A Catalogue of Human Saliva Proteins Identified by Free Flow Electrophoresis-based Peptide Separation and Tandem Mass Spectrometry*

Hongwei Xie‡, Nelson L. Rhodus§, Robert J. Griffin¶, John V. Carlis‖, and Timothy J. Griffin‡**

Human saliva has great potential for clinical disease diagnostics. Constructing a comprehensive catalogue of saliva proteins using proteomic approaches is a necessary first step to identifying potential protein biomarkers of disease. However, because of the challenge presented in cataloguing saliva proteins with widely varying abundance, new proteomic approaches are needed. To this end, we used a newly developed approach coupling peptide separation using free flow electrophoresis with linear ion trap tandem mass spectrometry to identify proteins in whole human saliva. We identified 437 proteins with high confidence (false positive rate below 1%), producing the largest catalogue of proteins from a single saliva sample to date and providing new information on the composition and potential diagnostic utility of this fluid. The statistically validated, transparently presented, and annotated dataset provides a model for presenting large scale proteomic datasets from future studies in saliva.  

Molecular & Cellular Proteomics 4:1826–1830, 2005.

Clinicians and researchers value human saliva as potentially the ultimate bodily fluid for clinical disease diagnosis and prognostic monitoring (1, 2). To identify potential protein biomarkers in saliva, construction of a comprehensive protein catalogue in this fluid is a necessary first step. Several mass spectrometry-based proteomic studies have begun this task. These studies used different strategies, including two-dimensional gel-based analysis (3–6) and liquid chromatography-based analysis coupled with MS/MS (7, 8). The wide range of protein abundance in whole saliva makes cataloguing of saliva proteins using any of these strategies a challenge (9).

Therefore, to obtain a more comprehensive catalogue of saliva proteins innovative proteomic approaches are needed.

Recently we described a new approach (10) to proteomic analysis that uses preparative IEF by free flow electrophoresis (FFE)† (11, 12) for a first dimension fractionation of complex peptide mixtures. The use of FFE not only provides a high resolution peptide separation, but also it adds a constraint of peptide pl information to the determination of peptide sequence matches in the sequence database search of the MS/MS data, significantly improving the confidence of the peptide sequence matches and effectively increasing the number of high confidence protein identifications (10, 13–15).

The goal of this study was to use peptide separation by FFE coupled with a linear ion trap mass spectrometer to comprehensively identify proteins in whole human saliva. We identified 437 proteins with high confidence, providing the largest catalogue of proteins from a single saliva sample to date. The protein catalogue provides new information on the composition of this bodily fluid and its potential utility in disease diagnostics. The statistically validated and transparently presented dataset (shown in the supplemental table) provides a model for presenting large, mass spectrometry-based proteomic data that should provide improved dissemination and comparison of datasets in this clinically important biological fluid.

EXPERIMENTAL PROCEDURES

Clinical Saliva Collection and Protein Preparation—Whole unstimulated saliva was collected from a healthy female subject in the University of Minnesota Oral Medicine Clinic using a protocol described previously (16). 1 ml of whole saliva was removed and centrifuged at 25,000 × g and 4 °C for 30 min. The supernatant was collected and quantified by using the BCA protein assay (Pierce), giving 1.05 mg of total soluble proteins. The saliva was brought to 100 mM with HEPES, pH 8.0 and 5 mM with Tris(2-carboxyethyl)phosphine and incubated overnight with 20 μg of trypsin (Promega, Madison, WI) at 37 °C. The resulting peptides were concentrated and desalted using a reverse-phase Sep-Pak cartridge (Waters, Milford, MA) and dried by vacuum centrifugation.

†The abbreviations used are: FFE, free flow electrophoresis; μLC, microcapillary LC.
FFE Fractionation of Peptides and Sample Processing—Preparative IEF of the peptide mixture was performed using a commercially available Pro Team free flow electrophoresis system (BD Biosciences) (11, 12). The saliva peptides were dissolved in 250 μl of FFE separation buffer and fractionated by FFE into a 96-well microtrict plate as described previously (10). Immediately after FFE separation, the pH of each FFE fraction was measured using a microelectrode (Accumet combination microelectrode, Fisher). A 50-μl aliquot (of ~500 μl total) was taken from each of the microtrict plate wells and processed as described previously (10) prior to mass spectrometric analysis.

μLC-ESI MS/MS Analysis—All μLC separations were done on an automated Paradigm MS4 system (Michrom Bioreources, Inc., Auburn, CA). Each processed FFE fraction was automatically loaded across a Paradigm Platinum Peptide Nanotrap (Michrom Bioreources, Inc., Burn, CA). Each processed FFE fraction was automatically loaded onto an automated Paradigm MS4 system (Michrom Bioresources, Inc., Auburn, CA) and Picofrit capillary tubing (New Objective, Cambridge, MA). Peptides were eluted using a linear gradient of 10–35% buffer B over 60 min followed by isocratic elution at 80% buffer B for 5 min with a flow rate of 0.25 μl/min across the column.

