research articles Journal of proteome research

Whole Genome Searching with Shotgun Proteomic Data: Applications for Genome Annotation

Joel R. Sevinsky,[†] Benjamin J. Cargile,[†] Maureen K. Bunger,[†] Fanyu Meng,[‡] Nathan A. Yates,[‡] Ronald C. Hendrickson,[‡] and James L. Stephenson, Jr.*[,†]

Mass Spectrometry Program, Research Triangle Institute, 3040 Cornwallis Road, Research Triangle Park, North Carolina 27709-2194, and Molecular Profiling Proteomics, Merck Research Laboratories, Merck and Company Inc., Rahway, New Jersey 08854

Received April 06, 2007

High-throughput genome sequencing continues to accelerate the rate at which complete genomes are available for biological research. Many of these new genome sequences have little or no genome annotation currently available and hence rely upon computational predictions of protein coding genes. Evidence of translation from proteomic techniques could facilitate experimental validation of protein coding genes, but the techniques for whole genome searching with MS/MS data have not been adequately developed to date. Here we describe GENQUEST, a novel method using peptide isoelectric focusing and accurate mass to greatly reduce the peptide search space, making fast, accurate, and sensitive whole human genome searching possible on common desktop computers. In an initial experiment, almost all exonic peptides identified in a protein database search were identified when searching genomic sequence. Many peptides identified exclusively in the genome searches were incorrectly identified or could not be experimentally validated, highlighting the importance of orthogonal validation. Experimentally validated peptides exclusive to the genomic searches can be used to reannotate protein coding genes. GENQUEST represents an experimental tool that can be used by the proteomics community at large for validating computational approaches to genome annotation.

Keywords: isoelectric focusing • mass spectrometry • peptide identification • genome annotation • database searching

Introduction

Genome wide sequencing efforts have produced a large and rich collection of complete genomes for which there is often a lack of corresponding protein data. Although computational predictions of protein coding genes have greatly improved, these techniques still struggle to balance quantity versus quality for protein coding gene predictions.[1] Newly sequenced genomes can create problems for gene prediction algorithms due to their unusual GC content, lack of training sets, few orthologous homologues from closely related species, and preference for mammalian structures, to name just a few.[1] This results in the use of multiple algorithms and hence multiple gene lists, a challenging situation for comparative analysis and functional annotation. Although the integration of proteomic data sets into genome annotation efforts would allow high-throughput experimental validation of gene predictions and annotations, several challenges within proteomics remain before this can be realized. For example, proteomic data sets are inherently unbiased and do not depend on a predefined set of proteins. Shotgun proteomics requires database search routines that match MS/MS spectra to amino acid sequences for high-throughput identification of peptide species.[2–4] In principle, these same search routines could be used to search amino acid sequences derived from six-frame translated genomes,[5] allowing experimental validation of protein coding genes. However, these techniques face great challenges due to the enormous database sizes of six-frame translated genomes, requiring extremely long search times, as well as reductions in the signal-to-noise ratio (S/N, or true positive/false positives for all candidate peptides of a given MS/MS spectrum) resulting in greatly decreased sensitivity.[6,7] Recent work using probability modeling has advanced these techniques,[8] but further improvements are still needed.

Recent innovations in instrumentation and peptide separation strategies have improved the fidelity and dimensionality of shotgun proteomics data sets. High-resolution mass spectrometers allow the m/z of peptide ions to be determined with parts per million (ppm) accuracy,[9] increasing confidence in database search results. Furthermore, novel methods that utilize immobilized pH gradient isoelectric focusing (IPG-IEF) of peptides can determine the peptide isoelectric point (pI) to within 0.03 pI units.[10] In addition to the analytical benefits of peptide IPG-IEF, accurate peptide pI can be used to filter database search results for increased confidence while simul-

* Correspondence should be addressed to Dr. James L. Stephenson, Jr., Senior Program director for Mass Spectrometry Research, Research Triangle Institute, Durham, NC 27709, USA. Phone: (919) 316-3978. Fax: (919) 541-6161. E-mail: HTUstephensonjl@rti.orgUTH.

[†] Research Triangle Institute.

[‡] Merck Research Laboratories.

10.1021/pr070198n CCC: $40.75

Published on Web 12/07/2007

taneously reducing false negative and false positive identification rates.[11–13]

Orthogonal peptide pI filtering is not limited to results though. Accurate peptide pI can be applied prior to database searching to constrain the database size and complexity. Here we demonstrate GENQUEST (GENome Queries Using Experimental SpecTra), a novel method using accurate peptide pI to constrain the database size of a six-frame translated human genome, resulting in accurate and sensitive genome searching comparable to searching the human protein databases. This can be achieved with minimal computing power along with currently available instrumentation.

Methods

Cell Culture and Sample Preparation. DU4475 cells were grown in RPMI (Invitrogen) supplemented with 10% fetal bovine serum (HyClone) at 37 °C at a density of 5×10^5 cells/mL. Cells were extracted in 8 M urea and 25 mM Tris pH 7.5, with 1 mM sodium orthovanadate, 1X phosphatase inhibitor cocktail (Sigma), and 1X Complete protease inhibitor cocktail (Roche). After centrifugation to remove insoluble cellular debris, soluble proteins were diluted to 1 M urea in 25 mM Tris pH 7.5 and digested with 20 μg of trypsin (Promega) overnight at 37 °C. Tryptic peptides were desalted using SepPak C_{18} Light cartridges (Waters) and dried in a speedvac.

Peptide IPG-IEF. Peptides were focused as previously described[13] with some modifications. Briefly, a 24 cm pH 3.5–4.5 IPG strip (GE Healthcare) was rehydrated overnight with 1 mg of peptides resuspended in 8 M urea, 0.5% ampholytes. The strip was focused for 112000 Vhrs using an IPGPhor II (GE Healthcare) according to the manufacturer protocol. The strip was cut into 60 fractions of 4 mm in width. Each fraction was sequentially extracted with 200 μL of 0.1% TFA, 200 μL of 0.1% TFA/50% ACN, and 200 μL of 0.1% TFA/100% ACN. Extracted peptides were dried, resuspended in 0.1% TFA, and then cleaned and desalted using Oasis HLB SPE (Waters). Peptides were finally resuspended in 40 μL of 0.1% TFA.

