Review

Global absolute quantification of a proteome: Challenges in the deployment of a QconCAT strategy

Philip Brownridge[1], Stephen W. Holman[2]*, Simon J. Gaskell[2]**, Christopher M. Grant[3], Victoria M. Harman[1], Simon J. Hubbard[3], Karin Lanthaler[3], Craig Lawless[3], Ronan O'cualain[2], Paul Sims[2], Rachel Watkins[3] and Robert J. Beynon[1]*

[1] Protein Function Group, Institute of Integrative Biology, University of Liverpool, UK
[2] Faculty of Life Sciences, Manchester Interdisciplinary Biocentre, University of Manchester, Manchester, UK
[3] Faculty of Life Sciences, University of Manchester, Manchester, UK

In this paper, we discuss the challenge of large-scale quantification of a proteome, referring to our programme that aims to define the absolute quantity, in copies per cell, of at least 4000 proteins in the yeast *Saccharomyces cerevisiae*. We have based our strategy on the well-established method of stable isotope dilution, generating isotopically labelled peptides using QconCAT technology, in which artificial genes, encoding concatenations of tryptic fragments as surrogate quantification standards, are designed, synthesised de novo and expressed in bacteria using stable isotopically enriched media. A known quantity of QconCAT is then co-digested with analyte proteins and the heavy:light isotopologues are analysed by mass spectrometry to yield absolute quantification. This workflow brings issues of optimal selection of quantotypic peptides, their assembly into QconCATs, expression, purification and deployment.

Received: January 24, 2011
Revised: April 8, 2011
Accepted: April 19, 2011

Keywords:
Absolute quantification / QconCAT / *Saccharomyces cerevisiae* / Stable isotope standards / Technology

1 Introduction

As proteomics has become increasingly quantitative, new approaches have been developed to measure the quantity of a protein in a cell. Many of these approaches were developed to allow relative quantification, in which the protein quantity in one cell/physiological state is expressed relative to a second state – for example, a diseased state relative to a normal control. These data are dimensionless and expressed as ratios – thus, a protein might be defined as being expressed as "2.4-fold higher in cell state A compared with cell state B". This is undoubtedly of value in the discovery of differentially expressed proteins, but the lack of formally quantitative data means that further interpretation is difficult. There is a rapidly developing need for absolute quantification that would allow statements such as "Protein X, in cell state A was present at 65 000 (± error) molecules per cell, whereas in cell state B, this reduced to 38 000 (± error) molecules per cell."

One goal of systems biology is to enable predictive biology, in which detailed knowledge of the cellular constituents, their quantities, dynamics and interactions can be embedded in mathematical models that permit simulation of cellular state changes, testable by experiment and leading to a formal definition of living processes. The strength of this approach lies with the elegance of the modelling that creates a conceptual scaffold upon which knowledge of the players, obtained by experimentation and observation, can

Correspondence: Professor Robert J. Beynon, Protein Function Group, Institute of Integrative Biology, University of Liverpool, L69 7ZB, UK
E-mail: r.beynon@liv.ac.uk

Abbreviations: COPY, census of the proteome of yeast; **GluFib,** Glufibrinopeptide B; **PRIDE,** proteome identifications database; **QRL,** QconCAT replication level; **SRM,** selected reaction monitoring

*These two authors contributed equally to this work.
**Current address: Queen Mary University of London, London E1 4TS, UK

be assembled. It follows that the strength of the model is only as good as the data embedded in it, and that these data must be rigorously quantitative. Our goal is to create accurate baseline values for the cellular quantities of most proteins in the proteome of the yeast *Saccharomyces cerevisiae*. The programme (census of the proteome of yeast, COPY), funded by the Biotechnology and Biological Sciences Research Council (BBSRC) in the UK, has set a target of absolute quantification of at least 4000 proteins employing quantitative MS. The availability of at least 4000 quantified proteins covering over four orders of magnitude in dynamic range will provide a fundamental resource for the future development of MS-based quantification approaches (label-mediated or label-free). It will generate a comprehensive, fully quantitative, statistically validated data set that could become a gold standard for validation of other quantification approaches, permitting direct comparison of methodologies that aim to relate protein abundance to transcriptional data. We have not addressed the added dimensionality of post-translational space [1, 2] and our quantification strategy is intended to be "blind" to these modifications. Future studies will be needed to partition the total protein pool into the different post-translational variants.

The significance of this programme is that it will generate accurate, absolutely quantitative data for an entire proteome; data essential for any systems level modelling of cellular or subcellular processes. It will provide a quantitative framework and resources that can be used by the entire community, and will serve as a reference for future studies in which changes in the proteome are determined, and provide a reliable and complex data set for the development of new approaches.

2 Strategies for absolute quantification in proteomics

The focus of our programme is on the quantification attained at the MS level. MS approaches, compared with protein-tagging approaches [3, 4], should allow quantification of native proteins in an unperturbed system. Alternative approaches, such as quantitative immunochemical methods are not currently feasible, since they would require a specific and high-affinity antibody to each protein in the proteome, and an accurately quantified internal standard comprising the pure protein in the same form as the analyte – the quantification of such standards returns us to the issue of optimal methods for absolute quantification – a circular argument.

MS-based methods employ stable isotope labels for differential analysis, but there is an emergent interest in methods that do not rely on isotopic labels. The additional complexity (and cost) of stable isotope labelling a proteome sample or standard (whether metabolically or chemically, at the protein or peptide level) has led to the search for alternative methods of quantification that are based on direct assessment of the signal (or ion current) that is acquired by the mass spectrometer – referred to collectively as "label-free" methods (e.g. [5–9]). These methods are based on the entirely reasonable observation that when a mixture of proteins is digested to constituent peptides, the most abundant proteins are expected to yield more detectable ions and with stronger signal intensities. While label-free methods are undoubtedly attractive because of their simplicity, this simplicity incurs serious penalties, of which uncertain linearity of response and poor accuracy are the most critical [5]. The real benefits of label-free methods might be in the assignment of approximate abundance classes for proteins within a proteome, aiding rational design of absolute quantification experiments based on MS, and in comparative (relative quantification) analyses once absolute values are known. In our experience, label-free methods (especially precursor ion intensity based) are acceptable for the high-abundance components of any proteome sample, but at column loadings at the single figure to sub-fmol levels, where signals are still detectable, the technical variance becomes limiting.

Tagging methods, in which an immunologically detectable or fluorescent tag is fused to each protein in the proteome, are limited to organisms for which genetic manipulation is feasible. Tagging methods have been applied to the eukaryotic organism, *S. cerevisiae* and data for many thousands of proteins have been derived, in addition to comprehensive MS-based analyses. Comparison of such data sets is difficult, but some analysis is possible if the data are reduced to ppm values ([10], PAXDB (pax-db.org)). We have collated these expression data and compared them (Fig. 1). The results are rather surprising. While methodologically related approaches (intensity-based label-free methods, GFP tagging) show acceptable consistency, the scatter of quantification values for comparisons of dissimilar approaches should give cause for concern. It is probably fair to say that until these discrepancies can be resolved, no one data set can yet be considered to be a gold standard.

3 Is global proteome quantification feasible?

The ultimate aim of any programme to fully quantify the proteome of a single organism is to encompass every gene product encoded in the genome. At present, this remains unrealistic. First, the proteome is dynamic and not every gene is expressed at any one time in a cell or tissue. Hence, 100% proteome coverage for a single yeast strain under a given set of growth conditions is clearly impossible – although knowledge that some proteins are *not* expressed has intrinsic value. Second, although great strides have been made in proteomics technology, both in terms of the preparative and separation stages (e.g. multi-dimensional chromatography) and resolution and the lower limit of

