A quantitative analysis software tool for mass spectrometry–based proteomics

Sung Kyu Park[1,2], John D Venable[1,2], Tao Xu[1] & John R Yates III[1]

We describe Census, a quantitative software tool compatible with many labeling strategies as well as with label-free analyses, single-stage mass spectrometry (MS^1) and tandem mass spectrometry (MS/MS) scans, and high- and low-resolution mass spectrometry data. Census uses robust algorithms to address poor-quality measurements and improve quantitative efficiency, and it can support several input file formats. We tested Census with stable-isotope labeling analyses as well as label-free analyses.

In recent years, global quantification using mass spectrometry has garnered a great deal of interest as a result of the emergence of fields that rely on large-scale profiling of peptides or proteins (proteomics) and small molecules (metabolomics). In the field of proteomics, the identification of large numbers of peptides has become commonplace with the advent of new instrumentation[1–7] and informatics tools[8–11], but progress with regard to the quantification process has been hampered by computational challenges.

In general, peptide or protein quantification by mass spectrometry is performed either by stable-isotope labeling or a label-free approach. Stable-isotope labeling has become the core technology for high-throughput peptide quantification efforts that use mass spectrometry[12]. Quantification is typically achieved by comparison of an unlabeled or 'light' peptide (comprised of naturally abundant stable isotopes) to an internal standard that is chemically identical with the exception of atoms that are enriched with a 'heavy' stable isotope. Although the stable-isotope labeling approach has been the most commonly used over the past several years, label-free approaches have been gaining momentum recently because of their inherent simplicity, increased throughput and low cost. Several strategies for label-free differential expression analysis have emerged and can generally be divided into two groups; those that are fundamentally based on MS/MS identification of peptides before quantification and those that rely on MS^1 data alone[13].

Here we describe a new software tool for quantitative analysis called Census and discuss its potential impact for analyzing quantitative mass spectrometry proteomic data. What differentiates Census from other quantitative tools is its flexibility to handle data generated using most types of quantitative proteomics labeling strategies such as ^{15}N, stable isotope labeling of amino acids in culture (SILAC), isobaric tags for relative and absolute quantitation (iTRAQ; Applied Biosystems) and others, as well as label-free experiments (**Fig. 1**). Census is based on a program previously written in our laboratory called RelEx[14], but was re-written with many new features that improve the accuracy and precision of resulting measurements and the computational performance (**Supplementary Methods** and **Supplementary Table 1** online). Census is capable of quantification from either MS^1 or MS/MS scans and is thus able to process data generated from data-independent acquisition[15] or single-reaction monitoring analyses. Other features incorporated into Census include the ability to use high-resolution and high-mass-accuracy mass spectrometry data for improved quantification as well as the ability to perform quantitative analyses based on both spectral counting and a liquid chromatography–mass spectrometry (LC-MS) peak area approach using chromatogram alignment. To minimize false positive measurements and improve protein/peptide ratio accuracy, Census incorporates multiple algorithms such as weighted peptide measurements, dynamic peak finding and post-analysis statistical filters. Census also has a feature to detect singleton peptides (in which one isotopomer signal is below the detection limit). Census currently supports several instrument-independent input file formats including MS1/MS2, DTASelect, mzXML and pepXML (**Supplementary Fig. 1** and **Supplementary Table 2** online).

It is often impossible to distinguish isotopes in low-resolution mass spectrometry data for large peptides or peptides with high charge states. Thus, it is common to simply sum up all ion intensities within the predicted isotope distribution's *m/z* range. However, Census can take advantage of high-resolution and high-accuracy data by accurately predicting peptide molecular weights and corresponding *m/z* values and using a mass accuracy tolerance (**Supplementary Methods**). By using this strategy, noisy peaks or co-eluting peptides can be excluded. The mass accuracy tolerance can be user-defined in the Census configuration file. To achieve this, Census uses two extraction methods: "whole isotope envelope" (**Fig. 2a**) and "individual isotopes" (**Fig. 2b**). The first method is used with low-resolution data and extracts all peaks within the *m/z* range defined by the isotope envelope with greater than 5% of the calculated isotope cluster base peak abundance. The second method is used with high-resolution data and extracts individual isotopes using a mass accuracy tolerance. Noise peaks are easily

[1]Department of Cell Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, SR11, La Jolla, California 92014, USA. [2]These authors contributed equally to this work. Correspondence should be addressed to J.R.Y. (jyates@scripps.edu).