Peptides were analyzed by MS/MS using a linear ion trap mass spectrometer system (LTQ, Thermo Electron Corp., San Jose, CA). The electrospray voltage was set to 2.0 kV using a collision energy setting of 29% and a data-dependent procedure that alternated between one MS scan (over the m/z range of 400–1800) followed by four MS/MS scans for the four most abundant precursor ions in the MS survey scan. Both the MS and MS/MS spectra were acquired using a single microscan with a maximum fill time of 50 ms in the ion trap. m/z values selected for MS/MS were dynamically excluded for 30 s.

Sequence Database Searching and Peptide Sequence Match Filtering—The MS/MS spectra were searched database-searched using TurboSEQUEST (17) (Thermo Finnigan, San Jose, CA). The MS/MS spectra were searched against the non-redundant human International Protein Index database (18) containing ~50,000 protein sequences with a reverse version of the same database attached at the end of the forward version. The search parameters used included a precursor ion mass accuracy tolerance of 2.0 with methionine oxidation specified as a differential modification. Tryptic cleavage sites were specified as described below. The peptide sequence match results were organized and viewed using the software tool Interact (19). False positive rates were calculated as described previously (10, 20). The predicted pl of peptide sequences was calculated according to Shimura et al. (21) using an automated script, and peptide pl values were automatically inputted into the Interact results file. For FFE fractions in the pH range of 6.5–8.0 (fraction numbers 46–58), the average peptide pl value was used rather than the measured fraction pH for filtering peptide sequence matches in steps one and two (see “Results”). The MS/MS spectra were filtered against the database with the enzyme trypsin specified, allowing up to two missed cleavage sites in the peptide sequence match. To identify non-tryptic peptides derived from proline-rich proteins, as have been found in other proteomic studies of saliva (7, 8), the MS/MS data were also searched with no enzyme specified, and the peptide matches were filtered by peptide pl and FFE fraction pH. This resulted in the identification of eight additional proteins, which were added to the protein results from the first filtering step described under “Results.”

RESULTS

Our approach yielded a wealth of peptide sequence matches requiring filtering and statistical validation. To filter the sequence matches based upon peptide pl, it is first necessary to confirm the correspondence of peptide pl and measured FFE fraction pH for the dataset being analyzed (10). To this end, the sequence matches were first filtered using Peptide Prophet (22), which assigns to each peptide sequence match a probability (p) score between 0 and 1. The peptide sequence matches were initially filtered using a stringent p score threshold of 0.9. Next the theoretical pl for each matched peptide sequence was calculated (21), and the average peptide pl for each FFE fraction was determined. Fig. 1A shows the results of these calculations. The top two lines in the plot show the correspondence of the average peptide pl versus the measured pH value for each FFE fraction. Overall the close correspondence justifies the use of FFE fraction pH, in addition to p score, as a filtering criterion of peptide sequence matches for this catalogue as we describe below. There is some discrepancy between the pl and pH values in the pH range ~6.5–8.0. The reason for this discrepancy is unknown and needs further investigation, although it may reflect an inaccuracy in the pl prediction algorithm as it has been observed regardless of the method used for IEF of peptide mixtures (10, 13, 14). The bottom line in the plot shows the distribution of matched peptide sequence across each FFE fraction. The majority of the peptides cluster in the pH ranges 3.5–5.0 with very few peptides detected in fractions with neutral pH values (pH ~7–8), similar to the distribution of tryptic peptides in other studies using preparative IEF (10, 13, 14).

Our approach to generating a high confidence catalogue of proteins and their supporting peptide matches consists of two steps with each filtering step based upon the difference (ΔpH) between the calculated peptide pl value for the matched sequence and the measured pH value of the FFE fraction from which the peptide was identified. True peptide sequence matches should have pl values very close to the measured fraction pH value, whereas false matches are expected to have random pl values and be eliminated when using the ΔpH filter (10, 15). The first step initially filtered the peptide sequence matches using a ΔpH tolerance of ±0.5, which we have shown to be the optimal ΔpH tolerance based upon the IEF resolution using FFE (10). This filtering step allows for the p score threshold to be reduced while still maintaining a false positive rate below 1% (10). The optimal p score threshold using ΔpH filtering will be different for each dataset being analyzed. As Fig. 1B shows, for this particular dataset the p score could be reduced to 0.76 when applying the ΔpH filter, decreased from the p score threshold of 0.96 needed to achieve the same confidence without considering peptide pl. The second step filtered the peptide sequence matches using a low stringency p score threshold of 0.2 and peptide pl, again using a ΔpH value of ±0.5, with the added proviso that a protein would be added to the catalogue only if it was matched by two or more unique peptide sequences. This step is based upon the assumption that when combined with the peptide pl constraint, multiple peptide sequence matches provide added confidence to protein identification.
even when the matches have a low $p$ score. Indeed using these criteria the calculated false positive rate for this filtering step was also below 1%.

