RPTU Rheinland-Pfälzische Technische Universität Kaiserslautern Landau

Navigating the Noise: Sparse Profile Analysis of Omics Data

vom Fachbereich Biologie

der Rheinland-Pfälzischen Technischen Universität Kaiserslautern-Landau

zur Verleihung des akademischen Grades „Doktor der Naturwissenschaften“
genehmigte

Dissertation

vorgelegt von

Benedikt Christoph Venn, M.Sc.

Wissenschaftliche Aussprache: 15.04.2024

Berichterstattende:
Prof. Dr. Timo Mühlhaus (Betreuer)
Prof. Dr. Michael Schroda
Prof. Dr. Thomas Nägele

Kaiserslautern im Jahr 2024

D 386

Abstract

Living systems incessantly engage in the regulation of their cellular processes to fulfill their biological functions. Beyond development-related adjustments or cell cycle oscillations, environmental fluctuations compel the system to reorganize metabolic pathways, structural components, or molecular repair and reconstitution mechanisms. These responses manifest across diverse temporal scales, necessitating an intricate regulatory orchestration. Time series experiments have become increasingly popular for charting the chronological order and elucidating the underlying mechanisms. In the era of high-throughput technologies, the majority of cellular molecules can be analyzed in one fell swoop, generating a comprehensive snapshot of the status quo of most present molecules. Methodological advancements also permit the monitoring not only of molecular abundances but also the functional status of transcripts and proteins. However, due to the still high efforts associated with such experiments, the number of measured time points and the replication of measurements remains limited. Resulting datasets contain signals from thousands of molecules, yet they are sparse in temporal resolution and are often imprecise due to biological variability and technical measurement inaccuracies.

This thesis explores the complexities arising from the examination of short time series data and introduces pioneering tools that offer fresh insights into the realm of biological time series analysis. The broad spectrum of analytic possibilities ranges from a molecule-centric investigation of individual time courses to a holistic aggregation of the system’s response to its main characteristics. By creating a modeling framework that applies domain-specific constraints, time-course signals can be transformed from a series of discrete data points into a continuous curve. These curves align with current biological conjectures about molecule kinetics being smooth and devoid of superfluous oscillations. Noise present at individual time points is judiciously accounted for during curve fitting, mitigating the impact of time points with high variance on the curve. Subsequent classification is based on the features of these curves (extreme points and inflection points) and ensures a reduction in data amount and complexity. Succinct labels assigned to each molecule's kinetics encapsulate the signal's most notable features. Besides this modeling approach, an innovative enrichment strategy is introduced, that is independent of prior data partitioning and capable of segregating the temporal response into its thermodynamically relevant components. This approach allows for a continuous assessment of each molecule's contribution to these components, obviating the need for exclusive allocation. The application of various analytical approaches to heat acclimation experiments in Chlamydomonas highlights the relevance and potential of time series experiments and specifically tailored analysis techniques. The integration of different system levels has led to the identification of regulatory peculiarities, such as an increased correlation between transcripts and corresponding proteins during acclimation responses. These and other insights may herald new avenues of research that could ultimately enhance plant robustness in the face of increasing environmental perturbations.

The growing popularity of time series experiments necessitates dedicated analytical approaches that empower researchers and analysts to decipher patterns, discern trends, and unravel the underlying structures within the data, facilitating predictions and the derivation of meaningful conclusions that could potentially build bridges between the interweaved systems levels.

Zusammenfassung

Ein biologisches System, sei es eine einzelne Zelle, ein Gewebe, Organ oder Organismus, beschäftigt sich unermüdlich mit der Regulierung zellulärer Prozesse, um seine biologischen Funktionen zu erfüllen. Über Anpassungen während der Entwicklung und Zellzyklus-Oszillationen hinaus, zwingen Schwankungen von Umweltfaktoren das System dazu, Stoffwechselwege, strukturelle Komponenten oder Reparationsmechanismen neu zu organisieren. Diese Reaktionen verlaufen in unterschiedlichen zeitlichen Abfolgen und Laufzeiten und erfordern dadurch eine komplexe regulatorische Orchestrierung. Zeitreihenexperimente erfreuen sich zunehmender Beliebtheit, um diese Reihenfolgen zu erfassen und die zugrunde liegenden Mechanismen aufzuklären. Die Ära der Hochdurchsatztechnologien ermöglicht Wissenschaftlern, einen Großteil der zellulären Moleküle in einem einzigen Durchgang zu quantifizieren, was eine umfassende Momentaufnahme des zellulären Zustands liefert. Methodische Fortschritte ermöglichen neben der Messung der Molekülabundanz auch eine Schätzung ihrer biologischen Aktivität. Aufgrund des nach wie vor hohen Aufwands solcher Experimente ist die Anzahl von vermessenen Zeitpunkten sowie die Replikatanzahl von Zeitserienexperimenten vergleichsweise gering. Die resultierenden Datensätze enthalten die Messwerte von Tausenden von Molekülen, sind jedoch in ihrer zeitlichen Auflösung spärlich und aufgrund biologischer Variabilität und technischer Messungenauigkeiten oftmals ungenau.

Diese Arbeit befasst sich mit den Herausforderungen, die sich mit der Analyse kurzen, verrauschten Zeitreihen ergeben, und präsentiert die Entwicklung innovativer Methoden, die neue Perspektiven in der biologischen Zeitreihenanalyse eröffnen. Das Spektrum der Analysemöglichkeiten reicht von einer molekül-zentrischen Untersuchung einzelner Zeitverläufe bis hin zu einer ganzheitlichen Aggregation der Reaktion des Systems auf seine Hauptcharakteristiken. Durch die Entwicklung einer Modellierungsstrategie, die domänenspezifische Annahmen durchsetzt, können Zeitseriensignale aus einer Reihe diskreter Datenpunkte in einen kontinuierlichen Abundanz-Verlauf umgewandelt werden. Die entstehenden Kurven entsprechen aktuell gültigen Annahmen über die Kinetik von biologischen Molekülen, indem ihr Verlauf glatt ist und keine unnötigen Oszillationen aufweisen. Vorhandenes Rauschen an einzelnen Zeitpunkten wird bei der Modellierung berücksichtigt, um die Auswirkungen von Zeitpunkten mit hoher Varianz auf die Kurve zu mildern. Eine anschließende Klassifizierung, die auf den Merkmalen dieser Kurven beruht (Lage und Beschaffenheit von Extrem- und Wendepunkten), ermöglicht eine Reduktion der Datenmenge und -komplexität. Jedem Molekül kann so eine Kennzeichnung seiner Kinetik zugewiesen werden, die die auffälligsten Merkmale des Signals zusammenfasst. Neben dieser Zeitserien-Modellierung wird außerdem eine *Label-Enrichment*-Strategie vorgestellt, die von einer vorherigen Aufspaltung des Datensatzes unabhängig ist und außerdem die biologischen Reaktionen in ihre markantesten Komponenten unterteilt. Diese Methodik ermöglicht eine gewichtete Zuordnung der Molekülrelevanz zu diesen Komponenten. Die Anwendung verschiedener analytischer Strategien auf Hitzeakklimatisierungs-Experimente in *Chlamydomonas* soll die Relevanz und das Potenzial von Zeitreihenexperimenten und speziell darauf zugeschnittenen Analysetechniken unterstreichen. Durch die Integration verschiedener Systemebenen konnten regulatorische Besonderheiten unter Hitze ermittelt werden, wie beispielsweise eine erhöhte Korrelation zwischen Transkripten und ihren entsprechenden Protein-Abundanzen. Diese und weitere Einblicke eröffnen neue Forschungsansätze, die angesichts zunehmender klimatischer Veränderungen letztendlich die Widerstandsfähigkeit von Pflanzen steigern könnten.

Die wachsende Popularität von Zeitreihenexperimenten erfordert spezielle analytische Methoden, die Forschende dazu befähigen, zugrunde liegende Muster und Strukturen in den Daten zu entschlüsseln. Dies trägt dazu bei, Vorhersagen zu ermöglichen und Schlussfolgerungen abzuleiten, die potenziell unerkannte Verbindungen zwischen miteinander verflochtenen Systemebenen sichtbar machen.

Eigenständigkeitserklärung

Die vorliegende Dissertation wurde an der Rheinland-Pfälzischen Technischen Universität Kaiserslautern-Landau in der Arbeitsgruppe von Prof. Dr. Timo Mühlhaus angefertigt. Hiermit erkläre ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und nur unter der Zuhilfenahme der angegebenen Quellen und Hilfsmittel angefertigt habe. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt. Die Bestimmungen der Promotionsordnung vom 10. Oktober 1996 und deren Änderung am 30. März 2001 des Fachbereichs Biologie sind mir bekannt.

Kaiserslautern, 19.01.2024

Benedikt Christoph Venn

Darlegung aller benutzten Hilfsmittel und Hilfestellungen

Zur Erstellung dieser Arbeit wurde MS Office genutzt. Das Literaturverzeichnis und Literaturverweise wurden mittels Citavi erstellt. Abbildungen wurden mit MS Office und Plotly.NET erstellt. Zur sprachlichen Verbesserung wurde DeepL genutzt.

Table of contents

1. Introduction

Cellular dynamics

Cellular dynamics, the orchestrated interplay of diverse elements within living cells, form the essence of biological systems. At the molecular level, the interactions between proteins, nucleic acids, and other biomolecules regulate the cellular choreography and affect vital functions. Moving up the hierarchy, cellular networks emerge, comprising signaling pathways, metabolic processes, and genetic regulations that collectively sustain life. These elements seamlessly interact, creating a delicate steady state that ensures the smooth progression of cellular activities (Alberts 2008; Sharom et al. 2004; Qi and Ge 2006).

Understanding the dynamics at different systems levels is crucial to elucidate complex behavior and function or prevent disruptions in this balance leading to disorders and even diseases. Therefore, unraveling the complexities of cellular dynamics holds significance beyond academic curiosity. It provides insights crucial for advancements in medicine, biotechnology, and environmental sciences. At every moment an organism is regulating thousands of processes to either maintain its intended function, processing information gathered by internal or external signals, repair or remodel cellular structures, or cope with new environmental conditions (Tyson et al. 2001). When cells are faced with environmental changes there are several scenarios for how to proceed. If the change drives the cells or cellular processes outside their biological niche for a prolonged period, then senescence or programmed cell death may eliminate damaged cells beyond repair (Kerr et al. 1972; Mittler 2002). These cells cannot fulfill their function and even could cause tumorigenic events leading to further damage to the tissue or organism (Galluzzi and Myint 2023; Galluzzi et al. 2018). Subtle changes in environmental conditions can already be compensated for by the dynamics of metabolism itself and may not require any further action to restore any set point of cellular homeostases. As soon as there is an active intervention to regulate processes in order to (i) navigate to the previous set point, or (ii) navigate to a newly set point, the response is called acclimation. Acclimation not only deals with the mitigation of detrimental effects, but also takes advantage of favorable conditions (Kleine et al. 2021; Kültz 2005; Vonk and Shackelford 2022; Tuteja and Singh Gill 2013). The distinction between beneficial and detrimental perturbation effects is not trivial as a positive condition change in one process can simultaneously be negative for others. Increased oxygen availability for example can be advantageous for the efficiency of the respiratory chain, while increasing DNA damage or protein oxidation (Postmus et al. 2011; Bergamini et al. 2004). The states of all homeostatic systems together define whether a biological system as a whole must take acclimation reactions.

Acclimation

The definition of the term acclimation has been under dispute for many decades. While adaption is widely accepted as a descriptor of genetic adjustments in a population that are inheritable, acclimation describes non-hereditary, usually reversible changes within the lifetime of a single organism to cope with fluctuating environmental conditions (Kleine et al. 2021; Lagerspetz 2006; Vonk and Shackelford 2019). In 1950 Prosser gave two definitions for different kinds of acclimatization. *Genetic acclimatization* which operates by selection, and *physiological acclimatization* in which individuals alter their resistance within genetic limitations (Bishop et al. 1950). Later he defined the term *acclimatization* as genotypic adaptive alterations and the term *acclimation* as phenotypic adaptive alterations of individual organisms (Prosser 1955). In 1968 Precht didn't differentiate genetic from non-genetic processes, but defined *performance acclimation* ("Leistungsakklimatisierung") as processes within normal temperatures, whereas *resistance adaptation* ("Resistenzadaptation") describes the organism's adjustment to extreme temperatures (Precht 1968). Darwin himself seems to have used the term as a general description of any adjustment to abiotic changes. Despite chapter 5 of Darwin's "The Origin of Species" being titled "Acclimatisation" no definition is given, and it is used as a generic description for "adaptation to particular climates" (Darwin 1859). In later studies, *acclimation* described the short-term adjustment to a single specific factor, while *acclimatization* was used to describe the adjustment to a multi-variate change like seasonality (Feder and Hofmann 1999; West 1972) or real-world conditions outside the controlled laboratory (Buguet et al. 2023). Other publications use both terms interchangeably (Acosta et al. 2023; Diego et al. 2023; Banerjee et al. 2023). In some more recent publications, acclimation is not only restricted to abiotic climatic factors, but also biotic interactions (Khlebovich 2017). In the same study, the time span in which acclimation usually occurs is described as seasonal, while other publications - in line with the definition used in this thesis - describe acclimation as a response to environmental fluctuations varying from minutes to days with effects ranging from metabolic adjustments to differential gene expression (Kleine et al. 2021). To further specify acclimation, the concept of homeostasis must be considered (Figure 1): Homeostasis, of which there are many different definitions in scientific literature, describes a set of regulatory machinery that can sense, compare, and control its respective variable (e.g. ion concentration, redox state, energy household, pH) (Modell et al. 2015; Borowitzka 2018; Koolhaas et al. 2011; Torday 2015; Chrousos 2009). Every homeostatic system must have a set point, which is rather a value range instead of an exact value and resembles the range that the variable can take without the biological system being at risk. Minor fluctuations in the cellular redox state can easily be compensated by the passive buffering capacity of the metabolism. Once the variable is outside of the set point (but still within limits that are viable in the short term), coping mechanisms are actively started to bring the variable back into the set point range. This process must be activated deliberately and therefore is part of an acclimation response. Alternatively, the set point may change, and acclimation assists in reaching the new homeostatic set point (e.g. fever). The restriction of acclimation being dependent on gene expression can be found in literature, but to difficulties in the real world. Sometimes acclimation responses that have a drastic effect to cope with environmental changes do not include differential gene expression. A photosynthetic cell, for example, whose redox state is disturbed (out of the homeostatic set point) due to intense illumination may respond by a decoupling of light harvesting complexes and potentially a physical reorganization of thylakoid membranes. These changes, while clearly being part of an acclimation response to increased light, may not require any expression of genes, but rely on signaling pathways that utilize cellular reserves to dissociate the complexes and cause structural adjustments. However, if gene regulation occurs in response to environmental changes, it definitely is part of an acclimation response.

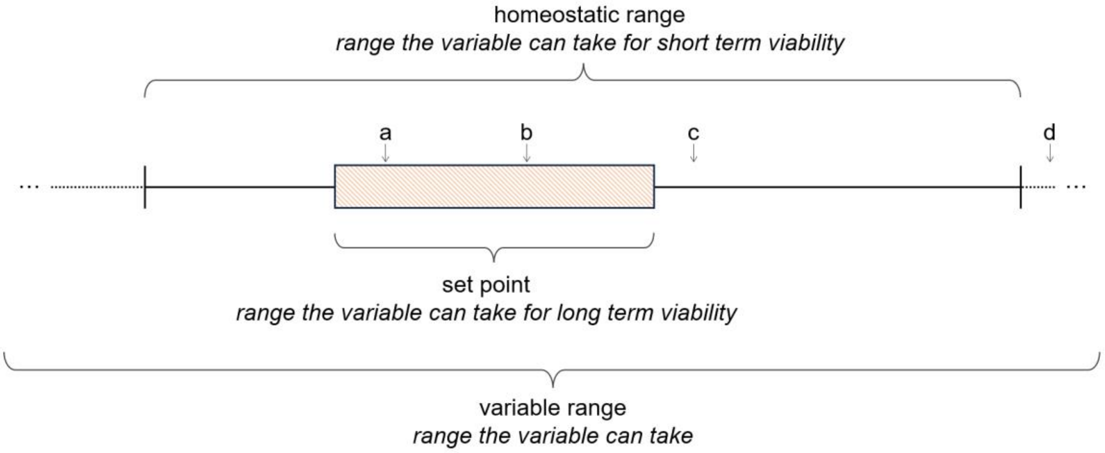

homeostatic range
range the variable can take for short term viability
a
b
c
d
…
…
set point
range the variable can take for long term viability
variable range
range the variable can take

Figure 1 A theoretical homeostatic system regarding intracellular pH. A biological system without effort can tolerate all pH values that lie within the set point (a and b). No acclimation is happening, and no acclimation responses are required when the system transitions from state a to state b. When the pH is outside of the set point (c), acclimation is required to direct the variable toward its set point. If the pH is outside of the range the homeostatic system can control, all processes that rely on the respective homeostatic system (all pH dependent processes and components) irreversibly fail. Alternative scenarios include a sudden relocation of the set point (e.g. fever for homeothermic organisms). If the current variable lies outside of this range, again acclimation responses are activated that aim to bring the variable towards this new set point.

Especially in the realm of short-term acclimation responses, the effectiveness of cellular dynamics relies on the precise coordination of molecular events. The regulatory networks governing these responses involve a myriad of signaling pathways, gene expressions, protein regulation, and metabolic adjustments. Identifying the central contributors within this complex interplay is crucial for deciphering acclimation responses, understanding the evolution of cellular processes, and in the future increasing the ability to predict and engineer cellular reactions (Lopes-Ramos et al. 2017; Kim et al. 2012). The simultaneous presence of identical players in multiple cellular processes makes it challenging to isolate and attribute functional significance to each participant. Nevertheless, understanding these nuanced contributions of individual molecules in acclimation responses becomes imperative to fully understand underlying regulatory dynamics.

As living systems rely on dynamic interactions and a constant control of the cell's state, all processes inevitably must be examined with respect of time. It plays a critical role in shaping the narrative of cellular events and hence requires a time resolved monitoring of cellular players. During acclimation, the biological system may trigger several reactions. Immediate life-threatening changes must be addressed quickly. Drastic changes in osmolarity, redox status, membrane fluidity, or a failure of essential pathways, for example, pose an acute threat. These require rapid recruitment of chaperones or structural components from cellular reserves that quickly can be mobilized (Schroda et al. 2015; Balogh et al. 2011, 2011; Csoboz et al. 2013). If the condition persists, a deeper reorganization could be required after this first aid to ensure continued survival. This can include a lasting change in the proteome. Instead of simply replacing aggregated proteins, efforts must be made to ensure that they do not immediately aggregate again. Either the processes these proteins are involved in are not sustainable under the given circumstances, or additionally synthesized folding capacities must ensure the stability of the proteins or enable their refolding. In the case of heat-induced influences, an accumulation of thermoprotective substances can help to maintain cellular processes, for which the expression of corresponding genes must be activated (Hemme et al. 2014). The entire biological system with all its components (e.g. energy management, anabolism, catabolism, cell cycle) could be affected and must be adapted to the new physiological conditions. Once the external influence has come to an end, the cell can revert the changes that have been made. As long-term changes, epigenetic changes may persist and lead to increased vigilance when the stimulus reappears (Boyko and Kovalchuk). To study these dynamics, many techniques were developed to perform real time in vivo measurements of metabolites, protein interactions, or gene expression to investigate these processes (Niemeyer et al. 2021; Xing et al. 2016; Hamada et al. 2016). However, these techniques are often limited to monitoring just a few molecules simultaneously. In order to get a comprehensive picture of cellular dynamics, huge parts of system levels need to be assessed in an efficient and reproducible way. This can be achieved by taking snapshots at various stages during acclimation, creating reference points that, when combined with high throughput technologies, enable time-resolved data acquisition of thousands of molecules. The resulting time series have interesting properties that can be used to obtain far more information than the individual data points provide in combination. This thesis aims to provide and apply analytical techniques that help to uncover these dynamics, with a focus on transcripts and their associated proteins. The developed tools intend to examine the existence of dynamics and their functions across various systems levels – from the time resolved classification of molecular kinetics to the broader behavior of systems subset in respect with their relevance during acclimation responses. Therefore, it is necessary to understand the types and properties of time series data and compare strategies to analyze these kinds of data in order to elucidate underlying dynamics.

Time series data

Time series consist of readings that have a natural temporal ordering and are used to track the change of these readings over time (Casella et al. 2008). The measurements could be either one dimensional data, e.g., ambient temperature measured over the course of a year, or multidimensional data, e.g. a weather station, that logs temperature data as well as humidity, air pressure, precipitation, and solar radiation intensity. A weather station is a common example of a time series in which time points are equally spaced, meaning that the time span between two adjacent measuring time points is fixed. In contrast, event-driven data collection results in irregular time point spacing (Salfner 2006). These include data logging devices that take a snapshot of the current readings when triggered by a threshold sensor. The same irregular time intervals occur in time series, for which prior knowledge of the expected data exists. To illustrate this subject, imagine an object to be heated to exactly 50 °C. An attached temperature sensor is rated for a limited number of measurement cycles only. To optimize its usage, the sampling rate has to be increased with higher readings. When the plate is cold, the readings can be rare, while in the range of the target temperature, the readings have to be frequent to not miss the point to stop the stove. Especially storage, computation, or cost constraints can lead to the generation of time series data with irregular time point spacings.

In contrast to other data point collections with no temporal ordering, time series data points often are highly dependent on each other while the ordering encodes valuable information (Leung et al. 2021; Jung and Tremayne 2003). Ten measurements of the ambient temperature taken in short succession are not limited to just accurately describing the temperature at each time point but enable the prediction of (i) temperature readings within two of the time points, and (ii) the temperature forecasting of time points that lie outside of the measured time span. Without additional environmental distortions, the temperature reading is constrained by its neighboring points, which limit the expected temperature range. The same is true for cellular protein amounts. If it is ensured that no new distortion of environmental conditions and internal processes occurs, a protein abundance measurement is going to lie between the prior and subsequent measurement time points. However, this statement is not helpful at all even if the environmental conditions are absolutely controllable. Living matter always undergoes some sort of regulation and ongoing processes constantly change the state of the intra-cellular environment. Additionally, expression noise cannot be adequately controlled or predicted (Chowdhury et al. 2021). Especially for biological molecules involved in regulatory processes that depend on signaling cascades or thresholds to be reached, abundance readings may result in unexpected patterns. Sudden expression bursts require complex regulatory machinery whose prediction involves multifaceted differential equation modeling (Luo et al. 2023; Beckman et al. 2021; Gardner et al. 2000). To understand these patterns or to be able to predict them, in theory, all factors influencing the state of a molecule must be known and measured. Quite some progress was made in the past in optimizing analysis and prediction techniques to cover a plethora of measurable molecules. Especially for isolated subprocesses of metabolite conversions it often is sufficient to measure a small set of enzymes that take place in these conversions and their activity status (Weaver et al. 2014). Thereby metabolic flux models can be created and used for predicting metabolite fluxes by varying abundances or activities for one or multiple players, be it proteins, co-factors, or the involved metabolite concentrations themselves (Kim et al. 2008). However, the data collected worldwide is still not sufficient to make reliable predictions about changes in the proteome or transcriptome on a global level, especially during acclimation responses (Lee et al. 2012). The reasons include (i) missing information about unknown protein-coding genes, (ii) unknown interactions between nucleic acids, proteins, and metabolites, (iii) inaccurate measurement techniques, (iv) phenotypic heterogeneity caused by genetic variability, and many more (Ghatak et al. 2019; Fröhlich et al. 2018). It is questionable anyway if a global model is reachable or even desirable. Predictive models – as introduced in the next chapter – must always represent a compromise between accuracy and universality. Accurately describing a specific relationship naturally is prone to overfitting and does not claim to be universal or generally applicable to any conditions (van Impe et al. 2013). However, the growing prevalence of machine learning approaches is undoubtedly revolutionizing the field enabling to connect data to predictive models of unseen accuracy and spanning various subdisciplines (Lopatkin and Collins 2020; Hassoun et al. 2022). But what is the connection of this kind of modeling to time series?

Besides simple data logging approaches that just present the measured data points, time series data can be used to build models and investigate underlying principles or identify factors that influence the measured feature. For the study of climate recordings, many of the factors that influence the readings are known, some may be unknown and none of them is controllable in the sense of performing multiple real-world scenarios. Laboratory experiment setups, however, should be designed to limit the number of factors influencing the sample to one at best. A highly controlled experiment, in which a temperature dependent chemical conversion takes place, is an example of an univariable experiment in which reaction rates can be precisely analyzed by comparing time series observations at different temperature conditions.

To be able to compare these changes in a quantitate manner, each time series can be condensed to a set of coefficients that describe the signal. These coefficients fully characterize the curve shapes and can be used to compare various conditions and infer conditions of e.g. maximal reaction speed. The next chapter deals with such models in general, how they are constructed and parameterized, and how they help to understand and predict the world around us.

Understanding time series requires a model-based representation

Models are simplified representations of real-world processes (Berry and Houston 1995). They capture the system's important aspects and enable the description, analysis, and prediction of system responses (Edwards and Hamson 1989). While models can be conceptual (flow charts and diagrams), computational (simulations), or physical (actual physical prototypes), here I focus on univariable mathematical models that can be expressed as functions taking numerical values as input parameters and output prediction. Factors that have no, or negligible influence are not taken into account by the model, thereby reducing the complexity to a minimum while preserving relevant dependencies. These models can range from single- to multi-coefficient complexity that incorporate a multitude of variables which in turn may influence themselves (e.g. the prior mentioned climate model).

$$f(t) = mt \qquad \text{(Equation 1)}$$

An example for a single coefficient model is a linear regression line through the origin (Equation 1). The function value $f(t)$ is determined by multiplying its input variable t by a scalar coefficient m which resembles the slope of the line. This model can be used to calculate the distance travelled by a car at a constant speed, where t is the time travelled in hours and m is the speed in kilometers per hour (Figure 2A). While it would be straightforward to just have the distance measured, or look at the pedometer, having this model at hand is handy if you want to predict how far you can drive within a specific amount of time. By rearranging the function, you can determine how fast you have to drive to reach your destination in time. Furthermore, you can apply calculus operations to determine the slope or the area under the curve for a specific time interval. While for the presented example this "absement" with its unit of km·h has no intuitive interpretation, it demonstrates the possibilities a model offers even if it just contains a single coefficient and a single input variable. A more complex model with multiple coefficients is the Verhulst growth model, which is used to describe the growth phases of organisms (Vogels et al. 1975).

$$P(t) = \frac{K}{1+\left(\frac{K-P_0}{P_0}e^{-rt}\right)} \text{ (Equation 2)}$$

$$with\ P(t) = population\ size\ at\ time\ point\ t$$

$$r = intrinsic\ growth\ rate$$

$$K = carrying\ capacity$$

$$P_0 = initial\ polulation\ size$$

The origin of the function may seem complex, but in fact, it derives from the easy to interpret differential equation $\frac{dP}{dt} = rP\left(1 - \frac{P}{K}\right)$. It describes that the slope of the growth curve depends on the (scaled) current population count itself (P) and a second term, that with constant K and increasing P reduces the overall slope (Figure 2B). Here the left part describes the rate of change the population experiences over time (the curves slope). Neglecting the second term, the first term on the right describes exponential growth, as the curve's slope just depends on its current reading multiplied with a growth constant that describes the population's average net growth. The second term, however, leads to a gradient descent when P approaches K. When integrated into the so called closed-form function (Equation 2) the three resulting coefficients (r, K, and P_0) have straightforward interpretations and especially r can be used directly to either determine generation times, or to compare populations regarding their maximal growth rate (Perni et al. 2005; Maier and Pepper 2015).

As demonstrated, models are theoretical templates you expect reality to fit in, that can be superimposed (fitted) to measured data. As researchers we observe a phenomenon, construct a mental model out of it and ultimately generate a mathematical model out of the theoretical constraints and the actual observations. The obtained coefficients associated with a model can be used to describe the reality or compare instances of the model. While the presented models have coefficients that have a direct interpretation (e.g. speed or carrying capacity), models of which relevant factors are unknown may have coefficients that lack a logical interpretation.

If the model is unknown, polynomials are practical tools for such signal modeling, in which no empirical or theoretical model is available (Greenland 1995; Royston et al. 1999). They are linear combinations of the input variable and coefficients. A cubic polynomial has a degree of three and consists of four coefficients (a, b, c, and d).

$$f(t) = at^3 + bt^2 + ct + d$$

While polynomials lack interpretable coefficient interpretations, they are convenient in modeling data. By adding additional terms, their flexibility is extendable as required, and calculus operations, as differentiation and integration, are easy and efficient (Edwards 1995) (Figure 2C).

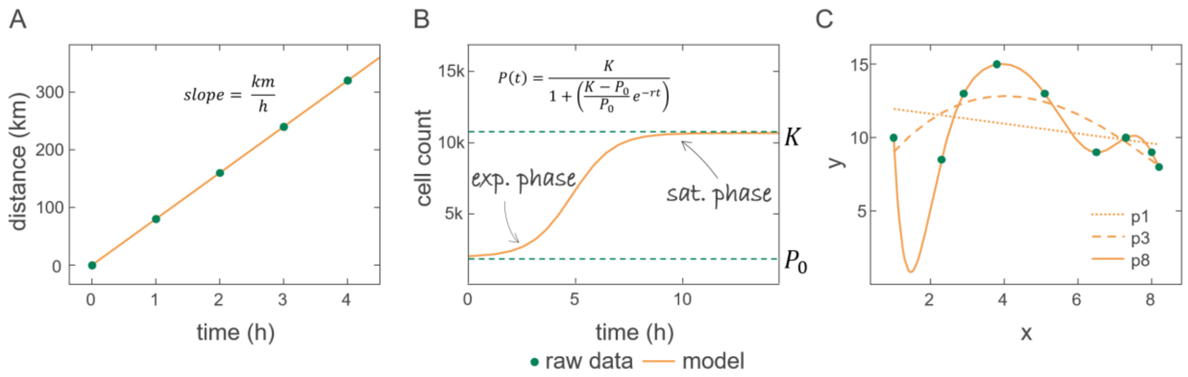


Figure 2 Model based data representation. (A) The distance travelled in km is given as function of time. Blue markers indicate raw data measurements while the orange line is the model representation of $distance = time \cdot slope$. Because the function starts at (0,0), no intercept term is required. (B) Verhulst growth model with three parameters (P_0, K and r) and the variable t. The cell count starts with exponential growth at P_0 before the growth is affected negatively by the carrying capacity (K), leading to a saturation phase in which the cell count asymptotically advances K. (C) The raw data (blue dots) are fitted by polynomials of different degree. While a straight regression line corresponds to a polynomial of degree 1 (p1: $ax + b$), a degree increase leads to increased flexibility (p3: cubic polynomial, $ax^3 + bx^2 + cx + d$). If the degree reaches the number of data points – 1, the polynomial interpolates all points (p8: $ax^8 + bx^7$ ...).

On the one hand, this easy extension of the mathematical formula is great to make the model more flexible, on the other hand, its prone to overfitting (Figure 2C). In the so called *kitchen sink regression* the analyst throws “everything but the kitchen sink” into the regression in hopes of finding the correct model, thereby maximizing the coefficient of determination (R^2). While the resulting model shows a higher fidelity to the data, its prediction capabilities are dramatically reduced, as the areas between measurements are unlikely to be observed (Yin 2022). If underlying models are unknown or not known to full extent, analysts should keep the number of coefficients and thereby the number of assumptions of a model low in order to reduce overfitting effects. This principle is called Occam’s razor and has a long tradition in model selection e.g. the most appropriate model for chemical reaction (Minkin and Carpenter; Carpenter 1984).

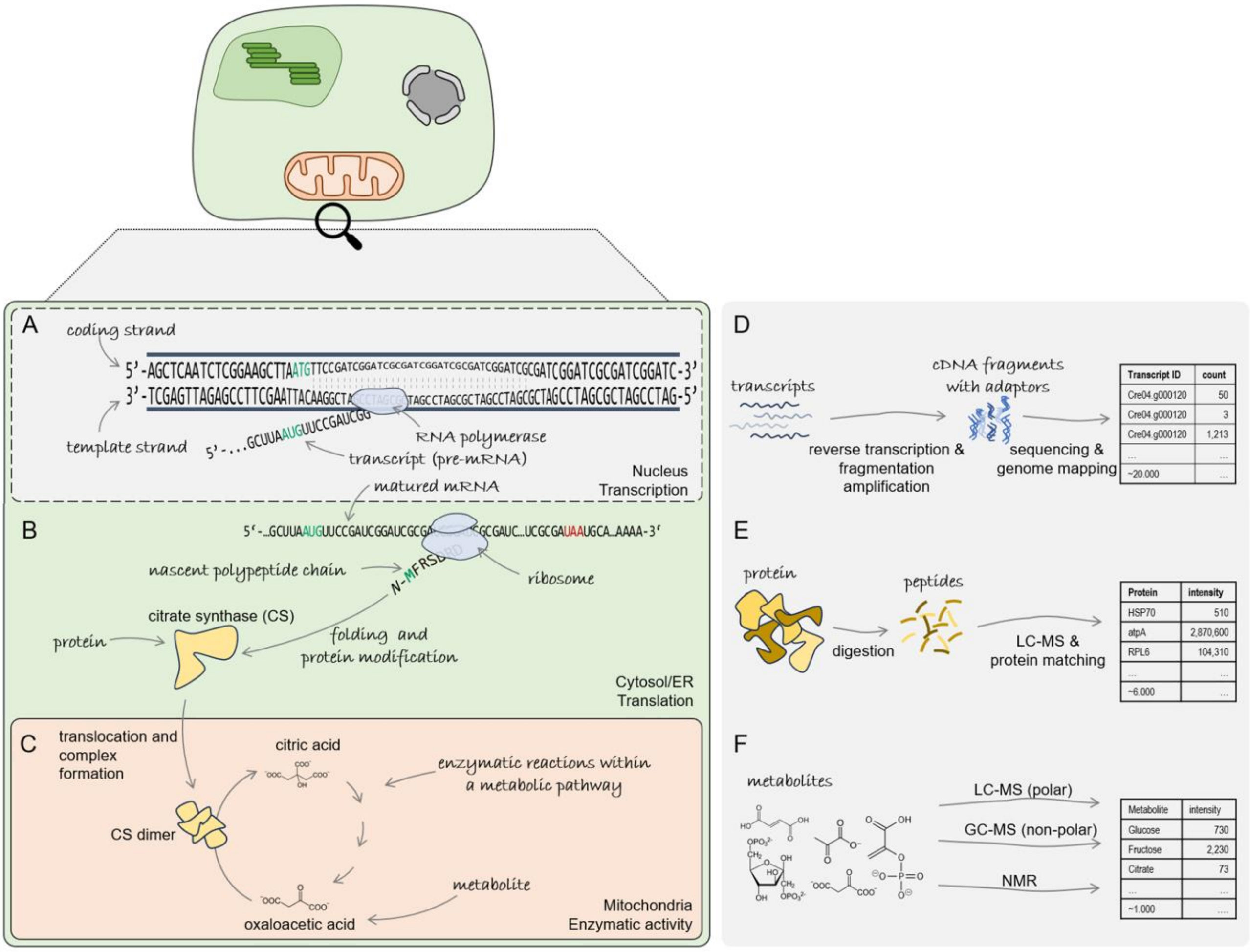
While it frequently is inaccurately paraphrased as “The simplest model is the correct one” (Domingos 1999), its more accurate translation is:

“If everything is equal, the simplest model is probably the correct one”, or more formally *“Entities should not be multiplied without necessity” – Clauberg, Logica vetus et nova. (1654).*

The addition of *necessity* is crucial as it isn’t necessarily beneficial to choose the option with the smallest number of coefficients just because it’s the simplest, speaking the one with the fewest assumptions. In the worst case, the model underfits the underlying phenomenon and therefore isn't useful as a representation (Figure 2C dotted line). Even if the exact mathematical equation is unknown and the model therefore only represents a rough approximation, restrictions are often possible that limit the capabilities of the model. Occam's razor therefore reminds us to select the number of assumptions in such a way that the anticipated properties are fulfilled, but to avoid making additional assumptions in the case of uncertainty. In the end, it - as often - condenses to "make as many assumptions as required, but as few as possible". It is important to note that few assumptions do not necessarily correspond to having few coefficients. In fact, many models just have one or two coefficients, but are very constrained to their specific use case, i.e. radioactive decay (2 coefficients λ and N_0) or Michaelis-Menten model (2 coefficients V_{max} and K_m) (Rutherford 1904; Michaelis and Menten 1913).

Excursus: Omics technologies

Models, however, require data to be modeled on. In modern biology, these data often originate from so called omics technologies that systematically aim to make life quantifiable. A living cell can be characterized by measuring the abundance and – if available – the activity status of its contained molecules. This molecule composition is different between cells of differing specialization (Melé et al. 2015). The plant's leaf cells that are involved in photosynthesis are likely to produce proteins that are crucial for light harvesting and sugar synthesis, while its root cells need a different set for starch storage and structural stability (Shi et al. 2021). The identification of such patterns is not always trivial, as proteins share functionalities, and many proteins are shared between cells of different tissues. Furthermore, the amount of molecule species within a cell is huge and the cell's function often is characterized by just a small subset of its proteome (Emig and Albrecht 2011; Uhlén et al. 2015). Cells consist of thousands of transcript- and protein species with hundreds or several thousand copies each (Ho et al. 2018). Omics technologies enable a holistic and undirected quantification of the molecules of a specific systems level (e.g., proteins, transcripts, or metabolites) (Subedi et al. 2022; Aebersold and Mann 2003; Winkler 1920). In contrast to analyses that rely on the purification of a single molecule species, all molecules are considered equally. Analyzing the number and composition of all proteins in a cell at a specific time point, i.e. the proteome, is the subject of proteomics. The same applies to transcripts (transcriptomics), metabolites (metabolomics), or lipids (lipodomics). There are now several other omics specializations, such as complexomics or interactomics, that deal with the identification and quantification of protein complexes and the interaction of proteins with other cellular molecules respectively (Narad and Kirthanashri 2019). While the comparison of the molecule composition of different tissues helps to analyze substrate flux within an organism or to describe tissue specialization, another interesting approach is to compare e.g. the proteome or transcriptome of the same tissue at differing development stages or environmental conditions. Hence, changes in molecule composition can be pinpointed to the reaction of the system to the stimulus faced during the experiment. The high accuracy necessary to detect subtle abundance changes and to cover the majority of the systems level is facilitated by modern high throughput methods. RNA sequencing (RNASeq) is a *next generation sequencing* (NGS) approach and deals with the quantification of transcript copies. At each moment in every cell, genes are transcribed to transcripts, which in turn are processed to mature mRNA and ultimately translated to proteins (Figure 3). While transcripts in most cases serve as an intermediate step, the resulting proteins act as executive elements that facilitate enzymatic reactions, ensure the cell's structural integrity, assist in transport activities, and serve in various other processes (Rost et al. 2003). However, transcripts can act on their own, e.g. as rRNA being involved in the synthesis of other proteins at ribosomes (Campbell and Farrell 2009). The transcriptome can be described as the entire RNA component of a cell, but often is defined as the polyadenylated products of RNA polymerase II (Tang et al. 2011; Wang et al. 2009). In RNASeq, purified transcripts are converted to cDNA and afterwards sequenced using NGS strategies (Figure 3D). The sequence reads are aligned to a reference genome which subsequently are aggregated and reported as *reads*, *counts*, or *read counts* per transcript in the original sample (Wang et al. 2009). Comparing read counts of different transcripts remains challenging, because of the formation of secondary structures and the probability of counts increasing with the transcript's length. While there are strategies to account for this bias (e.g. TPM normalization), it often is not required as counts of the same transcript are compared across different conditions. Here the length correction factor would be the same for both counts (Zhao et al. 2021). Ultimately protein coding transcripts are translated to proteins by ribosomes. Many ribosomes can attach to a single transcript resulting in a cascade effect during the transcription-translation procedure. Several thousand kinds of proteins exist in an organism, each of which exists in several thousand copies within a single cell (Ho et al. 2018). Using mass spectrometry-based proteomics (MS proteomics), a huge proportion of proteins can be measured in a single machine run (Figure 3E). Cell lysate containing all or specific factions of proteins is prepared by digesting them into peptides using endoproteases (e.g. Trypsin). Subsequently these peptides are separated by hydrophobicity on an HPLC and measured as mass over charge ratio using a mass spectrometer. If necessary, these peptides may be fragmented even further to identify the exact composition and order of the amino acids. By *in silico* analysis of the organism's theoretical proteome, these peptides can be associated with their originating proteins. The intensities determined for each peptide can then be aggregated to a global abundance estimate for each protein (Yates et al. 2009). It should be noted that the intensities across all proteins do not necessarily correlate with an exact abundance measure. This is due to differences in peptide observability leading to peptides with equal abundance showing differing intensities (Zimmer et al. 2018). While there are possibilities to quantify proteins in an absolute manner, e.g. using QconCATs, in comparative studies quantification is often performed relatively (Hammel et al. 2018). If changes in the same protein between different conditions are of interest, the absolute protein amount does not add further information since the intensity ratios stay constant. An extension of the proteomics workflow is called complexome profiling and deals with the analysis of protein complexes. While the peptide preparation and detection stay the same, an additional gel-based separation step enables the identification of complexes. Therefore, protein complexes are isolated and ran in a *blue native PAGE* before the individual lanes are cut into multiple slices. These slices contain native proteins in their original complex configuration and are separated by complex size. The mass-spectrometry based quantification of proteins allows for the identification of profiles that are separated by molecular weight or complex size complexes (Heide et al. 2012; Spaniol et al. 2022). Thereby ‘mer stages and the formation of protein complexes can be studied. A third major omics class deals with the measurement of metabolites. These intermediates or products of biochemical pathways play characteristic roles in cell physiology, development, and pathology and hence define the cell’s phenotype (Figure 3C). The metabolome resembles the complete set of all metabolites formed by the cell in association with its metabolism (Villas-Bôas et al. 2007). Unlike the direct connection that links genes to transcripts and ultimately to proteins, metabolites can have multiple sources, either by direct import or conversion out of precursors. Their quick turnover renders the prediction of metabolite levels out of prevalent transcript or protein abundances difficult. A high diversity in chemical composition which requires special handling and multiple spectrometry based analytical techniques to quantify the metabolome of a cell or tissue (Figure 3F).



Transcript ID	count
Cre04.g000120	50
Cre04.g000120	3
Cre04.g000120	1,213
...	...
~20.000	...

Protein	intensity
HSP70	510
atpA	2,870,600
RPL6	104,310
....	...
~6.000	...

Metabolite	intensity
Glucose	730
Fructose	2,230
Citrate	73
...	...
~1.000	....

Figure 3 Different system levels within a photosynthetic cell with compartments schematically reduced to a nucleus (grey), chloroplast (green), and mitochondrion (orange). (A) Genes are transcribed from the genome by an RNA polymerase. The resulting transcripts (pre-mRNA) are processed (e.g. capped, spliced, poly-A modified) and translocated into the cytosol. (D) The cellular transcript pool can be isolated, enriched for desired properties (poly-A purification or fractionation of compartments), and quantified by RNA-Sequencing. (B) Ribosomal complexes of proteins and rRNA form the ribosome, which translates the fully matured mRNA to sequences of amino acids, provided by tRNAs (not shown). Subsequent folding and post translational modification results in functional proteins. (E) The cellular or compartment specific protein pool can be isolated and quantified by MS-proteomics. (C) Metabolites are chemically diverse molecules that are produced or used during metabolism. Proteins that convert or produce metabolites are called enzymes and often are organized in metabolic pathways. (F) Due to their fast turnover and chemically diverse structure, the isolation and quantification of metabolites must be conducted with special care and do not follow a uniform protocol. Different metabolite kinds require different measuring devices.

When measuring molecules within cells, the choice and preparation of the sample material is crucial. As described earlier, different cell types can be characterized by differences in their ‘omes, so care must be taken to ensure that the samples being compared are of the same composition. When dealing with unicellular organisms, e.g. the green algae *Chlamydomonas reinhardtii*, no tissue specialization is present. Despite compartmentalization still separates important metabolic reactions, all processes, e.g. replication, motility, energy conversion, storage, and sensing of the environment, are facilitated by a single cell (Coates et al. 2014). While this may seem to complicate the analysis of general systems responses, it has the advantage of not being biased by tissue specific responses and not relying on an error-prone sampling strategy. For tissue specific studies it is crucial to sample comparable parts and amounts of an organism, but for unicellular organisms samples are taken from a uniform cell population. Although differences can also occur despite the same environmental conditions due to temporal synchronization or reaching critical cell densities, it is easier to control these factors and ensure a homogeneous sampling (Carrasco-Pujante et al. 2021). Despite its comfort and benefit of capturing a snapshot of the complete systems level, all presented omics techniques require a high level of skill, specialized equipment and are still cost and labor intensive. In practice, this commonly results in a certain sparsity within the datasets and renders data modeling even more necessary.

Omics driven time series measurements

Modern omics approaches allow to capture the *status quo* of the respective systems level experimentally. Based on the knowledge available from databases (e.g. ENA, PRIDE, GEO, MetaboLights) predictions of the cell's function and prevalent pathological conditions are possible if empirical data for the experimental circumstances exist (Chen and Zhou 2019). Using at least two samples allows for comparative analysis of the same tissue at differing conditions. By analyzing differential expression transcript and protein sets can be identified that potentially are important for the observed phenotype. It enables us to infer a variety of biological phenomena, such as protein function, involved signaling pathways, underlying pathophysiology, and the identification of biomarkers (Kline et al. 2022; Wang et al. 2014). Analyzing multiple samples taken over time allows additional analysis techniques to monitor acclimation responses and infer regulation properties and kinetics. These longitudinal studies are especially important when dealing with dynamic processes (Bar-Joseph et al. 2012; Desai et al. 2011). Therefore, samples are drawn from the same population in regular or irregular time intervals with subsequent isolation and detection of transcripts or proteins. If specific research questions have to be answered, the analysis of molecule subsets, e.g. by western blot, dot blot, northern blot or other biochemical methods may be sufficient, but if whole systems level are of interest or systems biology techniques should be applied, the beforementioned high throughput methods can be used to gain comprehensive insights into molecule kinetics. By concatenating the snapshots of biological statuses, time series analysis techniques can be exploited to gain insights into metabolic fluxes, regulatory processes, or signaling pathways.

Time point spacing

Biological systems encompass a myriad of processes occurring across various spatial dimensions and vastly disparate time scales. However, in many settings the time point spacing is fixed, so that samples are taken at regular intervals. This sampling pattern is used when studying cyclic processes or when the rate of change cannot be estimated in advance (Bar-Joseph et al. 2012; Spellman et al. 1998; Menges et al. 2002). This approach is particularly useful and cost effective when consistent changes are expected with no anticipated regulatory fluctuations, as is the case for cell cycle processes and circadian rhythms. Deviations from this uniform sampling scheme are particularly necessary when analyzing developmental processes or when samples are subjected to sudden perturbations. Cellular or organic differentiation is mainly driven by cytokines and hormones that trigger signaling cascades and a resulting reorganization of cellular tasks and processes. These differentiation reactions follow a strict temporal schedule in which important developmental stages are reached at different time intervals. Morphological markers that indicate the current development status can serve as a guide for choosing appropriate sampling time points (Mathavan et al. 2005; Gerstein et al. 2010; Bar-Joseph et al. 2012). Furthermore, empirical data shows that perturbations introduced by abiotic or biotic treatments trigger an acclimation response that predominantly happens directly after treatment onset and weakens over time (Gaucher et al. 2008; Mendoza-Parra et al. 2011). This is intuitive from a control system engineering perspective because when the change in conditions is beneficial, cells aim to take advantage of the new favorable conditions. When faced with detrimental influences, however, cells shift their focus to adjust their transcriptome, proteome, and metabolome to cope with the changes and prepare for the direct and indirect consequences of the treatment. If this fast initial regulation burst with a subsequent decreasing regulation rate is expected during an experiment, it is recommended to select sampling time points accordingly. A trivial solution is to capture homogeneous amplitude changes instead of homogeneous time spacings. Starting with the shortest sampling interval and doubling it with each iteration (e.g. 0 h, 1 h, 2 h, 4 h, 8 h, etc.) corresponds to an exponential regulatory pattern. However, the considered system level has to be kept in mind when selecting sampling times. While the metabolome may change within seconds after treatment, the degradation or synthesis and preparation of transcripts require energy investments and signaling pathways that activate transcription factors. It has been found that only a few minutes after treatment onset the first significant transcriptome changes can be observed (Ashburner and Bonner 1979; Lindquist 1986). As proteins are synthesized from transcripts, their response time is assumed to lag behind, at least for proteins that are synthesized *de novo*. Of course, besides transcription regulation, there are several alternative ways to influence protein activity. For example, by translational regulation biological systems can directly influence the protein synthesis processes while the transcript count remains the same. In general, it should be noted, that transcript counts and protein abundance are just proxies for transcriptional and protein activity (Furlan et al. 2021). Several regulation mechanisms exist, that alter the translational activity as well as protein activity (Figure 4). Just to mention a few, translation can be regulated by affecting the transcript stability, phosphorylation of initiation factors, modification of ribosome binding sites, or the formation of transcript structures that attract regulatory elements (Gebauer and Hentze 2004). An important acclimation reaction is the induction of aberrant expression of miRNAs that reduces the available mRNA pool (Ferrando et al. 2017). The activity of proteins can be regulated by conformation changes due to the binding of small molecules, in-/activation by post translational modification, protein-protein interactions, or spatial separation from the place of action (Cooper 2019). However, as most translational regulation relies on the modification of the available transcript pool, transcript counts and protein intensities are in most cases an appropriate estimator for increased or decreased activity of biological processes (Csárdi et al. 2015; Vogel and Marcotte 2012). This raises the question of whether there are regulatory techniques that are specific to certain stimuli. Depending on the source, the correlation between transcripts and corresponding proteins is between 0.6 and 0.9 (Csárdi et al. 2015). In this scenario, time series experiments help to shed light on translational regulation whether it is condition specific or even changes over time.

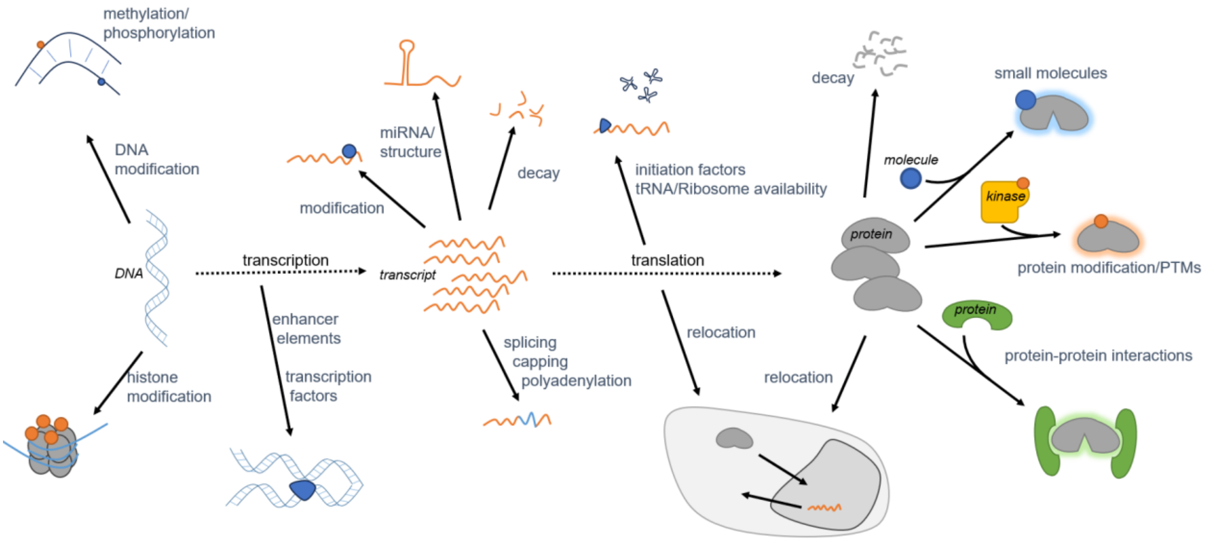


Figure 4 Collection of regulatory mechanisms from DNA to protein function. Some of the existing regulation mechanisms are listed, starting from chromatin modification (left) to protein activity tuning (right). Histone or DNA modification can alter the accessibility of underlying genes. Transcription is influenced by the presence of transcription factors and enhancer elements, that improve DNA accessibility and can bend the DNA to bring regulatory complexes in near proximity. The transcript (pre mRNA) can be inhibited by the formation of loops or the binding of complementary micro RNA. The degree of polyadenylation influences the mRNA's decay rate while alternative splicing can lead to the synthesis of proteins with differing functions. The translation of matured mRNA to proteins occurs in the cytosol or ER membrane associated and depends on generic and transcript specific initiation factors. Protein activity itself requires the protein to be correctly folded and located as well as post translational modifications, regulatory co-factors, or the formation of protein complexes.

Challenges of short time series data

When performing time series experiments the sampling time point number often ranges from four to ten. Even for experiments lasting several days with multiple condition changes, up to this day, sample counts rarely exceed 15 samples. Simultaneously the number of biological replicates remains small and often is lower than five (Kleyman et al. 2017; Garcia-Molina et al. 2020; Dudek et al. 2021; Zaza et al. 2023; Bakker et al. 2023). While challenging from a modeling and analysis perspective, this sparse sampling is due to the still high effort and cost associated with omics technologies and biological experiments in general. The initial organism population, may it be plants in a highly controlled environment, organisms in liquid culture, or animals, must be big enough to last throughout the sampling time course. With every measuring point, the remaining pool is reduced, at least if the organisms are not able to restore the same biomass within a measurement interval. Spatial or volumetric constraints as well as the limitation of analytical and preparational apparatuses limit the technically feasible number of samples. Furthermore, the sampling and sample preparation process often requires a high level of skill and takes minutes to hours of focused work which defines the lowest possible time interval between samplings. Nested teamwork is suboptimal as samples have to be highly consistent and laboratory devices may cause limited capacity. From an analytical perspective, several demands must be addressed.

Table 1 Analytical challenges of time series analysis

i	The identification of relevantly different from unaltered time series: This can employ a single molecule’s time course or the comparison of different molecules over time.
ii	The identification of shared responses by detection of similar kinetics.
iii	The condensation of obtained information to a level that is interpretable for researchers.
iv	The consideration and incorporation of biological variability as well as technical noise into the data analysis approach.
v	The modeling of an appropriate representation of the data received.

With improvements in automated sampling methodologies and decreasing costs of high-throughput technologies, the application of time series experiments in combination with omics technologies is becoming increasingly popular. They are useful to monitor, compare, and predict the transcript, protein, and metabolic fluxes and interpret acclimation reactions of biological systems. Due to the size and complexity of high-throughput data, dedicated analytic techniques are required to interpret the measured data.

Comparative analysis

The identification of relevant changes (Table 1i) can be realized with comparative approaches. Many study designs either focus on molecule abundances that are compared within a single sample, or a treated sample is compared against a control sample. A single molecule time series can be seen as a progressive shift from control to treatment (and eventually back). For any protein, transcript, or other molecule of interest, a statistical test can be applied to assess whether it underwent any statistically significant change. A commonly applied statistical test to check for a global change is an ANalysis Of VAriance (ANOVA, or more specifically one way ANOVA). It takes all replicates of all samples and checks if at any point there is a significant deviation of the sample means. An extension, called (multivariate) repeated measures ANOVA, incorporates the information that the samples are dependent on each other. If the abundance at time point 2 is higher than at time point 1, it is likely that time point 3 is higher as well. However, while often applied in behavioral science, in long lasting time series experiments this assumption not always holds true and the increased sensitivity to missing values or imputed data often leads to the decision for default ANOVA (Keselman et al. 2001; Park et al. 2009; Brillinger). Using post hoc tests, time points can be identified that differ from the overall mean. However, as many models and post hoc tests require the replicate variance at the individual time points to be equal or at least homoscedastic, this procedure is suited to just get a crude overview of the overall systems response and is not necessarily suited for in depth analysis of individual molecules. While non-parametric versions exist, the availability of just 3-4 biological replicates greatly reduces the test power and increases Type I errors (false positives). In general, the statistical test used should be chosen carefully, taking theoretical assumptions as well as the test's prerequisites into account. Tests to consider are Dunnett's test, variants of repeated measures ANOVA, or mixed-effect models in general (Dunnett 1955; Park et al. 2009; Laird and Ware 1982; Wood 2013).

Clustering

Shared responses can be identified by clustering or network-based techniques (Table 1ii), that can be applied to various data structures, including time series, often without the need for statistical tests. Clustering, in general, aims to group similar objects and form coherent groups that are similar within, but different between clusters. Popular algorithms include *k*-means clustering, hierarchical clustering, and density-based clustering (Warren Liao 2005). While not going into detail, all clustering techniques have in common that there is the need for a distance measure that describes the distance or (dis)similarity between two elements.

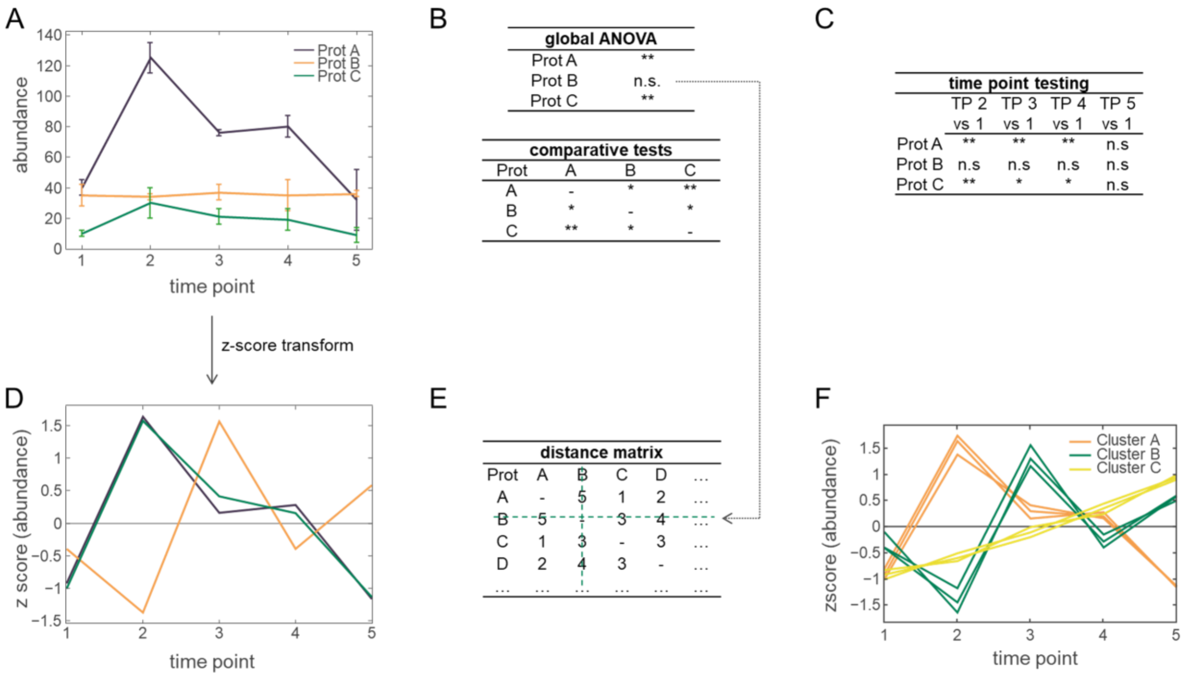
Distance measures and standardization

Most commonly the squared Euclidean distance is used for clustering approaches. It summarizes all squared distances between two elements at corresponding time points. In contrast to Euclidean distance, which contains an additional square root calculation, the squared variant behaves proportionally while being computationally more efficient.

$$E_{sq} = \sum_{t=1}^{n} (a_t - b_t)^2 \text{ (Equation 3)}$$

with a,b being time series signals of two molecules and n being the number of time points.

Other distance measures that are commonly used in clustering algorithms are Manhattan distance, Dynamic Time Warping, or even correlation-based measures such as Pearson's correlation coefficient. For clustering approaches, it is especially important to prepare the data appropriately. If two proteins are regulated the same way and show identical rates throughout the time course, they should be assigned a similarity of 1 or a distance of 0. Abundance discrepancies between both proteins however lead to high distances because the absolute deviation is measured rather than a relative one (Figure 5A). This holds true for most commonly used distance measures and clustering algorithms. While correlation measures are insensitive to intercept changes, they instead are outlier sensitive and by their nature assign higher weights to low or high values. To correct for differences in abundance, the signals can be standardized to lie within the same amplitude range (Figure 5D).



global ANOVA	
Prot A	**
Prot B	n.s.
Prot C	**

comparative tests

Prot	A	B	C
A	-	*	**
B	*	-	*
C	**	*	-

time point testing

	TP 2 vs 1	TP 3 vs 1	TP 4 vs 1	TP 5 vs 1
Prot A	**	**	**	n.s
Prot B	n.s	n.s	n.s	n.s
Prot C	**	*	*	n.s

distance matrix

Prot	A	B	C	D	…
A	-	5	1	2	…
B	5	-	3	4	…
C	1	3	-	3	…
D	2	4	3	-	…
…	…	…	…	…	…

Figure 5 Testing and clustering of time series data. (A) Normalized protein abundance data of a time series experiments. (B) Statistical testing of protein data reveals differences within and between protein time courses (n.s. not significant, * significant, ** highly significant). Note: Multiple testing corrections must be applied when multiple tests are conducted. (C) Using e.g. post hoc tests, each protein time course can be analyzed in detail. Here, time points 2-5 are tested against time point 1 within each respective protein. (D) The normalized protein data signals are standardized via z-score transformation to have zero mean and unit variance. The variance information is lost during this process. While proteins A and C showed high abundance differences, a z-score transformation yields almost identical signals, potentially indicating a shared regulator or similar functions. (E) From standardized signals distances (e.g. Euclidean distance) can be determined between each pair of proteins and written into a distance matrix. For high similarity, the distance measure must be minimal while similarity measures as e.g. Pearson correlation coefficient must be maximized. As protein B was identified to show mean differences during its time course, it can be excluded from downstream clustering analysis. (F) Based on the distance matrix clustering approaches can partition the data into groups of coherent signal behavior. These clusters subsequently can be characterized by label enrichments.

This preprocessing step is required for most clustering procedures as well as other distance-based approaches. While in many clustering approaches the features have to be standardized (data columns or here time points), for high throughput data a row wise standardization is applied (Tavazoie et al. 1999). A common technique called z transform, z-score transform, or just standard score, standardizes each time series signal to be centered at zero with a variance of one. The resulting signals are often referred to have zero mean and unit variance (Smet et al. 2002). When handling high throughput data, the amplitudes can span several orders of magnitude, interfering with the distance measures. Additionally, the transcript count or protein abundance is just a proxy for the underlying real measures and hence two equally abundant molecules may show different measurement values. However, these are not the only reasons why row-wise standardization is necessary.

Molecule activity status

Biological molecules do not have to be present to the same extent to have the same activity or a similar influence on processes. In signaling cascades, for example, proteins that are at the beginning of the pathway can have a much greater influence on the cell, even though they are much less abundant in their sheer number compared to players involved downstream of the cascade. Another example of this are transcription factors, which in relation to their abundance can be significantly more influential than highly abundant enzymes. In general, protein efficiency must always be considered when assessing their activity. While RubisCO, the most widespread carboxylating enzyme in autotrophic organisms, can account for up to 10% of a cell's abundance, its turnover is limited to a few reactions per second (Bathellier et al. 2018). Catalase, on the other hand, can carry out tens of thousands of reactions per second (Singh et al. 2008). These examples illustrate that in many cases it is not the absolute quantity (amplitude) that is of interest, but the change in it. Relative changes help us to investigate cellular regulation and the importance of individual molecules for certain processes. If co-regulation is of interest, candidates with similar rates of change (slopes) relative to their abundance could be identified. By performing a z-score transform, the amplitude information is lost but the relative changes are aligned. Distance calculations based on standardized signals allow these potentially co-regulated molecules to be identified. However, care must be taken with signals that do not differ over time! If a signal is theoretically constant and only affected by white noise, a z-score transformation would amplify the noise to appear as a real change (Figure 5). Clustering procedures would be corrupted by these originally constant signals, which often appear as oscillating spikes. Not only will these signals be misinterpreted as valid regulatory events, but because clustering procedures rely on grouping similar elements, the distinctiveness of these groups may be diluted. To address this problem, potentially constant signals can be filtered out prior to the transform and thus be excluded from the analysis (Figure 5B). A commonly used method is to use a statistical test to determine whether a signal changes in any way over time. One possibility is to perform an ANOVA, which examines the individual replicates of each time point and indicates without further specification whether the mean value of a time point differs significantly from the overall mean value of the time series signals (Askari 2021; Liu et al. 2005). The selection of a suitable p-value threshold must be made depending on the study design and is not generally valid. As this is only a first rough filter step and no further significances are of interest afterwards, a multiple testing correction can be dispensed with in most cases.

As with any analysis approach, assumptions must be made about the available data. Clustering techniques require some form of parameterization that affects either the clustering itself or the interpretation of the result. For the popular *k*-means clustering, the expected number of clusters must be specified in advance in addition to the distance measure used during the procedure. For known systems, this differentiation into a specific number of expected patterns is possible but proves to be difficult for global analyses of entire biological system levels. In hierarchical clustering, it may also be necessary to dissect the resulting dendrogram at a certain level after the actual clustering process to contain a certain number of resulting clusters. There are numerous strategies for determining a specific number of clusters (Xu et al. 2016; Kodinariya and Makwana 2013). Many methods use dispersion calculations of the resulting clusters to determine the number of clusters to be analyzed by visual methods (elbow method), cross validation (leave-one-out or k-fold cross validation), or comparison with a reference dataset (gap statistics or silhouette index). For dendrograms resulting from hierarchical clustering, the number of clusters is determined after the clustering itself by cutting the dendrogram at a certain height. Extensions of this approach allow the level height to be varied (Langfelder et al. 2008). Ultimately, the methods aim to minimize the mean dispersion within the clusters in relation to the mean dispersion between the clusters (Figure 5F). Experience has shown that typically cluster numbers between four and six are suggested for classical biological high-throughput studies. Density based clustering (DBSCAN) does not depend on the determination of cluster numbers. Instead, each signal is encoded as a single coordinate vector and thus represented as a single point in a multidimensional coordinate system. The number of coordinates corresponds to the number of measurement points. It is a non-exhaustive clustering technique, meaning that signals may not be assigned to a cluster (Ester et al. 1996; Pirim et al. 2012). In addition to identifying outliers, this method can generate an unlimited number of clusters. However, this advantage comes with the burden of determining two initial parameters that are difficult to estimate from the data. In addition to the minimum size of a cluster (minPts), the maximum distance between two points must be specified (ε). This distance is difficult to estimate intuitively in multidimensional spaces and the strong dependence of the two parameters on each other makes individual hyperparameter optimization difficult. OPTICS – as an alternative density-based clustering procedure – isn't sensitive to exact parameter choice, but as for hierarchical clustering makes it necessary to manually inspect the result and choose a specific cluster topology to identify distinct signal groups (Ankerst et al. 1999; Kriegel et al. 2011).

An alternative to clustering-based data partitioning is the application of network approaches. Correlation measures are insensitive to amplitude differences and therefore do not require prior standardization. Using pairwise correlations, co-expression networks can be generated. Without going through details, biological high throughput time series can as well be used to generate networks and afterwards detect communities that share similar behavior to the experimental conditions (Aoki et al. 2007; Rao and Dixon 2019; Luo et al. 2007; Blondel et al. 2008). In contrast to distance-based approaches, correlation measures additionally detect anti-correlation, meaning connections of molecules, that are negatively influenced by each other. In a broader sense, both clustering and network approaches lead to groups of molecules that may be connected in some regulatory way and that can be studied as a coherent element. The choice of partitioning approaches, however, is not trivial and differs in required user input, speed, outlier/tie handling, cluster topology, and the necessity for multi-cluster memberships (Andreopoulos et al. 2009; Jollyta et al. 2023). There is no strict guideline on when to use which approaches, but it is important to keep in mind the underlying partition strategy when interpreting the outcome.

Dealing with partitioned data

If conducted properly, clustering results offer a condensed view of the analyzed data. Considering the vast amount of signals, it is unfeasible to examine them all individually. However, identifying groups with high similarity permits the application of various techniques to extract valuable information representing shared properties among individuals in a group. Co-expressed genes or co-regulated proteins with consistent expression patterns may indicate shared functionality or involvement in the same pathway (Pirim et al. 2012; Ma and Chan 2009). Similarly, a partition of the dataset can be realized using aforementioned statistical analysis.

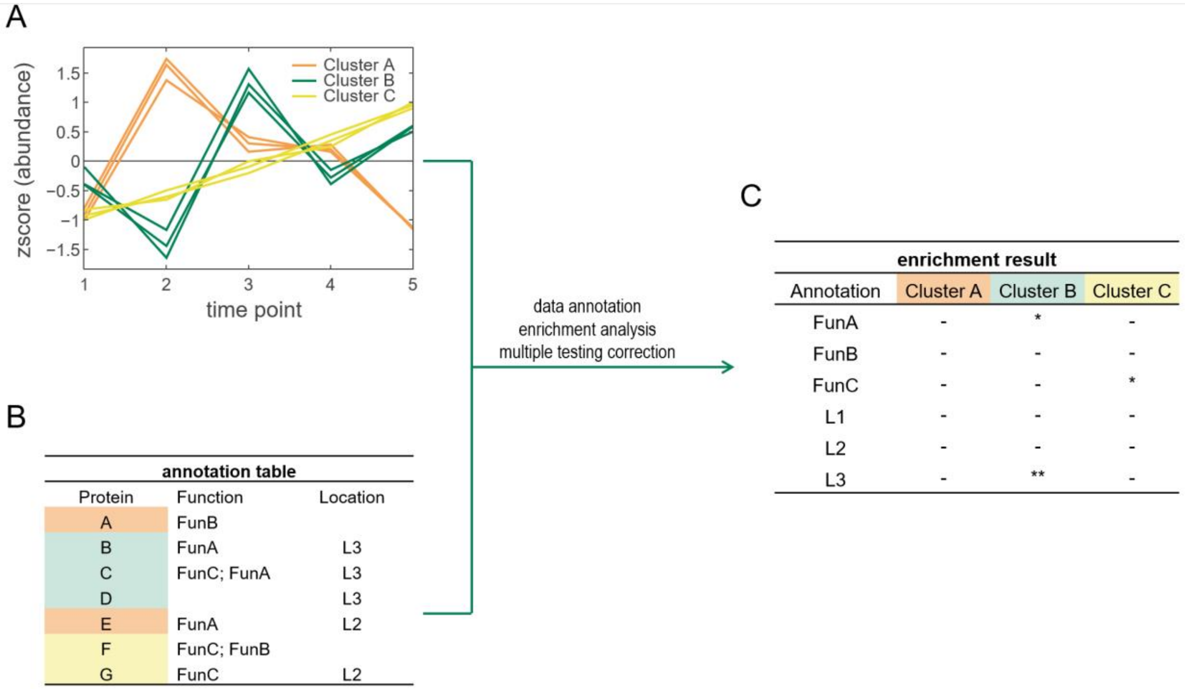
Enrichment analysis

For many biological molecules, additional information is available from experiments or predictions providing information about gene functions or physicochemical properties that can be exploited to characterize a cluster. Elements that show similar behavior and therefore are grouped together can be analyzed by gene set enrichment analysis (GSEA) of their annotation labels (Table 1iii, Figure 6). Most common enrichment methods include the comparison of identified labels within a cluster versus outside of the cluster (Subramanian et al. 2005). If there is a high discrepancy between the expected and observed label distribution, a label can be identified as significantly over- or underrepresented (enriched) for each cluster. The most common strategy to identify overrepresented labels is the application of a one-sided Fisher’s exact test that relies on the hypergeometric distribution. Here, for each label present within a cluster, a probability is reported that describes how likely it is to observe the actual label distribution or more extreme ones by chance (Rivals et al. 2007; Venn and Mühlhaus 2022). Low p-values indicate a high asymmetry between expected and observed label counts and therefore imply the cluster being enriched with molecules with the label in question. After multiple testing correction of the p-value list, the tandem analysis strategy of data-partitioning and label-enrichment enables the researcher to identify global trends and characterize common response kinetics.