RECEIVED 28 AUGUST 2007; ACCEPTED 4 FEBRUARY 2008; PUBLISHED ONLINE 16 MARCH 2008; DOI:10.1038/NMETH.1195

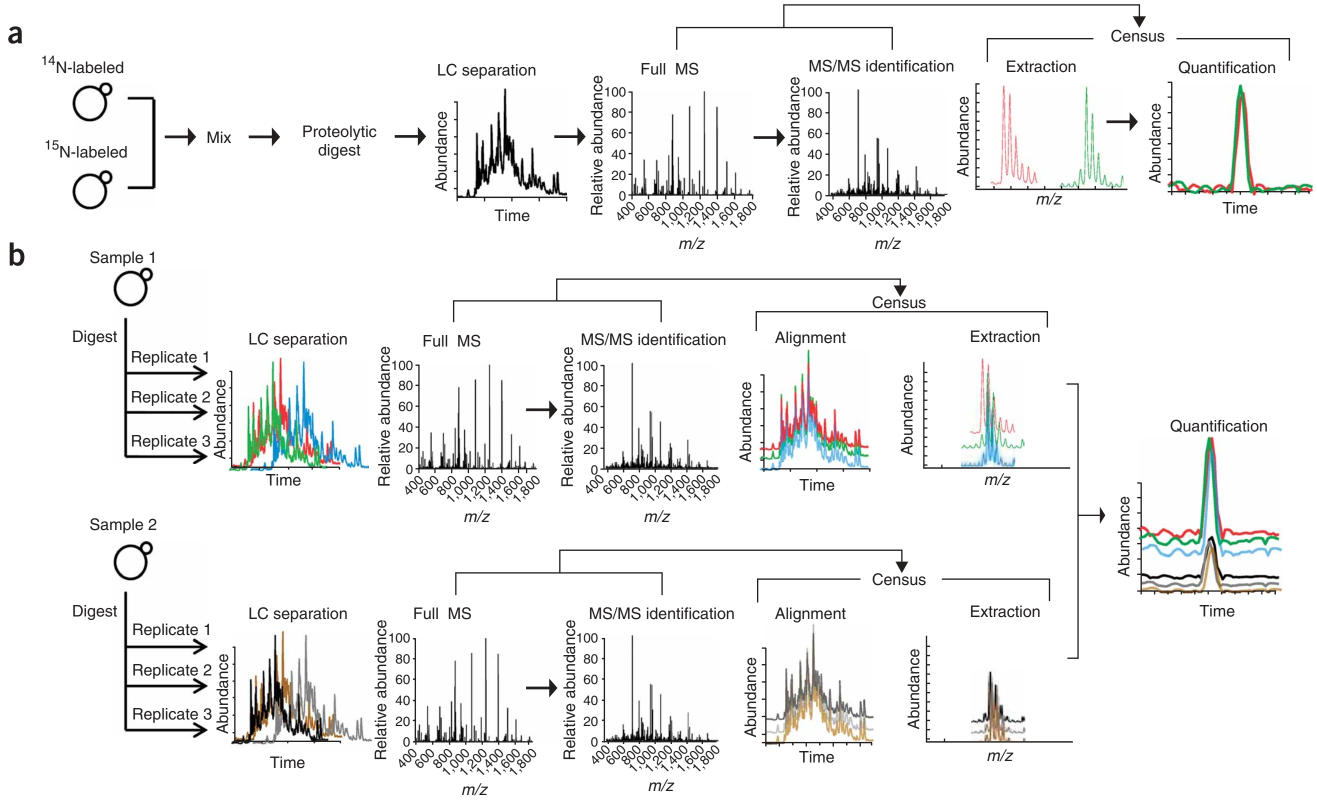

Figure 1 | Schematic detailing the quantitative analysis capabilities of Census. (**a**) Use of Census with isotopic labeling. (**b**) Use of Census with label-free analysis. LC, liquid chromatography.

excluded by these approaches; as result, the correlation becomes high, and the chromatograms are simple and track each other quite well (**Fig. 2**).

