Mol Cell Proteomics Papers in Press. Published on July 22, 2019 as Manuscript RA119.001472

Multi-batch TMT: false positives, missing values and batch effects

Multi-batch TMT reveals false positives, batch effects and missing values

Alejandro Brenes[1], Jens Hukelmann[1,2], Dalila Bensaddek[1,3], & Angus I. Lamond[1]*

[1]Centre for Gene Regulation and Expression, School of Life Sciences, University of Dundee, Dow St, Dundee, DD1 5EH, United Kingdom

[2]Present Address: Immatics, Paul-Ehrlich-Str. 15, Tuebingen, 72076, Germany

[3]Present Address: Biosciences Core labs, Proteomics, King Abdullah University of Science

and Technology, Thuwal 23955-6900, Saudi Arabia

[*]corresponding author

Multi-batch TMT: false positives, missing values and batch effects

Abstract

Multiplexing strategies for large-scale proteomic analyses have become increasingly prevalent, Tandem Mass Tags (TMT) in particular. Here we used a large iPSC proteomic experiment with twenty-four 10-plex TMT batches to evaluate the effect of integrating multiple TMT batches within a single analysis. We identified a significant inflation rate of protein missing values as multiple batches are integrated and show that this pattern is aggravated at the peptide level. We also show that without normalisation strategies to address the batch effects, the high precision of quantitation within a single multiplexed TMT batch is not reproduced when data from multiple TMT batches are integrated.

Furthermore, the incidence of false positives was studied by using Y chromosome peptides as an internal control. The iPSC lines quantified in this dataset were derived from both male and female donors, hence the peptides mapped to the Y chromosome should be absent from female lines. Nonetheless, these Y chromosome-specific peptides were consistently detected in the female channels of all TMT batches. We then used the same Y chromosome specific peptides to quantify the level of ion co-isolation as well as the effect of primary and secondary reporter ion interference. These results were used to propose solutions to mitigate the limitations of multi-batch TMT analyses. We confirm that including a common reference line in every batch increases precision by facilitating normalisation across the batches and we propose experimental designs that minimise the effect of cross population reporter ion interference.

Introduction

High-throughput, shotgun proteomics, using data dependent acquisition (DDA), now enables the comprehensive study of proteomes, allowing the identification of 10,000 or more proteins from cells and tissues(1-3). However, to achieve such deep proteome coverage using DDA,

extensive prefractionation of extracts prior to mass spectrometry (MS) analysis is frequently required(1,4). To evaluate statistically the significance of the resulting data, a minimum of 3 replicates for each sample/condition is also necessary(5,6). The data acquisition time involved is increased still further for experiments that analyse the multi-dimensional characteristics of the proteome; e.g. studying differences in protein subcellular localisation, turnover rates, post-translational modifications (PTMs) and protein-protein interactions(7-9).

To cope with the challenges of large-scale proteomics analyses, strategies have been developed to allow multiple samples to be analysed in parallel, through multiplexing isotopically tagged peptides (10,11). The most widely used MS multiplexing methods, TMT(12) and iTRAQ(13), use isobaric tags for simultaneous peptide identification and quantification. TMT in particular has increased in popularity and is now widely used (14,15). This reflects the ability of multiplexed TMT to increase sample throughput in proteomics studies and reduce the "missing values" problem that arises from the stochastic sampling inherent in DDA proteomics (16,17). Thus, within a single multiplex TMT batch, the number of missing values at the protein level is low, frequently <2%(14). Furthermore, the precision of the quantification within a multiplex TMT batch is high(18). However, it is less clear how well multiplexed TMT performs for very large-scale analyses, involving numerous TMT batches.

In this manuscript, we use a proteomic dataset of human iPSC cells, involving 24 separate 10-plex TMT batches(19). We compare the quantitation of data both within and between 10-plex batches and focus our analysis on 4 main issues: (i) missing values, (ii) accuracy of quantification, (iii) false positives and (iv) the effect of both reporter ion interference (RII) and co-isolation interference (CII).

We show there is an inflationary effect on missing values as data from multiple batches are integrated both at the protein and peptide level. We evaluated reproducibility both by studying the coefficient of variation (CV) within each 10-plex TMT batch, and by comparing a reference

line (technical replicates of the iPSC line "bubh_3") that were common to every batch. Furthermore, the incidence of false positives was studied by using Y chromosome peptides as an internal control. The iPSC lines quantified in this dataset were derived from 163 different donors including both male and female, hence the peptides mapped to the Y chromosome should be absent from female lines. Nonetheless, we confirm that these Y chromosome-specific peptides were consistently detected in the female channels of all TMT batches. Finally, by using these Y chromosome peptides, we quantified the effect of ion co-isolation and reporter ion interference upon TMT quantification accuracy.

Experimental procedures

Experimental design and statistics rationale

The study consists of 240 iPSC replicates, 217 biological replicates and 24 technical replicates, derived from 163 different donors. The study comprises twenty-four 10-plex TMT batches. Each batch consisted of 1 common reference line (technical replicates of iPSC cell line "bubh_3") and 9 different iPSC cell lines. The technical replicates were used for the data normalisation strategy described below. Out of the 240 replicates analysed, 142 were derived from female donors and from 98 male donors.

TMT Sample preparation

For protein extraction, iPSC cell pellets were washed with ice cold PBS and redissolved immediately in 200 μL of lysis buffer (8 M urea in 100 mM triethyl ammonium bicarbonate (TEAB)) and mixed at room temperature for 15 minutes. Cellular DNA was sheared using ultrasonication (6 X 20 s on ice). The proteins were reduced using tris-carboxyethylphosphine TCEP (25 mM) for 30 minutes at room temperature, then alkylated in the dark for 30 minutes using iodoacetamide (50 mM). Total protein was quantified using the

Multi-batch TMT: false positives, missing values and batch effects

EZQ assay (Life Technologies). For the first digestion with mass spectrometry grade lysyl endopeptidase, Lys-C (Wako, Japan), the lysates were diluted 4-fold with 100 mM TEAB then further diluted 2.5-fold before a second digestion with trypsin. Lys-C and trypsin were used at an enzyme to substrate ratio of 1:50 (w/w). The digestions were carried out overnight at 37ºC, then stopped by acidification with trifluoroacetic acid (TFA) to a final concentration of 1% (v:v). Peptides were desalted using C18 Sep-Pak cartridges (Waters) following manufacturer's instructions.

For tandem mass tag (TMT)-based quantification, the dried peptides were re-dissolved in 100 mM TEAB (50 µl) and their concentration was measured using a fluorescent assay (CBQCA, Life Technologies). For each 10-plex TMT batch 100 µg of peptides from each cell line to be compared, in 100 µl of TEAB, were labelled with a different TMT tag (20 µg ml^{-1} in 40 µl acetonitrile) (Thermo Scientific), for 2 h at room temperature. After incubation, the labelling reaction was quenched using 8 µl of 5% hydroxylamine (Pierce) for 30 min and the different cell lines/tags were mixed and dried *in vacuo*.

The TMT samples were fractionated using off-line, high-pH reverse-phase (RP) chromatography: samples were loaded onto a 4.6 × 250 mm Xbridge BEH130 C18 column with 3.5-µm particles (Waters). Using a Dionex bioRS system, the samples were separated using a 25-min multistep gradient of solvents A (10 mM formate at pH 9) and B (10 mM ammonium formate pH 9 in 80% acetonitrile), at a flow rate of 1 ml min^{-1}. Peptides were separated into 48 fractions, which were consolidated into 24 fractions. The fractions were subsequently dried and the peptides re-dissolved in 5% formic acid and analysed by LC–MS/MS.

Multi-batch TMT: false positives, missing values and batch effects

TMT LC–MS/MS

TMT-based analysis. Samples were analysed using an Orbitrap Fusion Tribrid mass spectrometer (Thermo Scientific), equipped with a Dionex ultra-high-pressure liquid-chromatography system (RSLCnano). RPLC was performed using a Dionex RSLCnano HPLC (Thermo Scientific). Peptides were injected onto a 75 μm × 2 cm PepMap-C18 pre-column and resolved on a 75 μm × 50 cm RP- C18 EASY-Spray temperature-controlled integrated column-emitter (Thermo Scientific), using a four-hour multistep gradient from 5% B to 35% B with a constant flow rate of 200 nl min^{-1}. The mobile phases were: 2% ACN incorporating 0.1% FA (solvent A) and 80% ACN incorporating 0.1% FA (solvent B). The spray was initiated by applying 2.5 kV to the EASY-Spray emitter and the data were acquired under the control of Xcalibur software in a data-dependent mode using top speed and 4 s duration per cycle. The survey scan is acquired in the orbitrap covering the *m/z* range from 400 to 1,400 Thomson with a mass resolution of 120,000 and an automatic gain control (AGC) target of 2.0×10^5 ions. The most intense ions were selected for fragmentation using CID in the ion trap with 30% CID collision energy and an isolation window of 1.6 Th. The AGC target was set to 1.0×10^4 with a maximum injection time of 70 ms and a dynamic exclusion of 80 s, the scan rate was set to 'Rapid'.