LC-FTMS Analysis of Complex Peptide Mixtures. An HP1100 capillary pump (Agilent) was used in this study. Solvent A was 0.1 M acetic acid in H_2O, and solvent B was 0.1 M acetic acid in 90% aceteonitrile/10% H_2O. The gradient used was 0–3.0 min, 100% A; 3.0–39.1 min, 100% A–70% A; 39.1–59.0 min, 70% A–10% A; 59.0–59.1 min, 10% A–100% A; and 59.1–75.5 min, 100% A. The flow rate was 1 μL/min for the peptidic region. A Famos autoinjector (Dionex Corp.) was used to load the samples onto a trap column (100 μm ID, 2.5 cm, New Objective) packed with ProteoPepII C_{18} media, and the injection volume was 1 μL per sample. The peptides were eluted out through a spraying column (100 μm ID, packed in house with POROS R2 media). A hybrid linear ion trap–FTMS (Thermo Electron) was used to acquire MS data. One full FTMS scan followed by three data-dependent ion trap MS/MS scans were continuously acquired. The key settings for FTMS were: AGC = 1×10^6; maximum injection time 1.0 s and resolution = 50000. A nanospray source with no sheath gas was used, and the spraying voltage was 3.0 kV. For MS/MS: 2 m/z isolation window, 35% normalized collision energy. The settings for dynamic exclusion were: repeat count, 2; repeat durations, 30 s; exclusion list, 50; and exclusion durations, 180 s.

In silico Processing of the Human Genome. The human genome sequence was downloaded from NCBI (Build 35, version 1). Python scripts were generated to perform six-frame translations of each contig for each chromosome by performing in silico trypsin digestion for each translated region, along with an in silico trypsin digestion of its reverse sequence, calculating the peptide MW and isoelectric point (pI) of each tryptic peptide, and assigning to the database all fully tryptic peptides (forward and reverse) between MW 800 and 3000. This database is known as the human genome peptide database (HGPdb). Narrow pI range peptide fasta files for SEQUEST searching were created one of two ways. The first was to query the HGPdb for all peptides within that narrow pI range, fasta format, and index with BioWorks Browser (Thermo Electron). The second way, and the faster way on most desktop systems, is to sort all peptides into fasta files representing 0.01 pI units for each file. When a narrow pI range is determined, the files covering this range can be quickly concatenated and then indexed with BioWorks Browser. All processing was done on a desktop computer with a Pentium 4 2.8 GHz processor and 3 GB RAM.

Database Searching and Data Analysis. The .RAW file for each fraction was searched against the EID protein database[14] with an appended reverse database[15,16] using the SEQUEST search algorithm.[3] The EID database, a fasta-formatted collection of sequences and annotations for all exons and introns obtained from human GenBank sequences, enabled us to easily distinguish peptides that were contained entirely within exon sequences ("exonic") from peptides whose sequence spanned an intron ("intron spanning"). For each fraction, all top-ranked peptide hits in each .out file, both forward and reverse, were collected and separated by charge state, and the pI for each peptide sequence was calculated using a novel pI prediction algorithm.[10] The peptide pI range for each charge state was calculated by first identifying all peptides with XCorrs greater than the highest reverse database hit and using quartile filtering of pI values to remove all outliers. Then the average pI value $\pm$ 3 standard deviations for these peptides was determined. All peptides were examined again, both forward and reverse, and peptides that had pI values outside of this range were excluded. True forward hits were those peptides that had XCorr values greater than the highest reverse database hit within this pI range and charge state. This high reverse database hit XCorr is referred to as the "XCorr cutoff" and produces false positive rates equal to 2 times the number of passing reverse identifications divided by the total number of identifications. Since the XCorr cutoff is the XCorr of the highest reverse identification, the false positive rate is less than 2 divided by the total number of identifications, or 1% to 2% for most fractions/charge states with 100–200 peptides identified. The pI range for the entire fraction was determined by using the highest and lowest pI values found for the three charge states within the fraction. In instances where there were too few peptides in the +1 charge state to accurately determine the pI range, only the +2 and +3 charge states were considered. The fraction pI range was used to query the HGPdb and create a fasta file of genome peptides to be indexed and searched again using the same .RAW file. These results were reverse database and peptide pI filtered as with the protein database searches, except a new pI range was not calculated for each charge state; rather the same pI range for each charge state as determined by the protein search was used. This method produced a false positive rate of 3.5% for the genome searches considered as a whole. To accurately and quickly compare the results from the two different database searches, when a particular fraction and charge state was examined, the highest XCorr cutoff between the two searches was used, and all peptides that were below this cutoff were not considered in the analysis. The peptides