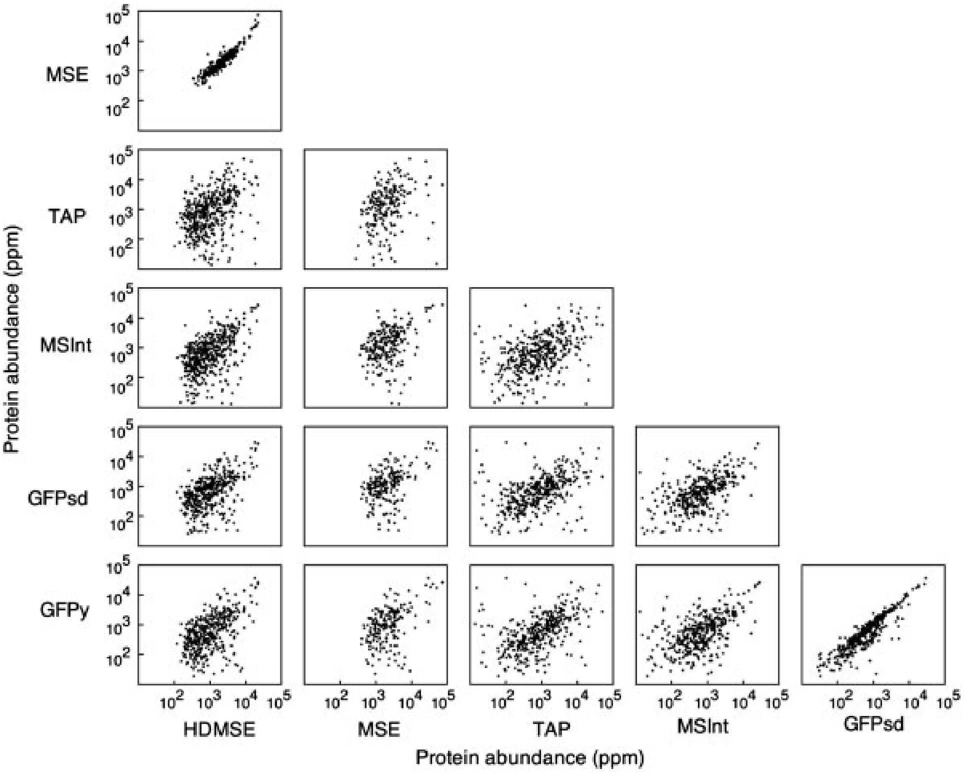

Figure 1. Comparison of protein quantification approaches. For *S. cerevisae*, there are multiple sets of protein quantification data, including those based on TAP-tagging [3] and GFP-tagging [4] under rich (GFPy) or nutrient-limited (GFPsd) conditions. We have also acquired label-free quantification data using the data-independent selection and fragmentation (MS^E) workflow, from which we derive intensity-based quantification data by comparison with an internal standard [8], either with ($HDMS^E$) or without (MS^E) ion-mobility to further resolve precursor/product ion association. A further MS data set based on precursor ion intensity (MSInt) was included [11]. For this comparison, we converted the MS^E and TAP-tagged data sets into ppm abundance, and recovered the equivalent data for the GFP-tagged data and MSInt data from the PAXDB database (pax-db.org). This permitted comparisons of the quantification data for about 400 proteins. The data should be interpreted solely in terms of the scatter, as the slopes of the scatter graphs may be modulated by the assumption in the conversion into ppm values.

quantification of the analytical stage (e.g. mass spectrometer), some proteins are missed during high-throughput experiments. Recent analyses suggest that the scanning speed of modern mass spectrometers is the principal rate-limiting factor, referring to this as ''sequencing speed'' [11, 12].

Given these caveats, a reasonable aim is to quantify a realistic ''detectable'' proteome. From large-scale proteome efforts to date, it has been estimated that only ~60% of the total yeast proteome is covered by extant MS data [13], although this extends to 76% for ORFs with SGD gene names. However, this number reflects the total coverage from 48 diverse experiments contained in the PeptideAtlas database [14], and not a single experiment. A more encouraging recent study [11] combined three parallel experimental strategies that facilitated relative quantification for over 4000 yeast proteins. To achieve an equivalent level of coverage for direct absolute quantification in a single study is a significant challenge, and hence we propose the realistic aim for this project is to generate quantitative data for the majority of the yeast proteome, covering about three-quarters of the predicted ORFs. This represents the optimal balance between feasibility and value to the community.

The dynamic range of protein expression (~10^5, from over one million to a few tens of copies per cell) in yeast [3] is substantially less than that for plasma (circa 10^{12}). Accurate absolute quantification will require that we span this range confidently and analysis of the extant proteomics data contained in PeptideAtlas suggests that MS-based approaches can reach sufficiently deeply into the proteome. Ranging studies have demonstrated convincingly that the yeast proteome is accessible to quantification by MS [15].

4 Multiplexed quantification by internal standard: QconCAT

The preferred methodology for absolute quantification by MS is based on the well-established (decades old) principle of stable isotope dilution. In this approach, a known amount of an isotopically labelled standard is added to the analyte preparation. Subsequently, the mixture of analyte and standard is analysed, or if required, analysis is preceded by proteolysis or purification steps. MS analysis of the standard:analyte mixture then yields the relative intensity of the standard and analyte ions, and since the standard is present in known amounts, the quantity of analyte can be calculated directly. The method is tolerant to manipulation or loss after the standard has been added since standard and analyte should be identical in their downstream behaviour. For example, the workflow is robust to sample fractionation and concentration at the peptide level but less so to pre-fractionation at the protein level.

In proteomics, it is rare that the standard is a stable isotope-labelled intact protein [16–19] although protein-level standards obviate some risks associated with peptide-based quantification [20]. Protein-level quantification would require (usually heterologous) biological synthesis of the standard isotope-labelled protein, and assumes that the standard and analyte occupy exactly the same post-translational state. Moreover, addition of the standard as a free protein may lead to behaviour that is different to the behaviour of, for example, the analyte embedded in a supra-molecular assembly. Finally, the challenge of generating several thousand full-length, accurately quantified protein standards for a large-scale proteome quantification is significant. More commonly, therefore, a second elaboration of the internal standard approach is that of surrogacy, where a peptide (usually tryptic) is used as a standard, with a 1:1 molar relationship to the protein from which it is derived. Analysis at the peptide level eliminates complexities associated with higher level organisation noted above. The same peptide can be synthesised chemically (often referred to as an AQUA peptide [21]), but this also brings several problems, including high cost per peptide, the difficulty of synthesising some peptides with intractable sequences and the tendency of peptides to adhere irreversibly to vessel walls. Moreover, if many proteins are to be quantified, each AQUA peptide must be separately quantified.

To circumvent many of the complications posed by large-scale AQUA-based quantification studies, we developed the QconCAT approach for multiplexed absolute quantification [22–26]. In brief, synthetic genes, optimised for heterologous expression in *Escherichia coli*, encode a single ORF that is a concatenation of tryptic peptides, each of which acts as an internal standard (a quantotypic or Qpeptide) for a different protein (of course it is equally feasible, or even desirable, to encode two or more peptides to report on each analyte protein). The gene design allows inclusion of N- and C-terminal sacrificial short peptides to protect the termini of true Qpeptides, purification sequences such as hexahistidine ("His-tag") motifs, and peptides that allow QconCAT to QconCAT comparison.

We have now designed, built and expressed over 120 QconCATs, providing over 6000 Qpeptides in a range of different studies. With a single exception, all of these artificial proteins have been expressed in inclusion bodies – a positive feature, since recovery of inclusion bodies gives an immediate tenfold purification and their dissolution in chaotropes prior to affinity chromatography ensures that there are no higher order structure constraints to impede proteolysis. Each of the QconCATs was expressed at levels more than adequate for quantification studies (typically, we prepare milligram quantities from 250 mL shake cultures), and was readily purified via the His-tag that is built into the QconCAT cassette. We hold the view that QconCATs of around 50–70 kDa, encoding 450–550 amino acids (~50 Qpeptides) are optimal in terms of density of Qpeptides and ease of downstream handling. Smaller QconCATs are imbalanced in terms of the effort of gene synthesis and preparation relative to the number of proteins quantified, whereas larger QconCATs are more prone to aggregation and to ectopic proteolysis during their preparation, which might compromise their use as standards.

QconCATs are mixed with the analyte protein preparation prior to digestion (usually by trypsin) in order to generate analyte peptides and a stoichiometric mixture of Qpeptides. It follows that proteolysis of both QconCAT and analyte must be complete, in order for quantification to be accurate (indeed, the same argument applies to AQUA studies, where complete proteolysis of analyte protein is often not evaluated). We have conducted extensive studies on the rate and completeness of proteolysis in QconCAT experiments and have demonstrated that QconCATs are digested rapidly and completely (within a few minutes) and that unless appropriate denaturation steps are used, analyte digestion is substantially slower [26]. The lack of higher order structure of the QconCATs undoubtedly contributes to their rapid proteolysis. Inherent digestibility of the standard is not commonly an issue, although we have encountered rare instances where the analyte peptide is released faster than the standard [27].

5 Choice of organism

The yeast *S. cerevisiae* offers an ideal model system for these studies since it is genetically tractable and has served as the organism of choice for most post-genomic studies. It was the first eukaryotic organism to have its entire genome sequenced and with the ready availability of a wide range of yeast transcriptome data (its messenger RNAs), it is now both logical and feasible to examine its cognate proteome (the protein complement). This is important since it is the proteins, and not the genes, which are the functional components of the cell. For the reference material, we are using a haploid yeast strain deleted for *ARG3*, encoding ornithine carbamoyltransferase, and *LYS2*, encoding α-aminoadipate reductase (*MATα leu2Δ0 lys2Δ0 ura3Δ0 his3Δ1 arg3::kanMX4*) that was produced by the Saccharomyces Deletion Project (EUROSCARF accession number Y11335). Yeast cultures are grown in glucose-limited chemostat cultures as we have previously described [28, 29]. Cells are grown in an eight-plex parallel system (www.dasgip.com) in

100 mL volumes, replicated at the inoculum level. At a dilution rate of $D = 0.1\,h^{-1}$, we generate 50×10^8 cells (yielding $\sim$20 000 µg of total extracted protein) per flask. For the most comprehensive studies, it is essential that protein extraction is quantitatively complete and reproducible and our current approach is to subject the cells to extensive disruption cycles, but not to invoke any centrifugal fractionation of soluble and insoluble fractions.