During the MS3 analysis for more accurate TMT quantifications, 5 fragment ions were co-isolated using synchronous precursor selection with a window of 2 Th and further fragmented using HCD collision energy of 55%. The fragments were then analysed in the orbitrap with a resolution of 60,000. The AGC target was set to 1.0×10^5 and the maximum injection time was set to 105 ms.

Machine, blanks and standards

All of the TMT batches were analysed on the same Orbitrap Fusion MS instrument. Between each individual TMT experiment, one blank was run, followed by analysis of a 15 peptide

6

Multi-batch TMT: false positives, missing values and batch effects

Retention Time Calibration (RTC) standard, to evaluate retention time drift. This was followed by analysis of an MCF10a total cell digest standard to evaluate peptide and protein identifications. The last step consisted of analysis of two blanks, one with an oscillating gradient and one with the gradient matching the samples to be run.

Identification & Quantification

The data from all twenty-four 10-plex TMT batches were collected and analysed simultaneously, using Maxquant (20,21) v. 1.6.3.3. The FDR threshold was set to 1% for each of the respective Peptide Spectrum Match (PSM) and Protein levels. The data was searched with the following parameters; type was set to Reporter ion MS3 with 10plex TMT, stable modification of carbamidomethyl (C), variable modifications, oxidation (M), acetylation (protein N terminus), deamidation (NQ), Glutamine to pyro-glutamate (N-terminus), with a 2 missed tryptic cleavages threshold, reporter mass tolerance set to 0.03 ppm. Minimum peptide length was set to 7 amino acids. Proteins and peptides were identified using UniProt (SwissProt December 2018). The run parameters are accessible at ProteomeXchange(22) via the PRIDE repository (23), along with the full MaxQuant(20) quantification output (PDX010557).

Filtering

All proteins that were marked as 'Reverse', 'Potential Contaminants' or 'Only identified by site' were discarded. The final subset comprised 9,640 proteins. Peptides marked as 'Potential contaminants' or 'Reverse' were also filtered from the analysis. The final peptide dataset comprised 178,491 peptides.

7

Copy number generation

Protein copy numbers were calculated following the proteomic ruler approach(24). For protein, p, $uCN_{b,c,p}$ is the uncorrected protein copy number:

$$uCN_{b,c,p} = protein\ MS3\ signal_{b,c,p} \times \frac{A}{M_p} \times \frac{6.85 * 10^{-12}}{\sum_{h \in b,c} histones\ MS3\ signal_h}$$

$$for\ batch\ b\ \in \{1,2,\dots24\}\ and\ channel\ c\ \in \{126C, 127N, \dots, 131N\}$$

where A is Avogadro's constant, M_p is the molar mass of the protein p, protein MS3 signal is the protein MS3 intensity and histones MS3 signal is the MS3 intensity for all histones, h.

These uncorrected copy numbers, which will be referred to here as "raw copy numbers", were used to study the coefficient of variation (CV). To control for technical variation between the 24 different 10-plex batches, a correction factor, cf, was applied to every protein, p, in every batch, b, to adjust the protein copy numbers.

$$cf_{b,p} = \frac{uCN_{b,126C,p}}{\sum_b uCN_{b,126C,p}/24}$$

$$for\ b\ \in \{1,2,\dots24\}$$

where $uCN_{b,126C,p}$ is the protein copy number derived from reporter channel 126C (the reference channel). The normalised copy number, $normCN$, is calculated for protein, p in all batches, b, and all channels c:

$$normCN_{b,c,p} = uCN_{b,c,p} \times cf_{b,p}$$

$$for\ batch\ b\ \in \{1,2,\dots24\}\ and\ channel\ c\ \in \{126C, 127N, \dots, 131N\}$$

8

Multi-batch TMT: false positives, missing values and batch effects

Missing value calculations

First, to estimate missing values within this DDA analysis, a list of unique proteins/peptides that were detected with at least 1 reporter intensity greater than zero were calculated for each batch. To determine the number of missing values within each 10-plex TMT batch, the number of unique proteins/peptides per reporter channel was compared to the number of unique proteins/peptides identified within the batch. This approach was applied to generate the missing value calculations for each of the 24 individual 10-plex TMT batches.

To assess the effect of integrating multiple TMT batches, random sampling was performed to estimate how missing values are affected by a progressive increase in the number of 10-plex TMT batches analysed. This was performed in an incremental fashion starting from 2 and finishing with 22 batches (PT6388 was not used for this analysis) , with 500 iterations per level. At the first level 2 batches would be selected at random with no replacement, and at the last level 22 batches would be selected at random, again with no replacement. This was performed with the R function "sample()" part of the base R-core package.

At each level a new list of proteins/peptides detected with at least 1 reporter ion intensity greater than zero within any of the integrated TMT batches was calculated, and the number of proteins/peptides with intensity greater than 0 per reporter channel was evaluated against the new list.

Coefficient of variation

The coefficient of variation (CV) in protein abundance levels was calculated using the log10 transformed protein copy numbers.

$$CV = \frac{S}{X} \times 100$$

For each protein the CV is equal to the copy number standard deviation (S) divided by the mean copy number (X) times 100. The protein CV within each 10-plex TMT batch was

calculated for all 10 cell lines within the same batch, using all proteins detected in every reporter channel. The reference line CV was calculated using proteins that were detected in the TMT[10] -126C (reference line) channel across all of the 24 10-plex TMT batches.

Correlation Clustering

For each 10-plex TMT batch, a concordance correlation value was calculated for all cell lines within the same batch. The calculations were performed using "correlation()" function from the R package "agricolae" version 1.2.8.

The same process was applied to calculate the concordance correlation values for the reference lines, i.e. using reporter channel 126C in all TMT batches.

Peptide intensity normalisation

The replicate normalised intensity, rni, was calculated per peptide, q:

$$rni_q = log_{10}(\frac{peptide\ MS3\ signal_{b,c,q}}{median(I_{b,c})})$$

$$I_{b,c} = \{peptide\ MS3\ signal_{b,c,q} : \forall q\}\ given\ batch, b\ and\ channel, c$$
$$for\ batch\ b\ \in \{1,2,\dots 24\}\ and\ channel\ c\ \in \{126C, 127N, \dots, 131N\}$$

The median normalised intensity, mni, for peptide, q, is the median of all batches, b, and channels, c:

$$mni_q = median(rni_{b,c,q})$$

$$for\ all\ batches\ b\ \in \{1,2,\dots 24\}\ and\ channels\ c\ \in \{126C, 127N, \dots, 131N\}$$

The *global median* is the median of *mni* for all peptides, q:

$$global\ median = median(mni_q)$$

Reporter ion interference classification

The reporter ion interference (RII) targets are based on a typical product data sheet for 10-plex TMT Label Reagents from ThermoFisher Scientific, as summarised in the table below:

10

Multi-batch TMT: false positives, missing values and batch effects

Mass Tag	Reporter channel	-1Da (secondary RII)	+1Da (primary RII)
TMT^{10}-126	0	--	127C
TMT^{10}-127N	1	--	128N
TMT^{10}-127C	2	126	128C
TMT^{10}-128N	3	127N	129N
TMT^{10}-128C	4	127C	129C
TMT^{10}-129N	5	128N	130N
TMT^{10}-129C	6	128C	130C
TMT^{10}-130N	7	129N	131
TMT^{10}-130C	8	129C	--
TMT^{10}-131	9	130N	--

Analysis of reporter ion interference

To study the effect of reporter ion interference across different TMT channels, we selected a subset of 69 peptides that were specific to the following list of protein coding genes uniquely located on the Y chromosome; "CDY1", "CDY2A", "DDX3Y", "EIF1AY", "KDM5D", "NLGN4Y", "PCDH11Y", "RPS4Y1", "TBL1Y", "USP9Y" & "UTY".

