

Mapping the Proteome of *Drosophila melanogaster*: Analysis of Embryos and Adult Heads by LC–IMS–MS Methods

John A. Taraszka,^{†,§} Ruwan Kurulugama,[†] Renā A. Sowell,[†] Stephen J. Valentine,[†]
 Stormy L. Koeniger,[†] Randy J. Arnold,[†] David F. Miller,[‡] Thomas C. Kaufman,[‡] and
 David E. Clemmer^{*,†}

Department of Chemistry, Indiana University, Bloomington, Indiana 47405, Department of Biology, Indiana University, Bloomington, Indiana 47405, and Novartis Institutes for Biomedical Research Inc., Cambridge, Massachusetts 02139

Received February 17, 2005

Multidimensional separations combined with mass spectrometry are used to study the proteins that are present in two states of *Drosophila melanogaster*: the whole embryo and the adult head. The approach includes the incorporation of a gas-phase separation dimension in which ions are dispersed according to differences in their mobilities and is described as a means of providing a detailed analytical map of the proteins that are present. Overall, we find evidence for 1133 unique proteins. In total, 780 are identified in the head, and 660 are identified in the embryo. Only 307 proteins are in common to both developmental stages, indicating that there are significant differences in these proteomes. A comparison of the proteome to a database of mRNAs that are found from analysis by cDNA approaches (i.e., transcriptome) also shows little overlap. All of this information is discussed in terms of the relationship between the predicted genome, and measured transcriptomes and proteomes. Additionally, the merits and weaknesses of current technologies are assessed in some detail.

Keywords: *Drosophila melanogaster* • ion mobility • proteomics • development

Introduction

Drosophila melanogaster (the fruit fly, hereafter referred to as *Drosophila*) displays four distinct developmental stages: the embryonic stage characterized by rapid mitotic activity and cell differentiation that extends from 0 to 22 h; the larval stage characterized by three molts encompassing 22 h to 7 d; the pupae stage where larval structures are replaced with adult structures (from 7 to 11 d); and, an adult that generates no new cells (except for cells associated with the gonads and gastrointestinal system) and lives for ~60 d.^{1,2} Throughout these stages, there are complex changes in the organism's morphology and physiology that can be initiated by peptide and steroid hormones.³ Such changes must involve a cascade of events that regulate the expression of the genome.³ Typically, genetic and immunohistochemical methods are used to study the development of various tissues; and, studies of the brain,^{4,5} eye,⁶ wings,⁷ and genitalia⁸ have been reported. It is also possible to investigate the regulation of gene expression as a function of development. To this end, small (20–30 nucleotides) messenger RNAs (mRNAs),⁹ cell cycle regulations,¹⁰ and global mRNA expressions using DNA microarrays have been investigated.¹¹

Although *Drosophila* proteins have been the subject of many reports,¹² few studies have characterized large numbers of

proteins. Vierstraete and his collaborators accumulated a database (containing about 40 entries) of larval hemolymph proteins identified from two-dimensional gels by mass spectrometry (MS) analysis.¹³ Hunt and co-workers presented a preliminary study of the proteins in the sperm of *Drosophila* and have identified 251 proteins (near the total number expected from two-dimensional gels).¹⁴ In addition, Heck and co-workers reported a quantitative metabolic labeling method for *Drosophila* (and *C. elegans*).¹⁵ Most recently our group has profiled the proteomes associated with three individual fly heads using techniques that are similar to those described below.¹⁶ Other recent studies have characterized peptides in the nervous system^{17–18,19} as well as peptides from larval hemolymph fluid (the fly equivalent of blood).²⁰ In addition, analysis of genome and transcriptome data has provided the first *Drosophila* protein interaction map.²¹

In this paper, we report the development of a multidimensional analytical approach for the direct characterization of proteomes. This approach involves the construction of what we refer to as a proteome map, where tryptic peptides associated with specific proteins are positioned at reproducible locations within an analytical space. This map makes it possible to assess the proteome at different developmental stages. Here, we report: (1) the construction of an initial tryptic peptide map for two states of the *Drosophila* proteome (the whole embryo and the adult head); (2) the direct identification of proteins; and, (3) a global comparison of the genome, transcriptome, and proteome of these two states. In total, we find evidence

[†] Department of Chemistry, Indiana University.

[‡] Department of Biology, Indiana University.

[§] Novartis Institutes for Biomedical Research Inc..

for 1133 unique proteins: 780 in the adult head and 660 from the embryo. Only 307 are common to both states. The advantages and limitations of techniques that are used to construct the map are considered in detail.

Many of the results presented here can be compared with the *Drosophila* genome database FlyBase.²² FlyBase predicts that the *Drosophila* genome contains 13 809 genes. One advantage of working with a model organism, for which the genome is sequenced,²³ is the ability to assign genes and gene products to known biological pathways. To accomplish this, a well-defined nomenclature for describing genes and their products that is general to any organism has been developed by the Gene Ontology (GO) consortium.²⁴ There are three general GO categories: *biological process* that describes the biological role of the gene or gene product; *molecular function* that defines the biochemical activity of the gene product; and *cellular component* that defines the cellular location of where a gene product is active. In *Drosophila*, 9159 out of 13 809 (66%) genes have GO entries, but only 4314 out of 13 809 (31%) are associated with cellular components. Below, we classify mRNAs and proteins by their GO cellular component, allowing us to compare how aspects of cells differ in alternate states of *Drosophila*—in this case the whole embryo (the first developmental stage) and the adult head (from the final developmental stage). When proteins are classified by their cellular component, we find that cells from the embryo and head show substantial differences in protein expression. Some of these variations can be rationalized in terms of the different functions of these states.

The studies reported here build on significant advances in instrumentation and experimental protocols.^{25,26} Analytical platforms for proteomics must offer high throughput and peak capacity.²⁵ One of the more common approaches combines multiple dimensions of condensed-phase separations with mass spectrometry.²⁶ Several groups are adopting this strategy to characterize post-translational modifications,^{27–30} assess relative and absolute peptide abundances,^{31–35} and analyze extracted cellular components.³⁶ Our group has worked to include an additional separation between the condensed phase and mass spectrometry analysis, where gas-phase ions are separated based on differences in the mobilities of the ions through a buffer gas.^{37–40} This approach is described below. We discuss for the first time the use of a combined mobility approach with other commercial approaches and describe the strengths and weaknesses of each strategy independently as well as the combined approach.

Experimental Section

Protein Isolation and Tryptic Digestion. Samples were prepared using the following procedures. Wild-type Oregon-R *Drosophila* were grown at 25 °C on standard media that was supplemented with bakers yeast.⁴¹ Adult heads were obtained from one week old females. In this study, we used a population of 166 adult heads. Populations of embryos (Oregon-R) that spanned the complete range of times associated with this developmental stage (0–22 h) were harvested. We estimate that the population of embryos included ~1000 individuals. Proteins from heads and embryos were extracted using a mortar and pestle into 500 μ L of phosphate buffered saline containing 8 M urea and 0.1 mM α -toluenesulfonyl flouride. A Bradford assay indicated that 2.8 mg and 7.5 mg of protein were recovered from the heads and embryos, respectively. Disulfide bonds were reduced and alkylated by addition of dithiothreitol

at a 1:40 mole ratio; after 2 h of incubation at 37 °C, iodoacetamide was added at a 1:80 mole ratio, and the sample incubated in darkness at 0 °C for 2 h. Cysteine was added at a 1:40 mole ratio to quench the reaction.

Tryptic peptides were produced as follows. The solution containing reduced and alkylated proteins was diluted to a final urea concentration of 2 M with 0.2 M Tris buffer (pH = 8.0, 10 mM CaCl₂), and TPCK-treated trypsin (2% of enzyme by mass to that of the protein) was added. Samples were incubated for 24 h at 37 °C. Tryptic peptides were desalted using Oasis hydrophilic–lipophilic balance (HLB) cartridges (Waters Inc., Milford, MA) and dried on a centrifugal concentrator.

Strong-Cation Exchange (SCX) Fractionation. Tryptic peptides were separated into fractions using SCX chromatography. A Waters system consisting of a 600 Pump and 2487 Dual Wavelength detector (Waters Inc., Milford, MA) was used. Separation was performed on a 100 \times 2.1 mm column packed with 5 μ m 200 Å Polysulfethyl A (PolyLC Inc., Columbia, MD). Peptides were fractionated at a flow of 0.2 mL·min⁻¹ into 96-well plates using 1 min intervals and the following gradient: 0% B for 5 min, 0–20% B in 40 min, 45–90% B in 45 min, 90–100% B in 10 min, 100% B for 10 min [A = 5 mM potassium phosphate, pH = 3 (75:25 water:acetonitrile); B = 5 mM potassium phosphate, 0.35 M potassium chloride, pH = 3 (75:25 water:acetonitrile)]. Fractionation was monitored by measuring the absorbance of the eluting peptides at λ = 214 nm. After fractionation, the individual wells from the 96 well plates were pooled into 10 fractions using chromatographic peak profiles. This was done to keep the concentration of peptides in each fraction relatively consistent. Individual fractions were desalted with Oasis HLB cartridges, dried on a centrifugal concentrator, and stored at –80 °C until further analysis. Examination of peptides that are identified in both embryo and head samples indicates that 73% elute in the same SCX fraction and 25% elute in adjacent SCX fractions. Only 2% of the peptides identified are observed to elute in nonadjacent fractions.

