Quantitative High-Resolution Analysis of Cell Cycle Regulated Protein and Phosphorylation Abundance in Non-Transformed Cells

By

Paul Anderson

A Master’s Thesis

Barts and The London School of Medicine and Dentistry

Queen Mary University of London

August 2025

Word count:

# Acknowledgements

I would like to express my sincere thanks to Doctor Tanya Soliman of the Barts Cancer Institute for providing the needed data for the project and for her helpful advice and supervision throughout.

Table of Contents

[Acknowledgements 2](#_Toc203464294)

[Abstract 3](#_Toc203464295)

[Introduction 4](#_Toc203464296)

[Aims and Objectives 5](#_Toc203464297)

[Materials and Methods 6](#_Toc203464298)

[Results 7](#_Toc203464299)

[Discussion 8](#_Toc203464300)

[Conclusion 9](#_Toc203464301)

[References 10](#_Toc203464302)

# Abstract

The cell cycle consists of a highly regulated sequence of events that result in the cell splitting into two daughter cells. It consists of several stages, each of which is covered by changes in protein and phosphorylation abundance.

Various papers have addressed changes in these abundance levels in cancerous cells. Here we analyse non-transformed cells to offer a baseline reference for these papers. We achieve this by analysing changes in abundance in proteomic and phosphoproteomic in TODO – copy from Tanya. with a higher greater temporal resolution than the bulk of other investigations. This has revealed various significant findings regarding proteins and phosphorylation events involved in the cell cycle, including oscillations

# Introduction

The reproduction of cells is achieved through a highly regulated sequence events known as the cell cycle [1]. In eukaryotic cells this process is highly complex and an area of ongoing study, having various aspects that still poorly understood. It consists of a sequence of discrete stages. The first and longest is G1, consisting of duplication of the majority of organelles (ribosomes, mitochondria etc.) and an increase in cell size [2]. Unless entering the quiescent G0 phase it then will proceed to the S phase and duplication of the chromosomes into two sister chromatids. The cycle then progresses to the G2 phase, a period of rapid growth, before proceeding to the highly complex M phase. The M (for mitosis) phase itself consists of five phases (prophase, prometaphase, metaphase, anaphase and telophase) culminating in splitting of the nucleus into two identical nuclei. This is then rapidly followed by fission of the cell during cytokinesis.

Progression through the cell cycle is controlled by interactions between a complex network of proteins and phosphorylation events. Protein abundances are tightly controlled by transcription, translation and degradation, with phosphorylation levels being regulated by kinases and phosphatases [3]. Amongst the most significant of these are cyclin dependent kinases (CDKs) [4] which are themselves subject to phosphorylation. Disruption of this intricate system can lead to cancer [5].

Since their discovery [6] the majority of investigations into cyclins have been low throughput, which whilst they have been highly revealing have left substantial areas uncovered. More recent investigations have addressed the issue of throughput using Mass Spectrometry [7], although the specific methods used are often not without drawbacks. They typically have not thoroughly investigated phosphorylation, an important process in the cell cycle [8]. They also often generally have low time resolution, not covering all the cell stages. Attempts at cell synchronisation often caused damage to the cells [9], and studies without synchronisation such as those using FUCCI examine a low number of cells which limits the amount of data that can be captured [10]. Palbociblib, a CDK4/CDK6 inhibitor that arrests cell development in stage G1, has proven effective and is used here along with RO3306, a CDK1 inhibitor that arrests cells at G2/M phase.

This study offers a high-resolution quantitative analysis of the proteome and phosphoproteome during cell cycle progression, incorporating data from previous research. We perform deep MS analysis at ten time points during the cell cycle, starting with TODO – explain what the various stages are about.

# Aims and Objectives

The aim of the study is to identify proteins and phosphorylation events relevant to the cell cycle process.

# Materials and Methods

TODO – rewrite these two paragraphs.

RPE1 cells were grown to 70% confluence before treatment with either Palbociclib (CDK4/6i) or RO3306 (CDK1i) for 18h. Cells were washed 3x in warm media before harvesting at the described timepoints. At these timepoints, cells were lysed, samples equalised to 1mg/mL total protein concentration and stored for transfer to the Centre for Genome Regulation Mass Spectrometry facility.