This approach of using peptide values from Y chromosome specific genes depends upon there being a diverse mixture of male and female donor-derived iPSC lines in each 10-plex TMT batch. However, two of the 24 TMT batches comprised exclusively female donor-derived iPSCs, which had been shown not to have Y chromosome derived DNA in QC analyses(25). For these female donor-specific batches, any peptide assigned to Y chromosome specific genes was excluded from the analysis. An additional batch, PT6388, displayed an irregular behaviour, and was hence also discarded from the posterior analysis. A final subset of 65 Y chromosome-specific peptides were used for this analysis (see supplemental data for list).

Peptide male vs reporter ion interference ratios

Multi-batch TMT: false positives, missing values and batch effects

The peptide ratios comparing male channels vs female channels, mpr, subjected to different reporter ion interference conditions, $cond$, were calculated per 10-plex TMT batch, b, for peptide, q, utilising the replicate normalised intensities:

$$mpr_{b,q} = \frac{median(RNI_{b,male,q})}{median(RNI_{b,cond,q})}$$

$$RNI_{b,male,q} = \{rni_{b,c,q} : \forall\ male\ channels, c\},$$

$$RNI_{b,cond,q} = \{rni_{b,c,q} : \forall\ cond\ channels, c\},$$

$$for\ b\ \in \{1,2,\dots 24\}\ and\ cond\ \in \{primary\ RII, secondary\ RII, double\ RII, no\ RII\}.$$

The box plot comparing male replicates to the different reporter ion interference conditions used these peptide batch ratios and was plotted using "ggplot2" version 3.0.0(26).

Peptide reporter ion interference ratios

The peptide ratios, npr, comparing different reporter ion interference (RII) conditions, $cond$, in female channels to female channels with no RII, $noRII$, were calculated within each 10-plex TMT batch, b, for peptide, q, utilising the replicate normalised intensities.

$$npr_{b,q} = \frac{median(RNI_{b,cond,q})}{median(RNI_{b,noRII,q})}$$

$$RNI_{b,cond,q} = \{rni_{b,c,q} : \forall\ cond\ channels, c\},$$

$$RNI_{b,noRII,q} = \{rni_{b,c,q} : \forall\ noRII\ channels, c\},$$

$$for\ b\ \in \{1,2,\dots 24\}\ and\ cond\ \in \{primary\ RII, secondary\ RII, double\ RII\}.$$

These results were stratified by the global median, where peptides with median normalised either intensity greater than or equal to the global median were considered 'High intensity' and those lower than the global median were considered 'Low intensity'. The box plot comparing

12

different reporter ion interference conditions to the replicates not affected by reporter ion interference used these peptide batch ratios and was plotted using "ggplot2" version 3.0.0(26).

Results

Missing values in TMT

A known advantage of using TMT is the low index of missing values that are present within a single TMT batch. Recent studies report as low as <1% missing values at the protein level (18), albeit data are usually not reported at the peptide level.

We started by analysing the iPSC 10-plex TMT data for the number of missing values at the protein level within each TMT batch (Fig. 1a). The preliminary results are consistent with previous reports, i.e. 92% of the 24 different 10-plex TMT batches show <1% missing values at the protein level, with only 1 outlier with missing protein values >1.5%. This was experiment PT6388, which is highlighted in red (Fig. 1a&b). Furthermore, when we analyse the data at the peptide level, there is very close agreement to the protein data, with 92% of the 24 different 10-plex TMT batches having <5% missing peptide values, however the outlier batch has an exacerbated effect and displays 9% missing values at the peptide level. We therefore excluded PT6388 from the rest of the analysis.

These previous results do not address the effect of integrating data from multiple, independent 10-plex TMT batches into a single analysis. To study the effect of data integration, we increased the number of batches selected, from 2 to 22 and recalculated the number of missing values that were present (Fig. 1c&d; see methods). At the protein level, the median number of missing values increases from 0.19% with one 10-plex TMT batch, to 6.35% when data from a second 10-plex TMT batch were integrated (Fig. 1c). When we integrate data from 5 different

13

Multi-batch TMT: false positives, missing values and batch effects

10-plex TMT batches, the median number of missing values at the protein level expanded to ~10%.

This situation was exacerbated when the analysis was performed at the peptide level (Fig. 1d). When integrating data from just two 10-plex TMT batches, the median number of missing peptide values was >23%. Even more striking, it only required integrating data from 5 different 10-plex TMT batches to produce ~40% missing values at the peptide level. The data suggest peptides are not reproducibly detected among batches, but is it only low abundance peptides? Based upon these results, we decided to perform a more in-depth analysis on the inflation rate of peptide missing values. We observed that the number of peptides identified within each 10-plex TMT batch is relatively constant (Fig. 2a), but quite variable across different batches. The median number of peptides identified per batch was 84,046 with a standard deviation of 11,354. To further analyse these peptide level data, we first median-normalised the MS3 intensities for all peptides in all cell lines (see methods). The $\log_{10}$ median normalised MS3 intensities spanned 6 orders of magnitude (see methods; Fig. 2b).

We next analysed the peptide dataset by quartiles, based on the $\log_{10}$ median normalised intensity values (Fig. 2c). The first quartile represented the 25% least abundant peptides and the fourth quartile the 25% most abundant peptides. There are only 11 peptides within the first quartile that are detected in all TMT channels and only 603 peptides that are seen in > 90% of the TMT channels. As DDA selects the n most abundant ions reaching the mass-spectrometer during a MS1 scan(27) (where n typically is 10-30), this bias is predictable. However, when we analyse the results from the fourth quartile (25% most abundant peptides), we see that 26% of these peptides are still only detected in <50% of the TMT channels. These high intensity peptides that are detected in less than half of all channels have a median normalised intensity of 0.32, representing the 84th percentile of abundance. Furthermore only 24% of the peptides from the 4th quartile were detected in all the TMT channels. In total there are 12,140 peptides

that were detected in all TMT channels, regardless of intensity classification, which represents 6.81% of all peptides.

Next, we analysed the data by comparing the identification quartiles, organised by the percentage of TMT channels in which they were detected (Fig. 2d). The first quartile represented the 25% of peptides that were detected least frequently, i.e. in less than 29 TMT channels (see supplemental data). Of these peptides >29% had a median normalised intensity higher than the global median (see methods; -0.184), highlighting that even relatively abundant peptides are not identified consistently. Overall, ~50% of peptides are detected in <40% of all TMT channels.

Variation between 10-plex TMT batches

Multiple studies have documented TMT as a method producing precise quantitation, in some cases having a coefficient of variation (CV) ~3x lower than comparable label free data(18). Most of these studies have focused on analysing quantitative precision within a single TMT batch, and do not explore the effect of integrating data from multiple TMT batches into one analysis. However, projects involving large scale proteomic analyses of multiple cell lines and/or conditions, need to employ multiple TMT batches in a single experiment(10).

We calculated protein copy numbers for 230 iPSC replicates, which included 208 biological replicates and 23 technical replicates of a control iPSC line (bubh_3), across 23 separate 10-plex TMT batches (PT6388 is once again removed from the analysis). We then proceeded to calculate Lin's concordance correlation coefficient, which measures the agreement between two variables and is proposed to evaluate reproducibility(28). This was done for every iPSC line within each TMT 10-plex batch, and for all the technical replicates of the control line, channel TMT[10] 126, across the 23 10-plex TMT batches.

The concordance correlation coefficient within each 10-plex TMT batch is very high, (median value 98% concordance), highlighting the precision of the quantitation within each single

15

batch. However, when the same calculation is applied to the technical replicates of the control iPSC line across 23 respective batches, the median concordance coefficient drops to 81%.

To explore this situation further, we calculated the CV for the $\log_{10}$ transformed protein copy numbers[29] (see methods), both within each 10-plex TMT batch, and across 23 controls (Fig. 3a). When we calculated the protein CV exclusively within each 10-plex TMT batch, the median was 1.72 with all 10-plexes showing a median protein CV <2.5. Accordingly, the data show that for every batch, proteins with a CV >7.5 were considered outliers (Fig. 3a). These data show high precision of quantitation within each individual multiplexed experiment.

Similarly, to evaluate accuracy we calculated the CV for all technical replicates of the reference iPS cell line, bubh_3, which were analysed in channel TMT[10] 126 in every 10-plex TMT batch. The median CV of all the proteins detected in the technical replicates, was 11.03, i.e. 6.4-fold higher than the median within-batch CV. The CV of over 80% of all proteins in the technical replicates would be considered outliers in all of the 10-plex TMT analyses.

