BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 23 2012, pages 3115–3122
doi:10.1093/bioinformatics/bts594

Systems biology

Advance Access publication October 11, 2012

Network-based inference from complex proteomic mixtures using SNIPE

David P. Nusinow[1], Adam Kiezun[1], Daniel J. O'Connell[1], Joel M. Chick[2], Yingzi Yue[1], Richard L. Maas[1], Steven P. Gygi[2] and Shamil R. Sunyaev[1,*]

[1]Division of Genetics, Brigham and Women's Hospital and [2]Department of Cell Biology, Harvard Medical School, Boston, MA 02115, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Proteomics presents the opportunity to provide novel insights about the global biochemical state of a tissue. However, a significant problem with current methods is that shotgun proteomics has limited success at detecting many low abundance proteins, such as transcription factors from complex mixtures of cells and tissues. The ability to assay for these proteins in the context of the entire proteome would be useful in many areas of experimental biology.

Results: We used network-based inference in an approach named SNIPE (Software for Network Inference of Proteomics Experiments) that selectively highlights proteins that are more likely to be active but are otherwise undetectable in a shotgun proteomic sample. SNIPE integrates spectral counts from paired case–control samples over a network neighbourhood and assesses the statistical likelihood of enrichment by a permutation test. As an initial application, SNIPE was able to select several proteins required for early murine tooth development. Multiple lines of additional experimental evidence confirm that SNIPE can uncover previously unreported transcription factors in this system. We conclude that SNIPE can enhance the utility of shotgun proteomics data to facilitate the study of poorly detected proteins in complex mixtures.

Availability and Implementation: An implementation for the R statistical computing environment named snipeR has been made freely available at http://genetics.bwh.harvard.edu/snipe/.

Contact: ssunyaev@rics.bwh.harvard.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 12, 2012; revised on September 18, 2012; accepted on September 27, 2012

1 INTRODUCTION

The proteins expressed in a tissue are critical to determining its identity and proper function. Characterizing these proteins is the goal of proteomics; however, current mass spectrometer technology is largely incapable of detecting many of the low-abundance proteins in complex mixtures, such as mammalian whole tissue lysates (Bantscheff *et al*., 2007; Gerber *et al*., 2003; Malmstrm *et al*., 2007). The common approaches to deal with this problem include enrichment assays or selective on-line monitoring of specific ions or reactions. The primary issue with these methods is that by design, they only capture a fraction of the proteome. As a result, comprehensive proteome analysis remains a difficult task (de Godoy *et al*., 2008; Nagaraj *et al*., 2011), and an as-yet unsolved problem for multi-cellular eukaryotes. Because of this, the biochemical state of a tissue must often be inferred from gene expression data and a few select trusted antibodies, leaving the vast majority of the proteome invisible and essentially unapproachable.

The ability to assay the complete proteome would be desirable for fields such as developmental biology. However, in addition to technical challenges discussed previously, developmental biologists often study highly complex tissues available in limited amounts. Also, many of the key proteins of interest, including transcription factors and signalling molecules, are present only at low abundance and are not well detected by whole proteome analysis. Gene expression microarrays or RNA-Seq are used as a stand-in for an effective whole-proteome assay, but it has been repeatedly established that quantitative proteomic measurements correlate poorly to gene expression levels (Greenbaum *et al*., 2003; Gygi *et al*., 1999; Schrimpf *et al*., 2009; Vogel and Marcotte, 2012), rendering this approach problematic for assaying the proteome. Thus, although a clear need exists for effective proteomics approaches, current technology does not meet the needs of researchers when tissue quantity or machine time is constrained.

A common goal in developmental biology is to define the mechanisms that determine the states of cells and tissues in time and space. These mechanisms are carried out by the coordinated tissue-specific action of a large number of proteins. The activity of these proteins can be differentially regulated in several ways. First, the proteins themselves can be expressed at different levels. Alternately, although being expressed at the same level, they can change their localization or function because of modifications. These proteins do not act individually but are members of genetic pathways, in which the regulation of one member will have an impact on other pathway members. Thus, the expression and function of other pathway members is indicative of the activity of a given protein. For example, the role of a transcription factor in specific tissue or developmental time point can be evident from the differential regulation of its transcriptional targets. Similarly, the role of a secreted signalling molecule can be inferred from the changes in protein levels or localization of other canonical pathway members.

*To whom correspondence should be addressed.

We hypothesized that the coordinated expression of associated proteins can be used to infer the identity of proteins involved in development. This approach can be applied to proteomic data, where the protein of interest cannot be readily identified or quantified. To test this hypothesis, we have developed a wholly computational method named SNIPE (Software for Network Inference of Proteomics Experiments). SNIPE integrates previous biological knowledge encoded in a pre-existing network with the simple spectral counts obtained from any mass spectrometry experiment to select proteins whose network neighbourhoods are enriched between samples. Proteins that are involved in development, but otherwise undetected in the sample, would be expected to have enriched surrounding subnetwork and may be thus detectable by SNIPE.