6 QconCAT design

In a typical quantification workflow, one or more stable isotope-labelled QconCATs (we refer to QconCATs associated with this study as COPYCATS) are mixed with the analyte broken cell preparation and the mixture is then co-digested to completion with trypsin. Subsequently, the levels of standard and unlabelled analyte peptides are determined by the most appropriate suitable MS method. Although simply stated, the number of pitfalls in this workflow is considerable. As we have embarked on this programme, it has become clear that many proteotypic peptides are not suitable for quantification studies; they can be the products of a missed cleavage, contain residues such as methionine, which when oxidised can split the ion current between the oxidised and non-oxidised form, be subject to PTMs that are variable or difficult to reproduce in the standard or are isobaric to other peptides in the proteome. As such, we have coined the term ''quantotypic'' for those peptides that are formally and quantitatively representative of the protein; the discovery of quantotypic peptides is substantially more challenging than the discovery of proteotypic peptides. Some of the factors requiring consideration are mapped in Fig. 2.

6.1 QconCAT replication level

Our goal is to generate the highest possible quality data set that will act as a sustained and respected resource. We therefore need to address the issue of the number of Qpeptides used for each protein, the QconCAT replication level (QRL). As can be seen from Table 1, the QRL is the largest influence on the number of QconCATs that will be required. At one extreme, a QRL of unity lacks any independent check and at the other extreme, the use of the entire recombinant protein as a standard is inherently unworkable in multiplexed studies, because of the excessive number of additional peptides that would be added to the analytical preparations. We favour a QRL of 2, since each protein

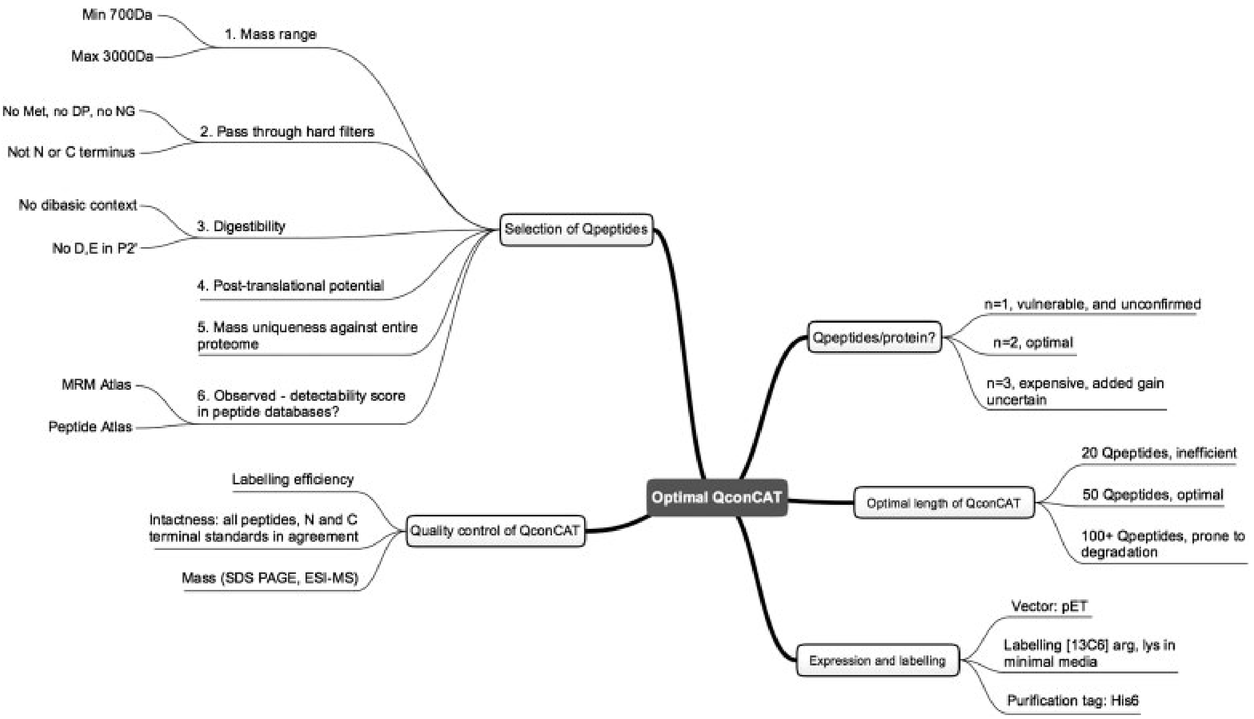
Min 700Da
Max 3000Da
1. Mass range
No Met, no DP, no NG
Not N or C terminus
2. Pass through hard filters
No dibasic context
No D,E in P2'
3. Digestibility
4. Post-translational potential
5. Mass uniqueness against entire proteome
MRM Atlas
Peptide Atlas
6. Observed - detectability score in peptide databases?
Selection of Qpeptides
Qpeptides/protein?
n=1, vulnerable, and unconfirmed
n=2, optimal
n=3, expensive, added gain uncertain
Labelling efficiency
Intactness: all peptides, N and C terminal standards in agreement
Mass (SDS PAGE, ESI-MS)
Quality control of QconCAT
Optimal QconCAT
Optimal length of QconCAT
20 Qpeptides, inefficient
50 Qpeptides, optimal
100+ Qpeptides, prone to degradation
Expression and labelling
Vector: pET
Labelling [13C6] arg, lys in minimal media
Purification tag: His6

Figure 2. Design considerations in development of a QconCAT strategy. The optimal construction and expression of QconCATs requires decisions to be made at multiple levels, from overall design principles, selection of quantotypic peptides, assembly into QconCATs, expression and purification. (Key: DP, Asp-Pro sequence; NG, Asn-Gly sequence; D, E in P2', no acidic residue two amino acids C-terminal to the scissile bond). The ''no dibasic context'' filter avoid any peptide that is flanked by two or more R, K residues at either end – the variability in cleavage at these positions can compromise quantification. Finally, we avoid the N- or C-terminal peptides because these could be prone to exoproteolytic fraying.

quantification is independently verified (the two Qpeptides do not even have to be in the same QconCAT), reconciling assurance with analytical complexity. We then have to factor in realistic estimates of failures, either in performance of a Qpeptide or in expression of a QconCAT, since these proteins are entirely novel their behaviour is inherently unpredictable. Even at a defaulter rate of one Qpeptide in ten (10%) the additional QconCATs required to pick up the defaulters are modest by comparison with the total number of QconCATs required, which is strongly driven by QRL. Further, higher values of QRL may be appropriate for proteins that present more of an analytical challenge – low abundance or membrane proteins. At present, our working assumption is that a QRL of 2 across the entire proteome quantification is appropriate. We will therefore require between 160 and 176 QconCATs. We must also allow for a finite failure rate for QconCAT synthesis or expression – and have aimed for 200 QconCATs, each containing ~50 Qpeptides, in additional to common peptide tags – a total coverage equivalent to 10 000 "protein quantifications" for our target of up to 4000 proteins (Table 1).

6.2 Qpeptide grouping strategy

From the outset, we have chosen to group proteins (and thus their quantotypic peptides) in COPYCATs according to functional relatedness. Functional grouping builds QconCATs that define specific pathways and reflect the interest of external research groups who will use the COPYCATs and generate comparative data – such as a group of proteins involved in a concerted pathway or functional process that comprises a limited sub-proteome; e.g. pentose phosphate pathway or those proteins responsible for the response to oxidative stress. A group can also comprise a set of proteins that share a common general functional theme, such as transcription factors. Alternatively, organisational grouping can take cognisance of the supramolecular/organellar structure of the cell, permitting approaches based on subcellular [30] or supramolecular isolation (such as the mitochondrion, vacuole, peroxisome, ribosome, proteasome, APC, etc). We do not favour abundance grouping, based on a consensus of existing data sets to cluster Qpeptides according to the abundance of the cognate proteins, obtained, for example, from tagging methods [3] or label-free quantification. A low/modest degree of overlap is anticipated, and even encouraged, as this will provide inter-QconCAT replication rather than intra-QconCAT replication.

Table 1. Scope and scale of a QconCAT approach to global quantification of a proteome

QRL	Quantotypic peptide defaulting rate			
	0%	1%	5%	10%
	Number of QconCATs needed to quantify 4000 proteins			
1	80	81	84	88
2	160	162	168	176
3	240	243	252	264

For the absolute quantification of 4000 proteins, it is possible to calculate the number of quantotypic peptides, and thus the number of QconCATs that would be needed. This parameter is most strongly controlled by the average number of quantotypic peptides that are used to quantify each protein (QRL), but also by the expectation of a finite rate of failure in the design and biosynthesis of wholly novel recombinant proteins.