It is commonly assumed in proteomics studies that variation predominantly affects low intensity proteins and peptides. We decided to test this assumption and focused on the extremely abundant proteins. We focused the analysis on the 23 reference line replicates, and selected the top 100 most abundant proteins where the CV was greater than 7.5 and created box plots based on their copy numbers across all reference line data (Fig. 3b).

We chose to highlight 5 cases; 60S ribosomal protein L35a (RPL35A; highlighted in blue), Histone H1.2 (HIST1H1C; highlighted in green), ATPase inhibitor, mitochondrial (ATP5IF1; highlighted in grey), Peptydyl-prolyl cis-trans isomerase H (PPIH; highlighted in purple) and Glutathione S-transferase omega-1 (GSTO1; highlighted in pink). RPL35A ranges in expression from ~16,786,000 to ~185,000 copies and was identified with 10 unique + razor peptides (URP). HIST1H1C ranges from ~4,847,000 to ~43,000 copies and was identified with 6 URP, ATP5IF1 from ~4,825,000 to ~84,000 copies and was identified with 6 URP, PPIH

16

from ~4,180,000 to ~45,000 copies and was identified with 14 URP and GSTO1 from ~3,920,000 to ~115,000 and was identified with 16 URP. All of these 5 proteins are highly abundant, with a median copy number > 500,000 within the 23 technical replicates of the reference line, have been identified with >6 URP and yet they vary drastically between the different multiplexed experiments, highlighting that the variation is not limited to low abundance proteins.

We also note that while the majority of iPSC lines in this study come from healthy donors, some of these TMT batches, e.g. PT6390, contain mixtures of iPSC lines derived from both healthy donors and donors with rare genetic diseases, including "Usher syndrome", "Monogenic Diabetes" and "Bardet-Biedl syndrome". Nonetheless, the median protein CV within PT6390 is still ~10 fold lower than the CV obtained from analysing the 23 technical replicates of bubh_3, indicating that TMT batch effects have a bigger influence on the proteomics data than a healthy vs diseased physiology.

Our results highlight that while multiplex TMT is a useful and precise methodology for quantitative proteomics, it is important to be aware also of its potential limitations, particularly when analysing data from multiple TMT batches. It is important to note that copy numbers already provide a layer of normalisation(24), however it proves insufficient to deal with batch variation. These findings underline that when conducting very large-scale proteomics analyses across multiple TMT batches, it is essential to be aware of the potential for batch variation to affect data quality. To reduce the batch effect several methods have been proposed, from protein expression inference(30), to a standard reference line(31). Here we used the technical replicates of a reference iPSC line as an internal reference standard to control for variation between batches (see methods). Using this normalisation method provided a median CV of 2.96% across all cell lines and technical replicates, making the results closer to the metrics obtained for each individual within-batch analysis.

Multi-batch TMT: false positives, missing values and batch effects

False positives

The iPSC dataset(19,25) provided us with an excellent opportunity to study the incidence of false positives within a multi-plex TMT batch. The study utilised iPSC lines derived from both male and female donors within twenty-two of the twenty-four 10-plex TMT batches analysed here. Since only the lines from male donors should include proteins encoded by genes exclusively on the Y chromosome, this provided effectively a set of endogenous "spike-in" peptides, which we could use to monitor the presence of false positives as well as exploring the effects of reporter ion interference (RII) between TMT channels and co-isolation interference (CII).

The dataset detected 11 proteins that were mapped specifically to the Y chromosome. Correspondingly, all unique peptides derived from these proteins should only be present in the TMT channels with male cell lines and, in theory, should be absent in the TMT channels with female cell lines. To avoid mismatches arising from shared peptides, we focussed our analysis on a subset of 69 peptides that mapped uniquely to the following Y chromosome specific genes; "CDY1", "CDY2A", "DDX3Y", "EIF1AY", "KDM5D", "NLGN4Y", "PCDH11Y", "RPS4Y1", "TBL1Y", "USP9Y", "UTY". Additionally, since two of the 10-plex TMT batches analysed (PT6384 and PT7422) had only female cell lines, any potential Y chromosome-specific peptides that were detected in these batches were treated as wrongly annotated and discarded from further analysis. Furthermore, batch PT6388 was also considered an outlier and not included for further analysis. As a result, we focussed on 65 unique, Y chromosome encoded peptides that were used as "male-specific" spike-in references for the analysis of false positives within the previously mentioned 21 TMT batches.

We evaluated false positives by exploring how frequently the respective female TMT channels were quantifying signal from Y chromosome-specific peptides (Fig. 4). Surprisingly, this showed that in all twenty-one 10-plex TMT batches considered here and in all reporter channels

18

containing a female cell line, a minimum of over 40% of the Y chromosome-specific peptides identified within the batch also had signal in the female channels. Remarkably, across all these batches, a median of 89% of Y chromosome-specific peptides quantified in each batch were quantified in TMT channels that contained female cell lines. The Y chromosome peptides should not be present in any female line, hence the level of detection mentioned earlier is unexpected. We infer that the appearance of signal for Y chromosome-specific peptides in the channels containing female cell lines likely results from a combination of co-isolation and reporter ion interference.

Reporter ion interference and co-isolation interference

Reporter ion interference, also known as cross-label isotopic impurity, can arise from manufacture level impurities and experimental error(32). Co-isolation interference is the effect caused by multiple labelled peptides being selected within the isolation window(33). To study both of these conditions we focused on the previously mentioned Y chromosome-specific peptides, as these should only be present in the male channels and absent in the female channels, therefore any signal detected in female channels should be artificial. We used the Y chromosome peptides to evaluate the difference in replicate normalised peptide intensities (see methods) between male and female lines, across all the twenty-one 10-plex TMT batches (Fig. 5a).

 The results of the analysis revealed significant variation between TMT batches. For example, some batches, such as PT6379 and PT6386, have 17 and 65-fold difference in intensity between male and female channels, simplifying the detection of false positives due to co-isolation interference. However, other batches, e.g. PT7430 and PT6391, only show a 2.5 and 4.4-fold difference, respectively making the detection of the false positives problematic. We note both

of the previously mentioned batches display low intensity peptides and hence low signal-to-noise ratios, making female channels more vulnerable to co-isolation interference.

To evaluate co-isolation ion interference, we selected female channels with no primary or secondary reporter ion interference (no cross-label isotopic impurity; see methods), as likely examples of co-isolation interference(34). Peptides in male channels show a median of 9.43-fold higher replicate normalised intensity (see methods) compared to female channels not affected by reporter ion interference. However, the effects vary depending on the peptide intensity thresholds. High intensity peptides, where the median normalised peptide intensity across male lines was greater than or equal to the global median (-0.184; see methods), displayed 12.8-fold higher intensities. Low intensity peptides, where the median normalised peptide intensity of male lines was lower than the global median, only displayed 2.14-fold increased intensity, revealing higher vulnerability to co-isolation interference.

We also examined the potential effects of reporter ion interference (RII). For this analysis, we calculated a peptide specific ratio for each condition (see methods), firstly comparing the male channels vs the different types of reporter ion interference present in female channels. These conditions were, 'primary RII' when a male channel affects by isotopic impurity the +1Da female channel of the same isotope (male 126C to female 127C). 'secondary RII' is when a male channel affects by isotopic impurity the -1Da female channel of the same isotope (male 127C to female 126C). 'double RII' is when a female channel is affected by both 'primary RII' and 'secondary RII' from male channels. Channels not affected by either primary, or secondary RII were labelled as 'no RII'. The ratios comparing males to the previous conditions were used to generate the box plot (Fig 5b).

The male lines were a median of 9.4 fold higher than female channels not affected by reporter ion interference ("no RII"), but only 4.9 fold higher than female channels subjected to primary and secondary reporter ion interference ("double RII"). The smallest effect was caused by

"secondary RII", where the male lines were 8.94 fold higher than the female lines, suggesting that "primary RII" is the main source of isotopic impurities and explaining why we see little difference between the "primary RII" and "double RII" conditions. We note that for both "primary RII" and "double RII" the false positives are within the 8-fold increase/decrease range for *bona fide* changes in protein/peptide expression levels often detected within proteomic datasets(35,36).