Excursus: Data Labeling and Ontologies

The labels themselves can describe various molecule characteristics. These annotations are organized in ontologies where terms of a distinct domain are formally specified along with their relationships. They help to standardize the terminology and provide a framework for organizing and integrating biological information (Noy, McGuiness 2001). These annotation domains range from generic descriptions to organism or even disease specificity (Jackson et al. 2021). One of the most common biological ontologies describes three categories summarized under the *Gene Ontology* (GO) consortium (Ashburner et al. 2000). Here, ontologies exist for (i) *biological processes*, (ii) *molecular functions*, and (iii) *cellular components*. Annotation terms may originate from empirical studies, predicted from computational models, or inferred from paralogs or orthologs based on sequence similarity. *Biological processes* involve some sort of transformation, may it be ‘*cell division*’, a specific metabolic pathway, or signal transduction, i.e. ‘*sterol metabolic process*’ or ‘*abscisic acid-activated signaling pathway*’. *Molecular functions* refer to biochemical activities, i.e. ’*protein kinase activity*’, ‘*protein binding*’, or ‘*glutathione transferase activity*’. *Cellular components* define the place of action of the protein (e.g. ‘*cytoplasm*’, ’*nucleus*’, or ’*chloroplast outer membrane*’) (Ashburner et al. 2000). Besides these generic ontologies that benefit from inferring functions from other species that already are experimentally analyzed, specialized ontologies exist for many model organisms and are tailored to the needs of the research conducted on with these organisms. Besides *Mouse Anatomy Ontology*, *Fungal Phenotype Ontology*, and *Zebrafish Anatomy Ontology*, *MapMan* is a hierarchical ontology that is specialized for photosynthetic organisms such as plants or algae (Thimm et al. 2004; Usadel et al. 2009). A special feature of MapMan is the hierarchy, which remains immediately visible for each term. Unlike ontologies that contain hidden *is_a* or *child_of* fields to represent the relationships, MapMan annotations are exhaustive and contain every higher level descriptor within the label itself. All proteins that have been found to be involved in light harvesting complex II are therefore labeled with

PS.lightreaction.photosystem II.LHC-II. By addition of terms separated by periods, the specificity level is increased. Like most ontologies, they tend to be highly specialized in domains that have been the focus of current or past research. Domains that do not lie within the current research focus are often summarized under superficial terms. Hierarchical ontologies allow such asymmetric research mapping by aiming for a similar level of specificity per annotation level. With the emergence of powerful machine learning models, ontologies can benefit and increasingly adopt robust predicted annotations to complement research areas with sparse coverage.



annotation table

Protein	Function	Location
A	FunB	
B	FunA	L3
C	FunC; FunA	L3
D		L3
E	FunA	L2
F	FunC; FunB	
G	FunC	L2

enrichment result

Annotation	Cluster A	Cluster B	Cluster C
FunA	-	*	-
FunB	-	-	-
FunC	-	-	*
L1	-	-	-
L2	-	-	-
L3	-	**	-

Figure 6 Enrichment analysis. (A) Data is partitioned either by hypothesis testing or by grouping based on similarity (clustering, network communities). (B) Annotation table that assigns generic or organism-specific terms to the proteins. Functions could include e.g. photosynthesis, structural component, cell cycle related. Subcellular locations may be the nucleus, thylakoid, or plasma membrane. (C) By performing a gene set enrichment analysis based on hypergeometric tests, overrepresented terms can be identified by comparing their occurrence within the cluster of interest against its occurrence in the background data. Here *cluster B* is overrepresented with elements located in *L3*.

If clustering and enrichment strategies are applied to time series, it is natural to (i) use significant labels to characterize the cluster progression, or (ii) attribute the average cluster progression to the labels. While it seems reasonable, caution is advised for both statements.

First, even if annotations are significantly enriched within a cluster, it does not necessarily indicate that the cluster consists primarily of elements associated with the respective annotations. An annotation can be significantly enriched while being present with only a few elements. The challenges are illustrated for the following example: In an experiment in which 1000 proteins were measured, 10 proteins were assigned to the term *cell cycle*. A cluster of size 350 which contains just 7 *cell cycle* proteins, already has a p-value of 0.025 and would therefore be significantly enriched. However, proteins of this term make up only 2% of the cluster and therefore only qualify to a limited extent as a characterization of the cluster.

On the other hand, the average cluster time course may not be the exclusive descriptor for all *cell cycle* proteins. There may even be multiple clusters that each report *cell cycle* proteins to be overrepresented. Even if there is a sole enriched cluster, the majority of proteins may not be part of it. Imagine the upper example with population size 1000, cluster size 10, 50 *cell cycle* proteins, and 3 *cell cycle* proteins within the cluster. This results in a p-value of 0.01, despite just 6% of *cell cycle* proteins being within the cluster. Assigning the average cluster time course to the *cell cycle* process could potentially lead to misinterpretation. The method could be improved by dissecting the cellular responses by their temporal progression and identifying the respective important processes.

Both cluster-interpretation examples show that conclusions on both cluster composition and protein group behavior should not be based on significance alone, but always together with considering the respective proportions. Furthermore, regulation of pathways may be facilitated by single activator or repressor proteins that alone are capable of enhancing or reducing metabolite turnovers. While just a few players undergo significant abundance changes complete metabolic cycles or regulatory pathways are affected, which cannot be identified by “majority vote” based enrichment approaches.

Especially when dealing with time series data, the results that are drawn from such analyses are often misinterpreted. As illustration, it would be natural to label a cluster with high amplitudes at the beginning of a time series that decreases just at the very end as *late down* response group. Accordingly, strong increases after treatment onset could be described as an *early up* response group. While in theory, it is possible to make rough categorizations with a given degree of caution, such an approach is prone to misclassification. Cluster progressions of the same clustering often are similar or lack a distinct intuitive categorization due to ambiguous progressions. Additionally, within a single cluster, signals can be highly variable leading to misclassification of enclosed signals and thereby further increasing the classification bias. In general, it is advisable not to mix clustering with classification approaches.

Clustering vs classification

Partitioning signals into clusters based on data point similarities can unveil inherent patterns or structures within datasets, without the need for pre-specified categories or labels. This method differs from classification, which assigns data points to specific categories by extracting features that differentiate between them. Clustering is especially useful when working with complex datasets or when patterns are unclear, as it enables the identification of natural groups based on similarities between data points. Due to its unsupervised nature, clustering may encounter difficulties in situations that necessitate explicit class labels or when the objective is to produce precise predictions for new, unseen instances. In contrast, classification excels in circumstances where distinct categories are available, and the goal is to accurately assign data points to predetermined classes (Figure 7). This makes it appropriate for tasks that prioritize accuracy and well-defined categorization.

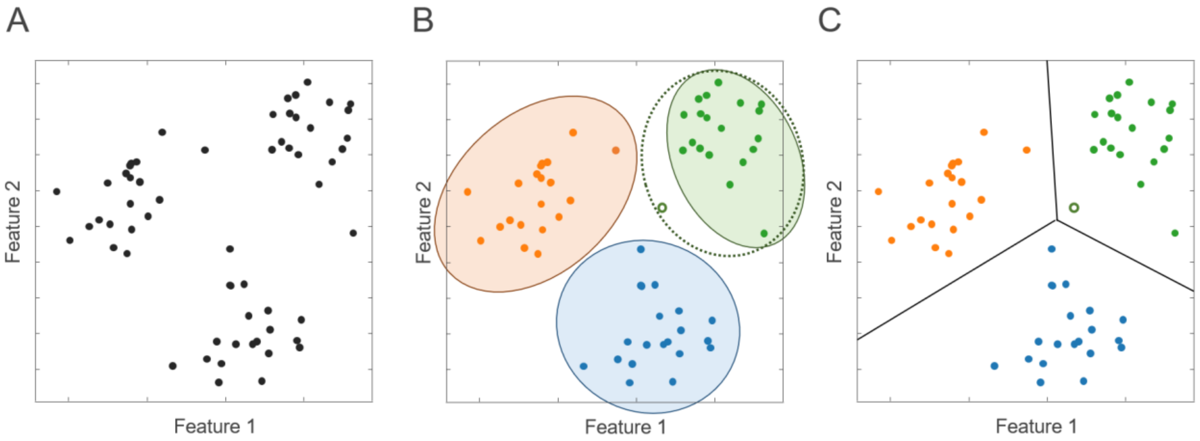


Figure 7 Clustering vs classification. (A) Two-dimensional raw data that shall be partitioned into groups. (B) Clustering based approaches (e.g. *k*-means) compare the distances between all points and partition the data into the desired number of clusters. If new data points occur (central circle), the cluster landscape will change to incorporate the additional point (dashed green cluster). (C) Classification based approaches define feature ranges that strictly separate the data into distinct groups. New points do not change the classification ranges and the new point is assigned with the label of the range it lies in.

The k-means clustering algorithm commonly yields cluster sizes that are similarly large, at least within an order of magnitude. This outcome is a result of the similarity-based grouping approach, wherein every element is averaged into new centroids during each iteration. Small clusters of molecules, which may have significant roles in acclimation responses, are not given preferential treatment and are instead included in the nearest cluster, thereby reducing conciseness. As a result of this averaging-out effect, the cluster mean becomes blurred, and relevant shapes are often covered up (Singh et al. 2011).

The lack of conciseness of the cluster shape, together with the reliance on prior significance filtering and standardization, isn't much of a problem for obtaining impressions of the experiment and a condensed average system response. However, it makes the methodology prone to error when it comes to identifying and characterizing temporal responses based on similarity partitioning. As discussed earlier, there are several reasons why transcript counts or protein abundances may change in similar ways without being related in any way. In particular, for sudden changes in conditions, as often encountered in time series experiments, regulatory, acclimation, and repair processes must occur simultaneously. The limitation to a small number of time points makes it difficult to identify these processes in the necessary detail and challenges the analysis approach to distinguish causal relationships from spurious correlations.

When similarity partitioning is performed on time series data, the ordered sequence of measurements for each molecule is used as input. Since distance measurements generally neglect the readings' order, the time dependence is not considered (Table 1v). This oversight is a major limitation of clustering approaches, especially when dealing with time series data. Traditional clustering methods, such as *k-means*, *hierarchical clustering*, or *density-based clustering* compare each measurement individually and treat them as independent entities. In time series data, however, the order and temporal dependencies between measurements are critical to capturing underlying patterns and dynamics. Neglecting the temporal order of observations can lead to suboptimal clustering results by ignoring important information encoded in the time-dependent relationships between data points (Aghabozorgi et al. 2015). Incorporating the temporal aspect into clustering approaches is essential for creating more robust and precise models. Researchers should carefully consider the temporal characteristics of their data and choose partitioning methods that account for the sequential nature of observations to obtain more meaningful insights and reliable results. Several techniques have been proposed to address this limitation and enhance clustering algorithms to account for the sequential nature of observations. By exchanging the popular *Euclidean distance* measure with *dynamic time warping* (DTW) distance, measurement order is taken into account. DTW is a technique designed to measure the similarity between two sequences while considering possible distortions in the time axis. By aligning time series based on their shapes, DTW allows clustering algorithms to capture similarities in temporal patterns, even if they exhibit variations in speed or phase. Unfortunately, a warp is not meaningful for a rough time series with not more than five measurement time points (Müller 2007).

Signal variation

An important aspect of biological time series is the noise introduced by either biological variability or technical variation (Table 1iv). Biological variability is a fundamental aspect of living systems, reflecting the inherent diversity and complexity of biological entities. When employing high-throughput techniques, understanding and accounting for this variability is critical for obtaining reliable and reproducible results. Biological variability refers to the natural differences observed among biological samples, even when they are derived from the same organism or genetically identical populations. Origins of biological variability are (i) genetic variation or epigenetic modification, (ii) cellular heterogeneity, (iii) environmental factors, or (iv) biological dynamics (Eling et al. 2019; Simpson et al. 2009; Arriaga 2009): (i) Ongoing mutations lead to *single nucleotide polymorphisms* (SNPs) or copy number variations that affect gene expression or protein function. Differences in epigenetic factors, such as DNA methylation or histone modification do not alter the underlying DNA but may regulate the gene expression in different ways. (ii) Biological samples are composed of diverse cell types, each with its own gene expression. Heterogeneity within a sample introduces variability as the contribution of each cell type is mixed. For liquid cultures of unicellular organisms with thorough stirring, this effect is mitigated. (iii) It is difficult to control all environmental factors in an experimental setup. Even the smallest differences in lighting, temperature, ventilation, vibration, or spatially varying cell densities can lead to differential expression of gene sets. (iv) Biological systems are dynamic, exhibiting temporal changes in response to different stimuli. The sampled snapshots of these dynamic processes may be subject to variability due to the inherent fluctuation of biological activities over time. Cells at different stages of their cell cycle naturally differ in their expressed gene set. In addition, random expression noise adds to the phenotypic heterogeneity (Chowdhury et al. 2021). By using cultures that are not synchronized in cell cycle and sampling large numbers of cells, the variability due to biological dynamics can be reduced.

These deviations should not be the target of normalization techniques. Biological variability provides valuable information on uncertainty that should not be overinterpreted. If perfectly controlled cell cultures differ significantly after exposure to a perturbing condition, a high measurement variance in certain transcripts or proteins may indicate a minor role of the respective molecules. Attempting to normalize this variability will impede all subsequent analysis strategies. Understanding and addressing biological variability in high-throughput techniques is essential for the accurate interpretation of experimental results.

Technical variance, on the other hand, is not caused by actual differences in the composition of individual samples. Rather, it arises from undesirable measurement distortion that is introduced by sampling and sample processing, measurement devices, or the application of inappropriate data mapping or normalization techniques (Piehowski et al. 2013; McIntyre et al. 2011): (i) Inhomogeneous sampling of tissue or culture medium can result in technical errors, leading to incorrect representations of the underlying population. Heterogeneous solubilization due to impurities in chemicals or inconsistent sample preparation can also interfere with measuring devices. (ii) High-throughput technology measurement devices are particularly delicate and precise instruments that require careful handling and knowledge of possible disruptive factors. Parameter settings, vibrations, humidity and temperature changes, or component wear during the measurement runs may impact both the accuracy and detectability of the analytes. (iii) Data generated by measurement of the samples often needs special attention before a comparison of molecules from different samples is possible. The most common deviation is due to inconsistencies of material amount in each sample. To account for these homogeneous deviations, linear correction techniques can be applied that overall aim to align expression profiles to an underlying average. Common techniques for normalization include aligning measurement sums and using the median of ratios (Love et al. 2014). Heterogeneous distortions may occur if there is disproportionate sensitivity for certain areas or if heteroscedasticity is present. This could require the utilization of nonlinear normalization methods that adjust their normalization intensity based on measurement intensity or other measurement features. Commonly used techniques include *variance stabilizing transformation* (VST) or *regularized logarithm* (rLog). In comparison to biological variability, technical variation between samples and their replicates needs to be addressed in the early steps of data analysis. Unfortunately, only in the rarest of cases can a distinction be made between biological and technical variance, so that a normalization step represents a compromise that deals with both uncertainties. Researchers need to acknowledge the multifaceted origins of variability and adopt strategies that not only minimize technical artifacts, but also embrace the inherent diversity of biological systems. By doing so, the reliability and robustness of high-throughput analyses can be significantly improved, paving the way for more accurate biological insights and discoveries (Sloutsky et al. 2013).

Signal fitting

Many of the analysis strategies presented focus solely on individual time points, without taking time into consideration (Table 1v). By considering the sequential order of time series, additional information can be incorporated to refine the modeling of the system response. Although connecting sequential data points in two dimensions may appear simple at first glance, it presents a wide range of potential techniques and approaches. When discussing fitting of biological signals, it involves defining a curve that closely approximates the trajectory transcript counts or protein amplitudes (y axis) over multiple time points (x axis). Several assumptions can be made to model the path, each of which has the potential to change its shape slightly or on a global level (Maeland 1988). After introducing models in general on page 7, the following section explores a range of fitting techniques for biological time series. The nature of the highly diverse and dynamic system does not allow for a universal mathematical model to be applied to each of these signals. Multifaceted regulation, diverse modes of action, and involvement in a variety of cellular processes contribute to a signal range that spans from constant to oscillating, from responsive to slow, and from subtle to dramatic. Appropriate models are necessary to capture all of these signal properties. The most intuitive approach is to just connect the data points of each measurement time point. This fitting approach is called *linear spline* and is in accordance with the recommendation to use as few assumptions as possible. But it lacks in fulfilling the second part of Occam's razor principle: "as many as necessary". While being easy, several problems become apparent when dots are connected. From a regulatory perspective it of course is unreasonable to assume the curve is straight between measurements. Additionally, it is impossible for the curves' slope to change instantaneously without any sort of curvature or smooth transition. Fitting a curve to a signal aims to enable amplitude predictions within intervals. However, utilizing linear splines offers only marginal benefits, as the predictive capabilities merely offer a vague idea of what occurs between measurements.

Replicate aggregation

Speaking about the connection of dots, most biological experiments rely on replication to either be able to perform statistical tests, or to assess present variability. Valid mathematical curves allow only one amplitude value per time point, so replicate measurements must be aggregated into a single reading. For many biological scenarios, features are expected to be normally distributed, hence, the most common and obvious aggregation strategy is to take the arithmetic mean of all replicates taken as a point estimate for each respective time point. If sufficient replicate counts are present, outlier insensitive estimates can be determined by geometric mean or trimmed mean. However, if feature distributions are not normal but skewed, then alternative estimation procedures are required. Count distributions from transcript counts typically follow a negative binomial distribution or lognormal distribution that includes no negative values and is highly right skewed. To account for this, the geometric mean can be used to estimate an average expression (Williams et al. 2014; Booeshaghi and Pachter 2021; Love et al. 2014). Equivalently you can apply a log transform to the count data and treat the transformed data as you would with normally distributed data. Therefore, count data often are represented with a logged y axis.

Modeling the time

Special attention should be given to the time axis. As discussed in section *Time point spacing*, researchers frequently choose the measurement intervals based on anticipated changes in rate. This practice can result in irregular time point intervals, requiring additional precautions when fitting the data. In cases where homogeneous changes are expected or the time points were chosen to match the expected response kinetics, it is recommended to have uniformly spaced x values (e.g., corresponding to the time point index, 0, 1, 2, 3 …) instead of treating the x values according to their original sampling time. While this has no impact on prediction results for linear splines, it should be considered for more intricate fitting procedures.

Curve fitting: Interpolating techniques

Interpolation deals with the construction of a curve, that passes through all available data points and thereby approximates the underlying function (Figure 8). Linear splines connect all data points using straight lines (Figure 8A). Sudden changes in slope are inconvenient for modeling biological time series. A curve is considered smooth if its derivatives are continuous throughout the time course. C^1 smoothness ensures the curve itself and its first derivative (slope) are continuous, while C^2 smoothness additionally constrains the second derivative to be continuous. Having a continuous curvature (C^2) is usually considered a condition that should apply to a smooth function as expected in modeled time series. There are computationally efficient ways to determine coefficients for interpolating polynomials (Figure 8B). A system of linear equations must be solved to determine the polynomial coefficients. As interpolating polynomials require the same number of coefficients as there are data points, constructing and solving the equation system is straightforward (Bjorck and Pereyra 1970). For time series fitting purposes they are unfeasible because of their high tendency of oscillations, called Runge's phenomenon, which are of course unfounded for the time series in question (Figure 2Figure 8B, D).

This oscillation can be reduced by utilizing *cubic splines* that ensure a smooth C^2 curve and avoid frequent oscillations. Within the framework of splines, knots represent the points at which the individual cubic polynomials "meet". These are typically the x-values of the given data points. The cubic spline function is defined using a distinct cubic polynomial between each pair of knots (Figure 8C). As a result, the number of coefficients needed to construct a cubic spline is considerably greater in comparison to linear splines or polynomials. Nevertheless, determining the coefficients is straightforward and has been studied for decades (Micula and Micula 1998; Dyer and Dyer 2001). For any interval, four coefficients are required to fit a cubic polynomial ($f(t) = at^3 + bt^2 + ct + d$). The computation of coefficients again involves solving a system of equations. For each unknown coefficient, a constraint function is required that gives some hint for the function shape. As a polynomial must be fitted for each interval the following conditions are imposed: (i) Interpolation: The function value at the knots must be identical for adjacent polynomials. This prevents sudden function steps. (ii) Slope continuity: The curve's slope at knots must be identical for adjacent polynomials. This prevents kinks that are unreasonable to occur at the measuring time points. (iii) Curvature continuity: Additionally to the slope, the second derivative has to be identical at the knots. Continuous curvature ensures a smooth transition between the intervals. Slopes and curvatures cannot be constrained for the first and last time points because there are no adjacent intervals and therefore no polynomial. These two missing constraints cannot be defined universally and require further user input. While several options are available, the most common approach is to set the curvature at the first and last knot to zero. By solving this system of equations, we obtain the coefficients for each cubic polynomial, fully defining the cubic spline. Numerical methods such as Gaussian elimination or tridiagonal matrix algorithms are often employed to efficiently solve these systems (Venn et al. 2022). If a cubic spline is constructed in the described way it's called a *natural cubic spline*, often to be found in computer graphics, numerical analysis, engineering, and scientific curve fitting research. Cubic splines offer a versatile and effective way to represent functions with a high degree of smoothness and continuity. Understanding the interplay between knots, cubic polynomials, and coefficients is crucial for successfully implementing cubic splines in various applications. For biological time series cubic splines provide an effective tool to model the time course signal and analyze the underlying kinetics. However, the most relevant drawback of this method is its reliance on interpolation. All interpolating methods have to pass through the measured data points regardless of their initial accuracy. However, the accuracy of transcript and protein measurements is often unsatisfactory for confidently designating the measurements as ground truth, due to the presence of biological and technical variation.

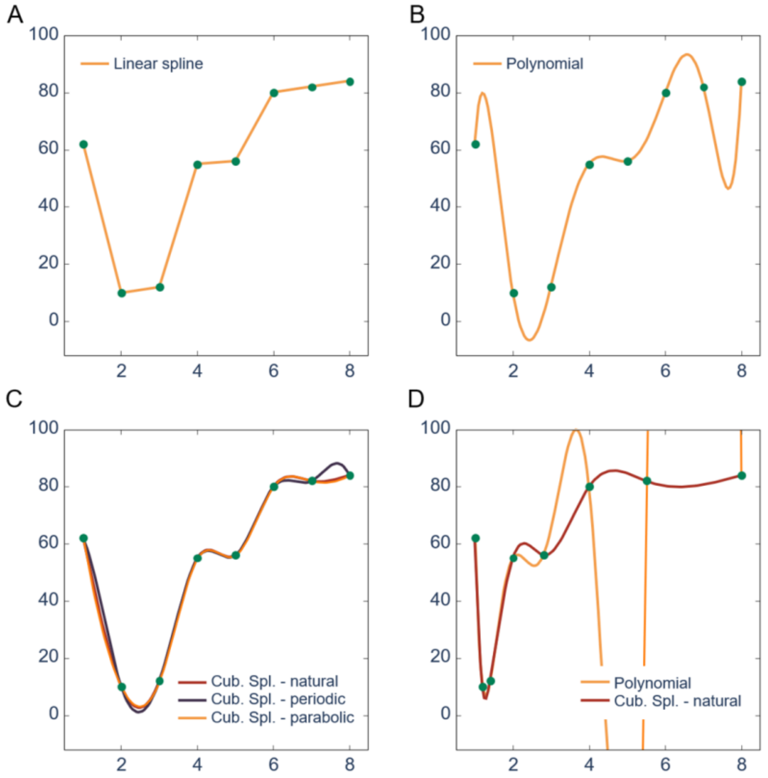


Figure 8 Signal interpolation. As interpolants always pass through all data points variances of individual points have no influence on the models. (A) Uniformly spaced data points are interpolated by connecting straight lines (linear splines). (B) 7th order interpolating polynomial. (C) Cubic splines with boundary conditions: For natural cubic splines the second derivative of the first and last point are set to 0. For periodic cubic splines the slope of the first point is equivalent to the slope of the last point. For parabolic cubic splines, the curvature of the first and second as well as the curvature of the last and penultimate points are equal. (D) The data points are rearranged to a typical experimental spacing and modeled by a polynomial and a natural cubic spline. Note that interpolants in B,C, and D add inappropriate oscillations even when there is a monotone increase.

Curve fitting: Regression techniques

Curves that are fitted to the time series data should not be constrained to match the data points exactly. Because of heavy noise that may be superimposed on the true time-abundance relationship, a curve should (i) be smooth with respect to the first and second derivative, (ii) not be constrained to pass through the measurements, and (iii) be outlier insensitive. The transition from interpolation to regression techniques comes with an intuitive objective change. Rather than constraining the curve to interpolate the data smoothly, in regression the curve should be smooth and minimize the distance of the predicted curve to the actual data points. Again, polynomials play a crucial role in regression. Regression with polynomials is also called linear regression as the relationship is modeled as a linear combination of the input variable. The goal is to find the best-fit polynomial of specified order, that minimizes the sum of squared differences between the observed and predicted values (Figure 9B). In contrast to polynomial interpolation, in regression, the polynomials order is not fixed and has to be defined by the user. This choice of an appropriate order cannot be made based on theoretical considerations, but often requires iterative fitting with varying order and comparing their performance (either by cross validation or coefficient of determination (R^2)). The determination of the polynomial coefficients again involves a system of equations that is constructed by setting all partial derivatives of the error term to zero. Without going into mathematical detail, this ensures the sum of squared differences between the prediction curve and actual measurement values, called residuals, is minimal (Gergonne 1974; Venn et al. 2022).

A convenient advantage of regression techniques in general is the possibility of individual data point weighting which comes in handy to account for variation between sample replicates. Two features of time series data of biological high throughput experiments were already discussed: (1) Description of the time axis with necessary interval spacing considerations and (2) the description of the y axis, be it protein abundances or transcript counts that may need to be transformed and normalized. The third property to consider is the measurement variation at each time point. Biological replicates are taken to assess the certainty of the respective readings and to evaluate their confidence. This variation can be used as additional information to strengthen the model’s validity and robustness against overfitting. By assigning weights to the data points, the impact of these points on the resulting curve can be influenced (Figure 9A). The implementation of the weighting is straightforward and discussed in the publication *Temporal classification of short time series data* (Venn et al. 2024). While the sum of squared residuals is minimized, weights can be added that in- or decrease the respective distance. In many cases, the inverse replicate variance is used as weighting. If the variance - and therefore its uncertainty - is high, its inverse is low, therefore the impact of the distance from the resulting curve to the actual data point is reduced. This leads to the curve not being forced to comply to this point. Wherever the variance is low, the curve should be near to the original data point as the certainty of the estimation is high. A low variance leads to a high inverse and therefore to a higher weighting of the distance, which should be minimized (Strutz 2016).

As mentioned, the choice of the optimal order is not trivial as with increasing order, the function starts to oscillate as it’s the case for polynomial interpolation (Figure 8). Fortunately, as for interpolating cubic splines, the same piecewise technique exists for cubic regression polynomials. *Smoothing splines* combine the smoothness possibilities of cubic splines with the benefits of regression and time point weighting (Figure 9C). They aim to find an optimal compromise of being flexible enough to match the data points, but simultaneously ensuring the necessary stiffness to prevent the curve from overfitting. The construction of these splines is part of this thesis and is discussed in the publication *Temporal classification of short time series data* (Venn et al. 2024).

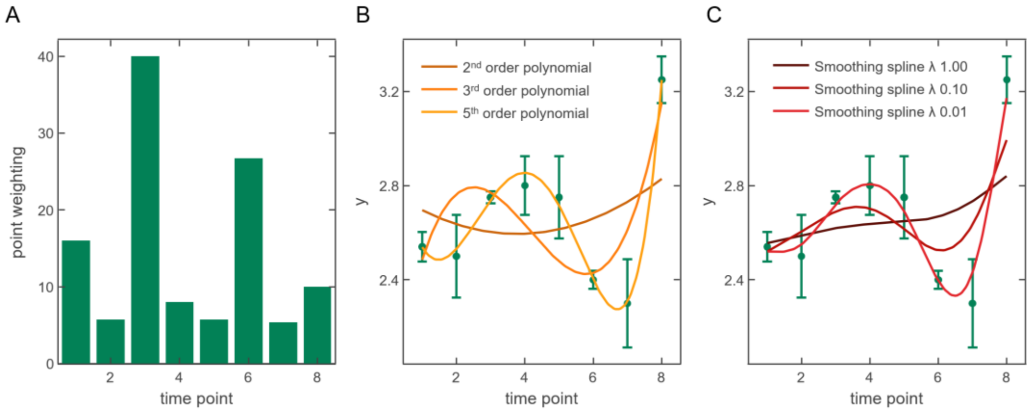


Figure 9 Regression with weighted polynomials and smoothing splines. (A) The weighting of point relevance can be realized by determining the reciprocal of the standard deviation of each time point. Points with high variance are assigned with low weights. (B) If the degree of polynomials is lower than data count - 1, the curve does not necessarily pass through the data points. The lower the degree, the stiffer the curve will become. (C) Smoothing splines consist of piecewise cubic polynomials and balance between fidelity to the data (residuals) and smoothness (curvature). The parameter λ defines the smoothing strength.

Although, or perhaps because, there are still so many hurdles to the analysis of short noisy time series, it is a rapidly developing field that holds great potential for extracting previously unrecognized knowledge from time series. As existing analysis methods evolve and new strategies are developed, the analysis becomes more robust, and the hurdles are overcome. Due to decreasing costs and standardized procedures, more and more time series experiments are being performed, providing more precise measurements at shorter intervals and with a higher number of replicates, to ultimately be able to model all large components of cellular processes and use them to enable simulations and unravel the regulatory secrets. In a broader sense, these methods can also be applied to experimental designs that match key properties with time series. For example, complexomics studies described earlier are suitable candidates. Although there is no temporal component, the individual band snippets are comparably dependent on each other. Technical variation is more common here, as even small differences in gel polymerization and (semi-) manual cutting can cause proteins to shift into adjacent bands.

2. Aims of this thesis

In the field of plant biology, time-series experiments that investigate acclimation reactions are becoming increasingly significant. While unraveling the underlying processes is an academic pursuit, comprehending acclimation reactions is imperative in light of inevitable climate changes. Plants, leading a sessile lifestyle, are already faced with intensifying climatic fluctuations, that demand further scientific inquiry of the molecular responses (Janni et al. 2020; Ortiz-Bobea et al. 2019). Time-series experiments coupled with high-throughput technologies have emerged as the preferred method to shed light on the global intricacies of cellular dynamics. These experiments enable time-resolved observations across multiple system levels, providing crucial insights into cellular regulation strategies. Within the realm of short time series data, new analysis possibilities emerge that use established strategies and adapt them to the special characteristics of short biological time series. Especially the measurement variance of high-throughput techniques combined with a limited number of replicates requires specialized approaches that cannot only handle uncertainties, but actively leverage them to attain more robust results without being prone to overinterpretation.

This thesis aims to apply various data partition and enrichment approaches to time series experiments conducted on the green algae *Chlamydomonas*. Moreover, this work presents novel methodologies that incorporate established approaches and merge them into new strategies. As described in previous chapters, drawing a curve through data points is not a trivial task. A plethora of techniques are available, each with its own benefits and challenges. It’s the responsibility of the researcher to choose appropriate methods for modeling their data. Keeping Occam’s razor in mind, assumptions can and should be made to help identify the best model and not get lost in the sheer number of fitting possibilities. This model choice does not only depend on the kind of data you are dealing with, but also on the question that should be answered by its application. In this work I extended the application of smoothing splines by additionally applying further constraints and assumptions to the curve’s shape.

The main goal is to extend the analytical toolkit by developing approaches that capture the temporal coordination of acclimation responses and enhance enrichment techniques by incorporating a temporal dimension into the analytical framework. Thereby, the comprehension of the temporal orchestration regulating acclimation responses can be studied with a new perspective and support researchers dissecting the huge amount of data into manageable groups of separated cellular responses.

3. Relevant publications for this thesis

This cumulative thesis is based on the following publications:

I. **Temporal classification of short time series data**

   Benedikt Venn, Thomas Leifeld, Ping Zhang, Timo Mühlhaus

   | Bioinformatics, 2024, doi: 10.1186/s12859-024-05636-6

II. **TMEA: A Thermodynamically Motivated Framework for Functional Characterization of Biological Responses to System Acclimation**

   Kevin Schneider*, **Benedikt Venn***, Timo Mühlhaus

   | Entropy, 2020, doi: 10.3390/e22091030

III. **Systems-wide analysis revealed shared and unique responses to moderate and acute high temperatures in the green alga *Chlamydomonas reinhardtii***

   Ningning Zhang, Erin M. Mattoon, Will McHargue, **Benedikt Venn**, David Zimmer, Kresti Pecani, Jooyeon Jeong, Cheyenne M. Anderson, Chen Chen, Jeffrey C. Berry, Ming Xia, Shin-Cheng Tzeng, Eric Becker, Leila Pazouki, Bradley Evans, Fred Cross, Jianlin Cheng, Kirk J. Czymmek, Michael Schroda, Timo Mühlhaus & Ru Zhang

   | Communications Biology, 2022, doi.org/10.1038/s42003-022-03359-z

IV. **Moderate high temperature is beneficial or detrimental depending on carbon availability in the green alga *Chlamydomonas reinhardtii***

   Ningning Zhang*, **Benedikt Venn***, Catherine E Bailey, Ming Xia, Erin M Mattoon, Timo Mühlhaus, Ru Zhang

   | Journal of Experimental Botany, 2023, doi.org/10.1093/jxb/erad405

* These authors contributed equally to this work

Article I: Temporal classification of short time series data

Summary

This publication introduces an innovative methodology for the analysis and interpretation of time series data, particularly those subject to replicate variability. The central aim is to conceptualize short time series as continuous functions, facilitating the extraction of extreme points to characterize molecular regulation dynamics. Preliminary analysis confirms that biological signals typically follow a smooth trajectory. Consequently, the fitting of these signals is performed with adjustable monotonicity constraints, effectively reducing unnecessary oscillations in the curves. This technique allows the generation of diverse shape configurations on individual protein or transcript signals. The incorporation of measurement variance is a pivotal aspect of this methodology, allowing for the fine-tuning of the impact attributed to each temporal data point. The parameter λ, denoting smoothness, plays a crucial role in determining the rigidity of the resultant curve. Under the given point weightings, the optimization of λ is achieved through *modified generalized cross-validation* (mGCV). Given the non-convex nature of the resultant objective function, a systematic grid search is employed to ascertain the most suitable value of λ. Following the establishment of a smoothing spline for each predefined shape configuration, the selection of an optimal model is governed by the minimization of the *Akaike Information Criterion* (AICc). The final model, characterized by its smooth and, where relevant, monotonic curve, allows for the isolation of extremal points via an analysis of the first and second derivatives. These distinct maxima and minima are indicative of significant regulatory events within the observed time series and thus serve a pivotal role in the classification process. In instances where a signal exhibits a monotonically increasing or decreasing trend, the classification is refined based on the characteristics of the signal's curvature, as determined by the second derivative. This refined classification scheme enables a comprehensive grouping of signals in accordance with their regulatory characteristics.

The approach allows the classification of data into distinct temporally resolved regulation classes. Thereby acclimation responses are provided as time resolved response groups, that can either be examined in isolation or grouped into coherent regulatory cohorts (e.g. group of early increasing and late decreasing elements). This nuanced classification facilitates a deeper understanding of the temporal dynamics governing molecular regulation.

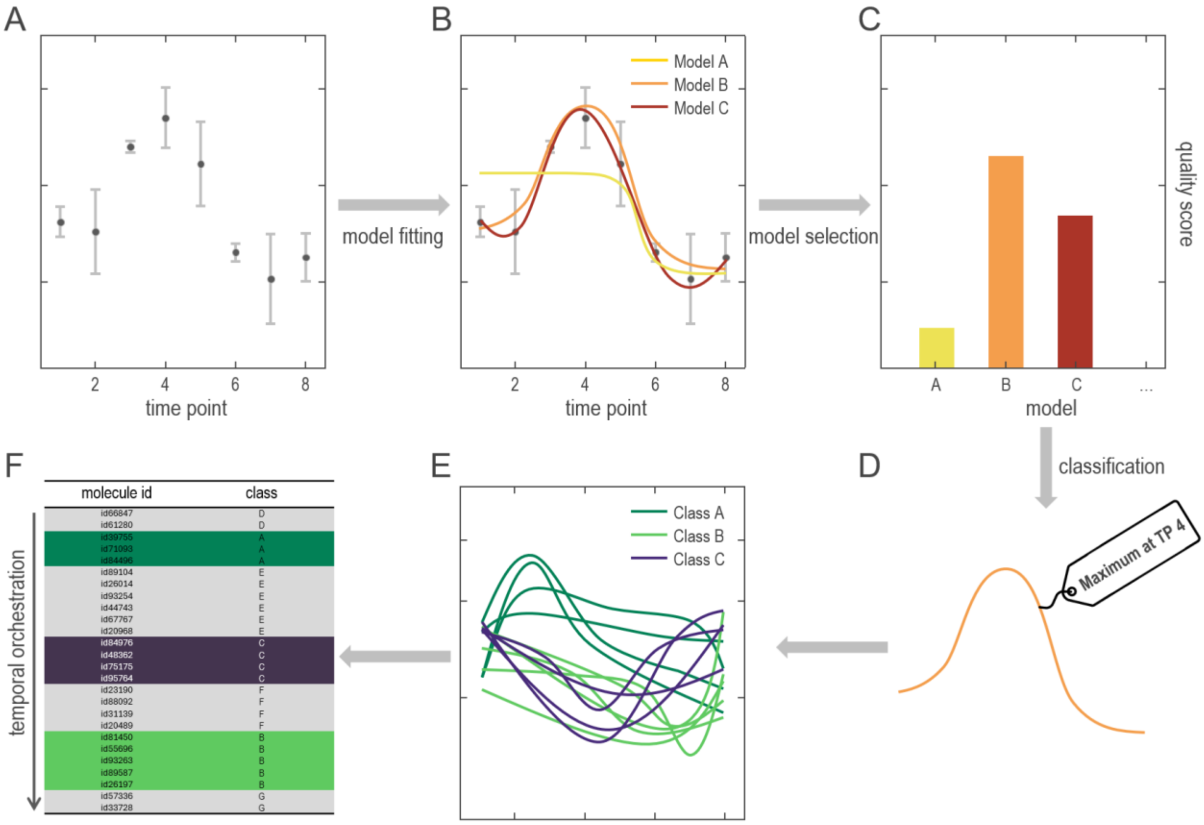
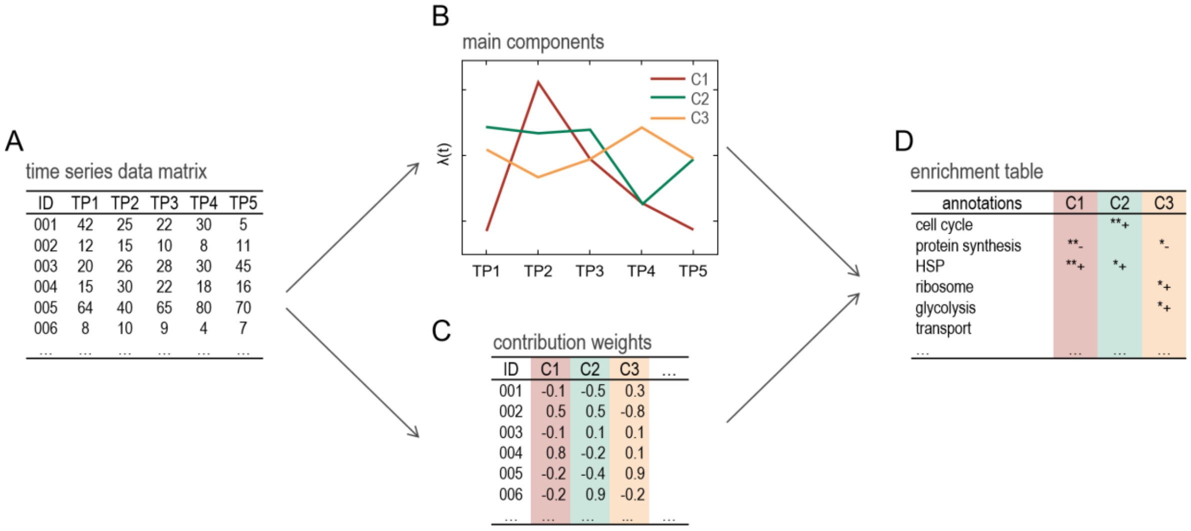


Figure 10 From raw data to temporal annotations. (A) Raw data of a typical time series experiment with eight measured time points that are affected by noise. (B) Several models are fitted onto the data, each with other curve characteristics (yellow: monotonically decreasing; orange: a single maximum; red: minimum followed by a maximum and a minimum). The models incorporate variance information by considering points of high variance less important. Where appropriate models are constrained to be monotone. (C) Based on a quality score that balances the fidelity to the data against the curve smoothness together with the number of extrema the optimal model is chosen. (D) Given the splines coefficients the curve characteristics (slope and curvature) are determined and used for signal labeling. After fitting of all signals measured during the experiment (E), temporal patterns can be analyzed and used for further functional characterization of the system response (F).

Article II: TMEA: A Thermodynamically Motivated Framework for Functional Characterization of Biological Responses to System Acclimation

Summary

The statistical evaluation of time series data, particularly those with a limited number of replicates, poses distinct challenges. Conducting gene set enrichment analysis (GSEA) can be approached in two ways: either for each individual time comparison or for the entire time series. However, this necessitates the establishment of a threshold value capable of distinguishing between significantly altered and unaltered molecules, a non-trivial task with underpowered tests. TMEA (thermodynamically motivated enrichment analysis) was created to enable a time-resolved enrichment analysis that removes the need for a preliminary significance threshold. It dissects the biological system's response into its main components, referred to as constraints. The framework then accurately assesses the significance of individual annotation groups in relation to these constraints. Annotations determined to be significant thus have a noteworthy impact on the respective constraint potential curve and can be characterized accordingly. Moreover, to groups coherently following the constraint potential’s time course, a strong response of an individual element can cause the whole group to become significant. This enables the identification of molecules whose sole regulation is responsible for the activity status of a pathway. A salient feature of TMEA is that each element within the dataset is ascribed a weight relative to the individual constraints, thereby circumventing the requirement for a priori significance testing. This methodology facilitates a more nuanced understanding of the data, enabling researchers to discern patterns and relationships that might otherwise be obscured. Furthermore, TMEA allows for the merging of different experimental conditions, offering insights into the shared and distinct responses elicited under varying conditions.



ID	TP1	TP2	TP3	TP4	TP5
001	42	25	22	30	5
002	12	15	10	8	11
003	20	26	28	30	45
004	15	30	22	18	16
005	64	40	65	80	70
006	8	10	9	4	7
...	...	...	...	...	...

ID	C1	C2	C3	...
001	-0.1	-0.5	0.3	
002	0.5	0.5	-0.8	
003	-0.1	0.1	0.1	
004	0.8	-0.2	0.1	
005	-0.2	-0.4	0.9	
006	-0.2	0.9	-0.2	
...	...	...	...	...

annotations	C1	C2	C3
cell cycle		**+	
protein synthesis	**-		*-
HSP	**+	*+	
ribosome			*+
glycolysis			*+
transport			
...	...	...	...

Figure 11 TMEA workflow. (A) Normalized time series abundances (e.g. transcripts or proteins). (B) Surprisal analysis dissects the dataset into its main components (constraints). Their potentials are depicted as function with respect to time. Here three major temporal responses were identified. (C) Weights are assigned to each molecule that corresponds to the contribution they possess to each constraint. Negative weights indicate high contributions to the inverse constraint. (D) For any annotation term included in the dataset, a permutation test on molecule weight sums is conducted and returns whether a function group has a significant contribution to the constraint time course. Significant negative contributions indicate high a high contribution to the inverse constraint time course.

Article III: Systems-wide analysis revealed shared and unique responses to moderate and acute high temperatures in the green alga *Chlamydomonas reinhardtii*

Summary

The study investigates the 24h acclimation and 48h recovery of Chlamydomonas reinhardtii to 35 °C and 40 °C respectively. The cells were cultivated mixotrophically in photobioreactors. Besides measurements of ploidy, photosynthetic efficiency, pigments, starch, and oxygen, proteomics and transcriptomics measurements were conducted at several time points during the heat acclimation and the recovery phase. It could be observed that cells stopped replicating at 40 °C, but were able to grow at 35 °C after an initial short cell cycle arrest. An active cellular response in both treatments was verified by increased levels of HSP's and HSF levels and by comparison using dimensionality reduction methods. As expected from the observed growth behavior, the transcript and protein analysis revealed an increased regulation in cells undergoing 40 °C heat treatment. A strong discrepancy was observed for transcripts and proteins involved in gluconeogenesis and the glyoxylate cycle. While under 35 °C there was an increase of transcripts and proteins, the same decreased at 40 °C. Interestingly it could be observed that under heat acclimation there is an elevated correlation between transcript count and protein abundance. This indicates a reduced translational regulation and instead a reliance on canonical direct transcription-translation coupling.

TMEA was conducted separately for both conditions and revealed similar constraint potential courses, indicating shared responses during heat acclimation. The most important constraint differentiates the conditions and undergoes a sign switch from acclimation to recovery. Constraint 2 deviates during the whole time course from the control time point and recovers after 8-24 hours of recovery. Through multiple time-point enrichments and a correlation network approach, a significant reorganization was observed in several key biological processes. These included photosynthesis, protein folding, redox reactions, lipid metabolism, and the gluconeogenesis/glyoxylate cycle, highlighting the complex adaptive mechanisms employed by *Chlamydomonas reinhardtii* in response to heat stress.

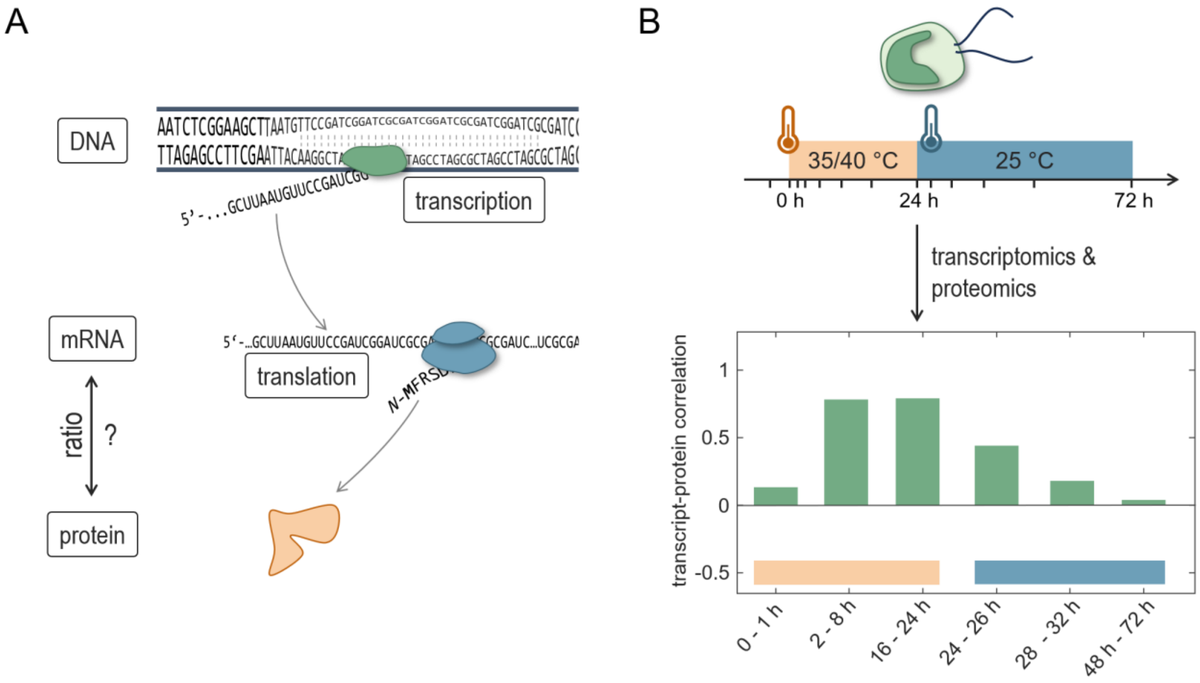
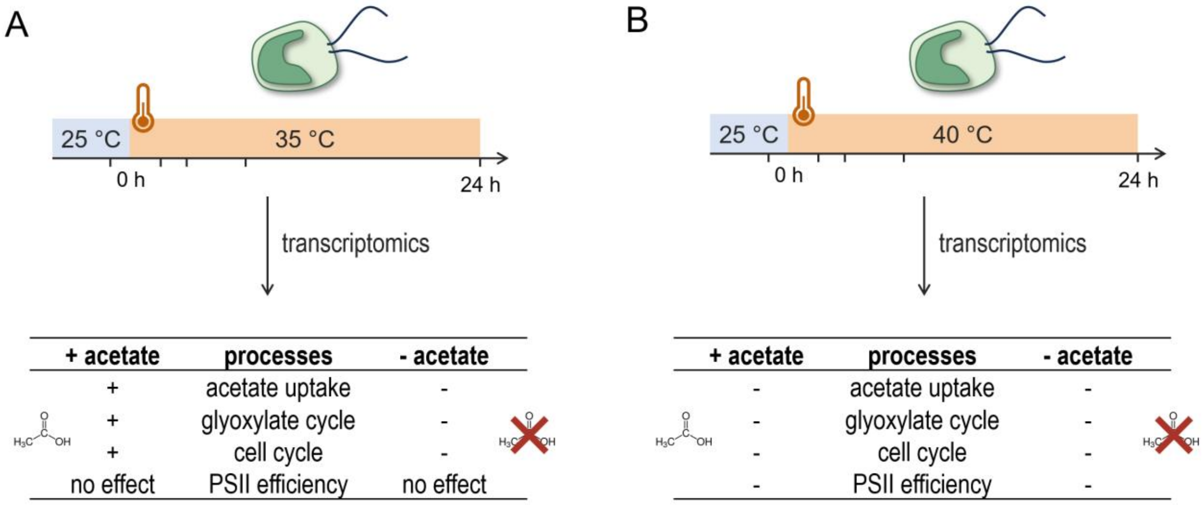


Figure 12 Transcript-protein correlation during heat acclimation and recovery. (A) DNA is transcribed into transcripts by RNA Polymerase II. The matured mRNA is translated into a polypeptide chain by the ribosome. After folding and post translational modification, the synthesis of a functional protein is completed. (B) *Chlamydomonas* cells are treated with 35 °C or 40 °C heat. During early heat acclimation the correlation between transcripts and their corresponding proteins increases drastically. In later phases of acclimation and recovery to 25 °C, the correlation declines to a low value similar to conditions prior to heat onset. This indicates a direct coupling between the transcript and corresponding protein without major translational regulation.

Article IV: Moderate high temperature is beneficial or detrimental depending on carbon availability in the green alga *Chlamydomonas reinhardtii*

Summary

This study deals with the response of *Chlamydomonas* to high temperatures under mixotrophic or semi-photoautotrophic conditions. Therefore, *Chlamydomonas* cells were grown at 25 °C, 35 °C, and 40 °C in media containing acetate as an organic carbon source, and media that was acetate-depleted during the time course. Besides a control, RNASeq samples were taken after 2 h, 4 h, 8 h, and 24 h respectively. Carbon sources play an important role in acclimation behavior to heat. While cells can grow at 35 °C with provided acetate, cells that depend on atmospheric carbon fixation cannot cope with the additional burden and stop growth. Cells faced with 40 °C heat stress could not grow, regardless of carbon supply. RNASeq analysis revealed proteins involved in acetate uptake, acetate metabolism, and carbon concentrating mechanism to be significantly increased at 35 °C whereas at 40 °C no such behavior could be observed. A decreased CO_2 solubilization capacity at elevated temperatures makes these changes necessary to sustain the cell's function. Transcripts involved in plastidic protein synthesis were strongly depleted during both temperatures and both media, suggesting a transient translation stop to reorganize the cellular energy resource management. For 40 °C conditions transcripts related to mitochondrial F1-ATPase were drastically downregulated within the first two hours but reached levels comparable to 25 °C and 35 °C during the rest of the time course. In summary, as seen before, 40 °C heat treatment leads to a cell cycle arrest and leads to cell death after 3-4 days. 35 °C conditions, however, can be tolerated, but its effect strongly depends on the presence of an organic carbon source. With acetate, cells even have an increased PSII efficiency and slightly increased growth rate, whereas acetate diminishing conditions lead to the depletion of cellular carbon reserves due to required carbon investments to acclimate to the increased temperature.



+ acetate	processes	- acetate
+	acetate uptake	-
+	glyoxylate cycle	-
+	cell cycle	-
no effect	PSII efficiency	no effect

+ acetate	processes	- acetate
-	acetate uptake	-
-	glyoxylate cycle	-
-	cell cycle	-
-	PSII efficiency	-

Figure 13 Key effects of acetate supply during heat acclimation. *Chlamydomonas* cells were treated with 35 °C (A) and 40 °C (B) with and without continuous acetate supply. Transcriptomics samples were taken at five individual time points. While 40 °C treatment was detrimental regardless of the energy source, the culture’s response at 35 °C differed strongly depending on the availability of acetate.

Statement of own contribution

Article I: Benedikt Venn, Thomas Leifeld, Ping Zhang, Timo Mühlhaus, *Temporal classification of short time series data*, Bioinformatics, 2024

The project's initial conceptualization was undertaken by other authors, upon which I significantly refined its objectives and strategic approach. Utilizing the programming language F#, I lead the method's implementation. Originally a component of the FSharp.Stats library the project evolved, leading to its establishment as an independent software library, now accessible at github.com/CSBiology/TempClass. I was involved in developing the majority of the methods required, specifically those relating to spline implementations, model selection, extrema extraction, and classification. For the purpose of the publication, I conducted an in-depth analysis of a protein dataset from Zhang et al. 2022. This involved a comparative assessment against various other fitting techniques, most of which were implemented by me and consumed from the FSharp.Stats package. In addition to the technical contributions, I was responsible for preparing the manuscript, including the composition of the text, the creation of figures, the result interpretation, and the collation of supplemental information.

Article II: Kevin Schneider*, Benedikt Venn*, Timo Mühlhaus, *TMEA: A Thermodynamically Motivated Framework for Functional Characterization of Biological Responses to System Acclimation*, Entropy, 2020

For the publication, I annotated the data retrieved from the Gene Expression Omnibus (GEO, specifically dataset GSE125950) using annotation labels derived from the MapMan Ontology and the KEGG Compound Database. For the purpose of GSEA, I performed the necessary hypothesis testing to identify differentially expressed genes. As a benchmark GSEA was performed based on hypergeometric tests with subsequent multiple testing correction via Benjamini-Hochberg. Subsequently, I engaged in a detailed interpretation of the results obtained from both TMEA and the conventional GSEA. This interpretation was not only grounded in the statistical outcomes but was also correlated with relevant biological processes. These processes were, in part, verified by metabolic measurements provided in the original publication of the data source. TMEA was found to be able to dissect a huge number of biological signals in its most relevant underlying response kinetics and connect biological processes to these traces.

Article III: Ningning Zhang, Erin M. Mattoon, Will McHargue, Benedikt Venn, David Zimmer, Kresti Pecani, Jooyeon Jeong, Cheyenne M. Anderson, Chen Chen, Jeffrey C. Berry, Ming Xia, Shin-Cheng Tzeng, Eric Becker, Leila Pazouki, Bradley Evans, Fred Cross, Jianlin Cheng, Kirk J. Czymmek, Michael Schroda, Timo Mühlhaus, Ru Zhang, *Systems-wide analysis revealed shared and unique responses to moderate and acute high temperatures in the green alga Chlamydomonas reinhardtii*, Communications Biology, 2022

After obtaining the imputed and normalized protein abundance matrix, I annotated the dataset and performed all subsequent proteome analyses. After hypothesis testing of the individual time series using Dunnett's multiple comparison test, gene set enrichments were carried out on proteins deemed significantly deregulated, followed by the application of multiple testing correction using the Benjamini-Hochberg FDR. For network generation, I determined a Pearson correlation matrix and an appropriate correlation coefficient threshold by employing random matrix theory. After dissecting the generated network into communities, I determined the eigenvectors of the communities and identified significantly enriched cellular processes associated with these. In addition, I conducted a thermodynamically motivated enrichment analysis (TMEA) and identified contributing gene sets. Together with the normalized transcript data I performed the correlation analysis between gene expression and protein abundance. By separating the time series into several chunks and performing linear regression on each MapMan annotation group it could be analyzed whether protein abundance correlated with observed transcript counts. By this, it could be shown, that the correlation of protein amount and transcript counts increases when Chlamydomonas cells are faced with abiotic stress. Finally, I visualized all signals and summarized them into an interactive spreadsheet containing relevant information to facilitate the scientist's analysis.

Article IV: Ningning Zhang*, Benedikt Venn*, Catherine E Bailey, Ming Xia, Erin M Mattoon, Timo Mühlhaus, Ru Zhang, *Moderate high temperature is beneficial or detrimental depending on carbon availability in the green alga Chlamydomonas reinhardtii*, Journal of Experimental Botany, 2023

After receiving the transcriptomics data, I performed quality control and prepared them for further analysis. This included the imputation of missing samples using additional measurements from a 25 °C time course and correlation-based guidance. The data was normalized and labeled with both MapMan and ChlamyCyc pathway annotations. A PCA ensured the data quality and enabled the verification of induced acclimation responses and constant behavior for 25 °C time courses. Statistical testing was conducted using DeSeq2 in a multifactor design followed by ontology annotations and subsequent multiple testing correction. The results were summarized in an interactive spreadsheet with combined visualization of transcript time courses and statistical results. Transcripts were grouped according to their annotation label, z-score transformed and visualized to interpret the responses of each group to the six conditions.

* These authors contributed equally to this work

Die vorliegende Einschätzung über die erbrachte Eigenleistung deckt sich mit den Angaben aus den jeweiligen Fachzeitschriften und wurde somit mit den an der Publikation beteiligten Ko-Autoren/Ko-Autorinnen einvernehmlich abgestimmt.

_19.01.2024__________________

Datum, Unterschrift Doktorand

_19.01.2024___________________

Datum, Unterschrift Betreuer

4. Discussion

The interplay of transcripts, proteins, and metabolites enables organisms to convert energy sources - be it food or sunlight – into the building blocks needed to sustain life. Nucleic acids, proteins, lipids, carbohydrates, and other metabolites all play a part in the organization of processes within either subcellular compartments, cells, tissues, organisms, or even populations (Shen 2020; Leoncini et al. 2004). As these processes are dynamic, time series experiments are crucial to investigate underlying molecule kinetics and connect them to draw an overarching picture of cellular orchestration. Understanding the dynamics helps to elucidate regulatory mechanisms and coordination, the crosstalk of metabolic pathways, as well as protein function at steady state or during acclimation conditions. Especially when faced with perturbations, a temporally resolved depiction of involved molecules helps in the identification of the regulatory organization of cellular responses. Consequently, time series experiments in combination with ‘omics technologies have emerged as a fundamental approach for studying cellular dynamics. Indeed, there has been a constant increase in annual time series dataset uploads in public databases like GEO for transcriptomics or PRIDE for proteomics (Edgar et al. 2002; Perez-Riverol et al. 2022). Due to declining costs and their high accuracy, high throughput measurements have become popular even when the focus of a study is on the quantification of only a few molecules. These datasets can be made available in specialized repositories for scientists around the world to answer further questions. This leads to a more efficient use of monetary and human resources, as well as increased recognition for experimenters who have shared their data (Barrett et al. 2011; Weil et al. 2023).

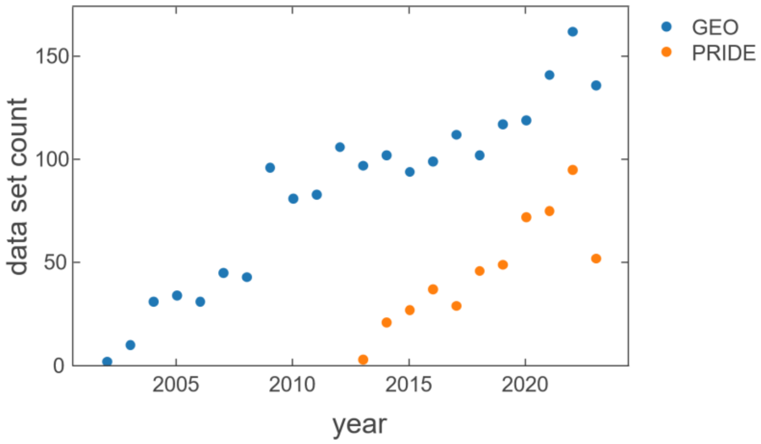


Figure 14 Number of datasets containing “time series” or “time course” for transcript data (GEO) and protein data (PRIDE).

Perspectives on the analysis of time series

Omics time-series experiments are characterized by their ability to simultaneously capture a plethora of molecules. Additionally, the time-resolved sampling allows for tracing the kinetics of these molecules. However, due to biological or technical reasons, the variance at individual time points can be high and should be addressed especially in global analyses in which the influence of individual signals cannot be interpreted afterwards. Given the multitude of measured samples, an appropriate normalization strategy is necessary to mitigate experimental and technical artifacts as effectively as possible. From molecule-centric analytical approaches, which examine the temporal profiles of individual molecules, to global analyses aiming to aggregate all temporal profiles into a few key statements, time-series experiments enable a broad range of analytical strategies that can vary significantly in detail.

Statistical methods enable the examination of individual profiles for changes or the identification of differences between different profiles, automatically considering any occurring variances. In many comparisons, a robust and high-quality normalization strategy is essential, as individual outliers can disrupt the analysis of specific signals. If an outlier specifically occurs at the beginning of the measurements, statistical tests may be flawed if they consider this value as a reference. Statistical analyses at the level of individual molecules are rightfully popular for detailed examination of the response of individual components.

On the other end of the spectrum are global analyses based on data aggregation to summarize the experiment with its key characteristics. Often, this approach involves separating the data into groups whose members exhibit similar temporal profiles. Difficulties in normalization play a subordinate role here, as they equally affect all molecules and thus the effect is at least numerically corrected. However, the consideration of time point variances receives little attention, and replicates are often summarized by determining the mean. This aggregation strategy is solely based on the replicates themselves and without the consideration of the surrounding signal topology. This carries the risk of overinterpreting oscillations caused by outliers. Standardization techniques (e.g., z-score) can promote this bias and distort both the clustering algorithm itself and the resulting groups. Particularly for time series, there exist suitable possibilities to model the temporal progression and thus the measurements at individual time points.

Between these two extremes, the observation of individual molecules in contrast to holistic, summarizing analyses, a wide range of analytical possibilities opens up, in which both perspectives can be balanced (Figure 14). Depending on the chosen aggregation degree, interactions between molecules or entire system levels can be examined to attribute commonalities and differences to shared effectors.

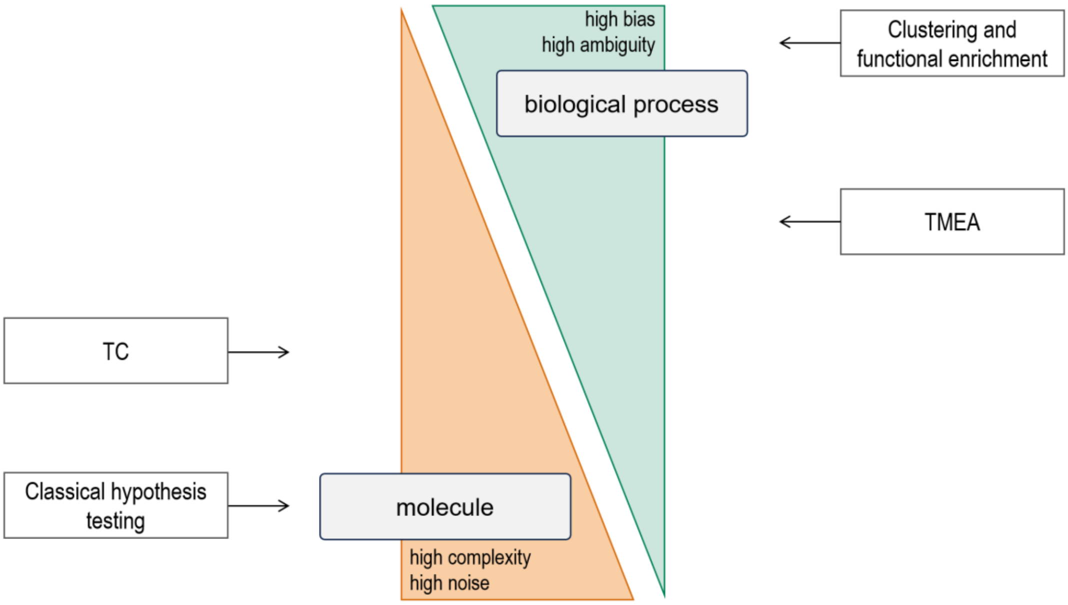


Figure 15 Analysis perspectives. Time series can be analyzed from molecule- to process-centric strategies. Approaches regarding individual molecules result in a multitude of detailed information whereas process-centric approaches lose these details in favor of a condensed representation of the system's response.

Experimental challenges

Capturing a whole system level, requires several purification and processing steps which leads to a relatively high amount of required biological mass especially at the start of the recording. Advances in sequencing methodologies and machine sensitivity have led to a decrease in required sampling amounts. For transcriptomics, single cell sequencing serves as an improvement to reduce the required cell count (Tang et al. 2011). For proteomics, improvements in sample preparation strategies and the development of single cell proteomics were driven by increasing the sensitivity and performance of instruments (Cai et al. 2022; Kassem et al. 2021). However, in many applications the limited number of available cells together with still high costs of high throughput studies constraints the number of sampling time points and measured replicates (Labib and Kelley 2020). Using bioreactors in combination with fast-replicating organisms ensures experimental conditions to be highly controllable and cell counts sufficient for many sampling time points. The sampling time points themselves must be chosen carefully. In contrast to data logging approaches, most biological experiments that involve repeated sampling investigate some sort of system reaction to environmental or developmental factors. In most cases educated guesses have to be made of how the organism is going to react and what is of interest to the researcher. For model fitting techniques, a spacing at which the expected amplitude change is uniformly distributed would be desirable, but as process reaction speeds vary drastically, the choice of sampling schemes always is a compromise.

Analytical challenges

General pitfalls in the analysis of biological time series include their preprocessing and normalization. This thesis does not cover different normalization techniques, however, briefly acknowledging the utter importance of proper normalization for interpreting omics data. Several factors, i.e. handling variations, changes during machine runs, or the measurement technique itself harbor sources for technical variation and were discussed earlier. If the input samples are of comparable composition and concentration, median of ratios (Love et al. 2014) or quantile normalization (Bolstad et al. 2003; Venn et al. 2022) are intuitive and powerful techniques to align abundances of multiple samples. When interpreting visual comparisons of transcript counts or protein abundances of a perfectly normalized dataset, error bars should indicate the biological variability of each element across all biological replicates. However, this term gives the impression that this variation is undesirable and should be avoided. Approaches such as the use of the *standard error of the mean* (SEM) instead of *standard deviation* are often used to keep these error bars small. When the labeling is unclear or the sole reason to favor SEM is to reduce the error bar, it is not only misleading to the reader, but even could deceive the researcher to overinterpret the data when no proper statistical testing is performed. A reconsideration is needed that does not condemn large error bars out of hand but encourages the results to be presented as they were measured to allow transparent interpretation. However, biological variability can never be analyzed in detail, as it always occurs in combination with technical variance.

Another important aspect regarding the time series visualization is the applied standardization technique. Since abundances can span several orders of magnitude and therefore signals may differ in amplitude while still behaving similarly over time, a standardization to the same amplitude range is necessary to be able to detect these similarities. A popular approach is to determine relative changes to the control time point (T0) and visualize these fold changes as logarithm of base 2. This is convenient as an amplitude change of +1 always indicates a doubling (+2 corresponds to a fold change of 4) and -1 indicates the reduction to half of the original reading. However, there is a flaw in this strategy. As every time point is related to the first, the accuracy of all data points depends on the accuracy of a single sample (the first one). Slight deviations in the first sample can have huge implications for the whole time course. A better strategy is to use a z-score transform as described earlier. Here, the impact of all samples is evened out by setting the time course of each molecule to zero mean and unit variance. A single outlier still has an impact on the values of all other points, although to a lesser degree.

When it comes to the detection of effects that occurred during a time course, the examination of gene expression or protein abundance data is crucial, as these changes encode most of the relevant cellular responses that are required when faced with the applied condition. The homogeneity of the chemical structure of the molecules led to the development of high throughput techniques that cover the majority of the cellular molecule pool and give accurate results even when using single cells as starting material (Jovic et al. 2022; Bennett et al. 2023).

When relating abundance differences with the observed variances between replicates, probabilities can be determined of how likely the observed difference is valid and relevant. This corresponds to a molecule-centric perspective which enables the examination focused on individual transcripts or proteins (Figure 14). As for any data analysis approach assumptions and prerequisites have to be discussed before employing statistical tests. Dozens of approaches and statistical packages are available to study these differences (e.g. repeated measures (two way) ANOVA, DESeq2, or limma) (Macey et al. 2016; Jones 1985; Love et al. 2014; Ritchie et al. 2015). Taken together, the aim of inferential statistical analyses is to support the researcher to focus on real differences and neglect differences that are most likely due to chance and probably would waste time and resources when investigated further.

However, it should always be kept in mind, that transcript counts as well as protein abundance are just estimators of their respective molecule activity. For transcripts, activity in most cases refers to the translation rate of mRNA to proteins. The activity of proteins that are involved in enzymatic reactions or signalling pathways on the other hand, is evaluated from their turnover rate. There are dozens of regulatory mechanisms that influence the activity, despite its abundance being constant. However, until now these abundance proxies are the method of choice to get closest to the actual activity level. Several methods have been developed to determine the activity status of proteins that are regulated by modifications. The most common modification that regulates protein activity is their phosphorylation by kinases. Phosphoproteomics is based on the enrichment of phosphopeptides using phosphospecific antibodies with subsequent mass spectroscopic analysis (Kratchmarova et al. 2005; Zhang et al. 2005). In redox proteomics, alkylating agents lead to a size difference between reduced and oxidized proteins, which can subsequently be separated by electrophoresis and measured individually (Zimmer et al. 2021). On the transcriptome side, it is possible to determine the read rate of individual transcripts using RiboSeq. Here, mRNA-ribosome complexes are isolated and the mRNA that is not occupied by ribosomes is digested. The remaining mRNA fragments can then be sequenced to obtain an overview of how many ribosomes were in the process of translation at any given time (Ingolia 2014; Dougherty 2017). RNA immunoprecipitation involves the targeted enrichment of transcripts to which ribosomal proteins or initiation factors are bound. Subsequent sequencing can also provide an overview of the translatome. A proteomics-based analysis of the transcript activity status is the so-called Puromycin-Associated Nascent Chain Proteomics (PUNCH-P) in which nascent chains emerging from the ribosome are labeled, isolated, and analyzed via mass spectrometry (Aviner et al. 2014). Using these techniques in combination with time resolved monitoring of cellular reactions will shed light on regulatory particularities and have the potential to give rise to new medication approaches or increase acclimation capabilities.

Development of analysis approaches specialized on omics time series

The wide range of analytical methods for short time series ranges from the analysis of individual molecules to a holistic view of the system response (Figure 14). The set of tools can be extended by developing intermediate approaches that aim for a certain balance between both extremes. Both presented approaches *temporal classification* (TC) as well as *thermodynamically motivated enrichment analysis* (TMEA) aim to occupy two of those balance niches that are further specified in the following.

The identification of groups that share the same regulator can be a promising approach when studying the response to stressors. Besides network analysis techniques, clustering algorithms are commonly applied to longitudinal data to identify signal groups that behave coherently over time. This approach is not limited to time series data, but can also be applied to data of several conditions instead of time points. Clustering challenges include the determination of the correct cluster number, missing point weighting based on replicate variability, and reliance on statistical tests for subsequent enrichment analyses. TC partially solves these issues and tends to be located in the molecular spectrum of Figure 14. By using predefined classes instead of arbitrary cluster numbers, signals do not influence the classifier. While in theory, every data addition requires the recomputation of clustering techniques, classification strategies are based on a feature extraction that is independent of other presented data. Low replicate counts combined with high biological and technical variance not only reduce statistical power but also degrade the performance of clustering approaches. TC employs constrained smoothing splines to model each molecule signal as a continuous and smooth curve. These curves are constrained to reduce oscillations where possible and incorporate the time point variance leading to a robust description of the time course. In regions of increased noise, the curve tends to be constant in order to reduce overinterpretation of measurement artefacts. This incorporation of measurement noise in signal modeling is exclusive to time series data. Neighboring time point readings are highly dependent on each other and thereby present unique characteristics. While for classical hypothesis testing, this dependence requires careful consideration of testing framework prerequisites, it enables curve fitting strategies to be applied to exploit information of measurement values in the near proximity to get a more robust estimate of the actual measurements. As proven in (Venn et al. 2024), biological time courses in general tend to be smooth. Sudden changes in slope or curvature are unlikely from a regulatory perspective as they require an energy-dependent decay or *de novo* synthesis of molecules. To ensure efficient use of energy resources, continuous modeling is preferable, especially in cases where available measurement time points are sparse and thus sudden regulatory events cannot be made visible at all. This leads to constraining assumptions that can be modeled using monotonicity constraints. By isolating features out of the modeled signals that can be used for subsequent classification, molecule kinetics can be sorted into predetermined groups with labels that accurately describe the feature of interest. Besides the function trajectory itself, its slope and curvature can be calculated efficiently and can be used e.g. for the determination of regulatory events that point to the molecule's function or regulation. By using extreme values as classifying features we chose events in which the molecule abundance trend switches from de- to increasing or vice versa. It is important to note, that these time points do not necessarily indicate time points of regulatory switches that affect the molecule in question. A switch in regulation in most cases would lead to a sign change in curvature rather than a sign change in slope. However, the modeling of the time series as continuous piecewise polynomials allows for the isolation of various characteristics. Besides the function trajectory itself, a comparison based on slope or curvature is trivial to establish. During the development of this method, it was found that comparing it with established clustering-based approaches is challenging. The number and topology of the resulting groups are too heterogeneous for an objective comparison. Clustering algorithms typically generate few and similarly sized clusters, whereas in TC, the number of classes is fixed before analysis, and the class size is distributed heterogeneously depending on the given stimulus. However, the question posed to both methods is fundamentally different. Clustering aims to group signals based on similarity, while classification determines relevant properties in advance and sorts signals into classes based on those properties, regardless of internal similarities. Although both methods separate data into groups, it is difficult to make an objective comparison. However, the analysis of the robustness of different fitting methods has shown that spline-based modeling has advantages over mean-based clustering. Time points of high variance, which are aggregated to their mean value during clustering, can be specifically taken into account by the modeling so that their relevance for the subsequent classification is reduced. A limitation that must be considered is the reliance on a smooth curve, preventing sudden slope changes. While this is an essential property for the description of molecule kinetics at constant environmental conditions, it prohibits the application of TC to time series experiments that undergo an additional change in condition other than the initial perturbation at the very first time point. Such condition changes during the time course could result in sudden reregulation of molecules that consequently would change the slope of molecule kinetics instantly. However, the modeling of short time series by constrained smoothing splines allows for a detailed and robust representation of the underlying kinetics. The possibility to aggregate signal kinetics based on selected features rather than non-reproducible similarity clustering, enables subsequent enrichment analysis that has an increased degree of conciseness, as group labels describe distinct kinetic features. In future, extensions of the current methodology could include quality measures of assigned class labels, that give hint to the label's robustness. By assessing mGCVs of all applied models and smoothing factors (λ), an empirical score distribution can be obtained and compared to the score of the final model. Furthermore, the isolation of signal features of each molecule could serve as input for neural networks, enabling a new perspective of time series analysis. In addition to identified extreme points, slope and curvature signals may provide valuable benefits as feature input vectors. A third application perspective is the comparison between two time courses. This could be established by developing distance measures that describe the differences between the two curves and report a significance of how likely both signals differ from each other. Using the current state of the methodology, classes resulting from the classification and concise labeling can be investigated individually by filtering specific events, or globally by enrichment analysis. Thereby, scientists are assisted in detecting interesting regulatory peculiarities and examining the regulatory cellular events in their temporal order. The subsequent analysis of enriched functional terms within a class helps to characterize the chronological response sequence of cellular processes. Searching for overrepresented functional terms is a common approach to condense the information of hundreds of individual signals to a few dozen statements. Over- or underrepresented groups can be used to characterize either network communities, clusters, or a group of differentially expressed genes. A necessary preprocessing step is to partition the data into two (or more) groups. While one serves as background knowledge, the group of interest is analyzed for deviations from the observed background term distribution.