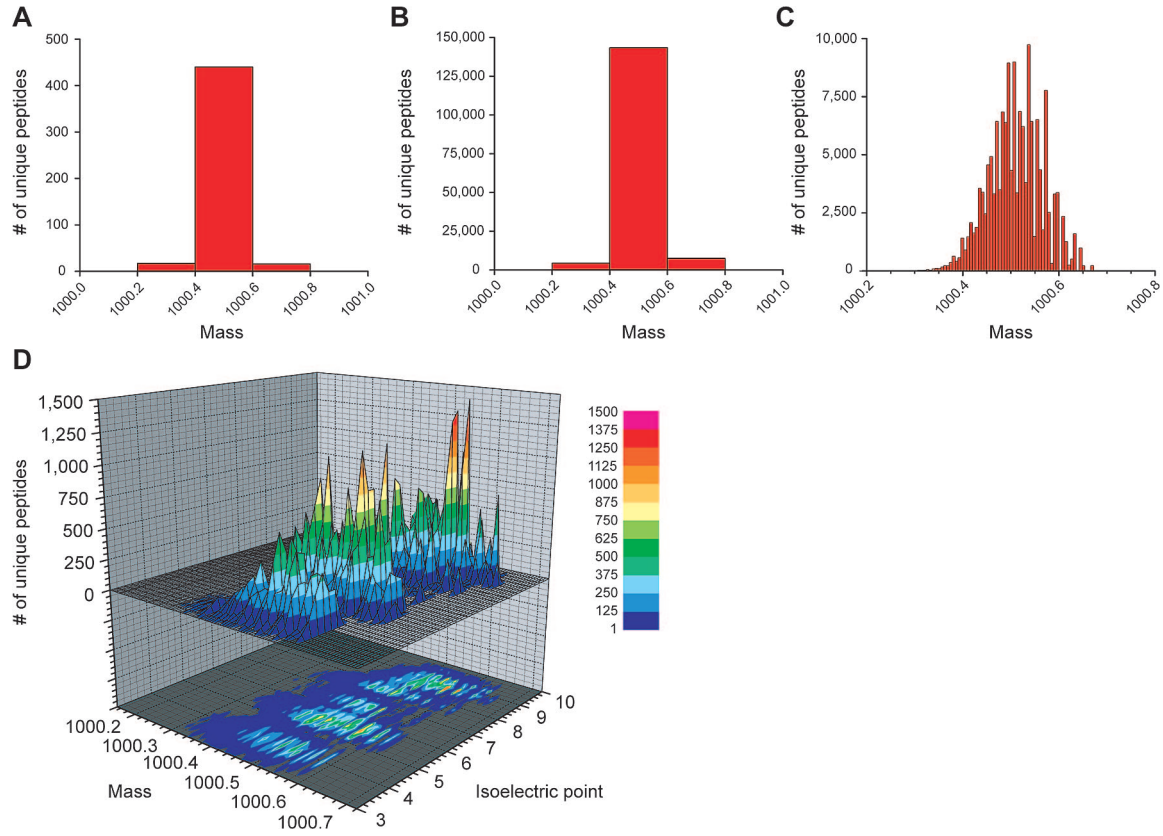

Figure 1. Theoretical considerations for GENQUEST. In silico trypsin digestion, pI calculations, and mass calculations between MW 1000 and 1001 were performed for both the EID protein database (A) and a six-frame translated human genome (HGPdb) (B–D). The number of peptides in each mass and pI bin was calculated for the EID protein database at 100 ppm mass accuracy (A), the HGPdb at 100 ppm mass accuracy (B), the HGPdb at 3 ppm mass accuracy (C), and the HGPdb at 3 ppm mass accuracy, and isoelectric point focusing in 0.2 pI unit bins (D).

above this "shared XCorr cutoff" are known as "shared XCorr cutoff peptides" and are the focus of this analysis. All processing was done on a desktop computer with a Pentium 4 2.8 GHz processor and 3 GB RAM.

BLAST Searching of Genome Peptides. The 540 unique genome peptides were searched against the human genome protein database (hs_genome/protein) from NCBI using BLAST. Peptides that had significant matches of less than 100% identity over their full length and peptides that had no significant matches were selected for manual examination. Peptides that had significant matches of less than their full length were not manually examined.

Peptide Synthesis and LC-MS/MS Analysis. Peptides were synthesized by Sigma using PepScreen technology. Peptides were analyzed by MS/MS using direct infusion in 50:50 ACN: H2O + 1% acetic acid on a linear ion-trap mass spectrometer using fragmentation parameters identical to those used in the LTQ-FT analysis. Peptides with cysteines were alkylated with iodoacetimide for 30 min prior to analysis.

Results

Theoretical Considerations for Searching the Human Genome with MS/MS Data. Most shotgun proteomics workflows search experimentally acquired MS/MS spectra against databases of known proteins using sophisticated computer algorithms. Although this technique has been very successful, it is not amenable for searching MS/MS spectra against large genomic sequences because of the decrease in S/N observed due to the large increase in database size, resulting in greatly reduced sensitivity when trying to maintain low false positive rates. For example, if MS/MS spectra were acquired on an ion trap instrument capable of 100 ppm mass accuracy and a particular tryptic peptide of mass 1000.5 Da was to be searched against the Exon–Intron Database[14,17] (EID, a secondary database of all human proteins from GenBank along with splicing annotations), a search space of 473 unique theoretical tryptic peptide candidates would have to be considered by the search algorithm (Figure 1A). The search space would increase to 155222 unique fully tryptic peptides if this same spectrum was searched against a six-frame translated raw genomic database (Figure 1B), greatly decreasing the S/N. Accurate mass at 3 ppm will decrease the search space to fewer than 10000 peptides (Figure 1C), but the database size will still remain the same and be difficult to manage (standard human genomic peptide databases are larger than 5 GB, while a human protein database is typically 25–50 MB). If peptide IPG-IEF is utilized for first dimension peptide separation, the pI range of each particular fraction can be used to prefilter the search space to fewer than 1500 peptides (90% of peptides found in search spaces fewer than 1000 peptides, 72% found in search spaces fewer than 500 peptides) as shown in Figure 1D. Furthermore, the database size will decrease significantly because the database need only consist of those peptides in a very narrow pI range. Thus, accurate pI simultaneously reduces the number of candidates and the size of the database, resulting in greatly increased S/N

comparable to that of the initial protein database searching as well as significantly shorter database search times.

Experimental Design of Raw Genomic Sequence Searching with MS/MS Spectra. Previous work by our laboratory and others established the significance of using peptide pI as an orthogonal discriminator for validating results from SEQUEST for shotgun proteomics data.[13,18] Briefly, when peptides are fractionated by isoelectric point using IPG strips, each fraction will contain peptides of a narrow pI range. This information can be used to filter the MS/MS search results, excluding all peptides that fall outside of this narrow pI range. As an extension to this work, once a peptide pI range is known for a particular peptide fraction, this information can then be used to restrict the peptide search space when searching much larger databases, such as a six-frame translated human genome. If a peptide is not within this narrow pI range, then there is no reason to compare the MS/MS spectra against it. This is the principle behind the GENQUEST method. For the experiments discussed here, we used the workflow presented in Figure 2.