To quantify the differences between primary and secondary reporter ion interference across peptide abundance categories, we now compared female channels affected by reporter ion interference ("primary RII", "secondary RII" and "double RII") to female channels not affected by reporter ion interference (Fig. 5c; see methods). We also stratified this analysis by the median normalised peptide intensity, with high intensity values being either greater than or equal the global median, and low intensity being lower than the global median. Within the analysis we confirmed that the smallest effect was caused by "secondary RII"; for high intensity Y chromosome-specific peptides it displayed only a 1.05-fold increase compared to the channels not affected by reporter ion interference and virtually no change in low abundance peptides (Fig 5c). The "primary RII" displayed a more pronounced effect, with a median increase of 1.57 fold in the high intensity peptides and 1.10 fold in the low intensity ones. The combination of primary and secondary RII produced a median of 1.7-fold increase in the high intensity peptides and a small reduction in the median for the low intensity peptides.

These results suggest low intensity peptides are mostly affected by co-isolation interference, as reporter ion interference has little to no effect on their ratios, but that reporter ion interference can have profound effect in quantification of high intensity peptides. This provides important practical information that aids the design of TMT experiment to help minimise the potential effects on data quantification of cross condition/population reporter ion interference.

21

Multi-batch TMT: false positives, missing values and batch effects

Optimising the experimental design

For all studies based on more than a single multi-plex TMT batch, we advocate at least one relevant internal reference sample should be included in each batch and assigned to either channel 126C, or 127N. These channels avoid "primary RII", the main cause of isotopic impurities, and are only affected by "secondary RII", which only causes a median increase of 2.2% in intensity. In contrast, placing the reference line at channel 131N, or 131C increases the impact of reporter ion interference by exposing them to "primary RII".

Our results also show TMT experimental designs that can help to minimise the effects of primary and secondary reporter ion interference between the different populations/conditions. For example, in a 10-plex TMT study, when two conditions are being analysed, each with 5 biological replicates, a 5-5 grouped layout would cause multiple channels to be affected by cross population/condition reporter ion interference (Fig. 6a). The optimal design would involve alternating the two conditions across the 10 channels (Fig. 6b).

If a control is included, or for studies analysing 3 different conditions in triplicate, (e.g. three time points, or a control and two different perturbations, etc.), we recommend using TMT 11-plex as all 10-plex TMT setups involve cross population/condition reporter ion interference. An 11-plex TMT set up enables a design without reporter ion interference between the 3 conditions/populations but requires two empty channels at 129C and 130N to achieve this (Fig. 6c). If a control channel is included, as advocated, then it should be placed in channel 126C, while locating the empty channel to position TMT[11] 130N, between the alternating experimental conditions and the final replicates of the 3rd condition (Fig. 6d). All of the suggested setups aim to reduce cross condition/population reporter ion interference, thereby avoiding decreases in quantification accuracy by isotopic impurities.

Discussion

Quantitative proteomic analysis using TMT labelling has become one of the most popular DDA methods currently used, thanks to its multiplexing capabilities, scalability, low missing values index and precision when a single multiplexed batch is analysed. However, when very large-scale studies that require the use of multiple parallel TMT batches are performed(37,38), the situation becomes more complicated. Here, we have used the analysis of data integrated from over 20, 10-plex TMT batches to investigate accuracy, missing values, false positives, co-isolation interference, reporter ion interference and experimental design within very large-scale proteomics experiments. We have focussed on a model data set derived from the analysis of human iPS cell lines, derived from both male and female donors(25).

The resulting data confirms that a single batch of multi-plex TMT experiments minimises the typical missing values issue associated in proteomics with Data Dependent Acquisition (DDA), both at the protein and peptide levels. However, this situation changes as data from two or more separate multi-plex TMT batches are integrated. As multiple batches are combined the missing values index inflates rapidly. This effect is particularly striking at the peptide level, where integrating data from only two different batches causes the missing values index to increase from <2% to ~24%. Even though the inflation rate at the protein level is lower, the integration of the second batch pushes missing protein values from <0.5% to >6%. This inflationary effect can decrease the accuracy of results derived from large-scale experiments that compare data generated from multiple TMT batches. One potential solution would be to utilise MS2-based TMT quantitation, as this has been reported to produce more total peptide identifications(39), however there is no guarantee this will detect peptides/proteins more reproducibly across batches. Furthermore MS2-based TMT quantitation will intensify the disruptive effect of the co-isolation interference.

Multi-batch TMT: false positives, missing values and batch effects

While single TMT batches can provide remarkably precise results within the multiplexed experiment, we found that this is often not reproducible across multiple batches. To study reproducibility, we normalised the data using the proteomic ruler(24) and for every protein we calculated the coefficient of variation (CV), both for the 23 technical replicates of the reference line, and within each of 10-plex TMT batches. The median copy number CV of the technical replicates was ~6.4-fold higher than data from different donors analysed within the same 10-plex TMT batch.

This also underlines the importance of normalising batch effects, in our case via a common control sample within each TMT batch which allows for objective data normalisation to minimise the batch effects, as has been reported(40,41). Copy numbers apply a layer of normalisation, but our data highlight they are insufficient to address batch effects. We showed that by introducing at least one suitable control (reference line) within each TMT batch the batch effects can be normalised effectively. The challenge lies in identifying a suitable control that is truly representative for most proteins being compared within the experiment and in creating a control that is highly reproducible across all the TMT batches.

This study has also highlighted the issue of false positives, reporter ion interference and co-isolation interference. The dataset we selected provided an ideal set up to analyse these factors, as it contained iPSC lines derived from both male and female donors. Thus, by identifying a set of peptides uniquely mapped to the male-specific Y chromosome, these provided a convenient set of internal controls to monitor the expression of false positives. The data showed that even for a 10-plex TMT batch with only two male channels (PT6380), the remaining 8 female channels still quantified 97.5% of all the Y chromosome-specific peptides that were detected in that batch. This means there are false positives being consistently detected in the female channels within the multiplexed experiments, which suggests there are severe limitations when analysing heterogenous populations with TMT. This issue we attribute mostly

to co-isolation interference and we note that the issue has been reduced, though not completely eliminated, with newer generation Orbitrap MS instruments, where the improved source has enhanced the signal to noise ratio(34). Furthermore new isobaric tagging methods have been developed which claim to be co-isolation free(42), however their multiplexing capability is currently limited to a 6-plex.

Furthermore, we used the previously mentioned Y chromosome peptides to study the different reporter ion interference (RII) conditions across 21 different 10-plex TMT batches with different number of male and female derived cell lines, as well as different channel combinations. The data highlighted the effects of primary (male channel isotopic contamination into a +1Da female channel), secondary (male channel isotopic contamination into a -1Da female channel). Reporter channels affected by both primary and secondary reporter ion interference showed a median signal increase in high intensity peptides of 1.7-fold compared to channels not subjected to reporter ion interference. This was found to be primarily caused by "primary RII" as the data also showed that "secondary RII" had the smallest effect with only a median 1.02-fold increase compared to channels with no reporter ion interference. To best avoid the effects of reporter ion interference, we have used these data to propose optimised experimental set ups for assigning samples to specific channels that can either minimize, or eliminate (when possible), the effect of primary and secondary reporter ion interference between conditions/populations. Nonetheless, we highlight again that mixing significantly different populations within a TMT batch, for example iPSCs and terminally differentiated somatic cells, will introduce false positives within the data, as illustrated here by the Y chromosome-specific peptides detected within all female cell lines.

For such large-scale experiments it is also vital to have strict quality control (QC) procedures in place to evaluate and maintain a constant performance within the instrument/s. In our case one of the TMT batches (PT6388) revealed poor performance in the QC run (see supplemental

figures). The failed QC run was not detected until after the samples were run, producing poor results within that batch. We therefore recommend that to execute large-scale TMT a rigorous QC procedure should be set in place before the start of the experiment.

In conclusion, TMT is a valuable methodology for DDA analysis and its potential to increase scalability and precise quantitation have made it a justifiably popular approach for high-throughput proteomic studies. Here, we have provided an in-depth, practical evaluation of parameters affecting the generation of high-quality quantitative data from very large-scale TMT-based proteomics analyses, and we highlight some of the limitations which should be carefully considered when planning these experiments. We hope the resulting information will prove useful for improving experimental design and resulting data quality for many future proteomics projects.

Author contributions

A.I.L. supervised the project. J.H. assisted in the interpretation and analysis of the data. D.B. designed and implemented the sample prep and proteomics experiments. A.B. conceived the study and performed the data analysis and interpretation. The paper was written by A.B. and edited by all authors.