Description of the LC Conditions Employed. Nanoflow reverse-phase separation was accomplished using an Agilent 1100 CapPump (Agilent Technologies Inc., Palo Alto, CA). In this setup, peptides from one of the fractions were loaded at 4 μ L·min⁻¹ onto a 1.5 cm \times 100 μ m i.d. trapping column (IntegraFrit from New Objectives Inc., Woburn, MA) packed with 5 μ m 200 Å Magic C18AQ (Microm BioResources Inc., Auburn, CA) stationary phase. After 12 min the flow was reduced to 250 nL·min⁻¹, and peptides were separated on a pulled-tip analytical column [15 cm \times 75 μ m i.d. packed with 5 μ m, 100 Å Magic C18AQ] using a gradient of 0–5% B in 5 min, 5–20% B in 50 min, 20–40% B in 40 min, 40–80% B in 5 min, 80% B for 10 min, 80–0% B in 5 min, 0% B for 15 min (A = 96.95% water, 2.95% acetonitrile, 0.1% formic acid; B = 99.9% acetonitrile and 0.1% formic acid). The pulled-tip column was made by heating 75 μ m i.d. fused silica (Polymicro Technologies LLC, Phoenix, AZ) in a microflame torch (Microflame Inc., Plymouth, MN); once pulled a methanol slurry of the stationary phase was packed into the column at a pressure of 69 bar.

General Overview of IMS–MS and IMS–(CID)–MS Experiments. A schematic diagram of the ion mobility spectrometry (IMS) instrument used in these experiments is shown in Figure 1. Many researchers have used ion mobility approaches as an analytical separations approach⁴² and a structural probe,⁴³ and authoritative reviews are available.⁴⁴ A detailed description of the instrument used in these studies has also been discussed

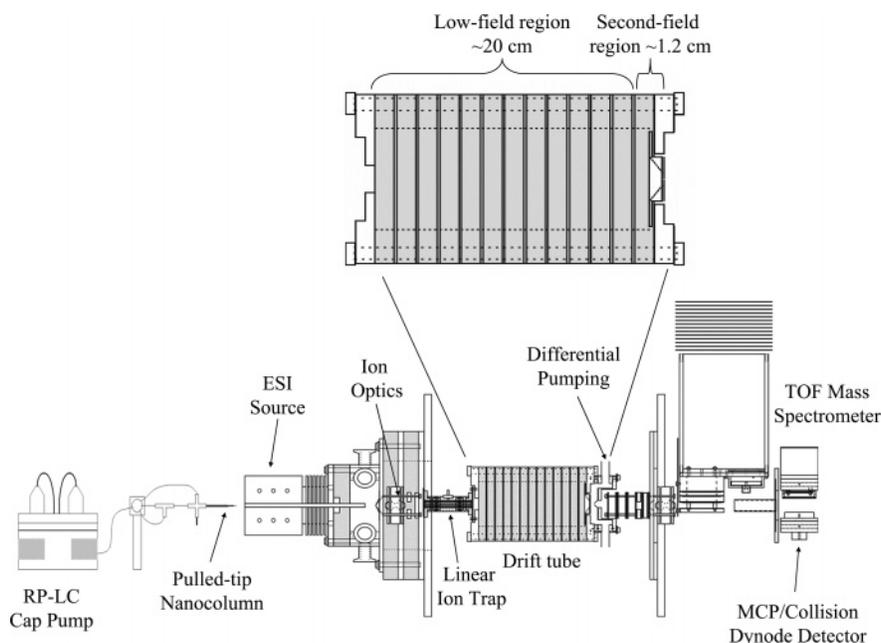


Figure 1. Schematic of the LC–IMS–MS instrument used in these experiments. This instrument consists of a nano LC column coupled to an ESI source. The continuous beam of ions is accumulated in a linear octopole trap. IMS–MS experiments are initialized by injecting pulses of ions from the trap into the drift tube. The drift tube incorporates a split-field design. The insert illustrates the different regions of the split-field drift tube. The first region is operated under low-field conditions, and the second region is modulated between low- and high-field conditions (see text).

previously.⁴⁵ Briefly, IMS–MS analysis is performed as follows. Peptides eluting from the pulled-tip nanocolumn are electro-sprayed into a linear octopole ion trap where ions are accumulated and are stored for the pulsed IMS experiments. IMS measurements are initiated by injecting a 100 μ s pulse of ions into a drift tube containing ~ 1.65 Torr of 300 K He buffer gas. The drift tube used is a new design that incorporates a split-field configuration.⁴⁵ Ions migrate across the first-field region under the influence of a weak applied electric field (~ 5 V \cdot cm⁻¹) and separate based on differences in their mobilities through the buffer gas. Compact ions (with small cross sections) have higher mobilities than more extended conformations.^{43,44} Also, high charge states have higher mobilities than low charge states because they experience a greater drift force (qeV).^{46,47} This region comprises most of the drift tube length (~ 20 cm). As ions exit the low-field region they enter a much shorter (~ 1.2 cm) second-field region (see Figure 1) that can be operated under low-field conditions (to transmit precursor ions) or high-field conditions (to induce fragmentation).⁴⁵

Under low-field conditions the internal temperature of the ion is characterized by the temperature of the buffer gas.⁴⁸ Under these conditions, the mobility is independent of the applied electric field, and the ion velocity is proportional to the applied field. Under high-field conditions, it is possible to collisionally activate the ions and induce fragmentation.⁴⁵ Here, we employ a high-field region at the back of the drift tube to activate ions. In this setup, the experimental drift time [$t_D(\text{total})$] is the sum of the time spent in both field regions [i.e., $t_D(\text{total}) = t_{D1} + t_{D2}$, where t_{D1} is the drift time in the first (low-field) region, and t_{D2} is the drift time in the second-drift region (which may correspond to motion of the precursor ion or its fragments)]. In these experiments, t_{D1} is constant, but t_{D2} is modulated between low- and high-field conditions. This modulation makes it necessary to calibrate $t_D(\text{total})$ between collision induced dissociation (CID) and precursor ion datasets.

This is done using a multipoint calibration to a known system. The values that are reported in the analytical map for drift times correspond to parent ion conditions and are effectively low-field mobility measurements. Experiments are normalized to a buffer gas pressure and temperature of 1.70 Torr and 300 K, respectively.

As the ions exit the drift tube they are extracted and focused into the source region of an orthogonal time-of-flight (TOF) reflectron MS. Because flight times (t_F) are much shorter than the drift times, hundreds of mass spectra can be collected for a single drift pulse.⁴⁹ Flight times are converted into mass-to-charge (m/z) values using simple calibration procedures.⁵⁰

LC–QIT Conditions. LC–MS experiments were performed on a LCQ Deca XP quadrupole ion trap (QIT) mass spectrometer (ThermoElectron Inc., Waltham, MA) coupled to a nano-flow LC system (Dionex Inc., Sunnyvale, CA). The LC gradient employed was the same as that used for the LC–IMS–MS experiments. The instrument was operated in a data dependent mode, in which a full-scan mass spectrum (m/z range = 250 to 1500) was followed by a MS/MS acquisition using the following instrument specific parameters: an isolation window of 2 m/z ; a normalized collision energy of 35%; and, a dynamic exclusion time of 1 min. Under these conditions, precursor ions are isolated using a 2 m/z width isolation window, and after a precursor ion peak is selected for MS/MS analysis, the same peak is not reselected for MS/MS analysis for a period of 1 min.

Calibration, Resolution, and Reproducibility of the LC–IMS–MS Map. Overall. To compare proteome states at the peptide level it is necessary to calibrate the positions of all peptide peaks in each of the dimensions of the analytical space used for the map. It is important for the context of this paper to describe: (1) the calibrations that are used for comparison of data from different instruments as well as different states of the proteome (2) the overall resolution of these dimensions;

Table 1. Abridged List of Peptides And Parent Ion Peak Positions $t_{\text{SCX}}\{\text{RRT}[t_{\text{D}}(m/z)]\}$ Contained in the *Drosophila* Trypic Peptide Map

FBgn no. ^a	mRNA ^b	protein ^c	cell component ^d	peptide sequence	t_{SCX}^e	RRT ^f	t_{D}^g	m/z^h
FBgn0000024	head	head	cytoplasm, plasma membrane	KPVPAEPWHGVLDATR	71–78	0.943		591.87
FBgn0000052	ND	embryo	not specified	GLLLDEALER	44–48	1.037	4.372	1130.50
FBgn0000053	ND	head	not specified	DSGVDDIDAGDALVQR	38–43	1.071		766.59
				EACQAVDEILGDLK	38–43	1.125		781.37
FBgn0000055	ND	both	not specified	AAVVNFTSSLAK	44–48	0.718	2.987	604.77
				AAVVNFTSSLAK	44–48	0.719	5.633	1207.17
				AIELNQNGAIWK	44–48	0.811	2.390	453.09
				DGCDFAK	16–37	0.568	3.233	812.94
				IENPAAIAELK	44–48	0.726	2.944	584.95
				LDLGTLEAIQWTK	44–48	1.405	3.243	744.88
				NVIFVAGLGGIGLDTSK	38–43	1.416	3.401	831.78

^a The FlyBase gene number is provided as protein identification. ^b mRNA is present in whole embryo (**embryo**), adult head (**head**), or both cDNA libraries (**both**). ND indicates that the mRNA is not detected. ^c The protein is identified in whole embryo (**embryo**), adult head (**head**) or both samples (**both**) in our proteomics experiments. ^d The cellular location is obtained from GO data accessed on FlyBase. ^e The range of SCX retention times (t_{SCX}) of the SCX fractions are listed in minutes. ^f The relative retention time (RRT) of the peptide is measured with respect to the retention time of leucine enkephalin. ^g Experimental drift times (t_{D}) have been normalized to a He pressure of 1.70 Torr. During the experiments the electric field in the low-field drift region is constant. Not all peptides have been assigned drift times; these peptides are only mapped in three dimensions (t_{SCX} , RRT, and m/z). ^h The experimental m/z ratio for the parent ion is provided.

and (3) reproducibility of each of the LC, IMS and MS measurements.