TMEA pursues a different strategy and can be located more towards the process-centric spectrum of Figure 14. Enrichment strategies often rely on a prior hypothesis testing to distinguish between significantly different molecules versus molecules that show a constant time course. Instead of partitioning the data into these two groups, TMEA associates weights to each molecule that indicate its relevance to the most prominent time courses. While this weighting is molecule-centric, the association of these weights to the most prominent system responses that originate from aggregation procedures is process-centric. By determining the weight sum for each functional term and conducting a permutation test, p-values are assigned to the terms. Instead of a minimum count threshold that must be exceeded in conventional enrichment tests, even single strong weights can lead to a significant contribution to the respective constraint. This is especially useful when searching for regulators that serve as on/off switches for regulatory or metabolic pathways. In these scenarios, abundances of many involved proteins stay the same while a single conversion step is inhibited to prevent further pathway activity. The regulation of switch-like molecules is efficient as it influences all downstream reactions without the need to regulate all involved enzymes. Naturally these switch-behaviors are common in pathways, whose metabolic intermediates are pathway exclusive and are not synthesized from multiple processes. Future applications of TMEA include the concatenation of time series data from differing experiments. Thereby, TMEA can extract and compare the main components driving the changes between different conditions.

There is no free lunch

TMEA as well as TC extend recognized analysis strategies to refine the interpretation capabilities that time series experiments provide. As omics technologies generate thousands of data points, an aggregation step is inevitable for the interpretation at a systems level. Established methods such as similarity-based clustering and gene set enrichment strategies enable the researchers to extract global effects of the experimental treatments. A more detailed dissemination of the data in which even subtle changes and regulatory peculiarities become visible inevitably requires a higher degree of manual examination of the results. Wealth of detail always goes hand in hand with a lower aggregation degree and therefore a higher interpretation effort and higher degree of error or miss interpretation. Both methods, TMEA and TC, were developed to balance the amount of additional information and necessary manual inspection.

As outlined in the introduction, Occam’s razor principle – or parsimony principle – states that when modeling unknown phenomena, the use of coefficients should be kept low where possible. However, in the presented case of modeling time series data using constrained smoothing splines, several constraints are imposed: Unlike simple interpolating polynomials with n coefficients, a smoothing spline of the same time series contains $4(n-1)$ coefficients. In addition, there are restrictions on the monotonicity and curvature of the function, as well as on the weighting of each time point depending on the variance. The reason why this is still a valid procedure lies in the motivation for the restriction. Both interpolating and regressing polynomials tend to oscillate strongly, especially in the bounds. However, the statement that little is known about the abundance curve of biological time series is only partially true. Although there are hardly any concrete biochemical formulas that can describe system-wide adaptive reactions, it was shown that the function progressions are generally both smooth and low in oscillation. As a result, models built under these constraints are more accurate, even though they require a higher number of coefficients and assumptions.

In articles III and IV some of the aforementioned methods were applied to time series data of heat acclimation experiments in *Chlamydomonas reinhardtii*. Besides classical hypothesis testing, ontology enrichment, and several clustering approaches, network-based community detection and transcript-protein correlation analysis were conducted. Because transcriptomics as well as proteomics samples were taken throughout the time course, both systems levels could be compared where transcripts were quantified (RNASeq) along with their corresponding proteins (MS proteomics). Note, that these correlations cannot be determined for individual transcript-protein pairs, but only for functional groups as defined in MapMan (Thimm et al. 2004). Therefore, the predictive capabilities for individual molecules must be questioned. However, it became apparent that transcript counts and protein abundances do not correlate well at control conditions (Article III Figure 2e). The average Pearson's correlation coefficient was approximately zero, indicating that the abundance of most proteins cannot be predicted from the transcript counts. While this may seem astonishing, this phenomenon is commonly observed (Bauernfeind and Babbitt 2017; Moritz et al. 2019; Edfors et al. 2016). It partly can be explained by the manifold regulation mechanisms that decouple protein levels from mRNA levels (Figure 4) (Wilhelm et al. 2017; Payne 2015; Gygi et al. 1999). In steady state with no acclimation pressure, cellular processes seem to rely on fine-tuned regulation and not on the relatively costly regulation by increasing translation rates to increase first transcription and finally protein abundances. Under stress, however, transcriptional regulation seems to have a major stake in regulating the pool of available protein. The Pearson correlation coefficient between transcripts and proteins strongly increases after heat onset (2-8 h) and steadily declines to zero after 48 h of recovery at control conditions. A direct link between gene expression and protein activity without reliance on post translational modifications could be interpreted as a conservative and robust backup system if fine adjustment via assembly regulation, allosteric inhibition, and other cross talk regulation is derailed, like for sudden environmental perturbations. Cells need to rapidly adjust their physiological and biochemical states, leading to more synchronized changes in both mRNA and protein levels. This coordination seems to ensure an effective response to mitigate damage or adapt to new conditions (Halbeisen and Gerber 2009; Lackner et al. 2012).

Besides these global analyses, several specific findings have been made that cannot be discussed in necessary depth here. One discovery, however, was the effect of supplied acetate during heat acclimation. As an organic carbon source, *Chlamydomonas* is able to grow heterotrophically without light driving photosynthesis by importing acetate from the surrounding environment and converting it to carbohydrates in the glyoxylate cycle and gluconeogenesis (Burlacot et al. 2019; Johnson and Alric 2012). Of course, any additional energy source can become vital under environmental perturbations that require major metabolic and structural reorganization. Consequently, an increased acetate uptake to supply thermotolerance processes during heat acclimation is not surprising (Zhang et al. 2023; Olas et al. 2021). Besides the import of external carbon sources, the energy distribution is actively reorganized during heat acclimation by the redirection of photosynthetic energy from the Calvin cycle to the remodeling of membrane composition to ensure an appropriate fluidity (Hemme et al. 2014). Stopping the Calvin cycle in an effective and quick manner is part of ongoing research. First glances in the evolutionary conserveness degree of RCA1, a regulator of RubisCO, hint to a deliberate thermolability that enables quick aggregation and thereby a stop of the Calvin cycle (unpublished data). It was shown that acetate may also play a critical role in protecting PSII against photoinhibition and support the accumulation of thermoprotective metabolites (Schroda et al. 2015; Roach et al. 2013; Hemme et al. 2014).

In summary, time series analysis provides a lens through which we can observe, quantify, and interpret the temporal patterns inherent in cellular processes. From the oscillating rhythm of molecular clocks to the orchestrated interaction of cellular networks, this experimental approach offers an analytical method to unravel the dynamic nature of biological systems. This allows for a more nuanced understanding of the regulatory mechanisms governing cellular dynamics. Due to the possibility applying modeling and prediction techniques, the total information gain of time series data is greater than the sum of the information drawn from their individual point comparisons. Moreover, time series analysis goes beyond retrospective examination. It empowers us to construct predictive models of cellular behavior. By deciphering temporal trends and correlations, we gain the ability to anticipate future states of cellular systems. This predictive capacity holds immense potential for understanding the consequences of perturbations, offering valuable insights for therapeutic interventions and environmental assessments.

5. Table of Abbreviations

AICc	corrected Akaike information criterion
ANOVA	analysis of variance
cDNA	complementary DNA
CS	citrate synthase
DBSCAN	density-Based Spatial Clustering of Applications with Noise
DTW	dynamic time warping
ENA	European Nucleotide Archive
ER	endoplasmic reticulum
E_{sq}	squared euclidean distance
FDR	false discovery rate
GC	gas chromatography
GEO	Gene Expression Omnibus
GO	Gene Ontology
GSE	Gene Expression Series
GSEA	gene set enrichment analysis
HPLC	high-performance liquid chromatography
HSF	heat shock factor
HSP	heat shock protein
KEGG	Kyoto Encyclopedia of Genes and Genomes
LC	liquid chromatography
LHC	light harvesting complex
mGCV	modified generalized cross validation
miRNA	micro RNA
mRNA	messenger RNA
MS	mass spectrometry
n.s.	not significant
NGS	next generation sequencing
NMR	nuclear magnetic resonance
PAGE	polyacrylamide gel electrophoresis
PCA	principle component analysis
PRIDE	proteomics identification database
PS	photosystem
PTM	post translational modification
PunchP	Puromycin-Associated Nascent Chain Proteomics
QconCAT	quantitative concatamers
R^2	coefficient of determination
rLog	regularized logarithm
RNASeq	RNA sequencing
rRNA	ribosomal RNA
RubisCO	Ribulose-1,5-bisphosphate carboxylase/oxygenase
SEM	standard error of the mean
SNP	single nucleotide polymorphisms
stDev	standard deviation
TC	Temporal classification
TMEA	thermodynamically motivated enrichment analysis
TPM	transcript per million
var	variance
VST	variance stabilizing transformation

6. Table of Figures

7. Curriculum Vitae

Benedikt Venn

Education

since 2018	PhD Student, Computational Systems Biology, RPTU Kaiserslautern
2015 – 2018	Master *Microbial and Plant Biotechnology*, TU Kaiserslautern
2012 – 2015	Bachelor *Biowissenschaften*, TU Kaiserslautern
2003 – 2012	Abitur, Geschwister-Scholl-Gymasium Daun

Engagement

since 2023	Member of the *FsLab* steering committee
2016 – 2018	Member of the department council biology, TU Kaiserslautern
2012 – 2017	Member of the student council biology, TU Kaiserslautern
2017	Member of the accreditation committee, bachelor program *Biowissenschaften* and master programm *Biology*, TU Kaiserslautern
2017	Member of the appointment committee *Molecular Ecology*

Profession

2018 – 2024	Pre-doctoral fellow, *Computational Systems Biology*, TU Kaiserslautern
2015 – 2017	Student assistant, *Mol. Biotechnology & Systems Biology*, TU Kaiserslautern
2016 – 2017	Student assistant, *Deanship Biology*, TU Kaiserslautern
2015	Student assistant, *Plant Physiology*, TU Kaiserslautern

Honors

2019	*Recognized F# Expert* at the *Applied F# Challenge 2019*
2018	Graduated with honors, Master *Microbial and Plant Biotechnology*
2014 – 2017	Scholarship holder *Deutschlandstipendium*

Selected software packages

1. **Venn B**, Mühlhaus T, Schneider K, Weil HL, Zimmer D. fslaborg/FSharp.Stats: Multipurpose project for statistical testing, linear algebra, machine learning, fitting and signal processing. Zenodo. 2022. doi: 10.5281/zenodo.6337056
2. Schneider K, **Venn B**, Mühlhaus T. Plotly.NET: A fully featured charting library for .NET programming languages. Zenodo. 2022. doi: 10.5281/zenodo.6344285
3. **Venn B**, Mühlhaus T. CSBiology/OntologyEnrichment: Command line tool to perform functional annotation and ontology enrichments for Chlamydomonas and Arabidopsis. Zenodo. 2022. doi: 10.5281/zenodo.6340412
4. **Venn B**, Mühlhaus T. CSBiology/TempClass: Package for fitting and classification of short and noisy time series using constrained smoothing splines. Zenodo. 2023. doi: 10.5281/zenodo.10040283
5. Schneider K, Weil HL, Zimmer D, **Venn B**, Mühlhaus T. CSBiology/BioFSharp: Open source bioinformatics and computational biology toolbox written in F#. Zenodo. 2022. doi: 10.5281/zenodo.6335372

Publications

1. **Venn B**, Leifeld T, Zhang P, Mühlhaus T. Temporal classification of short time series data. BMC Bioinformatics. 2024. doi: 10.1186/s12859-024-05636-6
2. Zhang N*, **Venn B***, Bailey CE, Xia M, Mattoon EM, Mühlhaus T, Zhang R. Moderate High Temperature is Beneficial or Detrimental Depending on Carbon Availability in the Green Alga Chlamydomonas reinhardtii. J Exp Bot. 2023. doi: 10.1093/jxb/erad405. Epub ahead of print. PMID: 37877811
3. Araguirang GE, **Venn B**, Kelber NM, Feil R, Lunn J, Kleine T, Leister D, Mühlhaus T, Richter AS. Spliceosomal complex components are critical for adjusting the C:N balance during high-light acclimation. bioRxiv. 2023. doi: 10.1101/2023.07.19.549727
4. Scherhag A, Räschle M, Unbehend N, **Venn B**, Glueck D, Mühlhaus T, Keller S, Pérez Patallo E, Zehner S, Frankenberg-Dinkel N. Characterization of a soluble library of the Pseudomonas aeruginosa PAO1 membrane proteome with emphasis on c-di-GMP turnover enzymes. Microlife. 2023. doi: 10.1093/femsml/uqad028. PMID: 37441524. PMCID: PMC10335732
5. Zhang N, Mattoon EM, McHargue W, **Venn B**, Zimmer D, Pecani K, Jeong J, Anderson CM, Chen C, Berry JC, Xia M, Tzeng SC, Becker E, Pazouki L, Evans B, Cross F, Cheng J, Czymmek KJ, Schroda M, Mühlhaus T, Zhang R. Systems-wide analysis revealed shared and unique responses to moderate and acute high temperatures in the green alga Chlamydomonas reinhardtii. Commun Biol. 2022. doi: 10.1038/s42003-022-03359-z. PMID: 35562408. PMCID: PMC9106746

6. Spaniol B*, Lang J*, **Venn B***, Schake L, Sommer F, Mustas M, Geimer S, Wollman FA, Choquet Y, Mühlhaus T, Schroda M. Complexome profiling on the Chlamydomonas lpa2 mutant reveals insights into PSII biogenesis and new PSII associated proteins. J Exp Bot. 2022. doi: 10.1093/jxb/erab390. PMID: 34436580. PMCID: PMC8730698
7. Schneider K, **Venn B**, Mühlhaus T. Plotly.NET: A fully featured charting library for .NET programming languages. F1000Research. 2022. doi: 10.12688/f1000research.123971.1
8. Garth C, Lukasczyk J, Mühlhaus T, **Venn B**, Krüger J, Glogowski K, Martins Rodrigues C, von Suchodoletz D. Immutable yet evolving: ARCs for permanent sharing in the research data-time continuum. In Heuveline V and Bisheh N (Eds): E-Science-Tage 2021: Share Your Research Data, Heidelberg. heiBOOKS, 2022, S. 366–373. doi: 10.11588/heibooks.979.c13751
9. Schneider K*, **Venn B***, Mühlhaus T. TMEA: A Thermodynamically Motivated Framework for Functional Characterization of Biological Responses to System Acclimation. Entropy (Basel). 2020. doi: 10.3390/e22091030. PMID: 33286800. PMCID: PMC7597090
10. Theis J, Niemeyer J, Schmollinger S, Ries F, Rütgers M, Gupta TK, Sommer F, Muranaka LS, **Venn B**, Schulz-Raffelt M, Willmund F, Engel BD, Schroda M. VIPP2 interacts with VIPP1 and HSP22E/F at chloroplast membranes and modulates a retrograde signal for HSP22E/F gene expression. Plant Cell Environ. 2020. doi: 10.1111/pce.13732. PMID: 31994740
11. Theis J, Lang J, Spaniol B, Ferté S, Niemeyer J, Sommer F, Zimmer D, **Venn B**, Mehr SF, Mühlhaus T, Wollman FA, Schroda M. The Chlamydomonas deg1c Mutant Accumulates Proteins Involved in High Light Acclimation. Plant Physiol. 2019. doi: 10.1104/pp.19.01052. PMID: 31604811. PMCID: PMC6878023
12. Leifeld T, **Venn B**, Cui S, Zhang Z, Mühlhaus T, Zhang P. Curve form based quantization of short time series data. 18th European Control Conference (ECC), Naples, Italy. 2019. doi: 10.23919/ECC.2019.8795870

8. Danksagung

Während meiner Zeit als Student und Doktorand durfte ich viele besonderen Menschen kennenlernen, die mir mit Rat und Tat zur Seite standen und diese Zeit besonders gemacht haben. Dafür möchte ich mich von Herzen bedanken!

Ein besonderer Dank gilt Prof. Timo Mühlhaus, nicht nur für die Möglichkeit meine Promotion in seiner Abteilung zu absolvieren, sondern auch für die vielen Diskussionen zu jeder erdenklichen Tageszeit, die sowohl meine Forschung beeinflusst als auch des Öfteren neue Perspektiven geschaffen haben. Weiterhin gilt mein Dank Prof. Michael Schroda für die Erstellung des Zweitgutachtens meiner Dissertation und die vertrauensvolle Zusammenarbeit. Prof. Stefan Kins danke ich herzlich für die Übernahme des Vorsitzes meiner Prüfungskommission.

Ein weiterer großer Dank geht an die Abteilung CSB, besonders an Lukas, David und Kevin für die spaßige Anfangszeit in den Gebäuden 56 und 23. Den Zusammenhalt in den - nicht allzu seltenen - heiklen Phasen habe ich sehr geschätzt und wenn es darauf ankam, konnte man sich immer auf euch verlassen.

Ebenso möchte ich mich bei ehemaligen Büronachbarn, insbesondere bei Crissi, Justus, Vinny, Claudia und Anna bedanken. Die Abende rund um den Bierkühlschrank in eurer Gesellschaft haben für die oftmals dringend notwendige Zerstreuung gesorgt 🍺

Ganz herzlich möchte ich auch Fred für die viele Frickelei danken, auch wenn sie nicht immer den skeptischen Blicken unseres Sicherheitsbeauftragten Gerhard standhielten 😉

Ein besonderer Dank gebührt meiner Familie und vor allem Anna. Deine ständige Unterstützung, die ermutigenden Worte und Geduld haben mir durch die Höhen und Tiefen der letzten Jahre geholfen.

Die Beiträge von euch allen haben meine Zeit in Kaiserslautern bereichert. Vielen Dank!

9. Publication bibliography

Acosta, Jesus A.; Chen, Juxing; Hancock, Deana (2023): 139 Trace Mineral Source Matters for Weaned Pigs Challenged with Escherichia Coli F18 in Diets without Pharmacological Zinc Oxide. In *Journal of Animal Science* 101 (Supplement_2), pp. 96–97. DOI: 10.1093/jas/skad341.107.

Aebersold, Ruedi; Mann, Matthias (2003): Mass spectrometry-based proteomics. In *Nature* 422 (6928), pp. 198–207. DOI: 10.1038/nature01511.

Aghabozorgi, Saeed; Seyed Shirkhorshidi, Ali; Ying Wah, Teh (2015): Time-series clustering – A decade review. In *Information Systems* 53, pp. 16–38. DOI: 10.1016/j.is.2015.04.007.

Alberts, Bruce (2008): Molecular biology of the cell. 5th ed. New York: Garland Science.

Andreopoulos, Bill; An, Aijun; Wang, Xiaogang; Schroeder, Michael (2009): A roadmap of clustering algorithms: finding a match for a biomedical application. In *Briefings in bioinformatics* 10 (3), pp. 297–314. DOI: 10.1093/bib/bbn058.

Ankerst, Mihael; Breunig, Markus M.; Kriegel, Hans-Peter; Sander, Jörg (1999): OPTICS. In *SIGMOD Rec.* 28 (2), pp. 49–60. DOI: 10.1145/304181.304187.

Aoki, Koh; Ogata, Yoshiyuki; Shibata, Daisuke (2007): Approaches for extracting practical information from gene co-expression networks in plant biology. In *Plant & cell physiology* 48 (3), pp. 381–390. DOI: 10.1093/pcp/pcm013.

Arriaga, Edgar A. (2009): Determining biological noise via single cell analysis. In *Analytical and bioanalytical chemistry* 393 (1), pp. 73–80. DOI: 10.1007/s00216-008-2431-z.

Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M. et al. (2000): Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. In *Nature genetics* 25 (1), pp. 25–29. DOI: 10.1038/75556.

Ashburner, M.; Bonner, J. J. (1979): The induction of gene activity in drosophilia by heat shock. In *Cell* 17 (2), pp. 241–254. DOI: 10.1016/0092-8674(79)90150-8.

Askari, S. (2021): Noise-resistant fuzzy clustering algorithm. In *Granul. Comput.* 6 (4), pp. 815–828. DOI: 10.1007/s41066-020-00230-6.

Aviner, Ranen; Geiger, Tamar; Elroy-Stein, Orna (2014): Genome-wide identification and quantification of protein synthesis in cultured cells and whole tissues by puromycin-associated nascent chain proteomics (PUNCH-P). In *Nature protocols* 9 (4), pp. 751–760. DOI: 10.1038/nprot.2014.051.

Bakker, Ruben; Ellers, Jacintha; Roelofs, Dick; Vooijs, Riet; Dijkstra, Tjeerd; van Gestel, Cornelis A. M.; Hoedjes, Katja M. (2023): Combining time-resolved transcriptomics and proteomics data for Adverse Outcome Pathway refinement in ecotoxicology. In *The Science of the total environment* 869, p. 161740. DOI: 10.1016/j.scitotenv.2023.161740.

Balogh, Gábor; Maulucci, Giuseppe; Gombos, Imre; Horváth, Ibolya; Török, Zsolt; Péter, Mária et al. (2011): Heat stress causes spatially-distinct membrane re-modelling in K562 leukemia cells. In *PloS one* 6 (6), e21182. DOI: 10.1371/journal.pone.0021182.

Banerjee, Dhrubajyoti; Singh, Vikram; Thakur, Ritik (Eds.) (2023): Micro Propagation on Strawberry: A Review: ICSDG.

Bar-Joseph, Ziv; Gitter, Anthony; Simon, Itamar (2012): Studying and modelling dynamic biological processes using time-series gene expression data. In *Nature reviews. Genetics* 13 (8), pp. 552–564. DOI: 10.1038/nrg3244.

Barrett, Tanya; Troup, Dennis B.; Wilhite, Stephen E.; Ledoux, Pierre; Evangelista, Carlos; Kim, Irene F. et al. (2011): NCBI GEO: archive for functional genomics data sets--10 years on. In *Nucleic acids research* 39 (Database issue), D1005-10. DOI: 10.1093/nar/gkq1184.

Bathellier, Camille; Tcherkez, Guillaume; Lorimer, George H.; Farquhar, Graham D. (2018): Rubisco is not really so bad. In *Plant, cell & environment* 41 (4), pp. 705–716. DOI: 10.1111/pce.13149.

Bauernfeind, Amy L.; Babbitt, Courtney C. (2017): The predictive nature of transcript expression levels on protein expression in adult human brain. In *BMC genomics* 18 (1), p. 322. DOI: 10.1186/s12864-017-3674-x.

Beckman, William F.; Jiménez, Miguel Ángel Lermo; Verschure, Pernette J. (2021): Transcription bursting and epigenetic plasticity: an updated view. In *Epigenetics Commun.* 1 (1). DOI: 10.1186/s43682-021-00007-1.

Bennett, Hayley M.; Stephenson, William; Rose, Christopher M.; Darmanis, Spyros (2023): Single-cell proteomics enabled by next-generation sequencing or mass spectrometry. In *Nature methods* 20 (3), pp. 363–374. DOI: 10.1038/s41592-023-01791-5.

Bergamini, Carlo M.; Gambetti, Stefania; Dondi, Alessia; Cervellati, Carlo (2004): Oxygen, reactive oxygen species and tissue damage. In *Current pharmaceutical design* 10 (14), pp. 1611–1626. DOI: 10.2174/1381612043384664.

Berry, John; Houston, Ken S. (1995): Mathematical modelling. Oxford, Burlington, MA: Elsevier (Modular mathematics series).

Bishop, David W.; Brown, Frank A., Jr; Jahn, Theodore L.; Prosser, C. Ladd; Wulff, Verner J. (1950): Comparative animal physiology. Edited by C. Ladd Prosser. Philadelphia: Saunders.

Bjorck, Ake; Pereyra, Victor (1970): Solution of Vandermonde Systems of Equations. In *Mathematics of Computation* 24 (112), p. 893. DOI: 10.2307/2004623.

Blondel, Vincent D.; Guillaume, Jean-Loup; Lambiotte, Renaud; Lefebvre, Etienne (2008): Fast unfolding of communities in large networks. In *J. Stat. Mech.* 2008 (10), P10008. DOI: 10.1088/1742-5468/2008/10/P10008.

Bolstad, B. M.; Irizarry, R. A.; Astrand, M.; Speed, T. P. (2003): A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. In *Bioinformatics (Oxford, England)* 19 (2), pp. 185–193. DOI: 10.1093/bioinformatics/19.2.185.

Booeshaghi, A. Sina; Pachter, Lior (2021): Normalization of single-cell RNA-seq counts by log(x + 1)† or log(1 + x)†. In *Bioinformatics (Oxford, England)* 37 (15), pp. 2223–2224. DOI: 10.1093/bioinformatics/btab085.

Borowitzka, Michael A. (2018): The ‘stress’ concept in microalgal biology—homeostasis, acclimation and adaptation. In *J Appl Phycol* 30 (5), pp. 2815–2825. DOI: 10.1007/s10811-018-1399-0.

Boyko, Alex; Kovalchuk, Igor: Epigenetic Modifications in Plants Under Adverse Conditions: Agricultural Applications. In, pp. 233–267.

Brillinger, David R.: 8 Analysis of variance and problems under time series models. In, vol. 1, pp. 237–278.

Buguet, Alain; Radomski, Manny W.; Reis, Jacques; Spencer, Peter S. (2023): Heatwaves and human sleep: Stress response versus adaptation. In *Journal of the neurological sciences* 454, p. 120862. DOI: 10.1016/j.jns.2023.120862.

Burlacot, Adrien; Peltier, Gilles; Li-Beisson, Yonghua (2019): Subcellular Energetics and Carbon Storage in Chlamydomonas. In *Cells* 8 (10). DOI: 10.3390/cells8101154.

Cai, Xue; Xue, Zhangzhi; Wu, Chunlong; Sun, Rui; Qian, Liujia; Yue, Liang et al. (2022): High-throughput proteomic sample preparation using pressure cycling technology. In *Nature protocols* 17 (10), pp. 2307–2325. DOI: 10.1038/s41596-022-00727-1.

Campbell, Mary K.; Farrell, Shawn O. (2009): Biochemistry. 6. ed., international student ed. Belmont, Calif: Thomson Higher Educaltion.

Carpenter, Barry K. (1984): Determination of organic reaction mechanisms. New York, Chichester: Wiley.

Carrasco-Pujante, Jose; Bringas, Carlos; Malaina, Iker; Fedetz, Maria; Martínez, Luis; Pérez-Yarza, Gorka et al. (2021): Associative Conditioning Is a Robust Systemic Behavior in Unicellular Organisms: An Interspecies Comparison. In *Frontiers in microbiology* 12, p. 707086. DOI: 10.3389/fmicb.2021.707086.

Casella, George; Fienberg, Stephen; Okin, Ingram; Cryer, Jonathan D.; Chan, Kung-Sik (2008): Time Series Analysis. New York, NY: Springer New York.

Chen, Wei; Zhou, Xiaobo (2019): Drug Effect Prediction by Integrating L1000 Genomic and Proteomic Big Data. In *Methods in molecular biology (Clifton, N.J.)* 1939, pp. 287–297. DOI: 10.1007/978-1-4939-9089-4_16.

Chowdhury, Debajyoti; Wang, Chao; Lu, Aiping; Zhu, Hailong (2021): Cis-Regulatory Logic Produces Gene-Expression Noise Describing Phenotypic Heterogeneity in Bacteria. In *Frontiers in genetics* 12, p. 698910. DOI: 10.3389/fgene.2021.698910.

Chrousos, George P. (2009): Stress and disorders of the stress system. In *Nature reviews. Endocrinology* 5 (7), pp. 374–381. DOI: 10.1038/nrendo.2009.106.

Coates, Juliet C.; Umm-E-Aiman; Charrier, Bénédicte (2014): Understanding "green" multicellularity: do seaweeds hold the key? In *Frontiers in plant science* 5, p. 737. DOI: 10.3389/fpls.2014.00737.

Cooper, Geoffrey M. (2019): The cell. A molecular approach. International 8th ed. New York, Oxford: Sinauer Associates.

Csárdi, Gábor; Franks, Alexander; Choi, David S.; Airoldi, Edoardo M.; Drummond, D. Allan (2015): Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. In *PLoS genetics* 11 (5), e1005206. DOI: 10.1371/journal.pgen.1005206.

Csoboz, Balint; Balogh, Gabor E.; Kusz, Erzsebet; Gombos, Imre; Peter, Maria; Crul, Tim et al. (2013): Membrane fluidity matters: hyperthermia from the aspects of lipids and membranes. In *International journal of hyperthermia : the official journal of European Society for Hyperthermic Oncology, North American Hyperthermia Group* 29 (5), pp. 491–499. DOI: 10.3109/02656736.2013.808765.

Darwin, Charles (1859): The Origin of Species: Cambridge University Press.

Desai, Keyur H.; Tan, Chuen Seng; Leek, Jeffrey T.; Maier, Ronald V.; Tompkins, Ronald G.; Storey, John D. (2011): Dissecting inflammatory complications in critically injured patients by within-patient gene expression changes: a longitudinal clinical genomics study. In *PLoS medicine* 8 (9), e1001093. DOI: 10.1371/journal.pmed.1001093.

Diego, Tee-Jay A. San; Al-Shaer, Mustafa Abdu Qader; Jamali, Ebrahim Abdulla Husain Al (2023): Re-evaluating eurahaline nature of Nile tilapia, Oreochromis niloticus: A hatchery perspective. In *Int. J. Fish. Aquat. Stud.* 11 (5), pp. 223–231. DOI: 10.22271/fish.2023.v11.i5c.2869.

Domingos, Pedro (1999): The Role of Occam's Razor in Knowledge Discovery. In *Data Mining and Knowledge Discovery* 3 (4), pp. 409–425. DOI: 10.1023/A:1009868929893.

Dougherty, Joseph D. (2017): The Expanding Toolkit of Translating Ribosome Affinity Purification. In *The Journal of neuroscience : the official journal of the Society for Neuroscience* 37 (50), pp. 12079–12087. DOI: 10.1523/JNEUROSCI.1929-17.2017.

Dudek, M.; Angelucci, C.; Pathiranage, D.; Wang, P.; Mallikarjun, V.; Lawless, C. et al. (2021): Circadian time series proteomics reveals daily dynamics in cartilage physiology. In *Osteoarthritis and cartilage* 29 (5), pp. 739–749. DOI: 10.1016/j.joca.2021.02.008.

Dunnett, Charles W. (1955): A Multiple Comparison Procedure for Comparing Several Treatments with a Control. In *Journal of the American Statistical Association* 50 (272), pp. 1096–1121. DOI: 10.1080/01621459.1955.10501294.

Dyer, S. A.; Dyer, J. S. (2001): Cubic-spline interpolation. 1. In *IEEE Instrum. Meas. Mag.* 4 (1), pp. 44–46. DOI: 10.1109/5289.911175.

Edfors, Fredrik; Danielsson, Frida; Hallström, Björn M.; Käll, Lukas; Lundberg, Emma; Pontén, Fredrik et al. (2016): Gene-specific correlation of RNA and protein levels in human cells and tissues. In *Molecular systems biology* 12 (10), p. 883. DOI: 10.15252/msb.20167144.

Edgar, Ron; Domrachev, Michael; Lash, Alex E. (2002): Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. In *Nucleic acids research* 30 (1), pp. 207–210. DOI: 10.1093/nar/30.1.207.

Edwards, Dilwyn; Hamson, Mike (1989): Guide to Mathematical Modelling. London: Macmillan Education UK.

Edwards, Harold M. (1995): Linear Algebra. Boston, MA: Birkhäuser Boston.

Eling, Nils; Morgan, Michael D.; Marioni, John C. (2019): Challenges in measuring and understanding biological noise. In *Nature reviews. Genetics* 20 (9), pp. 536–548. DOI: 10.1038/s41576-019-0130-6.

Emig, Dorothea; Albrecht, Mario (2011): Tissue-specific proteins and functional implications. In *Journal of proteome research* 10 (4), pp. 1893–1903. DOI: 10.1021/pr101132h.

Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996): A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Evangelos Simoudis, Jiawei Han, Usama Fayyad (Eds.): 2nd International Conference on Knowledge Discovery and Data Mining. KDD'96. Portland Oregon.

Feder, M. E.; Hofmann, G. E. (1999): Heat-shock proteins, molecular chaperones, and the stress response: evolutionary and ecological physiology. In *Annual review of physiology* 61, pp. 243–282. DOI: 10.1146/annurev.physiol.61.1.243.

Ferrando, Alejandro; Castellano, M. Mar; Lisón, Purificación; Leister, Dario; Stepanova, Anna N.; Hanson, Johannes (2017): Editorial: Relevance of Translational Regulation on Plant Growth and Environmental Responses. In *Frontiers in plant science* 8, p. 2170. DOI: 10.3389/fpls.2017.02170.

Fröhlich, Holger; Balling, Rudi; Beerenwinkel, Niko; Kohlbacher, Oliver; Kumar, Santosh; Lengauer, Thomas et al. (2018): From hype to reality: data science enabling personalized medicine. In *BMC medicine* 16 (1), p. 150. DOI: 10.1186/s12916-018-1122-7.

Furlan, Mattia; Pretis, Stefano de; Pelizzola, Mattia (2021): Dynamics of transcriptional and post-transcriptional regulation. In *Briefings in bioinformatics* 22 (4). DOI: 10.1093/bib/bbaa389.

Galluzzi, Lorenzo; Myint, Melissa (2023): Cell death and senescence. In *Journal of translational medicine* 21 (1), p. 425. DOI: 10.1186/s12967-023-04297-y.

Galluzzi, Lorenzo; Yamazaki, Takahiro; Kroemer, Guido (2018): Linking cellular stress responses to systemic homeostasis. In *Nature reviews. Molecular cell biology* 19 (11), pp. 731–745. DOI: 10.1038/s41580-018-0068-0.

Garcia-Molina, Antoni; Kleine, Tatjana; Schneider, Kevin; Mühlhaus, Timo; Lehmann, Martin; Leister, Dario (2020): Translational Components Contribute to Acclimation Responses to High Light, Heat, and Cold in Arabidopsis. In *iScience* 23 (7), p. 101331. DOI: 10.1016/j.isci.2020.101331.

Gardner, T. S.; Cantor, C. R.; Collins, J. J. (2000): Construction of a genetic toggle switch in Escherichia coli. In *Nature* 403 (6767), pp. 339–342. DOI: 10.1038/35002131.

Gaucher, Denis; Therrien, René; Kettaf, Nadia; Angermann, Bastian R.; Boucher, Geneviève; Filali-Mouhim, Abdelali et al. (2008): Yellow fever vaccine induces integrated multilineage and polyfunctional immune responses. In *The Journal of experimental medicine* 205 (13), pp. 3119–3131. DOI: 10.1084/jem.20082292.

Gebauer, Fátima; Hentze, Matthias W. (2004): Molecular mechanisms of translational control. In *Nature reviews. Molecular cell biology* 5 (10), pp. 827–835. DOI: 10.1038/nrm1488.

Gergonne, J.D (1974): The application of the method of least squares to the interpolation of sequences. In *Historia Mathematica* 1 (4), pp. 439–447. DOI: 10.1016/0315-0860(74)90034-2.

Gerstein, Mark B.; Lu, Zhi John; van Nostrand, Eric L.; Cheng, Chao; Arshinoff, Bradley I.; Liu, Tao et al. (2010): Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. In *Science (New York, N.Y.)* 330 (6012), pp. 1775–1787. DOI: 10.1126/science.1196914.

Ghatak, Sankha; King, Zachary A.; Sastry, Anand; Palsson, Bernhard O. (2019): The y-ome defines the 35% of Escherichia coli genes that lack experimental evidence of function. In *Nucleic acids research* 47 (5), pp. 2446–2454. DOI: 10.1093/nar/gkz030.

Greenland, S. (1995): Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. In *Epidemiology (Cambridge, Mass.)* 6 (4), pp. 356–365. DOI: 10.1097/00001648-199507000-00005.

Gygi, S. P.; Rochon, Y.; Franza, B. R.; Aebersold, R. (1999): Correlation between protein and mRNA abundance in yeast. In *Molecular and cellular biology* 19 (3), pp. 1720–1730. DOI: 10.1128/MCB.19.3.1720.

Halbeisen, Regula E.; Gerber, André P. (2009): Stress-dependent coordination of transcriptome and translatome in yeast. In *PLoS biology* 7 (5), e1000105. DOI: 10.1371/journal.pbio.1000105.

Hamada, Toshiyuki; Sutherland, Kenneth; Ishikawa, Masayori; Miyamoto, Naoki; Honma, Sato; Shirato, Hiroki; Honma, Ken-Ichi (2016): In vivo imaging of clock gene expression in multiple tissues of freely moving mice. In *Nature communications* 7, p. 11705. DOI: 10.1038/ncomms11705.

Hammel, Alexander; Zimmer, David; Sommer, Frederik; Mühlhaus, Timo; Schroda, Michael (2018): Absolute Quantification of Major Photosynthetic Protein Complexes in Chlamydomonas reinhardtii Using Quantification Concatamers (QconCATs). In *Frontiers in plant science* 9, p. 1265. DOI: 10.3389/fpls.2018.01265.

Hassoun, Soha; Jefferson, Felicia; Shi, Xinghua; Stucky, Brian; Wang, Jin; Rosa, Epaminondas (2022): Artificial Intelligence for Biology. In *Integrative and comparative biology* 61 (6), pp. 2267–2275. DOI: 10.1093/icb/icab188.

Heide, Heinrich; Bleier, Lea; Steger, Mirco; Ackermann, Jörg; Dröse, Stefan; Schwamb, Bettina et al. (2012): Complexome profiling identifies TMEM126B as a component of the mitochondrial complex I assembly complex. In *Cell metabolism* 16 (4), pp. 538–549. DOI: 10.1016/j.cmet.2012.08.009.

Hemme, Dorothea; Veyel, Daniel; Mühlhaus, Timo; Sommer, Frederik; Jüppner, Jessica; Unger, Ann-Katrin et al. (2014): Systems-wide analysis of acclimation responses to long-term heat stress and recovery in the photosynthetic model organism Chlamydomonas reinhardtii. In *The Plant cell* 26 (11), pp. 4270–4297. DOI: 10.1105/tpc.114.130997.

Ho, Brandon; Baryshnikova, Anastasia; Brown, Grant W. (2018): Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome. In *Cell systems* 6 (2), 192-205.e3. DOI: 10.1016/j.cels.2017.12.004.

Ingolia, Nicholas T. (2014): Ribosome profiling: new views of translation, from single codons to genome scale. In *Nature reviews. Genetics* 15 (3), pp. 205–213. DOI: 10.1038/nrg3645.

Jackson, Rebecca; Matentzoglu, Nicolas; Overton, James A.; Vita, Randi; Balhoff, James P.; Buttigieg, Pier Luigi et al. (2021): OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. In *Database : the journal of biological databases and curation* 2021. DOI: 10.1093/database/baab069.

Janni, Michela; Gullì, Mariolina; Maestri, Elena; Marmiroli, Marta; Valliyodan, Babu; Nguyen, Henry T.; Marmiroli, Nelson (2020): Molecular and genetic bases of heat stress responses in crop plants and breeding for increased resilience and productivity. In *Journal of experimental botany* 71 (13), pp. 3780–3802. DOI: 10.1093/jxb/eraa034.

Johnson, Xenie; Alric, Jean (2012): Interaction between starch breakdown, acetate assimilation, and photosynthetic cyclic electron flow in Chlamydomonas reinhardtii. In *The Journal of biological chemistry* 287 (31), pp. 26445–26452. DOI: 10.1074/jbc.M112.370205.

Jollyta, Deny; Efendi, Syahril; Zarlis, Muhammad; Mawengkang, Herman (2023): Analysis of an optimal cluster approach: a review paper. In *J. Phys.: Conf. Ser.* 2421 (1), p. 12015. DOI: 10.1088/1742-6596/2421/1/012015.

Jones, R. H. (1985): Repeated measures, interventions, and time series analysis. In *Psychoneuroendocrinology* 10 (1), pp. 5–14. DOI: 10.1016/0306-4530(85)90035-6.

Jovic, Dragomirka; Liang, Xue; Zeng, Hua; Lin, Lin; Xu, Fengping; Luo, Yonglun (2022): Single-cell RNA sequencing technologies and applications: A brief overview. In *Clinical and translational medicine* 12 (3), e694. DOI: 10.1002/ctm2.694.

Jung, Robert C.; Tremayne, A. R. (2003): Testing for serial dependence in time series models of counts. In *Journal Time Series Analysis* 24 (1), pp. 65–84. DOI: 10.1111/1467-9892.00293.

Kassem, Sara; van der Pan, Kyra; Jager, Anniek L. de; Naber, Brigitta A. E.; Laat, Inge F. de; Louis, Alesha et al. (2021): Proteomics for Low Cell Numbers: How to Optimize the Sample Preparation Workflow for Mass Spectrometry Analysis. In *Journal of proteome research* 20 (9), pp. 4217–4230. DOI: 10.1021/acs.jproteome.1c00321.

Kerr, J. F.; Wyllie, A. H.; Currie, A. R. (1972): Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. In *British journal of cancer* 26 (4), pp. 239–257. DOI: 10.1038/bjc.1972.33.

Keselman, H. J.; Algina, J.; Kowalchuk, R. K. (2001): The analysis of repeated measures designs: a review. In *The British journal of mathematical and statistical psychology* 54 (Pt 1), pp. 1–20. DOI: 10.1348/000711001159357.

Khlebovich, V. V. (2017): Acclimation of animal organisms: basic theory and applied aspects. In *Biol Bull Rev* 7 (4), pp. 279–286. DOI: 10.1134/S2079086417040053.

Kim, Dongsan; Kim, Man-Sun; Cho, Kwang-Hyun (2012): The core regulation module of stress-responsive regulatory networks in yeast. In *Nucleic acids research* 40 (18), pp. 8793–8802. DOI: 10.1093/nar/gks649.

Kim, Hyun Uk; Kim, Tae Yong; Lee, Sang Yup (2008): Metabolic flux analysis and metabolic engineering of microorganisms. In *Molecular bioSystems* 4 (2), pp. 113–120. DOI: 10.1039/b712395g.

Kleine, Tatjana; Nägele, Thomas; Neuhaus, H. Ekkehard; Schmitz-Linneweber, Christian; Fernie, Alisdair R.; Geigenberger, Peter et al. (2021): Acclimation in plants - the Green Hub consortium. In *The Plant journal : for cell and molecular biology* 106 (1), pp. 23–40. DOI: 10.1111/tpj.15144.

Kleyman, Michael; Sefer, Emre; Nicola, Teodora; Espinoza, Celia; Chhabra, Divya; Hagood, James S. et al. (2017): Selecting the most appropriate time points to profile in high-throughput studies. In *eLife* 6. DOI: 10.7554/eLife.18541.

Kline, Rachel A.; Lößlein, Lena; Kurian, Dominic; Aguilar Martí, Judit; Eaton, Samantha L.; Court, Felipe A. et al. (2022): An Optimized Comparative Proteomic Approach as a Tool in Neurodegenerative Disease Research. In *Cells* 11 (17). DOI: 10.3390/cells11172653.

Kodinariya, Trupti M; Makwana, Prashant R (2013): Review on determining number of Cluster in K-Means Clustering 1 (6).

Koolhaas, J. M.; Bartolomucci, A.; Buwalda, B.; Boer, S. F. de; Flügge, G.; Korte, S. M. et al. (2011): Stress revisited: a critical evaluation of the stress concept. In *Neuroscience and biobehavioral reviews* 35 (5), pp. 1291–1301. DOI: 10.1016/j.neubiorev.2011.02.003.

Kratchmarova, Irina; Blagoev, Blagoy; Haack-Sorensen, Mandana; Kassem, Moustapha; Mann, Matthias (2005): Mechanism of divergent growth factor effects in mesenchymal stem cell differentiation. In *Science (New York, N.Y.)* 308 (5727), pp. 1472–1477. DOI: 10.1126/science.1107627.

Kriegel, Hans-Peter; Kröger, Peer; Sander, Jörg; Zimek, Arthur (2011): Density-based clustering. In *WIREs Data Min & Knowl* 1 (3), pp. 231–240. DOI: 10.1002/widm.30.

Kültz, Dietmar (2005): Molecular and evolutionary basis of the cellular stress response. In *Annual review of physiology* 67, pp. 225–257. DOI: 10.1146/annurev.physiol.67.040403.103635.

Labib, Mahmoud; Kelley, Shana O. (2020): Single-cell analysis targeting the proteome. In *Nature reviews. Chemistry* 4 (3), pp. 143–158. DOI: 10.1038/s41570-020-0162-7.

Lackner, Daniel H.; Schmidt, Michael W.; Wu, Shuangding; Wolf, Dieter A.; Bähler, Jürg (2012): Regulation of transcriptome, translation, and proteome in response to environmental stress in fission yeast. In *Genome biology* 13 (4), R25. DOI: 10.1186/gb-2012-13-4-r25.

Lagerspetz, Kari Y.H. (2006): What is thermal acclimation? In *Journal of Thermal Biology* 31 (4), pp. 332–336. DOI: 10.1016/j.jtherbio.2006.01.003.

Laird, Nan M.; Ware, James H. (1982): Random-Effects Models for Longitudinal Data. In *Biometrics* 38 (4), p. 963. DOI: 10.2307/2529876.

Langfelder, Peter; Zhang, Bin; Horvath, Steve (2008): Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. In *Bioinformatics (Oxford, England)* 24 (5), pp. 719–720. DOI: 10.1093/bioinformatics/btm563.

Lee, Dave; Smallbone, Kieran; Dunn, Warwick B.; Murabito, Ettore; Winder, Catherine L.; Kell, Douglas B. et al. (2012): Improving metabolic flux predictions using absolute gene expression data. In *BMC systems biology* 6, p. 73. DOI: 10.1186/1752-0509-6-73.

Leoncini, Isabelle; Le Conte, Yves; Costagliola, Guy; Plettner, Erika; Toth, Amy L.; Wang, Mianwei et al. (2004): Regulation of behavioral maturation by a primer pheromone produced by adult worker honey bees. In *Proceedings of the National Academy of Sciences of the United States of America* 101 (50), pp. 17559–17564. DOI: 10.1073/pnas.0407652101.

Leung, Kin Kwan; Rooke, Clayton; Smith, Jonathan; Zuberi, Saba; Volkovs, Maksims (2021): Temporal Dependencies in Feature Importance for Time Series Predictions.

Lindquist, S. (1986): The heat-shock response. In *Annual review of biochemistry* 55, pp. 1151–1191. DOI: 10.1146/annurev.bi.55.070186.005443.

Liu, Hua; Tarima, Sergey; Borders, Aaron S.; Getchell, Thomas V.; Getchell, Marilyn L.; Stromberg, Arnold J. (2005): Quadratic regression analysis for gene discovery and pattern recognition for non-cyclic short time-course microarray experiments. In *BMC bioinformatics* 6, p. 106. DOI: 10.1186/1471-2105-6-106.

Lopatkin, Allison J.; Collins, James J. (2020): Predictive biology: modelling, understanding and harnessing microbial complexity. In *Nature reviews. Microbiology* 18 (9), pp. 507–520. DOI: 10.1038/s41579-020-0372-5.

Lopes-Ramos, Camila M.; Paulson, Joseph N.; Chen, Cho-Yi; Kuijjer, Marieke L.; Fagny, Maud; Platig, John et al. (2017): Regulatory network changes between cell lines and their tissues of origin. In *BMC genomics* 18 (1), p. 723. DOI: 10.1186/s12864-017-4111-x.

Love, Michael I.; Huber, Wolfgang; Anders, Simon (2014): Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. In *Genome biology* 15 (12), p. 550. DOI: 10.1186/s13059-014-0550-8.

Luo, Feng; Yang, Yunfeng; Zhong, Jianxin; Gao, Haichun; Khan, Latifur; Thompson, Dorothea K.; Zhou, Jizhong (2007): Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. In *BMC bioinformatics* 8, p. 299. DOI: 10.1186/1471-2105-8-299.

Luo, Songhao; Wang, Zihao; Zhang, Zhenquan; Zhou, Tianshou; Zhang, Jiajun (2023): Genome-wide inference reveals that feedback regulations constrain promoter-dependent transcriptional burst kinetics. In *Nucleic acids research* 51 (1), pp. 68–83. DOI: 10.1093/nar/gkac1204.

Ma, Patrick C. H.; Chan, Keith C. C. (2009): An iterative data mining approach for mining overlapping coexpression patterns in noisy gene expression data. In *IEEE transactions on nanobioscience* 8 (3), pp. 252–258. DOI: 10.1109/TNB.2009.2026747.

Macey, Paul M.; Schluter, Philip J.; Macey, Katherine E.; Harper, Ronald M. (2016): Detecting variable responses in time-series using repeated measures ANOVA: Application to physiologic challenges. In *F1000Research* 5, p. 563. DOI: 10.12688/f1000research.8252.2.

Maeland, E. (1988): On the comparison of interpolation methods. In *IEEE transactions on medical imaging* 7 (3), pp. 213–217. DOI: 10.1109/42.7784.

Maier, Raina M.; Pepper, Ian L. (2015): Environmental Microbiology (Third Edition). Chapter 3 - Bacterial Growth.

Mathavan, Sinnakaruppan; Lee, Serene G. P.; Mak, Alicia; Miller, Lance D.; Murthy, Karuturi Radha Krishna; Govindarajan, Kunde R. et al. (2005): Transcriptome analysis of zebrafish embryogenesis using microarrays. In *PLoS genetics* 1 (2), pp. 260–276. DOI: 10.1371/journal.pgen.0010029.

McIntyre, Lauren M.; Lopiano, Kenneth K.; Morse, Alison M.; Amin, Victor; Oberg, Ann L.; Young, Linda J.; Nuzhdin, Sergey V. (2011): RNA-seq: technical variability and sampling. In *BMC genomics* 12, p. 293. DOI: 10.1186/1471-2164-12-293.

Melé, Marta; Ferreira, Pedro G.; Reverter, Ferran; DeLuca, David S.; Monlong, Jean; Sammeth, Michael et al. (2015): Human genomics. The human transcriptome across tissues and individuals. In *Science (New York, N.Y.)* 348 (6235), pp. 660–665. DOI: 10.1126/science.aaa0355.

Mendoza-Parra, Marco A.; Walia, Mannu; Sankar, Martial; Gronemeyer, Hinrich (2011): Dissecting the retinoid-induced differentiation of F9 embryonal stem cells by integrative genomics. In *Molecular systems biology* 7, p. 538. DOI: 10.1038/msb.2011.73.

Menges, Margit; Hennig, Lars; Gruissem, Wilhelm; Murray, James A. H. (2002): Cell cycle-regulated gene expression in Arabidopsis. In *The Journal of biological chemistry* 277 (44), pp. 41987–42002. DOI: 10.1074/jbc.M207570200.

Michaelis, L; Menten, ML (1913): Die Kinetik der Invertinwirkung (49), pp. 333–369.

Micula, Gheorghe; Micula, Sanda (1998): Handbook of Splines. Cham: Springer International Publishing; Springer Nature.

Minkin, Vladimir I..; Carpenter, Barry K..: Ockham's Razor and Chemistry. In.

Mittler, Ron (2002): Oxidative stress, antioxidants and stress tolerance. In *Trends in plant science* 7 (9), pp. 405–410. DOI: 10.1016/s1360-1385(02)02312-9.

Modell, Harold; Cliff, William; Michael, Joel; McFarland, Jenny; Wenderoth, Mary Pat; Wright, Ann (2015): A physiologist's view of homeostasis. In *Advances in physiology education* 39 (4), pp. 259–266. DOI: 10.1152/advan.00107.2015.

Moritz, Christian P.; Mühlhaus, Timo; Tenzer, Stefan; Schulenborg, Thomas; Friauf, Eckhard (2019): Poor transcript-protein correlation in the brain: negatively correlating gene products reveal neuronal polarity as a potential cause. In *Journal of neurochemistry* 149 (5), pp. 582–604. DOI: 10.1111/jnc.14664.

Müller, M (2007): Information Retrieval for Music and Motion. Dynamic Time Warping: Springer, Berlin, Heidelberg.

Narad, Priyanka; Kirthanashri, S. V. (2019): Omics - Approaches, Technologies And Applications. Introduction to Omics: Springer, Singapore.

Niemeyer, Justus; Scheuring, David; Oestreicher, Julian; Morgan, Bruce; Schroda, Michael (2021): Real-time monitoring of subcellular H2O2 distribution in Chlamydomonas reinhardtii. In *The Plant cell* 33 (9), pp. 2935–2949. DOI: 10.1093/plcell/koab176.

Noy, Natalya F; McGuiness, Deborah L (2001): Ontology Development 101: A Guide to Creating Your First Ontology. Stanford University, Stanford, CA, 94305.

Olas, Justyna Jadwiga; Apelt, Federico; Annunziata, Maria Grazia; John, Sheeba; Richard, Sarah Isabel; Gupta, Saurabh et al. (2021): Primary carbohydrate metabolism genes participate in heat-stress memory at the shoot apical meristem of Arabidopsis thaliana. In *Molecular plant* 14 (9), pp. 1508–1524. DOI: 10.1016/j.molp.2021.05.024.

Ortiz-Bobea, Ariel; Wang, Haoying; Carrillo, Carlos M.; Ault, Toby R. (2019): Unpacking the climatic drivers of US agricultural yields. In *Environ. Res. Lett.* 14 (6), p. 64003. DOI: 10.1088/1748-9326/ab1e75.

Park, Eunsik; Cho, Meehye; Ki, Chang-Seok (2009): Correct use of repeated measures analysis of variance. In *The Korean journal of laboratory medicine* 29 (1), pp. 1–9. DOI: 10.3343/kjlm.2009.29.1.1.

Payne, Samuel H. (2015): The utility of protein and mRNA correlation. In *Trends in biochemical sciences* 40 (1), pp. 1–3. DOI: 10.1016/j.tibs.2014.10.010.

Perez-Riverol, Yasset; Bai, Jingwen; Bandla, Chakradhar; García-Seisdedos, David; Hewapathirana, Suresh; Kamatchinathan, Selvakumar et al. (2022): The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. In *Nucleic acids research* 50 (D1), D543-D552. DOI: 10.1093/nar/gkab1038.

Perni, Stefano; Andrew, Peter W.; Shama, Gilbert (2005): Estimating the maximum growth rate from microbial growth curves: definition is everything. In *Food Microbiology* 22 (6), pp. 491–495. DOI: 10.1016/j.fm.2004.11.014.

Piehowski, Paul D.; Petyuk, Vladislav A.; Orton, Daniel J.; Xie, Fang; Moore, Ronald J.; Ramirez-Restrepo, Manuel et al. (2013): Sources of technical variability in quantitative LC-MS proteomics: human brain tissue sample analysis. In *Journal of proteome research* 12 (5), pp. 2128–2137. DOI: 10.1021/pr301146m.

Pirim, Harun; Ekşioğlu, Burak; Perkins, Andy; Yüceer, Cetin (2012): Clustering of High Throughput Gene Expression Data. In *Computers & operations research* 39 (12), pp. 3046–3061. DOI: 10.1016/j.cor.2012.03.008.

Postmus, Jarne; Tuzun, Işil; Bekker, Martijn; Müller, Wally H.; Teixeira de Mattos, M. Joost; Brul, Stanley; Smits, Gertien J. (2011): Dynamic regulation of mitochondrial respiratory chain efficiency in Saccharomyces cerevisiae. In *Microbiology (Reading, England)* 157 (Pt 12), pp. 3500–3511. DOI: 10.1099/mic.0.050039-0.

Precht, H. (1968): Der Einfluß „normaler" Temperaturen auf Lebensprozesse bei wechselwarmen Tieren unter Ausschluß der Wachstums- und Entwicklungsprozesse. In *Helgolander Wiss. Meeresunters* 18 (4), pp. 487–548. DOI: 10.1007/BF01611681.

Prosser, C. Ladd (1955): PHYSIOLOGICAL VARIATION IN ANIMALS. In *Biological Reviews* 30 (3), pp. 229–261. DOI: 10.1111/j.1469-185X.1955.tb01208.x.

Qi, Yuan; Ge, Hui (2006): Modularity and dynamics of cellular networks. In *PLoS computational biology* 2 (12), e174. DOI: 10.1371/journal.pcbi.0020174.

Rao, Xiaolan; Dixon, Richard A. (2019): Co-expression networks for plant biology: why and how. In *Acta biochimica et biophysica Sinica* 51 (10), pp. 981–988. DOI: 10.1093/abbs/gmz080.

Ritchie, Matthew E.; Phipson, Belinda; Di Wu; Hu, Yifang; Law, Charity W.; Shi, Wei; Smyth, Gordon K. (2015): limma powers differential expression analyses for RNA-sequencing and microarray studies. In *Nucleic acids research* 43 (7), e47. DOI: 10.1093/nar/gkv007.

Rivals, Isabelle; Personnaz, Léon; Taing, Lieng; Potier, Marie-Claude (2007): Enrichment or depletion of a GO category within a class of genes: which test? In *Bioinformatics (Oxford, England)* 23 (4), pp. 401–407. DOI: 10.1093/bioinformatics/btl633.

Roach, Thomas; Sedoud, Arezki; Krieger-Liszkay, Anja (2013): Acetate in mixotrophic growth medium affects photosystem II in Chlamydomonas reinhardtii and protects against photoinhibition. In *Biochimica et biophysica acta* 1827 (10), pp. 1183–1190. DOI: 10.1016/j.bbabio.2013.06.004.

Rost, B.; Liu, J.; Nair, R.; Wrzeszczynski, K. O.; Ofran, Y. (2003): Automatic prediction of protein function. In *Cellular and molecular life sciences : CMLS* 60 (12), pp. 2637–2650. DOI: 10.1007/s00018-003-3114-8.

Royston, P.; Ambler, G.; Sauerbrei, W. (1999): The use of fractional polynomials to model continuous risk variables in epidemiology. In *International journal of epidemiology* 28 (5), pp. 964–974. DOI: 10.1093/ije/28.5.964.

Rutherford, E (1904): The Succession of Changes in Radioactive Bodies. In *Nature* 70 (1807), pp. 161–162. DOI: 10.1038/070161a0.

Salfner, Felix (2006): Modeling Event-driven Time Series with Generalized Hidden Semi-Markov Models. With assistance of Humboldt-Universität zu Berlin.

Schroda, Michael; Hemme, Dorothea; Mühlhaus, Timo (2015): The Chlamydomonas heat stress response. In *The Plant journal : for cell and molecular biology* 82 (3), pp. 466–480. DOI: 10.1111/tpj.12816.

Sharom, Jeffrey R.; Bellows, David S.; Tyers, Mike (2004): From large networks to small molecules. In *Current opinion in chemical biology* 8 (1), pp. 81–90. DOI: 10.1016/j.cbpa.2003.12.007.

Shen, Gangxu (2020): Campbell biology (edited by Lisa Urry, Michael Cain, Steven Wasserman, Peter Minorsky and Jane Reece). In *Journal of biological research (Thessalonike, Greece)* 27 (1), p. 19. DOI: 10.1186/s40709-020-00127-0.

Shi, Dongbo; Jouannet, Virginie; Agustí, Javier; Kaul, Verena; Levitsky, Victor; Sanchez, Pablo et al. (2021): Tissue-specific transcriptome profiling of the Arabidopsis inflorescence stem reveals local cellular signatures. In *The Plant cell* 33 (2), pp. 200–223. DOI: 10.1093/plcell/koaa019.

Simpson, Michael L.; Cox, Chris D.; Allen, Michael S.; McCollum, James M.; Dar, Roy D.; Karig, David K.; Cooke, John F. (2009): Noise in biological circuits. In *Wiley interdisciplinary reviews. Nanomedicine and nanobiotechnology* 1 (2), pp. 214–225. DOI: 10.1002/wnan.22.

Singh, Kehar; Malik, Dimple; Sharma, Naveen (2011): Evolving limitations in K-means algorithm in data mining and their removal 12, pp. 105–109. Available online at https://ijcem.org/papers42011/42011_26.pdf.

Singh, Rahul; Wiseman, Ben; Deemagarn, Taweewat; Jha, Vikash; Switala, Jacek; Loewen, Peter C. (2008): Comparative study of catalase-peroxidases (KatGs). In *Archives of biochemistry and biophysics* 471 (2), pp. 207–214. DOI: 10.1016/j.abb.2007.12.008.

Sloutsky, Roman; Jimenez, Nicolas; Swamidass, S. Joshua; Naegle, Kristen M. (2013): Accounting for noise when clustering biological data. In *Briefings in bioinformatics* 14 (4), pp. 423–436. DOI: 10.1093/bib/bbs057.

Smet, Frank de; Mathys, Janick; Marchal, Kathleen; Thijs, Gert; Moor, Bart de; Moreau, Yves (2002): Adaptive quality-based clustering of gene expression profiles. In *Bioinformatics (Oxford, England)* 18 (5), pp. 735–746. DOI: 10.1093/bioinformatics/18.5.735.

Spaniol, Benjamin; Lang, Julia; Venn, Benedikt; Schake, Lara; Sommer, Frederik; Mustas, Matthieu et al. (2022): Complexome profiling on the Chlamydomonas lpa2 mutant reveals insights into PSII biogenesis and new PSII associated proteins. In *Journal of experimental botany* 73 (1), pp. 245–262. DOI: 10.1093/jxb/erab390.

Spellman, P. T.; Sherlock, G.; Zhang, M. Q.; Iyer, V. R.; Anders, K.; Eisen, M. B. et al. (1998): Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. In *Molecular biology of the cell* 9 (12), pp. 3273–3297. DOI: 10.1091/mbc.9.12.3273.

Strutz, Tilo (2016): Data fitting and uncertainty. A practical introduction to weighted least squares and beyond. 2nd, revised and extended edition. Wiesbaden: Springer Vieweg.

Subedi, Prabal; Moertl, Simone; Azimzadeh, Omid (2022): Omics in Radiation Biology: Surprised but Not Disappointed. In *Radiation* 2 (1), pp. 124–129. DOI: 10.3390/radiation2010009.

Subramanian, Aravind; Tamayo, Pablo; Mootha, Vamsi K.; Mukherjee, Sayan; Ebert, Benjamin L.; Gillette, Michael A. et al. (2005): Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. In *Proceedings of the National Academy of Sciences of the United States of America* 102 (43), pp. 15545–15550. DOI: 10.1073/pnas.0506580102.

Tang, Fuchou; Lao, Kaiqin; Surani, M. Azim (2011): Development and applications of single-cell transcriptome analysis. In *Nature methods* 8 (4 Suppl), S6-11. DOI: 10.1038/nmeth.1557.

Tavazoie, S.; Hughes, J. D.; Campbell, M. J.; Cho, R. J.; Church, G. M. (1999): Systematic determination of genetic network architecture. In *Nature genetics* 22 (3), pp. 281–285. DOI: 10.1038/10343.

Thimm, Oliver; Bläsing, Oliver; Gibon, Yves; Nagel, Axel; Meyer, Svenja; Krüger, Peter et al. (2004): MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. In *The Plant journal : for cell and molecular biology* 37 (6), pp. 914–939. DOI: 10.1111/j.1365-313x.2004.02016.x.

Torday, John S. (2015): Homeostasis as the Mechanism of Evolution. In *Biology* 4 (3), pp. 573–590. DOI: 10.3390/biology4030573.

Tuteja, Narendra; Singh Gill, Sarvajeet (2013): Plant Acclimation to Environmental Stress. New York, NY: Springer New York.

Tyson, J. J.; Chen, K.; Novak, B. (2001): Network dynamics and cell physiology. In *Nature reviews. Molecular cell biology* 2 (12), pp. 908–916. DOI: 10.1038/35103078.

Uhlén, Mathias; Fagerberg, Linn; Hallström, Björn M.; Lindskog, Cecilia; Oksvold, Per; Mardinoglu, Adil et al. (2015): Proteomics. Tissue-based map of the human proteome. In *Science (New York, N.Y.)* 347 (6220), p. 1260419. DOI: 10.1126/science.1260419.

Usadel, Björn; Poree, Fabien; Nagel, Axel; Lohse, Marc; Czedik-Eysenberg, Angelika; Stitt, Mark (2009): A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, Maize. In *Plant, cell & environment* 32 (9), pp. 1211–1229. DOI: 10.1111/j.1365-3040.2009.01978.x.

van Impe, J. F.; Vercammen, D.; van Derlinden, E. (2013): Toward a next generation of predictive models: A systems biology primer. In *Food Control* 29 (2), pp. 336–342. DOI: 10.1016/j.foodcont.2012.06.019.

Venn, Benedikt; Leifeld, Thomas; Zhang, Ping; Mühlhaus, Timo (2024): Temporal classification of short time series data. In *BMC bioinformatics* 25 (1). DOI: 10.1186/s12859-024-05636-6.

Venn, Benedikt; Mühlhaus, Timo (2022): CSBiology/OntologyEnrichment: Release 0.0.1: Zenodo.

Venn, Benedikt; Mühlhaus, Timo; Kevin Schneider; Lukas Weil; David Zimmer; Selina Ziegler et al. (2022): fslaborg/FSharp.Stats: Release 0.4.7: Zenodo.

Villas-Bôas, Silas G.; Roessner, Ute; Hansen, Michael A. E.; Smedsgaard, Jørn; Nielsen, Jens (2007): Metabolome Analysis: Wiley.

Vogel, Christine; Marcotte, Edward M. (2012): Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. In *Nature reviews. Genetics* 13 (4), pp. 227–232. DOI: 10.1038/nrg3185.

Vogels, Maurice; Zoeckler, Rita; Stasiw, Donald M.; Cerny, Lawrence C. (1975): P. F. Verhulst's “notice sur la loi que la populations suit dans son accroissement” from correspondence mathematique et physique. Ghent, vol. X, 1838. In *J Biol Phys* 3 (4), pp. 183–192. DOI: 10.1007/BF02309004.

Vonk, Jennifer; Shackelford, Todd (Eds.) (2019): Encyclopedia of Animal Cognition and Behavior. Cham: Springer International Publishing.

Vonk, Jennifer; Shackelford, Todd K. (2022): Encyclopedia of Animal Cognition and Behavior. Cham: Springer International Publishing.

Wang, Kui; Huang, Canhua; Nice, Edouard Collins (2014): Proteomics, genomics and transcriptomics: their emerging roles in the discovery and validation of colorectal cancer biomarkers. In *Expert review of proteomics* 11 (2), pp. 179–205. DOI: 10.1586/14789450.2014.894466.

Wang, Zhong; Gerstein, Mark; Snyder, Michael (2009): RNA-Seq: a revolutionary tool for transcriptomics. In *Nature reviews. Genetics* 10 (1), pp. 57–63. DOI: 10.1038/nrg2484.

Warren Liao, T. (2005): Clustering of time series data—a survey. In *Pattern Recognition* 38 (11), pp. 1857–1874. DOI: 10.1016/j.patcog.2005.01.025.

Weaver, Daniel S.; Keseler, Ingrid M.; Mackie, Amanda; Paulsen, Ian T.; Karp, Peter D. (2014): A genome-scale metabolic flux model of Escherichia coli K-12 derived from the EcoCyc database. In *BMC systems biology* 8, p. 79. DOI: 10.1186/1752-0509-8-79.

Weil, Heinrich Lukas; Schneider, Kevin; Tschöpe, Marcel; Bauer, Jonathan; Maus, Oliver; Frey, Kevin et al. (2023): PLANTdataHUB: a collaborative platform for continuous FAIR data sharing in plant research. In *The Plant journal : for cell and molecular biology* 116 (4), pp. 974–988. DOI: 10.1111/tpj.16474.

West, G. C. (1972): The effect of acclimation and acclimatization on the resting metabolic rate of the common redpoll. In *Comparative biochemistry and physiology. A, Comparative physiology* 43 (2), pp. 293–310. DOI: 10.1016/0300-9629(72)90188-0.

Wilhelm, Mathias; Hahne, Hannes; Savitski, Mikhail; Marx, Harald; Lemeer, Simone; Bantscheff, Marcus; Kuster, Bernhard (2017): Wilhelm et al. reply. In *Nature* 547 (7664), E23. DOI: 10.1038/nature22294.

Williams, Alexander G.; Thomas, Sean; Wyman, Stacia K.; Holloway, Alisha K. (2014): RNA-seq Data: Challenges in and Recommendations for Experimental Design and Analysis. In *Current protocols in human genetics* 83, 11.13.1-20. DOI: 10.1002/0471142905.hg1113s83.

Winkler, Hans (1920): Verbreitung und Ursache der Parthenogenesis im Pflanzen- und Tierreiche / von Dr. Hans Winkler. Jena: G. FIscher.

Wood, S. N. (2013): A simple test for random effects in regression models. In *Biometrika* 100 (4), pp. 1005–1010. DOI: 10.1093/biomet/ast038.

Xing, Shuping; Wallmeroth, Niklas; Berendzen, Kenneth W.; Grefen, Christopher (2016): Techniques for the Analysis of Protein-Protein Interactions in Vivo. In *Plant physiology* 171 (2), pp. 727–758. DOI: 10.1104/pp.16.00470.

Xu, Shuo; Qiao, Xiaodong; Zhu, Lijun; Zhang, Yunliang; Xue, Chunxiang; Li, Lin (2016): Reviews on Determining the Number of Clusters. In *Appl. Math. Inf. Sci.* 10 (4), pp. 1493–1512. DOI: 10.18576/amis/100428.

Yates, John R.; Ruse, Cristian I.; Nakorchevsky, Aleksey (2009): Proteomics by mass spectrometry: approaches, advances, and applications. In *Annual review of biomedical engineering* 11, pp. 49–79. DOI: 10.1146/annurev-bioeng-061008-124934.

Yin, Anwen (2022): Does the kitchen-sink model work forecasting the equity premium? In *Int Rev Finance* 22 (1), pp. 223–247. DOI: 10.1111/irfi.12352.

Zaza, Gianluigi; Neri, Flavia; Bruschi, Maurizio; Granata, Simona; Petretto, Andrea; Bartolucci, Martina et al. (2023): Proteomics reveals specific biological changes induced by the normothermic machine perfusion of donor kidneys with a significant up-regulation of Latexin. In *Scientific reports* 13 (1), p. 5920. DOI: 10.1038/s41598-023-33194-z.

Zhang, Ningning; Venn, Benedikt; Bailey, Catherine E.; Xia, Ming; Mattoon, Erin M.; Mühlhaus, Timo; Zhang, Ru (2023): Moderate High Temperature is Beneficial or Detrimental Depending on Carbon

Availability in the Green Alga Chlamydomonas reinhardtii. In *Journal of experimental botany*. DOI: 10.1093/jxb/erad405.

Zhang, Yi; Wolf-Yadlin, Alejandro; Ross, Phillip L.; Pappin, Darryl J.; Rush, John; Lauffenburger, Douglas A.; White, Forest M. (2005): Time-resolved mass spectrometry of tyrosine phosphorylation sites in the epidermal growth factor receptor signaling network reveals dynamic modules. In *Molecular & cellular proteomics : MCP* 4 (9), pp. 1240–1250. DOI: 10.1074/mcp.M500089-MCP200.

Zhao, Yingdong; Li, Ming-Chung; Konaté, Mariam M.; Chen, Li; Das, Biswajit; Karlovich, Chris et al. (2021): TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. In *Journal of translational medicine* 19 (1), p. 269. DOI: 10.1186/s12967-021-02936-w.

Zimmer, David; Schneider, Kevin; Sommer, Frederik; Schroda, Michael; Mühlhaus, Timo (2018): Artificial Intelligence Understands Peptide Observability and Assists With Absolute Protein Quantification. In *Frontiers in plant science* 9, p. 1559. DOI: 10.3389/fpls.2018.01559.

Zimmer, David; Swart, Corné; Graf, Alexander; Arrivault, Stéphanie; Tillich, Michael; Proost, Sebastian et al. (2021): Topology of the redox network during induction of photosynthesis as revealed by time-resolved proteomics in tobacco. In *Science advances* 7 (51), eabi8307. DOI: 10.1126/sciadv.abi8307.

10. Appendix

Article I: Temporal classification of short time series data

Benedikt Venn, Thomas Leifeld, Ping Zhang, Timo Mühlhaus

| Bioinformatics, 2024, 10.1186/s12859-024-05636-6

Venn *et al. BMC Bioinformatics* (2024) 25:30
https://doi.org/10.1186/s12859-024-05636-6

BMC Bioinformatics

RESEARCH **Open Access**

Temporal classification of short time series data

Benedikt Venn[1], Thomas Leifeld[2], Ping Zhang[2] and Timo Mühlhaus[1]*

*Correspondence:
timo.muehlhaus@rptu.de

[1] Computational Systems Biology, RPTU Kaiserslautern, 67663 Kaiserslautern, Germany
[2] Institute of Automatic Control, RPTU Kaiserslautern, 67663 Kaiserslautern, Germany

Abstract

Motivation: Within the frame of their genetic capacity, organisms are able to modify their molecular state to cope with changing environmental conditions or induced genetic disposition. As high throughput methods are becoming increasingly affordable, time series analysis techniques are applied frequently to study the complex dynamic interplay between genes, proteins, and metabolites at the physiological and molecular level. Common analysis approaches fail to simultaneously include (i) information about the replicate variance and (ii) the limited number of responses/shapes that a biological system is typically able to take.