Data availability

The raw files used for this analysis are accessible at ProteomeXchange(22) via the PRIDE repository (23), along with the full MaxQuant(20) processed output (PDX010557).

Acknowledgments

References

1.   Bekker-Jensen, D. B., Kelstrup, C. D., Batth, T. S., Larsen, S. C., Haldrup, C., Bramsen, J. B., Sorensen, K. D., Hoyer, S., Orntoft, T. F., Andersen, C. L., Nielsen, M. L., and Olsen, J. V. (2017) An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. *Cell Syst* **4**, 587-599 e584

2.   Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., Herzog, F., Rinner, O., Ellenberg, J., and Aebersold, R. (2011) The quantitative proteome of a human cell line. *Mol Syst Biol* **7**, 549

3.   Meier, F., Geyer, P. E., Virreira Winter, S., Cox, J., and Mann, M. (2018) BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat Methods*

4.   Camerini, S., and Mauri, P. (2015) The role of protein and peptide separation before mass spectrometry analysis in clinical proteomics. *J Chromatogr A* **1381**, 1-12

5.   Rost, H. L., Malmstrom, L., and Aebersold, R. (2015) Reproducible quantitative proteotype data matrices for systems biology. *Mol Biol Cell* **26**, 3926-3931

6.   Turck, C. W., Falick, A. M., Kowalak, J. A., Lane, W. S., Lilley, K. S., Phinney, B. S., Weintraub, S. T., Witkowska, H. E., Yates, N. A., and Association of Biomolecula Resource Facilities Proteomics Research, G. (2007) The Association of Biomolecular Resource Facilities Proteomics Research Group 2006 study: relative protein quantitation. *Mol Cell Proteomics* **6**, 1291-1298

7.   Larance, M., and Lamond, A. I. (2015) Multidimensional proteomics for cell biology. *Nat Rev Mol Cell Biol* **16**, 269-280

8.   Boisvert, F. M., Ahmad, Y., Gierlinski, M., Charriere, F., Lamont, D., Scott, M., Barton, G., and Lamond, A. I. (2012) A quantitative spatial proteomics analysis of proteome turnover in human cells. *Mol Cell Proteomics* **11**, M111 011429

9.   Larance, M., Kirkwood, K. J., Tinti, M., Brenes Murillo, A., Ferguson, M. A., and Lamond, A. I. (2016) Global Membrane Protein Interactome Analysis using In vivo Crosslinking and Mass Spectrometry-based Protein Correlation Profiling. *Mol Cell Proteomics* **15**, 2476-2490

10.   Hennrich, M. L., Romanov, N., Horn, P., Jaeger, S., Eckstein, V., Steeples, V., Ye, F., Ding, X., Poisa-Beiro, L., Lai, M. C., Lang, B., Boultwood, J., Luft, T., Zaugg, J. B., Pellagatti, A., Bork, P., Aloy, P., Gavin, A. C., and Ho, A. D. (2018) Cell-specific proteome analyses of human bone marrow reveal molecular features of age-dependent functional decline. *Nat Commun* **9**, 4004

11.   Munoz, I. M., Morgan, M. E., Peltier, J., Weiland, F., Gregorczyk, M., Brown, F. C., Macartney, T., Toth, R., Trost, M., and Rouse, J. (2018) Phosphoproteomic screening identifies physiological substrates of the CDKL5 kinase. *EMBO J*

12.   Thompson, A., Schafer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., Johnstone, R., Mohammed, A. K., and Hamon, C. (2003) Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* **75**, 1895-1904

13.   Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., and Pappin, D. J. (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* **3**, 1154-1169

14. Isasa, M., Rose, C. M., Elsasser, S., Navarrete-Perea, J., Paulo, J. A., Finley, D. J., and Gygi, S. P. (2015) Multiplexed, Proteome-Wide Protein Expression Profiling: Yeast Deubiquitylating Enzyme Knockout Strains. *J Proteome Res* **14**, 5306-5317

15. McAlister, G. C., Nusinow, D. P., Jedrychowski, M. P., Wuhr, M., Huttlin, E. L., Erickson, B. K., Rad, R., Haas, W., and Gygi, S. P. (2014) MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal Chem* **86**, 7150-7158

16. Lazar, C., Gatto, L., Ferro, M., Bruley, C., and Burger, T. (2016) Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *J Proteome Res* **15**, 1116-1125

17. Webb-Robertson, B. J., Wiberg, H. K., Matzke, M. M., Brown, J. N., Wang, J., McDermott, J. E., Smith, R. D., Rodland, K. D., Metz, T. O., Pounds, J. G., and Waters, K. M. (2015) Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *J Proteome Res* **14**, 1993-2001

18. O'Connell, J. D., Paulo, J. A., O'Brien, J. J., and Gygi, S. P. (2018) Proteome-Wide Evaluation of Two Common Protein Quantification Methods. *J Proteome Res* **17**, 1934-1942

19. Brenes, A., Bensaddek, D., Hukelmann, J. L., Afzal, V., and Lamond, A. I. (2018) The iPSC proteomic compendium. *bioRxiv*

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

21. Tyanova, S., Temu, T., and Cox, J. (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc* **11**, 2301-2319

22. Vizcaino, J. A., Deutsch, E. W., Wang, R., Csordas, A., Reisinger, F., Rios, D., Dianes, J. A., Sun, Z., Farrah, T., Bandeira, N., Binz, P. A., Xenarios, I., Eisenacher, M., Mayer, G., Gatto, L., Campos, A., Chalkley, R. J., Kraus, H. J., Albar, J. P., Martinez-Bartolome, S., Apweiler, R., Omenn, G. S., Martens, L., Jones, A. R., and Hermjakob, H. (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol* **32**, 223-226

23. Vizcaino, J. A., Csordas, A., del-Toro, N., Dianes, J. A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., Xu, Q. W., Wang, R., and Hermjakob, H. (2016) 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res* **44**, D447-456

24. Wisniewski, J. R., Hein, M. Y., Cox, J., and Mann, M. (2014) A "proteomic ruler" for protein copy number and concentration estimation without spike-in standards. *Mol Cell Proteomics* **13**, 3497-3506

25. Kilpinen, H., Goncalves, A., Leha, A., Afzal, V., Alasoo, K., Ashford, S., Bala, S., Bensaddek, D., Casale, F. P., Culley, O. J., Danecek, P., Faulconbridge, A., Harrison, P. W., Kathuria, A., McCarthy, D., McCarthy, S. A., Meleckyte, R., Memari, Y., Moens, N., Soares, F., Mann, A., Streeter, I., Agu, C. A., Alderton, A., Nelson, R., Harper, S., Patel, M., White, A., Patel, S. R., Clarke, L., Halai, R., Kirton, C. M., Kolb-Kokocinski, A., Beales, P., Birney, E., Danovi, D., Lamond, A. I., Ouwehand, W. H., Vallier, L., Watt, F. M., Durbin, R., Stegle, O., and Gaffney, D. J. (2017) Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* **546**, 370-375

26. Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York

27. Hu, A., Noble, W. S., and Wolf-Yadlin, A. (2016) Technical advances in proteomics: new developments in data-independent acquisition. *F1000Res* **5**

28. Lin, L. I. (1989) A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**, 255-268

29. Limpert, E., Stahel, W., and Abbt, M. (2001) Log-normal Distributions across the Sciences: Keys and Clues. *BioSciences* **51**, 341-352

30. Schwacke, J. H., Hill, E. G., Krug, E. L., Comte-Walters, S., and Schey, K. L. (2009) iQuantitator: a tool for protein expression inference using iTRAQ. *BMC Bioinformatics* **10**, 342

31. Herbrich, S. M., Cole, R. N., West, K. P., Jr., Schulze, K., Yager, J. D., Groopman, J. D., Christian, P., Wu, L., O'Meally, R. N., May, D. H., McIntosh, M. W., and Ruczinski, I. (2013) Statistical inference from multiple iTRAQ experiments without using common reference standards. *J Proteome Res* **12**, 594-604

32. Ow, S. Y., Salim, M., Noirel, J., Evans, C., Rehman, I., and Wright, P. C. (2009) iTRAQ underestimation in simple and complex mixtures: "the good, the bad and the ugly". *J Proteome Res* **8**, 5347-5355