GENQUEST Provides Sensitive Peptide Identification. To test the effectiveness of this technique, DU4475 cell extracts were digested with trypsin; the peptides were focused on a 24 cm pH 3.5–4.5 IPG strip; the strip was cut into 60 fractions; and 40 of these fractions were analyzed on a hybrid linear ion trap Fourier transform ion cyclotron resonance mass spectrometer (LTQ-FT). The RAW files were searched using the GENQUEST method using a precursor ion mass accuracy of 20 ppm due to mass drift on the LTQ-FT for this particular analysis. As shown in Table 1, this technique produces peptide fractions with very narrow pI ranges. The average pI range for each fraction was 0.25 ± 0.08 pI units. Furthermore, the search times were very short. SEQUEST searches against the EID and IPI protein databases averaged 18.2 ± 1.8 and 15.2 ± 1.5 spectra per second, respectively, while searching the genome peptide database averaged 8.3 ± 4.5 spectra per second. This is less than a 3-fold increase in search time for a greater than 100-fold increase in search space using modest computing resources.

The sensitivity of GENQUEST was comparable to that of protein database searches. In Figure 3A and in Supplemental Table 1 is a summary of the peptides identified from the EID database and the HGPdb, indicating the number of shared and unique peptides. There were 2953 EID peptides identified and 2749 HGPdb peptides identified, with 2209 identified in both searches (Table 2). Of these, the vast majority are peptides completely contained within exons. There were 744 peptides unique to the EID searches and 540 unique to the HGPdb searches. There were also peptides that did not meet the shared XCorr cutoff criteria (see Methods) and were not considered in the analysis described here. These can be considered false negatives for their respective searches, 438 peptides for the EID search and 105 for the HGPdb search.

The 744 proteome specific peptides that were not found in the HGPdb search should predominately consist of intron spanning peptides. A total of 595 of these peptides were intron spanning, but there were 149 exonic peptides (peptides for which the nucleotide coding sequence is contained completely within an exon) that were not identified, and these were investigated further. The original HGPdb construction did not include N- and C-terminal peptides, and hence 39 of these peptides were not identified. Another seven peptides were not identified because of I/L redundancy, where two peptides are identical except for their isoleucine and leucine residues (these peptides are indistinguishable in the experiments performed here), and hence will score identically but not necessarily rank reproducibly between searches against different databases (Figure 3B). Four peptides had better matches in the genome for the respective MS/MS spectra used to identify them. The largest class of exon peptides not identified were those where the splice between the two exons fell immediately 3′ to any nucleotide of a K or R codon. This results in a peptide for which, although the sequence is completely within an exon, the peptide sequence will be represented differently in the genome (Figure 3C, D). Hence, for the purposes of determining whether a tryptic peptide is exonic or intron spanning, the K/R immediately N-terminal to the peptide should be considered part of the peptide since it defines the boundaries of the peptide in both protein and genomic sequences.

Examination of MS/MS Spectra with Unique Matches within the Human Genome. The 540 genome specific peptides were considered to be potential new open reading frames and were examined further. BLAST[19] was used to search all 540 peptides against the NCBI human genome protein database (hs_genome\protein) to find peptides from human proteins that might not be represented in the GenBank files used to create the EID protein database. Forty peptides were found to have matches of 100% identity over their full length. Similar searches in the human IPI protein database (version 3.03) compared to the genome searches found 21 protein peptides that were not identified (data not shown). Of the remaining 500 peptides, 377 had significant matches less than 100% over the full length of the peptide, and 123 had no significant matches. A subset of 166 peptides were examined manually (see Supplemental Table 2), and 97 were rejected because they did not meet the criteria for peptide MS/MS spectra (lack of contiguous b/y ion series and/or high S/N ions). This number is artificially high because the 166 peptide identifications selected for manual analysis contained all 123 peptides that had no detectable protein database similarities and hence were more likely than peptides with protein database similarities to be false positives. Four identifications were discarded because the peptide sequence was found in a fully tryptic IPI peptide database, and nine were found to have protein identifications that were not identified due to I/L redundancy. Two peptide sequences were potentially due to a +14 methylation of D to E, and one peptide was from a repetitive element, leaving 53 peptides. Of these 53 peptides, 24 were high-quality spectra but did not match the peptide identified. The remaining 29 peptides were high-quality spectra and matched the peptide identified.

Identification of a Novel Protein Coding Region. These 29 peptides were synthesized and analyzed on an LTQ mass spectrometer, and 12 out of the 29 produced MS/MS spectra matching those determined previously (see Figure 4A). Upon more thorough examination of these 12 peptides, six were found in nonspecific cleavage searches of the human IPI database; two peptides were found to be ambiguous with known protein peptides due to amino acid rearrangements (rearrangements did not significantly affect pI, and the specific MS/MS ions for these fragments were not present); and one peptide had multiple genome locations. RT-PCR experiments were carried out to identify transcripts from genomic regions coding for the final three peptides, and mRNA was successfully amplified and sequenced for the peptide 'EFVQAYEDVLER' within the MACF1 gene (data not shown). There were 12 other peptides identified for the MACF1 protein (see Supplemental