Results: We present a novel approach to model and classify short time series signals, conceptually based on a classical time series analysis, where the dependency of the consecutive time points is exploited. Constrained spline regression with automated model selection separates between noise and signal under the assumption that highly frequent changes are less likely to occur, simultaneously preserving information about the detected variance. This enables a more precise representation of the measured information and improves temporal classification in order to identify biologically interpretable correlations among the data.

Availability and implementation: An open source F# implementation of the presented method and documentation of its usage is freely available in the *TempClass* repository, https://github.com/CSBiology/TempClass [58].

Keywords: Time series analysis, Smoothing spline, Profile classification, Omics analyis

Introduction

Biological systems are constantly regulating their genes, proteins, and metabolites to maintain an optimal internal state. Optimal, however, is context-dependent and contingent upon prevailing environmental factors. Disturbances such as alterations in light, temperature, moisture, or mineral concentrations necessitate metabolic adjustments to mitigate potential stress and restore optimal conditions according to the external influence. These acclimation responses are meticulously orchestrated and follow a defined sequence. Their primary objectives are to mitigate the adverse effects of unfavourable environmental conditions or optimally exploit positive alterations.

With the declining costs associated with high throughput technologies like RNA-Seq and MS proteomics, the utilization of time series analyses has gained popularity as a valuable tool to study the kinetics/temporal dynamics of biological molecules. Nonetheless, challenges complicate comprehensive analyses of such data. Time series datasets often comprise a limited number of measurement points, and due to the substantial investment required in growing biological material and the still relatively high costs of these analyses, experiments are typically designed with a modest number of measurement points (typically 4 to 8) and a few replicates (2–4).

When characterizing the cellular response characteristics for individual molecules, it is imperative to assign lower significance to measurement points characterized by elevated uncertainty [1]. While this assignment is often intuitive when performed manually, an automated evaluation method necessitates the explicit incorporation of this consideration.

Simultaneously, the biological response capacity is constrained. High amplitude fluctuations are improbable from a regulatory standpoint, and they would entail substantial synthesis and degradation costs for the biological system in question. In the absence of new stimuli, one can reasonably anticipate smooth kinetics in biological molecules, especially for more complex molecules like proteins. This assumption provides valuable additional information that enhances the precision and utility of biological models. However, it mandates the development of novel analytical techniques to effectively incorporate such information.

Our proposed approach addresses the dual challenges of variable measurement uncertainties and the expectation of low signal fluctuation. We have employed smoothing splines, which impose continuity in function, slope, and curvature, while also permitting the weighting of individual measurement points and the imposition of shape constraints. Unlike existing methods that use predefined profiles [2], or require a preselected number of clusters [3], our approach classifies the data where a single classification is uncoupled from the remaining data.

Methods

Time series data can be assumed as functions of time with superimposed heteroscedastic biological variance and technical noise [4].

$$y_i = f(t_i) + \varepsilon_i, i = 1, \ldots, n$$

where f(·) resembles the function of the true abundance time course at the ith of n time points. The error term ε_i combines biological variation and noise introduced by sample processing and measurement devices, leading to the blurring of the true relationship f(·) to the final reading y at time point i. The interval widths between measuring time points are defined by:

$$h_i = y_{i+1} - y_i, i = 1, \ldots, n - 1$$

While in many analysis strategies, e.g. common clustering procedures or statistical testing frameworks, interval widths are not taken into account, they may possess valuable information regarding the dynamic of the underlying kinetic [5, 6]. High amplitude changes within a short time period can be considered unlikely and thus penalized by the model. However, for biological regulatory responses, the model time point spacing may vary from the actual experimental time point spacing as discussed below.

Time point spacing

As indicated by its name, the independent variable is time (e.g. hours since experiment start). For experiments with no perturbation, these time intervals can be directly used for curve fitting. Cell cycle regulation may be measured with fixed time intervals because the expected rate of change is evenly distributed between the time points. If the biological system, however, faces sudden condition perturbation, the spacing according to time intervals is insufficient for modelling. A perturbation causes the biological system to react immediately. This regulation of molecular processes has to occur quickly with regulation in later time points being less fluctuating. To account for this asymmetric regulation, samples are taken according to the expected rate of system response. The measurements of the presented experiment were taken by doubling the time interval at each measurement. This is according to the estimated change apparent within two measurements. Hence, samples can be spaced uniformly in time. The presented approach is not restricted to uniformly spaced time points and works with any univariate time series.

Smoothing spline

To investigate the underlying kinetic, the application of smoothing splines offers a valuable approach to model the data, striking a balance between data fidelity and the smoothness of the fit. While various other fitting techniques exist, i.e. interpolation strategies or (non-) linear regression, most of them rely on a predefined template function or ignore point uncertainty as they are forced to interpolate the sample means. Further explanations for choosing smoothing splines over other fitting techniques are given in the discussion. In this method, piecewise cubic polynomials are employed to model each subinterval of the data while smooth transitions are ensured by enforcing the equality of function values, slopes, and curvatures at each designated knot. Considering that knots are positioned at each time point, there are $n-1$ intervals to analyze.

Splines within the interval $[t_i, t_{i+1}]$ are defined as

$$f_i(t) = a_i\phi_{0i}t + a_{i+1}\phi_{1i}t + c_i\gamma_{0i}t + c_{i+1}\gamma_{1i}t \tag{1}$$

with $a_i = f_i(t_i)$, $c_i = f''_i(t_i)$, and basis functions ϕ_0, ϕ_1, γ_0, and γ_1 defined in [7] and listed in the supplement. The vector a is the vector that only contains the function values at the measured time points, c contains the function curvature (second derivative) at the measured time points and the basis functions describe how the adjacent knots influence the curve shape between them.

The estimation of spline segments is subject to the minimization of the following cost function [7, 8]:

$$\int_{t_1}^{t_n} \left[f''(t)\right]^2 dt + \frac{\lambda}{n}\|W(a-y)\|^2, W_{i,i} = w_i \tag{2}$$

W is a diagonal matrix of observation weights introduced in Eq. 10, a is a rowvector of the splines function values at the knots, y is a rowvector of the observation values, and

$|| \cdot ||$ denotes the Euclidean norm. While the first term serves as a roughness penalty, the second term ensures the required fidelity to the data [9]. A smoothing factor λ mediates between these opposing error terms. When $\lambda = 0$, the resulting spline results in a straight least squares regression line, while $\lambda \to \infty$ leads to an interpolating cubic spline.

In-depth spline theory is given in [8–12]. The minimization of Eq. 2 can be rewritten as a quadratic optimization problem [7].

$$\min_{a} \frac{1}{2} a^T G_\lambda a + c_\lambda^T a, \tag{3}$$

with $G_\lambda = 2\left(H^T D^{-1} H + \frac{\lambda}{n} W^T W\right)$ and $c_\lambda = -2\frac{\lambda}{n} y^T W^T W$. Band matrices D and H are defined as:

$$D_{n-2\times n-2} = \begin{bmatrix} d_1^a & d_2^b & 0 & 0 & \cdots & 0 & 0 & 0 \\ d_2^b & d_2^a & d_3^b & 0 & \cdots & 0 & 0 & 0 \\ 0 & d_3^b & d_3^a & d_4^b & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & d_{n-3}^b & d_{n-3}^a & d_{n-2}^b \\ 0 & 0 & 0 & 0 & \cdots & 0 & d_{n-2}^b & d_{n-2}^a \end{bmatrix} \tag{4}$$

$$d_i^a = (h_i + h_{i+1})/3 \tag{5}$$

$$d_i^b = h_i/6 \tag{6}$$

$$H_{n-2\times n} = \begin{bmatrix} e_1^b & e_1^a & e_2^b & 0 & \dots & 0 & 0 & 0 \\ 0 & e_2^b & e_2^a & e_3^b & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & e_{n-2}^b & e_{n-2}^a & e_{n-1}^b \end{bmatrix} \tag{7}$$

$$e_i^a = -(h_i + h_{i+1})/(h_i \cdot h_{i+1}) \tag{8}$$

$$e_i^b = 1/h_i \tag{9}$$

Measurement weighting

A crucial part of spline smoothing is the determination of the weighting matrix W. For each signal, the time point weighting w_i relies on the signal's standard deviation that is divided by the average standard deviation of all time points. As smoothing splines—under the given smoothness constraints—aim to minimize the distance of the original data points to the resulting prediction (sum of squares), outlier values would negatively impact the prediction function. As time points with outliers often are affected by high uncertainty, this variance can be exploited to reduce the outlier impact. When calculating the sum of squares, high variances are encoded as low weights that reduce the impact of points, which should have reduced influence on the fit (Eq. 2).

$$W = \begin{bmatrix} w_1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & w_i & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & w_n \end{bmatrix}, w_i = \frac{w_{init}(i)}{\overline{w_{init}}} \tag{10}$$

$$w_{init}(i) = \sigma_{y_i} \tag{11}$$

Shape constraints

To control the spline's shape, monotonicity can be enforced for each interval separately. To obtain a single maximum within interval h_i, combinations of monotonicity constraints are applied, so that intervals within $[t_0, t_{i-1}]$ are monotonically increasing, and intervals in $[t_{i+1}, t_n]$ are monotonically decreasing. Monotonicity constraints are well studied and can be enforced by the following conditions [13–16].

For monotonically increasing polynomials in $[t_i, t_{i+1}]$ ensure that

$$a_{i+1} - a_i \geq 0 \tag{12}$$

$$f'_i \geq 0 \tag{13}$$

$$f'_{i+1} \geq 0 \tag{14}$$

$$3\frac{a_{i+1} - a_i}{h_i} - f'_i \geq 0 \tag{15}$$

$$3\frac{a_{i+1} - a_i}{h_i} - f'_i \geq 0 \tag{16}$$

For monotonically decreasing polynomials signs are switched from $\geq$ to $\leq$. The derivative f'_i is calculated from $B_i a$ where B is a $n \times n$ Matrix defined by $B = P^{-1}U$. P and U are described in [7] and can be seen in the supplement. The conditions for every interval are summarized in a constraint matrix $Ca \geq [0]$ representing linear inequality constraints. When solving (12–16) with respect to $Ca \geq [0]$ some constraints will be satisfied as equality constraints. This constraint set where $C_A a = [0]$ is termed 'Active set'. Matrix Z which columns form a basis for the null space of C_A is used for determining the smoothing factor λ [7, 17]. Using smoothing splines in combination with the described monotonicity constraints allows the construction of smooth curves with oscillations allowed in specified regions that are not constrained to be monotone.

For every signal to fit, several monotonicity constraints can be applied, resulting in a range of potential parent shapes. These parent shapes span a spectrum from monotonically in- or decreasing curves to those featuring 1, 2, 3, or 4 extrema, with each type starting either with a maximum or a minimum. This results in a total of 10 distinct parent shapes, along with a single unconstrained scenario, that can be applied to a single time series signal.

Under the specified slope constraints, the case of monotonically increasing or decreasing curves offers only one shape, while curves containing at least one extremum necessitate the consideration of multiple shape possibilities (as summarized in Additional file 1: Table S1). Consequently, it is necessary to fit curves corresponding to all conceivable shapes to each signal. Subsequently, the most appropriate shape is selected as the descriptor for elucidating the underlying molecule kinetics.

Model selection

Besides the data points with an associated weighting matrix, smoothing splines rely on a smoothness parameter λ that controls how curved the resulting curve is going to be. The smoothing strength must be determined individually for each signal. The ideal λ value is estimated by minimizing the modified generalized cross validation (mGCV), which is an estimate for the model prediction error and—in contrast to other cross validation techniques—only requires a single passage. A minimal mGCV hints at the optimal compromise between over- and underfitting respectively [12, 18, 19].

$$\underset{\lambda}{min}\, n\frac{\|W(a-y)\|^2}{(Tr(I-\rho A_{\lambda}))^2} \tag{17}$$

Matrix I is a $n \times n$ identity matrix and $A_{\lambda} = 2\frac{\lambda}{n}Z(Z^TG_{\lambda}Z)^{-1}Z^TW^TW$ an influence matrix, so that $a = Ay$ and $\rho = 1.3$ as compensatory factor for small sample sizes [7, 8, 20]. Because the minimization of Eq. 17 with respect to Eq. 3 leads to a non-convex and non-continuous optimization problem, a grid search approach is used to choose from a wide range of λ.

For every parent shape, a single candidate is reported and selected by minimizing mGCV. When a maximum of four extrema is allowed, this results in 11 final fits to choose from ((i) 5 fits with 0–4 extrema starting with a positive slope, (ii) 5 fits with 0–4 extrema starting with a negative slope, and (iii) an unconstrained spline that has no assumptions regarding its monotonicity). While mGCV proves suitable for making reasonable determinations of the smoothing strength within the parent shape class, it encounters challenges when confronted with the task of selecting the correct shape from the remaining 11 shape options. As mGCV does not include information about the number of allowed extrema it tends to favour fits with increased flexibility over conservative ones.

An adapted version of the Akaike information criterion (AIC) is used to select the final shape [21]. This modified approach incorporates a correction factor tailored for cases with limited sample sizes (AICc) as originally proposed in [22] and a term that incorporates the number of extrema present in the curve.

$$AIC_c = n \cdot \ln\left(\frac{\|W(a-y)\|^2}{n}\right) + 2k + \frac{2k^2+2k}{n-k-1} \tag{18}$$

Here k equals the enforced number of extrema.

Among all shape assumptions, the model that minimizes AIC_c (Eq. 18) is assumed to represent the underlying function at best and is used for temporal classification.

Extrema extraction and classification

It is trivial to identify the splines extrema, since all polynomial coefficients can easily be obtained for every interval from function values a and their second derivatives c at adjacent knots. Polynomial template:

$$s(t) = k_1 t^3 + k_2 t^2 + k_3 t + k_4, \tag{19}$$

with $s(t_i) = a_i$, $s(t_{i+1}) = a_{i+1}$, $s''(t_i) = c_i$, and $s''(t_{i+1}) = c_{i+1}$

Basic calculus leads to the polynomial coefficients:

$$k_1 = -\frac{1}{6}(c_i + c_{i+1}) \tag{20}$$

$$k_2 = \frac{1}{2}c_i \tag{21}$$

$$k_3 = -\frac{1}{3}c_i - \frac{1}{6}c_{i+1} - a_i + a_{i+1} \tag{22}$$

$$k_4 = a_i \tag{23}$$

Extreme points are determined by setting $s'(t) = 0$ and $s''(t) < 0$ for maxima and $s''(x) > 0$ for minima within $[t_i, t_{i+1}]$. Additional file 1: Fig. S6 gives a visual impression of the extrema extraction process. The location of extreme points is used to group similar shaped time series. Although the location of extreme values is determined by the model, it is necessary to determine the exact position after a model is fitted to the data. This downstream identification of extrema enables previous filtering of quasi-constant signals and ensures that extrema can be assigned to their nearest knots. The classifier may result in \`Min3,Max4\` indicating the spline having a minimum at the third time point and a maximum at the fourth time point. If there is no extremum, the classifier results in either \`I\` or \`D\`, depending on whether the spline is monotonically in- or decreasing. If necessary, these two monotone classes may be further refined using the second derivative to identify prominent changes in curvature (plateau regions).

Iterative clustering

As a common time-series clustering procedure, k-means clustering is used and compared to the presented classification method. The k-means algorithm iteratively recalculates the position of k initial centroids. All points are assigned to the nearest centroid, while the squared Euclidean distance is used as distance measurement. Convergence is reached when no reallocation of points to different centroids occurs [23, 24].

The optimal cluster number k is determined by the gap statistics method [25].

If relative changes of two signal slopes behave the same but their absolute values differ, most distance measures would report high distances even if their parallel change would suggest a difference of 0 (or similarity of 1). To prevent dominance of variables with differing ranges a normalization step often is needed to equalize all amplitudes. A popular normalization is the z-score, which transforms a single time series to have zero mean and unit variance by

$$y'_k = \frac{y_k - \mu}{\sigma} \tag{24}$$

when μ denotes the arithmetic mean and σ is the standard deviation of the given data y_k [26]. This analysis was performed using FSharp.Stats v0.5.0 [27].

Comparison of leave one out cross validation

In order to assess the robustness of the used constrained smoothing spline, protein time series were fitted using four different curve fitting methods: (i) constrained smoothing spline as presented in this work, (ii) polynomial interpolation of sample averages, (iii) linear spline interpolation of sample averages, (iv) cubic spline interpolation of sample averages. All but the constrained smoothing spline procedures are available at FSharp. Stats v0.5.0. For leave one out cross correlation all three replicates of an internal sample were deleted from the time series.

After deletion all four fitting procedures were applied to the modified time series. The distance of the prediction at the time point of missing data to the original prediction is determined. For each protein signal, this leads to 6 distances. The same procedure was applied using the distance from the prediction at the time point of missing data to the sample mean (Additional file 1: Figure S1).

Visualization

All visualizations presented in this manuscript were prepared using Plotly.NET v4.0.0 [28].

Enrichment analysis

Gene set enrichment analysis is performed using extended MapMan annotations (for example "PS.lightreaction.LHC" becomes "PS.lightreaction.LHC", "PS.lightreaction" and "PS") [29–31]. All 1292 proteins served as background if overrepresentation of functional annotations is studied within classes. Enrichment was performed using multiple hypergeometric tests while p values were corrected for multiple testing using Storey's q value method [32, 33].

Results

Our approach assumes biological molecule kinetics have the intrinsic constraint to avoid curvature and therefore be smooth. To underpin this assumption with data, a time series dataset with 12 time points was analysed, which were analysed with 7 replicates each. Naturally, the more replicates measured, the more accurate the abundance estimator is supposed to be. To validate the smoothness assumption 2, 3, 5, and 7 replicates of each protein were randomly selected and analysed. The signals were interpolated with both linear splines and cubic splines, and the slope (linear spline) and curvature (cubic spline) at each knot were extracted. Both slopes and curvatures of the signals decrease with increasing number of replicates (Fig. 1). The higher the measurement accuracy (increased number of replicates), the higher

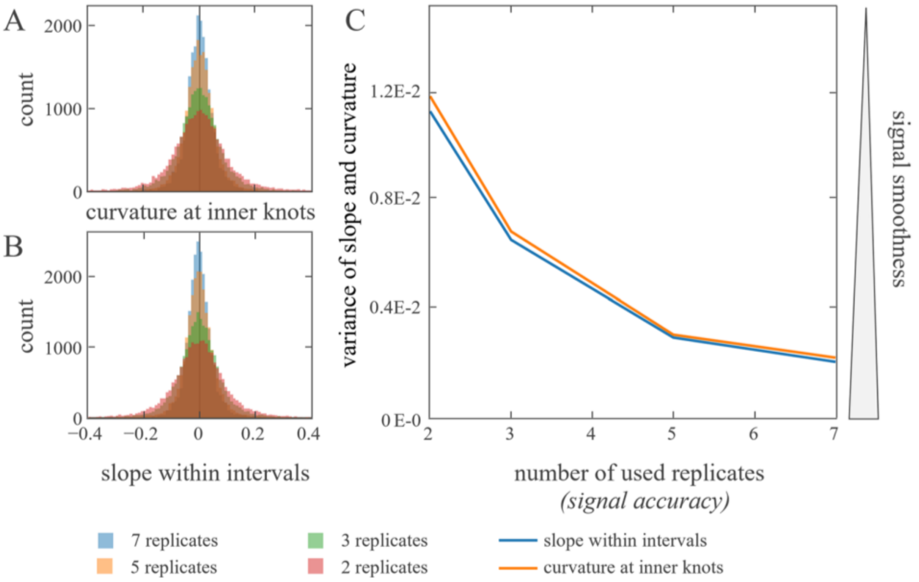


Fig. 1 In a circadian time series proteomics experiment, protein intensities were measured every four hours for two days (PXD019431). To investigate the smoothness of the signal, replicates at each time point were shuffled randomly and reduced to a replicate number given in the legend. In each data reduction iteration curvature and slopes were determined. **A**: Protein intensity means at 12 time points were interpolated with cubic splines using natural boundary conditions [27]. The second derivative of the spline at the inner knots was determined as smoothness measure. **B**: Protein intensity means at 12 time points were interpolated with linear splines. The first derivative of the lines within the knot intervals was determined as smoothness measure. **C**: Variances of curvatures **A** and slopes **B** of the signal interpolation are calculated and plotted against the number of used replicates

the signals smoothness (reduced variance in slope and curvature) supporting the assumption of biological molecule kinetics showing the tendency of being smooth.

Based on the constrained smoothing splines, we classified protein abundances from a heat acclimation experiment conducted in the green algae *Chlamydomonas reinhardtii* [34]. The utilized data subset consists of measurements taken at 8 time points during 40 °C heat treatment (0 h, 0.5 h, 1 h, 2 h, 4 h, 8 h, 16 h, and 24 h respectively). 1,292 proteins are measured in three biological replicates. An exemplary comparison of four fitting methods is given in Fig. 2. Constrained smoothing splines ensure smooth curves while being flexible enough to recognize relevant changes.

The signals were classified by the location of their extrema as described in *Extrema extraction and classification*. For 8 measured time points, 327 possible curve configuration possibilities (shapes) exist (Additional file 1: Table S1).

The analysis of the smoothed protein signals resulted in 77 classes to be filled with at least one protein signal. As expected, the classes become smaller with increasing specificity (Fig. 3). Classes with a high number of extreme points must have strong evidence of such behaviour in the form of low variances and high amplitude differences. The smoothness constraint therefore leads to less class occupancies the more complex the class gets, despite the fact, that the shape possibilities increase with more extrema (Fig. 3B).

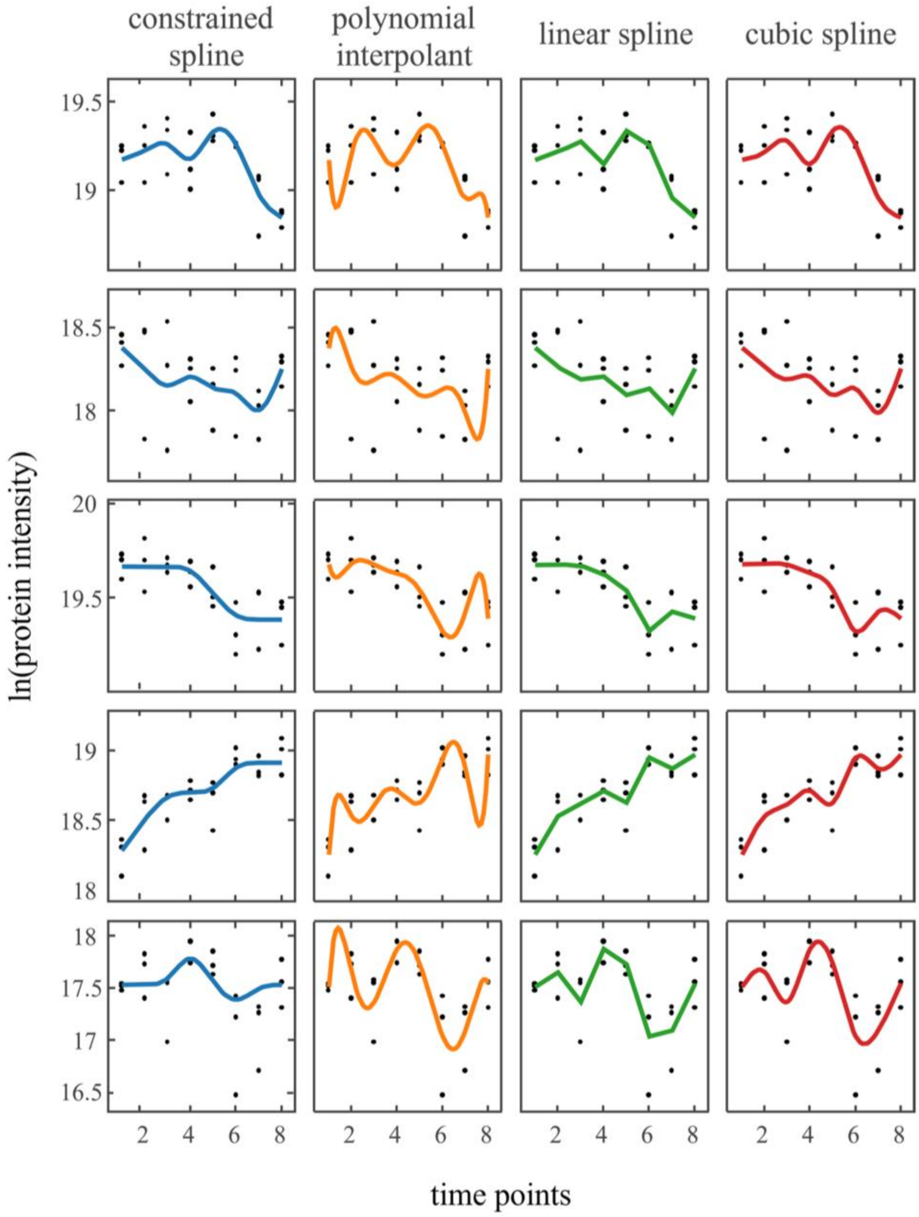


Fig. 2 Four curve fitting approaches on five proteins. Five exemplary protein abundance signals (top to bottom) modelled using four approaches: (i) constrained smoothing spline, (ii) polynomial interpolation of arithmetic means, (iii) linear spline interpolation of arithmetic means, and (iv) cubic interpolating splines with natural boundary conditions

Determination of fit robustness

To compare the use of constrained smoothing splines with other fitting methods, the data were cross-validated as explained in *Comparison of leave one out cross validation*. All replicates of an inner time point were removed, and the remaining time points fitted with the presented constrained smoothing spline, interpolating polynomial, linear spline, and cubic spline. The distance of the predicted value at the time of the missing data to the original prediction serves as a robustness measure. If a data point is deleted from the time series, the model curve may change in shape.

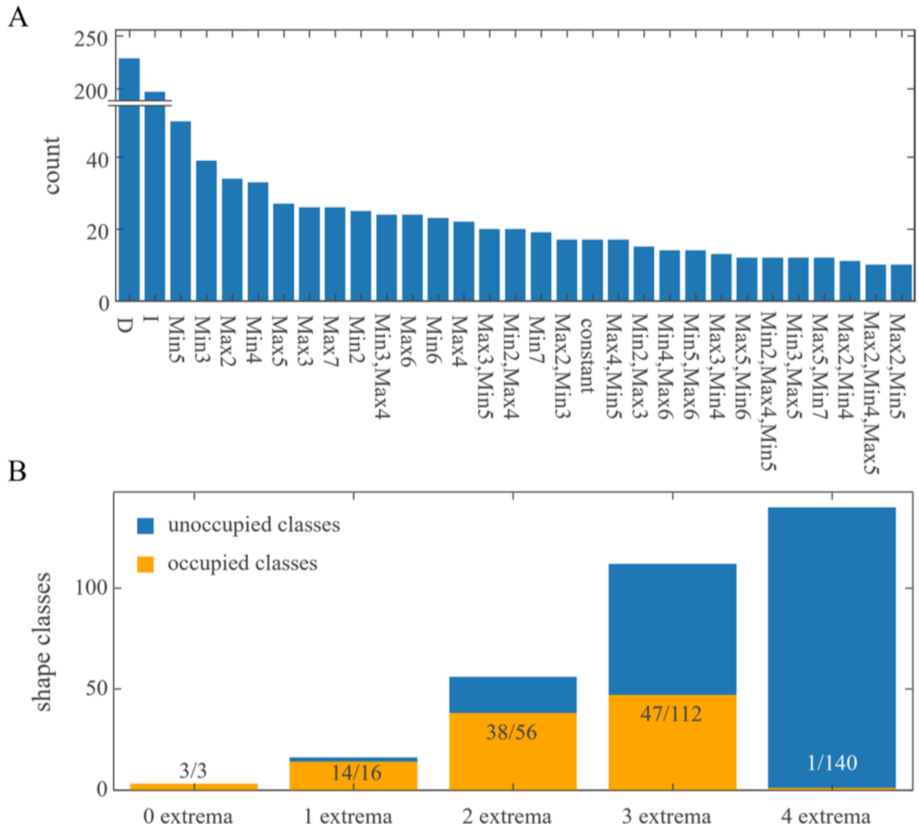


Fig. 3 Class occupancy **A** All classes that had more than 10 proteins assigned are depicted. Min3 indicates a single minimum at the third time point. Min3, Max4 indicates a minimum at the third time point followed by a maximum at the fourth time point. **B** Number of possible classes (blue) and number of occupied classes in the presented data set. No extremum is present in constant, monotonically in- and decreasing signals

The higher this change, the more prone to overfitting a model is. Simultaneously high distances indicate a high influence of the particular point for the model. The protein signals range from 16 to 24 with a median standard deviation of 0.156. As expected, polynomial interpolation leads to massive overfitting (compare curve shapes in Fig. 2). With increasing variance at the point of interest, the overfitting tendencies of linear and cubic splines tend to increase (Fig. 4B, C).

Especially when variance is high in the missing time point replicates, the constrained spline assigned lower weightings to this point and shows a reduced distance to the original curve (Fig. 4). To examine whether this robustness is solely due to a conservative fitting, the same procedure was performed using the distance of the prediction of the sparse data at the time point of missing replicates to the original sample mean instead of the original prediction. Low distances in this measure would hint at underfitting, indicating that the model isn't at all influenced by the signal manipulation. Because the polynomial as well as the linear and cubic spline interpolates the mean, the distances stay the same. The constrained spline shows similar distances as the other fitting techniques, indicating a comparable fidelity to the data and not giving suspicion for underfitting tendencies (Additional file 1: Fig. S2).

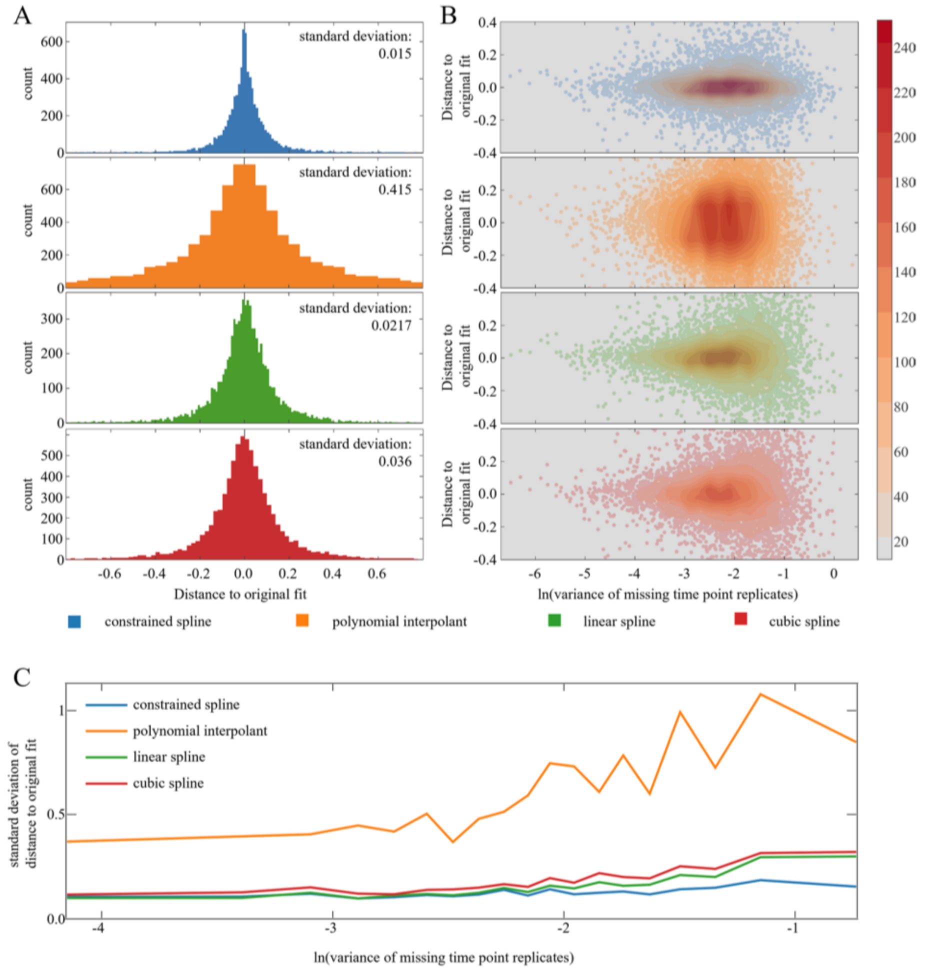


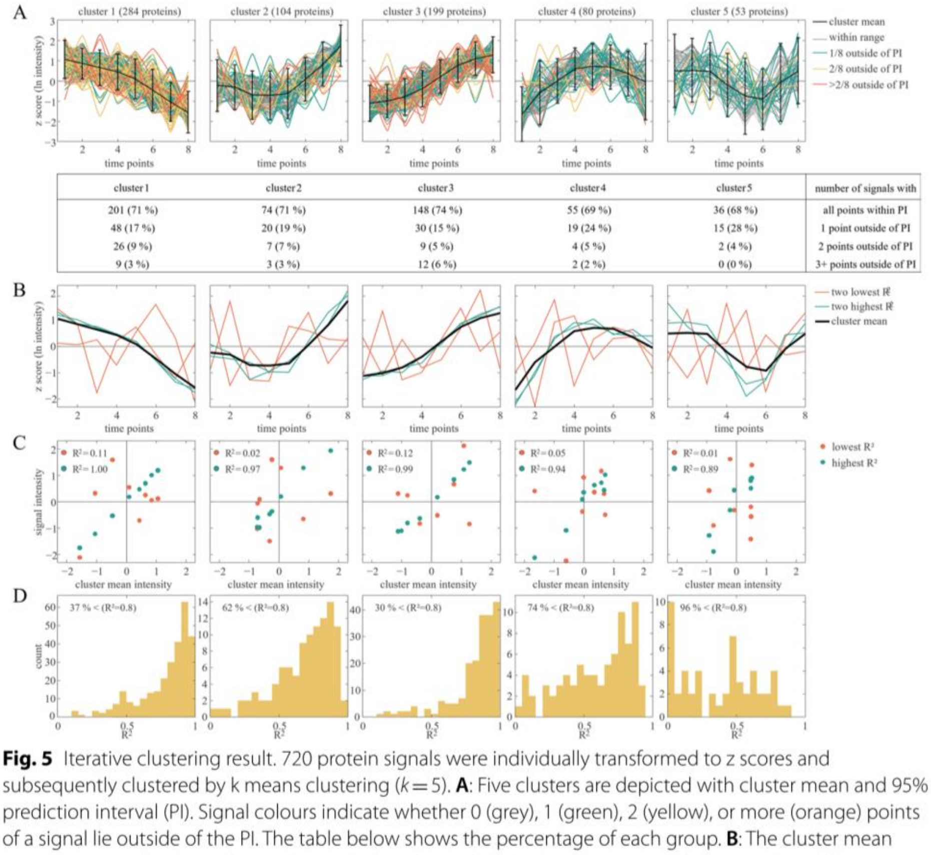
Fig. 4 Robustness analysis. **A**: Four fitting techniques were applied to each protein signal. After deleting every inner time point once, the distance from the original prediction to the prediction using the sparse signal is measured. For each protein, 6 distances are reported (number of inner time points) and summarized in a histogram. The histogram's standard deviation is given in the top right. **B**: The same data was used as in **A** but additionally separated by the variance of the time point replicates that were deleted. **C**: The distance data is separated in 20 equally large bins depending on the variance of the missing time point replicates (**B** x axis). The standard deviation of the distances within each bin is calculated and plotted against the average time point variance

Comparison to clustering approaches

Besides statistical methods, clustering approaches are the most common analyses of biological time series and thus are a genuine reference for our approach.

For this purpose, the signals should be transformed in advance so that they have zero mean and unit variance. However, signals whose abundance does not change would be strongly distorted by this transformation. In order to increase the clustering quality, such signals can be filtered out. A simple to use quality filtering approach often seen on biological time series data is the application of a one-way ANOVA. Its main purpose in clustering approaches is not the detection of significances, but to filter signals whose average did not change during the time course. As a common threshold a p value of 0.05 was chosen. Note that the presented method of temporal classification does not require this step, since signal classification is not affected by the presence of other signals.

Clustering of the remaining 720 protein signals was performed using the k means algorithm with Euclidean distance [27]. The optimal number of clusters was determined to be 5 (Additional file 1: Figure S3). It is obvious that five clusters do not sufficiently represent all regulatory responses within a biological system. It is a suitable methodology for obtaining a global impression of the data set and a summary of the major protein kinetic groups. But for a detailed description of the response of fine-tuned biological processes, this approach is too rough. Interesting regulatory details are blurred by the sheer amount of data within a single cluster (Fig. 5, Additional file 1: Fig. S4). Regression analysis reveals that for every cluster 30% of the protein signals are not within the 95% prediction interval of the cluster mean. If the regulatory response of proteins is studied, in at least 30% of the cases a classification based on the cluster would be inaccurate at best.



cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	number of signals with
201 (71 %)	74 (71 %)	148 (74 %)	55 (69 %)	36 (68 %)	all points within PI
48 (17 %)	20 (19 %)	30 (15 %)	19 (24 %)	15 (28 %)	1 point outside of PI
26 (9 %)	7 (7 %)	9 (5 %)	4 (5 %)	2 (4 %)	2 points outside of PI
9 (3 %)	3 (3 %)	12 (6 %)	2 (2 %)	0 (0 %)	3+ points outside of PI

Fig. 5 Iterative clustering result. 720 protein signals were individually transformed to z scores and subsequently clustered by k means clustering ($k=5$). **A**: Five clusters are depicted with cluster mean and 95% prediction interval (PI). Signal colours indicate whether 0 (grey), 1 (green), 2 (yellow), or more (orange) points of a signal lie outside of the PI. The table below shows the percentage of each group. **B**: The cluster mean intensity is visualized together with signals that showed the highest (green) and lowest (red) coefficient of determination (R^2) to the cluster mean. **C**: The cluster mean intensity is plotted against the intensities of the signal of highest (green) and lowest (red) R^2 within the cluster. R^2 of both signals is given in each panel. **D**: Histogram of all R^2 values between signals and cluster mean. The percentage in each panel depicts the percentage of signals whose R^2 is lower than 0.8

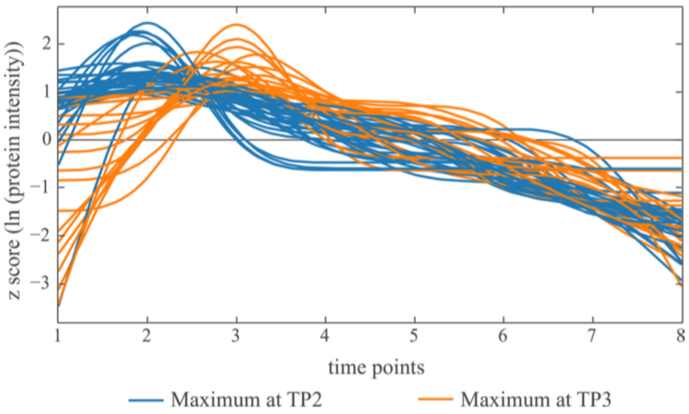


Fig. 6 Visualization of two classes that show a single maximum at time points 2 or 3 respectively. The intensity signals of 50 proteins are transformed to have zero mean and unit variance

Table 1 Enrichment result of early responder class

Functional term	q value	Trivial names (if annotated)
Protein.synthesis.ribosomal protein	0.0164	rps9; PRPL1; MRPL1; RPL23A; UBQ2; RPL40; UBQ1; RPS7; RPL18; RPL12; RPS27E1; RPL7; PRPL19;
Transport	0.0462	ATPVH;ATPVH;MPC1;AAA1;
Protein.synthesis.ribosomal protein. eukaryotic.60S subunit	0.0472	RPL23A; UBQ2; RPL40; UBQ1; RPL18; RPL12; RPL7
Protein.degradation.ubiquitin	0.0161	PKL1; UBC2; UBQ2; RPL40; UBQ1; UBQ2; RPL40; UBQ1; EIF3F; RPT4
RNA.RNA binding	0.0472	UBC2;REF1;HNR1
Transport.metabolite transporters at the mitochondrial membrane	0.0472	MPC1;AAA1
Polyamine metabolism.synthesis	0.0161	SPS1;SPD1

Functional annotations based on the MapMan ontology are listed together with the associated q value and trivial names if proteins had such

Biological interpretation

Besides using the smoothed signals for comparative analysis (e.g. co-expression networks), the smoothed and classified protein signals can be used for exploratory data analysis. To elucidate early, but short-term responders of the heat treatment, signals of the classes “Maximum at 2 or 3" can be isolated and used for categorizing the acclimation response.

Furthermore, global analysis strategies can be applied to classified signals. Gene set ontology enrichments of molecular functions can be applied to identify function overrepresentation.

Obviously, due to the sensitivity of the classification, the number of classes is by far greater than the number of clusters using common clustering strategies. This results in sparse occupancy of shape classes, which impedes enrichment strategies. However, the possibility of subsequent class aggregation presents valuable opportunities for analysing different combinations of response types. A gene set enrichment analysis was conducted on the early responder classes (Fig. 6) using MapMan functional annotations for *Chlamydomonas reinhardtii* genome version 5.5 [29, 30, 32]. Functional annotations that were overrepresented within the early responders can be seen in Table 1.

All functional annotations that are overrepresented in regulation shortly after heat onset were previously described to be involved in early heat acclimation regulation. (i) ribosomal proteins are required for the fast production of proteins; (ii) the transport group contains proteins predominantly involved meeting the increased demand of energy [35]; (iii) ubiquitin related proteins are necessary to both, degrade proteins that interfere with a heat acclimation, and remove proteins that aggregated due to the increased heat [36]; (iv) RNA binding proteins are involved in processing, stabilizing and exporting newly transcribed mRNA [37, 38]; (v) proteins of the polyamine synthesis group have been described to increase thermos-tolerance in algae [39].

The biological dissemination reveals that the classification approach is capable of elucidating the time-resolved orchestration of cellular responses and differentiating between different forms of regulation within a functional set of biological molecules.

Discussion

The era of high throughput technologies enabled researchers to analyse the abundance of thousands of molecules in a time-resolved manner. Scientists once needed days to take samples and measure the signals of a few proteins individually. Nowadays, it is possible to quantify the entire transcriptome or proteome in one fell swoop. Although it is possible to measure the kinetics of hundreds of proteins at a time, the number of robust strategies for an analysis of temporally resolved cell responses remains small [40–42].

Clustering methods have always been a popular tool to analyse time series experiments, as these are great options for unsupervised methods that work well with only a few assumptions to be made. For example, k-means clustering of time point averages represents the most commonly used algorithm for unsupervised analysis as its computation is efficient and the resulting clusters can easily be interpreted [43–45].

Although great findings have been made by this approach [46–48], it poses two problems when used for temporal characterization of regulation responses: (i) Most commonly used distance measures consider each coordinate separately. The time series vectors can be shuffled in pairs and still obtain the same distance. This behaviour is contrary to biological intuition because transcript or protein quantities are strongly dependent on the previous time points. This dependence is not taken into account in the model and inevitably leads to a decrease in the quality of the signal-to-noise separation [49, 50]. (ii) Although there are also distance measures that act in an environment-dependent manner (Dynamic Time Warping), in clustering methods it is necessary to specify the number of clusters in advance. This reduces accuracy, since small definable groups may be sorted into large groups, and their identification is thus only possible manually. Numerous ways of determining the optimal number of clusters have been developed (Elbow criterion, Silhouette index, Gap statistics) [25, 51], but in most cases these underestimate the number of biological response forms present. Furthermore, clustering approaches often are used to subsequently classify the data based on features that are visible when looking at whole clusters, but not necessarily are valid for individual cluster elements [52, 53]. As shown, this is prone to result in a huge number of misclassifications (Fig. 5). These and other similarity-based techniques are not able to dissect delicate regulation responses, but instead, these signals might be blurred by averaging effects.

Our approach models the kinetic response by constrained smoothing splines with the incorporation of measurement variance. Several other fitting techniques can be applied, each of which addresses different assumptions. Linear splines are the least complex fitting model for time series as they just connect the point estimates (median or average) at each time point. Point weighting is not possible and there is no separation of signal and noise. For the same reasons, other interpolating methods such as interpolating polynomials or cubic splines are not suited.

An exception is a polynomial-based function approximation with Chebyshev knots. A major problem with interpolating polynomials is Runge's Phenomenon. This is manifested by high frequency oscillation of the function in the outer knot intervals [54]. By clever rearrangement of the knots away from the curve centre and towards the peripheral areas such an oscillation can be prevented [55]. At the same time, the function no longer passes through the original data points. However, disadvantages here are both the non-obvious selection of knots and their weighting, as well as the lack of methods to comply with monotonicity constraints. Linear or nonlinear regression techniques seem inappropriate as they require either (i) the selection of a polynomial degree that does not represent any meaningful biological interpretation, or (ii) an already predefined function that is fitted to the signal.

The modelling of the time series by constrained smoothing splines is based on smoothness assumptions and the consideration of measurement point variances. Additionally, this regression approach preserves the existing dependence between neighbouring time points and therefore enforces monotonicity where excessive oscillation is unlikely. When compared to other interpolation methods (polynomials, linear splines, cubic splines) our approach showed high robustness while being flexible enough to capture characteristic events during the time course (Figs. 2, 4). Due to the overfitting tendency and presence of Runge's spike oscillation, polynomial interpolation is unsuited for classification analysis that handles extrema position as its primary characterization criterion. Linear splines show a high sensitivity for false declaration of extrema and perform poorly when it comes to predicting function values between the measured time points. Despite that cubic interpolating splines inherently aim to reduce heavy oscillations and perform great when it comes to predicting within intervals, their interpolating nature and inability to be weighted lead to little, but noticeable oscillations that interfere with an extrema-based classification strategy. While the number of shape classes can go to the hundreds, we could show that shapes with high flexibility and oscillations are found rarely (Fig. 3). This corresponds to the biological intuition of smooth protein regulation which was confirmed in Fig. 1.

Hybrid approaches are available that extend clustering approaches with prior smoothing regression [56], or by selecting meaningful expression profiles [1]. These approaches still rely on unsupervised approaches to group the data without predefining group labels. If faced with short time series data not exceeding 10 measurement time points, we propose that the number of group labels is manageable, hence all possible response shapes could be examined. However, a comparison of classification and clustering approaches remains difficult since the ground truth is unknown and both approaches address different questions. Most clustering approaches measure distances between signals, while classifications are concerned with the dissection of signal features. The incorporation of the information that biological signals tend to be smooth and not oscillating leads to a feature extraction that corresponds to the intuition regarding the regulation of biological molecules.

With our temporal classification approach for studying time resolved regulation, it is possible to not only find an optimal fit of the data, but also assign shape classes to large time series data sets. This makes it possible to analyse the temporal orchestration of acclimation response and actively search for patterns of interest. We were able to show that our method provides a robust estimator when faced with sparse data. Furthermore, a well-studied process of heat acclimation of *Chlamydomonas reinhardtii* was presented as an example of the method enabling a detailed and supervised analysis of specific acclimation responses.

The smoothing and classification algorithms can be accessed as F# implementation at https://github.com/CSBiology/TempClass (Additional file 1: Figure S5) [58].

Limitations

This classification strategy is based on the selection of the optimal combination of monotone regions and extreme points. This approach requires an inner optimization to obtain the best fit of any enforced combination of monotonicity constraints, and an outer optimization to select the most ideal of the best shapes. With an increasing number of measured time points, there is a combinatorial explosion of the number of potential curve configurations. This not only increases the calculation time exponentially, but also the large number of resulting classes becomes unmanageable. Therefore, we have limited the number of allowed extreme points to 4 and recommend temporal classification for time series with 4 to 12 measurement time points.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05636-6.

Additional file 1: Spline basis functions, Table S1, and Figures S1 - S6.

Author contributions

BV and TM wrote the manuscript text. BV, TL, and TM implemented the method. All figures were prepared by BV. TM and PZ. supervised the work. All authors reviewed the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. This work was supported by Deutsche Forschungsgemeinschaft (TR175, project D02).

Availability of data and materials

The data set the method was applied on is available in the repository: ***Systems-wide investigation of responses to moderate and acute high temperatures in the green alga Chlamydomonas reinhardtii***, accessible at https://doi.org/10.60534/9e5jx-75d83 [57].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 26 October 2023 Accepted: 3 January 2024
Published online: 17 January 2024

References

1. Ernst J, Nau GJ, Bar-Joseph Z. Clustering short time series gene expression data. Bioinformatics. 2005;21(Suppl 1):i159–68. https://doi.org/10.1093/bioinformatics/bti1022.
2. Ernst J, Bar-Joseph Z. STEM: a tool for the analysis of short time series gene expression data. BMC Bioinformatics. 2006;7:191. https://doi.org/10.1186/1471-2105-7-191.
3. Lloyd S. Least squares quantization in PCM. IEEE Trans Inform Theory. 1982;28:129–37. https://doi.org/10.1109/TIT.1982.1056489.
4. Bathia N, Yao Q, Ziegelmann F. Identifying the finite dimensionality of curve time series. Ann Statist. 2010. https://doi.org/10.1214/10-AOS819.
5. Huang X, Ye Y, Xiong L, Lau RY, Jiang N, Wang S. Time series k-means: a new k-means type smooth subspace clustering for time series data. Inf Sci. 2016;367–368:1–13. https://doi.org/10.1016/j.ins.2016.05.040.
6. Warren LT. Clustering of time series data—a survey. Pattern Recogn. 2005;38:1857–74. https://doi.org/10.1016/j.patcog.2005.01.025.
7. Wood SN. Monotonic smoothing splines fitted by cross validation. SIAM J Sci Comput. 1994;15(5):1126–33. https://doi.org/10.1137/0915069.
8. Leifeld T, Venn B, Cui S, Zhang Z, Mühlhaus T, Zhang P. Curve form based quantization of short time series data. In: pp. 3710–3715. doi:https://doi.org/10.23919/ECC.2019.8795870.
9. de Boor C. A practical guide to splines. New York, N.Y.: Springer; 2001.
10. Lancaster P. Curve and surface fitting: an introduction. London: Academic Press; 1986.
11. Eubank RL. Nonparametric regression and spline smoothing. 2nd ed. Boca Raton: Chapman and Hall/CRC; 1999.
12. Fahrmeir L, Kneib T, Lang S. Regression: modelle, methoden und anwendungen. Berlin: Springer; 2007.
13. Fn F. Monotone piecewise cubic interpolation. SIAM J Numer Anal. 1980;17:238–46.
14. Ramsay JO. Monotone regression splines in action. Stat Sci. 1988;1:425–41.
15. Meyer MC. Constrained penalized splines. Can J Stat. 2012;40:190–206. https://doi.org/10.1002/cjs.10137.
16. Turlach BA. Constrained smoothing splines revisited. Statistics Research Report SRR 008-97. Center for Mathematics and Its Applications. Australian National University Canberra. 1997.
17. Anderson E, Bai Z, Bischof C, Blackford S, Demmel J, Dongarra J, et al. LAPACK users' guide. 3rd ed. Philadelphia: Society for Industrial and Applied Mathematics; 1999.
18. Craven P, Wahba G. Smoothing noisy data with spline functions. Numer Math. 1978;31:377–403. https://doi.org/10.1007/BF01404567.
19. Hutchinson MF, Gessler PE. Splines—more than just a smooth interpolator. Geoderma. 1994;62:45–67. https://doi.org/10.1016/0016-7061(94)90027-2.
20. Lukas MA. Robust generalized cross-validation for choosing the regularization parameter. Inverse Prob. 2006;22:1883–902. https://doi.org/10.1088/0266-5611/22/5/021.
21. Akaike H (1998) Information theory and an extension of the maximum likelihood principle. In: Parzen E, Tanabe K, Kitagawa G, editors. Selected papers of Hirotugu Akaike. Springer Series in Statistics. New York: Springer. https://doi.org/10.1007/978-1-4612-1694-0_15
22. Hurvich CM, Tsai C-L. Regression and time series model selection in small samples. Biometrika. 1989;76:297–307. https://doi.org/10.1093/biomet/76.2.297.
23. MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1: Statistics: The Regents of the University of California; 1967.
24. Hartigan JA, Wong MA. A K-means clustering algorithm. Appl Stat. 1979;28:100. https://doi.org/10.2307/2346830.
25. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc Ser B Stat Methodol. 2001;63:411–23. https://doi.org/10.1111/1467-9868.00293.
26. Jain A, Nandakumar K, Ross A. Score normalization in multimodal biometric systems. Pattern Recogn. 2005;38:2270–85. https://doi.org/10.1016/j.patcog.2005.01.012.
27. Venn B, Mühlhaus T, Schneider K, Weil L, Zimmer D. fslaborg/FSharp.Stats: release 0.5.0: Zenodo; 2023.
28. Schneider K, Venn B, Mühlhaus T. Plotly. NET: a fully featured charting library for NET programming languages. F1000Res. 2022; 11: 1094. https://doi.org/10.12688/f1000research.123971.1.
29. Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, et al. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J. 2004;37:914–39. https://doi.org/10.1111/j.1365-313x.2004.02016.x.
30. Usadel B, Poree F, Nagel A, Lohse M, Czedik-Eysenberg A, Stitt M. A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species. Maize Plant Cell Environ. 2009;32:1211–29. https://doi.org/10.1111/j.1365-3040.2009.01978.x.
31. Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, et al. The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science. 2007;318:245–50. https://doi.org/10.1126/science.1143609.
32. Venn B. CSBiology/OntologyEnrichment: release 0.0.1: Zenodo; 2022.
33. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA. 2003;100:9440–5. https://doi.org/10.1073/pnas.1530509100.
34. Zhang N, Mattoon EM, McHargue W, Venn B, Zimmer D, Pecani K, et al. Systems-wide analysis revealed shared and unique responses to moderate and acute high temperatures in the green alga Chlamydomonas reinhardtii. Commun Biol. 2022;5:460. https://doi.org/10.1038/s42003-022-03359-z.

35. Vale RD. AAA proteins. Lords of the ring. J Cell Biol. 2000;150:F13–9. https://doi.org/10.1083/jcb.150.1.f13.
36. Galves M, Rathi R, Prag G, Ashkenazi A. Ubiquitin signaling and degradation of aggregate-prone proteins. Trends Biochem Sci. 2019;44:872–84. https://doi.org/10.1016/j.tibs.2019.04.007.
37. Pokora W, Tułodziecki S, Dettlaff-Pokora A, Aksmann A. Cross talk between hydrogen peroxide and nitric oxide in the unicellular green algae cell cycle: how does it work? Cells. 2022. https://doi.org/10.3390/cells11152425.
38. Pandey M, Stormo GD, Dutcher SK. Alternative splicing during the chlamydomonasreinhardtii cell cycle. G3 Bethesda. 2020;10:3797–810. https://doi.org/10.1534/g3.120.401622.
39. Liu S, Zhang J, Sun X, Xu N. Characterization of spermidine synthase (SPDS) gene and RNA—Seq based identification of spermidine (SPD) and spermine (SPM) involvement in improving high temperature stress tolerance in gracilariopsis lemaneiformis (Rhodophyta). Front Mar Sci. 2022. https://doi.org/10.3389/fmars.2022.939888.
40. Tripto NI, Kabir M, Bayzid MS, Rahman A. Evaluation of classification and forecasting methods on time series gene expression data. PLoS ONE. 2020;15: e0241686. https://doi.org/10.1371/journal.pone.0241686.
41. Androulakis IP, Yang E, Almon RR. Analysis of time-series gene expression data: methods, challenges, and opportunities. Annu Rev Biomed Eng. 2007;9:205–28. https://doi.org/10.1146/annurev.bioeng.9.060906.151904.
42. Wang X, Wu M, Li Z, Chan C. Short time-series microarray analysis: methods and challenges. BMC Syst Biol. 2008;2:58. https://doi.org/10.1186/1752-0509-2-58.
43. Jain AK, Dubes RC. Algorithms for clustering data; 1988.
44. Maigné É, Noirot C, Henry J, Adu Kesewaah Y, Badin L, Déjean S, et al. Asterics: a simple tool for the ExploRation and Integration of omiCS data. BMC Bioinformatics. 2023;24:391. https://doi.org/10.1186/s12859-023-05504-9.
45. Datta S, Datta S. Evaluation of clustering algorithms for gene expression data. BMC Bioinformatics. 2006;7(Suppl 4):S17. https://doi.org/10.1186/1471-2105-7-S4-S17.
46. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nat Genet. 1999;22:281–5. https://doi.org/10.1038/10343.
47. Saadeh H, Fayez RQA, Elshqeirat B. Application of K-means clustering to identify similar gene expression patterns during erythroid development. IJMLC. 2020;10:452–7. https://doi.org/10.18178/ijmlc.2020.10.3.956.
48. Nies H, Zakaria Z, Mohamad M, Chan W, Zaki N, Sinnott R, et al. A review of computational methods for clustering genes with similar biological functions. Processes. 2019;7:550. https://doi.org/10.3390/pr7090550.
49. Abanda A, Mori U, Lozano JA. A review on distance based time series classification. Data Min Knowl Disc. 2019;33:378–412. https://doi.org/10.1007/s10618-018-0596-4.
50. Bagnall A, Lines J, Bostrom A, Large J, Keogh E. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Min Knowl Discov. 2017;31:606–60. https://doi.org/10.1007/s10618-016-0483-9.
51. Kodinariya TM, Makwana PR. Review on determining number of cluster in K-means clustering. Int J. 2013;1(6):90–5.
52. Babichev S, Škvor J. Technique of gene expression profiles extraction based on the complex use of clustering and classification methods. Diagnostics. 2020. https://doi.org/10.3390/diagnostics10080584.
53. Datta S, Datta S. Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. BMC Bioinformatics. 2006;7:397. https://doi.org/10.1186/1471-2105-7-397.
54. Boyd JP. Defeating the Runge phenomenon for equispaced polynomial interpolation via Tikhonov regularization. Appl Math Lett. 1992;5:57–9. https://doi.org/10.1016/0893-9659(92)90014-Z.
55. Trefethen LN. Approximation theory and approximation practice. Philadelphia: Society for Industrial and Applied Mathematics; 2013.
56. Déjean S, Martin PG, Baccini A, Besse P. Clustering time-series gene expression data using smoothing spline derivatives. EURASIP J Bioinform Syst Biol. 2007;2007:1. https://doi.org/10.1155/2007/70561.
57. Zhang N, Mattoon E, McHargue W, Venn B, Zimmer D, Pecani K, Jeong J, Anderson C, Chen C, Berry J, Xia M, Tzeng SC, Becker E, Pazouki L, Evans B, Cross F, Cheng J, Czymmek K, Schroda M, Mühlhaus T, Zhang R. Systems-wide investigation of responses to moderate and acute high temperatures in the green alga Chlamydomonas reinhardtii [Data set]. DataPLANT. 2023. https://doi.org/10.60534/9e5jx-75d83
58. Venn B, Mühlhaus T. CSBiology/TempClass: release 0.0.1: Zenodo; 2023.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplemental Information

Spline basis functions:

$$\phi_{0i}(t) = (t_{i+1} - t_i)/h_i$$
$$\phi_{1i}(t) = (t_i - t_{i+1})/h_i$$
$$\gamma_{0i}(t) = (\phi_{0i}(t)^3 - \phi_{0i}(t))h_j^2/6$$
$$\gamma_{1i}(t) = (\phi_{1i}(t)^3 - \phi_{1i}(t))h_j^2/6$$

Table S1 Number of curve configuration possibilities at 2 - 8 timepoints. Additionally, to 1, 2, 3, or 4 consecutive extrema, two monotone in/decreasing curves and a constant, never changing configuration are possible. Since an extremum is assigned to its nearest knot, a extremum can occur at every knot.

Extrema count / Knot count	constant	0	1	2	3	4	Sum
2	1	2	4	2	-	-	**9**
3	1	2	6	6	2	-	**17**
4	1	2	8	12	8	2	**33**
5	1	2	10	20	20	10	**63**
6	1	2	12	30	40	30	**115**
7	1	2	14	42	70	70	**199**
8	1	2	16	56	112	140	**327**

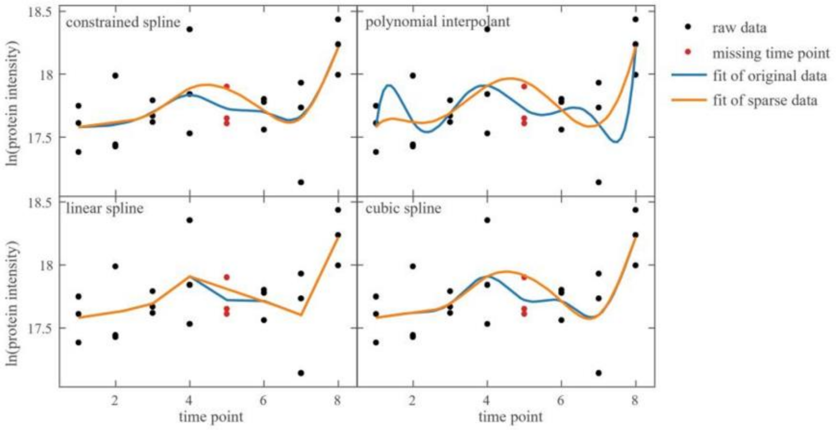


Figure S1. Exemplary leave one out cross validation using four fitting techniques. The blue curves represent the original fit using all available data points. The orange curves represent the fit when all, but the red data points are used. Vertical distances from the blue to the orange curve are measured at time point 5 to assess the robustness of the different fitting techniques.

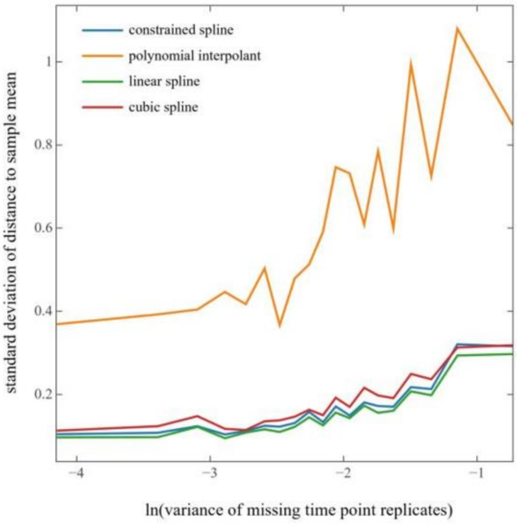


Figure. S2. Fidelity analysis. Four fitting techniques were applied to each protein signal. After deleting every inner time point once, the distance from the sample mean to the prediction using the sparse signal is measured. This data was grouped in 20 equally large bins by the variance of the original time point replicates. The standard deviation of the distances within each bin are calculated and plotted against the average time point variance.

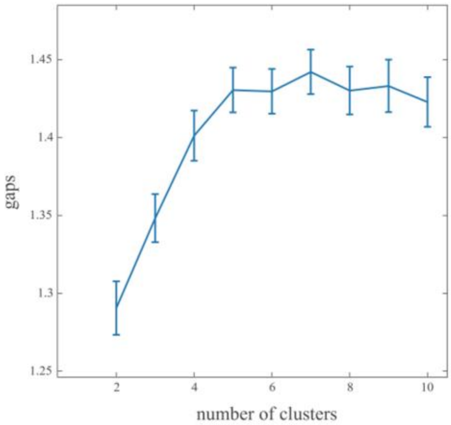


Figure S3. Determination of optimal cluster number via gap statistics. The optimal cluster number is the smallest k such that gap(k)≥ gap(k+1)-s(k+1) where s is the standard deviation of the reference dispersion multiplied by sqrt(1+1/bootstraps). 400 bootstrap iterations where performed. The optimal cluster number was determined to be 5.

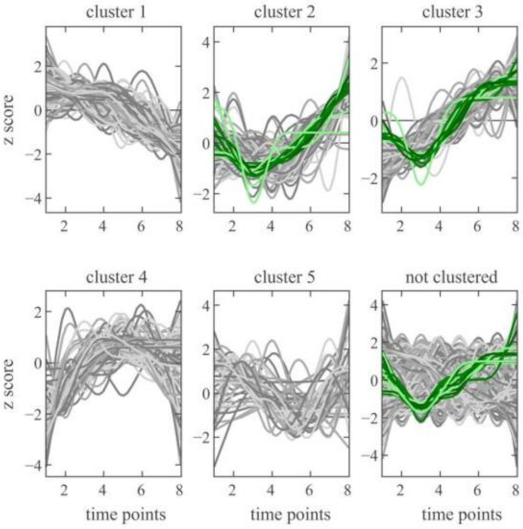


Figure S4. Example of distribution of smoothed protein signals that are member of class "Minimum at TP3" in different clusters.

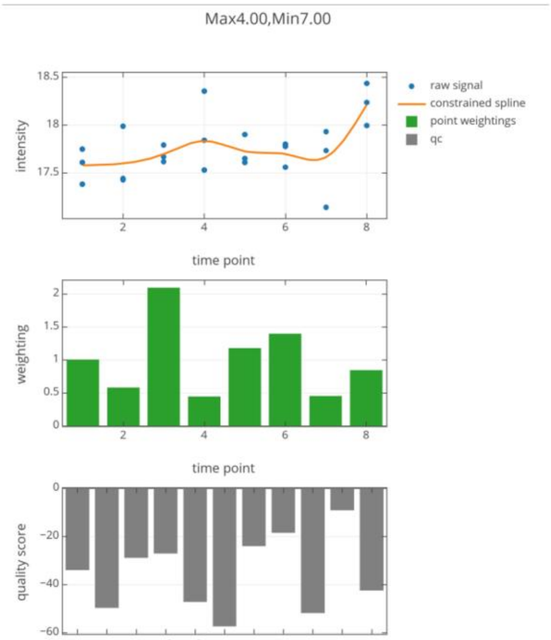


Figure S5 Example of protein-wise result visualization facilitated by the TempClass package available at https://github.com/CSBiology/TempClass.

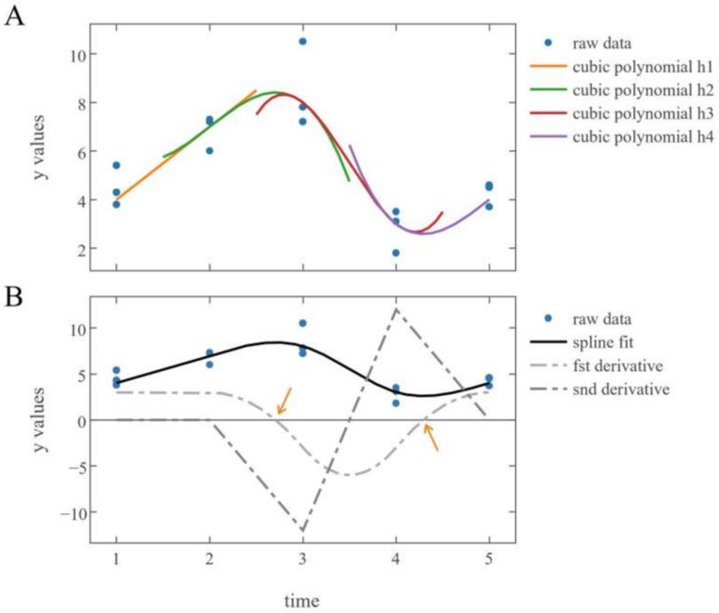


*Figure S6. Extrema extraction **A** The constrained smoothing spline of five knots consists of four cubic polynomials that are joined at the knots. **B** From the resulting spline (black) the first and second derivative are determined and visualized (dashed lines). Orange arrows mark time points of zero slope that indicate a maximum if the second derivative is negative or a minimum if the second derivative is positive.*

Article II: TMEA: A Thermodynamically Motivated Framework for Functional Characterization of Biological Responses to System Acclimation

Kevin Schneider*, Benedikt Venn*, Timo Mühlhaus

| Entropy, 2020, doi: 10.3390/e22091030

entropy

IMPACT FACTOR 2.7

Indexed in: PubMed

CITESCORE 4.7

Article

Editor’s Choice

TMEA: A Thermodynamically Motivated Framework for Functional Characterization of Biological Responses to System Acclimation

Kevin Schneider, Benedikt Venn and Timo Mühlhaus

Special Issue
Thermodynamics and Information Theory of Living Systems

Edited by
Dr. Bernat Corominas-Murtra, Dr. Artemy Kolchinsky and Prof. Rudolf Hanel

MDPI

https://doi.org/10.3390/e22091030

entropy

MDPI

Article

TMEA: A Thermodynamically Motivated Framework for Functional Characterization of Biological Responses to System Acclimation

Kevin Schneider †, **Benedikt Venn** † and **Timo Mühlhaus** *

Computational Systems Biology, University of Kaiserslautern, 67663 Kaiserslautern, Germany; schneike@rhrk.uni-kl.de (K.S.); venn@rhrk.uni-kl.de (B.V.)

* Correspondence: muehlhaus@bio.uni-kl.de

† These authors are equally contributed.

Received: 16 August 2020; Accepted: 11 September 2020; Published: 15 September 2020

Abstract: The objective of gene set enrichment analysis (GSEA) in modern biological studies is to identify functional profiles in huge sets of biomolecules generated by high-throughput measurements of genes, transcripts, metabolites, and proteins. GSEA is based on a two-stage process using classical statistical analysis to score the input data and subsequent testing for overrepresentation of the enrichment score within a given functional coherent set. However, enrichment scores computed by different methods are merely statistically motivated and often elusive to direct biological interpretation. Here, we propose a novel approach, called Thermodynamically Motivated Enrichment Analysis (TMEA), to account for the energy investment in biological relevant processes. Therefore, TMEA is based on surprisal analysis, which offers a thermodynamic-free energy-based representation of the biological steady state and of the biological change. The contribution of each biomolecule underlying the changes in free energy is used in a Monte Carlo resampling procedure resulting in a functional characterization directly coupled to the thermodynamic characterization of biological responses to system perturbations. To illustrate the utility of our method on real experimental data, we benchmark our approach on plant acclimation to high light and compare the performance of TMEA with the most frequently used method for GSEA.

Keywords: GSEA; gene set enrichment analysis; pathway analysis; surprisal analysis; information theory; thermodynamics; free energy; acclimation response; transcription levels

1. Introduction

Within the frame of their genetic capacity, organisms are able to acclimate to changes in environmental conditions. Acclimation responses thereby represent a complex dynamic adjustment of the entire molecular cellular network. The ability to acclimate ensures the survival of all living organisms and is therefore fundamental for the understanding of biological systems. Due to their mainly sessile lifestyle, plant systems particularly have to face fluctuating environmental conditions, including biotic and abiotic stresses [1,2]. Detailed knowledge about how plants acclimate to a changing environment is crucial especially in times of global climate changes, as plants are of great importance for our quality of life as a key source of food, shelter, fiber, medicine, and fuel [3,4]. A comprehensive understanding of plant acclimation responses allows the development of strategies to stabilize or enhance yields in increasingly hostile environments. Acclimation dynamics occur on different time scales—from minutes to days—and act on all system levels involving the modification of gene expression, protein activity, and metabolite profiles.

To elucidate these dynamics and to describe the different phases of acclimation, multiple time course experiments recording changes on various system levels have been performed in the past [5–13].

However, the identification and functional characterization based on these measurements remains a non-trivial task. Typically, these experiments result in huge lists of different molecules such as transcripts, metabolites, and proteins modified over the time course of the acclimation process. Therefore, gene set enrichment analysis (GSEA) has become an important approach to interpret these resulting lists. The principle of GSEA is to identify sets of biological molecules that are significantly overrepresented in a functional coherent set in a known biological pathway, compared to a background set of measured entities. Usually, the grouping is derived from functional gene and pathway annotation databases such as MapMan [14], GO [15], KEGG [16], Reactome [3], Wikipathways [17], BioCyc [18], or others.

One of the most frequently used approaches to perform GSEA is a one-sided hypergeometric or Fisher's exact test that detects overrepresented functional sets derived from an experiment [19–25]. Therefore, every measured molecule is assigned a *p*-value or label that indicates whether it showed a (significant) change during a time course and/or compared to a reference. A subsequent hypergeometric test identifies functional sets that are significantly overrepresented in the data [26]. Every term leads to an individual test, leading to the necessity for multiple testing corrections. The drawback of this method is that it relies on applying a *p*-value cutoff to define the boundary between included and excluded molecules. This arbitrary distinction leads to a discretization of the information that dramatically influences the outcome of a GSEA [27] and is particularly difficult in time-series analysis. This problem is addressed by several methods that can be categorized into Functional Class Scoring (FCS) and Single-Sample (SS) methods. While FCS calculates scores (*p*-values or ranks) for every entity within a given set, SS aims to score every gene set per sample according to its importance [28–31]. In addition, multiple methods have been proposed to integrate multiple annotation databases or address the problem of overlapping set annotations due to molecules playing a role in different pathways and processes [32]. In addition, network-based approaches are available; however, they are restricted to biological systems where a deeper understanding of the molecular interaction is already available [33–35]. The existence of different counting or ranking metrics, enrichment statistics, and several variants on significance estimation demonstrates the difficulty of finding a single, optimal statistic due to the complexity, heterogeneity, and multi-modal distribution within the data [36]. Currently, the definition of an enriched pathway is predominantly of statistical nature due to an a priori defined set of interest. From a biological perspective, that might not always be an ideal scenario, especially if the pathways of interest are not regulated by a majority but rather a few or even a single key enzyme.