As part of an effort to use systems biology approaches for mammalian organ engineering, we focused on a mouse model for the development of SNIPE. The developing mouse tooth is a premier model for organogenesis through epithelial–mesenchymal interactions, which comprise a fundamental principle for the development of numerous mammalian organs. These interactions are characterized in all tissues, including the tooth, by the exchange of signalling molecules leading to the induction of transcription factors and other signalling molecules that are responsible for the development of the organ (Bei, 2009). As these low-abundancy proteins form the core of the developmental mechanism driving the formation of many organs, it is important to be able to assay them. Applying SNIPE to the study of the developing tooth would, therefore, be an optimal test-case for the method, as its performance could be judged by its ability to assay these critical, but challenging to detect, proteins. Here, we report the SNIPE algorithm and software implementation and its ability to correctly infer the presence of proteins known to be expressed in and functionally important in tooth development, as well as its ability to infer the presence of transcription factors previously undescribed in the developing tooth.

2 ALGORITHM

2.1 The SNIPE algorithm

The key feature of SNIPE is its ability to assay the entire proteome despite the challenges of current mass spectrometry technology to detect many proteins in complex samples. SNIPE uses simple spectral counts from mass spectrometry data paired with a known existing network based on available knowledge to increase statistical power and to make inferences (Fig. 1). For each protein in the network, SNIPE sums the number of spectral counts corresponding to that protein and all of that protein's immediate neighbours in the network. This is done for two sample sets that act as a case–control design (such as tooth germ and non-tooth oral tissue), and then these sums are normalized and compared by a scoring function (see Algorithm). To calculate a *P*-value for differential enrichment, the individual spectral counts of the two samples are randomly permuted at each node without disturbing the network architecture, effectively simulating the null hypothesis. Each node is given a score under each permutation, and the *P*-value is assigned given the position of the observed score in the distribution of simulated scores. To control for multiple testing, the best score in the entire

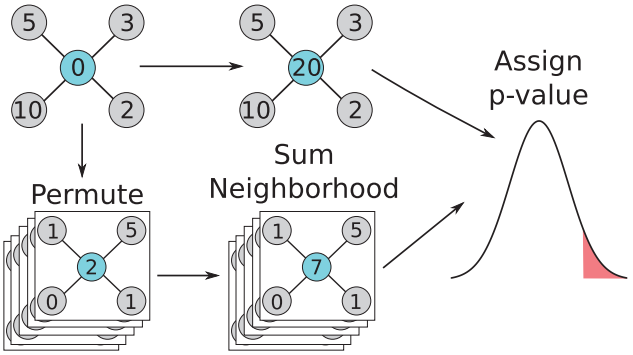

Fig. 1. Diagram of the SNIPE method. Spectral counts for each protein are matched to their nodes in a given network. A score for the protein is calculated by summing the node (blue) and its immediate neighbours (grey). A *P*-value is assigned by permuting the counts in the network and generating a distribution of scores for that node and comparing the observed score to that distribution

network for each permutation is stored, and the position of each observed score is compared with this distribution of extreme value scores, providing a *P*-value that is corrected while explicitly taking into account both network architecture and simulated null hypothesis data rather than fully theoretical distributions. The case–control set-up eliminates several potential problems with this application, including differential peptide ionization efficiencies and protein lengths, as these are assumed to be constant between the case and control samples. It also removes the bias because of network architecture, as this feature remains constant between samples and during permutations.

The fundamental feature of SNIPE is that it is not limited to assaying only the proteins that are observed in the dataset. As in Figure 1, SNIPE can also act on 'empty nodes' that have observed spectral counts of 0. These empty nodes may have neighbours that are observed at high levels in a dataset, which then serve to inform the presence of the empty node protein. This guilt-by-association idea has been used effectively in many other areas (Deo *et al.*, 2010; di Bernardo *et al.*, 2005; Lee *et al.*, 2010; Nibbe *et al.*, 2010; Vanunu *et al.*, 2010; Wolfe *et al.*, 2005), but to the best of our knowledge, this is the first time it has been applied in this manner. A somewhat related method named as clique-enrichment approach was recently used to attempt to rescue low-confidence proteins identified in a search and uncover Gene Ontology (GO) categories known to be involved in a given phenotype (Li *et al.*, 2009). Clique-enrichment approach attempts to find cliques in the network architecture, whereas SNIPE relies on network neighbourhood and uses semi-quantitative data (in the form of spectral counts) in a case–control experimental design. As a result, the main benefit of SNIPE is its design as an enrichment test between the case and control sample to bias towards functionally relevant proteins, as described previously.

SNIPE relies on having a network of protein–protein associations available. The current SNIPE implementation uses the STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) network for this purpose. For this work, the STRING version 8.2 protein links file was downloaded from http://string.embl.de/. Links between proteins in STRING are determined

using a variety of sources, including protein–protein interaction, high-throughput gene expression, protein co-evolution and literature text mining datasets (Jensen *et al.*, 2009). As a result, STRING gives a broad, but inexact, view of whether two proteins associate in any way. Although SNIPE currently uses STRING, any similar network could potentially be used.