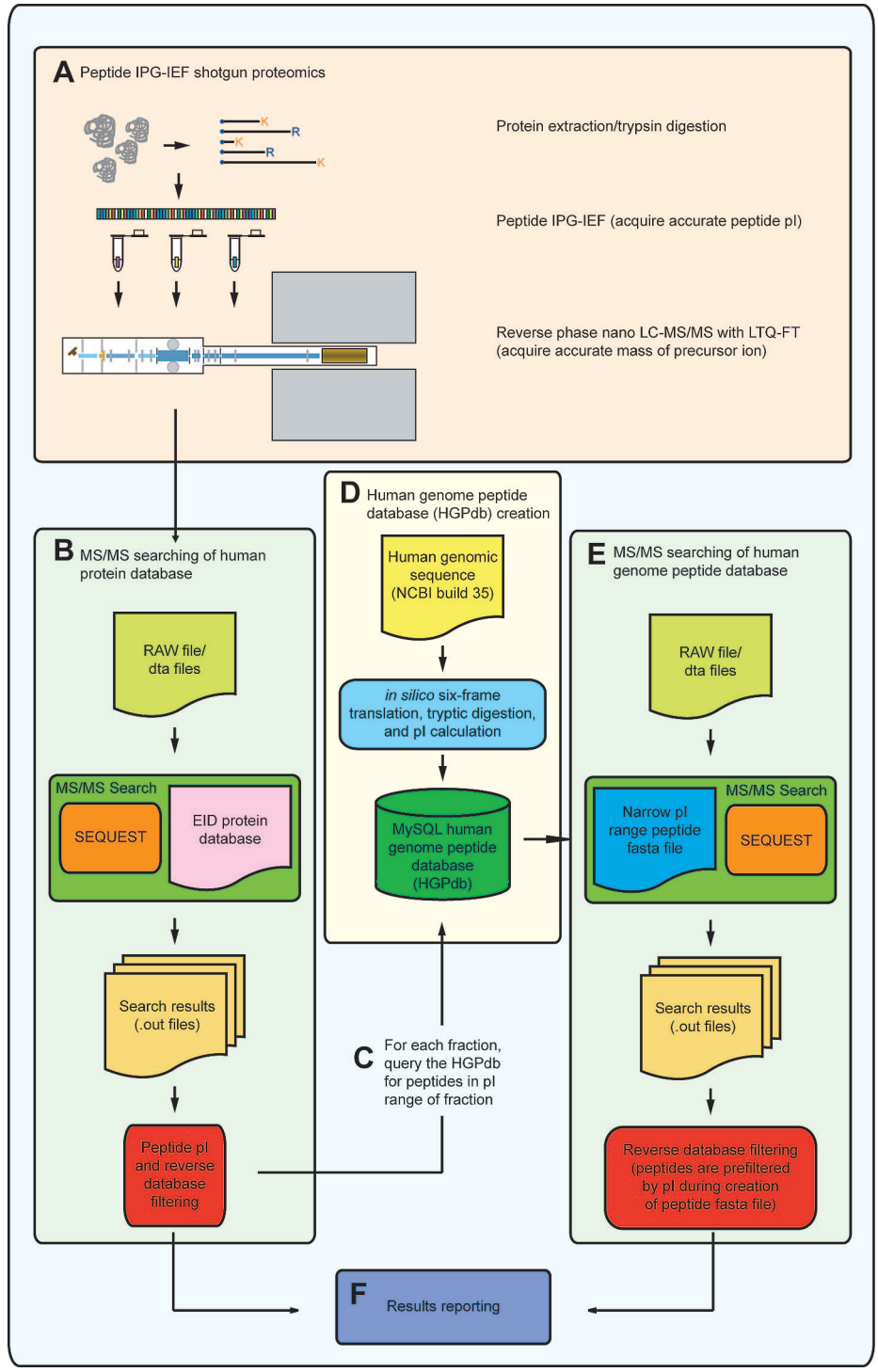

Figure 2. GENQUEST workflow. (A) Protein samples are digested with trypsin and undergo peptide IPG-IEF fractionation and reverse-phase nano-LC-MS/MS using a high mass accuracy mass spectrometer. (B) After data collection, RAW files for each fraction are searched against the EID protein database using SEQUEST, and the results for each fraction undergo pI and reverse database filtering. (C) The narrow pI range for each fraction is used to query the HGPdb for all peptides in this pI range. (D) The HGPdb was created from a six-frame translation of the NCBI human genome. Each open reading frame was fully digested in silico with trypsin, and all peptides between MW 800 and 3000 were indexed in a MySQL relational database. (E) The query results are fasta formatted, indexed with BioWorks, and searched using the original RAW file from (B). These results undergo reverse database filtering only because the peptide database has been prefiltered by pI before searching. (F) The EID and HGPdb results for each fraction are compared identifying peptides common and exclusive to each search.

Table 1, EID identifier 545A_NT_004511), 11 exonic and 1 intron in the EID search and 11 exonic in the HGPdb search, but this novel peptide is located within an intron where there is evidence for protein coding within the EST databases and Genscan predictions (see screenshot of UCSC genome browser[20] in Figure 4B). We believe that this peptide is part of a previously unannotated exon identified using the GENQUEST technique.

Table 1. pI Range Calculations and SEQUEST Search Times

	pI values			SEQUEST search (spectra/second)		
fraction	low	high	range	NCBI	IPI	genome
1	3.26	3.86	0.61	13.7	12.3	20.5
2	3.31	3.66	0.36	16.7	16.7	19.5
3	3.41	3.66	0.26	18.9	16.2	16.2
4	3.44	3.69	0.26	21.4	15.3	13.3
5	3.46	3.69	0.24	18.0	18.0	15.4
6	3.53	3.66	0.14	16.1	16.1	18.8
7	3.43	3.79	0.37	19.1	19.1	9.6
8	3.52	3.69	0.18	18.6	16.0	16.0
9	3.43	3.86	0.44	16.9	16.9	7.2
10	3.54	3.72	0.19	17.5	17.5	13.1
12	3.56	3.78	0.23	18.5	15.9	10.1
13	3.60	3.81	0.22	18.5	15.9	8.5
14	3.57	3.83	0.27	17.2	15.1	7.5
15	3.62	3.85	0.24	19.2	16.4	8.2
16	3.63	3.85	0.23	20.1	15.1	7.5
17	3.61	3.90	0.30	17.8	13.8	6.9
18	3.65	3.88	0.24	17.0	14.9	6.6
19	3.66	3.92	0.27	22.6	18.8	7.1
20	3.70	3.91	0.22	18.2	14.2	6.7
21	3.69	3.93	0.25	18.1	14.1	6.0
22	3.69	3.98	0.30	15.5	13.8	5.4
23	3.76	3.93	0.18	18.1	14.1	6.7
24	3.74	3.97	0.24	14.2	14.2	5.8
25	3.76	4.01	0.26	17.4	13.5	5.8
26	3.75	4.01	0.27	17.3	15.2	5.5
27	3.82	4.00	0.19	17.2	13.4	5.7
28	3.81	4.05	0.25	18.0	14.0	5.2
29	3.87	4.04	0.18	16.3	14.5	5.9
30	3.85	4.08	0.24	18.7	14.5	5.5
31	3.87	4.10	0.24	18.2	14.2	5.5
32	3.91	4.13	0.23	19.7	14.8	5.4
33	3.95	4.12	0.18	18.2	14.2	5.5
34	3.94	4.17	0.24	22.1	17.7	6.3
35	3.97	4.14	0.18	17.9	15.6	5.0
36	4.00	4.14	0.15	19.6	14.7	4.9
37	4.01	4.18	0.18	18.3	13.7	5.0
38	4.01	4.22	0.22	18.5	15.9	5.1
40	4.01	4.31	0.31	21.9	13.7	4.4
41	4.08	4.28	0.21	18.5	15.8	5.0
42	4.12	4.29	0.18	16.8	13.1	4.7
	average		0.25 ± 0.08	18.2 ± 1.8	15.2 ± 1.5	8.3 ± 4.5

Discussion

The motivation of this work was to overcome the current limitations of searching whole genomes with shotgun MS/MS data, thus allowing the experimental validation of protein coding genes to eventually complement genome annotation efforts. Previous attempts to search large and complex genomic sequences with shotgun MS/MS data have not been nearly as sensitive or accurate as searching the available protein databases due to the exponential increase in database size. Here we have shown that GENQUEST is a very accurate technique with high sensitivity and low false positives as demonstrated by the high correlation between results obtained when searching an EID database versus searching the HGPdb database. From a limited sample, over 2200 peptides from known, annotated human genes were identified from searching the human genomic sequence. This data would be for annotating the human genome if minimal annotation was available.

There are several advantages for using the GENQUEST technique for genome searching. First, due to the decrease in database size and the use of accurate mass, the HGPdb search times are short. Searches against the HGPdb take less than three times as long as searches against the EID or other protein databases, enabling users to perform the searches faster than the data is acquired on the instrument, all with modest computing resources. Second, the technique has accuracy and sensitivity comparable to protein database searching. When a shared XCorr cutoff is used for both searches, almost all peptides (99.8%) from the protein database searches that are present in the HGPdb are identified. The shared XCorr cutoff did produce additional false negatives for each search (438 for the EID search and 105 for the HGPdb search), and the larger number of false negatives for the EID search was most likely due to the increase in the search mass window used for this study (20 ppm rather than 3 ppm). Last and most importantly, it requires no expensive specialized hardware or computing power. All equipment used for these studies, with the exception of the LTQ-FT, is readily available in most proteomic facilities and relatively inexpensive. The increasing proliferation of accurate mass analyzers amenable to high-throughput proteomics will likely facilitate access to these instruments and

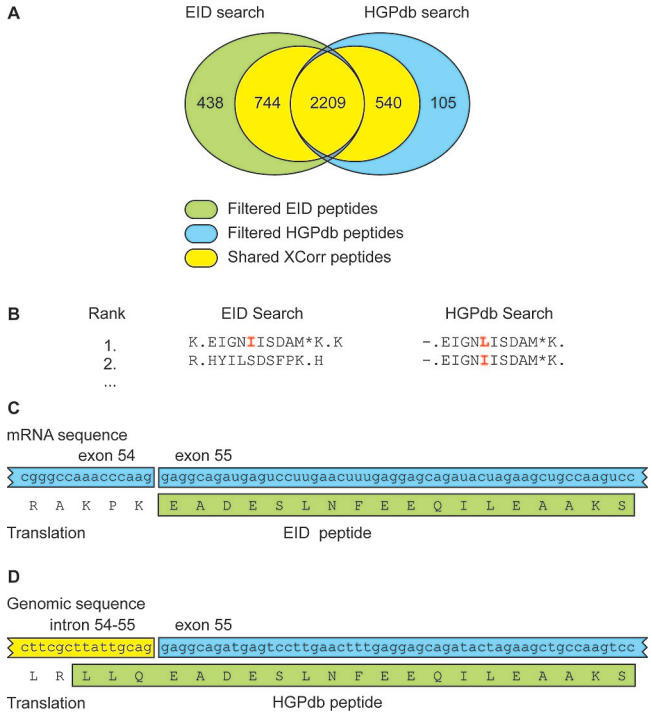

Figure 3. Peptides identified from both EID and HGPdb searches and unidentified exonic peptides. (A) Filtered EID peptides are all peptides identified from searching the EID database after pI and reverse database filtering. Filtered HGPdb are all peptides identified from searching the HGPdb after pI and reverse database filtering. Shared XCorr peptides are filtered peptides from both the EID and HGPdb search that score higher than the highest XCorr cutoff between the two searches for each fraction and charge state (see Methods). (B) Searching the EID database identifies the peptide EIGNIISDAMK from heat shock protein HSPD1. This same MS/MS spectrum when searched against the HGPdb identifies EIGN**L**ISDAMK as the first ranked identification and EIGN**I**ISDAMK as the second. (C) mRNA sequence and exonic peptide from TLN1. (D) Genomic sequence from TLN1 demonstrating an N-terminal extension of three amino acids when the tryptic peptide is translated from genomic sequence.

provide most laboratories the opportunity to perform these techniques. Additional development of this technique will be used to facilitate genome annotation projects for newly sequenced genomes that rely on computational predictions and comparative genomics. Since a first round of searching against a protein database is necessary to determine the pI range of the fraction, the first round of searching for novel genomes would be performed against a protein database of conservative computational predictions to establish an empirically determined pI range for each fraction. The second round of searching would be performed against the genomic sequence. An alternative method would be to rely on the pH gradient profile provided by the manufacturer to determine the pI range of the fraction. Our pI prediction algorithm closely profiles this gradient[10] and increasing the pI window to 0.3–0.4 pI units around this value would be an acceptable substitute for empirically determining the range. This whole genome searching would improve gene predictions as demonstrated by the identification of a previously unidentified exon in the MACF1 gene.

The HGPdb searches were able to identify almost all of the exonic peptides from the EID but were unable to identify intron spanning peptides which account for approximately 25% of all

Table 2. Summary of Peptide Identifications

peptides identified in both databases	**2209**
exonic	2092
exonic and intron spanning	17
intron spanning	100
proteome only peptides	**744**
intron spanning	595
exonic	149
splice on K/R	93
N/C-terminus	39
I/L redundancy	7
better match in genome	4
other	6
genome only peptides	**540**
BLAST hit found	40
no BLAST hit	500
manually checked	166
not realistic upon manual inspection	97
IPI peptide	4
I/L redundancy	9
potential methylation (D to E)	2
repetitive element	1
good spectra/does not match peptide identified	24
good spectra/matches peptide identified	29

tryptic protein peptides as noted in our work and others.[21] This was expected since the major goal of our work was to develop a method that allowed sensitive searching of raw genomic sequence for which there was minimal supporting experimental data. This is the case with many newly sequenced novel genomes, especially microbial genomes where annotations are sparse and nearly all genes are exonic. The results of using GENQUEST could then be used to guide computational predictions of gene structure and location. On the other hand, if thorough genomic annotation is desired in a system where extensive experimental data are available, for instance EST databases, then GENQUEST could be used to reduce the complexity of the EST databases for sensitive searching in much the same way as demonstrated here for the raw genomic sequence. These experiments are ongoing in our laboratory and will provide additional information for genome annotation with regards to alternative splicing and SNPs that are represented in the EST database. Alternatively, two elegant computational solutions are available for searching the EST databases and genomic databases for alternative splicing, SNPs, and novel peptides. Edwards[22] recently demonstrated a 35-fold reduction in EST databases for MS/MS searching, while Tanner et al.[23] created novel "exon graphs" from EST and gene prediction algorithms for MS/MS searching.

Although the emphasis of this research was on searching genomic sequences with MS/MS data for genome annotation applications, a further advantage of this research is in the ability to restrict the MS/MS search space in other proteomics experiments. As such, it has applications much broader than searching complex genomic sequences. Any application where the current search space is too great and reduces the S/N to a point where sensitivity is reduced unacceptably can benefit from this technique. For instance, peptide databases can be generated for all potential SNPs within the protein coding regions of the human genome, without any preconceived notions of what SNPs are actually present. The size of this

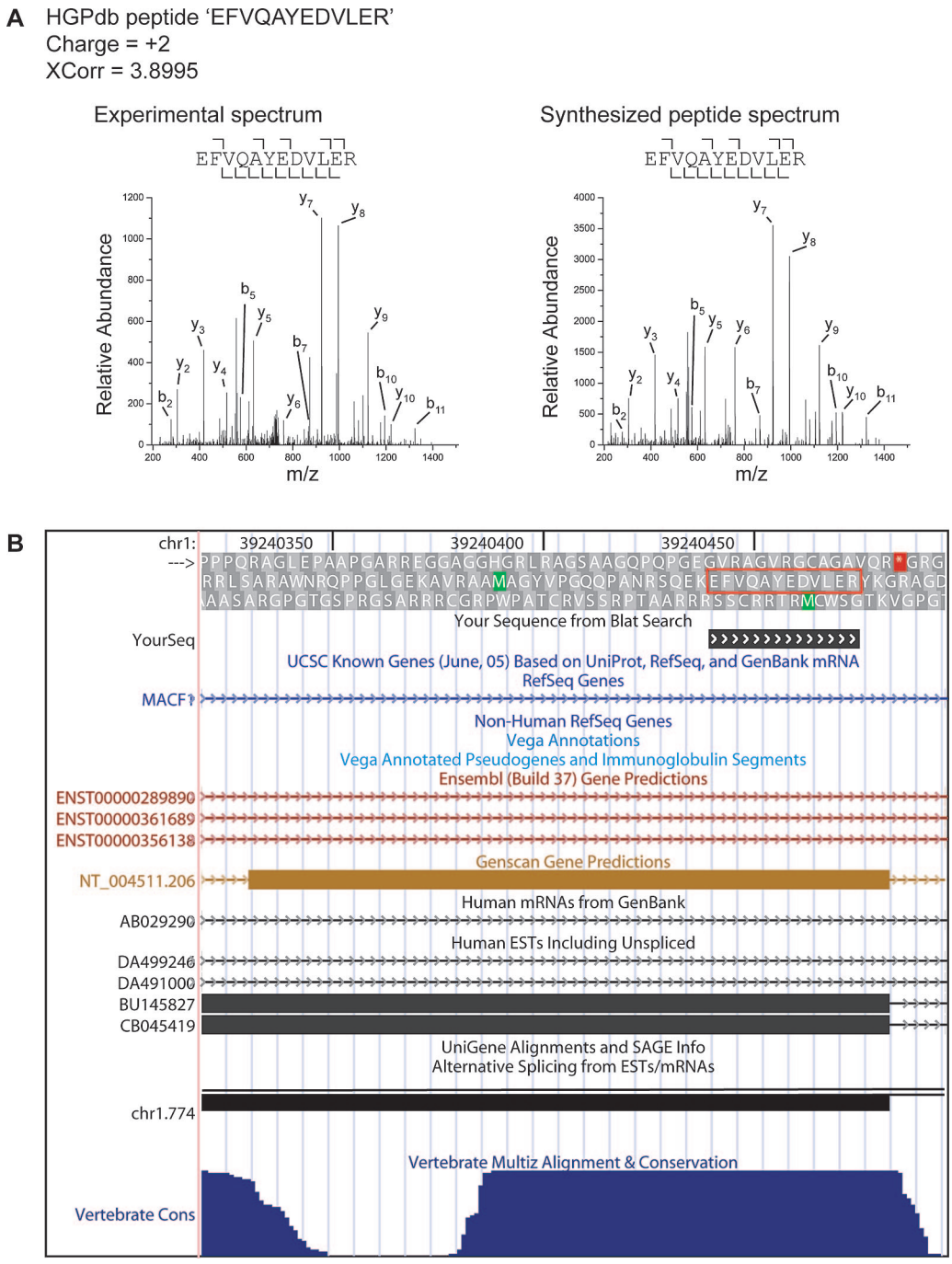

Figure 4. Identification of a novel exon within MACF1. (A) Experimental peptide versus synthesized peptide MS/MS spectra comparison. (B) UCSC genome browser screenshot (http://genome.ucsc.edu) showing alignment of peptide EFVQAYEDVLER (boxed in red) within the human genome (May 2004 build).

peptide database would be almost 2 orders of magnitude greater than current protein databases and as such is an excellent candidate for GENQUEST-like techniques. Furthermore, once the effects of PTMs on the pI of peptides can be accurately predicted, peptide databases of all potential PTMs can be constructed and searched as shown in this research. We expect that the currently available MS/MS search engines will begin to take advantage of peptide isoelectric point in reducing the MS/MS search space and provide faster, higher fidelity searching than is currently available.

Lastly, we wish to emphasize the importance of manual and experimental validation of proteomic results when attempting to identify novel genes. It would have been convenient for us to state that there were potentially 540 novel exons discovered in this study or to rely upon representation in EST databases alone, but our manual and experimental analysis has only identified one potential new exon within the 166 MS/MS spectra examined thus far. We expect the remaining 374 unexamined MS/MS spectra specific to the HGPdb searches to uncover only a modest number of novel exons or open reading frames. Since over 2200 peptides arising from a known protein were identified when searching the HGPdb and so few novel peptides have been validated, it suggests that the human genome is fairly well annotated with respect to the location of

exons and ORFs, but much work still remains to fully annotate the genome with respect to alternative splicing and SNPs.

Abbreviations: D, aspartic acid; E, glutamic acid; I, isoleucine; L, leucine; K, lysine; R, arginine; MB, megabyte; GB, gigabyte.

Acknowledgment. The authors wish to acknowledge Brett Phinney for preliminary experiments prior to this work, David Kroll and Nick Oberlies for DU4475 extracts, and Jonathan Bundy for critical review of the manuscript. This work was supported by the RTI Internal Research and Development program, a contract (HHSN2662004000670) from the National Institute of Allergy and Infectious Disease, and a research grant from Merck & Co.

Supporting Information Available: Supplemental Table 1, the summary of all peptide identifications, and Supplemental Table 2, the summary of manual analysis of 166 MS/MS spectra found in the HGPdb search. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Elsik, C. G.; Mackey, A. J.; Reese, J. T.; Milshina, N. V.; Roos, D. S.; Weinstock, G. M. *Genome Biol.* **2007**, *8*, (1),R13.
(2) Craig, R.; Beavis, R. C. *Bioinformatics* **2004**, *20* (9), 1466–1467.
(3) Eng, J. K.; McCormack, A. L.; Yates, J. R. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (11), 976–989.
(4) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20* (18), 3551–3567.
(5) Yates, J. R.; Eng, J. K.; McCormack, A. L. *Anal. Chem.* **1995**, *67* (18), 3202–3210.
(6) Choudhary, J. S.; Blackstock, W. P.; Creasy, D. M.; Cottrell, J. S. *Proteomics* **2001**, *1* (5), 651–667.
(7) Colinge, J.; Cusin, I.; Reffas, S.; Mahe, E.; Niknejad, A.; Rey, P. A.; Mattou, H.; Moniatte, M.; Bougueleret, L. *J. Proteome Res.* **2005**, *4* (1), 167–174.
(8) Fermin, D.; Allen, B. B.; Blackwell, T. W.; Menon, R.; Adamski, M.; Xu, Y.; Ulintz, P.; Omenn, G. S.; States, D. J. *Genome Biol* **2006**, *7* (4), R35.
(9) Syka, J. E.; Marto, J. A.; Bai, D. L.; Horning, S.; Senko, M. W.; Schwartz, J. C.; Ueberheide, B.; Garcia, B.; Busby, S.; Muratore, T.; Shabanowitz, J.; Hunt, D. F. *J. Proteome Res.* **2004**, *3* (3), 621–626.
(10) Cargile, B. J.; Sevinsky, J. R.; Essader, A. S.; Eu, J. P.; Stephenson, J. L., Jr. *Electrophoresis* **2006**, in press.
(11) Cargile, B. J.; Bundy, J. L.; Freeman, T. W.; Stephenson, J. L., Jr. *J. Proteome Res.* **2004**, *3* (1), 112–119.
(12) Cargile, B. J.; Bundy, J. L.; Stephenson, J. L., Jr. *J. Proteome Res.* **2004**, *3* (5), 1082–1085.
(13) Cargile, B. J.; Talley, D. L.; Stephenson, J. L., Jr. *Electrophoresis* **2004**, *25* (6), 936–945.
(14) Shepelev, V.; Fedorov, A. *Briefings Bioinf.* **2006**, *7* (2), 178–185.
(15) Moore, R. E.; Young, M. K.; Lee, T. D. *J. Am. Soc. Mass Spectrom.* **2002**, *13* (4), 378–386.
(16) Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. *J. Proteome Res.* **2003**, *2* (1), 43–50.
(17) Saxonov, S.; Daizadeh, I.; Fedorov, A.; Gilbert, W. *Nucleic Acids Res.* **2000**, *28* (1), 185–190.
(18) Cargile, B. J.; Sevinsky, J. R.; Essader, A. S.; Stephenson, J. L., Jr.; Bundy, J. L. *J. Biomol. Technol.* **2005**, *16* (3), 181–189.
(19) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. *J. Mol. Biol.* **1990**, *215* (3), 403–410.
(20) Kent, W. J.; Sugnet, C. W.; Furey, T. S.; Roskin, K. M.; Pringle, T. H.; Zahler, A. M.; Haussler, D. *Genome Res.* **2002**, *12* (6), 996–1006.
(21) Kuster, B.; Mortensen, P.; Andersen, J. S.; Mann, M. *Proteomics* **2001**, *1* (5), 641–650.
(22) Edwards, N. J. *Mol. Syst. Biol.* **2007**, *3*, 102.
(23) Tanner, S.; Shen, Z.; Ng, J.; Florea, L.; Guigo, R.; Briggs, S. P.; Bafna, V. *Genome Res.* **2007**, *17* (2), 231–239.

PR070198N