In this paper, we propose to account for the energy investment driving the required process to understand acclimation responses at the systems level. For this objective, we developed a novel approach called Thermodynamically Motivated Enrichment Analysis (TMEA). Plant systems are maintained in individual states far from thermodynamic equilibrium and fuel all biogeochemical processes by the absorption of incoming sunlight. Entropy production is a general consequence of these processes and allows computing their free energy. The principle of minimum entropy production states that systems are driven to steady states that are characterized by a minimum value of entropy production rate given the prevailing constraints [37].

Motivated by information theory, surprisal analysis offers a very compact, thermodynamic-free, energy-based representation of the biological steady state and of the biological change, the so-called unbalanced processes [38]. Therefore, we use surprisal analysis to compute free energy changes throughout the course of the specific acclimation response. Surprisal analysis identifies both a baseline state of maximum entropy and constraints that prevent the system from reaching it [39,40]. Molecules contribute to these constraints, and the difference in their contributions makes it possible to characterize different states of the system as patterns that collectively cause deviations from the baseline state. Associated with the constraints are time-dependent state variables that reflect the importance of the constraints and therefore carry information of how energy is invested over time [41,42]. In TMEA, we use the intensive variable *G*, which quantifies the contribution of each molecule underlying the free energy change as the basis for a Monte Carlo resampling procedure resulting in a functional characterization directly coupled to the thermodynamic characterization of biological responses to system perturbations, which is not yet addressed by conventional methods.

Finally, we demonstrate the application of our methods to light acclimation in *Arabidopsis thaliana* and evaluate the knowledge that we can recover solely from transcriptional changes compared to the current literature knowledge.

2. Materials and Methods

2.1. Dataset

The transcriptomics data used in this study were obtained from (NCBI Gene Expression Omnibus, Accession GSE125950) a high light experiment conducted with *Arabidopsis thaliana* [43]. First, 14-day-old Col-0 seeds were treated with 450 µmol photons m^{-2} s^{-1} for 4 days under long-day conditions (18 h d^{-1}). After 4 days of acclimation, the light was reduced to control conditions (80 µmol photons m^{-2} s^{-1}) for another 4 days. Entire shoots were harvested at 11 time points (0 min, 1 min, 15 min, 3 h, 2 days, 4 days for acclimation and de-acclimation, where 4 days of acclimation equals 0 min of de-acclimation). Transcripts were measured from three biological replicates for every time point by RNA-Seq using an Illumina HiSeq 2500 system (Illumina, San Diego, CA, USA). Metabolomics data for the verification of selected transcripts were obtained from the Supplemental Table S2 of the same study [43]. Metabolites were sampled at 13 time points (0 min, 5 min, 15 min, 3 h, 1 day, 2 days, and 4 days for acclimation and de-acclimation, respectively) [43].

2.2. Surprisal Analysis

Surprisal analysis (SA) assumes that a system will decrease its free energy spontaneously unless constrained [38]. It provides a method to determine a small set of state variables λ_α, which are dependent on time and determine the deviations of the observed process from a balance state of minimal free energy. For every constraint, a weight is assigned to each measured entity (e.g., transcript, metabolite, or protein), which describes the influence of this molecule to the constraint.

The surprisal of each individual observation $X_i(t)$ is defined as the deviation from the steady state X_i^0:

$$I(x_i) = -ln\left[\frac{X_i(t)}{X_i^0}\right]. \tag{1}$$

Then, SA fits the surprisal by a sum of terms:

$$-\sum_{\alpha=1} G_{i\alpha}\lambda_\alpha(t), \tag{2}$$

where α is the index of the constraint, $G_{i\alpha}$ is the weight of the event X_i in constraint G_α, and $\lambda_\alpha(t)$ is the Lagrange multiplier for G_α that is being varied to find the best fit. This is practically achieved by singular value decomposition, simultaneously yielding a baseline state of minimum free energy for $\alpha = 0$ [39,40,44].

Free energy changes can be determined for each constraint as work available to the molecular system under investigation from the results of surprisal analysis; the total work done on the system is the sum of these terms [45,46]:

$$\begin{gathered} F_\alpha(\mathrm{t}) = -\lambda_\alpha(t)\sum_i X_i(t)\, G_{i\alpha}, \\ F_{total}(\mathrm{t}) = -\textstyle\sum_{\alpha=1}(\lambda_\alpha(t)\sum_i X_i(t)\, G_{i\alpha}). \end{gathered} \tag{3}$$

With an increasing constraint index, the contribution to the deviations from the baseline state drastically decreases. SA was computed using our implementation provided within the TMEA package [47] written in F# based on LAPACK Version 3.8 [48].

2.3. Functional Annotation and Pathway Database

Functional annotations for each transcript were obtained from MapMan ontology. MapMan is a plant specific ontology that covers functional annotations and pathway information in great detail. Entities sharing functional properties are summarized as a functionally annotated set (FAS) Mapping files are available at [49] for a collection of all MapMan terms and [50] for Arabidopsis-specific annotations. Metabolite annotations for each transcript were obtained from the KEGG Compound Database [51]. Compound-involved enzymes were mapped to transcript identifiers (TAIR 10) by using KEGG Orthology for *Arabidopsis thaliana* [52].

2.4. Gene Set Enrichment Analysis Based on Hypergeometric Function

Several methods for the identification of enriched FAS are summarized under the concept of gene set enrichment analysis (GSEA). One of the most established and frequently applied methods is a one-sided hypergeometric test, which detects overrepresented FAS in all FASs derived from the experiment [24,25]. For enrichment analysis based on hypergeometric tests, all genes were tested for significant differential expression during the time course. Differentially expressed genes (DEGs) were obtained using DESeq2 [53] by a comparison of transcripts at each time point of the high light treatment with the initial time point. Transcripts are labeled as DEGs if their abundance fold change is >2 with a false discovery rate (FDR) ≤ 0.05. A subsequent hypergeometric test identifies the FASs with a minimal size of 5 that are significantly overrepresented in the data [26]. Since one test is performed for each annotation, a multiple testing correction is performed by controlling the FDR by the Benjamini–Hochberg method [25,54,55].

2.5. Further Statistical Analysis and Visualization

All computational analyses were conducted using the open source F# libraries FSharp.Stats [56] and BioFSharp [57]. Linear regression, Benjamini–Hochberg correction, and clustering were conducted using the FSharp.Stats version 0.2.1-beta. For ontology annotation and GSEA based on hypergeometric tests, we used BioFSharp version 2.0.0-beta4 [57]. Data visualization was performed using the FSharp.Plotly version 2.0.0 chart library built on plotly.js [58].

3. Results

3.1. A Thermodynamic-Free Energy-Based Framework for the Functional Description of Biological Systems Not in Equilibrium Named TMEA

We present Thermodynamically Motivated Enrichment Analysis (TMEA), which coupled with surprisal analysis (SA) provides an unbiased functional description for the thermodynamic constraints prevailing on a biological system. It is based on thermodynamic and information theoretic principles and reduces the complexity of a given dataset using Monte Carlo simulation to a level that is both easier to manage and interpret from a biological point of view. Our open source implementation of TMEA in the functional programming language F# is freely available at https://github.com/CSBiology/TMEA [47].

TMEA applies three distinct steps: (i) the computation of SA to identify the constraints and contributing weights; (ii) the annotation and grouping of entities in the dataset using a given biological function pathway annotation databases, and (iii) a Monte Carlo permutation test performed by resampling of the weight sums as a test statistic for all functional sets. Testing assesses if the weight sum of each category is observed due to chance given the distribution of weight contributions provided by SA. We designed step (iii) specifically for the functional analysis of constraints reported by SA and here provide both a mathematical formulation and rationale of the design decisions.

Let $E = \{w_1, \ldots, w_s\}$ denote a set of cardinality s, containing weighted contributions w_i of entities to the constraint G_α. Let $E^+ = \{w^+ \in E : w^+ > 0\}$ and $E^- = \{w^- \in E : w^- < 0\}$ denote the directional subsets of E with either positive or negative sign of cardinalities s^+/s^-. For the observed directional sums of contribution weights in E^+/E^-:

$$\hat{w^+} = \sum E^+; \; \hat{w^-} = \sum E^-, \tag{4}$$

we want to compute the p-values

$$p^+ = \mathrm{P}\left(\mathrm{W}^+ \geq \hat{w^+}\right); \; p^- = \mathrm{P}\left(\mathrm{W}^- \leq \hat{w^-}\right), \tag{5}$$

which determine how likely it is to observe contribution weight sums at least as extreme as $\hat{w^+}/\hat{w^-}$ for E^+/E^- given the distribution of the test statistic for directional contribution weight sums W^+ and W^-. However, we do not know the exact distributions of W^+/W^-, which may also not be normal depending on the dataset. Additionally, estimating W^+ and W^- by full permutation testing also proves impractical due to the size of the datasets typically used in modern biology. Therefore, we employ a Monte Carlo resampling procedure, which consists of resampling b independent replicates

$$E_1^{*+}, \ldots, E_b^{*+}; \; E_1^{*-}, \ldots, E_b^{*-} \tag{6}$$

from G_α with cardinality s^+ and s^- and aggregating the sum of these samples as:

$$W_1^+, \ldots, W_b^+; \; W_1^-, \ldots, W_b^-, \tag{7}$$

where

$$W_i^+ = \sum E_i^{*+}, \; W_i^- = \sum E_i^{*-}; \; i \in \{1, \ldots, b\}, \tag{8}$$

and using an empirical estimator for p^+/p^-:

$$\begin{aligned} p^+_{empirical} &= \tfrac{1}{b} \sum_{i=1}^{b} \mathbf{1}\left\{W_i^+ \geq \hat{w^+}\right\} \\ p^-_{empirical} &= \tfrac{1}{b} \sum_{i=1}^{b} \mathbf{1}\left\{W_i^- \leq \hat{w^-}\right\} \end{aligned} \tag{9}$$

where **1** is the indicator function. Note that b should be high, as the minimal p-value that can be obtained is $\frac{1}{b}$ [59]. After subsequently correcting $p^+_{empirical}/p^-_{empirical}$ based on FDR using the Benjamini–Hochberg method [55], the corresponding annotations can be assumed to have a significant influence on the respective constraint based on a confidence threshold of e.g., 0.05. A visual representation of the algorithm is depicted in Figure 1.

TMEA yields two functional descriptors for each constraint G_α: one for positively contributing entities, and one for inversely contributing entities. These descriptors report what kind of functional information is overrepresented in either part of the constraint. Coupled with the constraint potentials λ_α obtained by SA, TMEA results can be used to further characterize the thermodynamic state transitions that the biological system undergoes while responding to a perturbation.

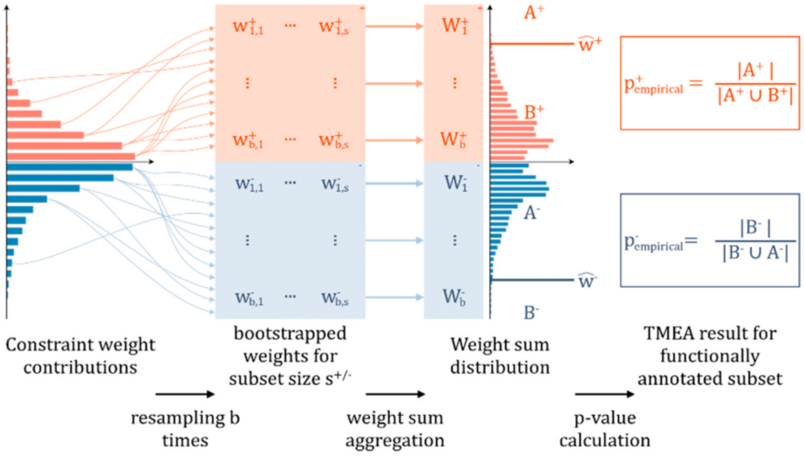

$w^+_{1,1}$ ⋯ $w^+_{1,s}$
$w^+_{b,1}$ ⋯ $w^+_{b,s}$
$w^-_{1,1}$ ⋯ $w^-_{1,s}$
$w^-_{b,1}$ ⋯ $w^-_{b,s}$
W^+_1
W^+_b
W^-_1
W^-_b
A^+
$\widehat{w}^+$
B^+
A^-
$\widehat{w}^-$
B^-
$p^+_{empirical} = \frac{|A^+|}{|A^+ \cup B^+|}$
$p^-_{empirical} = \frac{|B^-|}{|B^- \cup A^-|}$
Constraint weight contributions
bootstrapped weights for subset size s+/-
Weight sum distribution
TMEA result for functionally annotated subset
resampling b times
weight sum aggregation
p-value calculation

Figure 1. Schematic overview of the Monte Carlo permutation testing procedure used in Thermodynamically Motivated Enrichment Analysis (TMEA). Left to right: For a functionally annotated set of size s ($s > 5$) in the original dataset, the size of the positively and negatively contributing subsets is determined ($s^{+/-}$). Subsequently, b random samples are resampled from the weight distribution of the original constraint yielded by surprisal analysis from either the positive or negative part respectively, to generate b bootstrapped samples of sizes $s^{+/-}$. Then, these samples are aggregated to generate b weight sums for positive and negative weights each. Then, the frequency distributions of these weight sums are used to report empirical p-values, which inform how likely it is to observe the given positive or negative weight sum for bin sizes $s^{+/-}$ in the original constraint by chance based on the values above ($A^{+/-}$) and below ($B^{+/-}$) the observed value.

3.2. Contribution Weight Sums as Test Statistic

Ranking entities in a biological dataset from a thermodynamic point of view leads to a different perspective than applying purely statistical methods based on some form of majority voting [38]. The latter tend to reliably report FAS that show an overall consistent change but often fail to detect the importance of single or a small group of entities corresponding to a potential key regulator of the pathway. When statistically analyzing constraints reported by SA, it is important to select a test statistic that reflects this property. We applied TMEA to our high light acclimation benchmark dataset and treated positive and inverse weights separately after pooling the dominant constrains. Here, the first three constraints ($\alpha = 1, \ldots, 3$) were considered to contain sufficient information to depict the characteristics of the high light response by an elbow criterion based on "importance loss" (Figure A2) between the singular values obtained by the singular value decomposition (SVD) procedure. Together with the baseline state (the "zeroth" constraint for $\alpha = 0$), these patterns are sufficient to recover 98.6% of the original data (Figure A2).

To quantify how counting extreme values might relate to the sum of weight contributions, we then calculated the weight threshold for all quantiles between 1% and 99%, and for all those thresholds, both the ratios of the sum of contribution weights (weight ratio (WR)) and the amount of weights above/below the threshold (count ratio (CR)) for all annotated sets (Figure 2 top). Subsequent investigation of the 15% trimmed mean of R^2 of linear regression of WRs by CRs revealed that CRs can be used to explain 67.8% of the variance of WR for positively and 65.6% for negatively contributing subsets (Figure 2 bottom right and left, respectively), which indicates an importance of considering weights rather than just relying on counts. This observation supports the selection of the weight sum as the tests statistic for functionally describing constraints obtained by SA. Here, the directional sums of contribution weights $\hat{w}^+/\hat{w}^-$ can partially be explained with the count of extreme values suggesting that TMEA covers the classical scenario. However, a considerable amount of variance remains unexplained, pointing to the requirement to consider the influence of weights.

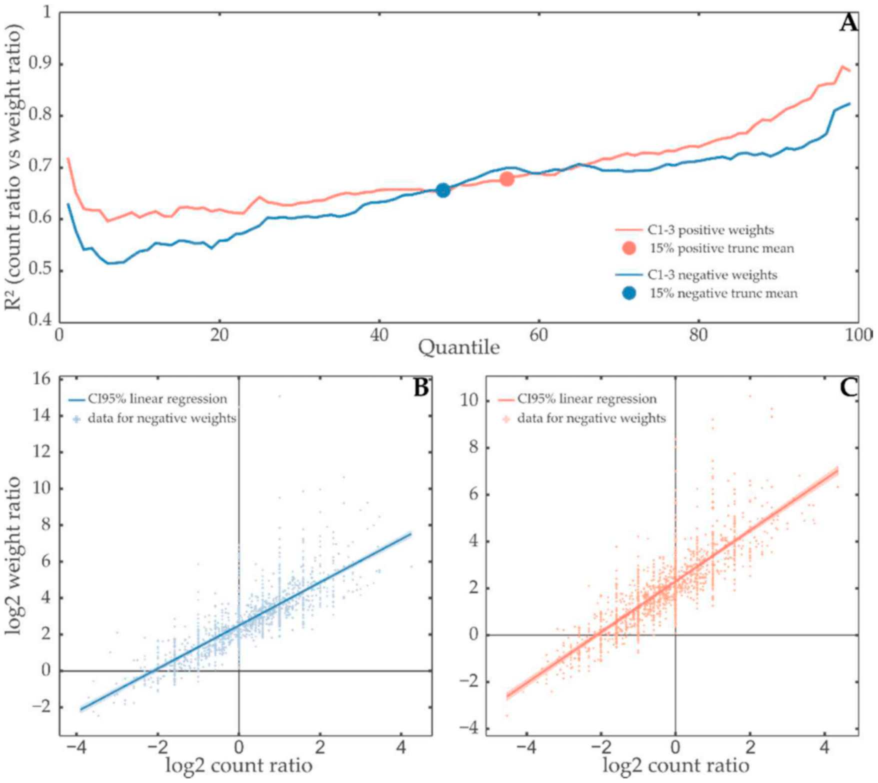


Figure 2. Contribution weights in constraints carry information beyond the count of extreme values. (**A**) R^2 as a measure of linear regression quality of weight sum ratio (WR) by count ratio (CR) is shown in dependence of the quantile used to split the weight distributions of annotated sets to generate these ratios in either all positive (red) or all negative (blue) weight distributions for annotated subsets of constraints 1–3. The ±15% truncated mean of each is shown as a point of the same color. The quantile that separates weights in constraints 1–3 so that it produces the 15% truncated mean R^2 regression quality shown in the upper part of the figure was used to split the weights in positively (56% quantile, (**C**)) and negatively (48% quantile, (**B**)) contributing parts of annotated subsets of constraints 1–3. Subsequently, both WR and CR were calculated for all the annotated subsets in the dataset. These values are shown as either red (right) or blue (left) points on the scatter plots. Linear regression was performed, and the resulting line was plotted with a 95% confidence band. These plots correspond to the regressions for a single y-value on the top plot. The existence and increase of outliers in the high weight/count ratio region suggests that high weight items carry an especially large amount of information that is lost when using traditional methods.

Based on these considerations, we can qualitatively classify three kinds of weight contributions: (1) cases where the overall distribution is shifted to more extreme values (i.e., the ‘majority vote’ case), (2) cases where a single or small amount of entities causes a whole functionally annotated set (FAS) to be reported as significantly altered, and (3) cases where a subset of the FAS is strongly skewed to extreme values, with cases (2) and (3) representing the aforementioned complementary results. Practical examples for each case are displayed in Figure 3. (1) The FAS *protein.synthesis.ribosomal protein* is reported to be significantly positively contributing to Constraint 1, with most of the entities being slightly more extreme than the overall weight distribution (Figure 3A), satisfying stoichiometric requirements during the regulation of large protein complexes [60]. Conversely, (2) *signaling.light* has a low amount of extreme contributions to Constraint 2, but two of them are sufficient to make the whole subset be reported as significant (Figure 3B). Finally, (3) the weights of a medium-sized subgroup of transcription factors in *RNA.regulation of transcription.MYB-related transcription factor family protein* show a distribution that is not reflected in the rest of the FAS (Figure 3C).

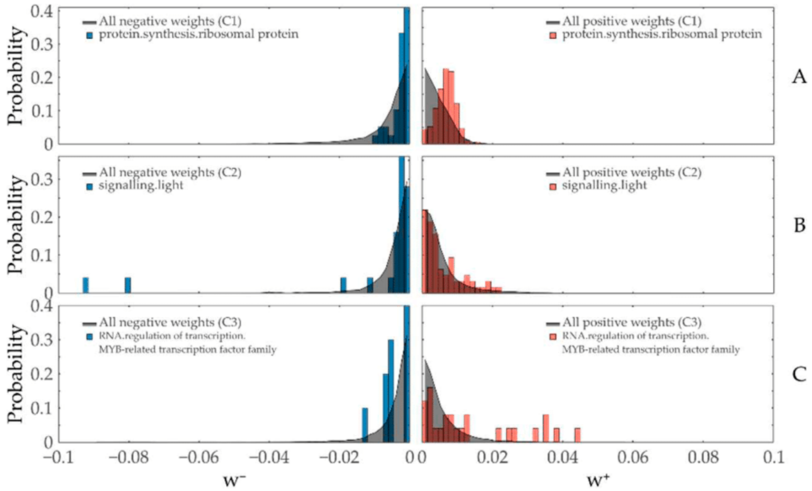


Figure 3. TMEA reports different weight distribution shapes for annotated subsets as significant. Histograms with a bin width of 0.0015 of both negatively (left part of the plots, blue) and positively (right part of the plots, red) contributing entities in the functionally annotated sets of (**A**) *protein.synthesis.ribosomal protein* for Constraint 1, (**B**) *signaling.light* for Constraint 2, and (**C**) *RNA.regulation of transcription.MYB-related transcription factor family protein* for Constraint 3 are plotted together with the respective overall distribution of weights (gray area plot) of the respective sign and constraint.

3.3. Comparison with Hypergeometric Test Based GSEA

In order to demonstrate the performance of the presented tandem approach, we compared our results from applying TMEA to transcriptomics data of a high light acclimation experiment to standard enrichment analysis based on hypergeometric distribution (hypGSEA). hypGSEA was performed for terms of transcripts that showed differential expression during the experiment time course (see Section 2.4). For TMEA, a statistical pre-analysis for binary entity labeling is not necessary, thereby eliminating bias resulting from preparatory analysis of the input. The size of entities grouped by one shared functional annotation often lies in the range of 5–50. Especially in small bin sizes (<50), the discrete nature of the hypergeometric distribution used in hypGSEA potentially leads to a lower significance level than intended (Figure A1). This loss of power could be mitigated by using a mid-*p*-value, which entails a risk of a significance level that is above the intended one [26,61] and therefore was not applied in this study.

On our light acclimation benchmark dataset, hypGSEA yields a set of 74 significant FASs. TMEA identified 103 FASs with significant contributions to constraints 1–3 and 97 FASs with a significant influence on constraints 4–10. Fifty-nine of the significant FASs are reported by both TMEA for constraints 1–3 and hypGSEA, leading to 15 FASs (12.7% of all reported FASs by hypGSEA and TMEA) exclusively reported by GSEA, and 44 exclusively reported by TMEA (37.3%) (Figure 4).

Although the intersect of TMEA and hypGSEA significant FASs is large, no strong correlation between both *p*-values can be seen (Figure 4B,C). Especially, FASs that are reported to be significant in constraints with lower priority (constraints 2 or 3) show increased *p*-values for respective GSEA tests and vice versa. With an increasing constraint index, the relevance of FASs significantly contributing to the respective constraint diminishes. While the reported FASs show significant impact to these constraints, the constraints themselves may be of minor importance to the current condition. In a comparison without threshold, 39 unique FASs are reported by constraints 4–10 that are not contained in constraints 1–3 (Figure 4A). More than half (51.3%) of these FASs show a high functional similarity and differ only in the level of detail encoded by the depth within the ontology tree (Table S4). However, it is currently common practice to only consider constraints that account for the majority of information in the dataset (Figure A2) [38,41,44].

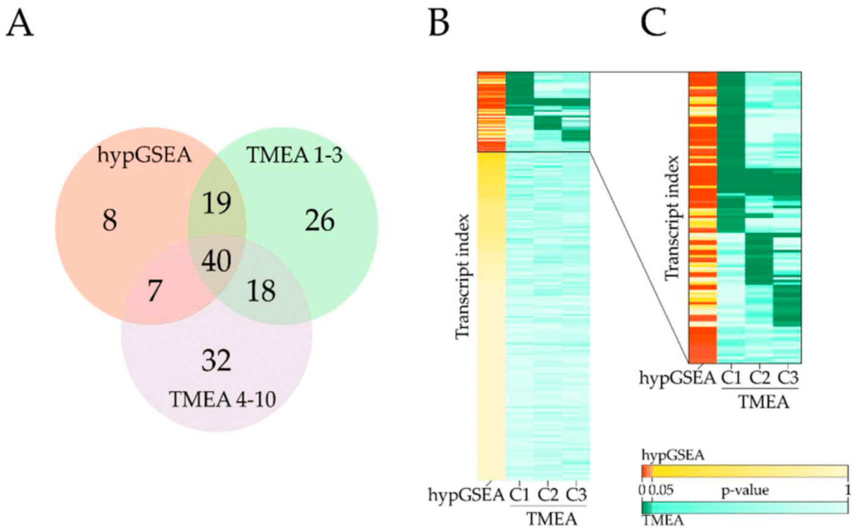


Figure 4. Comparison of significant functionally annotated sets (FASs) obtained by hypergeometric distribution (hypGSEA) and TMEA. (**A**) Venn diagrams of significant FASs with a minimum size of 5 reported by hypGSEA, TMEA constraints 1–3, and TMEA constraints 4–10 for a comparison without threshold. (**B**) Heatmap of adjusted *p*-values obtained by hypGSEA and TMEA. Measured transcripts were labeled with their respective hypGSEA *p*-value and the minimal TMEA *p*-value obtained within the first three constraints. All TMEA-significant bins are clustered by k-means clustering with k = 6. (**C**) Visualization of all FAS reported significant by hypGSEA and/or TMEA. Detailed cluster information is given in Table A1. Bins that are not reported by TMEA are appended to the end of the heatmap with increasing hypGSEA *p*-values.

3.4. Case Study: Characterization of Light Acclimation in *Arabidopsis thaliana*

Since the understanding of a plant's light response is of fundamental importance for future crop breeding and cultivation strategies, there has been a research focus on the acclimation to various light conditions, making light acclimation a suitable benchmark dataset. Furthermore, we focus on the transcripts as a proxy that influences the state of all levels: the proteome and, linked by proteins, the metabolome, lipidome, and even the phenome to some extent. So, most energy-consuming reactions or transitions are relying on transcripts, which makes them a feasible entry point to benchmark TMEA by relating observations previously not discovered on transcript but rather different system levels.

TMEA analysis based on transcript amounts measured during light acclimation reveals functional descriptions for the different thermodynamic states of the biology identified by SA. The dominant state variable (λ_1) indicates the existence of two major states by undergoing a state transition (changing its sign) between two and four days of high light acclimation. This coincides with an energy investment governed by the first constraint (Figure 5B). Here, TMEA identifies major metabolic functions such as amino acid, lipid, and nucleotide metabolism as well as protein transport to be characteristic processes significantly contributing to energy investments. Calcium signaling shows the inverse contribution regarding the identified states of Constraint 1. In state variable λ_2, two state transitions seem to occur during the early phases of acclimation and de-acclimation, respectively (15 min to 3 h of treatment). A local energy minimum for this constraint can be observed at the same time as the state transition described by λ_1. The functional characterization of Constraint 2 by TMEA reveals a positive contribution of photosystem light reaction, sugar transport, and trehalose metabolism and an inverse contribution of light signaling. Three state transitions in λ_3 point to a more refined state shifting that subdivides the experimental time course into (1) an immediate acclimation response (0–15 min), (2) early acclimation (3 h), (3) late acclimation and condition change (2 days of acclimation to 15 min of de-acclimation), and (4) central de-acclimation (3 h to 4 days of de-acclimation). Naturally, the contributions of the third constraint to the overall free energy are low, but they are sufficient to be responsible for a third overall energy minimum at 3 h of de-acclimation. The dominant processes that characterize this constraint are major carbon degradation, sulfate transport, transcriptional regulation, and phenylpropanoid synthesis. In the following biological examination, we demonstrate that TMEA results obtained in our benchmark dataset seem to be biologically sound according to the current biological understanding of light acclimation.

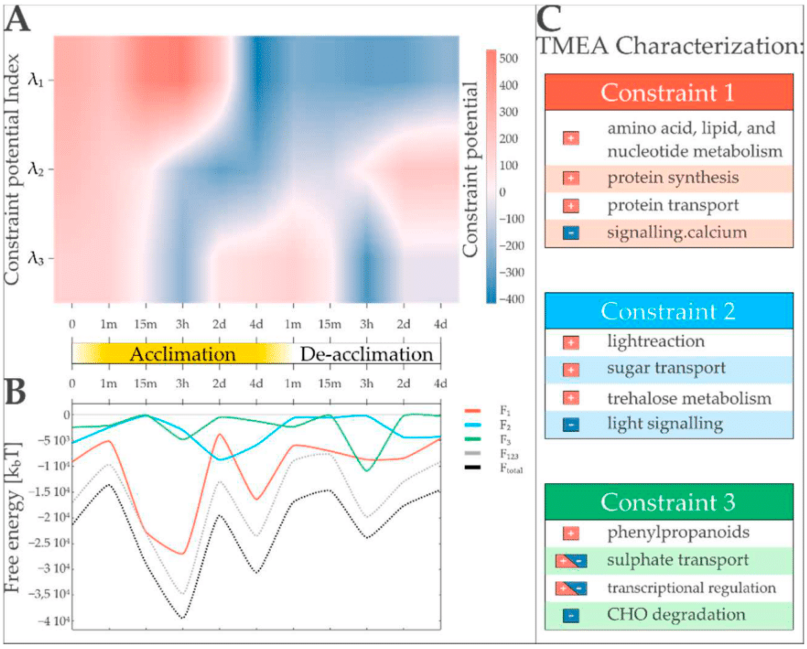


Figure 5. TMEA and surprisal analysis identify three major transcription patterns governing high light acclimation in *Arabidopsis thaliana* and provide a concise functional description for them. (**A**) Time course of the three major constraint potentials (λ_α for $\alpha = 1,2,3$) indicate the importance of the respective transcription pattern. The potentials of the first three constraints (λ_1–λ_3) are shown for four days of acclimation and four days of subsequent de-acclimation. While λ_1 separates the experiment in two major phases, λ_2 and λ_3 show more fluctuating patterns, defining three or four states, respectively. (**B**) Free energy landscapes defined by the three major state variables. Energy levels are plotted for transcription patterns (F_1–F_3), their sum (F_{123}), and the total free energy when using all constraints for free energy calculation (F_{total}). The dominant pattern is responsible for two of the three visible local energy minima. The least weighted pattern of the three is responsible for an energy minimum at the end of the time course. (**C**) Selected FASs reported by TMEA with significant influences on the respective constraints are listed. Directional influence (+ for positive, – for inverse) on the respective pattern is indicated.

3.4.1. Anthocyanins

A well-known response to high light treatment in plants is the accumulation of anthocyanins, preventing photoinhibitory damage caused by high irradiance [62,63]. In photosynthetic active tissue, the dyes absorb excess radiation, thereby minimizing oxidative damage for e.g., the photosystems or DNA [63–66]. After onset of the highlight treatment, a significant anthocyanin accumulation was observed that increased during the 4 days of acclimation from ≈2 to 20 $A{\cdot}g\ FW^{-1}$ before decreasing to a constant level of ≈8 $A{\cdot}g\ FW^{-1}$ during de-acclimation (Figure 6A).

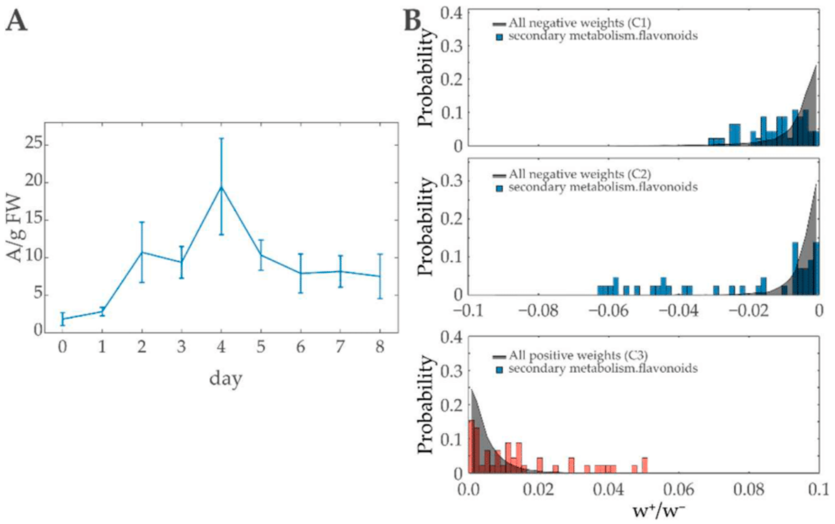


Figure 6. The role of Anthocyanins during high light treatment: (**A**) Anthocyanin content in *Arabidopsis thaliana* under 4 days of high light treatment (days 0–4) and 4 days of de-acclimation at ambient light condition (days 4–8). (**B**) Weight distributions of transcripts included in *secondary metabolism.flavonoids* demonstrating significant influences for constraints 1–3. TMEA reports a significance for the weight sums of all three constraints.

The enrichment analysis in previous work [43] identified *flavonoid biosynthesis* to be significantly overrepresented in the same transcriptomics data utilized in this publication. Anthocyanins thereby are included due to the fact that flavonoids is a collective term for a huge variety of chemical compounds including anthocyanins [67]. hypGSEA using MapMan-Ontology also indicates an enrichment of the FAS *secondary metabolism.flavonoids, secondary metabolism.flavonoids.anthocyanins*, and further related FASs (Table S1). Light-protecting dyes have a significant role during high light response, ensuring the survival of the plant. TMEA recovers this importance by reporting anthocyanin and flavonoid-related FASs to be of significant importance in all considered major constraints (Figure 6B).

3.4.2. Myb-Related Transcription Factor Family

A FAS solely detected by TMEA is *RNA.regulation of transcription.MYB-related transcription factor family*. Although based on the same dataset, neither the published enrichment [43] nor hypGSEA detected the respective FAS; however, biological relevance in high light response was discovered in previous studies. In [43], a motif search was performed within the 1000-bp promotor sequences of 456 genes and identified an overrepresented motif, which is bound by the members of Myb, and Myb-related-TF families, indicating a role in acclimation responses. The weights of the transcripts associated to this FAS were sufficient to report the importance in Constraint 3 using TMEA (see Figure 3C). The TF family is involved in the regulation of phenylpropanoid biosynthesis, which in turn is linked to lignin synthesis and UV protection [68,69]. Both hypGSEA and TMEA reported the phenylpropanoid biosynthesis to be enriched only taking transcripts into account. Particularly to Constraint 3, high weights are associated to both FASs (Table S2). As described in Section 3.4, the potential time course of Constraint 3 subdivides acclimation and de-acclimation in an early and late response (respectively).

One of the major metabolites that is required for phenylpropanoid synthesis and therefore is linked to Myb TF families is phenylalanine [69]. The metabolomics analysis conducted in parallel to the transcriptomics sampling reveals a distinct/prominent signal shape that quadrupled during the first day of acclimation, prior to returning to its original state during the high light phase. In the first day of the de-acclimation, the amount of phenylalanine quadrupled again and remained at high levels until the end of four days of de-acclimation. This characteristic shape resembles the time course of the potential of Constraint 3 (Figures 5A and 7), where both phenylpropanoid biosynthesis and the Myb family show a significant importance. Of the 22 transcripts that can be assigned to phenylalanine metabolism by KEGG, 14 are directly associated to amino acid metabolism. Of the remaining eight transcripts, four can be assigned to phenylpropanoid synthesis.

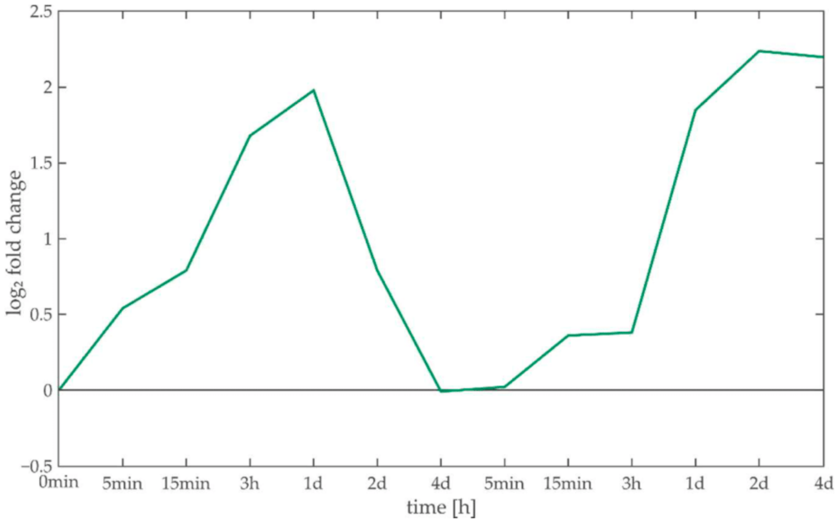


Figure 7. Phenylalanine time course. Phenylalanine fold changes during 4 days of high light acclimation and 4 days of de-acclimation under ambient conditions show increased abundance 3 h to 1 day after condition change.

3.4.3. Ribosomes

Changes in environmental conditions make it necessary to rearrange the cellular proteome, which partially must be facilitated by the synthesis of new proteins at ribosomes. MapMan is exhaustive in the characterization and subdivision of ribosomal protein families. The measured transcripts are linked to 20 FASs related to *protein.synthesis.ribosomal protein*. Eight of these are associated to significantly enriched FASs in the TMEA analysis (Table S1) with nuclear as well as plastidic ribosome annotations among them. The third level FAS *protein.synthesis.ribosomal protein* contains 384 transcripts, of which 345 with positive weights to Constraint 1 show a characteristic shape (Figure 3A). Most of the weights show a constant shift toward higher influence, which is characteristic for protein complexes that rely on a stoichiometric relationship.

3.4.4. Light/Calcium Signaling

Changes in the environment are perceived by plants and must be passed onto the responsible organs in order to take appropriate measures. Sometimes, it is sufficient to perform all steps within a single cell, so that the environmental information is perceived, processed, and reacted to without multi-cell communication [70]. Hormones and other signaling molecules serve as messengers for changes that must be communicated across several tissue types and functional units such as the shoot, root, or stem [71]. While the importance of three signaling-related FASs were identified by both hypGSEA and TMEA (*signlling*, *signaling.in sugar and nutrient physiology*, and *signaling.receptor kinases*), two additional FASs were reported exclusively by TMEA. Namely, *signaling.calcium* and *signaling.light* showed significant importance to constraint 2 or 3 respectively.

In FAS *signaling.light*, two genes were given particularly high weights. These two genes are *early light-induced protein 1* (ELIP1) and ELIP2 (AT3G22840 and AT4G14690), which both show a high upregulation upon high light treatment [72,73]. In fact, ELIP2 shows the overall highest negative weight in Constraint 2. Both are regulated by UVR8 [74] and CRY1 [73]. They are supposed to protect the plant cells from photo-oxidative stress [75,76] and play an important role in chlorophyll synthesis regulation [77].

Calcium ions are one of the most used intracellular second messengers in plants. Many environmental conditions trigger calcium-dependent signaling cascades, eventually leading to the activation of kinases responsible for appropriate stress responses [78,79]. TMEA identified the negative FAS weights to be significant in the most contributing constraint (constraint 1).

4. Discussion

Evaluating the performance of a GSEA method is challenging, as it is difficult to know which gene sets should be considered as true positives. A common approach is to simulate data to validate a particular method [80–82]. However, the validity of this approach is debatable, as the model used for the simulation strongly influences the results [28].

In this paper, we presented a novel approach to gene set enrichment analysis that is based on surprisal analysis (SA) and captures both biological functional knowledge and thermodynamic state description. We presented our rationale and formulation of the approach and applied it comparatively to hypergeometric test-based GSEA on a large transcriptomic dataset. To that extend, we could show that our proposed method can recover the functional knowledge extracted by the GSEA methods most frequently applied in comparable studies. Furthermore, we were able to report an array of additional biologically relevant findings based on transcriptional changes only that are in line with current literature knowledge and evidently emerge from its thermodynamic substantiation. For systemic acclimation responses, a proteome rearrangement is fundamental and well-studied. While under high light conditions, light harvesting is of minor importance, energy handling, energy distribution, and light protection become critical. Photoprotective mechanisms must be activated immediately without transcriptional reorganization and an extensive loss of time, so prearranged mechanisms are activated by post-translational modifications [83,84]. On the other hand, long-term and non-vital responses required within seconds can be regulated translationally. Most if not all reactions/transitions within an organism have their fundamental cause in the generation of catalyzing enzymes, whose abundances are in turn realized by transcriptional changes. It should be stressed though that this approach to validate TMEA is by no means perfect, as the process of previous knowledge discovery can also be biased by the methods applied by the different authors; however, it is thoroughly manually evaluated by an expert community.

Additionally, we believe that our approach is especially suited to analyze acclimation response on a systems level. Since biological systems always are under change, e.g., because of developmental issues or circadian rhythms, often a reference is desired to which the treated organism is compared. Two common procedures rely on (i) a control organism/culture monitored simultaneously to the treated one or (ii) a specific time point prior to the treatment that is taken as reference for the identification of condition responses. Both methods lack in robustness since (i) treated organisms behave in a different manner compared to control organisms, especially when treated with a systemic disturbance or during phases of development, and (ii) a single reference point can lead to massive misjudgments if the measurements are affected by an experimental bias. In previous studies, it could be shown that a thermodynamic viewpoint using SA alone already improves the understanding of responses to systems perturbation in plants [85–87]. However, we could demonstrate in this work that while SA is able to reveal states of the transcription system during acclimation, TMEA elucidates the subjacent pathways, contributing to these states. Thereby, TMEA provides a thermodynamic interpretation of the importance of functionally annotated sets (FASs).

In our transcript dataset, this leads to the novel finding of three stable states during light acclimation of *Arabidopsis thaliana* and allows for the distinction of functionally different phases during the acclimation response. The first stable state at 3 h of perturbation (Figure 5B) indicates an energy-intensive early acclimation phase, coinciding with the highest overall energy dissipation of the transcript system. To this state, only the first state variable is contributing meaningfully. TMEA characterization of the first transcription pattern informs that the energy sinks of the transcription system for this state are mainly metabolic pathways and protein synthesis, with a focus on ribosomal proteins (Figure 5 right, Table S1). The second stable state of the transcript system is identified at the last time point of acclimation treatment (4 days, Figure 5B) and can be interpreted as the acclimated state of the system, where energy is invested in the same pathways as in the first stable state, but possibly to maintain the long-term acclimation. The third stable state is reached in the early phase of de-acclimation (3 h, Figure 5B), with the third transcription pattern as the main energy sink. One of the central functions characterized to be significantly contributing to this transcription pattern is that of the various transcriptional regulators (Figure 5 right, Table S1). We hypothesize that this may be an indication for priming [88] of the transcript system for future responses to high light conditions. It is important to note that the energy investments in Transcription Pattern 2 are not leading to local energy minima. Interestingly, the time point at which the most work is done by this pattern (2 days into the acclimation phase of the experiment, Figure 5B) coincides with an overall local energy maximum, therefore lowering the overall energy level of the transcription system at this point. TMEA functionally associates this pattern mainly with light signaling and light reaction-related pathways (Figure 5 right, Table S1). These functional characterizations together with the fact that this pattern is not responsible for stable states leads us to the assumption that it is mainly responsible to lower the energy barriers that have to be overcome by the transcript system to reach its stable states, indicating that TMEA can separate regulatory patterns from enzymatic ones.

For future work, it might be beneficial to extent TMEA for the analysis of multivariate datasets using the multivariate version of the SA [89]. This would allow integrating information from different systems levels for the thermodynamically motivated functional characterization of biological responses to system acclimation. Furthermore, additional—and more practical—knowledge may be gained when comparing TMEA characterizations of different plants over the same condition, especially when applied to crop species or even organisms from another branch of life. So far, we provide an implementation of the whole analysis framework to facilitate the application of TMEA on different datasets using specific functional gene and pathway annotation databases. As more knowledge is collected and curated in those databases, we believe that TMEA will be increasingly useful for researchers especially studying systems acclimation responses.

Supplementary Materials: The following are available online at http://www.mdpi.com/1099-4300/22/9/1030/s1, Table S1: Comparison of significantly contributing FAS, Table S2: Detailed TMEA result, Table S3: Detailed hypGSEA result, Table S4: Comparison of significantly contributing FASs in TMEA.

Author Contributions: Conceptualization, K.S., B.V. and T.M.; methodology, K.S., B.V. and T.M.; software, K.S.; analysis and investigation, K.S. and B.V.; writing—original draft preparation, K.S., B.V. and T.M.; writing—review and editing, K.S., B.V. and T.M.; visualization, K.S. and B.V.; supervision, T.M.; funding acquisition, T.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the TRR 175 (project D02) and the Landesforschungsschwerpunkt BioComp.

Acknowledgments: We thank Raphael D. Levine for fruitful discussion and support. Additionally, we thank Antoni Garcia-Molina and Dario Leister for kindly providing the experimental data we used as the benchmark dataset.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DEG	Differentially expressed gene
FAS	Functionally annotated set
FCS	Functional Class Scoring
FDR	False discovery rate
GSEA	Gene set enrichment analysis
hypGSEA	Gene set enrichment analysis based on hypergeometric tests
TMEA	Thermodynamically motivated enrichment analysis

SA	Surprisal analysis
SS	Single-Sample
SVD	singular value decomposition

Appendix A

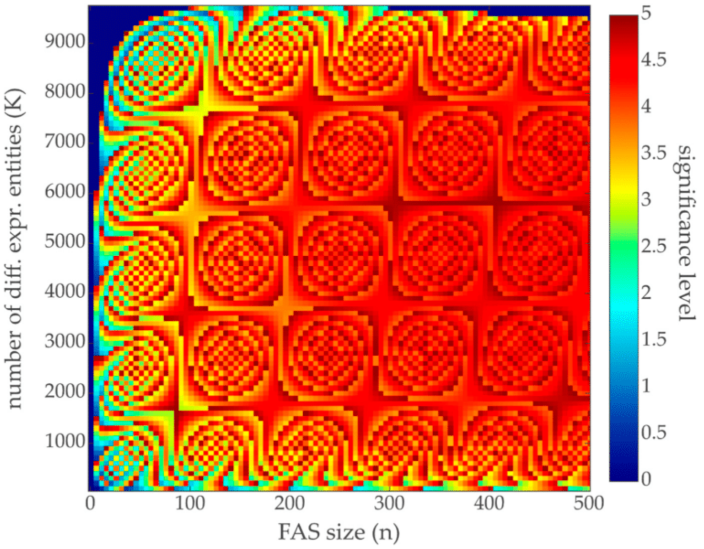


Figure A1. Maximal reachable α-level at a given α-level of 5%. The discrete nature of the hypergeometric distribution prevents the significance to reach 0.05 exactly. There always is a range of α-level space that must be sacrificed leading to a lower α than intended. The heatmap shows the maximal reachable α-level given: N = total number of genes = 10,000; K = number of differentially expressed genes; n = bin size; k = minimal number of differentially expressed genes needed for p-value < 0.05; intended α-level = 0.05. Especially when the bin size is low, even the half of the intended α-level often cannot be reached. Note that the bin size ranges from 1 to 500 in steps of 5.

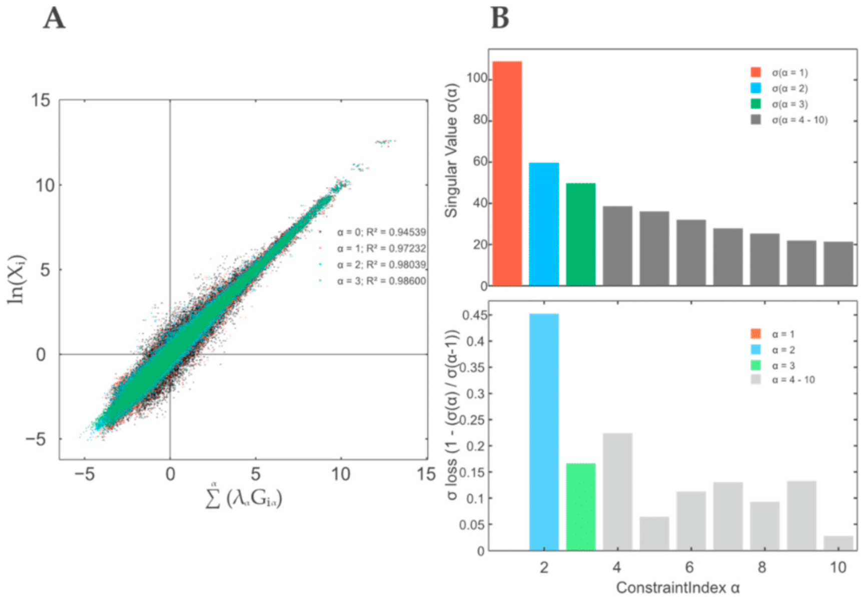


Figure A2. Constraint relevance. (**A**) Data reconstruction obtained by using (i) baseline state (constraint 0), (ii) Constraints 0–1, (iii) Constraints 0–2, and (iii) Constraints 0–3. (**B**) Singular values of constraints 1–10. The combination of a reconstruction efficiency of 98.6% and the singular value amplitude drop at α = 4 with no strong further decrease indicates a sufficient information supply by constraints 1–3.

Table A1. Significant FASs reported by TMEA in constraints 1–3. The *p*-values were clustered using the k-means clustering algorithm with a cluster number of 6 (cluster ID 1–6). The corresponding heatmap is depicted in Figure 4.

cID	MapMan Annotation (FAS)	cID	MapMan Annotation (FAS)
1	cell wall.cell wall proteins	4	major CHO metabolism
1	cell wall.cell wall proteins.AGPs	4	major CHO metabolism.degradation
1	cell wall.cell wall proteins.AGPs.AGP	4	major CHO metabolism.degradation.starch
1	cell wall.pectin*esterases.misc	4	misc.invertase/pectin methylesterase inhibitor family protein
1	lipid metabolism.FA desaturation	4	not assigned.no ontology.DC1 domain containing protein
1	lipid metabolism.FA desaturation.desaturase	4	not assigned.unknown
1	misc.beta 1,3 glucan hydrolases	4	RNA.regulation of transcription.AP2/EREBP, APETALA2/Ethylene-responsive element binding protein family
1	misc.beta 1,3 glucan hydrolases.glucan endo-1,3-beta-glucosidase	4	RNA.regulation of transcription.C2C2(Zn) CO-like, Constans-like zinc finger family
1	misc.glutathione S transferases	4	RNA.regulation of transcription.C2C2(Zn) DOF zinc finger family
1	misc.nitrilases, *nitrile lyases, berberine bridge enzymes, reticuline oxidases, troponine reductases	4	RNA.regulation of transcription.MYB-related transcription factor family
1	misc.O-methyl transferases	4	RNA.regulation of transcription.Psudo ARR transcription factor family
1	misc.protease inhibitor/seed storage/lipid transfer protein (LTP) family protein	4	secondary metabolism.isoprenoids.terpenoids
1	nucleotide metabolism.synthesis.purine	4	secondary metabolism.phenylpropanoids.lignin biosynthesis
1	protein	4	stress.abiotic
1	protein.degradation.AAA type	4	stress.abiotic.cold
1	protein.synthesis	4	stress.biotic.respiratory burst
1	protein.synthesis.ribosomal protein	4	transport.sulfate
1	protein.synthesis.ribosomal protein.eukaryotic	5	cell wall
1	protein.synthesis.ribosomal protein.eukaryotic.40S subunit	5	cell wall.modification
1	protein.synthesis.ribosomal protein.eukaryotic.60S subunit	5	misc
1	protein.synthesis.ribosomal protein.prokaryotic.chloroplast	5	secondary metabolism
1	protein.synthesis.ribosomal protein.prokaryotic.chloroplast.50S subunit	5	secondary metabolism.flavonoids
1	protein.synthesis.ribosome biogenesis	5	secondary metabolism.flavonoids.anthocyanins
1	protein.synthesis.ribosome biogenesis.Pre-rRNA processing and modifications	5	secondary metabolism.flavonoids.anthocyanins.anthocyanin 5-aromatic acyltransferase
1	protein.synthesis.ribosome biogenesis.Pre-rRNA processing and modifications.snoRNPs	5	secondary metabolism.flavonoids.dihydroflavonols

Table A1. *Cont.*

cID	MapMan Annotation (FAS)	cID	MapMan Annotation (FAS)
1	protein.synthesis.ribosome biogenesis.Pre-rRNA processing and modifications.WD-repeat proteins	5	stress
1	redox.glutaredoxins	5	stress.biotic
1	RNA.regulation of transcription.ARR	5	transport
1	RNA.regulation of transcription.NAC domain transcription factor family	6	cell wall.degradation
1	RNA.regulation of transcription.WRKY domain transcription factor family	6	cell wall.degradation.mannan-xylose-arabinose-fucose
1	secondary metabolism.simple phenols	6	DNA.synthesis/chromatin structure.retrotransposon/transposase
1	signaling	6	DNA.synthesis/chromatin structure.retrotransposon/transposase.gypsy-like retrotransposon
1	signaling.in sugar and nutrient physiology	6	hormone metabolism
1	signaling.receptor kinases.DUF 26	6	hormone metabolism.auxin
1	signaling.receptor kinases.misc	6	minor CHO metabolism
1	signaling.receptor kinases.wall associated kinase	6	minor CHO metabolism.trehalose
1	signaling.receptor kinases.wheat LRK10 like	6	minor CHO metabolism.trehalose.potential TPS/TPP
1	stress.biotic.PR-proteins.plant defensins	6	misc.gluco-, galacto- and mannosidases
1	transport.Major Intrinsic Proteins	6	not assigned.no ontology.glycine rich proteins
2	amino acid metabolism.synthesis	6	not assigned.no ontology.pentatricopeptide (PPR) repeat-containing protein
2	amino acid metabolism.synthesis.aspartate family	6	PS.lightreaction
2	development.storage proteins	6	PS.lightreaction.photosystem II
2	hormone metabolism.auxin.induced-regulated-responsive-activated	6	PS.lightreaction.photosystem II.LHC-II
2	nucleotide metabolism.synthesis	6	secondary metabolism.flavonoids.chalcones
2	protein.synthesis.ribosomal protein.eukaryotic.60S subunit.L7A	6	secondary metabolism.flavonoids.flavonols
2	protein.synthesis.ribosomal protein.prokaryotic	6	secondary metabolism.phenylpropanoids
2	signaling.calcium	6	signaling.light
2	stress.biotic.receptors	6	transport.ABC transporters and multidrug resistance systems
2	transport.Major Intrinsic Proteins.PIP	6	transport.sugars
3	misc.cytochrome P450		
3	misc.GDSL-motif lipase		
3	misc.peroxidases		
3	signaling.receptor kinases		
3	stress.biotic.PR-proteins		

References

1. Ruffel, S.; Krouk, G.; Coruzzi, G.M. A systems view of responses to nutritional cues in Arabidopsis: Toward a paradigm shift for predictive network modeling. *Plant Physiol.* **2010**, *152*, 445–452. [CrossRef]
2. Anjum, N.A. Plant acclimation to environmental stress: A critical appraisal. *Front. Plant Sci.* **2015**. [CrossRef]
3. Raza, A.; Razzaq, A.; Mehmood, S.S.; Zou, X.; Zhang, X.; Lv, Y.; Xu, J. Impact of Climate Change on Crops Adaptation and Strategies to Tackle Its Outcome: A Review. *Plants* **2019**, *8*, 34. [CrossRef] [PubMed]
4. Minorsky, P.V. Achieving the in Silico Plant. Systems Biology and the Future of Plant Biological Research. *Plant Physiol.* **2003**, *132*, 404–409. [CrossRef]
5. Beine-Golovchuk, O.; Firmino, A.A.P.; Dąbrowska, A.; Schmidt, S.; Erban, A.; Walther, D.; Zuther, E.; Hincha, D.K.; Kopka, J. Plant Temperature Acclimation and Growth Rely on Cytosolic Ribosome Biogenesis Factor Homologs. *Plant Physiol.* **2018**, *176*, 2251–2276. [CrossRef]
6. Brouwer, P.; Bräutigam, A.; Buijs, V.A.; Tazelaar, A.O.E.; van der Werf, A.; Schlüter, U.; Reichart, G.-J.; Bolger, A.; Usadel, B.; Weber, A.P.M.; et al. Metabolic Adaptation, a Specialized Leaf Organ Structure and Vascular Responses to Diurnal N_2 Fixation by Nostoc azollae Sustain the Astonishing Productivity of Azolla Ferns without Nitrogen Fertilizer. *Front. Plant Sci.* **2017**, *8*, 442. [CrossRef]
7. Hemme, D.; Veyel, D.; Mühlhaus, T.; Sommer, F.; Jüppner, J.; Unger, A.-K.; Sandmann, M.; Fehrle, I.; Schnfelder, S.; Steup, M.; et al. Systems-Wide Analysis of Acclimation Responses to Long-Term Heat Stress and Recovery in the Photosynthetic Model Organism Chlamydomonas reinhardtii. *Plant Cell* **2014**, *26*, 4270–4297. [CrossRef] [PubMed]
8. Mettler, T.; Mühlhaus, T.; Hemme, D.; Schöttler, M.-A.; Rupprecht, J.; Idoine, A.; Veyel, D.; Pal, S.K.; Yaneva-Roder, L.; Winck, F.V.; et al. Systems Analysis of the Response of Photosynthesis, Metabolism, and Growth to an Increase in Irradiance in the Photosynthetic Model Organism Chlamydomonas reinhardtii. *Plant Cell* **2014**, *26*, 2310–2350. [CrossRef]
9. Rademacher, N.; Wrobel, T.J.; Rossoni, A.W.; Kurz, S.; Bräutigam, A.; Weber, A.P.M.; Eisenhut, M. Transcriptional response of the extremophile red alga Cyanidioschyzon merolae to changes in CO_2 concentrations. *J. Plant Physiol.* **2017**, *217*, 49–56. [CrossRef] [PubMed]
10. Schmollinger, S.; Mühlhaus, T.; Boyle, N.R.; Blaby, I.K.; Casero, D.; Mettler, T.; Moseley Jeffrey, L.; Kropat, J.; Sommer, F.; Strenkert, D.; et al. Nitrogen-Sparing Mechanisms in Chlamydomonas Affect the Transcriptome, the Proteome, and Photosynthetic Metabolism. *Plant Cell* **2014**, *26*, 1410–1435. [CrossRef] [PubMed]
11. Valledor, L.; Furuhashi, T.; Hanak, A.-M.; Weckwerth, W. Systemic Cold Stress Adaptation of Chlamydomonas reinhardtii*. *Mol. Cell Proteom.* **2013**, *12*, 2032–2047. [CrossRef] [PubMed]
12. Zandalinas, S.I.; Sengupta, S.; Burks, D.; Azad, R.K.; Mittler, R. Identification and characterization of a core set of ROS wave-associated transcripts involved in the systemic acquired acclimation response of Arabidopsis to excess light. *Plant J.* **2019**, *98*, 126–141. [CrossRef] [PubMed]
13. Zuther, E.; Schaarschmidt, S.; Fischer, A.; Erban, A.; Pagter, M.; Mubeen, U.; Giavalisco, P.; Kopka, J.; Sprenger, H.; Hincha, D.K. Molecular signatures associated with increased freezing tolerance due to low temperature memory in Arabidopsis. *Plant Cell Environ.* **2019**, *42*, 854–873. [CrossRef]
14. Thimm, O.; Blasing, O.; Gibon, Y.; Nagel, A.; Meyer, S.; Kruger, P.; Selbig, J.; Mller, L.A.; Rhee, S.Y.; Stitt, M. MAPMAN: A user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J. Cell Mol. Biol.* **2004**, *37*, 914–939. [CrossRef]
15. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene ontology: Tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef]
16. Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [CrossRef] [PubMed]
17. Kelder, T.; van Iersel, M.P.; Hanspers, K.; Kutmon, M.; Conklin, B.R.; Evelo, C.T.; Pico, A.R. WikiPathways: Building research communities on biological pathways. *Nucleic Acids Res.* **2012**, *40*, D1301-7. [CrossRef] [PubMed]
18. Karp, P.D.; Ouzounis, C.A.; Moore-Kochlacs, C.; Goldovsky, L.; Kaipa, P.; Ahrén, D.; Tsoka, S.; Darzentas, N.; Kunin, V.; Lpez-Bigas, N. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.* **2005**, *33*, 6083–6089. [CrossRef] [PubMed]

19. Al-Shahrour, F.; Díaz-Uriarte, R.; Dopazo, J. FatiGO: A web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* **2004**, *20*, 578–580. [CrossRef]
20. Huang, D.W.; Sherman, B.T.; Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **2009**, *4*, 44–57. [CrossRef]
21. Zeeberg, B.R.; Feng, W.; Wang, G.; Wang, M.D.; Fojo, A.T.; Sunshine, M.; Narasimhan, S.; Kane, D.W.; Reinhold, W.C.; Lababidi, S.; et al. GoMiner: A resource for biological interpretation of genomic and proteomic data. *Genome Biol.* **2003**, *4*, R28. [CrossRef] [PubMed]
22. Zhong, S.; Storch, K.-F.; Lipan, O.; Kao, M.-C.J.; Weitz, C.J.; Wong, W.H. GoSurfer: A graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space. *Appl. Bioinform.* **2004**, *3*, 261–264. [CrossRef] [PubMed]
23. Zhou, X.; Su, Z. EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species. *BMC Genom.* **2007**, *8*, 246. [CrossRef]
24. Zhang, B.; Schmoyer, D.; Kirov, S.; Snoddy, J. GOTree Machine (GOTM): A web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinform.* **2004**, *5*, 16. [CrossRef]
25. Maere, S.; Heymans, K.; Kuiper, M. BiNGO: A Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **2005**, *21*, 3448–3449. [CrossRef] [PubMed]
26. Rivals, I.; Personnaz, L.; Taing, L.; Potier, M.-C. Enrichment or depletion of a GO category within a class of genes: Which test? *Bioinformatics* **2007**, *23*, 401–407. [CrossRef]
27. Pan, K.-H.; Lih, C.-J.; Cohen, S.N. Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 8961–8965. [CrossRef]
28. Tarca, A.L.; Bhatti, G.; Romero, R. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PLoS ONE* **2013**, *8*, e79217. [CrossRef]
29. Shen, H.; West, M. *Bayesian Modeling for Biological Pathway Annotation of Genomic Signatures*; Department of Statistical Science, Duke University: Durham, NC, USA, 2008.
30. Frost, H.R.; Li, Z.; Moore, J.H. Spectral gene set enrichment (SGSE). *BMC Bioinform.* **2015**, *16*, 70. [CrossRef]
31. Dinu, I.; Potter, J.D.; Mueller, T.; Liu, Q.; Adewale, A.J.; Jhangri, G.S.; Einecke, G.; Famulski, K.S.; Halloran, P.; Yasui, Y. Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinform.* **2007**, *8*, 242. [CrossRef]
32. Simillion, C.; Liechti, R.; Lischer, H.E.L.; Ioannidis, V.; Bruggmann, R. Avoiding the pitfalls of gene set enrichment analysis with SetRank. *BMC Bioinform.* **2017**, *18*, 151. [CrossRef]
33. Prifti, E.; Zucker, J.-D.; Clement, K.; Henegar, C. FunNet: An integrative tool for exploring transcriptional interactions. *Bioinformatics* **2008**, *24*, 2636–2638. [CrossRef] [PubMed]
34. Sun, C.-H.; Kim, M.-S.; Han, Y.; Yi, G.-S. COFECO: Composite function annotation enriched by protein complex data. *Nucleic Acids Res.* **2009**, *37*, W350-5. [CrossRef]
35. Vaske, C.J.; Benz, S.C.; Sanborn, J.Z.; Earl, D.; Szeto, C.; Zhu, J.; Haussler, D.; Stuart, J.M. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **2010**, *26*, i237–i245. [CrossRef] [PubMed]
36. Nilsson, B.; Håkansson, P.; Johansson, M.; Nelander, S.; Fioretos, T. Threshold-free high-power methods for the ontological analysis of genome-wide gene-expression studies. *Genome Biol.* **2007**, *8*, R74. [CrossRef] [PubMed]
37. Glansdorff, P.; Prigogine, I.V. *Thermodynamic: Theory of Structure, Stability*; Wiley: London, UK, 1971.
38. Zadran, S.; Arumugam, R.; Herschman, H.; Phelps, M.E.; Levine, R.D. Surprisal analysis characterizes the free energy time course of cancer cells undergoing epithelial-to-mesenchymal transition. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 13235–13240. [CrossRef]
39. Levine, R.D. Information Theory Approach to Molecular Reaction Dynamics. *Annu. Rev. Phys. Chem.* **1978**, *29*, 59–92. [CrossRef]
40. Agmon, N.; Alhassid, Y.; Levine, R.D. An algorithm for finding the distribution of maximal entropy. *J. Comput. Phys.* **1979**, *30*, 250–258. [CrossRef]
41. Kravchenko-Balasha, N.; Remacle, F.; Gross, A.; Rotter, V.; Levitzki, A.; Levine, R.D. Convergence of logic of cellular regulation in different premalignant cells by an information theoretic approach. *BMC Syst. Biol.* **2011**, *5*, 42. [CrossRef] [PubMed]
42. Gross, A.; Levine, R.D. Surprisal analysis of transcripts expression levels in the presence of noise: A reliable determination of the onset of a tumor phenotype. *PLoS ONE* **2013**, *8*, e61554. [CrossRef]

43. Garcia-Molina, A.; Kleine, T.; Schneider, K.; Mühlhaus, T.; Lehmann, M.; Leister, D. Translational Components Contribute to Acclimation Responses to High Light, Heat, and Cold in Arabidopsis. *iScience* **2020**, *23*, 101331. [CrossRef] [PubMed]
44. Remacle, F.; Kravchenko-Balasha, N.; Levitzki, A.; Levine, R.D. Information-theoretic analysis of phenotype changes in early stages of carcinogenesis. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 10324–10329. [CrossRef] [PubMed]
45. Gross, A.; Li, C.M.; Remacle, F.; Levine, R.D. Free energy rhythms in Saccharomyces cerevisiae: A dynamic perspective with implications for ribosomal biogenesis. *Biochemistry* **2013**, *52*, 1641–1648. [CrossRef]
46. Procaccia, I.; Levine, R.D. Potential work: A statistical-mechanical approach for systems in disequilibrium. *J. Chem. Phys.* **1976**, *65*, 3357–3364. [CrossRef]
47. CSBiology. TMEA Package. 8/16/2020. Available online: https://github.com/CSBiology/TMEA (accessed on 16 August 2020).
48. Anderson, E.; Bai, Z.; Bischof, C.; Blackford, S.; Demmel, J.; Dongarra, J.; Du Croz, J.; Greenbaum, A.; Hammarling, S.; McKenney, A.; et al. *LAPACK Users' Guide*, 3rd ed.; SIAM: Philadelphia, PA, USA, 1999.
49. NCBO BioPortal. GoMapMan—Summary. 2016. Available online: https://bioportal.bioontology.org/ontologies/GMM (accessed on 14 August 2020).
50. MapMan. MapManStore—Ath_AFFY_ATH1_TAIR10_Aug2012. 14/08/2020. Available online: https://mapman.gabipd.org/mapmanstore (accessed on 14 August 2020).
51. KEGG. KEGG COMPOUND Database. 14/08/2020. Available online: https://www.genome.jp/kegg/compound (accessed on 14 August 2020).
52. KEGG. KEGG BRITE: KEGG Orthology (KO)—*Arabidopsis thaliana* (thale cress). 14/08/2020. Available online: https://www.genome.jp/kegg-bin/get_htext?ath00001 (accessed on 14 August 2020).
53. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [CrossRef] [PubMed]
54. Young, A.; Whitehouse, N.; Cho, J.; Shaw, C. OntologyTraverser: An R package for GO analysis. *Bioinformatics* **2005**, *21*, 275–276. [CrossRef]
55. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **1995**, *57*, 289–300. [CrossRef]
56. CSBiology. FSharp.Stats. 13/08/2020. Available online: https://github.com/CSBiology/FSharp.Stats (accessed on 13 August 2020).
57. CSBiology. BioFSharp. 13/08/2020. Available online: https://github.com/CSBiology/BioFSharp (accessed on 13 August 2020).
58. Mühlhaus, T. FSharp.Plotly. 13/08/2020. Available online: https://github.com/muehlhaus/FSharp.Plotly (accessed on 13 August 2020).
59. Knijnenburg, T.A.; Wessels, L.F.A.; Reinders, M.J.T.; Shmulevich, I. Fewer permutations, more accurate P-values. *Bioinformatics* **2009**, *25*, i161–i168. [CrossRef]
60. Emmott, E.; Jovanovic, M.; Slavov, N. Ribosome Stoichiometry: From Form to Function. *Trends Biochem. Sci.* **2019**, *44*, 95–109. [CrossRef] [PubMed]
61. Agresti, A.; Min, Y. On small-sample confidence intervals for parameters in discrete distributions. *Biometrics* **2001**, *57*, 963–971. [CrossRef]
62. Harvaux, M.; Kloppstech, K. The protective functions of carotenoid and flavonoid pigments against excess visible radiation at chilling temperature investigated in Arabidopsis npq and tt mutants. *Planta* **2001**, *213*, 953–966. [CrossRef] [PubMed]
63. Trojak, M.; Skowron, E. Role of anthocyanins in highlight stress response. *World Sci. News* **2017**, *81*, 150–168.
64. Gould, K.S.; Dudle, D.A.; Neufeld, H.S. Why some stems are red: Cauline anthocyanins shield photosystem II against high light stress. *J. Exp. Bot.* **2010**, *61*, 2707–2717. [CrossRef] [PubMed]
65. Zeng, X.-Q.; Chow, W.S.; Su, L.-J.; Peng, X.-X.; Peng, C.-L. Protective effect of supplemental anthocyanins on Arabidopsis leaves under high light. *Physiol. Plant* **2010**, *138*, 215–225. [CrossRef]
66. Page, M.; Sultana, N.; Paszkiewicz, K.; Florance, H.; Smirnoff, N. The influence of ascorbate on anthocyanin accumulation during high light acclimation in *Arabidopsis thaliana*: Further evidence for redox control of anthocyanin synthesis. *Plant Cell Environ.* **2012**, *35*, 388–404. [CrossRef] [PubMed]
67. Williams, C.A.; Grayer, R.J. Anthocyanins and other flavonoids. *Nat. Prod. Rep.* **2004**, *21*, 539–573. [CrossRef] [PubMed]

68. Zhou, M.; Zhang, K.; Sun, Z.; Yan, M.; Chen, C.; Zhang, X.; Tang, Y.; Wu, Y. LNK1 and LNK2 Corepressors Interact with the MYB3 Transcription Factor in Phenylpropanoid Biosynthesis. *Plant Physiol.* **2017**, *174*, 1348–1358. [CrossRef]
69. Fraser, C.M.; Chapple, C. The phenylpropanoid pathway in Arabidopsis. *Arab. Book* **2011**, *9*, e0152. [CrossRef]
70. Lamers, J.; van der Meer, T.; Testerink, C. How Plants Sense and Respond to Stressful Environments. *Plant Physiol.* **2020**, *182*, 1624–1635. [CrossRef]
71. Bari, R.; Jones, J.D.G. Role of plant hormones in plant defence responses. *Plant Mol. Biol.* **2009**, *69*, 473–488. [CrossRef]
72. Rossini, S.; Casazza, A.P.; Engelmann, E.C.M.; Havaux, M.; Jennings, R.C.; Soave, C. Suppression of both ELIP1 and ELIP2 in Arabidopsis does not affect tolerance to photoinhibition and photooxidative stress. *Plant Physiol.* **2006**, *141*, 1264–1273. [CrossRef] [PubMed]
73. Kleine, T.; Kindgren, P.; Benedict, C.; Hendrickson, L.; Strand, A. Genome-wide gene expression analysis reveals a critical role for CRYPTOCHROME1 in the response of Arabidopsis to high irradiance. *Plant Physiol.* **2007**, *144*, 1391–1406. [CrossRef]
74. Brown, B.A.; Cloix, C.; Jiang, G.H.; Kaiserli, E.; Herzyk, P.; Kliebenstein, D.J.; Jenkins, G.I. A UV-B-specific signaling component orchestrates plant UV protection. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 18225–18230. [CrossRef] [PubMed]
75. Hayami, N.; Sakai, Y.; Kimura, M.; Saito, T.; Tokizawa, M.; Iuchi, S.; Kurihara, Y.; Matsui, M.; Nomoto, M.; Tada, Y.; et al. The Responses of Arabidopsis Early Light-Induced Protein2 to Ultraviolet B, High Light, and Cold Stress Are Regulated by a Transcriptional Regulatory Unit Composed of Two Elements. *Plant Physiol.* **2015**, *169*, 840–855. [CrossRef] [PubMed]
76. Hutin, C.; Nussaume, L.; Moise, N.; Moya, I.; Kloppstech, K.; Havaux, M. Early light-induced proteins protect Arabidopsis from photooxidative stress. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 4921–4926. [CrossRef]
77. Tzvetkova-Chevolleau, T.; Franck, F.; Alawady, A.E.; Dall'Osto, L.; Carrière, F.; Bassi, R.; Grimm, B.; Nussaume, L.; Havaux, M. The light stress-induced protein ELIP2 is a regulator of chlorophyll synthesis in *Arabidopsis thaliana*. *Plant J.* **2007**, *50*, 795–809. [CrossRef] [PubMed]
78. Tuteja, N.; Mahajan, S. Calcium signaling network in plants: An overview. *Plant Signal Behav.* **2007**, *2*, 79–85. [CrossRef]
79. Sanders, D.; Brownlee, C.; Harper, J.F. Communicating with calcium. *Plant Cell* **1999**, *11*, 691–706. [CrossRef]
80. Bauer, S.; Gagneur, J.; Robinson, P.N. GOing Bayesian: Model-based gene set analysis of genome-scale data. *Nucleic Acids Res.* **2010**, *38*, 3523–3532. [CrossRef]
81. Lu, Y.; Rosenfeld, R.; Simon, I.; Nau, G.J.; Bar-Joseph, Z. A probabilistic generative model for GO enrichment analysis. *Nucleic Acids Res.* **2008**, *36*, e109. [CrossRef]
82. Raghavan, N.; Amaratunga, D.; Cabrera, J.; Nie, A.; Qin, J.; McMillian, M. On methods for gene function scoring as a means of facilitating the interpretation of microarray results. *J. Comput. Biol.* **2006**, *13*, 798–809. [CrossRef] [PubMed]
83. Hashiguchi, A.; Komatsu, S. Impact of Post-Translational Modifications of Crop Proteins under Abiotic Stress. *Proteomes* **2016**, *4*, 42. [CrossRef] [PubMed]
84. Zhang, Q.; Bhattacharya, S.; Pi, J.; Clewell, R.A.; Carmichael, P.L.; Andersen, M.E. Adaptive Posttranslational Control in Cellular Stress Response Pathways and Its Relationship to Toxicity Testing and Safety Assessment. *Toxicol. Sci.* **2015**, *147*, 302–316. [CrossRef] [PubMed]
85. Bogaert, K.A.; Perez, E.; Rumin, J.; Giltay, A.; Carone, M.; Coosemans, N.; Radoux, M.; Eppe, G.; Levine, R.D.; Remacle, F.; et al. Metabolic, Physiological, and Transcriptomics Analysis of Batch Cultures of the Green Microalga Chlamydomonas Grown on Different Acetate Concentrations. *Cells* **2019**, *8*, 1367. [CrossRef]
86. Bogaert, K.A.; Manoharan-Basil, S.S.; Perez, E.; Levine, R.D.; Remacle, F.; Remacle, C. Surprisal analysis of genome-wide transcript profiling identifies differentially expressed genes and pathways associated with four growth conditions in the microalga Chlamydomonas. *PLoS ONE* **2018**, *13*, e0195142. [CrossRef]
87. Willamme, R.; Alsafra, Z.; Arumugam, R.; Eppe, G.; Remacle, F.; Levine, R.D.; Remacle, C. Metabolomic analysis of the green microalga Chlamydomonas reinhardtii cultivated under day/night conditions. *J. Biotechnol.* **2015**, *215*, 20–26. [CrossRef]

88. Ganguly, D.R.; Stone, B.A.B.; Bowerman, A.F.; Eichten, S.R.; Pogson, B.J. Excess Light Priming in *Arabidopsis thaliana* Genotypes with Altered DNA Methylomes. *G3* **2019**, *9*, 3611–3621. [CrossRef]

89. Remacle, F.; Goldstein, A.; Levine, R. Multivariate Surprisal Analysis of Gene Expression Levels. *Entropy* **2016**, *18*, 445. [CrossRef]

Article III: Systems-wide analysis revealed shared and unique responses to moderate and acute high temperatures in the green alga Chlamydomonas reinhardtii

Ningning Zhang, Erin M. Mattoon, Will McHargue, Benedikt Venn, David Zimmer, Kresti Pecani, Jooyeon Jeong, Cheyenne M. Anderson, Chen Chen, Jeffrey C. Berry, Ming Xia, Shin-Cheng Tzeng, Eric Becker, Leila Pazouki, Bradley Evans, Fred Cross, Jianlin Cheng, Kirk J. Czymmek, Michael Schroda, Timo Mühlhaus & Ru Zhang

| Communications Biology, 2022, doi.org/10.1038/s42003-022-03359-z

communications
biology

ARTICLE

https://doi.org/10.1038/s42003-022-03359-z OPEN

Systems-wide analysis revealed shared and unique responses to moderate and acute high temperatures in the green alga *Chlamydomonas reinhardtii*

Ningning Zhang[1,7], Erin M. Mattoon[1,2,7], Will McHargue[1,6], Benedikt Venn[3], David Zimmer[3], Kresti Pecani[4], Jooyeon Jeong[1], Cheyenne M. Anderson[1,6], Chen Chen[5], Jeffrey C. Berry[1], Ming Xia[1], Shin-Cheng Tzeng[1], Eric Becker[1], Leila Pazouki[1], Bradley Evans[1], Fred Cross[4], Jianlin Cheng[5], Kirk J. Czymmek[1], Michael Schroda[3], Timo Mühlhaus[3] & Ru Zhang[1✉]

Different intensities of high temperatures affect the growth of photosynthetic cells in nature. To elucidate the underlying mechanisms, we cultivated the unicellular green alga *Chlamydomonas reinhardtii* under highly controlled photobioreactor conditions and revealed systems-wide shared and unique responses to 24-hour moderate (35°C) and acute (40°C) high temperatures and subsequent recovery at 25°C. We identified previously overlooked unique elements in response to moderate high temperature. Heat at 35°C transiently arrested the cell cycle followed by partial synchronization, up-regulated transcripts/proteins involved in gluconeogenesis/glyoxylate-cycle for carbon uptake and promoted growth. But 40°C disrupted cell division and growth. Both high temperatures induced photoprotection, while 40°C distorted thylakoid/pyrenoid ultrastructure, affected the carbon concentrating mechanism, and decreased photosynthetic efficiency. We demonstrated increased transcript/protein correlation during both heat treatments and hypothesize reduced post-transcriptional regulation during heat may help efficiently coordinate thermotolerance mechanisms. During recovery after both heat treatments, especially 40°C, transcripts/proteins related to DNA synthesis increased while those involved in photosynthetic light reactions decreased. We propose down-regulating photosynthetic light reactions during DNA replication benefits cell cycle resumption by reducing ROS production. Our results provide potential targets to increase thermotolerance in algae and crops.