6.3 The QconCAT expression cassette

Each QconCAT is an assembly of tryptic peptides, the order of which is not critical, which gives us the opportunity to place the peptides in an order that serves experimental requirements. Where possible, we try to assemble the peptides in a pattern that preserves the P1' residue, in order to retain the local primary sequence context. Because the gene will be synthesised de novo, there are opportunities to introduce additional features (Fig. 3). Each QconCAT has common features at both the N- and C-terminus. At the N-terminus, we add a short sacrificial peptide that protects the N-terminus of the first true peptide, which is a quantification standard. We include Glufibrinopeptide B (GluFib) in every QconCAT in this position, as we can then quantify the heavy QconCAT by reference to an accurately quantified unlabelled GluFib standard. At the C-terminus we incorporate a sequence variant based on GluFib, allowing for two-point quantification and confirmation that the QconCAT is intact. The extreme C-terminus encodes a hexahistidine purification tag that is used for the purification of QconCATs on NiNTA columns.

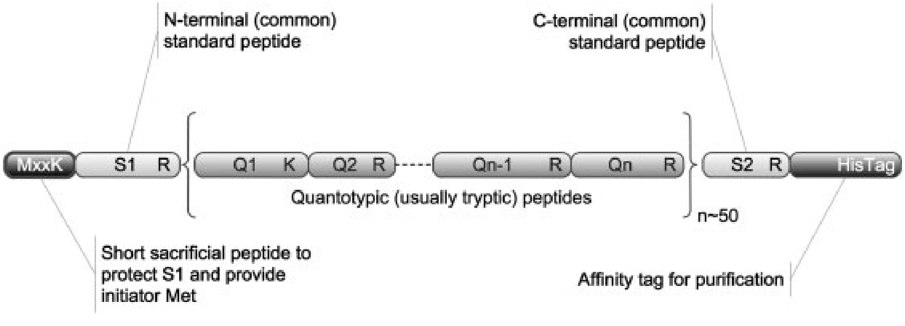
N-terminal (common) standard peptide
C-terminal (common) standard peptide
MxxK
S1 R
Q1 K
Q2 R
Qn-1 R
Qn R
S2 R
HisTag
Quantotypic (usually tryptic) peptides
n~50
Short sacrificial peptide to protect S1 and provide initiator Met
Affinity tag for purification

Figure 3. The QconCAT expression cassette. Each QconCAT is built within a common sequence context that provides for purification tags (HisTag), standardisation tags (peptides S1 and S2, one N-terminal and one C-terminal) and protective sacrificial peptides (MXXK, or additionally, MxxR) as well as the quantotypic peptides (Q_1–Q_n).

6.4 QconCAT assembly and design

The outcome of the functional grouping strategy is a list of gene names or a file of protein sequences from which an optimal QconCAT must be designed. On average, each of 25 proteins (mean molecular weight 35 kDa) would generate about 30 tryptic peptides, and at a QRL of 2, this yields 750 (30 × 25) peptides as the source peptide pool for design of a QconCAT of 50 Qpeptides, mapping to 25 analyte proteins. Although this appears at first glance to have a high level of redundancy, there are several constraints imposed on the process of gene design.

For this study, we have developed a customisable bioinformatic pipeline able to generate optimised candidate sets of Qpeptides and COPYCATs from a subset of yeast proteins or gene names. The pipeline selects peptides from the proteins meeting user-specified length and content criteria, ensures unambiguity in their m/z values for MS detection, ensures a high likelihood of ionisation and detectability by the mass spectrometer in the gas phase (“flyability”) and generates an optimal order of the Qpeptides within the QconCAT to suppress potential missed cleavages and optimise the encoding DNA sequence. Further optimisation occurs at the gene synthesis step that maximises the COPYCAT codon usage and minimises transcript secondary structure. Mass/charge ambiguity is often resolved by differences in the retention time on reversed-phase separations.

6.5 Peptide “detectability”

This is a key issue for Qpeptides; they must be readily cleaved enzymatically, both from the QconCAT and the endogenous proteins, and the peptide must have an MS response factor that is compatible with detection. The ability to detect such peptides reproducibly has prompted researchers to examine the “proteotypic” nature of such peptides in order to define signature peptides as standards for quantification. Recent efforts, including in our own group, have focused on developing machine-learning approaches to predict whether a peptide is likely to be observed in a given proteomics experiment, dependent on the separation method, ionisation, instrument and labelling [6]. We have continued to develop such predictive bioinformatics tools and this will also be continued in this project. At present, we are developing a consensus prediction pipeline, which predicts peptides from a candidate set of proteins based on their amino acid composition and associated physicochemical properties. Using support vector machines, Random Forests, artificial neural networks and genetic programming we are able to predict “detectability” with around 75% cross-validated accuracy (positive predictive value, PPV), at sensitivity over 50%. Although not perfect, it significantly enriches for peptides likely to be “quantotypic” and outperforms related tools on yeast and other organisms (Eyers et al., submitted). Typically, this approach provides around four peptides per protein on average from which to choose. Importantly, we also consider the likelihood of cleavage of the attendant tryptic peptide bonds; not only is this included as a prediction feature, but is used for optimal ordering of Qpeptides within a QconCAT, to reduce the potential for slow hydrolysis. Our experience confirms that efficiency of proteolytic cleavage is not a major issue (although see Section 6.6) but this step will provide additional assurance that potential problems are avoided.

6.6 Enzymatic cleavage of QconCATs and endogenous peptides

The concatenation of the Qpeptides into the QconCAT removes the native primary sequence context, which could influence quantification. Quantification is impaired if either the QconCAT or the analyte proteins are incompletely digested, such that the yield of either peptide is incomplete – indeed, this is not a problem unique to our workflow, but any quantitative approach using proteolytic digestion to generate peptides as analytes. Carefully controlled digestion protocols can ensure that these potential differential proteolysis problems are diminished. Furthermore, it has been well established that the main determinant of the rate of proteolysis of native proteins is higher order structure, not primary sequence context. Tightly folded proteins, particularly those with a high proportion of β sheet, are intrinsically resistant to proteolysis [31]. There is no reason, a priori, to expect that QconCATs would adopt such tightly folded structures. Indeed, their propensity to form insoluble inclusion bodies and their recovery by dissolution in strong chaotropes both diminish concerns about structural impediments to proteolysis. By contrast, unless care is taken in the prior denaturation of analyte proteins, their higher order structure would almost certainly influence proteolysis, impacting absolute quantification. The goal should be to make the primary sequence the only factor determining the rate of digestion. Trypsin makes interactions with three (or possibly four) residues around the scissile bond (numbered P4-P3-P2-P1//P1’-P2’-P3’-P4’ according to the nomenclature scheme of Schechter and Berger [32]).

There are some primary sequence considerations that can enhance the success of the quantification. Two of the strongest are the avoidance of an acidic residue close to the P1 site, particularly at P2’ either in the standard or analyte, and the avoidance of dibasic cleavage sites in the analyte. In the latter instance, the problem is not poor cleavage but that the proteolysis of analyte can be split between cleavage products at either of the two basic residues in P1. We have developed an approach using the information theory capable of predicting missed cleavage sites with over 90% accuracy [33] and have a prototype support vector machine method that increases this to over 95%. These algorithms are deployed in the Qpeptide selection pipeline, since we wish to avoid sites that are potential missed cleavage sites in both the native proteins (the two sites subtending the limit peptide) and the QconCAT itself. Thus candidate peptide selection can be driven by both "detectability" and "cleavability" for inclusion in a QconCAT.

6.7 Exploiting databases for peptide selection

In addition to the predictive tools available for selecting Qpeptides and designing QconCATs, there is naturally a wealth of data for yeast peptides/proteins in existing databases. Prior observation of a peptide in an experimental proteomics context is a further compelling reason to select it as a signature peptide. We also consider available yeast data from repositories, primarily PeptideAtlas [34], which is particularly useful for challenging proteins with few candidate proteins that pass all our selection criteria. We are currently building a QconCAT database to support collection of richer data for subsequent analyses, which we plan to deliver via a BIOMART interface. In particular it will be important to ensure sufficient experimental details are captured for each peptide identification, since ionisation method, instrument and labelling can affect peptide detectability and fragmentation patterns are equally important for later stages when selecting transitions for selected reaction monitoring (SRM)-based quantification.