SNIPE traverses each node in the STRING database, tallying the total spectral counts for each node and its immediate neighbours in the network for each of two samples given. A score $\chi = (x1 - E)/\sqrt{E}$ is calculated, where $E = ((x1 + x2) \times (x1 + y1))/(x1 + x2 + y1 + y2)$ and $x1$ and $x2$ are the sums of spectral counts for the neighbourhood of that protein in each sample, whereas $y1$ and $y2$ are the sums of all observed spectral counts in each sample. Thus, χ is closely related to the χ^2 statistic by design. In contrast to χ^2, χ allows for one-sided tests. A single χ score is calculated for the comparison of the two groups, as in the χ^2 test.

A *P*-value is calculated by permuting the spectral counts in the network to generate a distribution of χ scores under the null hypothesis. For this work, we used 1 million permutations. At each node, the sum of the counts in the neighbourhood for each sample is randomly distributed across the two sample types according to the binomial distribution using probability of success 0.5 because under the null hypothesis, we expect that the number of spectral counts at each node will be equal in the two sample types. Because the permutation uses the counts within the neighbourhood and thus the counts stay with that neighbourhood, this controls for different counts per node. Additionally, because of the case–control design and the design of the χ score, differing numbers of counts between the samples are dealt with naturally. The network structure is considered fixed and not permuted.

Multiple test correction is performed by storing the highest score for each permutation to create a distribution of best scores. A corrected *P*-value is calculated by comparing the observed score with this distribution.

Ideally, network neighbourhood enrichment detected by SNIPE implies the following: first, the protein is likely to be present. Second, the protein is either differentially expressed itself or expressed in equal amounts but is functionally involved in the coordinated enrichment of the network neighbourhood because of changes such as differential modification or localization. A negative SNIPE result, however, does not rule out protein presence, differential expression or a functional change. From a functional perspective, proteins highlighted by SNIPE would either cause the coordinated change of the network neighbourhood or would be downstream targets of an active pathway. Both of these possibilities are of interest to developmental biologists. In the real world, the analysis can be complicated by statistical noise, tissue- or time point-specific nature of protein associations in the network, as well as gross network inaccuracies. Despite these issues, as seen from the results presented later, SNIPE is able to correctly highlight proteins, such as transcription factors, that are necessary for tooth development.

3 IMPLEMENTATION

SNIPE is currently implemented in software as a package named snipeR for the R Statistical Computing Environment (R Development Core Team, 2011). Because of its size, STRING is not bundled with the package, but may be downloaded separately using a helper function included in the package or manually by the user from the STRING website.

4 RESULTS

4.1 The application of SNIPE to mammalian proteome data

We generated matched proteomic datasets for the developing mouse lower molars and surrounding non-dental oral tissue at embryonic day 13.5 (E13.5) (Table 1). The proteins identified in these samples showed significant overlap (Supplementary Fig. S1A), as was expected at this developmental stage. Although several thousand proteins in these datasets were identified by mass spectrometry, only a handful were found to be significantly enriched in the developing tooth compared with oral tissue using a statistical test of simple spectral counts [Bonferroni corrected *P*-value or false discovery rate (FDR) <0.05, Table 2]. This result was expected, given the limitations already described (Bantscheff *et al.*, 2007). Furthermore, most of the proteins that were known to play a role in tooth development were not detected in these datasets (Supplementary Fig. S1B), and none of them were found to be significantly enriched in the tooth germ, even in the many cases where they were known to be. This result was also expected, as these proteins are almost exclusively signalling pathway components and transcription factors and were thus likely to be found only at relatively low levels, if at all. GO category analysis found no significant enrichment of developmental pathways (Supplementary Table S1). These results are reflective of the inherent difficulties in using current proteomics technology to detect biologically important proteins in a complex sample.

Table 1. Peptide and protein recovery from tooth and non-tooth control sample

Dataset	Total peptides	Unique proteins
E13.5 tooth germ	27 810	2961
E13.5 oral tissue	26 889	2822

'Total peptides' is the total number of identified non-unique peptides discovered in each sample for two biological replicates. 'Unique proteins' is the total number of unique proteins identified by at least one spectrum in the same samples.

Table 2. Proteomics results with and without SNIPE

Multiple test correction	Fisher's test	SNIPE
Uncorrected (nominal $P<0.05$)	152	2588
Bonferroni corrected ($P<0.05$)	13	443
FDR <0.05, permutation-based	23	1534/514

FDR, Benjamini-Hochberg. Permutation-based multiple test correction is the SNIPE default (see Algorithm).

In contrast, SNIPE analysis of these datasets found a significant enrichment of the network neighbourhood for >10-fold more proteins out of the entire proteome (514 proteins with *P*-value of <0.05 after permutation-based multiple test correction and 1534 with FDR <0.05, Table 2). A greater fraction of the proteins originally detected only in the tooth germ were selected by SNIPE (4%) compared with those only in oral tissue or both samples (2% and 3%, respectively). This is consistent with expectations that SNIPE will highlight proteins uniquely expressed in the tissue of interest. GO category analysis revealed categories including transcription factor activity and embryonic development to be statistically overrepresented in this set (8.5×10^{-7} and 6.4×10^{-6} FDR, respectively, Supplementary Table S2). None of these categories were represented in the analysis of the raw proteomics data. This result indicates that our GO category enrichment through SNIPE was non-random, and that SNIPE was able to select truly relevant proteins where the raw proteomics results were unable to do so. To demonstrate this, we chose to focus on the 45 transcription factors in this set because of the importance of transcription factors to development and the difficulty in detecting them from complex protein mixtures. Of these factors, only two were detected in our original samples. Nine of the transcription factors were found to be expressed in the tooth during development by a literature search, and five of those were also known to have some functional role in tooth development. Importantly, among these five proteins were Pax9 and Msx1, which are both known to be essential for tooth formation (Bei, 2009) and which were not detected directly in our proteomic datasets. Thus, SNIPE highlights proteins previously characterized in the developing tooth including some known to be critical for organogenesis.