Quantification using MS/MS has been used extensively over the past three decades because of several key benefits including a reduction in chemical noise, increased specificity and increased sensitivity (when ion-trap mass spectrometry instruments are used) over MS[1]. In addition, the presence of multiple fragment ions has potential benefits for quantitative analysis, where ion intensities from several transitions can be summed to produce signal to noise enhancements[16] or averaged to obtain more accurate measurements[17]. Typically these experiments are performed in a directed fashion where precursor and MS/MS transitions are predetermined. One of the difficulties inherent to this approach is the selection of fragment ions that are to be monitored. Alternatively, full MS/MS scans can be acquired and chromatograms can be reconstructed. Census facilitates automated quantification from tandem mass spectra by optimizing the process of chromatogram reconstruction (**Supplementary Methods**). To do this, Census incorporates a filtering strategy that considers theoretical fragment ions and removes those that fall below a dynamic threshold. Remaining chromatograms are summed to increase sensitivity while selectivity is maintained by the filtering process. This strategy effectively filters noisy fragment ion chromatographic profiles in an automated fashion and can help to improve quantification of noisy or low-abundance peptides (**Supplementary Fig. 2** online).

As an initial evaluation of Census, we examined a collection of unlabeled and metabolically ^{15}N-labeled yeast cell lysates that we mixed in known ratios (1:1, 5:1 and 10:1; **Supplementary Methods**, **Supplementary Fig. 3**, **Supplementary Table 3** and **Supplementary Data** online). The ratios measured by Census were generally accurate for each of the standards analyzed (the average ratios were 1.07, 5.30 and 12.27 for the 1:1, 5:1 and 10:1 standards, respectively).

We also compared two different approaches for calculating protein ratios. For the first approach, we used the mean of all peptide measurements. The second strategy used a weighted average in which the individual peptide weights were determined by the inverse square of the standard deviation of the measurement (**Supplementary Methods** and **Supplementary Fig. 4** online). A comparison of these approaches revealed that the simple average approach underestimated the actual abundance for many measurements whereas the weighted average provided more accurate protein abundance measurements (**Supplementary Fig. 5** online). Census displays the weighted average for the protein abundance in a peptide distribution plot. This comparison shows how lower-quality measurements with low determinant factors have less impact on the calculated protein abundance than peptide measurements with high determinant factors.

The general strategy for the label-free quantification method used by Census is outlined in **Supplementary Methods** and in **Figure 1b**. We evaluated the peptides after first taking the union of search results so that a peptide need only be identified in one of the replicates to be quantified. Census uses a Pearson correlation between mass spectra and dynamic time warping (**Supplementary Methods** and **Supplementary Fig. 6** online) for chromatogram