As an example of the rate of attrition that operates in the design of suitable QconCATs, one of the earliest sets of standards that we designed was for the *Saccharomyces* chaperones. There are 63 proteins in this category, and they were divided between three QconCATs (Fig. 4). After selecting peptides on the basis of mass/length, the first uniqueness filter was based on sequence alone. This eliminated 238 (1635-1497) of the peptides within the chaperonin set, because they were found in a least two proteins within the whole proteome. A further series of composition and digestibility filters reduced the total peptide set to 618, of which 275 (618-343) of the peptides had the same mass as at least one other peptide within the whole proteome (though a different sequence). Two proteins having 99% sequence identity generated identical peptides, and are combined into a "protein group" as they cannot be distinguished by quantotypic peptides using trypsin.

7 QconCAT expression

The QconCATs are designed for high-level expression in *E. coli*. They are synthesised commercially (www.polyquant.com), having been through optimisation algorithms that emphasise high abundance codons and check for mRNA secondary structure features that might impair translation. Typically, the QconCATs are provided as sequence-verified genes, ready cloned in one of the T7 expression vectors; specifically, pET-21a. Expression is tightly regulated, and we prefer induction with IPTG to give precise control of the induction process. Expression is usually rapid and extensive (Fig. 5) and yields are typically of the order of 2–5 mg per 100 mL of culture. Given that a typical quantification experiment requires about 10 μg of the standard protein, we do not anticipate having to prepare each QconCAT more than once.

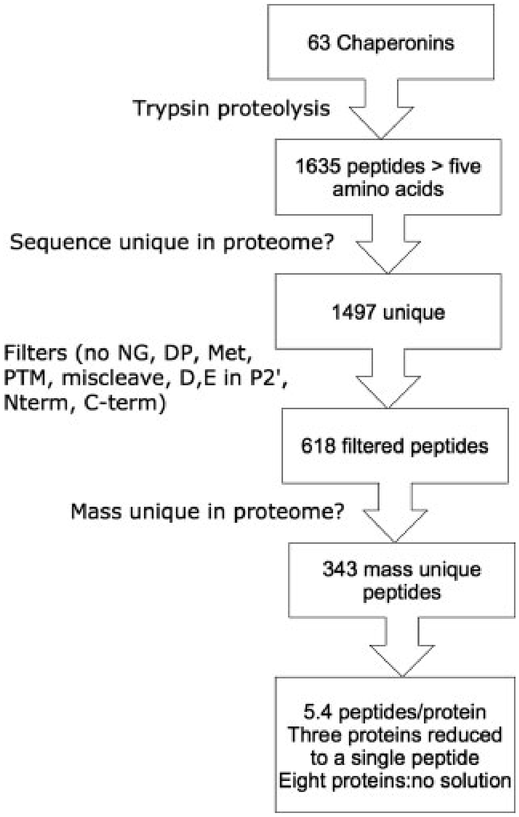
63 Chaperonins
Trypsin proteolysis
1635 peptides > five amino acids
Sequence unique in proteome?
1497 unique
Filters (no NG, DP, Met, PTM, miscleave, D,E in P2', Nterm, C-term)
618 filtered peptides
Mass unique in proteome?
343 mass unique peptides
5.4 peptides/protein
Three proteins reduced to a single peptide
Eight proteins:no solution

Figure 4. Attrition in the selection of optimal quantotypic peptides. The flowchart serves to illustrate how rapidly options fall away when seeking optimal quantotypic peptides. From a set of 63 chaperonins, after mass, composition, sequence and uniqueness filters have been applied, only 343 peptides remained as candidates. However, for three proteins, only one peptide remained and for a further eight, no peptides met all criteria. The filters are relaxed until an optimal solution is found for the residual proteins, and/or candidates selected, which are frequently observed in repositories such as PeptideAtlas (www.peptideatlas.org). Abbreviations for selection criteria are described in the legend of Fig. 2.

For quantitative studies, we use the QconCATs in stable isotope labelled form, and they are labelled by expression in minimal medium containing [$^{13}C_6$]arginine and [$^{13}C_6$]lysine. Bacterial growth is comparable to that obtained in rich media, and the yield of QconCAT is similar. Labelling is effective and tryptic fragments are labelled as extensively as the precursor amino acid.

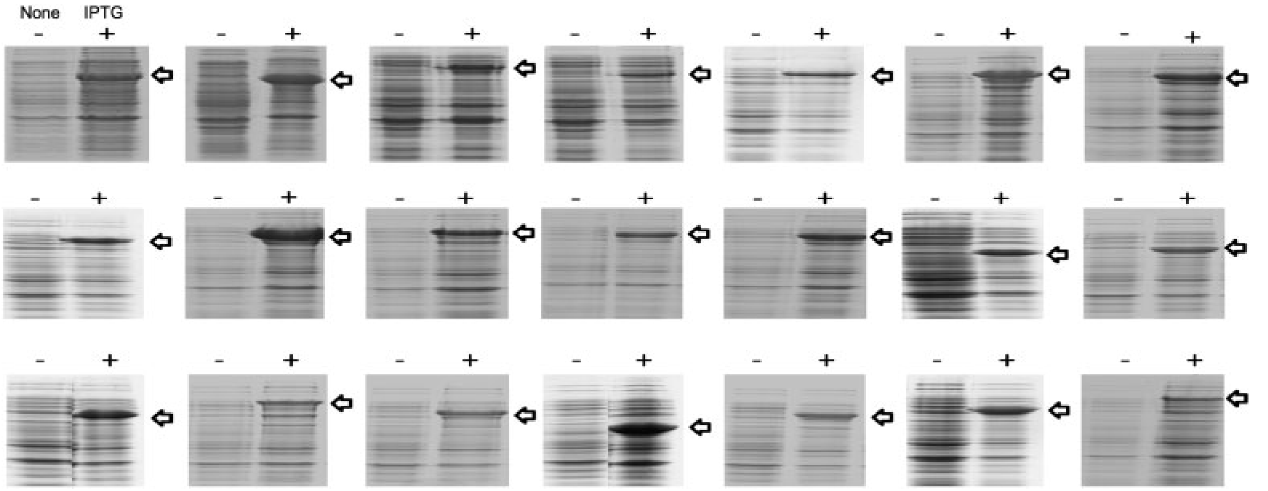

Figure 5. Expression of QconCATs. The QconCATs genes are synthesised and cloned into the pET21a vector, bringing expression under control of the *lacZ* promoter. Cells are grown to an OD of 0.6 (−) and at this stage, inducer (IPTG) is added to initiate QconCAT synthesis (+). For each of the QconCATs, a strong band (indicated by the arrow) is evident after a few hours of induction.

8 Quantification strategy

The objective of the MS analyses is to provide accurate and precise estimates of the abundance ratio of unlabelled and stable isotopically labelled peptide analogues, permitting the determination of the molar concentration of the proteins from which Qpeptides are derived. The mode of MS analysis employed is driven by the concentrations of the proteins of interest in the biological sample and by the precision required in the quantitative estimate.

While some peptides from the more abundant proteins might be analysed as heavy:light pairs of peptide ions in an accurate mass retention time strategy, we anticipate that the majority of quantification reactions will be obtained by MS/MS either through recording full product ion spectra [8] or (to achieve maximum sensitivity and hence precision) using SRM (for an excellent review, see [35]). The instrument configuration that provides the maximum duty cycle in SRM (assuming the co-detection of a small number of ions during a single chromatographic time window) is the tandem quadrupole (quadrupole/hexapole/quadrupole) instrument often referred to as a ''triple quad'' or QqQ instrument.

To achieve the lowest limit of quantification the entire duty cycle of the instrument must be focused on a single transition. This type of analysis would be far too slow to allow full proteome coverage. A balance of time and sensitivity can be achieved by the use of scheduled transitions where each peptide-specific set of transitions is only monitored in a time window around the chromatographic retention time of the peptide. A scheduled SRM workflow therefore requires more preparation because both the optimum transitions and chromatographic properties of each peptide must be determined prior to analysis. A QconCAT-based approach aids this preparation by providing an abundance of readily obtained material for characterisation of the heavy peptides. The target analyte light peptide will share identical chromatographic and fragmentation behaviour to the QconCAT-derived heavy peptide so light transitions can be generated from the optimised heavy transitions by simply adjusting the precursor and product ion m/z values to remove the stable isotope label contribution.

It is instructive to consider the limitations of a fully quantitative proteome analysis. Irrespective of the number of cells that are available at the start of the experiment the ''pinch point'' is the quantity of digest that can be applied to a one-dimensional reversed-phase chromatography column. Typically, this equates to the quantity of protein that is derived from ca. 200 000 yeast cells. A protein that is present at a level of ten copies per cell would therefore be applied on column at a total of 2 million copies, or just over 3 amol. Precise quantification of this level would require, minimally, a modest S/N ratio of at least 10:1. Therefore the peptides that are nominated for quantification should in principle be capable of such sensitive detection. If the lower limit of quantification is set to 100 amol (equivalent to 60 000 000 molecules on column), then the lowest limit of quantification will be equivalent to an average of 300 copies per cell (Table 2).