4.2 Bioinformatic validation of SNIPE

To estimate the rate of false-positive predictions made by SNIPE, we used the Helsinki database of tooth development (Kaski *et al.*, 1996), which contains annotations for many genes expressed at E13.5, including some genes that were marked as absent and hence could be considered as true-negative proteins. A particular challenge for this analysis is that most of the genes that are not expressed at E13.5 are expressed at some other stage of tooth development, and several of them are known to cause tooth defects when mutated. We expected the proteins coded by these genes to be associated with other tooth-related proteins in STRING, making it difficult for SNIPE to discern that they should not be present. However, contrary to this expectation, SNIPE was able to select far more true- than false-positive proteins (Fig. 2A). In contrast, increasing the FDR cut-off for the SEQUEST protein identifications in the raw signal did not increase this signal at all, indicating that these proteins were not identified at all in the original sample even at low confidence, but were inferred by SNIPE. Interestingly, the raw proteomics data detected a number of proteins that were annotated as being definitively absent from this particular developmental stage. This implies that our true negative dataset is flawed and that a base number of incorrectly annotated false-positive proteins should be assumed when using this database. Despite this, the Helsinki Tooth Database is an expert hand-curated database, and much of its contents

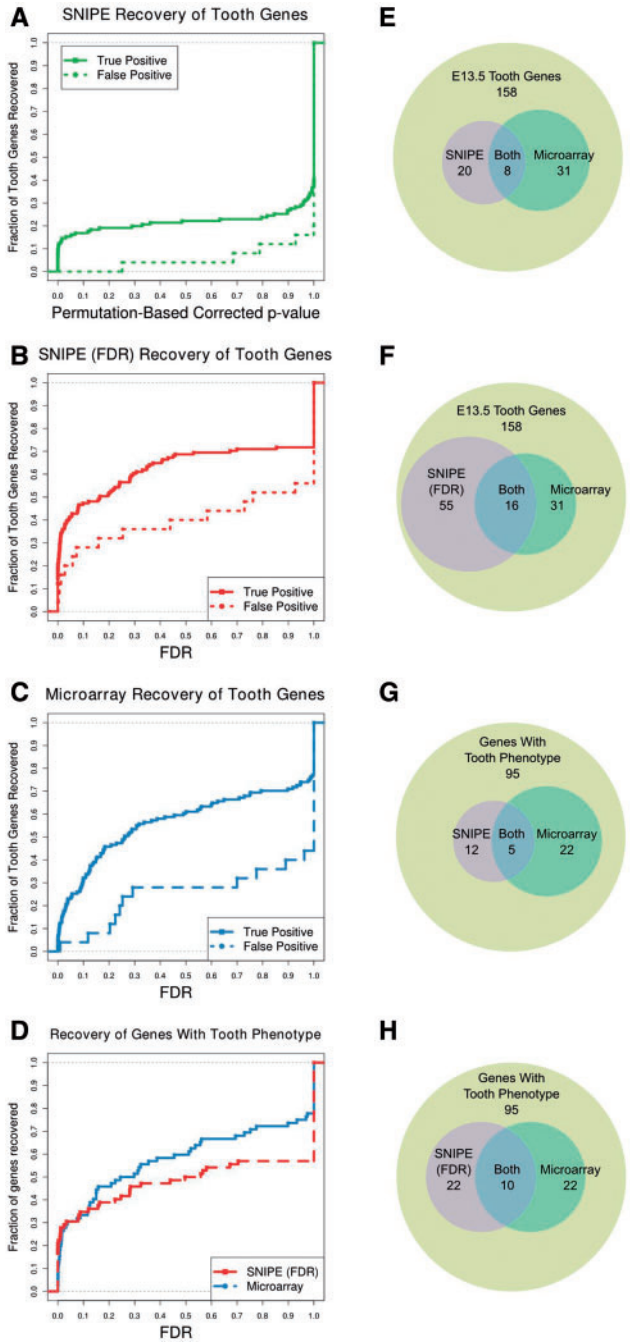

Fig. 2. SNIPE performance. (**A–D**) Fractions of true- and false-positive proteins recovered at various significance thresholds for (A) SNIPE using permutation-based multiple test correction, (B) SNIPE using FDR for multiple test correction and (C) microarray. (D) Fraction of genes in the MGI database annotated for causing a tooth phenotype recovered at different FDR for SNIPE and Microarray. (**E–G**) Numbers of proteins in the Helsinki database recovered by (E) SNIPE using permutation-based multiple test correction, (F) SNIPE using FDR compared with microarray. (**G** and **H**) Numbers of proteins known to cause a phenotype in the MGI database recovered by (G) SNIPE using permutation-based multiple test correction and (H) SNIPE using FDR compared with microarray

have been independently confirmed by authors of this article and others.