[1] Donald Danforth Plant Science Center, St. Louis, Missouri 63132, USA. [2] Plant and Microbial Biosciences Program, Division of Biology and Biomedical Sciences, Washington University in Saint Louis, St. Louis, Missouri 63130, USA. [3] TU Kaiserslautern, Kaiserslautern 67663, Germany. [4] The Rockefeller University, New York, New York 10065, USA. [5] University of Missouri-Columbia, Columbia, Missouri 65211, USA. [6] Present address: Plant and Microbial Biosciences Program, Division of Biology and Biomedical Sciences, Washington University in Saint Louis, St. Louis, Missouri 63130, USA. [7] These authors contributed equally: Ningning Zhang, Erin M. Mattoon. ✉email: rzhang@danforthcenter.org

High temperatures occur frequently in nature and impair crop yields and algal biofuel production[1,2]. Global warming increases the intensity, duration, and frequency of high temperatures above the optimal range for plant growth. It is projected that for every degree Celsius mean global temperature increases, the yields of major crop species will decrease by 3% ~ 8%[1,3]. Photosynthetic organisms experience different intensities of high temperatures in field conditions. Many crop species have threshold temperatures between 25 and 40°C, above which reduced growth is observed; most heat stress experiments have been conducted at acutely high temperatures near 42°C or above[4]. Plants frequently experience sustained moderate high temperatures around 35°C in nature, however these conditions have been largely understudied[5]. The acute high temperature at or above 40°C is more damaging but usually less frequent or shorter-lasting than the moderate high temperature in the field. We hypothesize that plants can acclimate to moderate high temperature but have reduced acclimation capacity to acute high temperature. Additionally, we propose that different levels of high temperatures induce shared and unique responses in photosynthetic cells. Understanding how photosynthetic cells respond to and recover from different intensities of high temperatures is imperative for improving crop thermotolerance[6].

High temperatures are known to have a wide variety of impacts on photosynthetic cells. Heat-increased membrane fluidity has been proposed to activate membrane-localized mechanosensitive ion channels leading to increased intracellular calcium concentrations, which may cause signaling cascades to activate heat shock transcription factors (HSFs)[7–9]. HSFs act in the nucleus to increase transcription of genes involved in heat response, e.g., heat shock proteins (HSPs)[10–12]. A recent work proposed that the accumulating cytosolic unfolded proteins, rather than changes in membrane fluidity, trigger the expression of HSPs in green algae[13]. Furthermore, high temperatures can decrease the stability of RNAs and alter the transcriptomic landscape of cells under heat stress[14]. Additionally, high temperature can cause damage to photosynthetic electron transport chains, reducing photosynthetic efficiency[15–19], and leading to increased reactive oxygen species (ROS) accumulation[4,20,21]. Heat-induced ROS production increases DNA damage and the need for DNA repair pathways, although the mechanisms of these processes are poorly understood[10,22]. In contrast to the extensive research on the effects during heat, how photosynthetic cells recover from heat is less studied[10].

Algae have great potential for biofuel production and bioproduct accumulation, but the knowledge surrounding mechanisms of algal heat responses are largely limited as compared to land plants[10]. Outdoor algal ponds frequently experience supraoptimal temperatures at or above 35°C during summer time[23], but how algal cells respond to moderate high temperatures remains largely understudied. Many previous algal heat experiments were conducted in flasks incubated in hot water baths (at or above 42°C) with sharp temperature switches, e.g., by resuspending centrifuged cells in prewarmed medium to initiate the heat treatments[13,18]. The previous research was valuable for paving the road to understand algal heat responses. Nevertheless, high temperatures in nature, especially in aquatic environments, often increase gradually and the rate of temperature increase may affect heat responses[6]. Acute high temperature at 39°C or 42°C results in algal cell cycle arrest[18,24–26]. Long-term experiments at moderate high temperatures that do not lead to a sustained cell cycle arrest cannot be conducted in flasks because cultures grow into stationary phase, causing nutrient and light limitation and therefore complicating analyses. Consequently, investigating algal heat responses under well-controlled conditions in photobioreactors (PBRs) with turbidostatic modes can mimic the heating speed in nature and reduce compounding factors during high-temperature treatments, improving our understanding of algal heat responses.

The unicellular green alga, *Chlamydomonas reinhardtii* (Chlamydomonas throughout), is an excellent model to study heat responses in photosynthetic cells for many reasons, including its fully sequenced haploid genome, unicellular nature allowing for homogenous treatments, generally smaller gene families than land plants, and extensive genetic resources[27–32]. At the cellular level, Chlamydomonas has many similarities with land plants, making it a powerful model organism to identify novel elements with putative roles in heat tolerance with implications for crops[10].

A previous transcriptome and lipidome-level analysis in Chlamydomonas under acute high temperature (42°C) over 1 hour (h) revealed changes in lipid metabolism and increased lipid saturation as one of the early heat responses[33]. Additionally, a proteome and metabolome-level analysis of acute high temperature (42°C) for 24-h followed by 8-h recovery demonstrated temporally resolved changes in proteins, metabolites, lipids, and cytological parameters in Chlamydomonas[18]. Both publications contributed to the foundational knowledge of how Chlamydomonas responds to acute high temperature. However, a temporally resolved transcriptome analysis during and after heat over a relatively long time is lacking and the correlation between transcriptome, proteome, and physiological responses to different intensities of high temperatures remains elusive. Integrating these multiomics approaches with physiological measurements under high-temperature conditions has great potential for improving algal thermotolerance[34]. Additionally, previous research showed increased starch accumulation when Chlamydomonas cells were switched from 30°C to 39°C heat, which is linked to cell cycle arrest and the shift from energy usage for cell cycle operation to chemical energy storage at 39°C[25,26]. However, the effects of moderate high temperature on starch accumulation remain elusive.

We investigated the response of wild-type Chlamydomonas cells to moderate (35°C) or acute (40°C) high temperatures at transcriptomic, proteomic, cytological, photosynthetic, and ultrastructural levels over a 24-h heat followed by 48-h recovery period in PBRs under well-controlled conditions. Our results showed that some of the responses were shared between the two treatments and the effects of 40°C were typically more extensive than 35°C; however, 35°C induced a unique set of responses that were absent under 40°C. Both 35 and 40°C induced starch accumulation but due to distinct mechanisms. We showed that 35°C transiently inhibited cell cycle followed by synchronization while 40°C halted the cell cycle completely. Heat at 40°C but not 35°C reduced photosynthetic efficiency, increased ROS production, and altered chloroplast structures. Furthermore, with the time-resolved paired transcriptome and proteome dataset, we demonstrated increased transcript and protein correlation during high temperature which was reduced during the recovery period. Additionally, we revealed up-regulation of genes/proteins related to DNA replication and down-regulation of those related to photosynthetic light reactions during early recovery after both heat treatments, suggesting potential crosstalk between these two pathways when resuming the cell cycle. These data further our understanding of algal heat responses and provide novel insights to improve thermotolerance of algae and crops.

Results

Heat at 35°C increased algal growth while 40°C largely reduced it. We cultivated Chlamydomonas cultures under well-controlled conditions (light, temperature, air flow, and

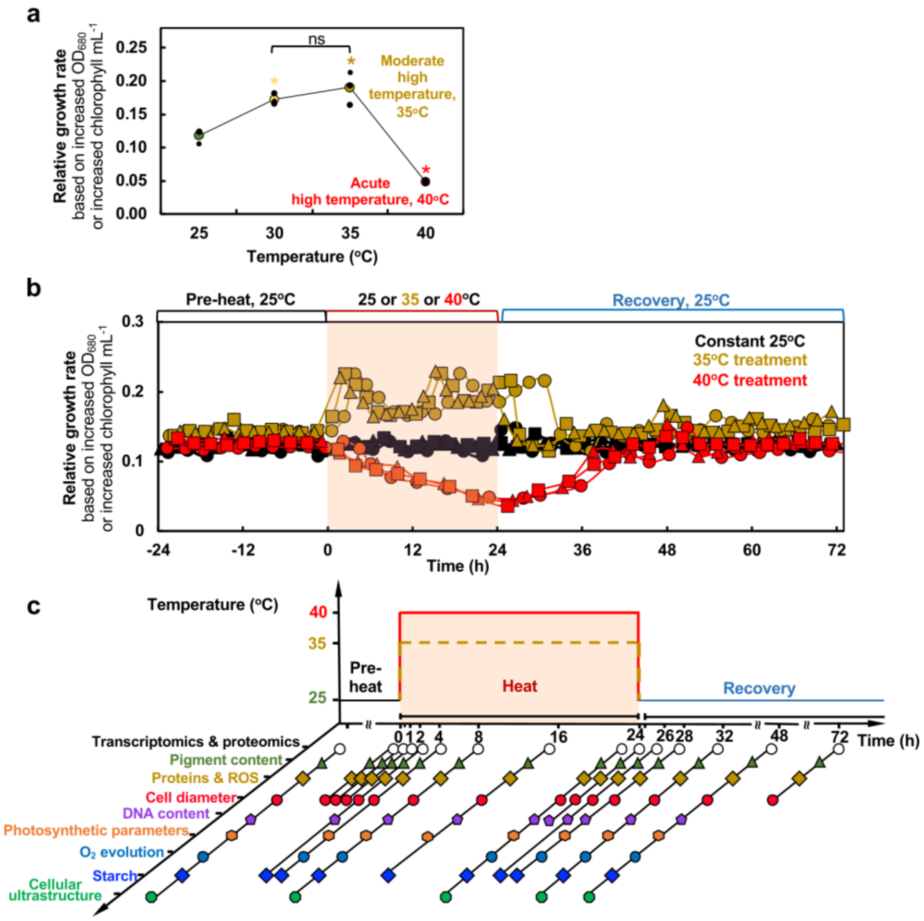


Fig. 1 Moderate (35°C) and acute (40°C) high temperatures had contrasting effects on the growth rates of Chlamydomonas cells. a Chlamydomonas growth rate plateaued around 35°C but was largely reduced at 40°C. Chlamydomonas cells (CC-1690, 21gr, wildtype) were grown in photobioreactors (PBRs) in Tris-acetate-phosphate (TAP) medium under turbidostatic conditions at different temperatures with a light intensity of 100 µmol photons m^{-2} s^{-1} and constantly bubbling of air. Relative growth rates were calculated based on the cycling of OD_{680}, see Supplementary Fig. 1 and methods for details. OD_{680} is proportional to total chlorophyll content in unit of µg chlorophyll mL^{-1}. Each temperature treatment was conducted in an individual PBR. Mean ± SE, $n = 3$ biological replicates. Statistical analyses were performed using two-tailed t-test assuming unequal variance by comparing treated samples with pre-heat or 30°C with 35°C (*, $p < 0.05$, the colors of asterisks match the treatment conditions). Not significant, ns. **b** Heat treatment at 35°C (brown) increased growth rates while 40°C (red) reduced it. Algal cultures in separate PBRs were first acclimated at 25°C for 4 days before the temperature was switched to 35°C or 40°C for 24 h, followed by recovery at 25°C for 48 h. Algal cultures grown constantly at 25°C (black) served as controls, which demonstrated steady growth without heat treatments. Three independent biological replicates for each condition were plotted. **c** PBR cultures with different treatments were sampled at a series of time points to study heat responses at multiple levels. Symbol colors match the colors of the parameters assayed. **b**, **c** The red shaded areas depict the duration of high temperature.

nutrient availability with supplied carbon source, acetate) by using turbidostatic mode based on OD_{680} in photobioreactors (PBRs). OD_{680} is proportional to total chlorophyll content in the unit of µg chlorophyll mL^{-1}. Fresh medium was added to the culture automatically by a peristaltic pump when the OD_{680} reached the defined maximum value to dilute the culture and the pump was stopped when the OD_{680} dropped to the defined minimum value (Supplementary Fig. 1a, b). Algal cultures then grew at approximately exponential rate to the defined maximum OD_{680} value before the next dilution cycle. The turbidostatic mode precisely controlled the growth condition. We calculated the doubling time or relative growth rates (inverse of the doubling time) based on the exponential growth phase. The doubling time and relative growth rates we refer to throughout are based on the increase of OD_{680} or the total chlorophyll per mL culture.

The relative growth rate of Chlamydomonas cells in PBRs grown mixotrophically increased with higher temperature between 25°C and 30°C, plateaued at 35°C, and largely decreased at 40°C, as compared to the control 25°C (Fig. 1a). Hence, we defined 35°C as moderate high temperature and 40°C as the acute high temperature under our experimental conditions. To investigate the systems-wide responses of Chlamydomonas to moderate and acute high temperatures, we acclimated algal cells in PBRs with well-controlled conditions at constant 25°C, followed by 24-h heat treatment at 35 or 40°C, and recovery at 25°C for 48 h (Fig. 1b, Supplementary Fig. 1a–c). Temperature increases from 25 to 35 or 40°C took about 30 minutes (min), and neither heat treatment affected cell viability (Supplementary Fig. 1d, e). In contrast, we found that a sharp temperature switch without gradual temperature increases reduced cell viability (Supplementary Fig. 1f), demonstrating heating speed affects thermotolerance. The transcript levels of selected circadian regulated genes, *LHCA1*[35] and *TRXF2*[36], did not change significantly under constant 25°C (Supplementary Fig. 1g, h), suggesting minimal circadian regulation existed under our experimental conditions with turbidostatic control and constant light. Our observed changes during and after heat were therefore most likely attributable to heat treatments.

Based on the rate of chlorophyll production, algal growth increased during cultivation at 35°C and decreased at 40°C compared to 25°C. The increased growth at 35°C was confirmed by medium consumption rates and growth on plates (Supplementary Fig. 2). We harvested the PBR cultures throughout the time-course experiment for systems-wide analyses, including transcriptomics, proteomics, cell physiology, photosynthetic parameters, and cellular ultrastructure (Fig. 1c). RT-qPCR analysis of select time points showed that both 35°C and 40°C induced heat stress marker genes (*HSP22A* and *HSP90A*) and *HSF*s, suggesting both heat treatments could induce heat responses, although the induction amplitude was much larger under 40°C than 35°C (up to 20-fold, Supplementary Fig. 3a–d). RNA-seq data were verified by testing select genes with RT-qPCR, with highly consistent results between the two methods (Supplementary Fig. 3).

Transcriptomic and proteomic analyses identified shared and unique responses during and after 35°C and 40°C treatments. Two-dimensional Uniform Manifold Approximation and Projection (UMAP) of Transcripts per Million (TPM) normalized RNA-seq data resulted in three distinct clusters (Fig. 2a). The cluster during heat had a temporally resolved pattern showing increasing variance between 35°C and 40°C time points throughout the high temperature treatment. The early recovery cluster consisted of 2- and 4-h recovery samples after 35°C heat, as well as 2-, 4-, and 8-h recovery samples after 40°C heat. Late recovery and pre-heat samples clustered together, suggesting transcriptomes fully recovered by 8-h following 35°C and 24-h following 40°C treatment. UMAP of proteomics data results in two distinct clusters, separating the 35°C and 40°C treated samples and demonstrating temporally resolved proteomes (Fig. 2b). However, the samples during heat, early, and late recovery did not fall into their own distinct clusters, consistent with resistance to rapid changes on the protein level as compared with the transcript level. The proteome recovered to pre-heat levels by the 48-h recovery after both 35°C and 40°C.

We employed differential expression modeling to identify differentially expressed genes (DEGs) that were overlappingly or uniquely up- or down-regulated during and after 35°C or 40°C heat treatments (Fig. 2c, Supplementary Data 1). The greatest number of DEGs were identified at 2- and 4-h recovery time points of 40°C treatment, while investigating the distribution of $\log_2$(fold-change) values for DEGs showed the greatest level of up-regulation at 0.5-h and down-regulation at 16-h of heat at 40°C (Supplementary Fig. 4a). Overall, there were more DEGs at most time points in the 40°C treatment as compared to the 35°C treatment (Fig. 2c). Analysis of differentially accumulated proteins (DAPs) that were overlappingly or uniquely up- or down-regulated during or after 35°C and 40°C showed a smoother distribution of expression pattern than transcriptomic data (Fig. 2d, Supplementary Data 2). Increasing numbers of DAPs were identified throughout the high-temperature period, followed by a gradual reduction throughout the recovery period. The distribution of $\log_2$(fold-change) for DAPs at each time point also showed a smoother pattern than for transcriptome data (Supplementary Fig. 4b). Through transcriptomic and proteomic analyses, we identified shared and unique responses for the 35°C and 40°C treatment groups (Supplementary Fig. 4c–f, Supplementary Data 1 and 2).

The global transcriptome analysis revealed the three most dominant transcriptional patterns during the two treatments (Supplementary Fig. 5a, b). The first constraint (λ1) divided the transcripts into acclimation and de-acclimation phases; the second constraint (λ2) separated control from disturbed conditions; the third constraint (λ3) showed a more fluctuating fine regulation. The amplitude difference between the treatments at 35°C and 40°C suggests an overall higher regulatory activity during the 40°C than 35°C treatment. However, there were a set of 108 genes uniquely upregulated during 35°C but not 40°C heat (Supplementary Fig. 4c, Supplementary Data 1), including *GAPDH* (involved in gluconeogenesis, glycolysis, and Calvin-Benson Cycle), *TAL1* (involved in pentose phosphate pathway), *COX15* (involved in mitochondrial assembly), and *CAV4* (encoding a putative calcium channel) (Supplementary Fig. 5c–f). Additionally, when investigating the $\log_2$(fold-change) ratios of overlapping up- or down-regulated genes in both treatment groups at the same time point, we found that although many genes had higher differential expression with 40°C than 35°C treatment, some genes were more highly differentially expressed in the 35°C than the 40°C treatment group (Supplementary Fig. 6, Supplementary Data 1). Taken together, these results indicate that 35°C induces a unique set of responses in Chlamydomonas that have not been previously described.

Of the 3,960 heat-induced genes (HIGs, up-regulated in at least one-time point of 35°C or 40°C), 2,754 were present in the JGI InParanoid ortholog list[37,38]. We used these data to investigate the conservation of Chlamydomonas HIGs with *Volvox carteri* (Volvox), *Arabidopsis thaliana* (Arabidopsis), *Oryza sativa* (rice), *Triticum aestivum* (wheat), *Glycine max* (soybean), *Zea mays* (maize), *Sorghum bicolor* (sorghum), and *Setaria viridis* (Setaria). For most plant species tested, approximately 1,000 Chlamydomonas HIGs have orthologs (Supplementary Fig. 5g). Between Chlamydomonas and the model plant Arabidopsis, there are 509 HIGs with a one-to-one orthologous relationship (Supplementary Fig. 5g, Supplementary Data 1).

In our transcriptome and proteome data, 44.7% of the transcripts that met minimum read count cutoffs have MapMan annotations, and 80.4% of proteins have MapMan annotations (Supplementary Fig. 4g, 4h). MapMan functional enrichment analysis of DEGs at each time point of the 35°C or 40°C treatment showed that early induced shared responses to both heat treatments included canonical heat response pathways, protein folding, and lipid metabolism (Supplementary Table 1, Supplementary Data 3). MapMan terms related to DNA synthesis, cell motility, protein processing, and RNA regulation were enriched in overlapping gene sets down-regulated during most time points of both 35°Cand 40°C heat treatments. DNA synthesis and repair MapMan terms were significantly enriched in genes up-regulated in both 35°C and 40°C treated samples during the 2- and 4-h recovery time points. MapMan terms related to amino acid metabolism, mitochondrial electron transport, and purine synthesis were enriched in gene sets uniquely up-regulated during 35°C heat. Carbon fixation (e.g., carbon concentrating mechanism) and starch synthesis related MapMan terms were significantly enriched in gene sets uniquely up-regulated during 40°C heat. MapMan terms related to amino acid metabolism and mitochondrial electron transport were enriched in gene sets uniquely down-regulated during early heat of 40°C, in contrast to 35°C.

Transcript/protein correlation increased during heat but decreased during recovery. Investigation of Pearson correlation coefficients between $\log_2$(fold-change) values for transcripts and proteins grouped by MapMan functional categories revealed higher positive correlation between transcriptome and proteome during heat than recovery for both 35°C and 40°C treatments (Fig. 2e, f, Supplementary Fig. 7). This indicates that

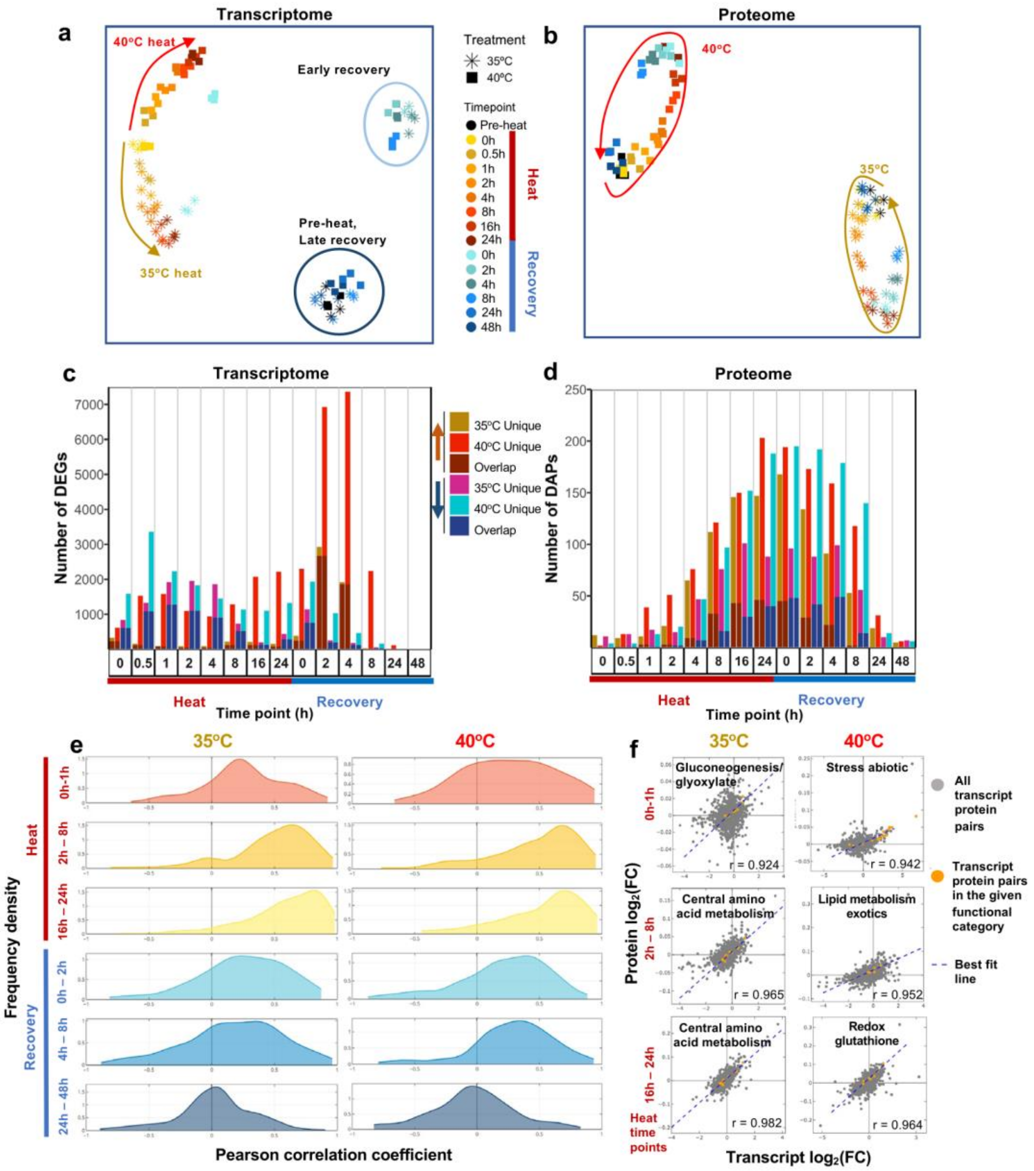


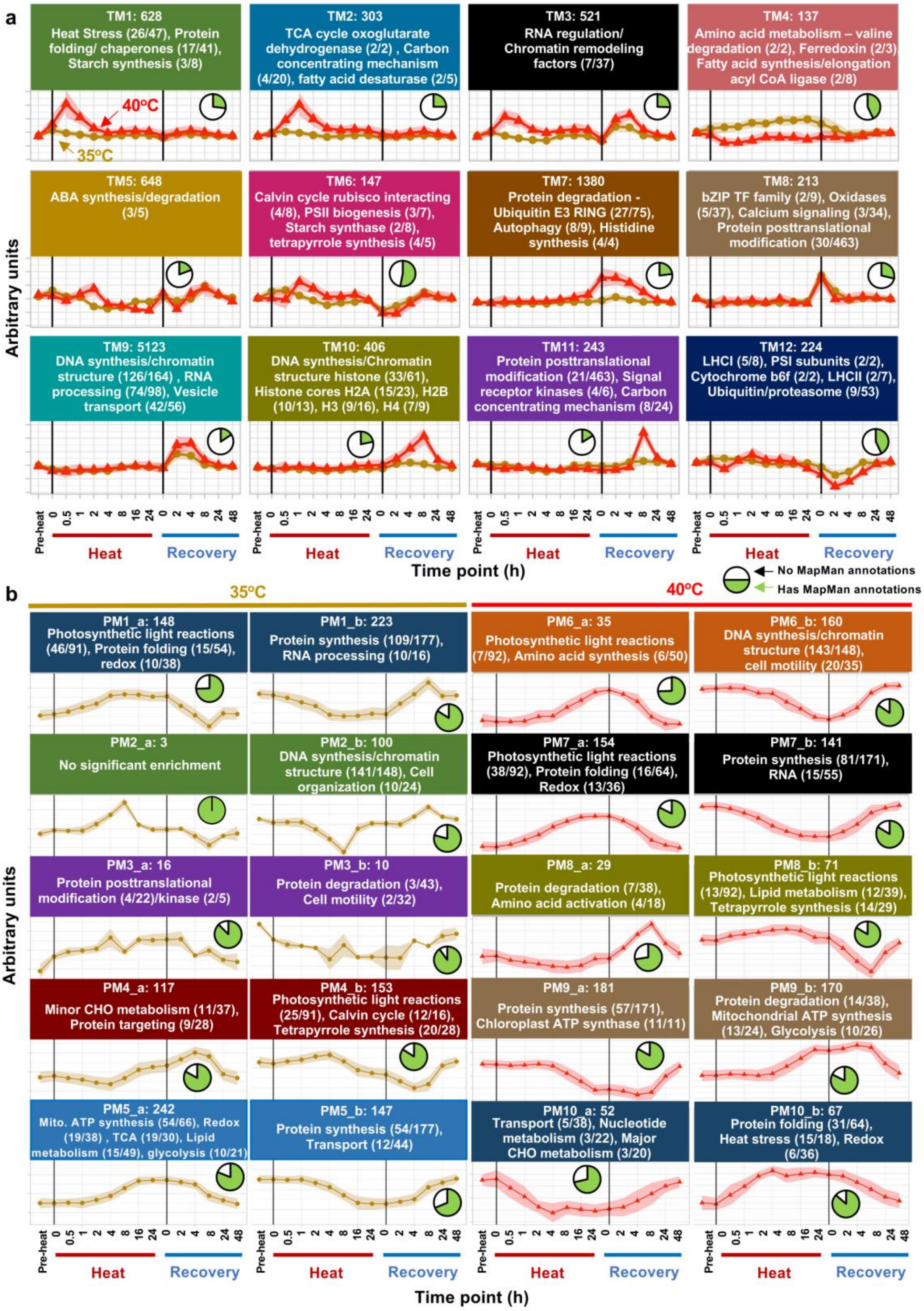
Fig. 2 Transcriptomics and proteomics revealed distinct and dynamic responses during and after heat treatments of 35°C and 40°C. a Uniform Manifold Approximation and Projection (UMAP) of Transcripts Per Million (TPM) normalized RNA-seq read counts and **(b)** UMAP of normalized protein intensities. Each data point represents all normalized counts from a single sample. Stars and squares represent algal samples with heat treatments of 35°C or 40°C, respectively. Different colors represent different time points. Brown and red arrows show the movement through time of the 35°C and 40°C treated samples, respectively. **c**, **d** Number of unique or overlapping Differentially Expressed Genes (DEGs) and Differentially Accumulated Proteins (DAPs) with heat treatment of 35°C or 40°C at different time points, respectively. Each time point has four bars: the first bar represents genes up-regulated in 35°C, the second represents genes up-regulated in 40°C, the third represents genes down-regulated in 35°C, and the fourth represents genes down-regulated in 40°C. The bottom portion of the stacked bars represents genes/proteins that are differentially expressed in both treatment groups and the top portion represents genes that are uniquely differentially expressed in the given treatment group at that time point. Significant differential expression in transcriptome data was defined as absolute values of $\log_2$(fold-change, FC) > 1, false discovery rate (FDR) < 0.05, and absolute difference of TPM normalized read counts between treatment and pre-heat control ≥ 1. Significant differential accumulation of proteins was defined by Dunnett's FWER < 0.05. **e** Analysis of correlation between transcripts and proteins revealed overall rising positive correlation during the heat treatment and a decreasing correlation during recovery. The time points were subdivided into three heat (red labels) and three recovery (blue labels) windows for 35°C (left) and 40°C (right) treated samples. The density plots of Pearson correlation coefficients between the fold-changes of transcripts and proteins are shown. X, Pearson correlation coefficient. Y, frequency density. Time points during heat: 0 h, reach high temperature of 35°C or 40°C; 1 h, heat at 35°C or 40°C for 1 h, similar names for other time points during heat. Time points during recovery: 0 h, reach control temperature of 25°C for recovery after heat; 2 h, recovery at 25°C for 2 h, similar names for other time points during recovery. See Supplementary Fig. 7 for more information. **f** Scatter plots of all transcript-protein pairs at 35°C (left) and 40°C (right) for the three heat time point bins shown in (**e**). X and Y, transcript and protein $\log_2$FC, compared with pre-heat, respectively. Transcript-protein pairs are shown as gray dots. Best fit lines are shown in blue. The Pearson correlation coefficient is shown at the bottom right corner of each scatterplot. Transcript-protein pairs belonging to MapMan functional categories with the highest Pearson correlation coefficient in the given time point bin are shown in orange. See the interactive figures in Supplementary Data 9: transcripts/proteins correlation.

transcriptional regulation dominates the heat period. Further investigation of Pearson correlation coefficients for individual MapMan terms showed that functional categories had varying correlation values throughout the course of high temperatures and recovery (Supplementary Data 4, 9). In early time points of the 35°C treatment, the gluconeogenesis/glyoxylate-cycle functional category had the highest correlation between transcripts and proteins, followed by amino acid and lipid metabolism at later heat time points (Fig. 2f). In early time points of the 40°C treatment, the abiotic stress functional category had the highest correlation between transcripts and proteins, followed by lipid metabolism and redox pathways at later heat time points (Fig. 2f). The proteins in the MapMan bin gluconeogenesis/glyoxylate cycle increased during 35°C but decreased during 40°C, suggesting elevated and reduced gluconeogenesis/glyoxylate-cycle activity during 35°C and 40°C heat treatment, respectively (Supplementary Fig. 8a–d). Isocitrate lyase (ICL1) is a key enzyme of the glyoxylate cycle in Chlamydomonas[39]. The transcript, protein, as well as the correlation of ICL1 increased during 35°C heat but decreased during 40°C heat (Supplementary Data 9,10, gluconeogenesis _ glyoxylate cycle.html). In Chlamydomonas, acetate uptake feeds into the glyoxylate cycle and gluconeogenesis for starch biosynthesis[40,41]. Several proteins related to acetate uptake/assimilation[41,42] were increased during 35°C heat but deceased during 40°C heat (Supplementary Fig. 8e–h, Supplementary Data 2). Our results suggested 35°C treatment increased acetate uptake/assimilation and glyoxylate cycle and gluconeogenesis pathways, which may be suppressed by the 40°C treatment.

Network modeling of transcriptome and proteome revealed expression patterns of key pathways during and after heat treatments. To investigate common transcriptional expression patterns throughout heat and recovery, we performed Weighted Correlation Network Analysis (WGCNA) on TPM-normalized RNA-seq data from 35°C and 40°C transcriptomes (Fig. 3a, Supplementary Data 5). Most genes with known roles in heat response belong to Transcriptomic Module 1 (TM1) with peak expression at 0.5-h heat, including *HSF*s and many *HSP*s. TM1 is also significantly enriched for MapMan terms related to protein folding, lipid metabolism, and starch synthesis. In total, there are 628 genes associated with this module, about 62% of which lack Chlamydomonas descriptions and 25% have no ontology annotations, indicative of novel genes with putative roles in heat response that have been previously undescribed. TM2 has peak expression at 1-h heat and is significantly enriched for MapMan terms relating to the TCA cycle and carbon concentrating mechanism. TM3 has peak expression at the beginning of both heat and recovery periods and is significantly enriched for MapMan terms related to RNA regulation and chromatin remodeling, highlighting the extensive alterations in transcriptional regulation caused by changing temperatures. TM4 contains 137 genes that have sustained increased expression throughout 35°C, with slightly reduced or no change in expression in 40°C. This module is significantly enriched for MapMan terms related to amino acid metabolism, ferredoxin, and fatty acid synthesis/elongation. TM4 represents genes that are unique to 35°C responses. TM5 shows increased expression throughout the heat period at 40°C. This module is enriched for ABA synthesis/degradation, suggesting potential interaction between the ABA pathway and algal heat responses. Previous reports showed that heat treated Arabidopsis and rice leaves had increased ABA contents and exogenous ABA application improved heat tolerance in WT rice plants[43,44]. ABA is reported to be involved in tolerance to oxidative stress[45], HCO3− uptake[46], and stress acclimation in Chlamydomonas[47]. TM6 displays similar expression patterns in both 35°C and 40°C treatments, with increased expression during early heat and reduced expression during early recovery. This module is enriched for genes in the Calvin-Benson cycle, PSII biogenesis, and tetrapyrrole synthesis. TM7 shows increased expression in 40°C uniquely in late heat, which continues through the early and mid-recovery periods. This module is notably enriched for protein degradation, ubiquitin E3 RING, and autophagy, which may indicate the need to degrade certain proteins and cellular components during and after prolonged acute high temperature at 40°C. TM8 shows increased expression only at the 0-h recovery time point immediately after the cooling from the 35°C and 40°C heat to 25°C. This module is notably enriched for bZIP transcription factors, oxidases, calcium signaling, and protein posttranslational modification, which likely contribute to the broad changes observed when recovering from high temperatures. TM9 and TM10 both have peaks in expression in recovery (2 h, 4 h for TM9; and 8 h for TM10) and have significantly enriched MapMan terms related to DNA synthesis. TM11 peaks in expression at 8-h recovery and is significantly enriched for MapMan terms related to protein posttranslational modification, consistent with the decreased correlation between transcript and protein levels at late recovery stages (Fig. 2e). TM12 shows reduced expression during the early recovery period and is significantly enriched for MapMan terms related to photosynthetic light reactions and protein degradation.

We verified the expression patterns of several key pathways of interest and transcription factors by visualizing $\log_2$(fold-change) values from differential expression modeling (Supplementary Fig. 9, 10, Supplementary Data 6). The RNA-seq results of select pathways from differential expression modeling were highly consistent with WGCNA modeling and provided gene-level resolution of interesting trends within these pathways. The down-regulation of select transcripts involved in photosynthetic light reactions during early recovery was also verified by RT-qPCR (Supplementary Fig. 9i, j).

Network modeling was performed separately for the proteomes of the 35°C and 40°C treated samples due to differences in peptides identified through LC-MS/MS between the two treatment groups and the relatively smaller number of proteins identified compared to transcripts. This analysis identified common expression patterns (Proteomics Module, PM) in the proteome data (Fig. 3b, Supplementary Data 7). Prominent functional terms that were enriched in the respective module are given for proteins correlating positively (PM_a) or negatively (PM_b) to their corresponding eigenvector (FDR < 0.05). Protein modules that increased during 35 and 40°C heat treatments are enriched for MapMan terms related to the part of photosynthetic light reactions, protein folding, and redox (PM1_a, PM6_a, PM7_a). Proteins that increased during the recovery phase after either heat treatments are enriched for MapMan terms related to protein synthesis (PM1_b, PM7_b), DNA synthesis and chromatin structure (PM2_b, PM6_b). Proteins related to photosynthetic light reactions first decreased and then increased during the recovery of both treatments (PM1_a, PM4_b, PM8_b). Unique responses for the 35°C treatment include proteins related to mitochondrial electron transport and lipid metabolism increased during heat (PM5_a) and proteins related to RNA processing and cell organization increased during the recovery phase after 35°C heat (PM1_b, PM2_b). Unique responses for the 40°C treatment include proteins related to abiotic stress increased during heat (PM10_b) and proteins related to cell motility and RNA increased during the recovery phase after 40°C heat (PM6_b, PM7_b). Network modeling of transcriptome and proteome data yielded consistent patterns for several key pathways during and



after heat treatments, e.g., heat responses, photosynthetic light reactions, and DNA synthesis.

Heat at 35°C synchronized the cell cycle while 40°C arrested it. The increased transcript and protein levels related to DNA synthesis during recovery (Fig. 3) prompted us to investigate the expression pattern of cell cycle related genes because DNA synthesis takes place immediately before and during cell division in Chlamydomonas[48]. Both 35°C and 40°C disrupted the expression pattern of cell cycle genes as compared to the pre-heat level, which recovered by 8-h heat treatment at 35°C but not

Fig. 3 Transcriptomic and proteomic network modeling revealed differential regulation of key biological pathways with treatments of 35°C or 40°C. **a** Weighted correlation network analysis (WGCNA) of transcriptome data identified gene modules with similar expression patterns. TM: transcriptomic module. The z-normalized consensus gene expression patterns for 40°C (red triangle) and 35°C (brown circle) for each module are displayed. **b** Correlation networks of protein abundance over time courses for 35°C and 40°C heat treatments. PM, proteomic module. For each module, the eigenvectors as aggregated signal shape is depicted. Prominent functional terms that were enriched in the respective module are given for proteins correlating positively (PM_a) or negatively (PM_b) to their corresponding eigenvector (FDR < 0.05). **a**, **b** Black vertical lines indicate the start and end of the heat treatment at either 35°C or 40°C. Pre-heat, before heat treatments. Time points during heat: 0 h, reach high temperature of 35°C or 40°C; 0.5 h, heat at 35°C or 40°C for 0.5 h, similar names for other time points during heat. Time points during recovery: 0 h, reach control temperature of 25°C for recovery after heat; 2 h, recovery at 25°C for 2 h, similar names for other time points during recovery. The y axes are in arbitrary units. Background shading indicates consensus expression pattern ± sd of eigenmodule members. The number at the top of each facet (e.g., TM1: 628, PM1_a:148) represents the total number of genes/proteins significantly associated with the given module (ANOVA, FDR < 0.05, genes/proteins can only belong to a single module). Select statistically significantly enriched MapMan functional terms are displayed in each facet. Ratios after each MapMan term (e.g., heat stress, 26/47) represent the number of genes/proteins with the assigned MapMan term in the given module relative to the number of genes/proteins with the assigned MapMan term in our entire dataset. Pie charts show the fraction of genes/proteins associated with the given module that have at least one assigned MapMan functional term (green) relative to those associated with the given module but without MapMan functional terms (white). Full functional enrichment analysis can be found in Supplementary Data 5 and 7.

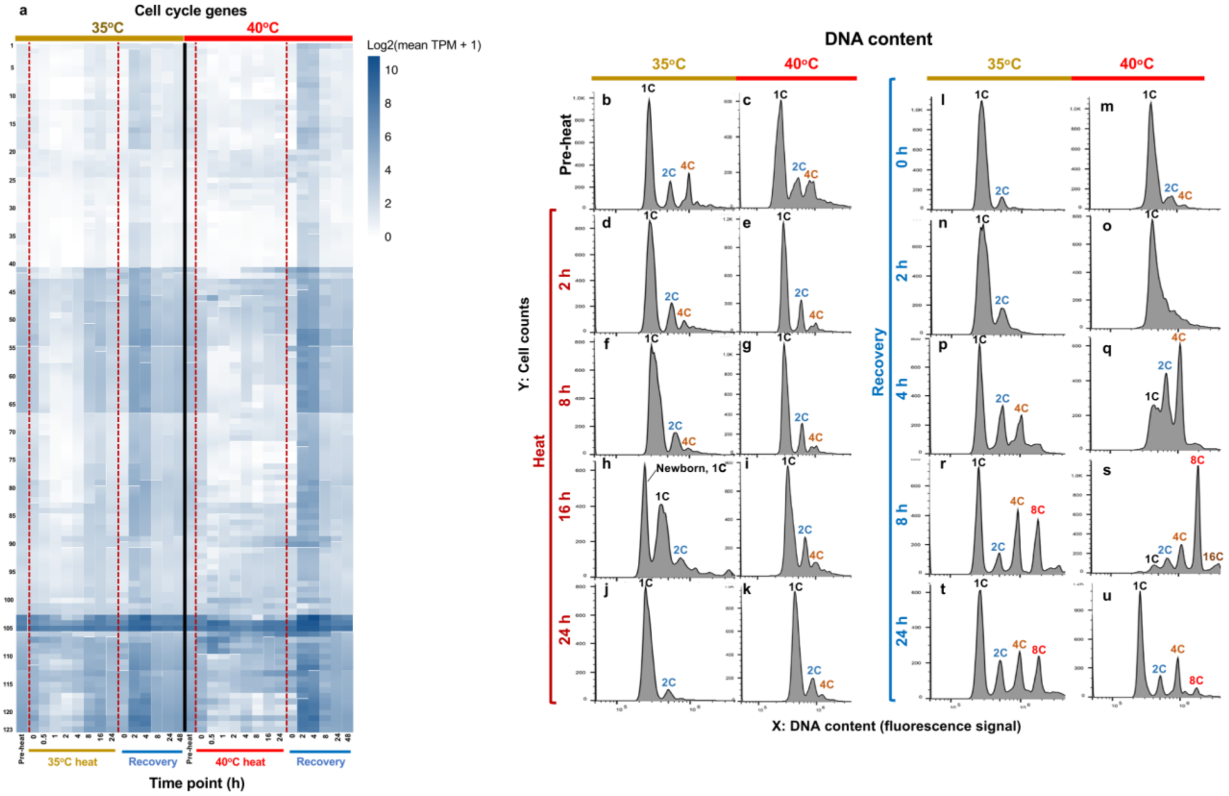


Fig. 4 Heat at 35°C synchronized the cell cycle while heat at 40°C arrested it. **a** Expression of cell cycle genes during and after heat treatment of 35°C or 40°C. Genes used for the expression pattern analysis are listed in Supplementary Data 6. Gene expression patterns for treatments of 35°C (left) and 40°C (right) are displayed as $\log_2$(mean TPM value + 1). TPM, transcripts per million. Darker blue colors indicate higher expression. Red dashed lines indicate the start and end of the heat treatments. Pre-heat, before heat treatment. Time points during heat: 0 h, reach high temperature of 35°C or 40°C; 2 h, heat at 35°C or 40°C for 2 h, similar names for other time points during heat. Time points during recovery: 0 h, reach control temperature of 25°C for recovery after heat; 2 h, recovery at 25°C for 2 h, similar names for other time points during recovery. **b**–**u** FACS (fluorescence-activated cell sorting) analysis of the DNA content in algal samples harvested at different time points before, during, and after heat treatment at 35°C or 40°C. For each figure panel, X axis is DNA content determined by plotting a histogram of fluorescence level (area of the fluorescence channel signal) in log-scale; Y axis is the cell counts in linear scale. DNA copy No. are labeled on top of each corresponding DNA content peak, 1 C (single DNA copy number), 2 C, 4 C, 8 C, 16 C.

during the entire 40°C heat treatment (Fig. 4a). Most cell cycle related genes had increased expression during 2- and 4-h of recovery in both treatment groups.

During the 40°C heat treatment, we observed irregular expression patterns of cell cycle related genes as compared to pre-heat, which led us hypothesize that the cell cycle had been arrested during 40°C. To investigate this hypothesis, we quantified cellular DNA content using flow cytometry (Fig. 4b±u). Pre-heat cultures showed typical asynchronous populations: most cells had 1 C (single DNA copy number) while a small fraction of cells had 2 C and 4 C (Fig. 4b, c). After 16-h at 35°C heat treatment, the broad 1 C size distribution from 8-h split; some of the bigger cells went to 2 C and the rest of the small cells stayed at 1 C (Fig. 4f, h). The 2 C population then divided by 24-h at 35°C, resulting in almost exclusively small 1 C cells with the same cell size as 1 C cells at pre-heat, suggesting culture synchrony (Fig. 4j). The cell division during 8-16 h of heat at 35°C was consistent with the recovery of cell cycle genes at the same time points (Fig. 4a). After 2-h of recovery at 25°C following 35°C heat, there were much fewer 2 C and 4 C cells than pre-heat, suggesting a partially synchronized population, until an almost complete recovery by 4-h at 25°C (Fig. 4l, n, p, r, t). These results indicated that the 35°C treatment synchronized cell division in Chlamydomonas.

Heat treatment at 40°C inhibited DNA replication and cell division, with three peaks of 1 C, 2 C, and 4 C persisting during all 40°C heat time points (Fig. 4c, e, g, i, k). By the end of the 24-h heat treatment at 40°C, the three cell populations had a much larger cell size, as evidenced by the right shift of the DNA peaks due to an increased cell size background effect (Fig. 4k, more information of the background effect can be found in the Methods). Cells started to replicate DNA between 2- and 4-h of recovery following 40°C heat, resulting in reduced 1 C but increased 2 C and 4 C cell populations (Fig. 4q). DNA replication continued until 8-h of recovery, resulting in the accumulation of high-ploidy level cells, ranging from 1 C to 16 C (Fig. 4s). By 24-h recovery, the cellular DNA content had almost recovered to the pre-heat level (Fig. 4u).

Cytological parameters confirmed cell cycle arrest during 40°C heat. Brightfield images and cell size quantification of algal cells showed that the 40°C treated cells had continuously increased cell size throughout the high-temperature period, followed by gradual recovery to the pre-stress level after returning to 25°C for 24 h (Fig. 5a, b, c). The quantity of chlorophyll, carotenoids, and protein per cell all increased during the 40°C heat treatment and gradually recovered after heat (Fig. 5d–g). ROS level per cell had a clear increasing tread in 40°C-treated cells (especially at the end of the heat and early recovery), although the changes were not significant after stringent statistical analysis with FDR correction (Fig. 5g). When normalized to cell volume, chlorophyll and carotenoid contents increased during 40°C heat treatment while no change of protein and ROS contents was observed (Supplementary Fig. 11a–d). The ratios of chlorophyll a/b and chlorophyll/carotenoid decreased during 40°C heat (Supplementary Fig. 11e, f). The changes of chlorophyll a and b were consistent with that of total chlorophyll (Supplementary Fig. 10g–j, Fig. 5d). Cell diameter, cell volume, chlorophyll and carotenoid contents only changed transiently after shifting to 35°C and after shifting back to 25°C (Fig. 5b–e, Supplementary Fig. 11).

Heat at 40°C impaired photosynthesis while the effects at 35°C were minor. We hypothesized that heat treatments might affect photosynthetic activities based on the changes of pigment contents (Fig. 5d, e), differentially regulated genes related to photosynthesis (Supplementary Fig. 9a, b), and the kinetics of transcripts and proteins related to the MapMan bin photosynthetic light reactions (Fig. 6, Supplementary Data 10). Most transcripts related to photosynthetic light reactions decreased during the early recovery from both heat treatments, followed by a gradual returning to the pre-stress levels (Fig. 6a, b, e, f, i, j), consistent with network modeling data (Fig. 3a). Proteins related to PSI, LHCII (light harvesting complex II), and LHCI (light harvesting complex I) increased during both heat treatments and decreased back to the pre-heat levels during the recovery (Fig. 6c, d, g, h). Proteins related to the ATP synthase decreased during the 40°C heat treatment and the early/middle recovery of both 35°C and 40°C heat treatments (Fig. 6k, l). Overall, the kinetics of transcripts and proteins related to the MapMan bin photosynthetic light reactions showed similar trends in both 35°C and 40°C treatments.

To investigate whether these pronounced changes of proteins related to LHCI, LHCII, PSI and ATP synthase under 35°C and 40°C affected photosynthesis, we measured various photosynthetic parameters during and after the heat treatments (Figs. 7–8). The PSII efficiency and linear electron flow rates decreased during 40°C heat (especially under light intensities exceeding the growth light of 100 μmol photons m^{-2} s^{-1}) while the 35°C heat treatment did not extensively affect these photosynthetic parameters (Fig. 7a–d). The Q_A redox state reflects the balance between excitation energy at PSII and the rate of the Calvin-Benson Cycle[49,50]. The amount of reduced Q_A is proportional to the fraction of PSII centers that are closed[51]. Under 35°C and 40°C, Q_A had no significant changes, although 40°C-treated cells showed the trend of increased redox status (Fig. 7e, f). Both 35°C and 40°C increased the formation of NPQ; however, the increased NPQ was steady during 4-24 h of heat at 35°C while during 40°C heat, NPQ first increased to a maximum at 8-h heat, then decreased by 24-h heat (Fig. 7g, h), suggesting that accumulative heat damages under prolonged exposure to 40°C eventually exceeded the photoprotective capacity of NPQ. Relative PSII antenna size increased during the 40°C heat treatment, while it increased transiently during the 35°C heat treatment (Fig. 7i, Supplementary Fig. 12).

Additionally, we performed electrochromic shift (ECS) measurements to monitor the effects of heat on the transthylakoid proton motive force (*pmf*, estimated by ECS_t) and proton conductivity (Fig. 8a, b, Supplementary Fig. 13a, b). No significant changes in *pmf* and proton conductivity were observed during and after the 35°C treatment (Fig. 8a, Supplementary Fig. 13a). The *pmf* increased particularly at late time points during the 40°C treatment, followed by a slow and partial recovery after shifting cells back to 25°C. Proton conductivity decreased during and after 40°C heat treatments, suggesting reduced or compromised ATP synthase activity (Supplementary Fig. 13b), consistent with reduced abundance of proteins related to ATP synthase (Fig. 6l). During both heat treatments, NPQ formation became more sensitive to *pmf*, with higher NPQ formed at a given *pmf* compared to the pre-heat condition (Fig. 8c, d), consistent with a previous report in tobacco plants[52]. The increased sensitivity of NPQ was collapsed by the end of the 24-h heat treatment at 40°C (Fig. 8d). P700 measurement revealed that the activity of cyclic electron flow around PSI (CEF) increased during both 35°C and 40°C heat, which recovered quickly after 35°C treatment but much more slowly after 40°C treatment (Fig. 8e). P700 appeared to be more reduced during 35°C and 40°C heat, although the changes were not significant with stringent FDR correction (Supplementary Fig. 13c). Furthermore, gross photosynthetic O_2 evolution rates and dark respiration rates had little changes during the 35°C treatment but dropped significantly during the 40°C heat treatment (Fig. 8f, g, h). Photosynthetic parameters had no significant changes in cultures maintained under constant 25°C (Supplementary Fig. 14).

Heat at 40°C altered thylakoid and pyrenoid ultrastructure. The effects of high temperatures on photosynthesis prompted us to investigate cellular ultrastructure using transmission electron microscopy (TEM) (Fig. 9a–r). Thylakoids became disorganized and loosely packed in cells treated with 40°C (Fig. 9f, g). Investigation of pyrenoid ultrastructure (Fig. 9j–r, s) showed that cells

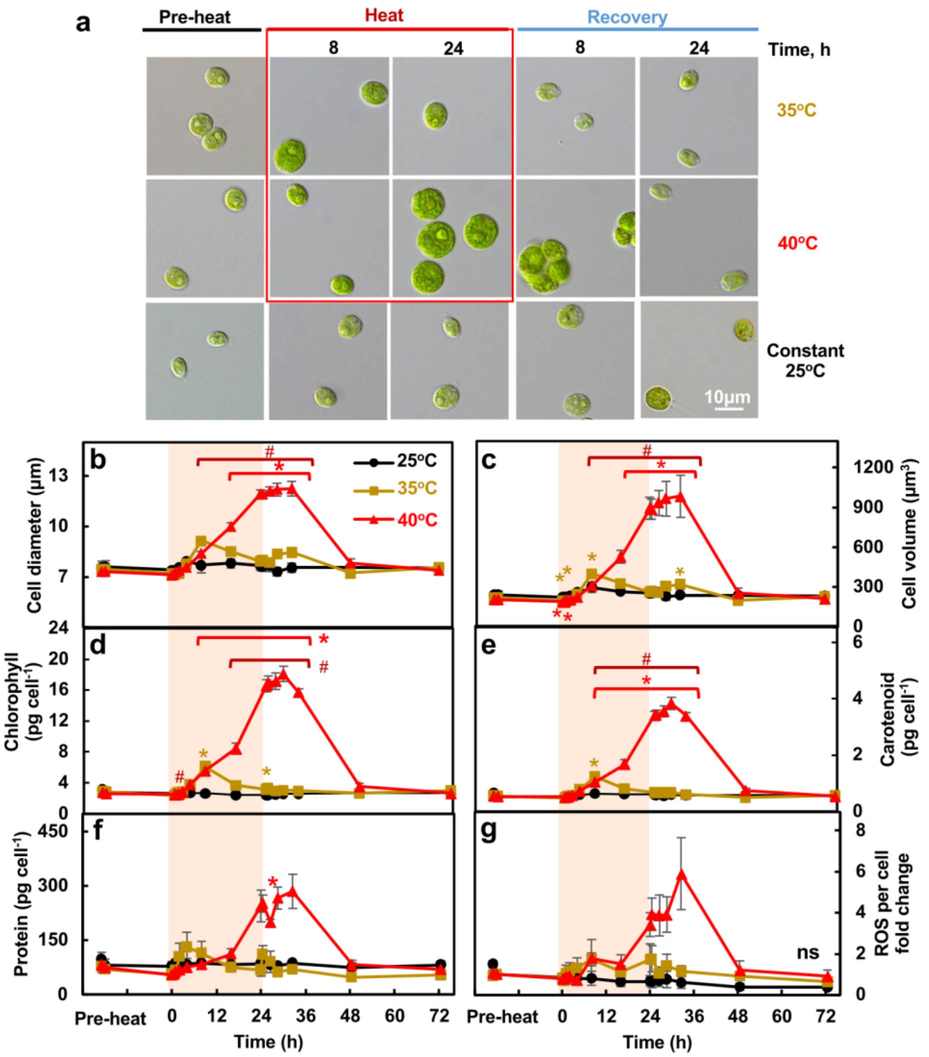


Fig. 5 Heat of 40°C persistently increased cell size, cellular levels of pigments, and proteins while these effects were transient with 35°C heat. a Light microscopic images of Chlamydomonas cells. **b**, **c** Cell diameters and volume determined using a Coulter Counter. **d**–**f** Total chlorophyll, carotenoid, protein content per cell. **g** Fold-change of reactive oxygen species (ROS) levels per cell quantified using CM-H2DCFDA ROS indicator. Mean ± SE, $n = 3$ biological replicates. Black, brown and red curves represent experiments with constant 25°C, treatments of 35°C or 40°C respectively. Red shaded areas depict the duration of high temperature. Statistical analyses were performed using two-tailed t-test assuming unequal variance by comparing treated samples with 25°C at the same time point (*, $p < 0.05$) or between 35°C and 40°C at the same time point (#, $p < 0.05$). **b**–**g** P values were corrected by FDR. The colors and positions of asterisks (*) match the treatment conditions and time points, respectively. The positions of pound signs (#) match the time points. (g) Not significant, ns, $p > 0.05$ after FDR correction.

treated with 40°C had altered pyrenoid matrices and absence of thylakoid tubules inside the pyrenoid (Fig. 9o, p), suggesting an inefficient carbon concentrating mechanisms (CCM). No changes in pyrenoid ultrastructure were observed in cells treated with 35°C (Fig. 9k, l). ImageJ quantification of pyrenoid structures showed that cells treated with 40°C had increased pyrenoid areas, which was attributed to increased areas of both the pyrenoid matrix and starch sheath (Fig. 9t, u, v). The increased pyrenoid size was abolished after 8-h of recovery. Biochemical quantification of starch contents showed that both 35°C and 40°C treatments increased starch levels per cell and per cell volume, which decreased during recovery (Fig. 9w, x). At the end of the heat treatments and early recovery, cells exposed to 40°C had a higher starch content per cell than those exposed to 35°C, but the differences between the two treatments were not significant per cell volume.

Discussion

We investigated how Chlamydomonas cells respond to moderate (35°C) and acute (40°C) high temperature at systems-wide levels (Fig. 1c). Our results show that 35 and 40°C triggered shared and unique heat responses in Chlamydomonas (Fig. 10).

Both high temperatures induced the expression of *HSF1* and *HSF2*, as well as canonical high-temperature response genes *HSP22A* and *HSP90A*, increased cell size, chlorophyll and carotenoid contents, PSII/PSI ratio, NPQ, CEF, PSI redox state, and starch formation (Fig. 10). The changes under 35°C were often transient and moderate while those under 40°C were sustained and dramatic. The correlation between transcripts and proteins increased during both heat treatments, suggesting that responses during heat were largely transcriptionally regulated. Functional categories of gluconeogenesis/glyoxylate-cycle and abiotic stress had the highest correlation between transcripts and proteins in

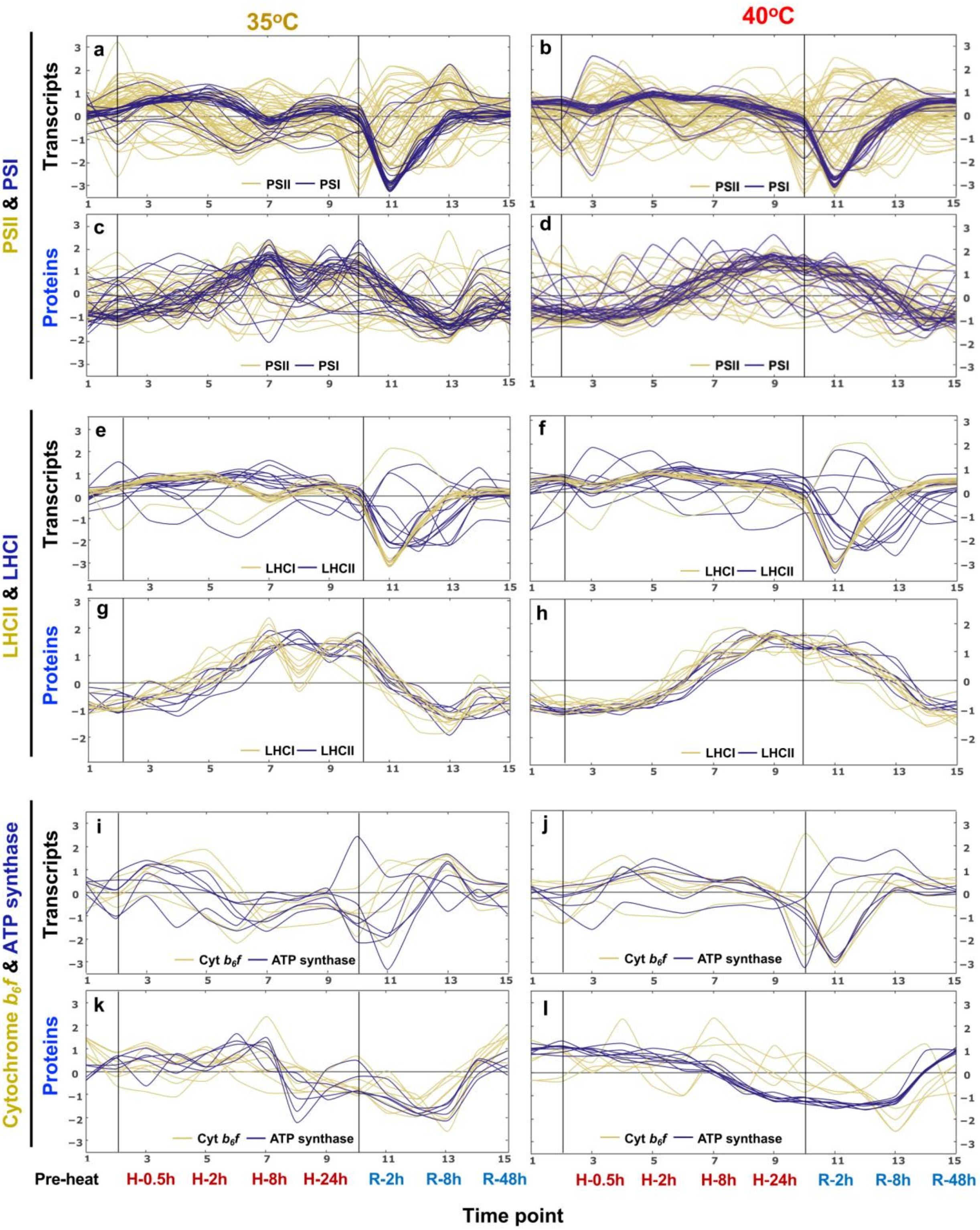


Fig. 6 Transcripts and proteins related to photosynthetic light reactions changed dynamically during and after 35°C and 40°C heat treatments. Signals of transcripts (**a**, **b**, **e**, **f**, **i**, **j**) and proteins (**c**, **d**, **g**, **h**, **k**, **l**) related to the MapMan bin photosynthetic light reactions, including PSII and PSI (**a–d**), LHCII and LHCI (**e-h**), Cytochrome b_6f and ATP synthase (**i–l**), were standardized to z scores (standardized to zero mean and unit variance) and plotted against equally spaced time point increments. The black vertical lines indicate the start and end of heat treatments at 35°C (**a**, **c**, **e**, **g**, **i**, **k**) and 40°C (**b**, **d**, **f**, **h**, **j**, **l**), respectively. Time points are labeled at the bottom. Timepoint 1: pre-heat. Time points 2-9, heat treatment at 35°C or 40°C, including reaching high temperature (0), 0.5, 1, 2, 4, 8, 16, 24 h during heat; time points 10–15, recovery phase after heat treatment, including reaching control temperature (0), 2, 4, 8, 24, 48 h during recovery. See the interactive figures with gene IDs and annotations in Supplementary Data 10 (transcript/protein dynamics), the groups of PS.lightreaction (PS for photosynthesis).

early time points of 35°C and 40°C treatment (Fig. 2f), respectively. High correlation values of these functional categories indicate that these responses may occur rapidly without much post-transcriptional regulation, which may help coordinate activities to adapt to high temperature quickly and efficiently. The decreased correlation between transcripts and proteins during both recoveries was consistent with increased protein posttranslational modification after heat treatments based on the RNA-seq network modeling results (Fig. 3a, TM8/11).

The increased NPQ in cells treated with 35°C and 40°C heat suggested that both heat treatments compromised photosynthetic efficiency (Fig. 7g, h). Heat at 40°C reduced photosynthetic efficiency much more than 35°C, especially when photosynthetic parameters were evaluated with light intensities higher than the growth light of 100 µmol photons m^{-2} s^{-1} (Figs. 7 and 8). Under the growth light we employed, the differences in photosynthetic parameters between 35°C and 40°C were smaller with comparable values, consistent with the similar kinetics of transcripts and proteins related to the MapMan bin photosynthetic light reactions (Fig. 6). We conclude that in algal cultures grown in PBRs under a growth light of 100 µmol photons m^{-2} s^{-1}, both 35°C and 40°C heat treatments affected photosynthetic efficiency but photosynthetic activity was maintained at a comparable level during both heat treatments; however, increasing light intensities exaggerated the heat-induced damages to photosynthesis, especially with the 40°C treatment.

During early recovery from both heat treatments, transcripts and proteins related to DNA synthesis increased while those related to photosynthetic light reactions decreased (Fig. 3). In synchronized algal cultures under day/night cycles, genes related to DNA synthesis and cell cycle peak during the early dark phase when the genes related to photosynthetic light reactions had minimal expression[53,54]. However, under constant light as in our experiment, genes related to DNA synthesis and photosynthetic light reactions may express simultaneously but their quantitative expression manner under constant light is understudied. The induction of cell cycle genes after recovery were comparable after both 35°C and 40°C heat (Fig. 4a), but the significant down-regulation of genes related to photosynthetic light reactions only occurred in the recovery following 40°C (Supplementary Fig. 9a). The photosynthetic light reactions are a major source of ROS production[55]. The measured ROS levels reflect the competition between ROS production and scavenging. Although the increase of the measured ROS level was not significant, the up-regulation of ROS scavenging transcripts (Supplementary Fig. 9e) during the heat and early recovery phase of 40°C supported the increased ROS production with 40°C treatment. Synchronized cells at the dark phase of day/night cycle most likely have minimal ROS production, as evidenced by the down-regulation of many ROS response genes[53]. Cells with 40°C heat had very different physiologies from synchronized culture at the early dark phase under the control temperature, considering the increased ROS production and heat damaged cellular structures in 40°C-treated cells. Thus, we suspect the up-regulation of transcripts related to cell cycle and down-regulation of transcripts related to photosynthetic light reactions during the recovery from 40°C under constant light and during the dark phase of day/night-cycle may be due to different mechanisms. ROS accumulation impairs DNA replication and induces DNA damage[56,57]. Mammalian cells are more sensitive to high temperatures during DNA replication than other cell cycle stages[58]. We propose that down-regulation of photosynthetic light reactions during DNA synthesis is beneficial to resuming the cell cycle by reducing ROS production during the early recovery. The mechanisms of the opposite transcriptional regulation of DNA replication and photosynthetic light reactions during the recovery from high temperatures is unknown but interesting for future research. One form of ROS is H_2O_2, which is highly diffusible and stable[55]. Recently, Niemeyer et al. employed hypersensitive H_2O_2 sensors in different compartments of Chlamydomonas cells and showed that H_2O_2 levels increased in the nucleus after heat treatment, suggesting diffusion of H_2O_2 from other cellular compartments to the nucleus[21]. H_2O_2 has been proposed as a secondary messenger in signal transduction[59]. Increased H_2O_2 in the nucleus may affect gene expression, e.g., those involved in photosynthetic light reactions.

Both high temperatures affected cell cycle genes, but during 35°C cells could recover the expression of cell cycle genes to the pre-heat level after 8-h of heat treatment with a partially synchronized cell cycle at the end of the 24-h heat treatment (Fig. 4). In contrast, under the 40°C heat treatment, the expression of cell cycle genes was disrupted. Additionally, we observed down-regulation of genes encoding cell wall proteins during both 35°C and 40°C heat treatments and many of these genes were up-regulated during the recovery of 35°C and 40°C; the differential regulation of cell wall genes was more dramatic with 40°C than 35°C treatment (Supplementary Fig. 9h). The Chlamydomonas cell wall protects cells from environmental challenges; it is proposed that osmotic/mechanical stresses and cell wall integrity regulate the expression of cell wall genes in Chlamydomonas[60]. However, the underlying mechanisms of how the cell wall responds to high temperatures are largely under-explored. Cell walls form around daughter cells during cell division. Under diurnal regulation, many cell wall genes are up-regulated following the up-regulation of many cell cycle genes during cell division[53]. The expression pattern of cell wall genes in our data may be related to the inhibited and resumed cell cycles during and after heat treatments, respectively.

Heat at 35°C and 40°C induced unique transcriptional responses (Fig. 3a, Supplementary Fig. 4, 5, 6, 9, 10). Heat at 35°C induced a unique gene set that was not induced under 40°C, including genes involved in gluconeogenesis and glycolysis, mitochondrial assembly, and a putative calcium channel (Supplementary Fig. 5c–f, Supplementary Data 1, 3). While most of the overlapping DEGs between 35°C and 40°C treatments were more strongly differentially expressed at 40°C than at 35°C, a fraction of the overlapping DEGs displayed a larger fold-change at 35°C than at 40°C (Supplementary Fig. 6). One group of genes with higher upregulation at 35°C than at 40°C are those low-CO_2 inducible (LCI) genes, e.g., *LCI26* (at reaching high temperature and 8-h heat), *LCI19* (16-h heat) (Supplementary Data 1), suggesting an effort to compensate for increased CO_2 demands with increased growth at 35°C. Many of these uniquely regulated genes in the 35°C treatment have unknown functions and may include novel candidates important for acclimation to moderate high temperature.