However, achieving an S/N ratio of 10:1 (which brings quantification data to within about 10% of the true value) is challenging at low loadings of standards or analytes, especially when delivered in a ''dirty'', complex analyte stream. We therefore assess ''operational S/N'' loadings as the value obtained as a specific loading of standard. For a typical set of about 170 peptides we find that notwithstanding the selection of ''optimal peptides'', many of these fail to give an S/N ratio of 10:1 at 100 amol loads (Fig. 6). Although it is difficult to obtain threshold S/N ratios at low loads, there is a

Table 2. Achieving deep proteome quantification

Cells on column (protein)	Lowest acceptable sensitivity			
	1 amol	10 amol	100 amol	1 fmol
	Limit of quantification (copies per cell)			
100 000 (500 ng)	6	60	600	6000
200 000 (1 μg)	3	30	300	3000
500 000 (2.5 μg)	2	12	120	1200
1 000 000 (5 μg)	<1	6	60	600

The table is constructed to reflect the requirements for deep proteome quantification for cells of the same size as *S. cerevisiae* (assuming 5 pg of total protein per cell). It correlates maximal capacity of a reversed-phase column (typical loadings of 1 μg are routine for most capillary columns) and the lower limit of quantification (as an unspecified *S/N* ratio, which will also dictate the quality of the quantification). For larger cells (e.g. mammalian cells, typically containing 250 pg protein) the sensitivity in copies per cell is commensurately reduced (50-fold).

grey area where quantification is achieved but at an *S/N* below threshold. These peptides represent the most challenging to quantify so even lower quality quantification data is likely to be beneficial with the caveat that the operational *S/N* must be reported for the data to be properly interpreted. The peptides in this grey area would also probably achieve the lower limit of quantification *S/N* ratios with additional sample prefractionation.

There are several explanations for poorly performing peptides. The first is that the peptide might not fragment well, which would lead to a lack of sensitivity in SRM-based assays even though the peptide may have acceptable performance at the MS level. Second, some peptides have poor chromatographic behaviour; they fail to elute in a sharp peak and instead elute in a broad "hump" that reduces sensitivity. A further possibility is that a combination of few candidate peptides passing our filters and protein homology has forced the selection of sub-optimal peptides. In its evolutionary history yeast has undergone genome duplications and so most genes have homologues that can reduce the selection of unique peptides.

It has been helpful for us to adopt a terminology that reflects the outcome of every peptide-level quantification. "Type A" (also known as "S^+/A^+", or "standard positive, analyte positive") quantifications reflect the optimal outcome, when both standard and analyte peptide deliver high quality quantification data. "Type B" quantification analyses (or "S^+/A^-") reflect a quantification run where the standard delivered an acceptable signal but no useable analyte signal was obtained – this sets the upper limit on the protein abundance, but the true value could range from zero to this value. "Type C" (or S^-/A^-) analyses refer to the rare situations where neither standard nor analyte reveal acceptable SRM data. Since peptides are chosen on the basis of their digestibility, ionisation and MS performance, it is perhaps unsurprising that sometimes the MS/MS performance is suboptimal. For one set of four QconCATs, we obtained, from 167 peptides: 114 Type A analyses (68%), 34 Type B analyses (21%) and 19 Type C analyses (11%). The Type B analyses set the lower limit of quantification and are in principle recoverable by enrichment strategies or increased instrument sensitivity. Type C analyses reflect a failure to select a peptide with high-quality fragmentation or chromatographic properties (Fig. 7). Some of these Type C peptides might be useable in accurate mass retention time quantifications, provided they map to relatively high abundance proteins. The shorthand nomenclature can also be applied to protein quantification. Thus, at a QRL of 2, a protein that is an "AA" has the highest confidence level, but "CC" protein quantifications have failed to deliver quantification data. Intermediate scores ("AB", "AC", "BB" or "BC") are interpreted and processed further according to the needs of the study.

Quantification performance can also be assessed by comparing quantification results from the sibling peptides (those peptides used in the QconCAT derived from the same protein). For example, if one peptide elicits a Type A performance and the second peptide elicits a Type B performance we will tend to emphasise the Type A quantification – it is simpler to explain the loss of analyte signal than the loss of standard signal. If the quantification values from each peptide are plotted against each other they should align along the equality diagonal, but as might be anticipated there is considerable scatter around the diagonal (Fig. 8). The main explanations for outliers are either low *S/N* ratio for one peptide or a diminution in peptide signal attributable to incomplete proteolysis. Miscleavage can occur in either the target protein, leading to apparently lower quantification values, or in the QconCAT, leading to apparently higher quantification values. Detection of miscleaves within the QconCAT is analytically simpler as increased column loadings of QconCAT can be used to detect miscleaves, which presumably are present at lower abundances. Finding miscleaves in the target protein is more challenging as it is difficult to increase column loading to bring the miscleaved peptide within the detection range of the analytical platform. A second cause of a low signal in the analyte peptide could be PTM of the target sequence, although the QconCAT design reduces the likelihood of this explanation by rejecting known PTM sites or consensus sites. With low abundance target proteins it is more likely that there will be an absence of confirmatory sequence data, making explanation of quantification differences difficult. It

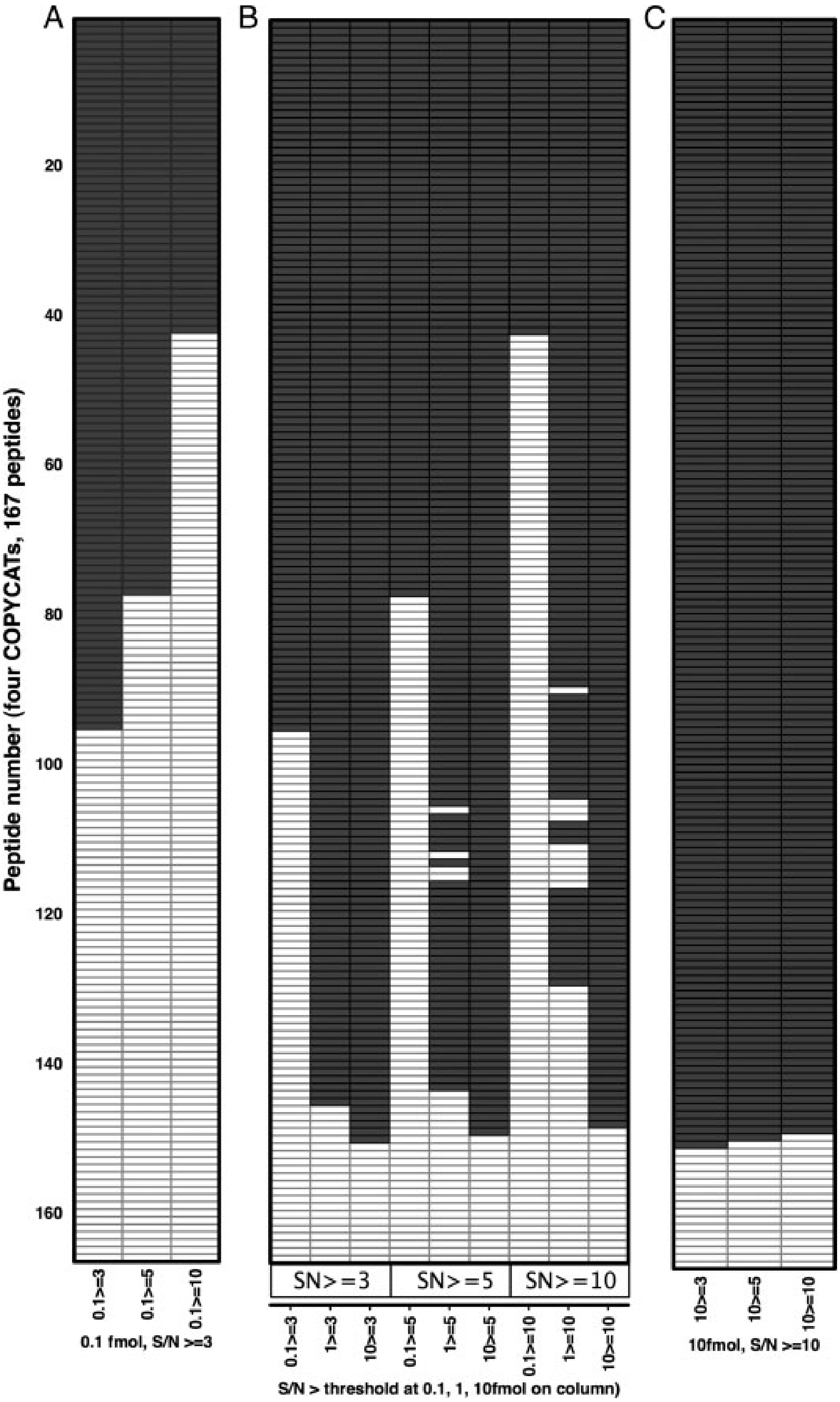

Figure 6. Performance of QconCAT peptides. A set of peptides from four QconCATs were evaluated for their performance, assessed as *S/N* ratio at three different loadings (10, 1, 0.1 fmol) applied in a background of 500 ng of yeast protein digest. Peptides are ranked according to their S/N, categorised as ≥3, ≥5 or ≥10 and peptides shaded black pass the threshold criterion. Panel (A) displays the peptides in relation to column load of 0.1 fmol for all *S/N* values, panel (C) does the same for 10 fmol column load. The centre panel (B) summarises the data according to *S/N* and load.

is still possible to obtain quantification even in situations where the peptides show disagreement; if high loadings of QconCAT reveal no miscleaves in the QconCAT then the problem can be assumed to be with the target peptide, reducing quantification, so the higher quantification of the two peptides can be accepted.