We next used the Mouse Genome Informatics (MGI) database (Bult *et al.*, 2008) to determine which of the proteins that SNIPE identified had some documented genetic effect on tooth development when their coding gene is mutated. This is critically important as a test of SNIPE's ability to highlight proteins that are not only present but are also functionally relevant. SNIPE is designed as an enrichment test to bias towards biologically relevant proteins. Thus, we would expect that SNIPE is able to predict the presence of these proteins in the tooth germ sample. These proteins were overwhelmingly selected by SNIPE (Fig. 2D).

4.3 Technical interpretation of the SNIPE output

An incompletely solved problem in shotgun proteomics is the process of assigning protein identities to peptides that are not unique in the proteome. Because SNIPE deals entirely with protein identifications, the assignment of these peptides could have significant effects on the output of the algorithm. To test this, we took the non-unique peptides identified in the tooth germ and oral tissue samples and determined the set of proteins they matched across the entire proteome. We then generated 10 sets of random protein identifications from the peptide matches in our samples, and subsequently ran SNIPE on each random set. We found that the random assignment of protein identities did not have a major impact on SNIPE's performance (Supplementary Fig. S2A and B, compare with Fig. 2A and B). We conclude that SNIPE is able to overcome problems with non-unique peptide assignment, without significant effects on its output.

Because of the large increase in fold-change caused by SNIPE, proteins that show little evidence of enrichment in the original sample will be overwhelmed by the signal from their network neighbours, leading to loss from the list of proteins under consideration. Although this is expected, as SNIPE is fundamentally an enrichment test, it will cause proteins that are identified by mass spectrometry to be given poor scores by SNIPE. The distribution of this effect for the pooled tooth germ samples is shown in Supplementary Figure S2C and D. The histogram in Supplementary Figure S2C shows the size of the effect of summing up the STRING network neighbourhood by SNIPE when compared with the originally observed spectral counts by mass spectrometry. This graphic does not include unobserved proteins. The two peaks show a divergent effect. The first peak, to the left, represents a ratio of 1, indicating that SNIPE provided no increased power to make any inferences about these proteins. The second peak, in the middle of the vaguely Gaussian distribution, shows that when SNIPE is able to bring the network to bear on the protein, the effect sizes tend to fall between 50- to 100-fold. Supplementary Figure S2D shows the distribution of the calculated χ score (see Algorithm) for all proteins in STRING, graphed against the observed spectral counts in the original mass spectrometry data. Proteins with low observed spectral counts have a generally equal probability of being assigned a low or high score. As the number of observed counts for a protein increases, however, the SNIPE score also begins to increase, as shown by the trend line, which begins to increase noticeably at a score of ~25. Thus, in cases where there is a good number of spectral counts for the protein already observed, SNIPE will tend to agree with the observed data.

4.4 Effects of the STRING network

SNIPE treats the underlying STRING network as fixed to capture the underlying information encoded in its architecture. The case–control design of the system also controls for a certain amount of bias in the network architecture. However, it is possible that because of the large number of nodes and connections in the network, a significant number of proteins were found simply by chance, despite attempts to correct for multiple hypothesis tests. To examine this, we randomly permuted the network using a standard permutation matrix, thereby maintaining the distribution of node degrees throughout the entire network. Under these conditions, few proteins in the Helsinki Tooth Database set passed the threshold of statistical significance, using either the permutation-based multiple test correction or FDR (Supplementary Fig. S3A and B). Additionally, there was little overlap between the proteins chosen by the normal and permuted networks. Accepting an equal number of proteins from both the normal and permuted network experiments, the number of those reaching statistical significance of corrected $P<0.05$ in the normal experiment and the same top number from the permuted experiment, found an overlap of ~5%, no better than statistical noise (Supplementary Fig. S3C). From this work, we can conclude that the specific architecture of the network is critical to the function of SNIPE, and that SNIPE's performance is not because of general features of STRING, such as distribution of node degree.

A distinctive feature of STRING is its use of multiple different data sources to build the final network. An interesting question was which of these data sources were contributing to the signal. Running SNIPE using subsets of STRING composed of only specific data sources (Supplementary Fig. S3E and F) revealed that the signal was almost exclusively coming from the 'text mining' and 'database' categories. Gene co-expression also provided a small amount of signal.

To further investigate the contributions to the observed signal, we looked at the proteins in the network that were contributing to the positive predictions for four transcription factors that are necessary for tooth development, Pax9, Msx1, Dlx2 and Barx1 (Supplementary Fig. S4). Although all of these proteins are critical for tooth development, none were directly observed in our proteomic data. In all four cases, the majority of the contributing spectral counts came from proteins not previously described in tooth development at all (Supplementary Fig. S4). Further, the known genes were unable to provide a significant *P*-value for enrichment by Fisher's exact test, whereas the unknown genes were. As in the global analysis, the signal for all four of these proteins came almost exclusively from the 'text mining' and 'database' STRING categories. Among the proteins contributing the signal from these four network neighbourhoods, only one (Msx1) could be considered a clique (Supplementary Fig. S5), demonstrating that clique finding approaches would be fundamentally unable to recover many of these necessary proteins from these data. Critically, these results demonstrate that SNIPE is not simply rediscovering previous findings from the tooth biology literature, but is instead using latent knowledge to make novel inferences about tooth development.