In your experiment, there were 30 samples from different phases of cell cycle. Samples were digested and their proteome was analysed in the Orbitrap Eclipse using a 90-min gradient with a DIA method (40 window of 10 Da). As a quality control BSA controls were digested in parallel and ran between each of your samples to avoid carryover and assess the instrument performance. Your samples were searched against SwissProt Human (July 2023) using DIA-NN (<https://www.nature.com/articles/s41592-019-0638-x#Sec1)>(v1.8). Peptides have been filtered based on FDR and only peptides showing an FDR lower than 1% have been retained. Find the information of the identified peptides and their corresponding proteins in the excel sheet of the results.

The resultant data consists of two spreadsheets of abundances, one for proteins and one for phosphopeptides. Each entry has readings for ten timepoints within the cell cycle. These data are analysed by software written in Python, Django, pytest, Postgres, Docker and D3.js.

TODO – rewrite? As changes in protein and phosphorylation levels drive the timing of the cell cycle, one focus of the study is on how these levels rise and fall during it. To find these patterns, we used a curve-fitting method along with ANOVA statistical tests. The curve-fitting finds the best-shaped line that matches the data over time and gives a fold-change score to show how much the levels increase or decrease. This method works well with time-based data, making it better at handling outliers and reducing random noise. As a result, it gives a more accurate picture of overall trends. We also consider the q-value, which is the false discovery rate (FDR) adjusted p-value. It estimates the proportion of false positives among all significant test results.

Relevant proteins were deemed to be those with a q value below 0.05 and a curve fold value over 1.2 (a 20% difference between the highest and lowest levels of a protein). Phosphopeptides were considered relevant with the same curve fold limit but a q value below 0.01. The value of 0.05 is a standard cutoff for q-values, analagous to the cutoff for p-values [11]. The lower value of 0.01 is more common for phosphopeptides as their identification and quantification is inherently more error prone due to lower abundances and site localisation ambiguity [12].

## Analysis steps

Quality control of proteins. All proteins with a protein id beginning with ‘CON\_’ were discarded.

Quality control of phosphopeptides. As with the proteins any protein id prepended with ‘CON\_’ was discarded, along with any phosphopeptide with a protein id not found in the spreadsheet of proteins.

The abundances for proteins and phosphopeptides for the ten stages and three replicates were read and inserted into the database. Any empty abundance cells were not entered.

After insertion of the data to the database the initial calculations were of medians, TODO – finish and means of the resultant data.

Calculate log2 means and various metrics (ANOVA and Fischer p value and q values TODO - confirm) for proteins and phosphopeptides.

Calculate protein and phosphopeptide oscillations.

Identify relevant proteins and phosphopeptides – those with low q values (TODO – confirm ideal q values) and high curve fold changes (> 1.2)

Classify by group, complex and subcellular localisation (TODO page 6 ‘Biological Role of cell-cycle dependent proteins – function written for CORUM).

Identify those of above that are high/low variability and oscillation.

Generated various charts from the data (Fig TODO)

SEE ALSO page 14, Defining cell cycle-dependent (CCD) proteins and phosphorylation sites

Outline how the curve fold threshold of 1.2 was determined, if done (Page 15, Protein curve fold change cut-offs were defined by quantifying theenrichment of the‘cell cycle’ term).

# Results

Protein quality control reduced the number of rows from 5416 to 5364. Elimination of phosphopeptides with no corresponding protein entry reduced the number of phosphopeptide rows from 34034 to 25847. Of the remaining rows 78% of the protein spreadsheet cells for readings were populated across the ten stages and three replicates, and 28% of the phosphopeptide cells for readings. All other cells for readings were empty. Spreadsheet cells that were empty were ignored, not considered to be zero, as declaring so large a quantity of empty cells to be zero resulted in medians of zero which are unusable. With row removal this reduced the total amount of usable data from the spreadsheets to 77% for proteins and 21% for phosphopeptides. The low quantity of data reduced the quality of results, most significantly for phosphopeptides – TODO – link to comparison heatmaps for ICR. TODO – link to heatmaps for SL and ICR for comparison

The 5,364 proteins quantified represent about 26% of the human proteome. Of these 367 are associated with the Genome Ontology (GO) term ‘cell cycle’. The 25,847 phosphopetides are from 2,634 proteins, 13% of the human proteome, with 232 having the ‘cell cycle’ GO term. This is a lower range of cell cycle regulators than Rega et al. TODO – add other studies - but with higher temporal resolution and so represents a detailed analysis of proteins and phosphorylation changes within the cell cycle.

List of relevant oscillation proteins

List of relevant oscillating phosphopeptides

Where the phosphorylation sites are and whether they’re disordered or accessible (end of page 8).

# Discussion

Page 12

# Conclusion

# Additional Information

# References