However, this comparison, obtained from an early set of QconCATs does raise a critical issue. For a substantial number of quantification analyses, the sibling peptides yielded discordant quantification data. Apart from casting some doubt over other quantification studies that are based on a single peptide, it could be argued that even two peptides are inadequate. Raising the QRL (Table 1) to 3 or higher has a major impact on the number of QconCATs that would be required. Moreover, it is much less probable that four peptides could be identified that were specific for each protein

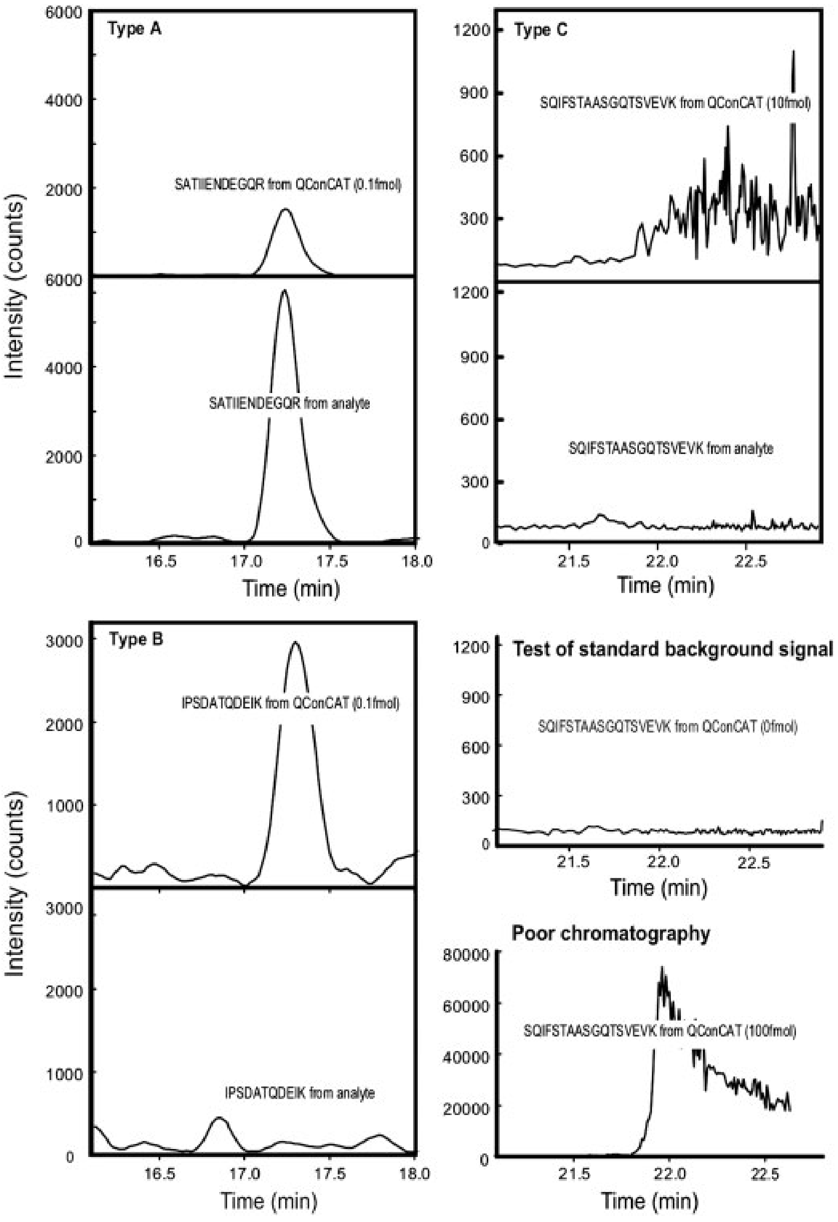

Figure 7. Classification of peptide performance in quantification. A peptide-level quantification can yield different outcomes, depending upon the performance of standard and analyte peptide. Type A quantifications are reliable, because both standard and analyte give confidently integrated peaks. Type B quantifications are the outcome of a well-behaved standard associated with a non-existent signal from the analyte – it is not possible to discern whether this is attributable to a very low level of protein expression or an unanticipated PTM or mutation that removes the analyte peptide from the SRM workflow. Type C peptides demonstrate poor performance, whether as standard or analyte. Two further tests (lower right are essential to demonstrate that the analyte background does not compromise the standard signal, and that the peptide has good chromatographic behaviour – the example (bottom right) is of a poorly performing peptide at high column load.

and well behaved. The extreme approach, of generating a protein-level standard for each analyte protein [16], giving the maximum opportunities for peptide-level quantification for each protein bring challenges of its own, not least the added complexity that a multiplexed approach would demand, and the need to quantify each intact protein standard.

There is one feature of this programme of research that sets it apart from most quantitative proteome studies. In this programme, we need to develop SRM conditions (standard peptides, their generation, optimisation of transitions) that, once optimised, are used for a very short time, for a limited number of samples, although they will remain in the public domain for others to use. This contrasts rather markedly with the typical development of an SRM assay, in which considerable time is devoted to the development of an assay that is then used extensively for many hundreds of subsequent analyses. This transient usage creates a rather different perspective on the balance between assay development and usage, a balance that will have to be addressed in any future quantification study.

9 Data verification, analysis and release

The project will generate a large volume of data for analysis and it is incumbent on us to verify the protein absolute quantitations and turnover rates where possible, and assign statistical significance to the values obtained. We can assign standard error and confidence values to Qpeptide-based quantifications, and these may be processed at the protein

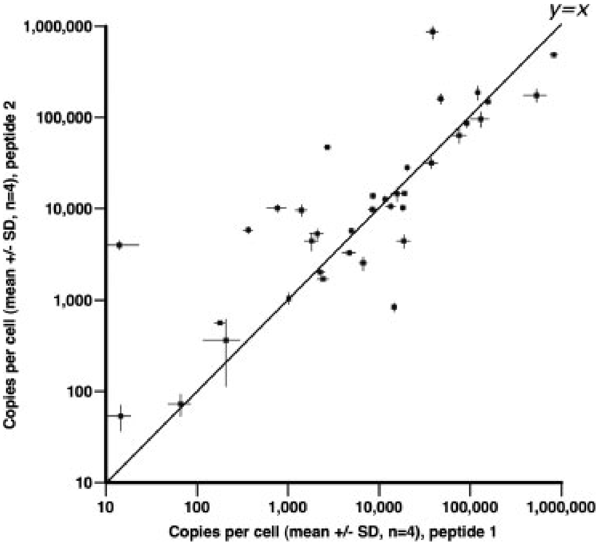

Figure 8. Correlation between sibling peptides in absolute quantification. For a set of 37 proteins from the chaperonin group, quantitative data were obtained from two different peptides derived from the same target protein (sibling peptides, mean ± SD, $n = 4$ four biological replicates). All quantification data should align along the equality diagonal. For off-diagonal points, three were caused by miscleaves in the COPYCAT, two by miscleaves in the target protein and eight were attributable to the uncertainty associated with a low *S/N* ratio.

level in a similar fashion to other groups (c.f. [36]) exploiting ion signal and peptide confidence scores to weight the attendant protein value over results from technical/biological replicates.

Data dissemination will be handled via two principal routes. We plan to make raw data available as soon as possible, once it has been quality controlled, by placing spectra and associated meta-data on a server for immediate access, using the Tranche platform (http://tranche.proteomecommons.org). Once processed and analysed, data will be delivered via the proteome identifications database (PRIDE, European Bioinformatics Institute, www.ebi.ac.uk). The PRIDE database is essential since it is the only repository compliant with HUPO PSI data standards and aims to capture the quantitative aspects [37]. A key aspect of PRIDE is its close links with the UniProt team – we anticipate a potential direct route for the quantitative and turnover data to reach UniProt/Swiss-Prot entries via the Feature Table or Protein Existence records. These records are classified as having five levels of confidence – we anticipate that the QconCAT MS data will achieve the highest level. This will also provide a dissemination route to reach biologists and bioinformaticians who will wish to use the data. Finally, the data will be shared with the Saccharomyces Genome Database team who will host the data and bring to bear their expertise in handling yeast data sets and integrating it into their knowledgebase, which is undoubtedly the best known to the biological community.