A feature of STRING is that each network edge includes a composite score of the confidence of the validity of that edge. Running SNIPE using versions of STRING filtered at specific cut-off values resulted in the true-positive recovery rate falling faster than the false-positive recovery rate as stringency is increased (Supplementary Fig. S3D). Once scores <0.5 are filtered out, the majority of the gain in the signal is lost. This indicates that the signal found in STRING is not primarily in the network edges with high scores, but in the composite of the lower-scoring edges.

4.5 Comparison of SNIPE to gene expression microarray analysis

The most common genome-scale method for evaluating transcription factors is differential gene expression analysis. To compare SNIPE with this standard, we generated microarray datasets in the same manner as the proteomic samples and looked at the set of genes enriched in the developing tooth (Fig. 2C–H). We compared these against the set of genes in the Helsinki tooth database. For microarray analysis, we chose to look for significant enrichment in the tooth germ sample compared with the non-tooth oral tissue rather than a simple detection above background level. The rationale for this is, as described previously, looking for signals enriched in the tissue of interest biases to search towards functionally relevant proteins. The microarray analysis predicts 677 proteins as significantly enriched in the tooth germ tissue, in comparison with SNIPE's 514. Using its normal multiple test correction, SNIPE selected far fewer proteins annotated as present in the tooth in the Helsinki database than microarray analysis (Fig. 2A and C). Suprisingly, the sets of proteins recovered by each analysis were relatively distinct (Fig. 2E). Correcting SNIPE's output using a standard FDR correction as in the microarray analysis (Fig. 2B and F) creates an alternative comparison of interest. In this case, SNIPE selected many more true-positive proteins, at the cost of many more false-positive proteins (Fig. 2B). A similar pattern held for the transcription factors alone as well as for the MGI mutant database set (Fig. 2D), indicating that developmentally important proteins are being highlighted. Again, SNIPE was able to select a distinct set of these proteins compared with microarray analysis (Fig. 2G and H). These results indicate that SNIPE can highlight relevant proteins and that these proteins are distinct and complementary to those found through microarray analysis.

4.6 Experimental validation of SNIPE predictions

A challenge with these findings is that it was possible that SNIPE was only selecting proteins already known to be involved in tooth development, whereas the remainder were false-positive proteins. Therefore, it was critical to determine whether SNIPE can correctly select novel proteins whose expression during organ development could be confirmed by an independent experimental method. Examining the set of transcription factors given *P*-values of <0.05 by SNIPE for which antibodies were available revealed a subset of these that had never been investigated or described in the tooth at all, and thus, presented an opportunity for new discovery through SNIPE. We tested 22 of these proteins by immunohistochemistry and found that 7 showed staining that was specifically localized to the developing tooth bud. These seven were chosen as candidates for further analysis. The proteins we did not select at this stage might have had antibodies that were not appropriate for immunohistochemistry or represent SNIPE false-positive proteins. Additionally, proteins with non–tooth-specific expression were eliminated at this stage. Of the seven, five showed nuclear localization, as typically observed for transcription factors (Fig. 3). Microarray data also showed RNA expression for all five of the nuclear-localized proteins save Steroidogenic factor 1 (SF1). These results were confirmed by PCR, with faint expression for SF1 mRNA detected by two primer pairs, potentially explaining the absent call in the microarray data (Supplementary Fig. S6). The two proteins that showed tissue specific non-nuclear staining were not confirmed by PCR, indicating that they were false-positive proteins. To the best of our knowledge, none of these proteins had been previously shown to be expressed in the developing tooth [Six1 was subsequently shown to be expressed transcriptionally (Nonomura *et al*., 2010)] nor have they been described functionally as of yet in this system. These results provide evidence that SNIPE can correctly infer the presence of novel proteins expressed in a developing organ.

5 DISCUSSION

SNIPE correctly highlights several proteins, including transcription factors, that are not normally detected by shotgun mass spectrometry of complex protein samples from whole tissues. Assaying these proteins has been a critical requirement for fields like developmental biology, and SNIPE broadens the applicability of proteomics for these areas.

SNIPE relies on the assumptions that associated proteins tend to be co-expressed and that these associations are not always tissue or time point specific. Current understanding of molecular pathway function is that coordinated activity among pathway members generally requires their co-expression, and that many of these pathways are used reiteratively across tissues and time points. The combinations of pathway activities serve to provide tissue and time point specificity from non-specific components. Recent analysis of protein–protein interactions in the human proteome provides further evidence that this is indeed the case globally, rather than just for well-studied pathways (Bossi and Lehner, 2009). SNIPE's explicit design around network neighbourhoods along with a semi-quantitative case–control experimental set-up allowed SNIPE to succeed where approaches relying only on network architecture would fail on this dataset (Supplementary Fig. S5). Despite this, the SNIPE model is not insensitive to the network inaccuracies, as we demonstrate in our analysis. Thus, we recommend that experimenters tailor their use of the *P*-values reported by SNIPE to their experimental needs, allowing more or less stringent cut-offs depending on their tolerance for error and willingness to subsequently validate the SNIPE inferences. Using SNIPE's permutation-based multiple test correction effectively suppresses false-positive proteins (Fig. 2A) and should be used when the experimenter has low tolerance for errors, whereas using a more traditional FDR calculation increases the error rate but also captures far more true-positive proteins (Fig. 2B).