IMS and TOFMS Calibration, Resolution and Reproducibility. The m/z resolving power [$m/\Delta m$, using the full width at half-maximum (fwhm) definition] ranges from ~600 to 1000 in these experiments; in the case of LC–IMS–MS studies, this is limited by our home-built electronics in the IMS-TOF detection system. The IMS resolving power [$t_{\text{D}}(\text{total})/\Delta t_{\text{D}}(\text{fwhm})$] ranges from ~17 to 35 for different ions across the spectrum, and drift times measured in a single-field drift tube for a single parent ion recorded in any two measurements are normally reproducible (after normalization of buffer gas pressure, temperature and drift field) to within 1% (relative uncertainty). In the split-field design incorporated here, minor changes in the focusing fields at the back cause any two measurements to be reproducible to within 2%.

LC Calibration, Resolution, and Reproducibility. Typical resolving powers for peaks in the LC separation range from $t_{\text{R}}/\Delta t_{\text{R}}(\text{fwhm}) = 100$ to 300, where t_{R} is the measured retention time of a peak. To construct the tryptic peptide map two types of calibrations across the LC separation are used. The first calibration is imposed so that LC–QIT data can be directly compared with data recorded for the same sample with the LC–IMS–MS instrument. This is accomplished using an empirical calibration between retention times recorded using the LC–QIT instrument and those recorded using the LC–IMS–MS instrument. Typically, we use 20 peaks within each dataset for this calibration. For any two measurements (involving the same sample, i.e., same SCX fraction for the same proteome state) it is possible to align LC runs such that peak positions are reproducible to within $\pm 2\%$ (relative uncertainty). In the tabulated map, only assigned peaks (from LC–QIT data) that can be unambiguously superimposed to a single parent ion peak (within the LC–IMS–MS data) at a unique location are included as assigned peaks in the LC–IMS–MS data. For example, if two peptides are mapped to the same LC–IMS–MS location, both are excluded unless one of the sequences is verified by analysis of the LC–IMS–(CID)–MS data. This leads to a number of cases where peak assignments are provided from LC–QIT data but are not mapped in the LC–IMS–MS data.

The second calibration uses relative retention times (RRT) to facilitate comparisons between samples (e.g., different SCX

fractions or proteome states). The retention times of peaks that are assigned to peptides in LC–IMS–MS datasets [either from LC–QIT data or from LC–IMS–MS data] are converted to RRTs that are determined with respect to an internal standard [in this case leucine enkephalin (Sigma, min. 95% purity), that has been spiked into all fractions at a 1 μM concentration]. Values of RRT are given by eq 1

$$\text{RRT} = \frac{t_{\text{Ri}}}{t_{\text{R(L-enk)}}} \quad (1)$$

here t_{Ri} is the retention time of peptide i and $t_{\text{R(L-enk)}}$ is the retention time of leucine enkephalin. The values of RRTs for the same peptides (found in different samples) for any two measurements are reproducible to within 7% (relative uncertainty) for all of the peaks between different SCX fractions and proteome states that we have examined. This uncertainty is relatively large because we have considered data from different columns. For any two back-to-back measurements on the same column, RRTs are typically reproducible to within 2%.

Nomenclature Associated with LC–IMS–(CID)–MS Analysis. Because of the large differences in time scales associated with each of the LC, IMS, and MS dimensions in these experiments, data are acquired in a nested fashion.³⁷ For raw data we report values of time scales associated with the different separation dimensions (e.g., t_{R} , t_{D} , and t_{F} , respectively) using a nomenclature that we have described previously.³⁹ In this system, values are bracketed in order to indicate the position of the peak within the nested measurement. For example, in a single three-dimensional LC–IMS–MS measurement, the position of a single peak would be indicated by the following: $t_{\text{R}}[t_{\text{D}}(m/z)]$ in units of min[ms(u/z)]. LC–MS data (from the commercial system) can be delineated with the same nomenclature; in this case there is no value for the drift time and peaks are given in values of $t_{\text{R}}(m/z)$. Here, we also include the range of times that fractions are collected from the SCX separation. In this case, the SCX retention times (t_{SCX}) are reported as the range of times (in minutes) over which the fraction was collected. Thus, a single peak would be delineated by $t_{\text{SCX}(\text{start})} - t_{\text{SCX}(\text{finish})}\{t_{\text{R}}[t_{\text{D}}(m/z)]\}$. In the proteome maps (Table S-1, Tables 1, 2) positions of peaks are reported as $t_{\text{SCX}(\text{start})} - t_{\text{SCX}(\text{finish})}\{\text{RRT}[t_{\text{D}}(m/z)]\}$, where RRT is defined from eq 3.

Table 2. List of Rhabdomere Peptides and Parent Ion Peak Positions $t_{SCX}\{RRT\{t_D(m/z)\}$ Contained in the *Drosophila* Tryptic Peptide Map

FBgn no. ^a	mRNA ^b	protein ^c	cell component ^d	peptide sequence	t_{SCX} ^e	RRT ^f	t_D ^g	m/z ^h
FBgn0000120	head	head	rhabdomere	AGIAVEGDIK	44–48	0.546	4.779	973.54
				DTALASTLLIASQDAR	38–43	0.891	3.430	817.32
				ELTLVSQQVCPQK	38–43	0.730	3.444	814.25
FBgn0000121	head	head	rhabdomere	HGIALDGHILK	79–98	0.455	2.581	531.33
				VFGQLATTYR	44–48	0.662	2.902	578.59
FBgn0000253	embryo	both	cytoplasm, rhabdomere	DGNGFISAAELR	38–43	0.924	4.009	1249.84
				DGNGFISAAELR	44–48	0.912	2.902	625.71
				EAFSLFDKGDGTITTK	54–62	1.039	2.686	616.14
				VFDKDGNGFISAAELR	54–62	1.028	3.628	870.47
				VFDKDGNGFISAAELR	54–62	1.028	2.686	580.56
FBgn0001263	head	head	rhabdomere	HAEVGSGIFISDLR	63–70	0.987		751.04
				NSTEQAVIDLK	16–37	1.314		666.35
FBgn0002938	head	head	cytoplasm, rhabdomere	ALGVLDTVIAR	44–48	1.002	2.902	564.7
				EPQHIVLSGESYSYGK	54–62	0.579	2.386	544.79
				EVNSSQLGPLPVPK	38–43	0.973	3.358	790.07
				LPFDEFRLR	44–48	1.270	2.646	519.23
				LVDFHNR	44–48	0.922	2.518	495.44
FBgn0002940	head	head	endoplasmic reticulum, rhabdomere	YYNDEFRLR	49–53	0.779	2.772	595.98
				SSDAQSQATASEAESK	44–48	0.474		799.17
				SSDAQSQATASEAESKA	44–48	0.578		834.58
FBgn0003861	head	head	rhabdomere	VGQSSAAAGGER	44–48	0.037		545.44
FBgn0004435	embryo	head	plasma membrane, rhabdomere	IEQADYLPTEQDILR	38–43	1.095	3.616	903.16
				YYLSDLAR	49–53	0.808	2.643	501.1
FBgn0004625	head	head	rhabdomere	EPPLVFEPVTLESRLR	38–43	1.447	3.647	863.79
				NDIEELFTSITK	44–48	1.430	3.200	706.25
				QIEEFSTDVQK	44–48	0.544	2.731	662.66
				VVLPDLAVLR	44–48	1.267	3.670	1095.06
				LDNILLDGEGHVK	54–62	0.893		712.28

^a The FlyBase gene number is provided as protein identification. ^b mRNA is present in whole embryo (**embryo**), adult head (**head**), or both cDNA libraries (**both**). ND indicates that the mRNA is not detected. ^c The protein is identified in whole embryo (**embryo**), adult head (**head**) or both samples (**both**) in our proteomics experiments. ^d The cellular location is obtained from GO data accessed on FlyBase. ^e The range of SCX retention times (t_{SCX}) of the SCX fractions are listed in minutes. ^f The relative retention time (RRT) of the peptide is measured with respect to the retention time of leucine enkephalin. ^g Experimental drift times (t_D) have been normalized to a He pressure of 1.70 Torr. During the experiments the electric field in the low-field drift region is constant. Not all peptides have been assigned drift times; these peptides are only mapped in three dimensions (t_{SCX} , RRT, and m/z). ^h The experimental m/z ratio for the parent ion is provided.

Criteria Used for Assignment of Peaks and Identification of Peptides (and Proteins) from Comparisons of MS–MS and CID–MS Data with Databases. Protein identifications rely on assignments of peptide sequences using m/z information from precursor and fragment ion datasets.²⁶ In this approach, the m/z value of the precursor ion is used in conjunction with the m/z values for fragment ions that are generated under well-defined collision conditions (either imposed by energizing collisions in the ion trap, or conditions encountered under high-field conditions associated with the second region of the drift tube). These values are then used as inputs for programs that search protein databases for probable tryptic peptide assignments. In these experiments, the MASCOT program (Matrix Science Ltd., London, UK) is used to search the National Center for Biotechnology Information *Drosophila* protein database.^{51,52}

A protein is considered identified only if at least one peptide having a sequence that is unique to a protein has a significant score. In this approach, the search considers expected tryptic peptides as well as sequences that would be created upon missing up to two cleavages. The precursor and fragment ion mass tolerances were set to ± 2.0 and ± 1.0 u, respectively (for both LC–QIT and LC–IMS–MS approaches). Carbamidomethylation of cysteine residues was specified as a fixed modification; no other variables were included for possible modifications. A significant score (in this case, a value of greater

than 37 as output from the search) indicates that the peptide match has a less than 5% chance of occurring at random.⁵¹ If a score from a search of the MS/MS data is not significant, then the peptide identification is discarded in an automated fashion using a Protein Results Parser program written in-house.⁵³ In all cases, MS and MS/MS spectra that yielded significant scores were also examined manually to check for any obvious false positives. Higher scores indicate a greater certainty of an actual (nonrandom) assignment. Also, the possibility of misidentifying a protein decreases when multiple peptides from the same protein are identified, or when replicate experiments lead to the same identification. At this point, we have carried out 10 replicate experiments involving the head and most proteins (>90%) that are identified for a single peptide have been confirmed in at least one of the replicate experiments. Similarly, replicate experiments involving the embryo (although fewer ~5) also confirm most (~90%) of the single hit assignments for this state.

Comparison of Identified Proteins to mRNA Expression Data (reported previously). It is often useful to compare identified proteins to mRNA transcripts. Transcript libraries were obtained by the Berkeley *Drosophila* Genome Project for the same embryo and head states of *Drosophila*.⁵⁴ These libraries were constructed from saturated sets of complementary DNA (cDNA)/expressed sequence tag (EST) clones that were recovered and sequenced. Although the mRNA recovery is

not complete, this analysis provides an initial estimate of the transcriptome. Identified head proteins are compared to GH, HL and RH cDNA libraries (mRNA source is adult head), and the embryonic proteins are compared to LD and RE libraries (mRNA source is 0–22 h embryos). cDNA technology has characterized 3775 and 4864 mRNA transcripts in adult heads and whole embryos, respectively.⁵⁴

Results

General Considerations of the LC–QIT and LC–IMS–MS Analysis for Mapping Proteomes: Differences in Approach and Utilizing Complementary Information. The results that are presented here are summarized as an analytical map of the proteome of these two states of *Drosophila*. Table 1 is presented to illustrate the general format of the map. The complete list of the positions of peaks for specific peptides and proteins that have been identified (as discussed below) as well as relevant genomic information are provided as Supporting Information (Table S-1). Many of the techniques that are described above for the LC–IMS–MS approach are motivated by the capabilities of the LC–QIT platform. During the course of a typical LC–QIT analysis more than 1500 precursor- and 1500 MS/MS-spectra are acquired. It is difficult to overstate the value and utility of this technology for proteomics analysis. This revolutionary technology is capable of rapidly providing a characterization of a proteome. Our intent in this section is 2-fold: to delineate some features of the LC–QIT approach technique that are not ideal, specifically aspects that may be addressed by the development of LC–IMS–MS strategies; and, to show how the LC–QIT and LC–IMS–MS techniques can complement one another to identifying specific proteins and obtain extensive proteome coverage.

Sampling Limitations Associated with the LC–QIT Analysis. The LC–QIT approach makes it possible to obtain MS and MS/MS information for peptides as they elute from the LC column. For very complex mixtures (such as those analyzed here) this approach is subject to sampling errors that influence the reproducibility. That is, the approach misses many components that are present in the sample during the time that some components are selected for MS/MS experiments. Even back-to-back runs of the same fraction of peptides in this study differ in the peptides that are identified by as much as 60%. Although faster scanning instruments (such as linear ion traps) are now available they still have the same fundamental limitations as a more traditional QIT instrument. However, the faster scan speed should decrease the difference in back-to-back experiments.

The LC–IMS–MS method combines dispersive technologies and, therefore provides a more comprehensive approach for analyzing complex mixtures. Back-to-back measurements often yield data that are nearly (>95%) identical with respect to those components that are clearly above the detection limit of the measurement. Thus, the approach appears to be well-suited for generating proteome maps with high coverage and high reproducibility -eliminating errors that are encountered from incomplete sampling by scanning-based technologies. This is especially important for comparing different states of the proteome.

Superimposing LC–QIT Data (and assignments) onto LC–IMS–MS Data. Our second aim is to show how LC–QIT information is used to aid in the assignment of peaks across the LC–IMS–MS map. At this point, many of the identifications of the LC–IMS–MS peaks were either made exclusively or

corroborated by comparison with the LC–QIT analysis. This is largely due to limitations of our in-house software (and size of the IMS datasets—the raw $t_R[m/z]$ data file for a single fraction ranges in size from ~0.8 to 20 GB). Although it is possible to identify peaks (and assign proteins) that are found from LC–IMS–(CID)–MS data directly, at this point it is more efficient to assign peaks by superimposing data (and assignments) from the LC–QIT experiments. This is done by first calibrating the LC dimensions of both datasets. Then, those LC–QIT MS and MS/MS spectra that can be assigned to a specific peptide sequence are aligned with the appropriate LC and MS regions of the LC–IMS–MS data and queried across the IMS dimension to find the corresponding spectral features in the LC–IMS–MS data. Upon finding the drift time maximum at which significant overlap exists between the assigned LC–QIT MS and MS/MS spectra and the LC–IMS–MS [and CID–MS] data, the latter data are assigned to the appropriate peptide sequence. The retention times, drift times, and precursor ion m/z values for the assigned peptide sequence (and corresponding protein) are then accumulated to create the analytical map of peptide positions in the proteome. In total ~50% of peaks that are included in the map are assigned with this approach (i.e., peptides assigned in the LC–QIT are assigned specific peak positions in the LC–IMS–MS map). In addition, the LC–IMS–MS and LC–IMS–(CID)–MS data can be directly used to assign some peaks that were not identified by LC–QIT analysis (~10% of those peptides that are listed). The remaining assignments (~40%) are made from only the LC–QIT analysis (discussed in more detail below).

Examples of Typical Datasets and Features Associated with LC–QIT Analysis. Figure 2 shows SCX and LC chromatograms as well as precursor and fragment ion mass spectra that are typical for peptides extracted from *Drosophila* heads. The base peak chromatogram corresponds to tryptic peptides that are collected from the SCX separation column from 44 to 48 min (one of the 10 fractions that were collected and analyzed). The mass spectrometer was operated in a data dependent mode, so that as peptides eluted from the LC column they were ionized, focused and accumulated in the trap and the trap was scanned to acquire a precursor ion mass spectrum. From this spectrum some precursor ions (usually several of the most abundant ions) were chosen for MS/MS analysis.

Several regions of the LC–QIT datasets illustrate both the utility of this approach to provide information that can be used to assign peptide sequences as well as the sampling issues that limit this approach. We have chosen precursor ion mass spectra from three different LC times for this discussion: $t_R = 69.6, 86.9, \text{ and } 109.8$ min. Other times yield similar results. Overall, it can be seen that most precursor ion spectra are dominated by a few intense peaks. For example, the spectrum at 69.6 min is dominated by an ion at $m/z = 478.7$ and several peaks that are less than 50% as intense are observed (e.g., at $m/z = 649.6, 668.6, \text{ and } 935.5$). The precursor ion spectra recorded at 86.9 and 109.8 are dominated by a few intense peaks ($m/z = 700.8 \text{ and } 896.3, \text{ and } m/z = 601.2, 705.8, \text{ and } 1053.2$ for these retention times, respectively). Some of these large features should be selected by the data-dependent peak-picking algorithm for subsequent MS/MS analysis. For example, the precursor ion at $t_R(m/z) = 109.8(601.2)$ is selected for MS/MS analysis. However, under the CID conditions employed the fragment ion coverage is insufficient to provide a significant score (from the MASCOT program) that makes it possible to assign the peptide. Some fragment ion assignments are shown

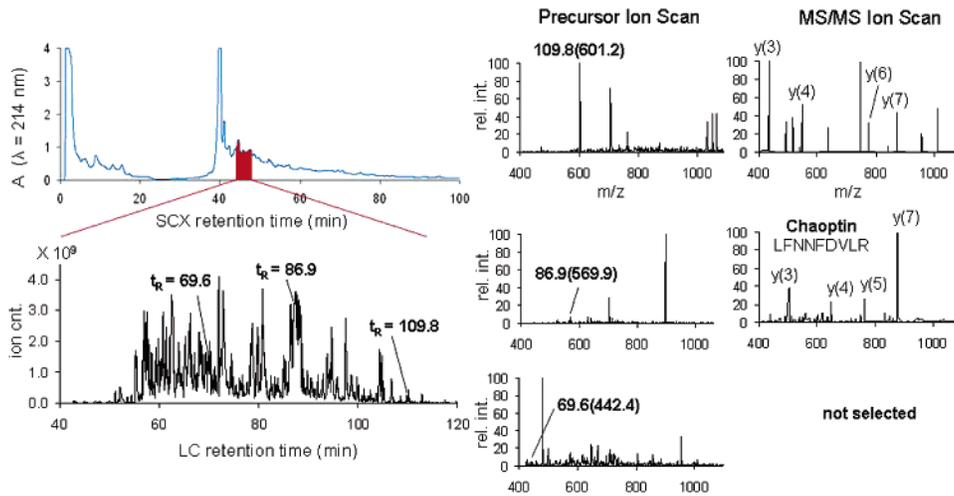


Figure 2. Example of typical SCX–LC–MS data from the ion trap experiments obtained from a population of adult heads. The upper left figure is the absorbance chromatogram from the SCX fractionation experiments, where the fraction collected between 44 and 48 min is highlighted in gray. The lower left plot shows the base peak chromatogram (BPC) obtained in LC–MS experiments where we have labeled the position of three peaks: $t_R = 69.6$, 86.9 , and 109.8 min. The right side of this figure illustrates the precursor and MS/MS ion scans for labeled BPC peaks. In the precursor ion scans we have labeled the peak positions using the $t_R(m/z)$ nomenclature. The MS/MS ion scan illustrates the MS data obtained from the selection of labeled precursor ion scan peaks; note that peak at $69.6(442.4)$ was not selected for MS/MS fragmentation. The identities of the fragment ion peaks shown for peak $109.8(601.2)$ were obtained from the LC–IMS–(CID)–MS analysis (see text).

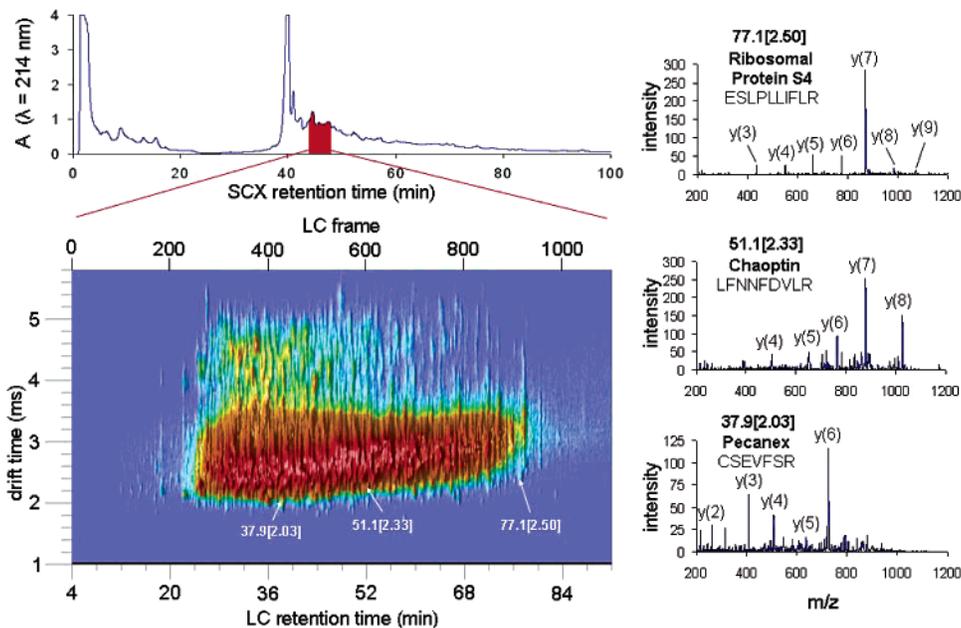


Figure 3. Example of a four-dimensional SCX–LC–IMS–(CID)–MS dataset obtained from the same SCX fraction shown in Figure 2. The lower left figures show a two-dimensional drift time versus LC frame plot for the SCX–LC–IMS–(CID)–MS dataset collected. Each spot on the two-dimensional plot contains complete MS information. The right side of the plot shows three examples of CID–MS spectra labeled using the $t_R[t_D]$ nomenclature. The CID–MS spectra correspond to the same MS/MS ion scan shown in Figure 2.

in Figure 2 [y(3), y(4), y(6), and y(7)] for this MS/MS scan; however, these assignments are from the LC–IMS–(CID)–MS dataset that is shown in Figure 3 (discussed below). This result illustrates that even for abundant precursor ions the fragment ion information that is generated may not be sufficient for identification. As noted by others,⁵⁵ we find that only ~5 to 10% of MS/MS spectra obtained with the LC–QIT lead to assignments. Recorded data may not lead to assignments because the fragmentation process did not yield spectra with sufficient information to allow for an unambiguous assignment

within the searching constraints used; or, the peptide may be modified such that the fragmentation pattern is not identified.

Figure 2 also shows data for a precursor ion at $86.9(569.9)$. This small peak has an intensity of only ~8% of the most intense ion. In this case, because there are only a few highly intense precursor ions, this relatively low abundant ion is selected for MS/MS analysis, and the spectrum provides enough information to propose the LFNFDVLR sequence, which is unique to the protein chaoptin. This type of peak illustrates another aspect of scanning approaches. There is a

significant variation associated with which peaks are selected from run to run, even for the same sample (as much as 60% of peaks picked in an initial run are not chosen in the second run of the same samples in the complex system studied here). Factors that influence this variability include shifts in retention times and changes in the relative intensities of the precursor ion (factors that are coupled). Hence, to obtain reproducible coverage of a complex proteome, multiple LC–QIT runs for the same sample need to be recorded. The number of peptides that are selected and identified with this sampling limitation will increase with the number of experiments that are carried out. This increase should approach the total number of peptides that can be detected with this approach in an exponential fashion. In many of the methods we have developed, it appears that at least 6 LC–QIT runs are required to approach 90% coverage of the peptides that could be detected with the data dependent peak picking approach employed. Thus, one sees that the use of scanning methodologies is quite inefficient for studies that aim to provide complete coverage of those peaks that are detectable.

The variability in precursor ion selection is further illustrated by a final example in Figure 2 where an observed precursor ion is not selected for MS/MS analysis. The precursor ion peak at 69.6(442.4) has a relative intensity of less than 5% of the most intense precursor ion. Because of its low signal and the presence of a number of more intense precursor ions in the same scan, e.g., 69.6(478.7), 649.6, 668.6, and 935.5), the 69.6-(442.4) ion is not selected for MS/MS analysis. The exclusion of low-intensity precursor ions is inherent to any MS/MS analysis that relies on data-dependent algorithms to select precursor ions for MS/MS analysis. The next section shows that this peak is observed (and the CID–MS data are sufficient to assign a sequence) using the dispersive LC–IMS–MS approach.

LC–IMS–MS and LC–IMS–(CID)–MS Analysis. Figure 3 shows a representation of the LC–IMS–(CID)–MS dataset obtained from the same SCX fraction shown in Figure 2. This discussion focuses on the same parts of the analysis that were highlighted in the discussion of the LC–QIT data. The two-dimensional $t_R[t_D]$ plot illustrates an aspect of the separation advantage that is gained from the LC–IMS combination. With the IMS separation many peaks that are not resolved by LC alone can often be resolved based on differences in their mobilities. The distribution of peptides that is observed extends across $t_R[t_D]$ values ranging from $\sim 20[1.7]$ to $84[5.5]$. Figure 3 also shows some examples of CID–MS information from the $t_R[t_D(m/z)]$ data. The precursor ions that produce these CID spectra correspond to the precursor ions discussed in Figure 2 [i.e., the ions at positions 109.8(601.2), 86.9(569.9), and 69.6-(442.4)].

In LC–IMS–(CID)–MS experiments, the peak at $t_R[t_D] = 77.1[2.50]$ represents a series of fragment ions positioned at $77.1[2.50(436.45, 549.50, 662.39, 775.39, 871.52, 986.07, \text{ and } 1073.22)]$. When used as input, the MASCOT search returns an identification of the ESLPLLIFLR sequence—a peptide unique to ribosomal protein S4. With this identification one sees that the $m/z = 436.45, 549.50, 662.39, 775.39, 871.52, 986.07, \text{ and } 1073.22$ values correspond to the $y(3)$ [m/z calc = 435.55], $y(4)$ [m/z calc = 548.71], $y(5)$ [m/z calc = 661.87], $y(6)$ [m/z calc = 775.03], $y(7)$ [m/z calc = 872.15], $y(8)$ [m/z calc = 985.31], and $y(9)$ [m/z calc = 1072.39] fragments, respectively. In the LC–QIT dataset that we discussed above, the information obtained from the MS and MS/MS datasets (Figure 2) was insufficient to make this assignment. Thus, this example illustrates an

assignment that was made directly from the LC–IMS–MS data and corroborated by the LC–QIT information (as is the case for $\sim 10\%$ of the identifications given in Table S-1).

Figure 3 also illustrates a peak at $51.1[2.33]$ that represents a series of fragment ions that identify the LFNNFDVLR peptide from the chaoptin protein (as observed and assigned above for the LC–QIT dataset). In this case, the LC–QIT data was used to identify this series of peaks; however, one can see by comparing Figures 2 and 3 that the CID results from the IMS data actually provides a slightly greater sequence coverage: LC–IMS–(CID)–MS results identify a series of $y(4)$ – $y(8)$ ions, while LC–QIT identifies $y(3)$, $y(4)$, $y(5)$, and $y(7)$ ions. This type of assignment (where information from LC–QIT is used to assign the LC–IMS–MS data) makes up $\sim 50\%$ of the assignments that are provided in Table S-1. We note that still $\sim 40\%$ of the assignments that are provided in Table S-1 come from LC–QIT data exclusively. This is the case when the LC–QIT spectra do not map uniquely onto a position within the LC–IMS–MS dataset.

Finally, we illustrate an example where LC–IMS–(CID)–MS method identifies a low intensity peak that was not selected for MS/MS analysis by LC–QIT method. IMS-based analysis indicates that the peak at $t_R[t_D] = 37.9[2.03]$ corresponds to series of fragment ions, $[y(2)$ – $y(6)]$ that identifies the corresponding precursor ion as $[CSEVFSR+2H]^{2+}$ —an ion that is unique to the pecanex protein.

Summary of Information that is Included (and missing) in the Tryptic Peptide Proteome Map of the Drosophila Embryo and Head Proteomes. In total, the map (Table S-1) provides information about 2457 peptides, corresponding to 1133 unique proteins. The tabulation also includes information that is useful for understanding genome expression, including: the FlyBase gene number for protein identification; the state (embryo, head, or both) in which corresponding mRNAs are detected; the state (embryo, head, or both) in which the protein is identified in our experiments; the GO cellular component in which the protein is assigned; and, the sequence of all of the peptides that have been mapped across the $t_{SCX}\{t_R[t_D(m/z)]\}$ analytical space.

At this stage, some information for specific peptides is incomplete. For example, drift times are not assigned to all peptides. In Table S-1, 1438 of 2457 (60%) of peptides included in the map have reported drift times. That is, 40% of identified peaks are represented based on information about their $t_{SCX}\{t_R(m/z)\}$ and no values for drift times are given. In many cases the inability to define a drift time comes about because we have not successfully mapped information from the LC–QIT analysis onto the LC–IMS datasets; drift times are provided only when the ion can be clearly mapped and we have taken a very conservative approach for the first draft of this map for the present system.

In other cases, we find no evidence for the peaks identified by LC–QIT analysis in the LC–IMS–MS dataset within the region of the map in which these ions are expected. The inability to find these features may come about for one of several factors (or a combination thereof). The size of the $t_{SCX}\{t_R[t_D(m/z)]\}$ dataset has led us to impose an intensity cutoff prior to analysis and it is possible that some features are present but fall below the imposed cutoff. This may be especially true of low intensity ions; upon dispersing these ions across another mobility dimension, intensities may fall below the critical level that allows us to find these features. Some of the peaks that are not mapped in the LC–IMS–MS datasets correspond to

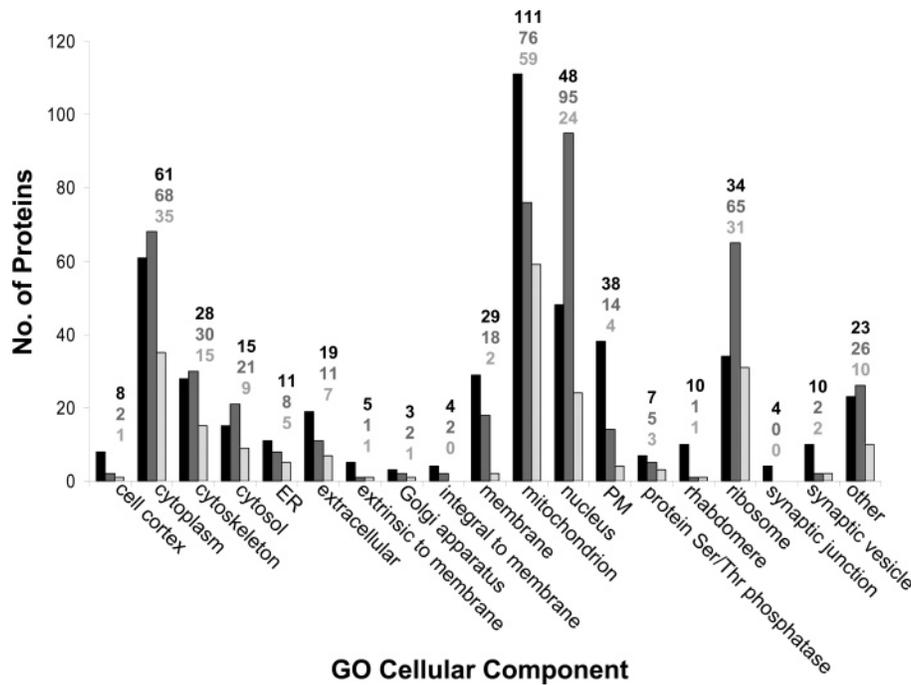


Figure 4. Bar graph representation of proteins identified as a function of their GO cellular components. Adult head proteins (solid black bars), whole embryo proteins (solid gray bars), and proteins found in common (light gray checkered bars) are illustrated; the number of the respective proteins in each component is also provided directly above the bars. The other category includes proteins that have known GO cellular components that are not a subset of the 18 cellular components listed. To make the graph more readable, we have omitted the not specified category. The abbreviations PM and ER refer to plasma membrane and endoplasmic reticulum, respectively.

intense peaks in the LC–QIT analysis. This observation suggests that other instrumental factors may be influencing the comparisons that we have made. For example, the peak may fall outside of the 7% reproducibility that we expect for most retention times. Or the peptide may exist as a different charge state or perhaps not produce abundant ion signals (i.e., differences in the ionization process). The LC–QIT instrument employs an ESI source that uses a heated capillary (operated at 150 °C in these experiments); in our current home-built systems no heated capillary is used. It is possible that the LC–IMS–MS experiments are simply not sensitive to those ions that are not readily desolvated at thermal energies because they exist across a range of hydration (or ion–solvent cluster) states.⁵⁶

Examples of Biological Results from the Map. Inspection of the data that are presented in Table 1 and (S-1) provides some interesting clues about the relationship of the mRNAs and proteins detected. Some proteins identified here do not have representative mRNA–cDNA clones. For example, as indicated in Table 1, the gene product FBgn0000055 was not recovered at the transcript level in either state, but six peptides are identified for the protein associated with this gene; and, this protein appears in both states of the organism. Thus, one immediate result is that the proteome cannot necessarily be predicted from the recovery of cDNA clones—even when transcripts and proteins are recorded from the same proteome states. This is most likely due to sampling limitations in cDNA and protein analysis. For other gene products, a protein is detected in a state different from the state in which the cDNA is recovered; or, a protein is detected in both states while the mRNA–cDNA clone is only detected in either the adult head or embryo. Only 432 out of 1133 (38%) of proteins and mRNAs contained in Table S-1 are found in identical states. To obtain

more insight it is useful to pursue additional information that is known about *Drosophila*. As mentioned in the Introduction, one of the powerful advantages of working with model systems is that substantial insight can be gained by considering information from the GO database. The proteome map that is tabulated in Table S-1 also includes information about the GO cellular component. This allows us to ask questions about what types of activities are carried out by different cell types (in this case those associated with the embryo and head).

Discussion

Overview of Proteome, Transcriptome and Genome Representations. Below we present Figures 4–7 which depict several different types of comparisons between the heads and embryos. Figures 4 and 5 show bar graph representations of the proteins and mRNAs, respectively, for different cellular components. We find this representation useful in visualizing the number of different proteins and transcripts that have been detected between these states. Figures 6 and 7 put these numbers into context with the entire genome by using a Venn diagram representation. In this case, the entire genome size (and sizes of individual GO genome components) can be compared directly to what has been detected in the transcriptome and proteome; and, the overlap between component transcriptomes and proteomes is represented. Depending on what type of comparison is made, it is often useful to compare results using several figures (as discussed in more detail below).

Expression of Proteins in Heads and Embryos: General Considerations. To begin understanding the similarities and differences at the protein level for cells associated with the head and embryo it is useful to classify the proteins according to their GO cellular component profiles available from FlyBase.²²

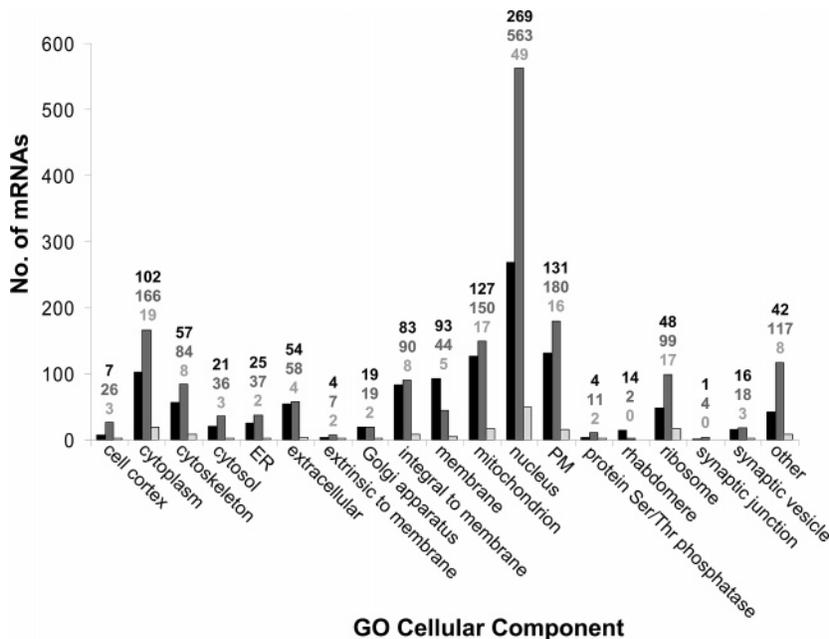


Figure 5. Bar graph representation of mRNA transcripts as a function of their GO cellular component. Adult head mRNAs (solid black bars), whole embryo mRNAs (solid gray bars), and mRNAs found in common (light gray checked bars) are illustrated; the number of the respective mRNA in each component is also provided directly above the bars. The other category includes mRNAs that have known GO cellular components that are not a subset of the 18 cellular components listed. To make the graph more readable, we have omitted the not specified category. The abbreviations PM and ER refer to plasma membrane and endoplasmic reticulum, respectively.

The data presented above provides evidence for 780 proteins associated with the head. Of these, 385 have been associated with specific cellular components. There is evidence for 660 proteins in the embryo, of which 383 have specified cellular components. In all, proteins detected in both the heads and embryos can be classified into over 150 different cellular components. The large number of different locations comes about because GO cellular components are divided into a hierarchy of categories. For example, there are 10 components that make up the more general mitochondrion component. To present a manageable discussion of our results in the context of the GO cellular components we have limited the number of components that are considered to 18 (in this case, these are the more general and highly populated cellular components in the adult head). Due to the nature of the classifications, it is also possible for proteins to be classified into multiple components. In a few cases, an individual protein is associated with more than one component; in the plots below, we represent it in all components that are specified. For example, a protein associated with the mitochondrial ribosome is classified under both mitochondrion and ribosome.

Figure 4 summarizes the number of different proteins that are associated with the following 18 cellular components: cell cortex; cytoplasm; cytoskeleton; cytosol; endoplasmic reticulum (ER); extracellular; extrinsic to membrane; Golgi apparatus; integral to membrane; membrane; mitochondrion; nucleus; plasma membrane (PM); protein Ser/Thr phosphatase; rhabdomere; synaptic junction, and synaptic vesicle. These 18 categories capture 758 of the 780 total proteins for the head and 634 of the 660 proteins associated with the embryo. The majority of proteins (in both *Drosophila* states) are associated with only a few cellular components: the mitochondrion (111 proteins from the head and 76 from the embryo, of which 59 are in common); the nucleus (48 proteins from the head and 95 from the embryo, of which 24 are in common); the

cytoplasm (61 from the head and 68 from the embryo, of which 35 are in common) and the ribosome (34 from the head and 65 from the embryo, with 31 in common to both). If we neglect those components with very little protein representation (e.g., those with fewer than 10 proteins from a state within a specific component), then we find that the overlap between individual components for the head and embryo states varies from a low value of 2 out of 29 (7%) for membrane proteins to as high as 31 of 34 (91%) for the ribosome. Additionally some components are more fully represented in the head (e.g., rhabdomere, mitochondrion, membrane, and plasma membrane) while others are more fully represented by the embryo (e.g., nucleus and the ribosome proteins). All of this indicates that there is a substantial change in the expression of proteins in cells associated with these two states.

Characterizing (and rationalizing) the Populations of Proteins Associated with Specific GO Cellular Components.

Further consideration of Figure 4 (and Figures 6 and 7) gives an idea about similarities and differences in the protein makeup of cellular components that are found in cells of the embryos and heads. Although the number of proteins in each state may potentially be used as a quantitative measure of each component, such quantitative inferences should be done cautiously (if at all). For example, the result that more mitochondrial proteins are detected in the head than in the embryo (111 in total for the head, compared with only 76 in the embryo) seems to indicate that proteins associated with this GO component are more abundant within cells of the head. This interpretation is consistent with a previous report that mitochondrial densities are higher in neurons than in other types of cells.⁵⁷ Thus, one could interpret our results to be in agreement with this report.

However, the following caveat is important to consider. Strictly, it is impossible to infer the abundance of individual proteins within the cell from a measure of the number of proteins that are detected from our analysis. That is, another

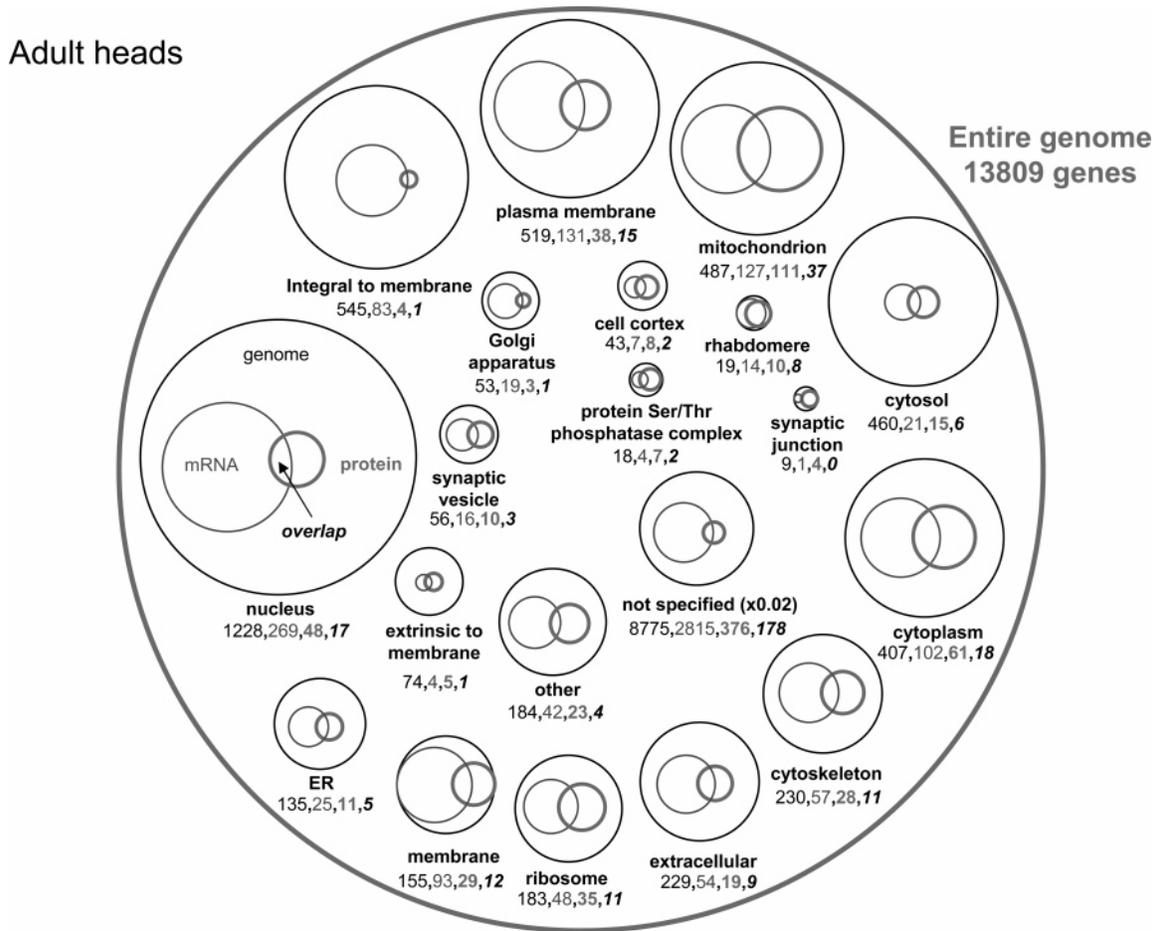


Figure 6. Venn diagram representations of genome, transcriptome, and proteome data for the *Drosophila* adult head state. The entire genome of 13809 genes is represented by the area of the large gray circle. Within this circle we represent predicted genes from the genome (thin black circles and black numbers), mRNA transcripts from cDNA libraries (thin gray circles and gray numbers), and proteins determined from our proteomics analysis (thick gray circles and bold gray numbers) that are associated with specified GO cellular components. The *not specified* category corresponds to gene products that do not have a specified cellular component. The areas of the circle correspond to the number of genes, transcripts, or proteins observed. We have also noted the number of proteins that overlap with the mRNA transcripts (bold black italic numbers). For example, in the nucleus genomic data indicates that there are 1228 genes (whose products are associated with the nucleus), and cDNA libraries indicate that 269 mRNA transcripts are present. Our proteome analysis identifies 48 proteins, of which 17 overlap with the transcriptome. The location endoplasmic reticulum is abbreviated as ER.

consistent interpretation is that although there are fewer proteins in the embryo associated with the mitochondria (compared with the head) they may be more abundant. In this interpretation we would state that the proteome associated with the mitochondria for the head and the embryo appears to change. Although we prefer the former explanation, we cannot rule the latter out.