10 Direct and indirect QconCAT

Although QconCAT or AQUA approaches are needed to quantify an organism in a single state, once that quantification is complete, the need for QconCATs diminishes. The first phase of the quantification experiment, using QconCATs, we refer to as a ''direct'' absolute quantification. Once the direct quantification phase is complete and the cellular concentration of each protein is known (together with the attendant variance), it becomes feasible to use the same material as an absolute reference for future studies, termed the ''indirect'' approach. There are a number of advantages of indirect absolute quantification. First, the organism, not the QconCATs becomes the reference, and a broad range of labelling strategies can be used to resolve standard from analyte – including labelling in vivo with other amino acids or labelling in vitro using reagents such as iTRAQ. Second, because the standard now comprises the endogenous protein, differentially labelled, any peptides derived from the analyte and reference can be used, resolved by any of the currently accepted methodologies. In order to build a resource of greatest value to the community, we will explore the optimal conditions for indirect, as well as direct QconCAT quantification. We will also make all QconCAT plasmids available to the yeast community via the EUROSCARF (European *S. cerevisiae* Archive for Functional Analysis) collection.

The COPY project is supported by the grants from the Biotechnology and Biological Sciences Research Council to Liverpool (BB/G009112/1) and Manchester (BB/G009058/1). We are pleased to acknowledge the involvement of Lynn McLean, Duncan Robertson, and Julian Selley and the support of Jim Langridge and colleagues at Waters Micromass, Manchester.

The authors have declared no conflict of interest.

11 References

[1] Jungblut, P. R., Holzhutter, H. G., Apweiler, R., Schluter, H., The speciation of the proteome. *Chem. Cent. J.* 2008, *2*, 16.

[2] Schwab, K., Neumann, B., Scheler, C., Jungblut, P. R., Theuring, F., Adaptation of proteomic techniques for the identification and characterization of protein species from murine heart. *Amino Acids* 2011, *41*, 401–414.

[3] Ghaemmaghami, S., Huh, W. K., Bower, K., Howson, R. W. et al., Global analysis of protein expression in yeast. *Nature* 2003, *425*, 737–741.

[4] Newman, J. R., Ghaemmaghami, S., Ihmels, J., Breslow, D. K. et al., Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* 2006, *441*, 840–846.

[5] Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., Kuster, B., Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* 2007, *389*, 1017–1031.

[6] Li, Y. F., Arnold, R. J., Tang, H., Radivojac, P., The importance of peptide detectability for protein identification,

quantification, and experiment design in MS/MS proteomics. *J. Proteome Res.* 2010, *9*, 6288–6297.

[7] Old, W. M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K. G. et al., Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell. Proteomics* 2005, *4*, 1487–1502.

[8] Silva, J. C., Gorenstein, M. V., Li, G. Z., Vissers, J. P., Geromanos, S. J., Absolute quantification of proteins by LCMS^{E}: a virtue of parallel MS acquisition. *Mol. Cell. Proteomics* 2006, *5*, 144–156.

[9] Zhang, Y., Wen, Z., Washburn, M. P., Florens, L., Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins. *Anal. Chem.* 2010, *82*, 2272–2281.

[10] Weiss, M., Schrimpf, S., Hengartner, M. O., Lercher, M. J., von Mering, C., Shotgun proteomics data from multiple organisms reveals remarkable quantitative conservation of the eukaryotic core proteome. *Proteomics* 2010, *10*, 1297–1306.

[11] de Godoy, L. M., Olsen, J. V., Cox, J., Nielsen, M. L. et al., Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* 2008, *455*, 1251–1254.

[12] Michalski, A., Cox, J., Mann, M., More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J. Proteome Res.* 2011, *10*, 1785–1793.

[13] King, N. L., Deutsch, E. W., Ranish, J. A., Nesvizhskii, A. I. et al., Analysis of the *Saccharomyces cerevisiae* proteome with PeptideAtlas. *Genome Biol.* 2006, *7*, R106.

[14] Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I. et al., The PeptideAtlas project. *Nucleic Acids Res.* 2006, *34*, D655–D658.

[15] Picotti, P., Bodenmiller, B., Mueller, L. N., Domon, B., Aebersold, R., Full dynamic range proteome analysis of *S. cerevisiae* by targeted proteomics. *Cell* 2009, *138*, 795–806.

[16] Brun, V., Masselon, C., Garin, J., Dupuis, A., Isotope dilution strategies for absolute quantitative proteomics. *J. Proteomics* 2009, *72*, 740–749.

[17] Dupuis, A., Hennekinne, J. A., Garin, J., Brun, V., Protein standard absolute quantification (PSAQ) for improved investigation of staphylococcal food poisoning outbreaks. *Proteomics* 2008, *8*, 4633–4636.

[18] Hanke, S., Besir, H., Oesterhelt, D., Mann, M., Absolute SILAC for accurate quantitation of proteins in complex mixtures down to the attomole level. *J. Proteome Res.* 2008, *7*, 1118–1130.

[19] Mirzaei, H., McBee, J. K., Watts, J., Aebersold, R., Comparative evaluation of current peptide production platforms used in absolute quantification in proteomics. *Mol. Cell. Proteomics* 2008, *7*, 813–823.

[20] Duncan, M. W., Yergey, A. L., Patterson, S. D., Quantifying proteins by mass spectrometry: the selectivity of SRM is only part of the problem. *Proteomics* 2009, *9*, 1124–1127.

[21] Kirkpatrick, D. S., Gerber, S. A., Gygi, S. P., The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications. *Methods* 2005, *35*, 265–273.

[22] Beynon, R. J., Doherty, M. K., Pratt, J. M., Gaskell, S. J., Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides. *Nat. Methods* 2005, *2*, 587–589.

[23] Johnson, H., Eyers, C. E., Eyers, P. A., Beynon, R. J., Gaskell, S. J., Rigorous determination of the stoichiometry of protein phosphorylation using mass spectrometry. *J. Am. Soc. Mass Spectrom.* 2009, *20*, 2211–2220.

[24] Johnson, H., Wong, S. C., Simpson, D. M., Beynon, R. J., Gaskell, S. J., Protein quantification by selective isolation and fragmentation of isotopic pairs using FT-ICR MS. *J. Am. Soc. Mass Spectrom.* 2008, *19*, 973–977.

[25] Pratt, J. M., Simpson, D. M., Doherty, M. K., Rivers, J. et al., Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes. *Nat. Protoc.* 2006, *1*, 1029–1043.

[26] Rivers, J., Simpson, D. M., Robertson, D. H., Gaskell, S. J., Beynon, R. J., Absolute multiplexed quantitative analysis of protein expression during muscle development using QconCAT. *Mol. Cell. Proteomics* 2007, *6*, 1416–1427.

[27] Brownridge, P. B., Beynon, R. J., The importance of the digest: proteolysis and absolute quantification in proteomics. *Methods* 2011, in press.

[28] Pratt, J. M., Petty, J., Riba-Garcia, I., Robertson, D. H. et al., Dynamics of protein turnover, a missing dimension in proteomics. *Mol. Cell. Proteomics* 2002, *1*, 579–591.

[29] Pratt, J. M., Robertson, D. H., Gaskell, S. J., Riba-Garcia, I. et al., Stable isotope labelling in vivo as an aid to protein identification in peptide mass fingerprinting. *Proteomics* 2002, *2*, 157–163.

[30] Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S. et al., Global analysis of protein localization in budding yeast. *Nature* 2003, *425*, 686–691.

[31] Hubbard, S. J., Beynon, R. J., Thornton, J. M., Assessment of conformational parameters as predictors of limited proteolytic sites in native protein structures. *Protein Eng.* 1998, *11*, 349–359.

[32] Schechter, I., Berger, A., On the size of the active site in proteases. I. Papain. *Biochem. Biophys. Res. Commun.* 1967, *27*, 157–162.

[33] Siepen, J. A., Keevil, E. J., Knight, D., Hubbard, S. J., Prediction of missed cleavage sites in tryptic peptides aids protein identification in proteomics. *J. Proteome Res.* 2007, *6*, 399–408.

[34] Deutsch, E. W., Lam, H., Aebersold, R., PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep.* 2008, *9*, 429–434.

[35] Lange, V., Picotti, P., Domon, B., Aebersold, R., Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* 2008, *4*, 222.

[36] Park, S. K., Venable, J. D., Xu, T., Yates, J. R. r., A quantitative analysis software tool for mass spectrometry-based proteomics. *Nat. Methods* 2008, *5*, 319–322.

[37] Lau, K. W., Jones, A. R., Swainston, N., Siepen, J. A., Hubbard, S. J., Capture and analysis of quantitative proteomic data. *Proteomics* 2007, *7*, 2787–2799.