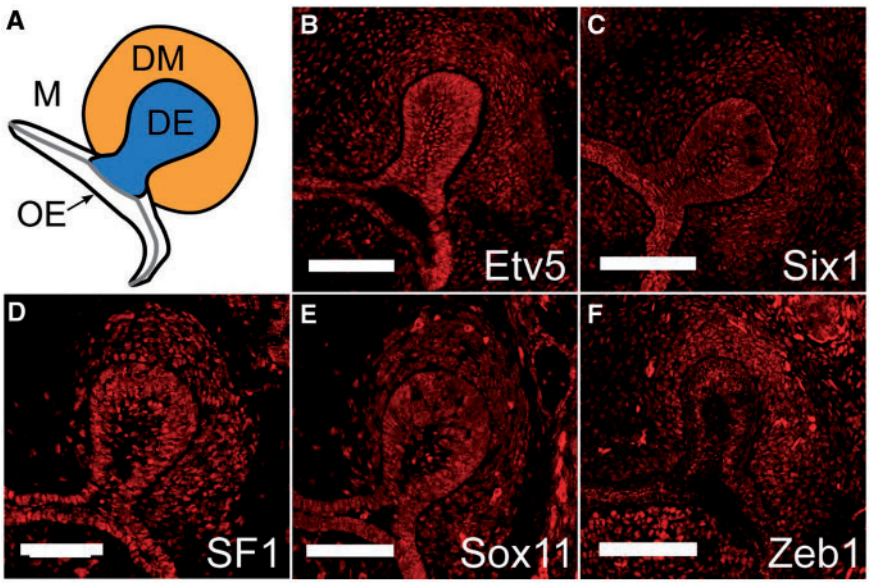

Fig. 3. Immunofluorescence of E13.5 tooth bud. (**A**) Diagram of the E13.5 murine first molar tooth bud. The tooth at this stage is composed of invaginating dental epithelium (DE, blue) and surrounding condensed dental mesenchyme (DM, orange). The condensed cells indicate differential cellular fate of the condensed dental mesenchyme from the surrounding non-dental mesenchyme (M). The non-dental oral epithelium (OE) is a cellular bilayer divided by the oral cavity (grey line). (**B**) Etv5 immunostain shows uniform ubiquitous nuclear expression in the tooth and non-tooth regions. (**C**) Six1 immunostain shows similarly ubiquitous nuclear expression in the tooth and non-tooth regions with a slight visible enrichment in the dental mesenchyme matching the region previously reported for its gene expression pattern (Nonomura *et al.*, 2010). (**D**) SF1 (Nr5a1) and (**E**) Sox11 immunostains show specific expression in the dental mesenchyme and all epithelial tissue. (**F**) Zeb1 immunostain shows a nuclear localized stain specifically in the mesenchymal tissue with slight upregulation in the dental mesenchyme. Scale bars: 88μm

A limitation of SNIPE is that it relies on integers rather than continuous variables. This dramatically simplifies the statistics, making simple permutations reasonable for generating *P*-values. However, there is significant general interest in using other non-integer measures of protein abundance besides spectral counts, such as spectral peak intensity. SNIPE cannot currently handle this form of data, as well as many types of non-proteomic data, such as microarray results. Given the broad move in the field towards quantitative mass spectrometry using methods such as isobaric tagging that generate continuous data, SNIPE will need to be adapted to be more applicable to these types of experiments. However, these quantitative methods do not solve the problem of dealing with low abundance proteins, and they can indeed exacerbate the problem by relying on methods such as triple-stage mass spectrometry (MS3), which decreases overall throughput (Ting *et al.*, 2011), thereby limiting the number of spectra assayed. Improved mass spectrometers will likely be the most important advance to address this problem, but it is unclear at what point they will be sufficient to detect and quantify a nearly complete proteome from restricted amounts of tissue. Even the most current mass spectrometers are challenged by the task of uncovering low abundance proteins in favourable conditions, such as large quantities of protein and high-quality enrichment procedures (Kim *et al.*, 2011; Nagaraj *et al.*, 2011). Because of these reasons, potential applications for SNIPE will exist for the foreseeable future, and will be enhanced if it is made to work with continuous data. This will likely be the subject of subsequent work on the method.