Both 35°C and 40°C induced starch accumulation but possibly for different reasons (Fig. 9w, x, 10). The increased starch in 35°C treated cells may be due to increased acetate uptake/assimilation, gluconeogenesis and the glyoxylate cycle, as evidenced by an induction of proteins related to these pathways (Fig. 2f, Supplementary Fig. 8). In Chlamydomonas, acetate uptake feeds into the glyoxylate cycle and gluconeogenesis for starch biosynthesis[40,41]. The increased starch in 40°C treated cells may be due to inhibited cell division (Figs. 4 and 5), resulting in starch storage exceeding its usage[25,26]. Starch accumulation could also be an electron sink to alleviate the over-reduced electron transport chain during 40°C heat treatment[18]. Heat treated Arabidopsis plants (42°C for 7 h) also had increased starch[43]. Several genes involved in starch biosynthesis were induced during 40°C heat and early recovery (Supplementary Fig. 9f). The over-accumulated starch during 40°C may also contribute to the downregulation of genes involved in acetate uptake and assimilation (Supplementary Fig. 8g, h).

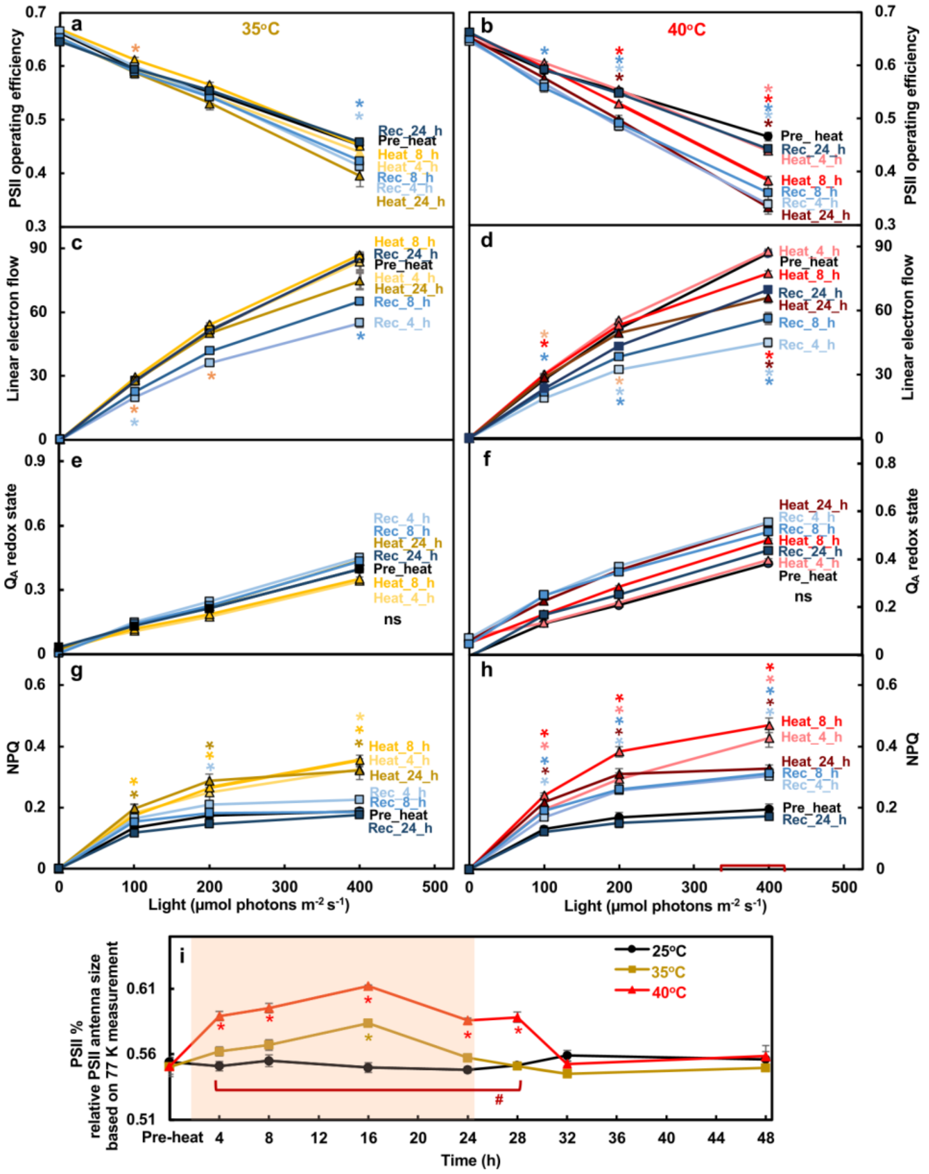


Fig. 7 Heat of 40°C impaired PSII efficiency more than heat of 35°C while both induced NPQ. Algal cultures harvested from PBRs before, during and after heat treatment at 35°C (**a**, **c**, **e**, **g**, **i**) or 40°C (**b**, **d**, **f**, **h**, **i**) were used for photosynthetic measurements. **a**–**h** Photosynthetic parameters measured using room temperature chlorophyll fluorescence. (**a**, **b**) PSII efficiency, the data at 0 μmol photons m^{-2} s^{-1} light are the maximum PSII efficiency in dark and data from light phase are PSII operating efficiency in light-adapted cells. **c, d** Linear electron flow, accounting for the changes of PSII antenna size during the treatments as in (**i**). **e**, **f** Q_A redox state, the redox state of chloroplastic quinone A (Q_A), the primary electron acceptor downstream of PSII; the bigger number of Q_A redox state means more reduced Q_A. **g**, **h** Nonphotochemical quenching, NPQ. **i** Relative PSII antenna fraction, percentage of light distributed to PSII measured by 77 K chlorophyll fluorescence. Mean ± SE, $n = 3$ biological replicates. Statistical analyses were performed using two-tailed t-test assuming unequal variance by comparing treated samples with the pre-heat samples under the same light (**a**–**h**, *) or constant 25°C samples at the same time point (**i**, *), or by comparing samples between 35°C and 40°C at the same time point (**i**, #). **a**–**i** *P* values were corrected by FDR. *, $p < 0.05$, the colors and positions of asterisks match the treatment conditions and time points, respectively. (**i**) #, $p < 0.05$, the positions of pound signs match the time points. **e**, **f** Not significant, ns.

Most of the reducing power from acetate assimilation is used in mitochondrial respiration[40,41]. However, the downregulated transcripts related to mitochondrial electron transport (Supplementary Fig. 15c), the heat sensitivity of mitochondrial respiration rates (Fig. 8h), and the over-accumulated starch may restrict acetate uptake and assimilation during 40°C heat treatment.

Heat at 35°C stimulated growth but 40°C decreased it (Fig. 1a, b). We quantified growth based on the rate of chlorophyll increase and medium consumption in PBRs under the turbidostatic mode, monitored by OD_{680} which is proportional to chlorophyll content (Supplementary Fig. 1a–c). Under our experimental conditions with little nutrient and light limitation, our results showed that cells exposed to 35°C reached the maximum OD_{680} faster than cell exposed to 40°C or kept at constant 25°C. The stimulated growth in liquid cultures under 35°C was confirmed by growth on plates (Supplementary Fig. 2) and was consistent with increased transcripts and proteins related to mitochondrial electron transport as well as increased mitochondrial relative volume in 35°C-treated cells

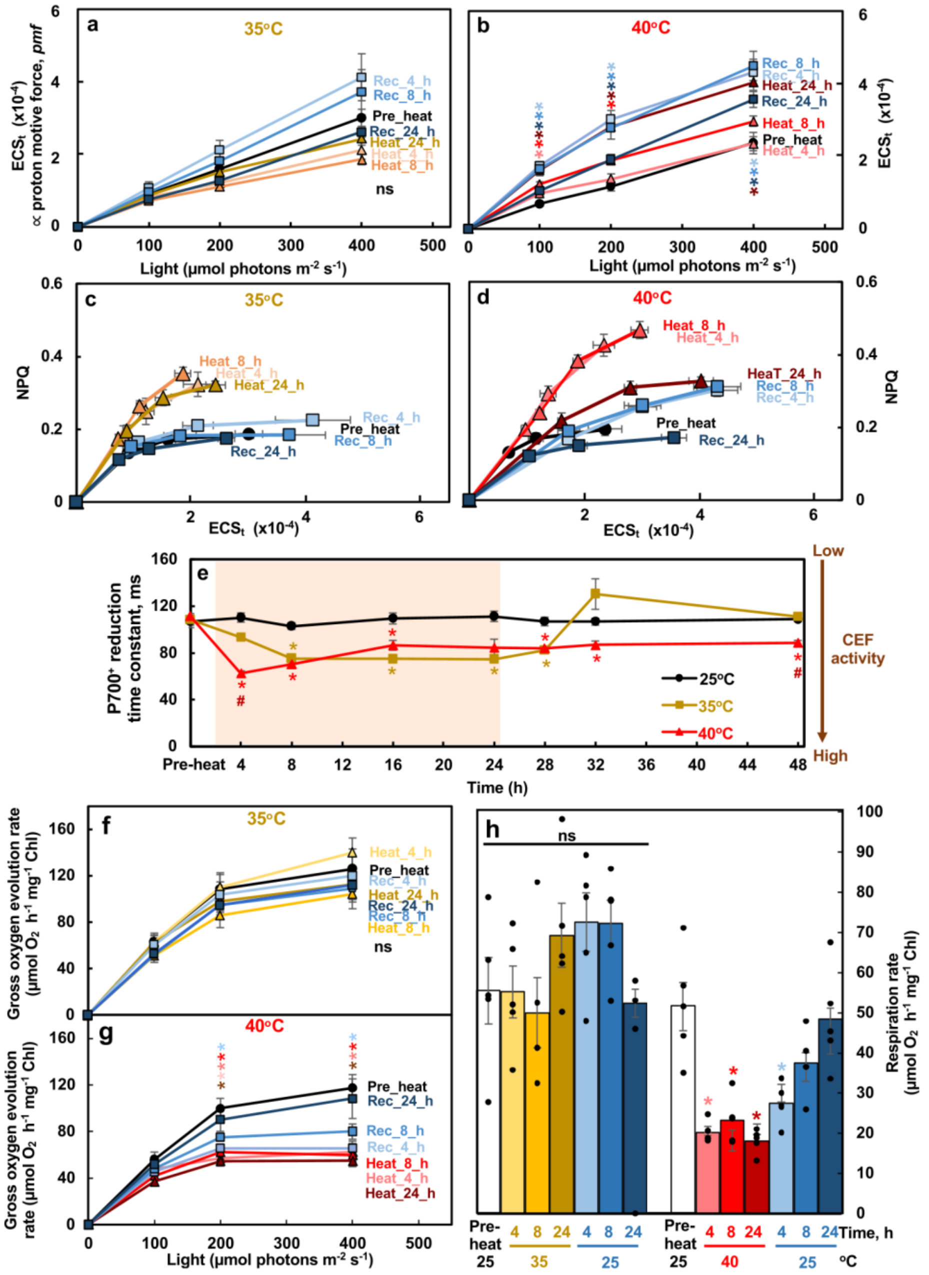


Fig. 8 Heat at 40°C induced transthylakoid proton motive force, NPQ, CEF, but reduced O2 evolution and respiration rates in Chlamydomonas cells. **a**, **b** Heat treatment of 40°C increased the transthylakoid proton motive force (*pmf*). ECS_t, measured by electrochromic shift (ECS), represents the transthylakoid *pmf*. **c**, **d** NPQ was more sensitive to ECS_t (or *pmf*) during both heat treatments, with higher NPQ produced at a given *pmf*. NPQ, non-photochemical quenching, measured using room temperature chlorophyll fluorescence. **e** Both 35°C and 40°C induced the activity of cyclic electron flow around PSI (CEF) although with different dynamics and reversibility. $P700^+$ reduction to measure CEF in the presence of 10 μmol DCMU to block PSII activity; the smaller $P700^+$ reduction time constant indicates faster $P700^+$ reduction and higher CEF activity. The red shaded area depicts the duration of the high temperature. **f**, **g**, **h** Gross O_2 evolution rates and respiration rates were reduced during the 40°C heat treatment, measured using a Hansatech Chlorolab 2 Clark-type oxygen electrode. Mean ± SE, $n = 5$ biological replicates. Statistical analyses were performed using two-tailed t-test assuming unequal variance by comparing treated samples with the pre-heat samples under the same light (**a**, **b**, **f**, **g**, **h**, *) or constant 25°C samples at the same time point (**e**, *), or by comparing samples between 35°C and 40°C at the same time point (**e**, #). **a**, **b**, **e**, **f**, **g** *P* values were corrected by FDR. *, $p < 0.05$, the colors and positions of asterisks match the treatment conditions and time points, respectively. (**e**) #, $p < 0.05$, the positions of pound signs match the time points. Not significant, ns.

(Supplementary Fig. 15). The increased protein levels in acetate uptake/assimilation and gluconeogenesis and glyoxylate cycles (Supplementary Fig. 8) may contribute to the faster growth under 35°C heat. Maize plants grown under moderate high temperature of 33°C had increased biomass but decreased biomass under higher temperature of 37°C as compared to controls at 31°C[61]. Similar temperature effects were also reported in synchronized algal cultures[62,63], consistent with our data. With increasing high

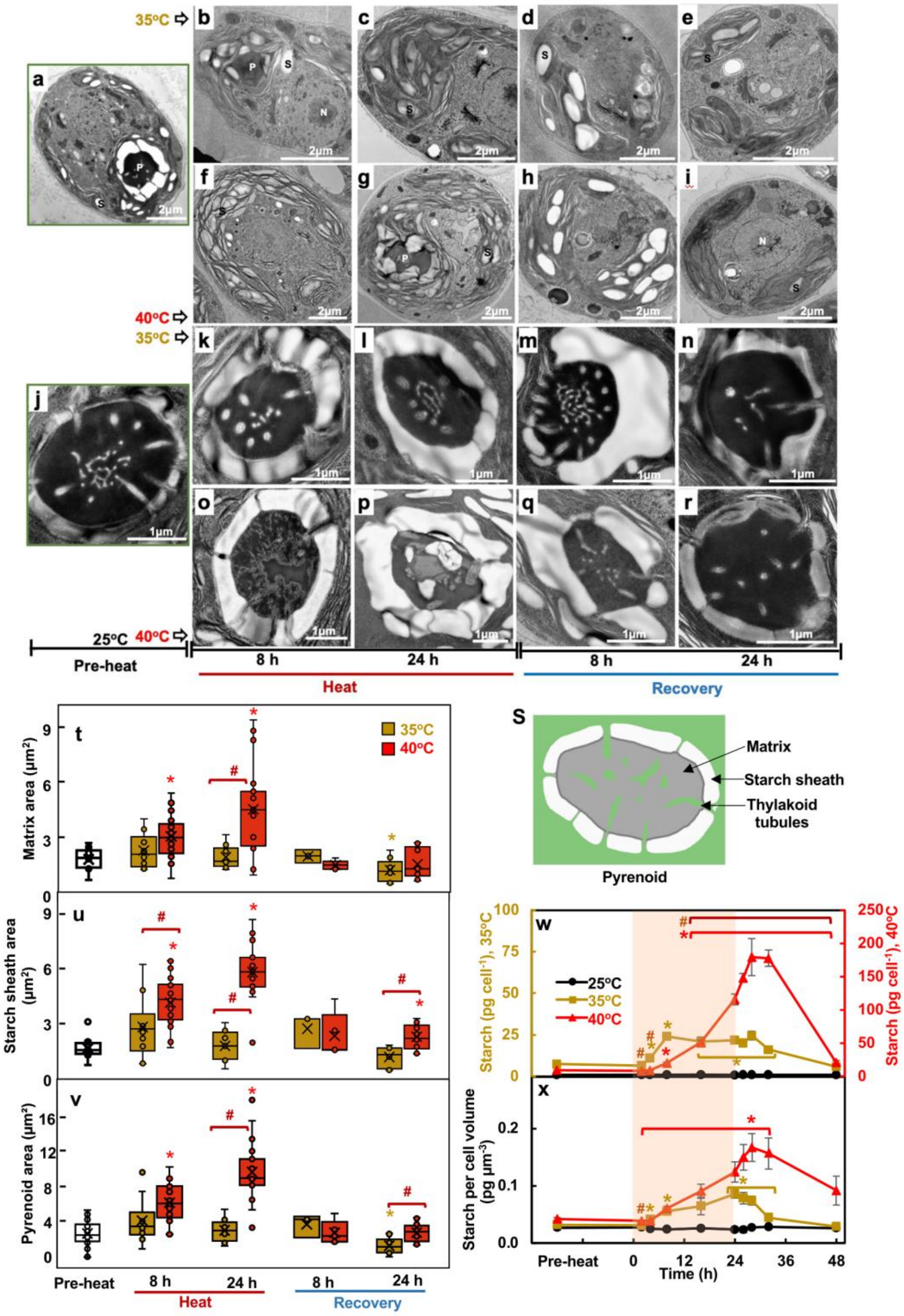


Fig. 9 Only 40°C altered thylakoid and pyrenoid ultrastructure while both 35°C and 40°C treatments stimulated starch accumulation. **a**–**r** Representative transmission electron microscopy (TEM) images of algal cells or pyrenoids at different time points before (**a**, **j**), during, and after heat treatment at 35°C (**b**–**e**, **k**–**n**) or 40°C (**f**–**i**, **o**–**r**). **s** Cartoon representation of Chlamydomonas pyrenoid structure. **t**, **u**, **v** Areas of pyrenoid matrix, starch sheath and the whole pyrenoids, respectively, quantified using ImageJ and TEM images from algal samples harvested before, during, and after heat treatments of either 35°C (brown) or 40°C (red) at the indicated time points. The data is presented as boxplot based on Tukey-style whiskers. Median values are represented by the horizontal black lines and mean values by the X sign inside each rectangular box. **w**, **x** Starch quantification using starch assay kits. Values are mean ± SE, $n = 3$ biological replicates. The red shaded areas depict the duration of the high temperature. **t**–**x** Statistical analyses were performed using two-tailed t-tests assuming unequal variance by comparing treated samples with pre-heat (**t**–**v**) or 25°C at the same time point (w, x) (*, $p < 0.05$, the colors of asterisks match the treatment conditions), or by comparing samples between 35°C and 40°C at the same time point (#, $p < 0.05$). (**w**, **x**) *P* values were corrected by FDR.

temperatures, the growth of photosynthetic organisms may accelerate first, then decrease when the high temperature exceeds a certain heat threshold.

The adaptive transcriptional changes in response to 40°C include rapid induction of transcripts encoding HSFs, HSPs, and ROS scavenging enzymes (Supplementary Fig. 9c, e, 10a). Chlamydomonas has two HSFs, HSF1 and HSF2[10]. HSF1 is a canonical HSF similar to plant class A HSFs and a key regulator of the stress response in Chlamydomonas[11], while the function of HSF2 is unclear. Our transcriptome data showed that both *HSF1* and *HSF2* were induced during early heat of 35°C and 40°C (Supplementary Fig. 10a), suggesting the potential role of HSF2 in heat regulation. Interestingly, *HSF1* was also induced during the early recovery phase after 40°C heat, possibly due to its potential roles in maintaining some heat responsive genes after heat treatment. HSF1 was shown to also be involved in altering chromatin structure for sustained gene expression[10,64]. HSP22E/F are small heat shock proteins targeted to the chloroplast, function in preventing aggregation of unfolded proteins, and are induced at temperatures at or above 39°C in Chlamydomonas[13,65]. The transcripts of *HSP22E/F* were induced transiently but strongly during 0.5- to 1-h heat of 40°C, but also during the first 4-h of recovery after 40°C heat (Supplementary Fig. 9c), suggesting their roles not only during heat but also the recovery from heat of 40°C.

Additionally, cells treated with 40°C heat had increased photoprotection (Figs. 7 and 8) and related transcripts were up-regulated (Supplementary Fig. 9a, e.g., *LHCSR*, *ELI*, and *PSBS*). With compromised photosynthesis under 40°C, the increased transthylakoid proton motive force (*pmf*), NPQ formation, sensitivity of NPQ to the *pmf*, and CEF activity were all helpful to dissipate excess light energy and reduce heat-induced oxidative stress. CEF generates only ATP but no NADPH, balances the ATP/NADPH ratio, contributes to the generation of *pmf*, and protects both PSI and PSII from photo-oxidative damage[66,67]. Increased CEF activity has been frequently reported under various stressful conditions in land plants[15,68,69] and algae[70–72]. It is proposed that the reduced plastoquinone (PQ) pool activates CEF in algae[72–74]. CEF is also proposed to provide the extra ATP needed for the carbon concentrating mechanisms (CCM) in Chlamydomonas[75]. Our results showed that increased CEF at 40°C (Fig. 8e) concurred with the induced transcripts involved in CCM (Supplementary Fig. 9b), and the increased proteins of PSI subunits (Fig. 6d).

Cells treated with 40°C heat had increased PSII/PSI ratio measured by 77 K chlorophyll fluorescence (Fig. 7i), as previously reported[18]. The 77 K chlorophyll fluorescence is often used to monitor the stoichiometries of PSII and PSI[76]. The ratios of PSII and PSI from the 77 K fluorescence emission is an indicator of the relative antenna size of each photosystem[77]. Hemme et al., (2014) reported that 42°C treated Chlamydomonas cells showed a blue shift of PSI emission peak from 713 to 710 nm, suggesting detachment of LHCI from PSI. In 40°C treated Chlamydomonas cells, we also observed the minor blue shift of the PSI emission peak but also an increased emission peak around 695 nm (Supplementary Fig. 12), which is associated with the PSII core antenna CP47[76]. Our spectral changes indicated reduced or detached PSI antenna but increased PSII antenna, thus an increased PSII/PSI ratio, suggesting relatively smaller antenna associated with PSI than PSII in cells treated with 40°C heat. Under high salt conditions, the Antarctic alga Chlamydomonas sp. UWO 241 forms a PSI-Cytochrome b_6f supercomplex with constitutively high rates of CEF but absence of a discernible PSI peak in 77 K chlorophyll fluorescence emission[78–80]. Chlamydomonas forms the PSI-Cytochrome b_6f supercomplex to facilitate CEF under anaerobic conditions[81,82]. Steinbeck et al. (2018) proposed that the dissociation of LHCA2/9 from PSI supported the formation of the PSI-Cytochrome b_6f supercomplex. In Chlamydomonas, the PSI core associates with LHCI which is comprised of ten LHCA subunits and LHCA2/9 are suggested to be weakly bound to the PSI core[83]. The PSI chlorophyll fluorescence under 77 K is mainly due to chlorophyll a in the LHCAs[76]. Combining our results with previous reports, we propose that heat-induced dissociation of LHCAs (possibly LHCA2/9) from the PSI core may facilitate the formation of the PSI-Cytochrome b_6f supercomplex and increase CEF activity.

Cells treated with 40°C heat had altered pyrenoid structures (Fig. 9j, o, p). Algae utilize pyrenoids to concentrate CO_2 around Rubisco through the CCM[84,85]. In Chlamydomonas, pyrenoids consist of three major components: starch sheath (a diffusion barrier to slow CO_2 escape), pyrenoid matrix (Rubisco enrichment for CO_2 fixation), and thylakoid tubules (delivery of concentrated CO_2 and diffusion path of Calvin-Benson Cycle metabolites) (Fig. 9s)[86]. Several pyrenoid-localized proteins sharing a conserved Rubisco-binding motif are proposed to mediate the assembly of the pyrenoid in Chlamydomonas:[87] the linker protein Essential Pyrenoid Component 1 (EPYC1) links Rubisco to form the pyrenoid matrix;[88,89] the starch-binding protein Starch Granules Abnormal 1 (SAGA1) mediates interactions between the matrix and the surrounding starch sheath;[90] the thylakoid-tubule-localized transmembrane proteins RBMP1/2 mediate Rubisco binding to the thylakoid tubules in the pyrenoid[87]. From our TEM images, thylakoid tubules appeared to be absent from the pyrenoid matrix in cells treated with 40°C heat, which may suggest that 40°C heat disrupts the interaction of thylakoid tubules with the pyrenoid matrix and compromises CCM efficiency. The transcripts of *EPYC1*, *SAGA1*, and *RBMP2* were induced during 40°C heat (Supplementary Fig. 9b). We propose that 40°C heat may increase the disorder of the pyrenoid structure and Chlamydomonas cells compensate for this by inducing transcripts encoding the pyrenoid-structure-maintaining proteins mentioned above. Several other transcripts related to the CCM, e.g., low CO_2 inducible proteins, LCIA/D/E/C (helping maintaining CO_2 concentration in pyrenoids), were all up-regulated during 40°C heat (Supplementary Fig. 9b), which may suggest the attempt to maintain the CCM and compensate for the heat induced photorespiration[18] as well as CO_2 leakage from pyrenoids. *SAGA1* and *LCIA/D/E* were also induced during early recovery, suggesting the efforts to recover the CCM and/or coordinate pyrenoid division with cell division after 40°C heat. The increased CCM transcripts during 40°C heat and early recovery may be an adaptive response to alleviate the over-reduced electron transport chain.

The increased chlorophyll during 40°C heat may be a maladaptive response. Cells treated with 40°C heat had more than 4x increased chlorophyll per cell (Fig. 5d), which could not be fully explained by increased cell volume (Supplementary Fig. 11a). Increased chlorophyll in heat treated Chlamydomonas cells has been reported previously[18], but the underlying mechanisms are unclear. Heat at 40°C appeared to promote chlorophyll biosynthesis. The gene encoding the key chlorophyll synthesis enzyme, porphobilinogen deaminase, *PBGD2*[91], was upregulated during 40°C heat (Supplementary Fig. 9d). Considering the compromised photosynthesis and decreased growth during 40°C heat, increasing chlorophyll levels to this extent is toxic. The elevated chlorophyll may lead to increased light harvesting with decreased photosynthesis in 40°C treated cells, resulting in ROS production. Chlorophyll contents positively correlate with nitrogen availability[92] and we found many genes related to the nitrogen assimilation pathways were up-regulated during 40°C heat (Supplementary Fig. 9g), providing a possible explanation for increased chlorophyll during 40°C heat. Maize plants showed greater sensitivity to high temperatures with increased

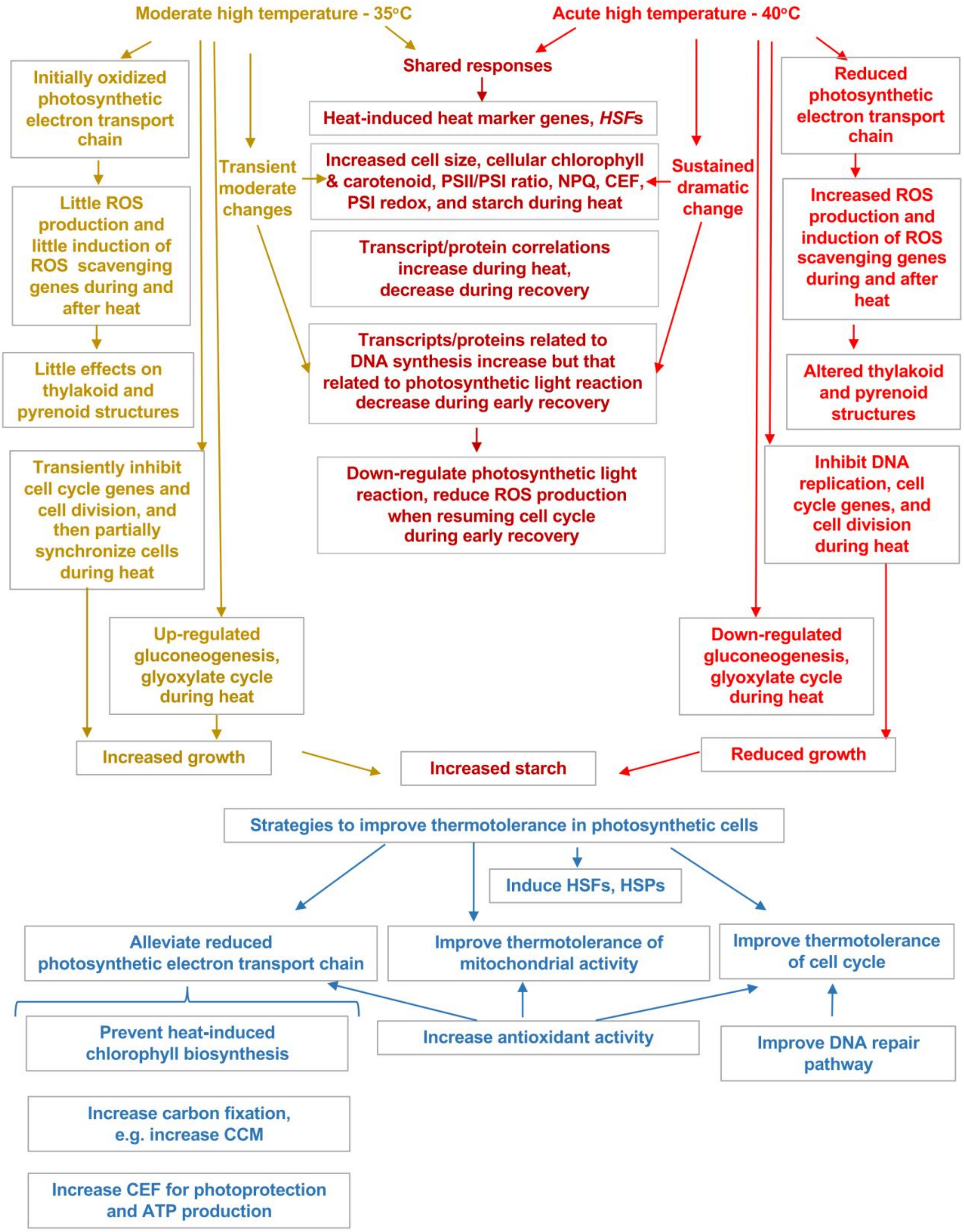

Moderate high temperature - 35°C
Acute high temperature - 40°C
Shared responses
Initially oxidized photosynthetic electron transport chain
Little ROS production and little induction of ROS scavenging genes during and after heat
Little effects on thylakoid and pyrenoid structures
Transiently inhibit cell cycle genes and cell division, and then partially synchronize cells during heat
Up-regulated gluconeogenesis, glyoxylate cycle during heat
Increased growth
Transient moderate changes
Heat-induced heat marker genes, HSFs
Increased cell size, cellular chlorophyll & carotenoid, PSII/PSI ratio, NPQ, CEF, PSI redox, and starch during heat
Transcript/protein correlations increase during heat, decrease during recovery
Transcripts/proteins related to DNA synthesis increase but that related to photosynthetic light reaction decrease during early recovery
Down-regulate photosynthetic light reaction, reduce ROS production when resuming cell cycle during early recovery
Increased starch
Sustained dramatic change
Reduced photosynthetic electron transport chain
Increased ROS production and induction of ROS scavenging genes during and after heat
Altered thylakoid and pyrenoid structures
Inhibit DNA replication, cell cycle genes, and cell division during heat
Down-regulated gluconeogenesis, glyoxylate cycle during heat
Reduced growth
Strategies to improve thermotolerance in photosynthetic cells
Induce HSFs, HSPs
Alleviate reduced photosynthetic electron transport chain
Improve thermotolerance of mitochondrial activity
Improve thermotolerance of cell cycle
Prevent heat-induced chlorophyll biosynthesis
Increase antioxidant activity
Improve DNA repair pathway
Increase carbon fixation, e.g. increase CCM
Increase CEF for photoprotection and ATP production

Fig. 10 Chlamydomonas has shared and unique responses to moderate and acute high temperatures of 35°C and 40°C. Top panel: Summarized results from multiple-level anlayses revealed the unique responses to 35°C (left, brown), unique to 40°C (right, bright red), and those shared between the two treatments (middle, dark red). **Bottom panel** (blue): Our results can be used to inform strategies to improve thermotolerance in photosynthetic cells.

nitrogen fertilization[93], which may support the possible links among nitrogen assimilation, chlorophyll biosynthesis, and heat responses. In land plants, long-term (e.g. several days) heat stress reduces chlorophyll content[94,95], however, the underlying mechanisms by which chlorophyll is degraded during long-term heat remain elusive[94]. Chlamydomonas cells treated at 39°C for more than one day had initially increased chlorophyll (8~16-h heat) followed by chlorophyll loss, cell bleaching, and death (33-h heat)[25]. It is possible that preventing chlorophyll increase during acute high temperature (especially early stage) could lead to improved thermotolerance in algae.

Combining our systems-wide analyses, we could distinguish adaptive versus maladaptive heat responses as mentioned above. The potential engineering targets for improved thermotolerance may include these adaptive heat responses, e.g., heat induced HSFs, HSPs, photoprotection, CEF, antioxidant pathways, and CCM transcripts (Fig. 10). The maladaptive responses could also be the targets for improved thermotolerance if we could find mediating solutions to reduce these changes, e.g., heat-induced chlorophyll. The cell cycle arrest induced by acute high temperature may be maladaptive but also adaptive: the halted cell division may be one of the main reasons for over-accumulated starch, reduced photosynthetic electron transport chain, and increased ROS production; on the other side, cell cycle arrest under acute high temperature may prevent damages/errors during DNA replication when DNA repair pathways are compromised by heat[58]. Thus, increasing thermotolerance of the DNA repair and cell cycle pathways may also be strategies to improve heat tolerance in photosynthetic cells. Furthermore, mitochondrial activity was stimulated slightly by 35°C heat but sensitive to 40°C heat (Fig. 8 h, Supplementary Fig. 15), suggesting mitochondrial activity could be another target to improve thermotolerance. Additionally, the genes associated with TM1 with early heat induction may include pathways that are essential for heat tolerance (Fig. 3a, Supplementary Data 5). Finally, we compared our algal heat transcriptome with that in Arabidopsis (heat at 42°C for 7 h)[43] and identified a set of highly conserved heat-induced genes sets (Supplementary Data 1), which may provide potential targets to improve heat tolerance in land plants.

In summary, Chlamydomonas is an excellent model to study the heat response and its regulation at the cellular level in photosynthetic cells. Our research helped fill the knowledge gaps regarding how algae respond to and recover from different intensities of high temperatures at multiple levels, discovered the increased transcript/protein correlation during heat treatments, showed the dynamics of photosynthesis in response to high temperatures, and revealed the antagonistic interaction between DNA replication and photosynthetic light reactions during the recovery from both moderate and acute high temperatures. Through systems-wide analyses, we advanced our understanding of algal heat responses and identified engineering targets to improve thermotolerance in green algae and land plants.

Methods

Strains and culture conditions. *Chlamydomonas reinhardtii* wildtype strain CC-1690 (also called *21gr*, *mt*$^+$, from the Chlamydomonas resource center)[96–99] was used in all experiments. CC-1690 were grown in standard Tris-acetate-phosphate (TAP) medium in 400 mL photobioreactors (PBRs) (Photon System Instruments, FMT 150/400-RB). Cultures were illuminated with constant 100 µmol photons m^2 s^{-1} light (50% red: 50% blue), mixed by bubbling with filtered air at a flow rate of 1 L/min. After PBR inoculation at initial cell density of 0.5×10^6 cells/mL, cultures were allowed to grow to a target cell density of 2.00×10^6 cells/mL corresponding to around 4.0 µg/mL chlorophyll content in log-phase growth at 25°C. Then, the target cell density was maintained turbidostatically using OD$_{680}$ by allowing the culture to grow to 8% above the target cell density before being diluted to 8% below the target cell density with fresh TAP medium provided through peristaltic pumps. Through the turbidostatic mode, the PBR cultures had exponential growth between dilution events. The OD$_{680}$ measurement during exponential growth phases in between dilution events was log$_2$ transformed, and the relative growth rate was calculated using the slope of log$_2$(OD$_{680}$), while the inverse of the slope yielded the doubling time of the culture (Supplementary Fig. 1a, b). All algal liquid cultivation used in this paper was conducted in PBRs with the conditions mentioned above.

High-temperature treatments in PBRs. Algal cultures in PBRs were maintained turbidostatically using OD$_{680}$ for 4 days at 25°C to allow cultures to adapt to steady growth conditions before heat treatments (Fig. 1b). PBR temperatures were then shifted to moderate or acute high temperature conditions (35°C or 40°C in different PBRs) for 24 h, then shifted back to 25°C for 48 h for recovery. PBR cultures grown under constant 25°C served as controls. Cultures were maintained turbidostatically during the entire experiment and harvested at different time points for various measurements. Cell density and mean cell diameter were measured using a Coulter Counter (Multisizer 3, Beckman Counter, Brea, CA). For the data in Fig. 1a, algal cultures were maintained in PBRs at 25°C for 4 days before switching to 30°C, 35°C, or 40°C for 2 days (different temperature switches in separate PBRs). The relative growth rates were calculated at the end of 2-day treatment of each temperature.

Spotting test for cell viability and growth. Cultures harvested from PBRs were diluted to 2×10^4 cells mL^{-1} or 1×10^5 cells mL^{-1} with TAP medium and 10 µL aliquots of the diluted cultures were spotted on 1.5% TAP agar plates and grown in temperature-controlled incubators under 25°C or 35°C with constant white LED light of 150 µmol photons m^{-2} s^{-1} for 44 h or 3 days. After 44-h growth, algal spots with 200 cells were imaged by a dissecting Leica microscopy and were used for growth quantification. Colony number and area were quantified using ImageJ. Viability was calculated as the number of colonies on plates divided by the number of cells spotted. Algal spots with 200 and 1000 cells were imaged after 3-day-growth for visual representations.

High temperature treatments in water bath. To measure the effects of heating speed on cell viability (Supplementary Fig. 1f), control PBR cultures without heat treatments were incubated in a water bath. Gradual heat treatment from 25°C to 41°C took place over 25 min, then cultures were kept at 41°C for 2 h. Directly heated samples were incubated in a water bath which was pre-heated to 41°C then kept at 41°C for 2 h (sharp temperature switch). Cell viabilities after 2-h 41°C heat treatment (either gradual or sharp heating) were assayed using the spotting test as above. Because PBRs cannot switch from the control to high temperatures in less 25 min, a water bath was used for this test, as previously reported[13,18,33].

RNA extraction and RT-qPCR. At each time point, 2 mL PBR cultures were pelleted with Tween-20 (0.005%, v/v) by centrifugation at 1,100 x g and 4°C for 2 min. The cell pellet was flash frozen in liquid nitrogen and stored at −80°C before processing. Total RNA was extracted with the TRIzol reagent (Thermo Fisher Scientific, Cat No. 15596026) as described before with some modifications[53]. RNA was purified by RNeasy mini-column (Qiagen, Cat No. 74106) after on column digestion with RNase-free DNase (Qiagen, Cat No. 79256) according to the manufacturer's instructions. RNA was quantified with Qubit™ RNA BR Assay Kit, (Life technology, Cat No. Q10210). Total 0.4 µg RNA was reverse transcribed with oligo dT primers using SuperScript® III First-Strand Synthesis System (Life technology, Cat No. 18080-051) according to the manufacturer's instructions. Quantitative real-time PCR (RT-qPCR) analysis was carried out using a CFX384 Real-Time System (C 1000 Touch Thermal Cycler, Bio-Rad, Hercules, California) using SensiFAST SYBR No-ROS kit (Bioline, BIO-98020). PCR was set up as follows: (1) 2 min at 95°C; (2) 40 cycles of 5 s at 95°C, 10 s at 60°C and 15 s at 72°C; (3) final melt curve at 60°C for 60 s, followed by continuous ramping of temperature to 99°C at a rate of 0.5°C s^{-1}. Melting curves and qPCR products was checked after PCR cycles to ensure there are no primer dimers or unspecific PCR products. All qPCR products were sequenced to verify their identifies. Expression of G-protein β-subunit-like polypeptide CBLP (Cre06.g278222) (Schloss, 1990) remain stable among all time points, and were used as internal controls[100]. The relative gene expressions were calculated relative to the gene's own expression in pre-heat by using the $2^{-\Delta\Delta CT}$ method as described previously[101–103]. Three biological replicates for each time point and treatment were conducted. The qPCR primers used are listed in Supplementary Table 2.

Transcriptomics. RNA libraries were prepared and sequenced by the Joint Genome Institute (JGI, Community Science Program) using the NovaSeq platform and generated 150-nt paired-end reads, with the goal of 20 million genome-mappable reads per sample. Samples were quality control filtered using the JGI BBDuk and BBMap pipelines[104]. BBDuk trims adapter sequences, and quality trims reads with low complexity, low quality scores, and reads with 1 or more "N" bases. BBMap removes reads that map to common contaminant genomes. Samples were quality assessed using FastQC[105] and mapped to the *Chlamydomonas reinhardtii* v5.6 genome[106] using STAR[107,108] with the maximum number of mapped loci set to 1 and the maximum mismatches per read set to 1, resulting in >92% of all reads being uniquely mapped to the *Chlamydomonas reinhardtii* v5.6 genome. Reads per feature were counted via HT-seq count[109]. The dataset was filtered for genes that met minimum read count cutoffs of at least 10 mapped reads in at least 10% of the samples, resulting in 15,541 genes for downstream analyses. Two-dimensional Uniform Manifold Approximation and Projection (UMAP) was used to reveal clusters of RNA-seq data[110].

Differential expression modeling was performed on transcripts per million (TPM) normalized read counts using a generalized linear mixed-effect model and a negative binomial distribution. Treatment time points were compared to pre-heat controls. Significant differential expression was defined as $|\log_2$ fold-change$| > 1$, FDR < 0.05, and | (control mean TPM)—(treatment mean TPM) | ≥ 1. The sign of | means absolute values. FDR correction was performed using the Benjamini-Hochberg method[111]. Heatmaps were generated using the R package pheatmap (version 1.0.12. https://CRAN.R-project.org/package=pheatmap). Weighted correlation network analysis (WGCNA) was performed on TPM normalized read counts that met minimum read count cutoffs[112]. Expression patterns from 35°C and 40°C were modeled together as a signed network, requiring at least 50 genes per module and combining modules with similarity >0.25. All genes were tested against all eigengene modules using ANOVA (FDR < 0.05) and were assigned to the module with highest significance. Functional enrichment analysis was performed using hypergeometric tests with subsequent FDR control based on MapMan annotations for Chlamydomonas reinhardtii (FDR < 0.05)[111]. Surprisal analysis was applied to dissect the transcriptional regulatory response into its major components[113]. Constraint potentials were calculated for each transcriptional data matrix, whose signals determine the deviations of the transcriptome from a theoretical balance state of minimal free energy[114]. The overall importance of the transcriptional pattern was estimated from its amplitude and decreases with increasing constraint index. Accordingly, the three major constrains were kept.

Proteomics. Algal samples (2 mL) harvested from PBRs were centrifuged to remove supernatant, flash frozen in liquid nitrogen, and stored at −80°C until use. Proteins were extracted using the IST sample preparation kit (PreOmics GmbH, Germany). Cell pellets (about 200 µg protein) were lysed in 50 µL lysis buffer, and 20% of the lysate was added into 40 µL lysis buffer and sheared in a sonicator (VWR Aquasonic 250D, 35Khz) for 3 min, then digested in a heating block at 37°C with gentle shaking for 3 h followed by stopping the digestion with the stop buffer. Peptides were purified using the PreOmics cartridges according to the manufacturer's instruction. Eluted peptides were transferred to 96 well plates, dried under vacuum for 2 h, and then dissolved in 30 µL of LC-loading buffer included in the kit. Finally, 5 µL of the suspension was used for LC-MS/MS analysis.

LC-MS/MS (Liquid chromatography–mass spectrometry) was carried out on an Orbitrap Fusion Lumos (Thermo Fisher Scientific, San Jose, CA) mass spectrometer coupled with a U3000 RSLCnano HPLC (Thermo Fisher Scientific, San Jose, CA). The peptide separation was carried out on a C18 column (Fritted Glass Column, 25 cm × 75 µm, Reprosil-Pur 120 C18-AQ, 1.9 µm, made by ESI Source Solution, LLC., Woburn, MA) at a flow rate of 0.3 µL/min and the following gradient: Time = 0–4 min, 2% B isocratic; 4–8 min, 2–10% B; 8–83 min, 10–25% B; 83–97 min, 25–50% B; 97–105 min, 50–98%. Mobile phase consisted of A, 0.1% formic acid; mobile phase B, 0.1% formic acid in acetonitrile. The instrument was operated in the data-dependent acquisition mode in which each MS1 scan was followed by Higher-energy collisional dissociation (HCD) of as many precursor ions in 2 second cycle (Top Speed method). The mass range for the MS1 done using the FTMS was 365 to 1800 m/z with resolving power set to 60,000 @ 200 m/z and the automatic gain control (AGC) target set to 1,000,000 ions with a maximum fill time of 100 ms. The selected precursors were fragmented in the ion trap using an isolation window of 1.5 m/z, an AGC target value of 10,000 ions, a maximum fill time of 100 ms, a normalized collision energy of 35 and activation time of 30 ms. Dynamic exclusion was performed with a repeat count of 1, exclusion duration of 30 s, and a minimum MS ion count for triggering MS/MS set to 5000 counts.

Protein identification and quantification. Quantitative analysis of MS measurements was performed using MaxQuant 1.6.12.0[115]. The library used to perform peptide spectrum matching was created based on version JGI5.5 of the Chlamydomonas reinhardtii genome. The search space was extended including methionine oxidation and acetylation of protein N-termini as variable modifications. The false discovery rate (FDR) thresholds for peptide spectrum matching and protein identification were set to 1%. Protein abundances were estimated using the Label free Quantification (LFQ) algorithm[116].

Data normalization and protein-level missing value imputation. Normalization and missing value imputation were performed independently for 35°C and 40°C time courses. Proteins were excluded from the data set if there was no biological replicate group with more than one value among different time points, as these proteins were considered not suitable for quantitative downstream analysis. Following normalization using the median-of-ratios method[117], we used global variance estimates and local gene wise mean estimates to impute missing data points as independent draws from normal distributions. If a protein showed no quantification in a group of biological replicates at one-time point, it was checked if the protein was present in the adjacent time points (the time point before and after the one in query). If this was the case, the protein mean was imputed using k-Nearest-Neighbour imputation, followed by sampling from a normal distribution. If adjacent time points had no values, no imputation was performed for the time point in the query. All further protein analyses were based on imputed values.

Network generation. All protein groups resulting from non-proteotypic peptides were duplicated to singletons and the intensities of each protein were log-transformed. To ignore proteins with constant abundance signals, they were filtered for significance by one-way ANOVA ($p < 0.05$). To generate a correlation matrix, the Pearson correlation coefficient was used. An absolute correlation threshold was determined by random matrix theory[118]. These thresholds were determined to be $\rho = 0.8194$ and $\rho = 0.8675$ for 35°C and 40°C, respectively. After filtering the correlation matrix accordingly, the nodes and edges were isolated and visualized in Gephi (www.gephi.org) using ForceAtlas2[119] and built-in modularity determination[120].

Statistical testing of proteins. Proteins that were identified by ambiguous (non-proteotypic) peptides or had missing values in at least one replicate were not considered for statistical testing. All time points were tested for significant accumulation or depletion in respect to a control measured prior to the start of the heat treatment. Dunnett's multiple-comparison test was applied with alpha levels of 0.05 and 0.01.

Proteomics enrichment analyses. To investigate the module compositions a term enrichment based on MapMan ontology was performed. The ontology tree of each term was expanded, so every protein could exist multiple times, corresponding to the number of ontology term levels. A hypergeometric test for each functional term was applied with subsequent FDR control by the Benjamini-Hochberg method[111]. To derive a representative signal shape of all proteins included in a module, its eigenvector was calculated based on complete time series, which were log-transformed and centered to an intensity of zero mean and unit variance[121]. For negatively correlating proteins, its corresponding eigenvector was inverted. Additional term enrichments were applied following the previous schema to gain insights into functions of positively or negatively correlating proteins of each module.

Correlation of transcripts and proteins. The $\log_2$(fold-change) of transcript reads and protein abundance were calculated in respect to the pre-heat samples. All available transcript/protein pairs were taken into consideration for correlation analysis. The heat period (HS) as well as the recovery period (RE) were split up into three windows each (HS: 0–1 h, 2–8 h, 16–24 h during the heat period; RE: 0–2 h, 4–8 h, 24–48 h during the recovery period after heat treatment). The average $\log_2$(fold-change) is determined for each window. Every identifier, that had both a transcript and protein associated with it, resulted in a transcript-protein fold-change pair for each window. By collecting all identifiers that are associated with a respective functional term a scatter plot of transcript-protein fold-change pairs were generated, and the Pearson correlation coefficient was calculated. By collecting the correlation coefficient for every present functional term, a density plot for each of the six windows was created. The resulting correlation coefficient histograms were smoothed using Silvermans rule of thumb for kernel density estimation[122]. See Supplementary Fig. 7 for the illustration.

All computational analyses on protein intensities were conducted using the open-source F# libraries FSharp.Stats, BioFSharp, and Plotly.NET. Linear regression, Benjamini–Hochberg correction, smoothing, correlation measures, and eigenvector calculations were performed using the FSharp.Stats version 0.4.1-beta. For ontology annotation and GSEA based on hypergeometric tests, we used BioFSharp version 2.0.0-beta4. Visualization of transcript-protein correlation was performed using Plotly.NET version 2.0.0-alpha5.

DNA content and ploidy. DNA content was analyzed by FACS (fluorescence-activated cell sorting) with modified protocol[123]. Cell cultures (10 mL) were collected and fixed in 30 mL ethanol:acetic acid (3:1) for 15 min at room temperature. Cells were spun down at room temperature at 4000 x g for 1 min and washed once with 1 mL phosphate-buffered saline (PBS). Then cells were collected by centrifugation, resuspended in 2 mL PBS with RNase A (100 µg/mL) for 2 h at 37°C, centrifuged again and finally resuspended in 2 mL PBS + 500 nM Sytox Green (Thermo Fisher Scientific, S7020). FACS was performed on an Accuri C6 instrument (BD Biosciences, Franklin Lakes, NJ), reading 20,000 cells per sample in the FL1 channel (488 nm exciting laser; emission filter: 530 ± 15 nm). A 90% attenuator was used to reduce signal below saturation levels. Data were analyzed using FlowJo software (BD Biosciences). Assignment of cell populations representing 1 C, or >1 C DNA content was determined. The raw fluorescence signal of 2 C was not two-times the signal at 1 C because there was background staining, which was cell-size-dependent[123]. Therefore, high-ploidy cells (which did not get much bigger during a multiple fission cycle) had DNA signal that was about proportional to actual DNA contents, while in low-ploidy cells the background contributed more. Thus, in the population of pre-heat samples, background is about 0.5 $\times 10^5$ of the DNA content fluorescence signal (x axis). Before background subtraction, 1 C, 2 C, 4 C cells peaked at 3 $\times 10^5$, 5.5 $\times 10^5$, 10 $\times 10^5$ of DNA content fluorescence signal, respectively. After background subtraction, 1 C, 2 C, 4 C cells were at 2.5 $\times 10^5$, 5 $\times 10^5$, 9.5 $\times 10^5$, respectively. There was also a small peak at 20 $\times 10^5$ which was

8 C. At 24 h heat of 40°C, cell size was about 2x bigger (Fig. 5a, b, c), with background of around 2 ×10^5, causing the right shift of the 1 C peak (Fig. 4k).

Cell imaging using light microscopy. Cultures harvested at select time points were fixed with 0.2% glutaraldehyde (VWR, Cat No. 76177-346). Cells were imaged with a Leica DMI6000 B microscope and a 63x (NA1.4) oil-immersion objective. Images shown are representative of results from at least three independent experiments (Fig. 5a).

Pigment analysis. Three biological replicates (each with two technical replicates) of 1 mL of PBR cultures were harvested at different time points, mixed with 2.5 µL 2% Tween20 (Sigma, P9416-100ML) to help cell pelleting, centrifuged at 18,407 g at 4°C, and stored in −80°C after removal of supernatant. Cell pellets were later thawed, resuspended in 1 mL of HPLC grade methanol (100%, Sigma, 34860-4L-R), vortexed for 1 min, incubated in the dark for 5 min at 4°C, and centrifuged at 15,000 g at 4°C for 5 min. Supernatant containing pigments was analyzed at 470, 652, 665 nm in a spectrophotometer (IMPLEN Nonophotometer P300) for carotenoids and chlorophyll a/b concentrations in µg mL^{-1} using the following equations: Chl a + Chl b = 22.12*A_{652} + 2.71*A_{665}, Chl a = 16.29*A665 – 8.54*A652, and Chl b = 30.66*A652 – 13.58*A665[124], and carotenoids = (1000* A470 – 2.86*Chl a – 129.2*Chl b)/221[125].

Protein concentration and ROS determination. Frozen sample pellets were thawed on ice and lysed by sonication in 10 mM Tris-HCl buffer[126]. Protein concentrations were determined by the method of Lowry (Lowry et al. 1951). The level of ROS was determined by the method described before with some modifications[127]. A 40 µg total protein extract was used for ROS measurement. Each sample was aliquoted and one of them was added with ascorbate (Thomas Scientific LLC, C988F55) to a final concentration of 100 mM. The samples containing ascorbate were used as background signal and subtracted from each experimental value later. The samples were incubated for 15 min at 25°C. The ROS indicator CM-H2DCFDA (Life Technologies, C6827) dissolved in 20% (v/v) DMSO was then added to a final concentration of 10 µM and incubated for 30 min at 30°C. The samples were transferred to 96 well microplates and ROS-related fluorescence was measured using Tecan Microplate reader M200 PRO with excitation at 485 nm and emission at 525 nm. The results were obtained from three biological replicates. Relative fold-change of ROS signal (compared to pre-heat) was either normalized to cell number or cell volume. Each of the three biological replicates included two independent measurements with two technical replicates.

Spectroscopic measurement of photosynthetic parameters. Photosynthetic measurements (chlorophyll fluorescence, electrochromic shift, and P700) were conducted using a multi-wavelength kinetic spectrophotometer/fluorometer with a stirring enabled cuvette holder (standard 1 cm pathlength) designed and assembled by the laboratory of Dr. David Kramer at Michigan State University using the method described before with some modifications[75]. A 2.5 mL volume (around 12~13 µg chlorophyll) of algal cells were sampled from the photobioreactors, supplemented with 25 µL of fresh 0.5 M $NaHCO_3$, loaded into a fluorometry cuvette (C0918, Sigma-Aldrich), and dark-adapted with a 10-min exposure to far-red light with peak emission of 730 nm at ~35 µmol photons m^{-2} s^{-1}. All photosynthetic measurements were performed at the room temperature at 25°C for algal cultures sampled at different time points with different temperature treatments to investigate how the changes induced by high temperatures affected algal photosynthetic performance by comparing with control algal cultures treated and measured at 25°C. The maximum efficiency of PSII (F_v/F_m) was measured with the application of a saturating pulse of actinic light with peak emission of 625 nm at the end of the dark adaptation period and after turning off the far-red light. Our tests indicated that far-red illumination could increase the values of F_v/F_m in dark-adapted algal cells. Far-red light was turned off during all chlorophyll fluorescence measurements to prevent its effect on chlorophyll fluorescence signals. Fluorescence measurements were taken with measuring pulses of 100 µs duration. The pulsed measuring beam was provided by a 505 nm peak emission (light emitting diode) LED filtered through a BG18 (Edmund Optics) color glass filter. After dark-adaptation, the algal sample was illuminated by a pair of LEDs (Luxeon III LXHL-PD09, Philips) with maximal emission at 620 nm, directed toward both sides of the cuvette, perpendicular to the measuring beam. Subsequent chlorophyll fluorescence and dark interval relaxation kinetic (DIRK) measurements were taken after 7.5-min adaptation of sequentially increasing light intensities of 100, 200, and 400 µmol photons m^{-2} s^{-1}. DIRK traces measure the electrochromic shift (ECS). ECS results from light-dark-transition induced electric field effects on carotenoid absorbance bands[51,128] and is a useful tool to monitor proton fluxes and the transthylakoid proton motive force (*pmf*) in vivo[129,130]. ECS measurements were taken at each light intensity following the chlorophyll fluorescence measurements. DIRK traces were run with measuring beam of peak 520 nm and pulse duration of 14 µs capturing absorption changes from a light-dark-light cycle with each phase lasting 300 ms. Three near-simultaneously DIRK traces were averaged for one measurement.

P700 measurement. Algal cultures were sampled directly from PBRs at different time points with different treatments, supplemented with 0.5 M $NaHCO_3$, incubated with 10 µM DCMU (3-(3,4-Dichlorophenyl)−1,1-dimethylurea, Sigma, D2425) to block PSII activity, and dark adapted for 5 min before the measurement. PSI redox kinetics were monitored using a 703 nm measuring beam pulsed at 10-ms intervals using the spectrophotometer mentioned above as described previously with some modifications[75]. After a 5 s baseline measurement in the dark, samples were exposed to 720 nm far-red light at an intensity of 30 µmol photons m^{-2} s^{-1} for 5 s to preferentially oxidize PSI before monitoring the reduction of oxidized P700 ($P700^+$) in dark for 5 s. Addition of saturating flash at the end of far-red illumination did not change the amplitude of $P700^+$ and the reduction time constant of $P700^+$ significantly so the saturating flash was skipped in the finalized P700 measurement. Five measurements were averaged into one trace, and the reduction time constant of PSI was calculated by fitting to a first order exponential function to the reduction phase of the averaged trace. The reduction time constant of $P700^+$ in the absence of PSII activity (with DCMU) can be used to estimate the activity of cyclic electron flow around PSI (CEF)[74].

77 K chlorophyll fluorescence measurement. Chlorophyll fluorescence emission spectra were monitored at 77 K to estimate antenna sizes[76,131]. Sampled algal cultures from PBRs were immediately loaded into NMR tubes (VWR, cat. No. 16004-860) and frozen in liquid nitrogen. While still submerged in liquid nitrogen, frozen samples were exposed to excitation LED light with peak emission at 430 nm through a bifurcated fiber optic cable coupled to an LED light source[75]. Components were held in alignment using a 3D printed device. Chlorophyll emission spectra were recorded on an Ocean Optics spectrometer (cat. No. OCEAN-HDX-XR), and three consecutive traces were averaged into one measurement. Further dilution of the PBR samples gave identical fluorescence emission peak distributions, indicating little distortion of the signals by reabsorption. Spectral data were normalized to the PSII spectral maximum value at 686 nm, and the relative PSII antenna size (PSII%) was calculated using the normalized PSII peak and PSI peak (714 nm) using the formula PSII% = normalized PSII peak / (normalized PSII peak + normalized PSI peak).

See Supplementary Table 3 for formulas to calculate photosynthetic parameters. All measurements were taken with at least three biological replicates. Statistical significance was assessed using two-tailed t-test assuming unequal variance. For more than 20 comparisons, FDR using Benjamin-Hochberg correction method was performed with adjusted p value < 0.05 as significance.

Oxygen evolution measurements. Oxygen evolution was measured with Hansatech Chlorolab 2 based on a Clark-type oxygen electrode at room temperature at 25°C, following the method described before with some modifications[132]. Two-mL of cells (around 10 µg chlorophyll) supplemented with 20 µL of 0.5 M $NaHCO_3$ were incubated in the dark for 10 min. The dark respiration rate was measured at the end of the dark incubation. The rate of oxygen evolution was measured at increasing light intensities (100, 200 and 400 µmol photons m^{-2} s^{-1}). Each light intensity lasted 5 min followed by 2 min dark. The rates of net oxygen evolution at each light intensity step were recorded for 1 min before the end of each light phase. Respiration rates were measured during each subsequent dark phase. For each time point, the respiration rates did not vary significantly after 10-min dark-adaptation or after different light intensities so only the respiration rates after 10-min dark-adaptation were plotted in Fig. 8h. The gross oxygen evolution rates for a given light intensity were calculated as the sum of the net oxygen evolution rate and the respiration rates following each light phase. The results were obtained from independent measurements of five biological replicates. Statistical analyses were performed using two-tailed t-test assuming unequal variance with FDR correction by comparing treated samples with the pre-heat samples under the same light.

Transmission Electron Microscopy (TEM). Algal samples (10 mL, around 2×10^6 cells/mL) harvested from PBRs were concentrated (1000 g, 2 min, room temperature), drawn into dialysis tubing (Spectrapor®, Spectrum Laboratories, Inc., Cat No. 132294), immersed in 20% (w/v) Bovine Serum Albumin (BSA. Sigma, Cat No. A7030-100G) and immediately frozen in a high-pressure freezer (Leica EM ICE; Leica Microsystems, IL, USA). Freeze substitution was carried out in 2% (v/v) osmium tetroxide in 97%/5% acetone/water (v/v) as follows: −80°C for 3 days, −20°C for 1 day, 4°C for 2 h, and then room temp for 2 h. Samples were washed 4 times in acetone, incubated in 1% thiocarbohydrazide (TCH) (EM Sciences, 21900) in acetone for 1 h, washed 4 times in acetone, incubated for 1 h in 2% (v/v) osmium tetroxide and washed 4 times in acetone. Samples were then laced in saturated uranyl acetate in 100% acetone overnight, washed in acetone, transferred to 1:1 ethanol:acetone and stained with saturated lead acetate in 1:1 ethanol: acetone for 2 h. Subsequently, cells were then dehydrated in 100% graded acetone and embedded in hard formulation Epon-Araldite (Embed 812, EM Sciences, Cat No. 14121). Embedments were cut into small blocks and mounted in the vise-chuck of a Reichert Ultracut UCT ultramicrotome (Leica, Buffalo Grove, IL, USA). Ultrathin sections (~60 to 70 nm) were cut using a diamond knife (type ultra 35°C, Diatome), mounted on copper grids (FCFT300-CU-50, VWR, Radnor, PA, USA), and counterstained with lead citrate for 8 min[133]. Sample were imaged with a LEO 912 AB Energy Filter Transmission Electron Microscope (Zeiss, Oberkochen, Germany). Micrographs were acquired with iTEM software (ver. 5.2) (Olympus Soft Imaging Solutions GmbH, Germany) with a TRS 2048 x 2048k slow-scan charge-coupled device (CCD) camera (TRÖNDLE Restlichtverstärkersysteme, Germany). Thirty electron micrographs were quantified for each time point and treatment. Each TEM image was acquired at 8,000X magnification and 1.37 nm pixel resolution. TEM images were analyzed with Stereology Analyzer software version 4.3.3 to quantify the relative volume mitochondria. Grid type was set as "point" with a sampling step of 500×500 pixels and pattern size of 15×15 pixels. The percent of relative volume for mitochondria was collected after identifying all grid points within one cell and further analyzed in excel[19]. TEM images with a magnification of 8 K were used in the Fiji (ImageJ, FIJI software, National Institutes of Health) analysis. The images were scaled to 0.7299 pixel/nm in ImageJ before the analysis of the pyrenoid area. Two different statistical tests were used to find the significance of p-values. The Kolmogorov-Smirnov test was used for relative volume data since it is commonly used to find significance between data in a form of ratios. A two-tailed t-test with unequal variance was used for area data from ImageJ. All statistical tests compared the treatment conditions to the pre-heat condition.

Starch quantification. Starch was extracted and quantified according to modified method from kit (Megazyme, K-TSTA-100A). Frozen sample pellets were homogenized using Qiagen TissueLyser and washed twice in 80% (v/v) ethanol at 85°C. The insoluble fraction was resuspended in DMSO and heated at 110°C for 10 min to improve starch solubilization. Starch hydrolysis and quantification were performed following kit protocol. Starch content was either normalized to cell number or cell volume. Each condition has three biological replicates.

Statistics and Reproducibility. All measurements had at least 3 biological replicates. Statistical analyses were conducted as mentioned in each method above. For more than 17 comparisons, FDR using Benjamin-Hochberg correction method was performed with adjusted *p* value < 0.05 as significance. Source data and *p* values were included in Supplementary Data 11.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The datasets analyzed in this paper are included in this published article and supplementary information files. The RNA-seq data discussed in this publication have been deposited in the NCBI Gene Expression Omnibus[134] and are accessible through GEO accession number GSE182207. The mass spectrometric proteomic data is available via the ProteomeXchange Consortium partner repository, PRIDE[135,136] with the dataset identifier PXD027778. Other information is available from the corresponding author on request.

Received: 1 October 2021; Accepted: 12 April 2022;
Published online: 13 May 2022

References

1. Zhao, C. et al. Temperature increase reduces global yields of major crops in four independent estimates. *PNAS* **114**, 9326–9331 (2017).
2. Siebert, S., Ewert, F., Rezaei, E. E., Kage, H. & Graß, R. Impact of heat stress on crop yield—on the importance of considering canopy temperature. *Environ. Res. Lett.* **9**, 044012 (2014).
3. Lobell, D. B., Schlenker, W. & Costa-Roberts, J. Climate trends and global crop production since 1980. *Science* **333**, 616–620 (2011).
4. Janni, M. et al. Molecular and genetic bases of heat stress responses in crop plants and breeding for increased resilience and productivity. *J. Exp. Bot.* **71**, 3780–3802 (2020).
5. Sharkey, T. D. Effects of moderate heat stress on photosynthesis: importance of thylakoid reactions, rubisco deactivation, reactive oxygen species, and thermotolerance provided by isoprene. *Plant, Cell Environ.* **28**, 269–277 (2005).
6. Mittler, R., Finka, A. & Goloubinoff, P. How do plants feel the heat? *Trends Biochem Sci.* **37**, 118–125 (2012).
7. Saidi, Y. et al. The heat shock response in moss plants is regulated by specific calcium-permeable channels in the plasma membrane. *Plant Cell* **21**, 2829–2843 (2009).
8. Wu, H.-C., Luo, D.-L., Vignols, F. & Jinn, T.-L. Heat shock-induced biphasic Ca2+ signature and OsCaM1-1 nuclear localization mediate downstream signalling in acquisition of thermotolerance in rice (Oryza sativa L.). *Plant, Cell Environ.* **35**, 1543–1557 (2012).
9. Königshofer, H., Tromballa, H.-W. & Löppert, H.-G. Early events in signaling high-temperature stress in tobacco BY2 cells involve alterations in membrane fluidity and enhanced hydrogen peroxide production. *Plant, Cell Environ.* **31**, 1771–1780 (2008).
10. Schroda, M., Hemme, D. & Mühlhaus, T. The Chlamydomonas heat stress response. *Plant J.* **82**, 466–480 (2015).
11. Schulz-Raffelt, M., Lodha, M. & Schroda, M. Heat shock factor 1 is a key regulator of the stress response in *Chlamydomonas*. *Plant J.* **52**, 286–295 (2007).
12. Schmollinger, S., Strenkert, D. & Schroda, M. An inducible artificial microRNA system for *Chlamydomonas reinhardtii* confirms a key role for heat shock factor 1 in regulating thermotolerance. *Curr. Genet* **56**, 383–389 (2010).
13. Rütgers, M. et al. Not changes in membrane fluidity but proteotoxic stress triggers heat shock protein expression in *Chlamydomonas reinhardtii*. *Plant Cell Environ.* **40**, 2987–3001 (2017).
14. Su, Z. et al. Genome-wide RNA structurome reprogramming by acute heat shock globally regulates mRNA abundance. *PNAS* **115**, 12170–12175 (2018).
15. Zhang, R. & Sharkey, T. D. Photosynthetic electron transport and proton flux under moderate heat stress. *Photosynth Res* **100**, 29–43 (2009).
16. Sharkey, T. D. & Zhang, R. High temperature effects on electron and proton circuits of photosynthesis. *J. Integr. Plant Biol.* **52**, 712–722 (2010).
17. Song, Y., Chen, Q., Ci, D., Shao, X. & Zhang, D. Effects of high temperature on photosynthesis and related gene expression in poplar. *BMC Plant Biol.* **14**, 111 (2014).
18. Hemme, D. et al. Systems-wide analysis of acclimation responses to long-term heat stress and recovery in the photosynthetic model organism *Chlamydomonas reinhardtii*. *Plant Cell* **26**, 4270–4297 (2014).
19. Anderson, C. M. et al. High light and temperature reduce photosynthetic efficiency through different mechanisms in the C_4 model Setaria viridis. *Commun. Biol.* **4**, 1092 (2021).
20. Pospíšil, P. Production of reactive oxygen species by photosystem II as a response to light and temperature stress. *Front. Plant Sci.* **7**, 1950 (2016).
21. Niemeyer, J., Scheuring, D., Oestreicher, J., Morgan, B. & Schroda, M. Real-time monitoring of subcellular H2O2 distribution in *Chlamydomonas reinhardtii*. *The Plant Cell.* **33**, 2935–2949 (2021).
22. Velichko, A. K., Petrova, N. V., Kantidze, O. L. & Razin, S. V. Dual effect of heat shock on DNA replication and genome integrity. *MBoC* **23**, 3450–3460 (2012).
23. Mata, T. M., Martins, A. A. & Caetano Nidia. S. Microalgae for biodiesel production and other applications: A review. *Renew. Sustain. Energy Rev.* **14**, 217–232 (2010).
24. Mühlhaus, T., Weiss, J., Hemme, D., Sommer, F. & Schroda, M. Quantitative shotgun proteomics using a uniform ^{15}N-labeled standard to monitor proteome dynamics in time course experiments reveals new insights into the heat stress response of *Chlamydomonas reinhardtii*. *Mol. Cell Proteom.* **10**, M110.004739 (2011).
25. Zachleder, V., Ivanov, I., Vítová, M. & Bišová, K. Cell cycle arrest by supraoptimal temperature in the alga *Chlamydomonas reinhardtii*. *Cells* **8**, 1237 (2019).
26. Ivanov, I. N., Zachleder, V., Vítová, M., Barbosa, M. J. & Bišová, K. Starch production in *Chlamydomonas reinhardtii* through supraoptimal temperature in a pilot-scale photobioreactor. *Cells* **10**, 1084 (2021).
27. Jinkerson, R. E. & Jonikas, M. C. Molecular techniques to interrogate and edit the *Chlamydomonas* nuclear genome. *Plant J.* **82**, 393–412 (2015).
28. Crozet, P. et al. Birth of a photosynthetic chassis: A MoClo toolkit enabling synthetic biology in the microalga *Chlamydomonas reinhardtii*. *ACS Synth. Biol.* **7**, 2074–2086 (2018).
29. Zhang, R. et al. High-throughput genotyping of green algal mutants reveals random distribution of mutagenic insertion sites and endonucleolytic cleavage of transforming DNA. *Plant Cell* **26**, 1398–1409 (2014).
30. Li, X. et al. An indexed, mapped mutant library enables reverse genetics studies of biological processes in *Chlamydomonas reinhardtii*. *Plant Cell* **28**, 367–387 (2016).
31. Li, X. et al. A genome-wide algal mutant library and functional screen identifies genes required for eukaryotic photosynthesis. *Nat. Genet* **51**, 627–635 (2019).
32. Dhokane, D., Bhadra, B. & Dasgupta, S. CRISPR based targeted genome editing of *Chlamydomonas reinhardtii* using programmed Cas9-gRNA ribonucleoprotein. *Mol. Biol. Rep.* **47**, 8747–8755 (2020).
33. Légeret, B. et al. Lipidomic and transcriptomic analyses of *Chlamydomonas reinhardtii* under heat stress unveil a direct route for the conversion of membrane lipids into storage lipids. *Plant, Cell Environ.* **39**, 834–847 (2016).