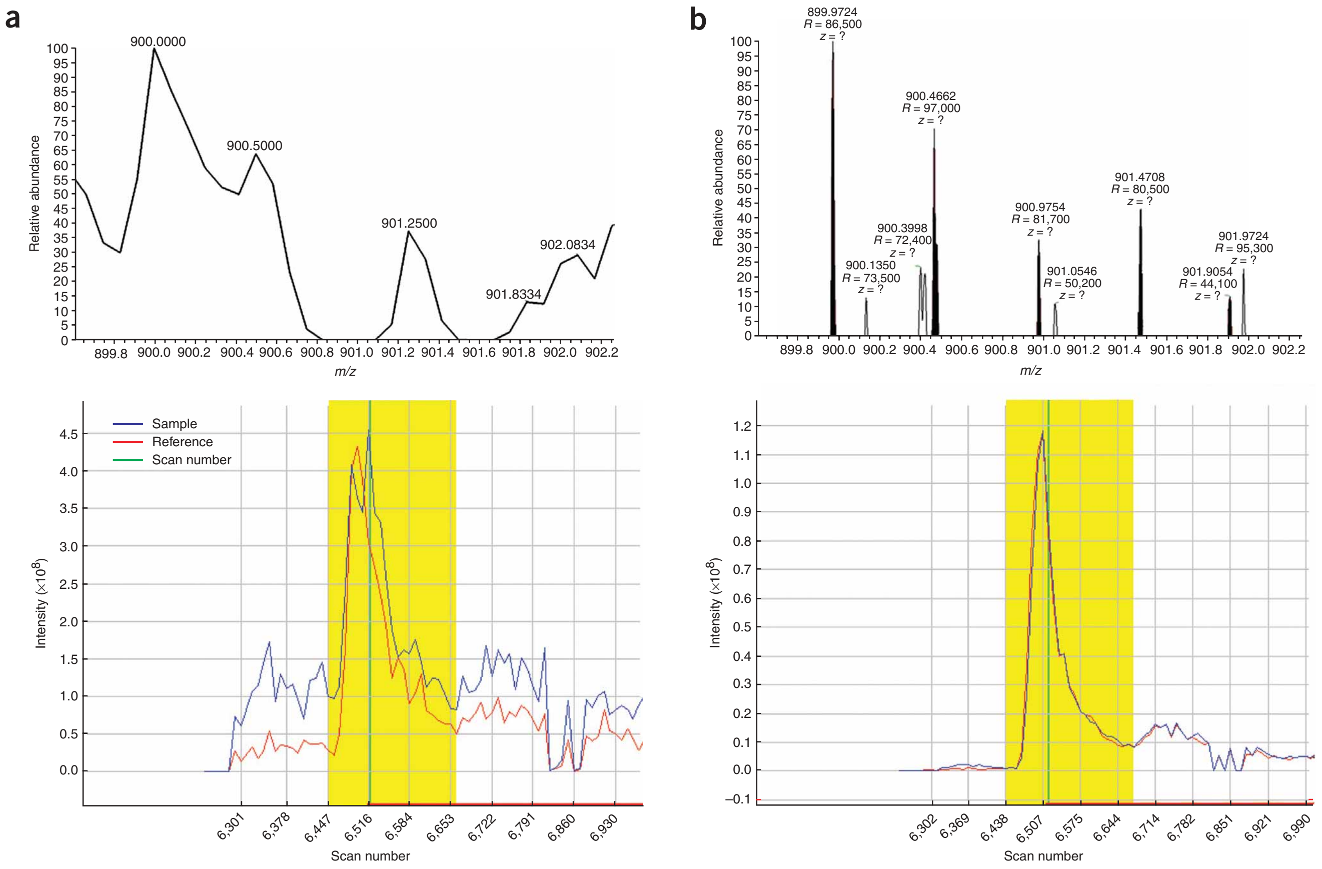

Figure 2 | Use of high mass accuracy for improved quantification with Census. (**a**) The 'whole isotopic envelope' method for extraction, using low-resolution (LTQ) mass spectrometry scans. (**b**) The 'individual isotope' extraction method, using high resolution (Orbitrap) mass spectrometry scans and a mass accuracy tolerance of 5 p.p.m. The green line in the chromatogram represents the identified scan, a close up of which is shown directly above the chromatograms. Yellow area shows scan range detected by peak-finding algorithm. White peaks in the spectrum are noise. *z*, charge state, is undefined.

alignment. Census can perform quantitative analyses based on both spectral counting and an LC-MS peak area approach using chromatogram alignment. To test the label-free quantification capabilities of Census, we analyzed four technical replicates of each sample of the 10-protein standard mixture with reversed-phase chromatography coupled to a linear trap quadrupole (LTQ) Orbitrap (Thermo Fisher; **Supplementary Methods** and **Supplementary Table 4** online). Using the mass spectrometry–based spectral alignment strategy, Census accurately quantified ∼70% of peptides (within a factor of 2 of the expected relative abundances). Protein abundance measurements were typically within 25% of the expected relative abundances, although the deviations for ovalbumin and β-casein were larger for unknown reasons (**Fig. 3**). Spectral counting has been shown to be useful as a semiquantitative measure of protein abundance. We found that protein relative abundances obtained from spectral counts generally correlated well with those obtained from peak areas, although the accuracy tended to be slightly worse for the former approach (**Fig. 3**).

In addition to profiling-type experiments, Census can also perform quantification from tandem mass spectra. To test this capacity, we mixed two standard peptides labeled with iTRAQ reagents and quantified them in three different mixtures (1:1, 1:4 and 4:1) (**Supplementary Methods**). Because iTRAQ is often used in the context of a data-dependent experiment, Census can be configured to calculate relative abundances using either reconstructed chromatograms or the relative intensities of reporter ions from a specific identified tandem mass spectrum. In general, the results obtained using the chromatographic profiles were more reproducible and reliable, and led to more accurate quantification (**Supplementary Fig. 7** online).

In cases in which the intensity of either the light or heavy isotopomer is below the detection limit, the correlation coefficient is typically low. As a consequence, proteins with very large differences in abundance can be penalized by the low determinant scores (R^2) of their respective peptide measurements. To address this limitation, Census uses a linear discriminant analysis to detect such singleton peptides (**Supplementary Methods**). To test this approach, we analyzed a sample from a two-step affinity purification strategy targeting human RNA polymerase II, which had been differentially labeled using SILAC.

We expected RNA polymerase and its associated proteins to be preferentially enriched, which would lead to a large abundance difference between the light and heavy isotopomers. Peptides derived from nonspecific interactions would not be enriched and would have similar abundances. Common contaminants (such as keratin proteins) would be enriched in the light sample because they are not derived from the cell lines used. Notably, over 60% of peptides from the RNA polymerase isoforms identified had

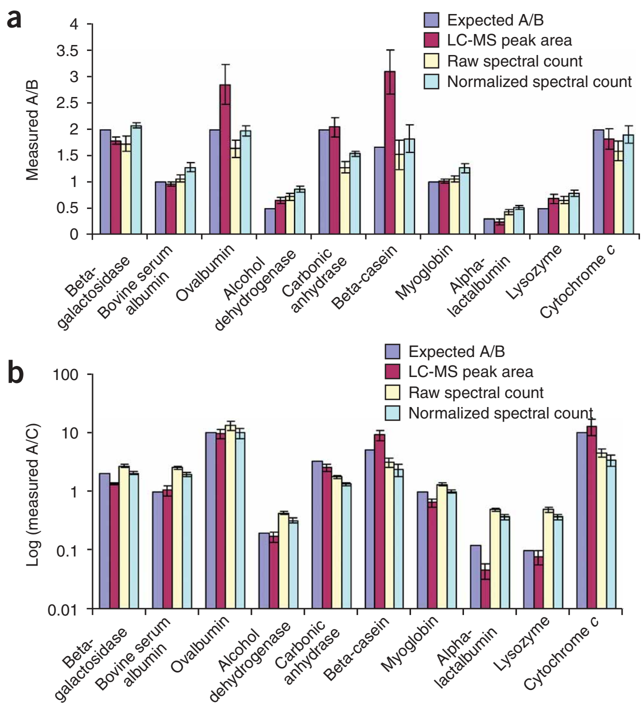

Figure 3 | Expected and measured relative abundances of technical replicates of a 10-protein mix dataset using Census. (**a**) Ratio of the signals measured for a mixture of sample A over sample B. (**b**) Ratio of the signals for a mixture of sample A over that of sample C using different strategies including LC-MS peak areas, spectral counting without normalization and spectral counting with normalization. A total of four replicate analyses were performed for each mixture and variance was determined as the standard deviation.

determinant scores (R^2) > 0.5 suggesting that the linear regression technique is often applicable even when the signal-to-noise ratio of the isotopomers is extremely low (**Supplementary Fig. 8** online). Consequently all 6 identified RNA polymerase proteins were quantified as having large abundance changes even without the singleton strategy. However, when we used the linear discriminant analysis algorithm to detect singleton peptides, we were able to detect 12 of 15 keratin proteins and isoforms, whereas we detected only 3 using the linear regression approach. Using the linear discriminant analysis approach with a threshold for the discriminant score of 0.94, we detected 153 true singleton peptides, 6 false ones (4% false positive rate) and no non-singleton proteins (proteins besides RNA polymerase and keratins). We are working on improving the singleton peptide detection methodology (**Supplementary Methods**, and **Supplementary Tables 5** and **6** online).

Quantitative analysis has become increasingly popular and important in the field of proteomics research and has fostered the development of quantitative software to expedite and validate the data generated. Census is flexible enough to use with various types of quantitative experiments; it is fast and accurate and should be a valuable tool for quantitative analysis. This software is available at http://fields.scripps.edu; academic and nonprofit use is free of charge.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank J. Wohlschlegel and D. Cociorva for their comments and discussions about the manuscript, M. MacCoss for his insights and work developing RelEx, S. Agarwalla and D. McMullan for sample preparation. S.K.P. is supported by US National Institutes of Health (NIH) National Institute of Allergy and Infectious Diseases (UOM/DMID-BAA-03). J.D.V. is supported by a National Research Service Award (NIH) fellowship. T.X. is supported by NIH (DE016267). J.R.Y. is supported by NIH (P41 RR011823, R01 MH067880 and R01 HL079442).

AUTHOR CONTRIBUTIONS

S.K.P. developed the algorithm and applications; J.D.V. prepared yeast cell lysates, 10 protein standard mixtures, iTRAQ samples and developed algorithms; T.X. contributed to discussions on algorithms; J.R.Y. led and coordinated the project.

Published online at http://www.nature.com/naturemethods/
Reprints and permissions information is available online at
http://npg.nature.com/reprintsandpermissions

1. Makarov, A. *et al. Anal. Chem.* **78**, 2113–2120 (2006).
2. Olsen, J.V. *et al. Mol. Cell. Proteomics* **4**, 2010–2021 (2005).
3. Yates, J.R., Cociorva, D., Liao, L. & Zabrouskov, V. *Anal. Chem.* **78**, 493–500 (2006).
4. Denison, C. *et al. Mol. Cell. Proteomics* **4**, 246–254 (2005).
5. Dieguez-Acuna, F.J. *et al. Mol. Cell. Proteomics* **4**, 1459–1470 (2005).
6. Foster, L.J., Hoog, C.L.d. & Mann, M. *Proc. Natl. Acad. Sci.* **100**, 5813–5818 (2003).
7. Venable, J.D., Wohlschlegel, J., McClatchy, D.B., Park, S.K. & Yates, J.R. III. *Anal. Chem.* **79**, 3056–3064 (2007).
8. Eng, J.K., McCormack, A.L. & Yates, J.R., III. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
9. Sadygov, R.G. & Yates, J.R., III. *Anal. Chem.* **75**, 3792–3798 (2003).
10. Tabb, D.L., Saraf, A. & Yates, J.R., III. *Anal. Chem.* **75**, 6415–6421 (2003).
11. Geer, L.Y. *et al. J. Proteome Res.* **3**, 958–964 (2004).
12. Ong, S.E. & Mann, M. *Nat. Chem. Biol.* **1**, 252–262 (2005).
13. Domon, B. & Aebersold, R. *Science* **312**, 212–217 (2006).
14. MacCoss, M.J., Wu, C.C. & Yates, J.R., III. *Anal. Chem.* **75**, 6912–6921 (2003).
15. Venable, J.D., Dong, M.Q., Wohlschlegel, J., Dillin, A. & Yates, J.R. *Nat. Methods* **1**, 39–45 (2004).
16. Owens, K.G. *Appl. Spectrosc. Rev.* **27**, 1–49 (1992).
17. Arnott, D. *et al. Mol. Cell. Proteomics* **1**, 148–156 (2002).