An important finding that was repeated throughout this work is that SNIPE not only highlights undetected proteins that are present in a given tissue sample, but that it focuses attention on proteins that are important for the developmental process. Although the number of proteins identified by the raw proteomic data was much higher than that selected by SNIPE, SNIPE was able to select dramatically more factors that were critical for the process of interest, including transcription factors that were known to be necessary by mutational analysis. This was reflected in the results from two different databases (Helsinki and MGI) as well as GO category analysis, demonstrating that this is a non-random effect. Our finding that the signal for known transcription factors comes from associations with proteins that were not studied in the tooth development field indicates that SNIPE is effectively able to bring evidence from other fields to bear on making tissue-specific predictions. Because of SNIPE's reliance on non–tissue-specific previous knowledge, it is unclear how powerful it will be in making novel discoveries across different tissues and experimental systems. However, our experimental validation of the expression of several predicted transcription factors never before described in the tooth indicates that SNIPE can be used to make novel discoveries (Fig. 3). Unfortunately, our experimental validation does not extend to functional studies of the found proteins, making it currently impossible to say that SNIPE is capable of biasing towards biologically relevant novel proteins. However, our finding that it can recover previously known functionally important proteins (Fig. 2D) and that the signal from these proteins comes from proteins not known to be related to tooth development (Supplementary Fig. S4) provides evidence that it will do so as designed. Additionally, our comparison with differential gene expression analysis suggests that gene expression analysis and SNIPE are complementary approaches, detecting

distinct, although partially overlapping, sets of genes/proteins (Fig. 2E–H). This indicates that researchers interested in a more complete global picture stand to gain by using both methods in concert.

ACKNOWLEDGEMENTS

The authors thank the members of the Sunyaev, Maas, and Gygi laboratories for thoughtful discussion. They also thank the SysCODE consortium, most notably Dr Peter Park, Dr Joshua Ho and Dr James Costello, for helpful advice. They also thank Dr Andrej Shevchenko and Dr Soumya Raychaudhuri for feedback on the method and its manuscript during preparation.

Funding: This work was supported by the National Institutes of Health Common Fund [5RL1DE019021, 5RL1DE019022]. Microarray expression experiments were performed by the Microarray Core Facility of the Molecular Genetics Core Facility at Childrens Hospital Boston supported by NIH-P50-NS40828 and NIH-P30-HD18655.

Conflict of Interest: none declared.

REFERENCES

Bantscheff,M. *et al.* (2007) Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.*, **389**, 1017–1031.

Bei,M. (2009) Molecular genetics of tooth development. *Curr. Opin. Genet. Dev.*, **19**, 504–510.

Bossi,A. and Lehner,B. (2009) Tissue specificity and the human protein interaction network. *Mol. Syst. Biol.*, **5**, 260.

Bult,C.J. *et al.* (2008) The mouse genome database (MGD): mouse biology and model systems. *Nucleic Acids Res.*, **36** (**Suppl 1**), D724–D728.

Csárdi,G. and Nepusz,T. (2006) The igraph software package for complex network research. *Inter J. Complex Syst.*, **1695**.

de Godoy,L.M.F. *et al.* (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*, **455**, 1251–1254.

Deo,R.C. *et al.* (2010) Interpreting metabolomic profiles using unbiased pathway models. *PLoS Comput. Biol.*, **6**, e1000692.

di Bernardo,D. *et al.* (2005) Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotech.*, **23**, 377–383.

Du,P. *et al.* (2008) lumi: a pipeline for processing illumine microarray. *Bioinformatics*, **24**, 1547–1548.

Gerber,S.A. *et al.* (2003) Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl Acad. Sci. USA*, **100**, 6940–6945.

Greenbaum,D. *et al.* (2003) Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol.*, **4**, 117.

Gygi,S.P. *et al.* (1999) Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.*, **19**, 1720–1730.

Jensen,L.J. *et al.* (2009) STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37** (**Suppl 1**), D412–D416.

Kaski,M. *et al.* (1996) *Gene Expression in Tooth*. Developmental Biology Programme of the University of Helsinki, http://bite-it.helsinki.fi (15 October 2012, date last accessed).

Kim,W. *et al.* (2011) Systematic and quantitative assessment of the Ubiquitin-Modified proteome. *Mol. Cell*, **44**, 325–340.

Lee,I. *et al.* (2010) Rational association of genes with traits using a genome-scale gene network for arabidopsis thaliana. *Nat. Biotech.*, **28**, 149–156.

Li,J. *et al.* (2009) Network-assisted protein identification and data interpretation in shotgun proteomics. *Mol. Syst. Biol.*, **5**, 303.

Malmstrm,J. *et al.* (2007) Advances in proteomic workflows for systems biology. *Curr. Opin. Biotechnol.*, **18**, 378–384. PMID: 17698335.

Nagaraj,N. *et al.* (2011) Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.*, **7**, 548.

Nibbe,R.K. *et al.* (2010) An integrative-omics approach to identify functional sub-networks in human colorectal cancer. *PLoS Comput. Biol.*, **6**, e1000639.

Nonomura,K. *et al.* (2010) Dynamic expression of six family genes in the dental mesenchyme and the epithelial ameloblast stem/progenitor cells during murine tooth development. *J. Anat.*, **216**, 80–91.

R Development Core Team (2011) R: a language and environment for statistical computing, Vienna, Austria.

Schrimpf,S.P. *et al.* (2009) Comparative functional analysis of the caenorhabditis elegans and drosophila melanogaster proteomes. *PLoS Biol.*, **7**, e1000048.

Ting,L. *et al.* (2011) MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat. Methods*, **8**, 937–940.

Vanunu,O. *et al.* (2010) Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.*, **6**, e1000641.

Vogel,C. and Marcotte,E.M. (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.*, **13**, 227.

Wolfe,C. *et al.* (2005) Systematic survey reveals general applicability of 'guilt-by-association' within gene coexpression networks. *BMC Bioinformatics*, **6**, 227.