34. Yang, Y. et al. Applications of multi-omics technologies for crop improvement. *Front Plant Sci.* **12**, 563953 (2021).
35. Hwang, S. & Herrin, D. L. Control of lhc gene transcription by the circadian clock in *Chlamydomonas reinhardtii*. *Plant Mol. Biol.* **26**, 557–569 (1994).
36. Barajas-López, J. et al. Circadian regulation of chloroplastic f and m thioredoxins through control of the CCA1 transcription factor. *J. Exp. Bot.* **62**, 2039–2051 (2011).
37. Remm, M., Storm, C. E. V. & Sonnhammer, E. L. L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314**, 1041–1052 (2001).
38. Goodstein, D. et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–D1186 (2012).
39. Plancke, C. et al. Lack of isocitrate lyase in *Chlamydomonas* leads to changes in carbon metabolism and in the response to oxidative stress under mixotrophic growth. *Plant J.* **77**, 404–417 (2014).
40. Johnson, X. & Alric, J. Interaction between starch breakdown, acetate assimilation, and photosynthetic cyclic electron flow in *Chlamydomonas reinhardtii*. *J. Biol. Chem.* **287**, 26445–26452 (2012).
41. Johnson, X. & Alric, J. Central carbon metabolism and electron transport in *Chlamydomonas reinhardtii*: metabolic constraints for carbon partitioning between oil and starch. *Eukaryot. Cell* **12**, 776–793 (2013).
42. Durante, L., Hübner, W., Lauersen, K. J. & Remacle, C. Characterization of the GPR1/FUN34/YaaH protein family in the green microalga Chlamydomonas suggests their role as intracellular membrane acetate channels. *Plant Direct* **3**, e00148 (2019).
43. Balfagón, D. et al. Jasmonic acid Is required for plant acclimation to a combination of high light and heat stress. *Plant Physiol.* **181**, 1668–1682 (2019).
44. Li, G. et al. Abscisic acid negatively modulates heat tolerance in rolled leaf rice by increasing leaf temperature and regulating energy homeostasis. *Rice* **13**, 18 (2020).
45. Yoshida, K., Igarashi, E., Mukai, M., Hirata, K. & Miyamoto, K. Induction of tolerance to oxidative stress in the green alga, *Chlamydomonas reinhardtii*, by abscisic acid. *Plant, Cell Environ.* **26**, 451–457 (2003).
46. Al-Hijab, L. et al. Abscisic acid induced a negative geotropic response in dark-incubated *Chlamydomonas reinhardtii*. *Sci. Rep.* **9**, 12063 (2019).
47. Colina, F. et al. Genome-wide identification and characterization of CKIN/SnRK gene family in *Chlamydomonas reinhardtii*. *Sci. Rep.* **9**, 350 (2019).
48. Cross, F. R. & Umen, J. G. The *Chlamydomonas* cell cycle. *Plant J.* **82**, 370–392 (2015).
49. Fu, H.-Y. et al. Redesigning the QA binding site of Photosystem II allows reduction of exogenous quinones. *Nat. Commun.* **8**, 15274 (2017).
50. Głowacka, K. et al. Photosystem II Subunit S overexpression increases the efficiency of water use in a field-grown crop. *Nat. Commun.* **9**, 868 (2018).
51. Baker, N. R., Harbinson, J. & Kramer, D. M. Determining the limitations and regulation of photosynthetic energy transduction in leaves. *Plant, Cell Environ.* **30**, 1107–1125 (2007).
52. Zhang, R. et al. Moderate heat stress reduces the pH component of the transthylakoid proton motive force in light-adapted, intact tobacco leaves. *Plant, Cell Environ.* **32**, 1538–1547 (2009).
53. Zones, J. M., Blaby, I. K., Merchant, S. S. & Umen, J. G. High-resolution profiling of a synchronized diurnal transcriptome from *Chlamydomonas reinhardtii* reveals continuous cell and metabolic differentiation. *Plant Cell* **27**, 2743–2769 (2015).
54. Strenkert, D. et al. Multiomics resolution of molecular events during a day in the life of *Chlamydomonas*. *Proc. Natl Acad. Sci. USA* **116**, 2374–2383 (2019).
55. Hasanuzzaman, M. et al. Reactive oxygen species and antioxidant defense in plants under abiotic stress: revisiting the crucial role of a universal defense regulator. *Antioxidants* **9**, 681 (2020).
56. Anderson, A. P., Luo, X., Russell, W. & Yin, Y. W. Oxidative damage diminishes mitochondrial DNA polymerase replication fidelity. *Nucleic Acids Res* **48**, 817–829 (2020).
57. Robert, G. & Wagner, J. R. ROS-induced DNA damage as an underlying cause of aging. *Advances in Geriatric Medicine and Research* **4**, e200024 (2020).
58. Velichko, A. K., Markova, E. N., Petrova, N. V., Razin, S. V. & Kantidze, O. L. Mechanisms of heat shock response in mammals. *Cell Mol. Life Sci.* **70**, 4229–4241 (2013).
59. Waszczak, C., Carmody, M. & Kangasjärvi, J. Reactive oxygen species in plant signaling. *Annu Rev. Plant Biol.* **69**, 209–236 (2018).
60. Cronmiller, E. et al. Cell wall integrity signaling regulates cell wall-related gene expression in *Chlamydomonas reinhardtii*. *Sci. Rep.* **9**, 12204 (2019).
61. Li, Z., Tang, J., Srivastava, R., Bassham, D. C. & Howell, S. H. The transcription factor bZIP60 links the unfolded protein response to the heat stress response in maize. *Plant Cell* **32**, 3559–3575 (2020).
62. Lien, T. & Knutsen, G. Synchronous growth of *Chlamydomonas reinhardtii*(chlorophyceae): a review of optimal conditions. *J. Phycol.* **15**, 191–200 (1979).
63. Vítová, M. et al. *Chlamydomonas reinhardtii*: duration of its cell cycle and phases at growth rates affected by temperature. *Planta* **234**, 599–608 (2011).
64. Strenkert, D., Schmollinger, S., Sommer, F., Schulz-Raffelt, M. & Schroda, M. Transcription factor–dependent chromatin remodeling at heat shock and copper-responsive promoters in *Chlamydomonas reinhardtii*. *Plant Cell* **23**, 2285–2301 (2011).
65. Rütgers, M. et al. Substrates of the chloroplast small heat shock proteins 22E/F point to thermolability as a regulative switch for heat acclimation in *Chlamydomonas reinhardtii*. *Plant Mol. Biol.* **95**, 579–591 (2017).
66. Yamori, W. & Shikanai, T. Physiological functions of cyclic electron transport around Photosystem I in sustaining photosynthesis and plant growth. *Annu. Rev. Plant Biol.* **67**, 81–106 (2016).
67. Johnson, G. N. Physiology of PSI cyclic electron transport in higher plants. *Biochimica et. Biophysica Acta (BBA) - Bioenerg.* **1807**, 384–389 (2011).
68. He, Y. et al. Increasing cyclic electron flow is related to Na+ sequestration into vacuoles for salt tolerance in soybean. *J. Exp. Bot.* **66**, 6877–6889 (2015).
69. Huang, W., Yang, S.-J., Zhang, S.-B., Zhang, J.-L. & Cao, K.-F. Cyclic electron flow plays an important role in photoprotection for the resurrection plant Paraboearufescens under drought stress. *Planta* **235**, 819–828 (2012).
70. Johnson, X. et al. Proton gradient regulation 5-mediated cyclic electron flow under ATP- or redox-limited conditions:a study of ΔATPase pgr5 and ΔrbcL pgr5 mutants in the green alga *Chlamydomonas reinhardtii*. *Plant Physiol.* **165**, 438–452 (2014).
71. Saroussi, S. I., Wittkopp, T. M. & Grossman, A. R. The type II NADPH dehydrogenase facilitates cyclic electron flow, energy-dependent quenching, and chlororespiratory metabolism during acclimation of *Chlamydomonas reinhardtii* to nitrogen deprivation. *Plant Physiol.* **170**, 1975–1988 (2016).
72. Aihara, Y., Takahashi, S. & Minagawa, J. Heat induction of cyclic electron flow around Photosystem I in the symbiotic dinoflagellate symbiodinium. *Plant Physiol.* **171**, 522–529 (2016).
73. Alric, J., Lavergne, J. & Rappaport, F. Redox and ATP control of photosynthetic cyclic electron flow in *Chlamydomonas reinhardtii* in aerobic conditions. *Biochimica et Biophysica Acta (BBA). Bioenergetics* **1797**, 44–51 (2010).
74. Alric, J. Cyclic electron flow around photosystem I in unicellular green algae. *Photosynth Res* **106**, 47–56 (2010).
75. Lucker, B. & Kramer, D. M. Regulation of cyclic electron flow in *Chlamydomonas reinhardtii* under fluctuating carbon availability. *Photosynth Res* **117**, 449–459 (2013).
76. Lamb, J. J., Røkke, G. & Hohmann-Marriott, M. F. Chlorophyll fluorescence emission spectroscopy of oxygenic organisms at 77 K. *Photosynthetica* **56**, 105–124 (2018).
77. Wood, W. H. J., Barnett, S. F. H., Flannery, S., Hunter, C. N. & Johnson, M. P. Dynamic thylakoid stacking Is regulated by LHCII phosphorylation but not its interaction with PSI. *Plant Physiol.* **180**, 2152–2166 (2019).
78. Szyszka-Mroz, B., Pittock, P., Ivanov, A. G., Lajoie, G. & Hüner, N. P. A. The Antarctic psychrophile *Chlamydomonas* sp. UWO 241 preferentially phosphorylates a Photosystem I-Cytochrome b6/f supercomplex. *Plant Physiol.* **169**, 717–736 (2015).
79. Cook, G. et al. The Antarctic psychrophiles *Chlamydomonas* spp. UWO241 and ICE-MDV exhibit differential restructuring of photosystem I in response to iron. *Photosynth Res* **141**, 209–228 (2019).
80. Kalra, I. et al. Chlamydomonas sp. UWO 241 exhibits high cyclic electron flow and rewired metabolism under high salinity. *Plant Physiol.* **183**, 588–601 (2020).
81. Iwai, M. et al. Isolation of the elusive supercomplex that drives cyclic electron flow in photosynthesis. *Nature* **464**, 1210–1213 (2010).
82. Steinbeck, J. et al. Structure of a PSI–LHCI–cyt b6f supercomplex in *Chlamydomonas reinhardtii* promoting cyclic electron flow under anaerobic conditions. *PNAS* **115**, 10517–10522 (2018).
83. Su, X. et al. Antenna arrangement and energy transfer pathways of a green algal photosystem-I–LHCI supercomplex. *Nat. Plants* **5**, 273–281 (2019).
84. Wunder, T., Oh, Z. G. & Mueller-Cajar, O. CO_2 -fixing liquid droplets: Towards a dissection of the microalgal pyrenoid. *Traffic* **20**, 380–389 (2019).
85. Meyer, M. T., Whittaker, C. & Griffiths, H. The algal pyrenoid: key unanswered questions. *J. Exp. Bot.* **68**, 3739–3749 (2017).
86. Hennacy, J. H. & Jonikas, M. C. Prospects for engineering biophysical CO_2 concentrating mechanisms into land Plants to enhance yields. *Annu. Rev. Plant Biol.* **71**, 461–485 (2020).
87. Meyer, M. T. et al. Assembly of the algal CO_2-fixing organelle, the pyrenoid, is guided by a Rubisco-binding motif. *Sci. Adv.* **6**, eabd2408 (2020).
88. Mackinder, L. C. M. et al. A repeat protein links Rubisco to form the eukaryotic carbon-concentrating organelle. *PNAS* **113**, 5958–5963 (2016).
89. He, S. et al. The structural basis of Rubisco phase separation in the pyrenoid. *Nat. Plants* **6**, 1480–1490 (2020).
90. Itakura, A. K. et al. A Rubisco-binding protein is required for normal pyrenoid number and starch sheath morphology in *Chlamydomonas reinhardtii*. *PNAS* **116**, 18445–18454 (2019).

91. Lohr, M., Im, C.-S. & Grossman, A. R. Genome-based examination of chlorophyll and carotenoid biosynthesis in *Chlamydomonas reinhardtii*. *Plant Physiol.* **138**, 490–515 (2005).
92. Bassi, D., Menossi, M. & Mattiello, L. Nitrogen supply influences photosynthesis establishment along the sugarcane leaf. *Sci. Rep.* **8**, 2327 (2018).
93. Ordóñez, R. A., Savin, R., Cossani, C. M. & Slafer, G. A. Yield response to heat stress as affected by nitrogen availability in maize. *Field Crops Res.* **183**, 184–203 (2015).
94. Wang, Q.-L., Chen, J.-H., He, N.-Y. & Guo, F.-Q. Metabolic reprogramming in chloroplasts under heat stress in plants. *Int. J. Mol. Sci.* **19**, 849 (2018).
95. Rossi, S., Burgess, P., Jespersen, D. & Huang, B. Heat-Induced Leaf Senescence Associated with Chlorophyll Metabolism in Bentgrass Lines Differing in Heat Tolerance. *Crop Sci.* **57**, S-169–S-178 (2017).
96. Sager, R. Inheritance in the green alga *Chlamydomonas reinhardtii*. *Genetics* **40**, 476–489 (1955).
97. Pröschold, T., Harris, E. H. & Coleman, A. W. Portrait of a species: *Chlamydomonas reinhardtii*. *Genetics* **170**, 1601–1610 (2005).
98. Zhang, N. et al. Comparative Phenotyping of Two Commonly Used Chlamydomonas reinhardtii Background Strains: CC-1690 (21gr) and CC-5325 (The CLiP Mutant Library Background). *Plants* **11**, 585 (2022).
99. Schloss, J. A. A. *Chlamydomonas* gene encodes a G protein β subunit-like polypeptide. *Mol. Gen. Genet* **221**, 443–452 (1990).
100. Xie, B. et al. *Chlamydomonas reinhardtii* thermal tolerance enhancement mediated by a mutualistic interaction with vitamin B12-producing bacteria. *ISME J.* **7**, 1544–1555 (2013).
101. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**, 402–408 (2001).
102. Hellemans, J., Mortier, G., De Paepe, A., Speleman, F. & Vandesompele, J. QBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. *Genome Biol.* **8**, R19 (2007).
103. Remans, T. et al. Reliable gene expression analysis by reverse transcription-quantitative PCR: reporting and minimizing the uncertainty in data accuracy. *Plant Cell* **26**, 3829–3837 (2014).
104. Bushnell, B. BBMap. *BBMap short read aligner, and other bioinformatic tools* https://sourceforge.net/projects/bbmap/.
105. Andrews, S. Babraham Bioinformatics - FastQC a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).
106. Merchant, S. S. et al. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* **318**, 245–250 (2007).
107. Dobin, A. & Gingeras, T. R. Mapping RNA-seq reads with STAR. *Curr. Protoc. Bioinformatics* **51**, 11.14.1–11.14.19 (2015).
108. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
109. Anders, S., Pyl, P. T. & Huber, W. HTSeq- a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2014).
110. Ma, F., Salomé, P. A., Merchant, S. S. & Pellegrini, M. Single-cell RNA sequencing of batch Chlamydomonas cultures reveals heterogeneity in their diurnal cycle phase. *Plant Cell* **33**, 1042–1057 (2021).
111. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.: Ser. B (Methodol.)* **57**, 289–300 (1995).
112. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinforma.* **9**, 559 (2008).
113. Remacle, F., Kravchenko-Balasha, N., Levitzki, A. & Levine, R. D. Information-theoretic analysis of phenotype changes in early stages of carcinogenesis. *PNAS* **107**, 10324–10329 (2010).
114. Schneider, K., Venn, B. & Mühlhaus, T. TMEA: A thermodynamically motivated framework for functional characterization of biological responses to system acclimation. *Entropy* **22**, 1030 (2020).
115. 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).
116. Cox, J. et al. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell Proteom.* **13**, 2513–2526 (2014).
117. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
118. Luo, F. et al. Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. *BMC Bioinformatics* **17**, 299 (2007).
119. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLOS ONE* **9**, 12 (2014).
120. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).
121. Langfelder, P. & Horvath, S. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst. Biol.* **17**, 54 (2007).
122. Silverman. Density estimation for statistics and data analysis. Book. Published in Monographs on Statistics and Applied Probability (London: Chapman and Hall, 1986).
123. Tulin, F. & Cross, F. R. A microbial avenue to cell cycle control in the plant superkingdom. *Plant Cell* **26**, 4019–4038 (2014).
124. Porra, R. J., Thompson, W. A. & Kriedemann, P. E. Determination of accurate extinction coefficients and simultaneous equations for assaying chlorophylls a and b extracted with four different solvents: verification of the concentration of chlorophyll standards by atomic absorption spectroscopy. *Biochimica et. Biophysica Acta (BBA) - Bioenerg.* **975**, 384–394 (1989).
125. Wellburn, A. R. The spectral determination of chlorophylls a and b, as well as total carotenoids, using various solvents with spectrophotometers of different resolution. *J. Plant Physiol.* **144**, 307–313 (1994).
126. Joo, J. H., Wang, S., Chen, J. G., Jones, A. M. & Fedoroff, N. V. Different signaling and cell death roles of heterotrimeric G protein α and β subunits in the Arabidopsis oxidative stress response to ozone. *Plant Cell* **17**, 957–970 (2005).
127. Pérez-Pérez, M. E., Couso, I. & Crespo, J. L. Carotenoid deficiency triggers autophagy in the model green alga *Chlamydomonas reinhardtii*. *Autophagy* **8**, 376–388 (2012).
128. Witt, H. T. Energy conversion in the functional membrane of photosynthesis. Analysis by light pulse and electric pulse methods: The central role of the electric field. *Biochim Biophys Acta.* **73**, 355–427 (1979).
129. Kramer, D. M., Avenson, T. J. & Edwards, G. E. Dynamic flexibility in the light reactions of photosynthesis governed by both electron and proton transfer reactions. *Trends Plant Sci.* **9**, 349–357 (2004).
130. Cruz, J. A. Plasticity in light reactions of photosynthesis for energy production and photoprotection. *J. Exp. Bot.* **56**, 395–406 (2004).
131. Murakami, A. Quantitative analysis of 77K fluorescence emission spectra in Synechocystis sp. PCC 6714 and *Chlamydomonas reinhardtii* with variable PS I/PS II stoichiometries. *Photosynthesis Res.* **53**, 141–148 (1997).
132. Jeong, J., Baek, K., Kirst, H., Melis, A. & Jin, E. Loss of CpSRP54 function leads to a truncated light-harvesting antenna size in *Chlamydomonas reinhardtii*. *Biochimica et Biophysica Acta (BBA). Bioenergetics* **1858**, 45–55 (2017).
133. Reynolds, E. S. The use of lead citrate at high pH as an electron-opaque stain in electron microscopy. *J. Cell Biol.* **17**, 208–212 (1963).
134. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207–210 (2002).
135. Perez-Riverol, Y. et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2019).
136. Deutsch, E. W. et al. The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics. *Nucleic Acids Res.* **48**, D1145–D1152 (2020).

Acknowledgements

This research was supported by the start-up funding from the Donald Danforth Plant Science Center (DDPSC), the Department of Energy (DOE) Basic Energy Sciences (BES) Photosynthetic Systems (PS) grant (Award No. 0019464), the DOE Biological & Environmental Research (BER) grant (Award No. 0020400), and the RNA-seq service support from the Department of Energy (DOE) Joint Genome Institute (JGI) Community Science Program (CSP, 503414) to R.Z.. The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No DE-AC02-05CH11231. E.M.M. was supported by the William H. Danforth Fellowship in Plant Sciences, DDPSC start-up funding (to R.Z.) and Washington University in St. Louis. M.S. was funded by Deutsche Forschungsgemeinschaft (TRR175, projects C02 and D02). We would like to thank researchers from the Kramer Laboratory (Drs. David Kramer, Jeffrey Cruz, Robert Zegarac, Ben Lucker, Geoffry Davis) for their helpful suggestions and discussion to set up and optimize the IDEAspec for spectroscopic measurements in algal cells. Dr. Geoffry Davis is also thanked for his suggestions for 77 K chlorophyll fluorescence measurement. We would like to thank Drs. Jeremy Schmutz, Susannah Tringe, Christa Pennacchio, Chris Daum, and Ronan O'Malley for RNA-seq service at DOE JGI. We acknowledge imaging support from the Advanced Bioimaging Laboratory (RRID:SCR_018951) at DDPSC and the usage of the LEO 912AB Energy Filter TEM acquired through a National Science Foundation (NSF) Major Research Instrumentation grant (DBI-0116650). We also acknowledge proteomics support from the Proteomics and Mass Spectrometry at DDPSC and usage of the Orbitrap Fusion Lumos LC-MS/MS, which was funded by the National Science Foundation under Grant No. DBI-1827534. We would like to thank Dr. James Umen for helpful discussion about our research and Sarah Rommelfanger for suggestions on RNA-seq analysis. We also want to thank Drs. Xin Wang and Blake Meyers for valuable feedback of the manuscript.

Author contributions

R.Z. supervised the whole project. R.Z., N.Z. and E.M.M. designed and planned all the experiments. R.Z. and M.S. discussed and designed the time points for high temperature treatments. N.Z. led the project, characterized cell physiologies (including cell density, size, viability, protein, ROS, starch, light microscopic images), conducted RT-qPCRs and extracted RNA for RNA-seq analysis. W.M. and M.X. grew and maintained algal cultures in photobioreactors. N.Z, W.M., C.A. J.J, M.X. and E.M.M. harvested algal samples from photobioreactors with different treatments at different time points. E.M.M. led RNA-seq data analysis and generated all related figures. J.C.B. provided suggestions for RNA-seq analysis and statistical analysis. N.Z. extracted protein for proteomics. S.T. and B.E. quantified protein abundance using LC-MS/MS. B.V., D. Z., T.M., E.M.M., M.S., and N.Z. analyzed the proteomics data. C.C. and J.C. provided suggestions for transcriptomic and proteomic analysis. K.P., F.C and N.Z. performed DNA content analysis. C.A. and M.X. quantified chlorophyll and carotenoid contents. W.M. performed all spectroscopic measurements of photosynthetic parameters. J.J. and L.P. performed O_2 evolution measurements. N.Z. and J.J. optimized the ROS protocol. N.Z., E.B. and K.J.C. performed TEM analysis. R.Z., E.M.M., and N.Z. led the writing of the manuscript with the contribution of all other co-authors. R.Z., E.M.M., N.Z., M.S., B.V., M. X., T.M., J.J, K.J.C., J.C.B., K.P, B.E. revised the manuscript. All the authors discussed the results and contributed to data interpretation.

Competing interests

The authors declare no competing interests.

Additional information

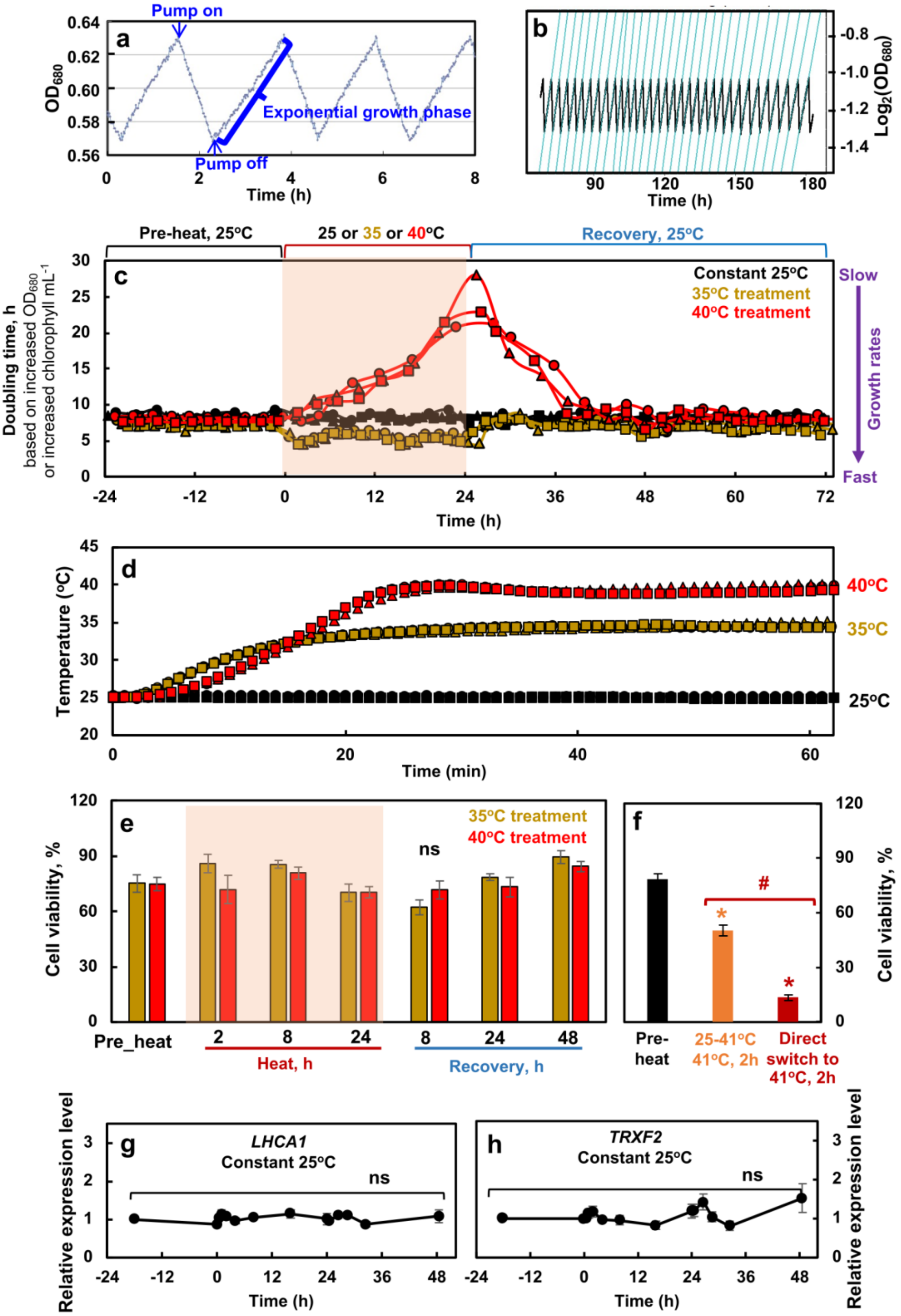
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-022-03359-z.

Correspondence and requests for materials should be addressed to Ru Zhang.

Reprints and permission information is available at http://www.nature.com/reprints

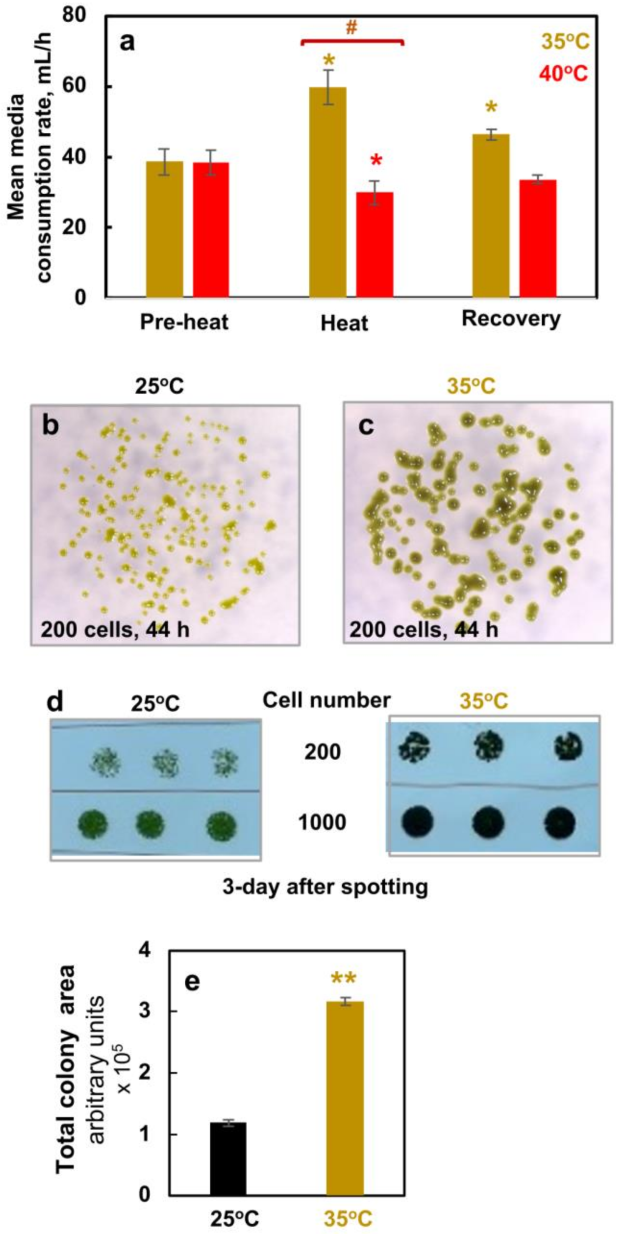
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary Fig. 1



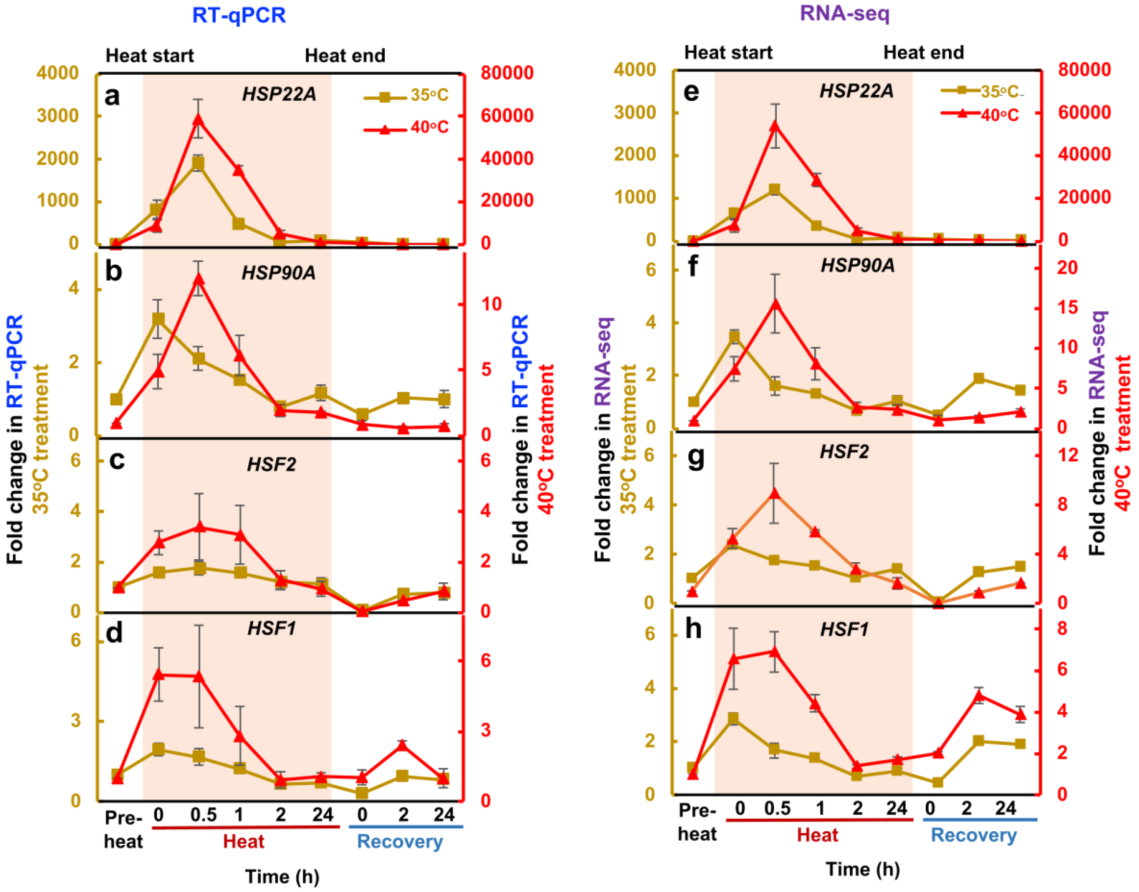
Supplementary Fig. 1: Algal cultures were grown in photobioreactors (PBRs) under well-controlled conditions with turbidostatic mode for different temperature treatments. (a) PBR cultures were grown with air level CO2 in Tris-acetate-phosphate (TAP) medium under constant light (100 µmol photons m-2 s-1 with equal amount of blue and red light) turbidostatically within a small range of OD680 which monitors chlorophyll content. When PBR cultures grew to the set maximum OD680, pumps turned on to add fresh medium and dilute the cultures to the set minimal OD680, then pumps turned off to allow for exponential growth to the maximum set OD680. **(b)** The doubling time of the PBR cultures were calculated by fitting the OD680 curves during exponential growth. **(c)** Doubling time of the PBR cultures before, during and after treatment at 35 °C, 40 °C, or constant 25 °C. Doubling time is inverse of relative growth rates and smaller doubling time represents faster growth based on the rate of chlorophyll increase. Three independent biological replicates are plotted for each treatment. Constant 25 °C served as controls which showed steady growth without heat treatment. **(d)** PBR heating profiles. PBR temperatures changed from 25 °C to 35 °C or 40 °C gradually within 30 min. Three independent biological replicates are plotted for each temperature treatment. **(e)** Heat treatment in PBRs at 35 °C or 40 °C up to 24 h did not affect cell viability. Algal cells with different heat treatments were diluted and spotted on TAP plates, grown under 150 µmol photons m-2 s-1 white LED light, 25°C for 44 h before microscopic imaging. Colony numbers were quantified using ImageJ. Cell viability was calculated as the number of colonies on plates divided by the number of cells spotted. Values are mean ± SE, n = 3 biological replicates. Statistical analyses were performed with two-tailed t-test assuming unequal variance by comparing different time points with the pre-heat samples. No significance (ns) among different time points (p>0.05). **(f)** Heating speed affected algal cell viability and direct heating at 41 °C in water bath significantly reduced algal cell viability. Algal cultures were harvested from PBRs before heat treatment (pre-heat, black bars), or incubated in a water bath which was heated from 25°C to 41 °C gradually in 25 min then kept at 41oC for 2 h (orange bars), or directly heated in a water bath which was pre-heated to 41oC (sharp temperature switch) then kept at 41 °C for 2 h (red bars). Cell viability was quantified as in **(e)**. Values are mean ± SE, n = 3 biological replicates. Statistical analyses were performed using two-tailed t-tests assuming unequal variance by comparing treated samples with pre-heat (*, p<0.05, the colors of asterisks match the treatment conditions) or between the two heating methods (#, p<0.05). **(g, h)** The circadian regulated genes LHCA1 and TRXF2 had constant expression levels without heat treatments. The relative expressions were calculated from RT-qPCR by normalizing to the reference gene CBLP and pre-heat-stress level. Mean ± SE, n = 3 biological replicates. Statistical analyses were performed with two-tailed t-test assuming unequal variance by comparing different time points with the first time point. No significance (ns) among different time points (p>0.05). **(c, e)** Red shaded area depicts the duration of high temperatures.

Supplementary Fig. 2



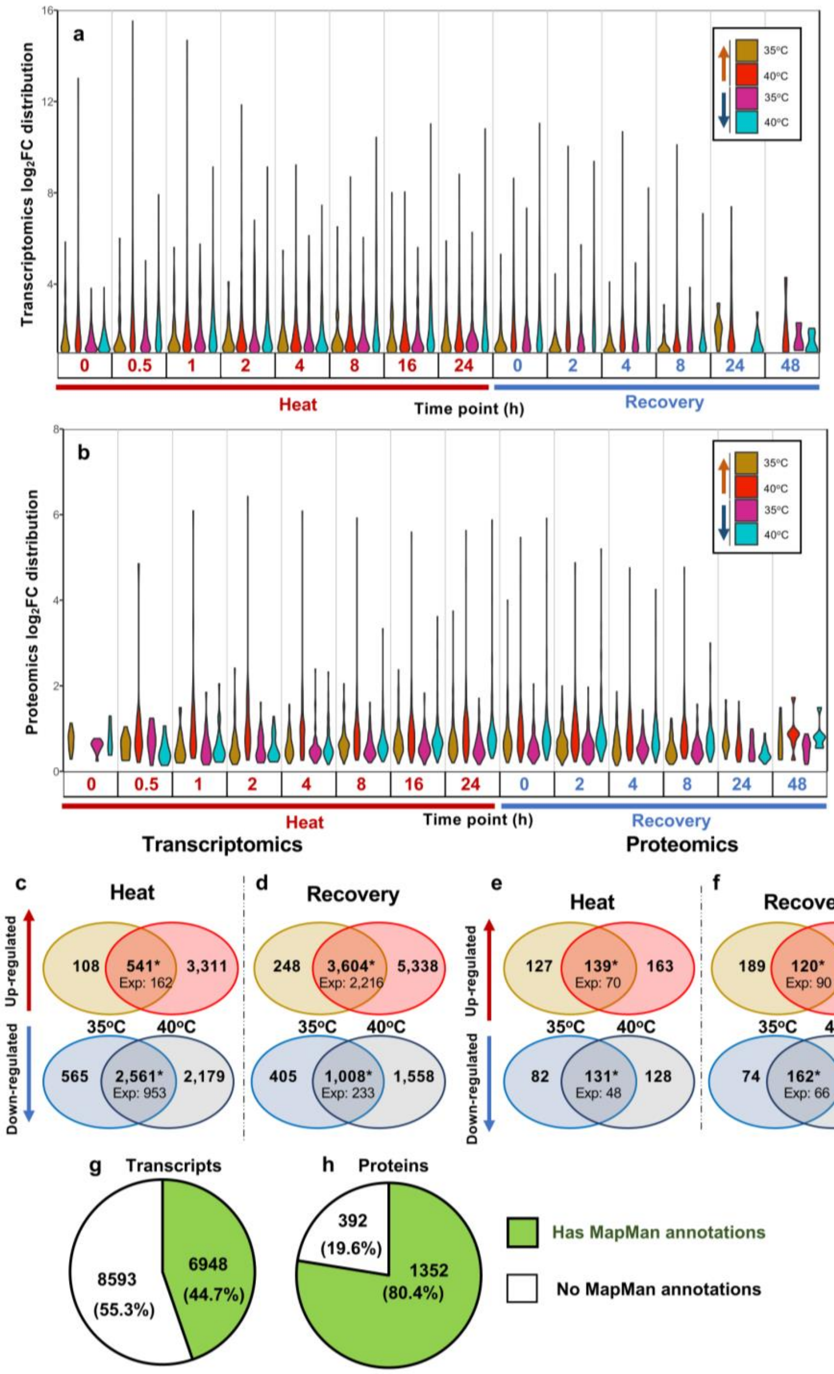
Supplementary Fig. 2. Moderate high temperature at 35 °C increased algal growth rates. (a) PBR mean media consumption before heat, at the end of 24-h heat of 35 °C or 40 °C, and at the end of the recovery at 25 °C, total media consumption volume divided by time. Cell growth induced culture dilution through turbidostatic control and consumed medium. Mean ± SE, n = 4. **(b-e)** The increased growth rates under 35 °C were confirmed by spotting tests on plates. Algal cells harvested from PBRs at 25 °C without heat treatments were diluted and spotted on TAP plates, grown under 150 µmol photons m-2 s-1 white LED light, in incubators of 25 °C or 35 °C for 44 h **(b, c)** or 3 days **(d)** before imaging. **(b, c, d)** One of the three biological replicates was shown. **(e)** Algal spots with 200 cells were imaged after 44-h growth and colony areas were quantified using ImageJ. Values are mean ± SE, n = 3 biological replicates. **(a, e)** Statistical analyses were performed with two-tailed t-test assuming unequal variance by comparing treated conditions with the pre-heat **(a)** or the 25 °C conditions **(e)** (*, p<0.05; **, p<0.01; the colors of asterisks match treatment conditions) or between 35 °C and 40 °C heat treatment (#, p<0.05).

Supplementary Fig. 3



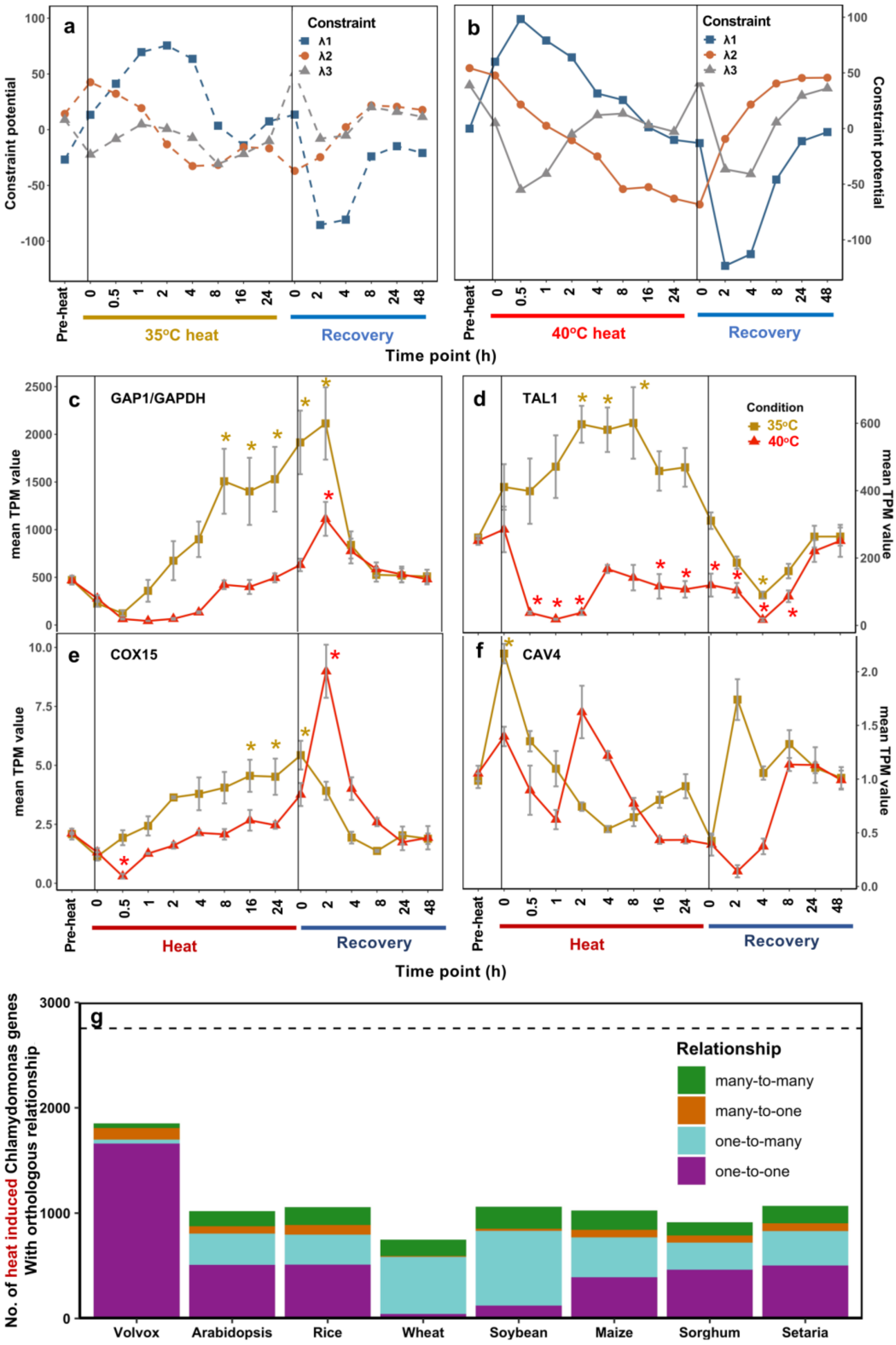
Supplementary Fig. 3. RT-qPCR analysis was consistent with RNA-seq results. Transcript fold-changes of two heat stress marker genes (HSP22A, HSP90A) and two heat shock transcription factors (HSF1 and HSF2) were calculated based on RT-qPCR (a-d) or RNA-seq (e-h) results. Different y scales were used for samples with 35oC (left) or 40oC (right) treatments. Red shaded area depicts the duration of high temperature. For RT-qPCR results, the fold-changes were calculated by normalizing the relative expression values at different time points with different treatments to the reference gene CBLP and the pre-heat time point. For RNA-seq results, the fold-changes were calculated based on Transcripts Per Million (TPM) normalized RNA-seq read counts. Values are mean ± SE, n = 3 biological replicates.

Supplementary Fig. 4



Supplementary Fig. 4: Transcriptomic and proteomic analyses revealed shared and unique regulation of transcripts and proteins during and after heat treatments of 35oC or 40oC. (a, b) Log2(fold-change, FC) distribution of Differentially Expressed Genes (DEGs) and Differentially Accumulated Proteins (DAPs) at different time points, respectively. For each time point, the first two violins represent up-regulated transcripts/proteins, and the last two violins represent down-regulated transcripts/proteins. For down-regulated transcripts/proteins, the inverse of the log2FC was displayed. The width of the violins is proportional to the fraction of transcripts/proteins at a certain fold-change value out of the total DEGs/DAPs at a given time point. Time points during heat: 0 h, reach high temperature of 35oC or 40oC; 0.5 h, heat at 35oC or 40oC for 0.5 h, similar names for other time points during heat. Time points during recovery: 0 h, reach control temperature of 25oC for recovery after heat; 2 h, recovery at 25oC for 2 h, similar names for other time points during recovery. (c-f) Venn diagrams of transcripts (c, d) and proteins (e, f) differentially expressed in at least one time point during heat treatment (c, e) or recovery (d, f). For each panel: top, up-regulated transcripts/proteins; bottom: down-regulated transcripts/proteins. Only transcripts and proteins identified in both 35oC and 40oC treatment groups were used for this analysis. Expected values are the number of transcripts/proteins expected to have overlapping differential expression between the 35ºC and 40ºC treatment groups based on random chance (Fisher's exact test, *: p< 1.29 x10-226). (g, h) Pie chart of transcripts (g) and proteins (h) in our analyses that have at least one MapMan annotation (green) versus no MapMan annotations (white).

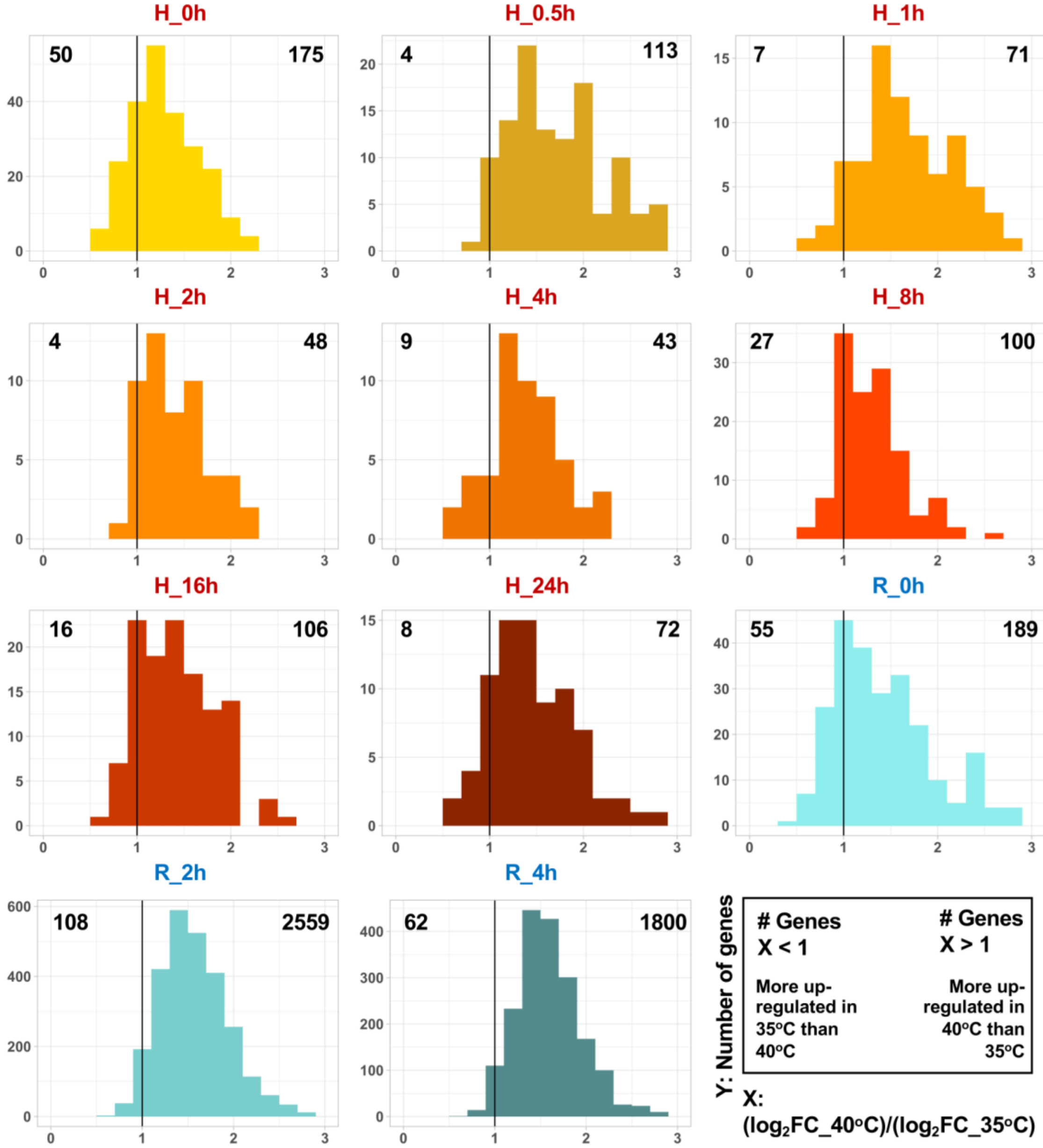
Supplementary Fig. 5



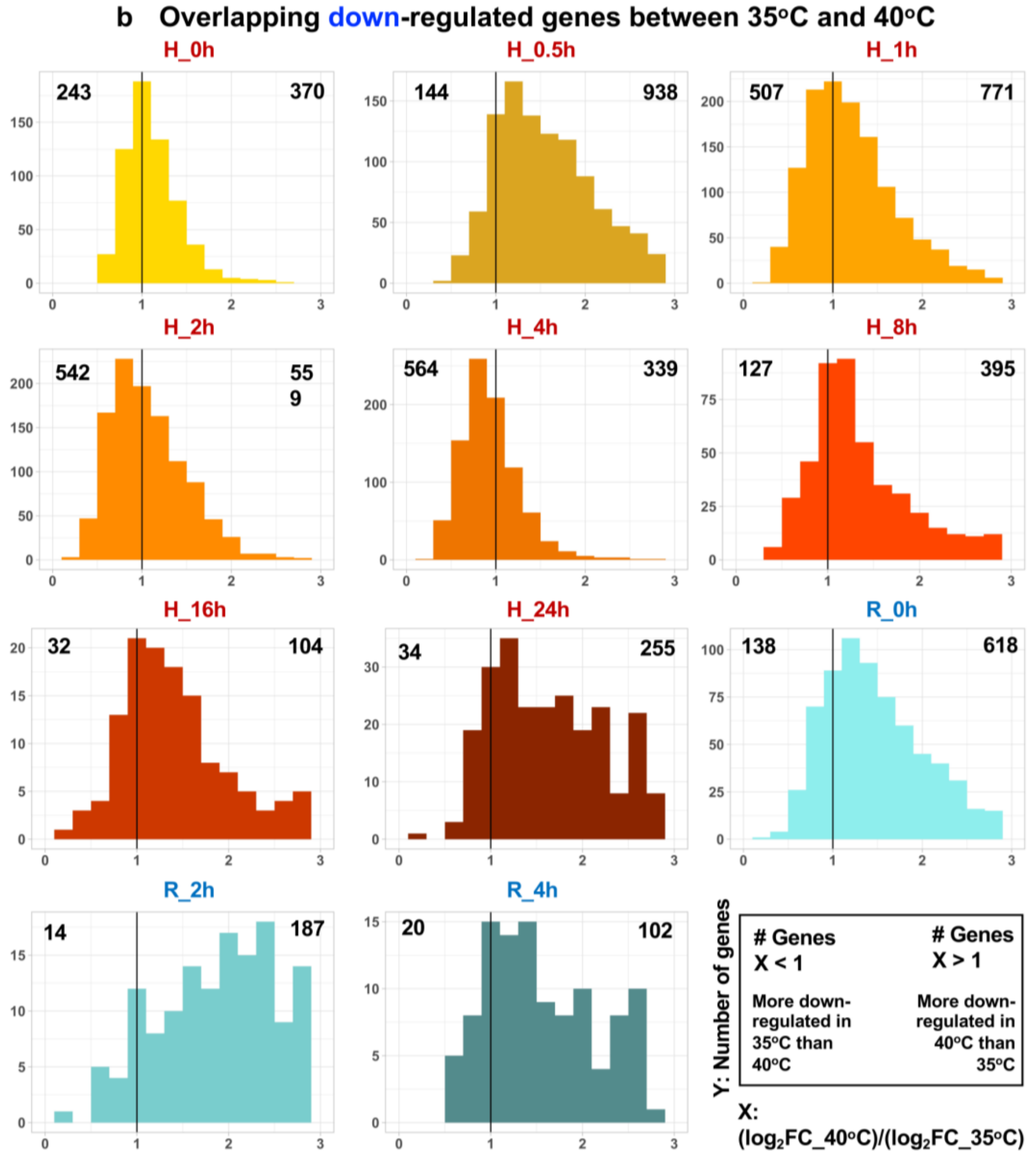
Supplementary Fig. 5: Global transcription patterns were similar between 35oC and 40oC treatments, but detailed analysis revealed uniquely differentially expressed genes in the 35oC treatment. Time course of the three major constraint potentials (l1- 3) derived from surprisal analysis for 35 °C (a) and 40°C (b) experiments, respectively. The constraint potentials indicate the most important transcriptional patterns during the time course. (c-f) Mean transcript per million (TPM) read counts at each time point for select genes that were uniquely up-regulated during 35 °C (brown) but not 40 °C (red) heat treatment period. Values are mean ± SE, n = 3 biological replicates, asterisks indicate significance in differential expression modeling. (c) GAP1/GAPDH: Cre12.g485150, Glyceraldehyde 3-phosphate dehydrogenase, involved in gluconeogenesis, glycolysis, and Calvin-Benson Cycle; (d) TAL1: Cre01.g032650, transaldolase, involved in the pentose phosphate pathway, which acts upstream of the glycolytic and gluconeogenic pathways; (e) COX15: Cre02.g082700, encoding mitochondrial cytochrome c oxidase assembly factor; (f) CAV4: Cre11.g467528, encoding a putative voltage-gated calcium channel. Vertical black lines indicate the start and end of heat treatments. Time points were labeled at the bottom. (a-f) Pre-heat, before heat treatments. Time points during heat: 0 h, reach high temperature of 35 °C or 40 °C; 0.5 h, heat at 35 °C or 40 °C for 0.5 h, similar names for other time points during heat. Time points during recovery: 0 h, reach control temperature of 25 °C for recovery after heat; 2 h, recovery at 25 °C for 2 h, similar names for other time points during recovery. (G) Conservation of Chlamydomonas heat induced genes (HIGs), which are up-regulated in at least one time point of 35 °C or 40 °C high temperature period, with select land plant species. Orthologous relationships were determined using JGI InParanoid data. Dashed horizontal line indicates the number of Chlamydomonas HIGs that were present in the JGI InParanoid data (2754 genes out of 3960 HIGs total). Abbreviated species names on x-axis correspond with the following JGI genomes: Volvox (Volvox carteri, Vcarteri_v2.1), Arabidopsis (Arabidopsis thaliana, Athaliana_TAIR10), Rice (Oryza sativa, Osativa_v7.0), Wheat (Triticum aestivum, Taestivum_v2.2), Soybean (Glycine max, Gmax_Wm82.a2.v1), Maize (Zea mays, Zmays_RefGen_V4), Sorghum (Sorghum bicolor, Sbicolor_v3.1.1), Setaria (Setaria viridis, Sviridis_v2.1).

Supplementary Fig. 6

a Overlapping up-regulated genes between 35°C and 40°C

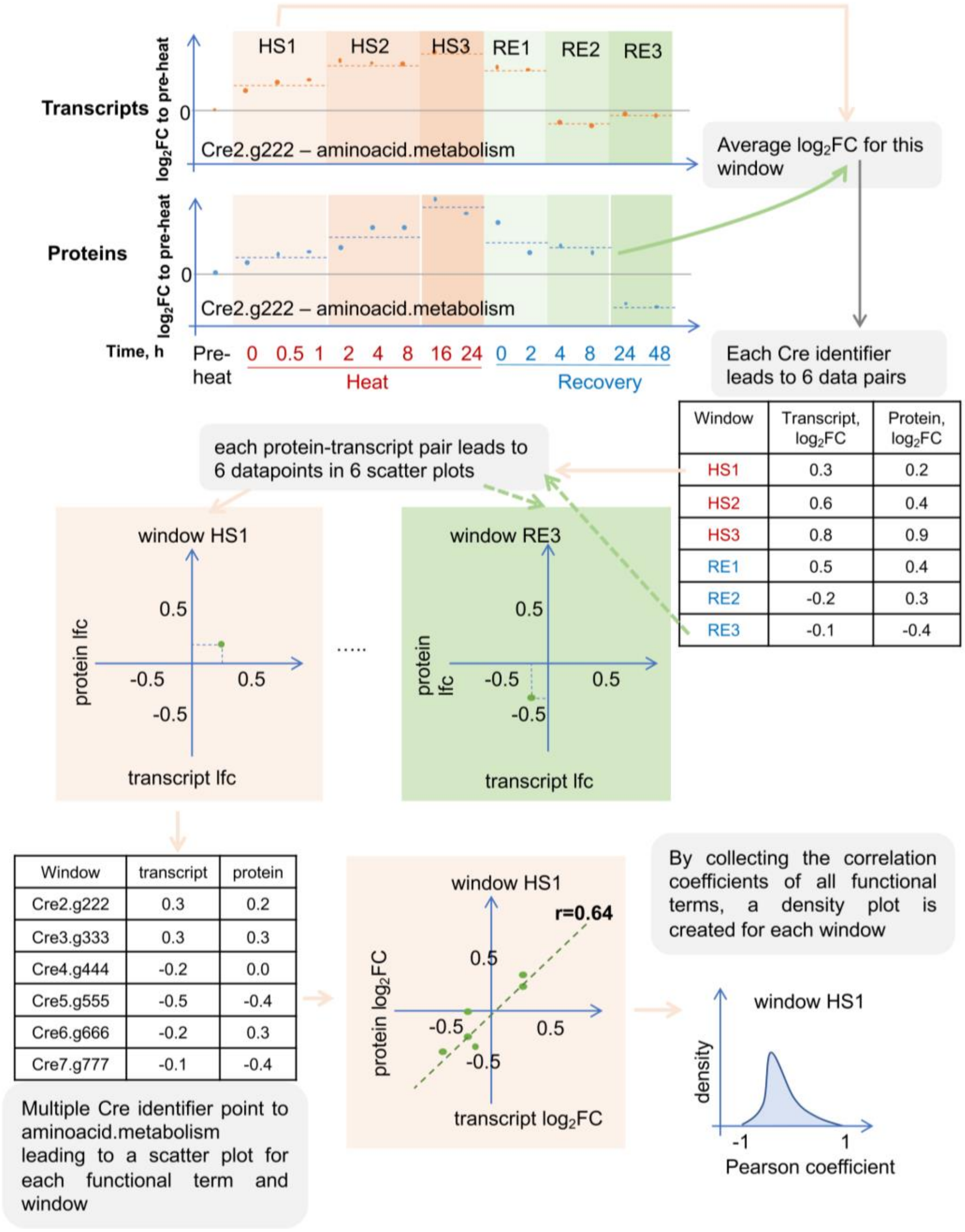


Supplementary Fig. 6



Supplementary Fig. 6: Some overlapping DEGs between 35 °C and 40 °C were more differentially regulated with 35°C than 40°C treatment. DEGs, differentially expressed genes. FC, fold-change. Histograms of log2(FC in 40 ºC)/log2(FC in 35 ºC) for overlapping up-regulated **(a)** and down-regulated **(b)** genes between 35 °C and 40 °C are displayed for each time point. Very few overlapping DEGs between 35 °C and 40 °C were identified at 8, 24, and 48 h of recovery, which were thus omitted. Black vertical lines indicate equal differential expression between 35 °C and 40 °C treatments. Bars to the left of the black line indicate genes more differentially expressed in the 35 ºC treatment group while bars to the right of the black line indicate genes more differentially expressed in the 40 ºC treatment group. Numbers in the top left and right corners of each histogram represent the number of genes with higher fold change values in 35 °C or 40 °C, respectively. H_0h, reach high temperature of 35 °C or 40 °C. H_0.5h, heat at 35 °C or 40 °C for 0.5 h, similar names for other time points during heat. R_0h, reach control temperature of 25 °C for recovery after heat. R_2h, recovery at 25 °C for 2 h, similar names for other time points during recovery.

Supplementary Fig. 7



Average log$_2$FC for this window

Each Cre identifier leads to 6 data pairs

Window	Transcript, log$_2$FC	Protein, log$_2$FC
HS1	0.3	0.2
HS2	0.6	0.4
HS3	0.8	0.9
RE1	0.5	0.4
RE2	-0.2	0.3
RE3	-0.1	-0.4

each protein-transcript pair leads to 6 datapoints in 6 scatter plots

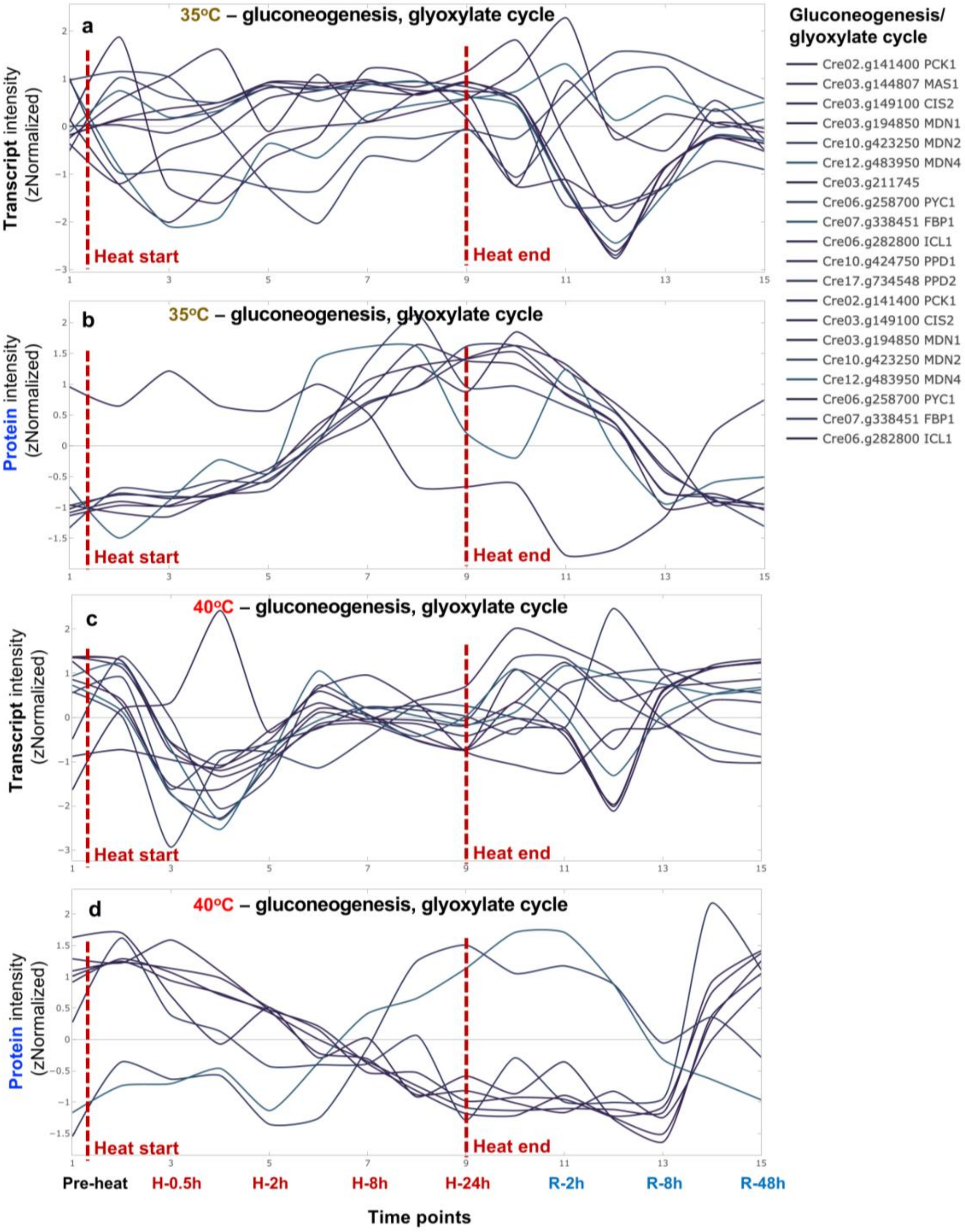
Window	transcript	protein
Cre2.g222	0.3	0.2
Cre3.g333	0.3	0.3
Cre4.g444	-0.2	0.0
Cre5.g555	-0.5	-0.4
Cre6.g666	-0.2	0.3
Cre7.g777	-0.1	-0.4

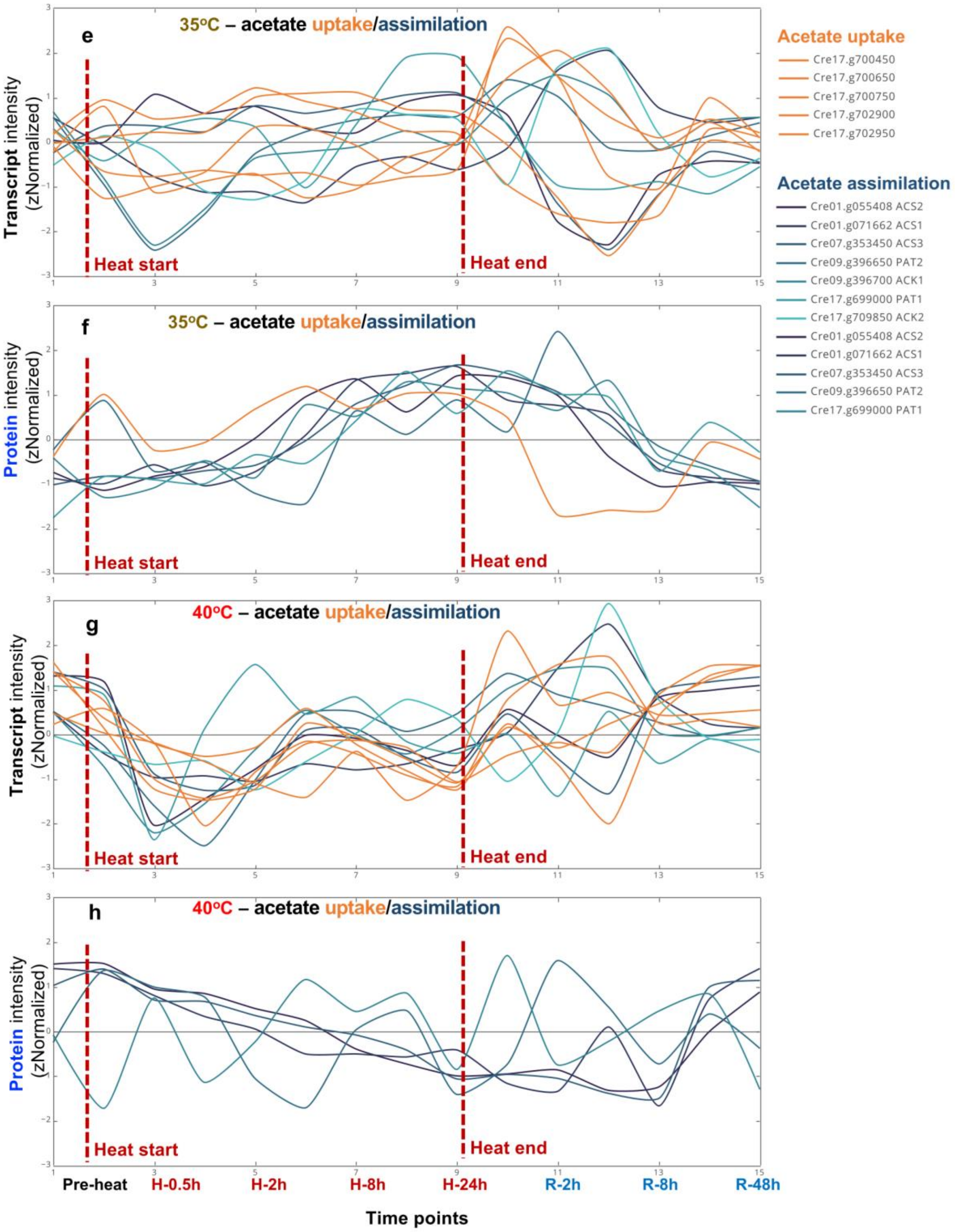
Multiple Cre identifier point to aminoacid.metabolism leading to a scatter plot for each functional term and window

By collecting the correlation coefficients of all functional terms, a density plot is created for each window

Supplementary Fig. 7: The fold-change correlation between transcripts and proteins were investigated. Transcript and protein correlation analysis by using Cre2.g222 (hypothetical gene ID as Cre identifier) of functional term aminoacid.metabolism as an example. Log$_2$(fold-changes, FC) of transcript reads and protein abundance are calculated in respect the pre-heat sample. The high temperature period (HS) as well as the recovery period (RE) are split into three windows each (HS1-3 and RE1-3). HS1-3 windows: 0-1 h, 2-8 h, 16-24 h during the heat period; RE1-3 windows, 0-2 h, 4-8 h, 24-48 h during the recovery period after heat treatment. Every identifier that has transcripts as well as proteins associated to it, results in a transcript-protein fold-change pair for each window. The average Log$_2$FC is determined for each transcript-protein pair in each window. By collecting all Cre identifiers that are associated with aminoacid.metabolism, a scatter plot of transcript-protein fold-change pairs is generated and the Pearson correlation coefficient is calculated. By repeating the workflow for all functional terms for each window, a density plot for each of the six windows is created to describe the overall correlation of transcript reads and protein abundance for that window.

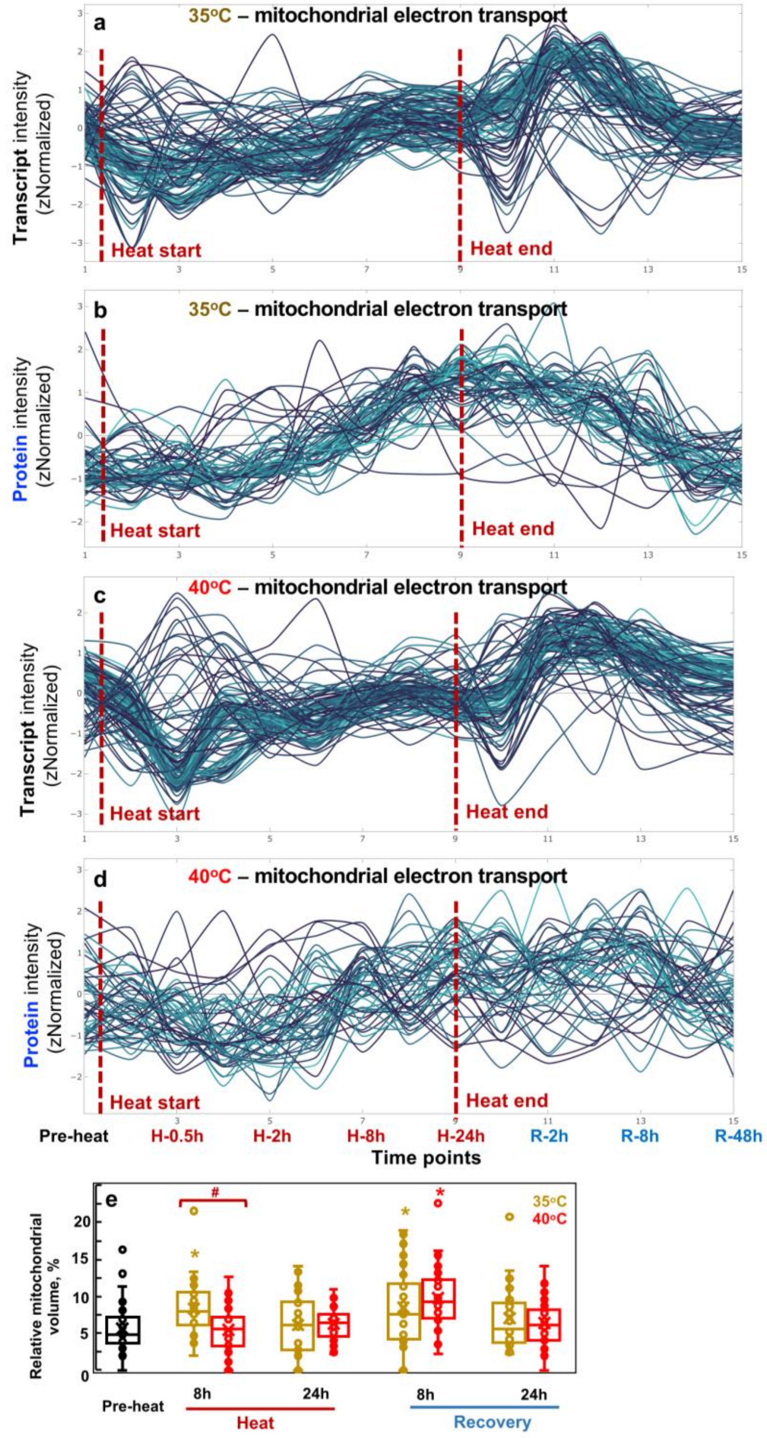
Supplementary Fig. 8





Supplementary Fig. 8. The kinetics of transcripts and proteins suggest gluconeogenesis/glyoxylate cycle and acetate uptake/assimilation increased during 35 °C but decreased during 40 °C heat. Transcript (a, c, e, g) and protein (b, d, f, h) signals related to the gluconeogenesis/glyoxylate cycle and acetate uptake/assimilation were standardized to z-scores (standardized to zero mean and unit variance) and are plotted against equally spaced time point increments. The red dashed lines indicate the start and end time of heat treatment for 35 °C (a, b, e, f) and 40 °C (c, d, g, h) respectively. Time points are labeled at the bottom. Timepoint 1: pre-heat. Time points 2-9, heat treatment at 35 °C or 40 °C, including reaching high temperature (0), 0.5, 1, 2, 4, 8, 16, 24 h during heat; time points 10-15, recovery phase after heat treatment, including reaching control temperature (0), 2, 4, 8, 24, 48 h during recovery. Genes involved in gluconeogenesis/glyoxylate cycle were based on MapMan function annotation; genes involved in acetate uptake/assimilation were manually curated based on Durante et. al. (2019) and Johnson et. al. (2013). See the interactive figures with gene IDs and annotations in Supplementary Data 10, gluconeogenesis_glyoxylate cycle.html. More information about genes involved in acetate uptake/assimilation can be seen in Supplementary Data 1 and 2 using the gene IDs on the figure.

Supplementary Fig. 15



Supplementary Fig. 15. Transcript/protein kinetics and TEM analysis suggested increased and reduced mitochondrial electron transport during 35 °C and 40 °C heat treatments, respectively. Transcript (a, c) and protein (b, d) signals related to the MapMan bin mitochondrial electron transport were standardized to z-scores (standardized to zero mean and unit variance) and are plotted against equally spaced time point increments. The red dashed lines indicate the start and end time of heat treatment for 35 °C (a, b) and 40 °C (c, d), respectively. Time points are labeled at the bottom. Timepoint 1: pre-heat. Time points 2-9, heat treatment at 35 °C or 40 °C, including reaching high temperature (0), 0.5, 1, 2, 4, 8, 16, 24 h during heat; time points 10-15, recovery phase after heat treatment, including reaching control temperature (0), 2, 4, 8, 24, 48 h during recovery. See the interactive figures with gene IDs and annotations in Supplementary Data 10, mitochondrial electron transport _ ATP synthesis.html. (e) Relative volume fractions of mitochondria were quantified using TEM images and Stereo Analyzer with Kolmogorov–Smirnov test for statistical analysis compared to the pre-heat condition (*, p<0.05, the colors of asterisks match treatment conditions) or between 35 °C and 40 °C at the same time point (#, p< 0.05). Each treatment had three biological replicates, total 30 images per treatment.

Article IV: Moderate high temperature is beneficial or detrimental depending on carbon availability in the green alga Chlamydomonas reinhardtii

Ningning Zhang*, Benedikt Venn*, Catherine E Bailey, Ming Xia, Erin M Mattoon, Timo Mühlhaus, Ru Zhang

| Journal of Experimental Botany, 2023, doi.org/10.1093/jxb/erad405

Aus rechtlichen Gründen ist es nicht möglich, den Artikel in der Dissertation abzudrucken.