Each filtering step added to the catalogue. The first step identified 433 proteins from peptide matches with a $p$ score at or above the 0.76 threshold; each was added to the catalogue. 181 of these proteins had at least two peptide sequence matches, and the remainder had one peptide match. The second step identified and added to the catalogue another four proteins. At least one additional peptide sequence match was also added to 101 proteins (as indicated in the supplemental table) already in the catalogue, increasing the proteins identified by two or more peptide sequence matches to 221 of 437 total proteins. The supplemental table provides detailed information on this dataset, including all peptide sequence matches and the known biochemical functions and localizations of the identified proteins.

**DISCUSSION**

The use of peptide $p$I maximized the number of high confidence proteins identified in this study. Using $p$ score filtering alone, without the use of peptide pl information, the minimum $p$ score threshold is 0.96 to obtain a false positive rate below 1% (see Fig. 1B). Such a threshold would have resulted in the identification of only 385 proteins. The use of peptide pl and FFE fraction pH in our two filtering steps allowed for a decrease in the $p$ score threshold, thereby producing a significantly larger catalogue of high confidence proteins. These additional peptide sequence matches would otherwise be false negative matches when using $p$ score filtering alone that are sequence matches that are actually correct but do not pass the set scoring threshold (10, 15). The combined filtering steps using peptide pl and FFE fraction pH also increased the sequence coverage of identified proteins with about half of the catalogued pro-
teins having two or more peptide sequence matches.

Our approach identified 437 proteins with high confidence (false positive rate below 1%). We compared our catalogue to those from other proteomic studies of saliva attempting to comprehensively identify proteins in saliva using non-gel electrophoresis-based strategies. One recent study using multidimensional liquid chromatography and tandem mass spectrometry identified 102 proteins in whole human saliva (8). These protein matches were statistically validated using reversed database searching, providing an estimated false positive rate below 1%. Most of their catalogue’s proteins are contained in ours but not vice versa. Another recent report used both liquid chromatography-based separations and also two-dimensional gel separations to identify a combined 309 proteins from saliva (7). The overlap between their catalogue and ours was relatively small with most of the common proteins being those that have also been found in other proteomic studies, most likely indicative of their high abundance and housekeeping functions in saliva. By comparison with these other studies, our catalogue of proteins is the largest obtained from a single saliva sample to date, thereby providing new information on its composition.

Comparison of other catalogues with ours highlights an ongoing problem in the proteomics community: a lack of standards in publishing mass spectrometry-derived proteomic datasets (23, 24). For example, in the case of the study described in Ref. 7, the dataset was non-transparently presented with little information on the criteria for determining correct peptide sequence matches provided and no estimate of false positive rates or detailed information on the scoring of peptide sequence matches. Furthermore the protein sequence database used outputted protein accession numbers (e.g. Uniprot) for identified proteins. Collectively these factors make comparison of these large proteomic datasets difficult. As such, we hope that the dataset of saliva proteins we present here will serve as a model for publishing large scale proteomic data to the growing number of research groups investigating this clinically important bodily fluid, helping the dissemination and comparison of proteomic datasets obtained from future studies.

Acknowledgments—We gratefully acknowledge the Mass Spectrometry and Proteomics Center at the University of Minnesota for access to the mass spectrometer used in this work. We thank Patton Fast at the Minnesota Supercomputing Institute for help in setting up and maintaining the computer cluster used for sequence database searching.

* This work was supported in part by funding from the Minnesota Medical Foundation. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.

** To whom correspondence should be addressed: Dept. of Biochemistry, Molecular Biology, and Biophysics, University of Minnesota, 6-155 Jackson Hall, 321 Church St. S.E., Minneapolis, MN 55455. Tel.: 612-624-5249; Fax: 612-624-0432; E-mail: griffin@umn.edu.