33. Muntel, J., Kirkpatrick, J., Bruderer, R., Huang, T., Vitek, O., Ori, A., and Reiter, L. (2019) Comparison of Protein Quantification in a Complex Background by DIA and TMT Workflows with Fixed Instrument Time. *J Proteome Res* **18**, 1340-1351

34. Paulo, J. A., O'Connell, J. D., and Gygi, S. P. (2016) A Triple Knockout (TKO) Proteomics Standard for Diagnosing Ion Interference in Isobaric Labeling Experiments. *J Am Soc Mass Spectrom* **27**, 1620-1625

35. Hukelmann, J. L., Anderson, K. E., Sinclair, L. V., Grzes, K. M., Murillo, A. B., Hawkins, P. T., Stephens, L. R., Lamond, A. I., and Cantrell, D. A. (2016) The cytotoxic T cell proteome and its shaping by the kinase mTOR. *Nat Immunol* **17**, 104-112

36. Ly, T., Endo, A., and Lamond, A. I. (2015) Proteomic analysis of the response to cell cycle arrests in human myeloid leukemia cells. *Elife* **4**

37. Plubell, D. L., Wilmarth, P. A., Zhao, Y., Fenton, A. M., Minnier, J., Reddy, A. P., Klimek, J., Yang, X., David, L. L., and Pamir, N. (2017) Extended Multiplexing of Tandem Mass Tags (TMT) Labeling Reveals Age and High Fat Diet Specific Proteome Changes in Mouse Epididymal Adipose Tissue. *Mol Cell Proteomics* **16**, 873-890

38. Maes, E., Hadiwikarta, W. W., Mertens, I., Baggerman, G., Hooyberghs, J., and Valkenborg, D. (2016) CONSTANd : A Normalization Method for Isobaric Labeled Spectra by Constrained Optimization. *Mol Cell Proteomics* **15**, 2779-2790

39. Myers, S. A., Klaeger, S., Satpathy, S., Viner, R., Choi, J., Rogers, J., Clauser, K., Udeshi, N. D., and Carr, S. A. (2018) Evaluation of Advanced Precursor Determination for Tandem Mass Tag (TMT)-Based Quantitative Proteomics across Instrument Platforms. *J Proteome Res*

40. Liu, P., Beer, L. A., Ky, B., Barnhart, K. T., and Speicher, D. W. (2017) Quantitative Comparisons of Large Numbers of Human Plasma Samples Using TMT10plex Labeling. *Methods Mol Biol* **1619**, 319-337

41. Ping, L., Duong, D. M., Yin, L., Gearing, M., Lah, J. J., Levey, A. I., and Seyfried, N. T. (2018) Global quantitative analysis of the human brain proteome in Alzheimer's and Parkinson's Disease. *Sci Data* **5**, 180036

42. Virreira Winter, S., Meier, F., Wichmann, C., Cox, J., Mann, M., and Meissner, F. (2018) EASI-tag enables accurate multiplexed and interference-free MS2-based proteome quantification. *Nat Methods* **15**, 527-530

Figure legends

Figure 1 – Protein and peptide missing values: (a) Percentage of missing values for each TMT batch calculated at the protein level. (b) Percentage of missing values for each TMT batch calculated at the peptide level. (c) Box plot showing the results for protein missing values as a function of the number of 10-plex TMT batches (see methods). (d) Box plot showing the results for peptide missing values as a function of the number of 10-plex TMT batches (see methods). For both c & d the lower and upper hinges represent the 1st and 3rd quartiles. The upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge, the lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge

Figure 2 – Peptide identifications and intensities: (a) Number of peptides identified with MS3 intensity greater than zero in all TMT channels, coloured by TMT batch. (b) Histogram of the median normalised peptide intensity (see methods). (c) Stacked density plot showing peptides grouped by median normalised peptide intensity quartiles and their percentage of detection across all TMT channels. (d) Stacked density plot showing quartiles of identification rates for each peptide and their corresponding $\log_{10}$ normalised MS3 intensity.

Figure 3 – Variation: (a) Box plots showing the protein copy number coefficient of variation for all proteins detected in each 10-plex TMT batch as well as all proteins detected in all the reference line replicates (TMT channel 126C in all batches). (b) Box plots showing the protein copy numbers of the 100 most abundant proteins with a coefficient of variation greater than or equal to 7.5 across all reference line replicates (TMT channel 126C in all batches). For both boxplots the lower and upper hinges represent the 1st and 3rd quartiles. The upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge, the lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge

Figure 4 – Y chromosome peptides in female channels: Scatter plot showing the gender/incidence of false positives across 21 TMT batches and their reporter ion mass tags. Male cell lines are shown as a grey square, female cell lines are represented by a circle. The female lines (circles) are shaded to indicate the percentage of Y chromosome specific peptides that were detected in their channel within each TMT batch.

Figure 5 – TMT channel leakage analysis: (a) Box plot showing the median normalised intensity of Y chromosome specific peptides for both female and male cell lines across 21 TMT batches. (b) Box plot of ratios for Y chromosome specific peptides, comparing male channels vs female channels affected by different reporter ion interference type. (c) Box plot of ratios for Y chromosome specific peptides, stratified by the median normalised intensity, comparing female channels affected by different types of reporter ion interference vs female channels not subjected to reporter ion interference. For all 3 boxplots the lower and upper hinges represent the 1st and 3rd quartiles. The upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge, the lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge

Figure 6 –TMT experimental design from reporter ion interference analysis: (a) 5-5 grouped layout for a 10-plex TMT batch with 5 replicates of two conditions. Two channels are being affected by cross population primary and secondary reporter ion interference. (b) optimal layout for a 10-plex TMT batch with 5 replicates of two conditions with no cross population primary or secondary reporter ion interference. (c) optimal 11-plex configuration for 3 populations with two empty channels, no channels suffer cross population reporter ion interference. (d) optimal 11-plex configuration for 3 populations with one empty channel and one control channel. Only two channel suffers primary and secondary reporter ion interference.
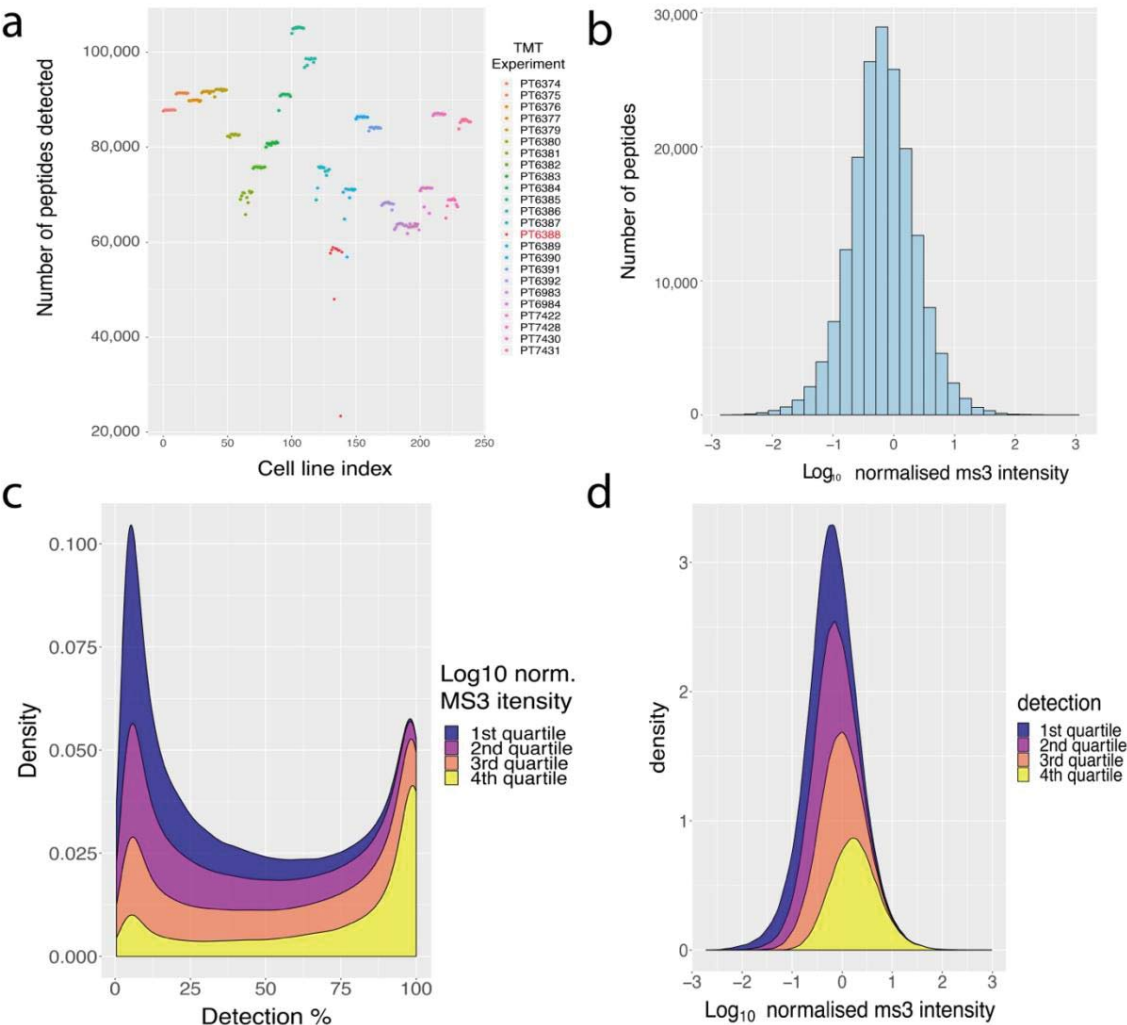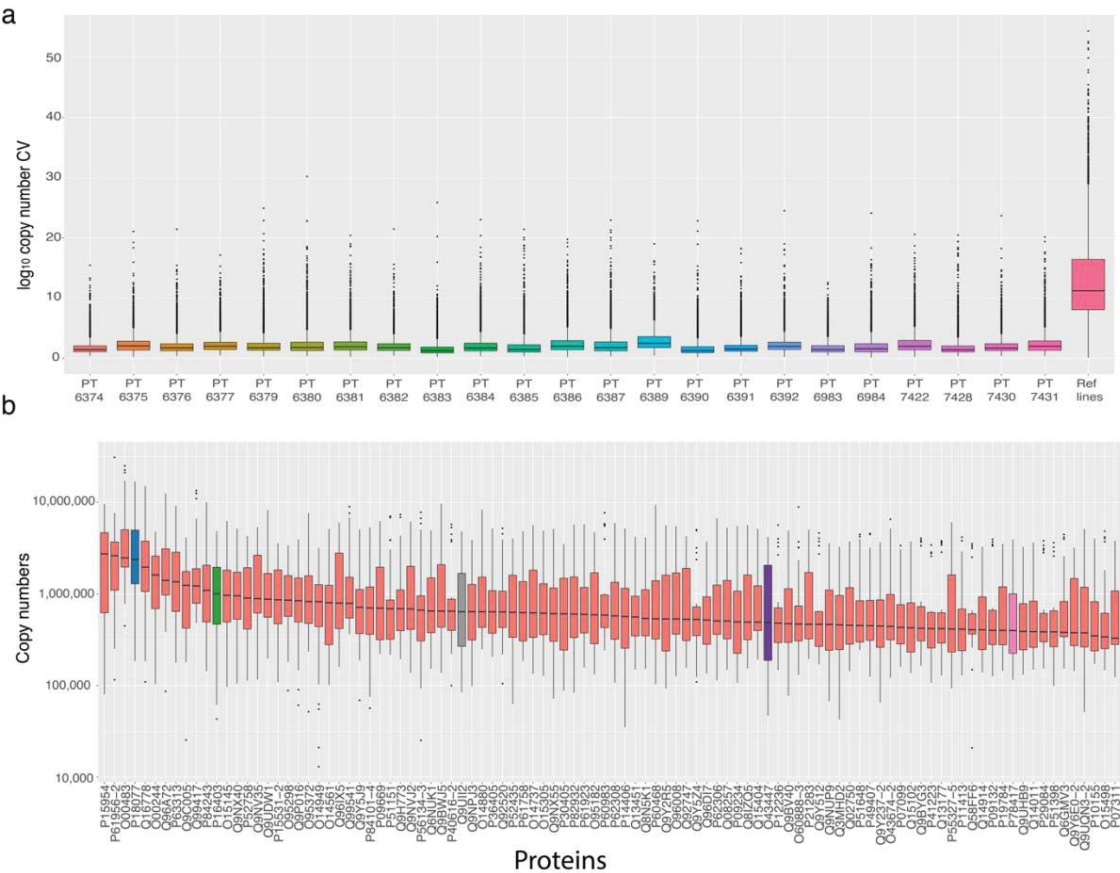
Figures:

Figure 1:
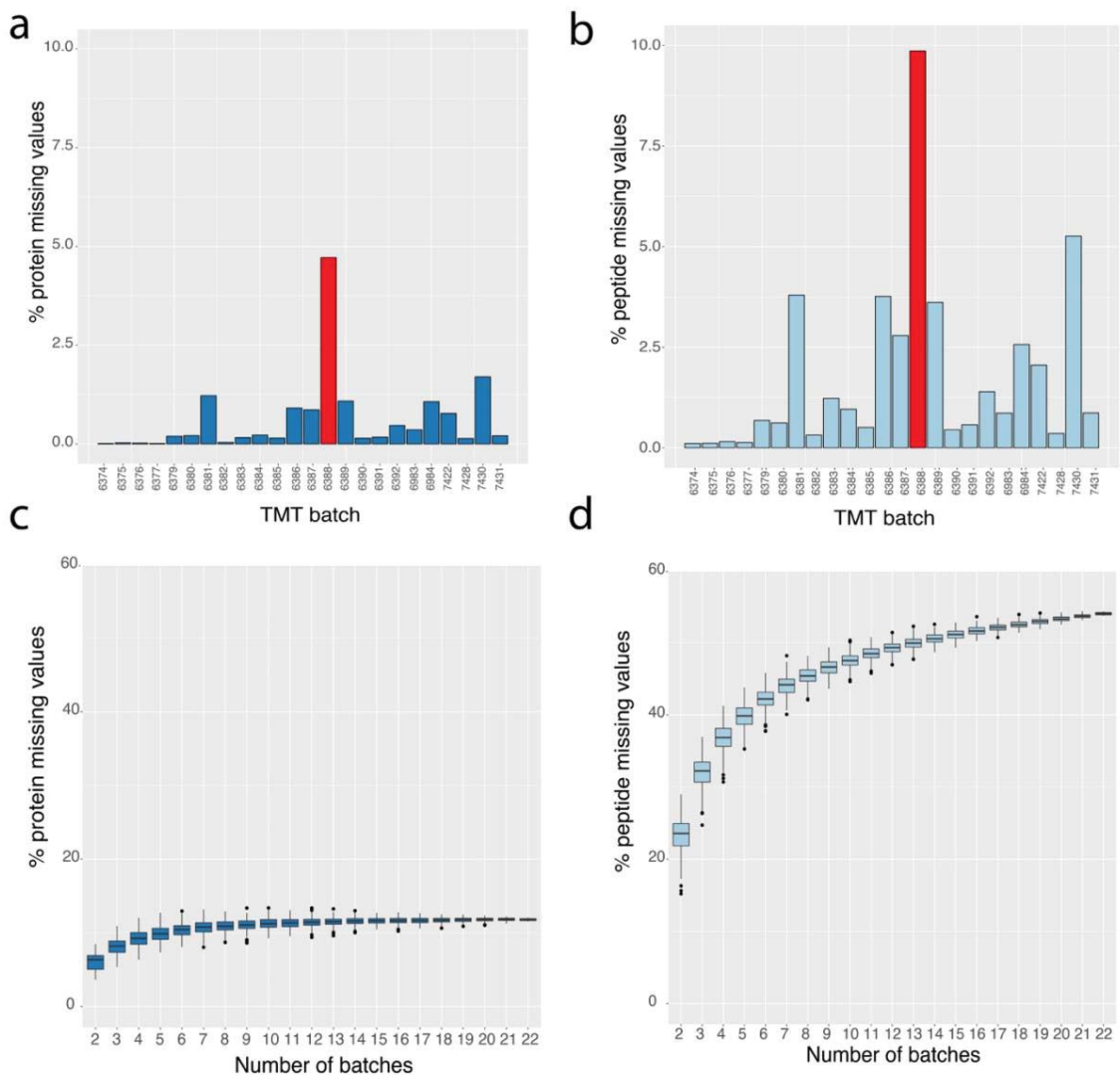
Figure 2:

Figure 3:

Figure 4:

Figure 5:

Figure 6: