistics* 20 (2): 281–300.

List of published manuscripts which are part of the dissertation:

1. Kustatscher, Georg, Piotr Grabowski, and Juri Rappsilber. 2017. “Pervasive Coexpression of Spatially Proximal Genes Is Buffered at the Protein Level.” *Molecular Systems Biology* 13 (8): 937. https://doi.org/10.15252/msb.20177548

2. Grabowski, Piotr, Georg Kustatscher, and Juri Rappsilber. 2018. “Epigenetic Variability Confounds Transcriptome but Not Proteome Profiling for Coexpression-Based Gene Function Prediction.” *Molecular & Cellular Proteomics: MCP*, July. https://doi.org/10.1074/mcp.RA118.000935

3. Kustatscher, Georg, Piotr Grabowski, and Juri Rappsilber. 2016. “Multiclassifier Combinatorial Proteomics of Organelle Shadows at the Example of Mitochondria in Chromatin Data.” *Proteomics* 16 (3): 393–401. https://doi.org/10.1002/pmic.201500267

4. Kustatscher, Georg, Piotr Grabowski, Tina A. Schrader, Josiah B. Passmore, Michael Schrader, Juri Rappsilber. 2019. “The human proteome co-regulation map reveals functional relationships between proteins” *bioRxiv* 582247. https://doi.org/10.1101/582247

5. Grabowski, Piotr and Juri Rappsilber. 2019. “A Primer on Data Analytics in Functional Genomics: How to Move from Data to Insight?” *Trends in Biochemical Sciences* 44 (1), 21-32. https://doi.org/10.1016/j.tibs.2018.10.010