REFERENCES

1. Hofman, L. F. (2001) Human saliva as a diagnostic specimen. J. Nutr. 131, 1621S–1625S
2. Lawrence, H. P. (2002) Salivary markers of systemic disease: noninvasive diagnosis of disease and monitoring of general health. J. Can. Dent. Assoc. 68, 170–174
3. Yao, Y., Berg, E. A., Costello, C. E., Troxler, R. F., and Oppenheim, F. G. (2003) Identification of protein components in human acquired enamel pellicle and whole saliva using novel proteomics approaches. J. Biol. Chem. 278, 5300–5308
4. Vitorino, R., Lobo, M. J., Ferrer-Correa, A. J., Dubin, J. R., Tomer, K. B., Domingues, P. M., and Amado, F. M. (2004) Identification of human whole saliva protein components using proteomics. Proteomics 4, 1109–1115
5. Shefouri, B., Tagesson, C., and Lindahl, M. (2003) Mapping of proteins in human saliva using two-dimensional gel electrophoresis and peptide mass fingerprinting. Proteomics 3, 1003–1015
6. Hardt, M., Thomas, L. R., Dixon, S. E., Newport, G., Agabian, N., Prakobphol, A., Hall, S. C., Witkowska, H. E., and Fisher, S. J. (2005) Toward defining the human parotid gland salivary proteome and peptidome: identification and characterization using 2D SDS-PAGE, ultratranslational, HPLC, and mass spectrometry. Biochemistry 44, 2885–2899
7. Hu, S., Xie, Y., Ramachandran, P., Ogorzalek Loo, R. R., Li, Y., Loo, J. A., and Wong, D. T. (2005) Large-scale identification of proteins in human salivary proteome by liquid chromatography/mass spectrometry and two-dimensional gel electrophoresis-mass spectrometry. Proteomics 5, 1714–1728
8. Wilmarsh, P. A., Riviere, M. A., Rustvold, D. L., Lauten, J. D., Madden, T. E., and David, L. L. (2004) Two-dimensional liquid chromatography study of the human whole saliva proteome. J. Proteome Res. 3, 1017–1023
9. Honore, B., Ostergaard, M., and Vorum, H. (2004) Functional genomics studies by proteomics. Bioessays 26, 901–915
10. Xie, H., Bandhakavi, S., and Griffin, T. J. (2005) Evaluating preparative isoelectric focusing of complex peptide mixtures for tandem mass spectrometry-based proteomics: a case study in profiling chromatin-enriched subcellular fractions in Saccharomyces cerevisiae. Anal. Chem. 77, 3198–3207
11. Loseva, O. I., Gavryushkin, A. V., Osipov, V. V., and Vanyakin, E. N. (1998) Application of free-flow electrophoresis for isolation and purification of proteins and peptides. Electrophoresis 19, 1127–1134
12. Moritz, R. L., Ji, H., Schultz, F., Connolly, L. M., Kapp, E. A., Speed, T. P., and Simpson, R. J. (2004) A proteome strategy for fractionating proteins and peptides using continuous free-flow electrophoresis coupled off-line to reversed-phase high-performance liquid chromatography. Anal. Chem. 76, 4811–4824
13. Cargille, B. J., Talley, D. L., and Stephenson, J. L., Jr. (2004) Immobilized pH gradients as a first dimension in shotgun proteomics and analysis of the accuracy of pl predictability of peptides. Electrophoresis 25, 936–945
14. Cargille, B. J., Bundy, J. L., Freeman, T. W., and Stephenson, J. L., Jr. (2004) Gel based isoelectric focusing of peptides and the utility of iso-electric point in protein identification. J. Proteome Res. 3, 112–119
15. Cargille, B. J., Bundy, J. L., and Stephenson, J. L., Jr. (2004) Potential for false positive identification from large databases through tandem mass spectrometry. J. Proteome Res. 3, 1082–1089
16. Rhodes, N. L., Cheng, B., Myers, S., Bowles, W., Ho, V., and Ondrey, F. (2005) A comparison of the pro-inflammatory, NF-kB-dependent cytokines: TNF-alpha, IL-1-beta, IL-6, and IL-8 in different oral fluids from oral lichen planus patients. Clin. Immunol. 114, 278–283
17. Eng, J., McCormack, A. L., and Yates, J. R., III (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. 5, 976–989
18. Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E., and Apweiler, R. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics* 4, 1985–1988

19. Han, D. K., Eng, J., Zhou, H., and Aebersold, R. (2001) Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat. Biotechnol.* 19, 946–951

20. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., and Gygi, S. P. (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* 2, 43–50

21. Shimura, K., Kamiya, K., Matsumoto, H., and Kasai, K. (2002) Fluorescence-labeled peptide pI markers for capillary isoelectric focusing. *Anal. Chem.* 74, 1046–1053

22. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* 74, 5383–5392

23. Carr, S., Aebersold, R., Baldwin, M., Burlingame, A., Clauser, K., and Nesvizhskii, A. (2004) The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. *Mol. Cell. Proteomics* 3, 531–533

24. Ravichandran, V., and Sriram, R. D. (2005) Toward data standards for proteomics. *Nat. Biotechnol.* 23, 373–376