Assuming that the number of detectable proteins in a component reflects the abundances of these components, Figure 4 (and Figures 6 and 7) shows that there are more nuclear and ribosomal proteins detected from analysis of the embryo than the head. This result suggests that these GO components are more abundant in the embryo and this seems rational since the level of activity and change associated with insect embryogenesis should be much higher in this state. During embryogenesis embryos undergo rapid mitosis, germ layer formation and extensive cellular differentiation -processes that should involve extensive protein synthesis and genome regulation (presumably requiring substantial nuclear and ribosomal protein machinery). We note also that the large numbers of nuclear and ribosomal proteins found in embryos

compared with heads is consistent with results from DNA microarray studies (Figures 5, 6, and 7); genes associated with transcription factors and protein synthesis appear to be highly expressed as mRNAs and proteins in embryos relative to adults (Figures 5, 6, and 7).¹¹

In several other cases the differences that are measured between protein expression in the head and embryo can be understood in perhaps the simplest of terms. One such case involves the proteins associated with the rhabdomere, a structure that is found in the eye.⁵⁸ A question that can be asked is as follows: when are transcripts and proteins associated with eye tissues synthesized? In this case, we use the GO rhabdomere component (and the data in Figures 4–7) to begin to address this question. There are 19 total genes associated with this component. Of these, we have detected 10 proteins as listed in Table 2 that are associated with the rhabdomere; 16 mRNAs have been detected. While all 10 of the detected proteins are found in the head, only one, calmodulin, is found in the embryo; and, the presence of this protein in the embryo can be rationalized, because calmodulin is also associated with the cytoplasm and is involved in several cellular

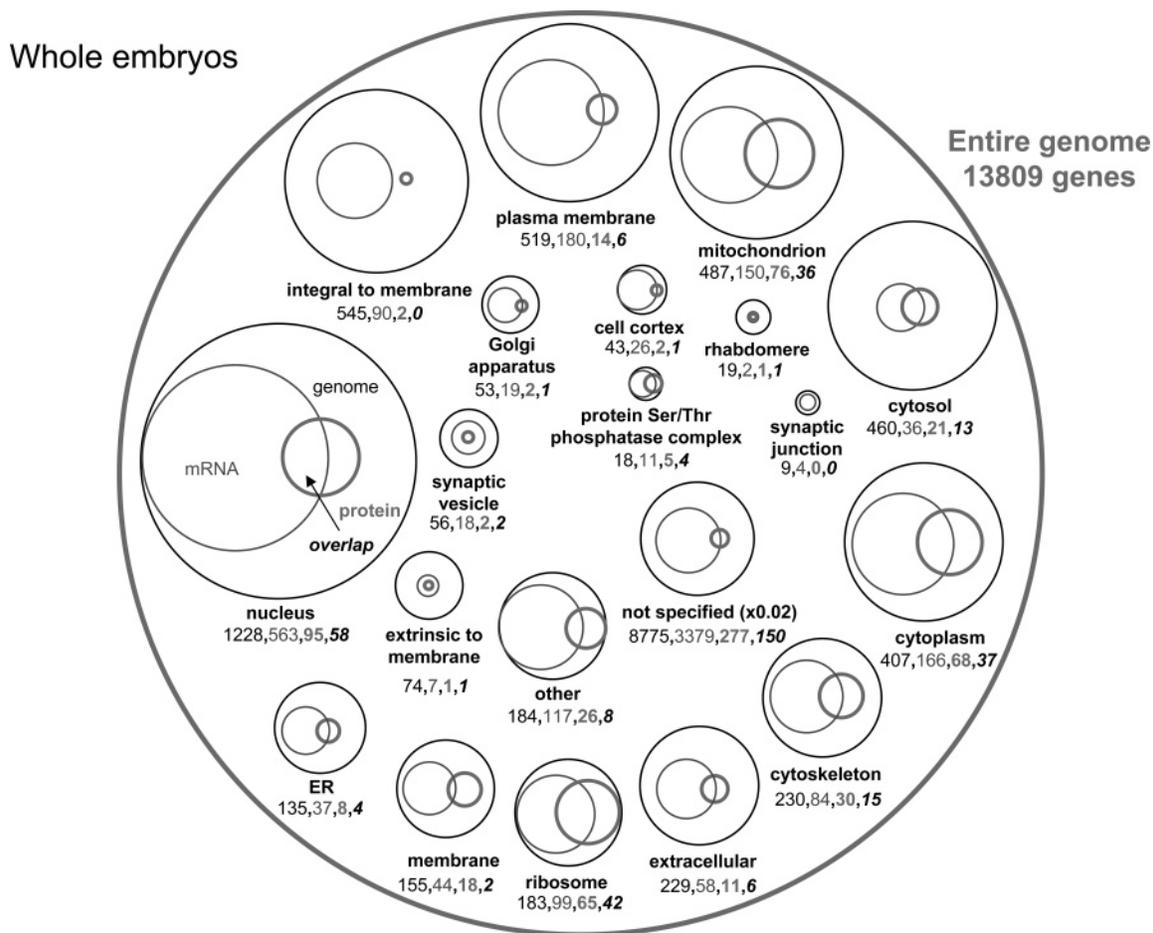


Figure 7. Venn diagram representations of genome, transcriptome, and proteome data for the *Drosophila* whole embryo state. The entire genome of 13809 genes is represented by the area of the large gray circle. Within this circle we represent predicted genes from the genome (thin black circles and black numbers), mRNA transcripts from cDNA libraries (thin gray circles and gray numbers), and proteins determined from our proteomics analysis (thick gray circles and bold gray numbers) that are associated with specified GO cellular components. The *not specified* category corresponds to gene products that do not have a specified cellular component. The areas of the circle correspond to the number of genes, transcripts, or proteins observed. We have also noted the number of proteins that overlap with the mRNA transcripts (bold black italics numbers). The location endoplasmic reticulum is abbreviated as ER.

pathways.⁵⁹ At the mRNA level, 16 transcripts associated with the rhabdomere have been detected (14 in the head and 2 in the embryo). It appears that because an embryo has not developed a differentiated eye, it expresses no proteins and few mRNAs associated with the rhabdomere. While this result may or may not have been obvious, it is satisfying as it appears that some structures across different states will be useful as internal controls. As we develop more robust means of quantifying proteins it should be possible to characterize proteins in the rhabdomere in significant detail. For example, studies as a function of development, or aging, or genetic mutations could be carried out.

Similar to the results for the rhabdomere Figures 4, 6, and 7 show that synaptic proteins are also more prevalent in adult heads; four proteins associated with synaptic junctions (the junction between neurons and the site for interneuronal communication) are identified from the heads, whereas none is found in the embryo. Ten synaptic vesicle proteins are detected in the head while only two are present in the embryo. The synaptic vesicle is an organelle secreted between neurons, and it is reasonable that this organelle is found in higher densities in the head than in the embryo. It is worthwhile to make several additional comments about these results (Figures

4, 6, and 7). Few integral membrane proteins are observed in both states. At least in part, this is because of the extraction procedure that is employed. Other procedures that incorporate the use of detergents should make it possible to sample more of these proteins.⁶⁰

Comparison of mRNAs Detected in Adult Heads and Whole Embryos. It is also interesting to compare the transcriptomes (mRNAs detected thru cDNA analysis) between the adult head and embryo (i.e., the existing data from other studies that can be extracted from the Berkeley *Drosophila* genome database).⁵⁴ Application of cDNA techniques have led to the detection of 3775 and 4860 mRNA transcripts in adult heads and whole embryos, respectively. The larger mRNA coverage in the embryo compared with the head is the opposite of what is seen for the proteins (as noted above, 780 proteins were found in the head while only 660 were found in the embryo). Only 565 mRNAs are found in both states; this number corresponds to a 15% (relative to the head) overlap. The percent overlap of mRNAs is significantly smaller than that observed in the proteome, where 39% of proteins (relative to the head) overlap.

Transcripts can also be classified according to their GO cellular component (as was done above for the proteins). Of the 3775 mRNAs that are found in cells associated with the

head, 950 transcripts have specified cellular components; for embryo cells, 1481 transcripts have specific cellular components. Figure 5 shows the number of mRNAs that are associated with the 18 cellular components for cells from the head and embryo. In the case of the transcripts, these 18 categories capture 908 out of the 3775 of the mRNAs for the head and 1344 of the 4864 mRNAs associated with the embryo. If we neglect components containing few mRNAs (e.g., those with fewer than 10 mRNAs from a state within a specific component) we find that the overlap between individual components for the head and embryo states varies from a low value of 5 out of 93 (~6%) for membrane proteins to as high as 17 out of 48 (35%) for the ribosome—less than the fractional overlap of the number of proteins associated with these components.

As was done above for proteins it is helpful to consider how many mRNAs are observed within the different cellular components of the embryo and head. In the case of mRNAs, the component with largest number of detected mRNAs is the nucleus, where 269 mRNAs are detected in the head, and 563 are found in the embryo. Only a remarkably small number of these are in common between these states (49 total). This result suggests that while the nuclear mRNAs are present in both states, their expressions have changed. The observation that there is little overlap in the mRNAs that are detected for the embryo and head appears to be a general trend for nearly all of the components (at least those that are present in significant numbers to allow comparisons). Of those mRNAs that have been assigned to cellular components (960 for the head and 1481 for the embryo) only 170 overlap. This is much lower than the overlap associated with what is detected at the protein level for these components. In the case of proteins with specified cellular components (404 in the head and 382 in the embryo), the overlap is 200 (a substantially larger fraction). Overall, it appears that across these two states there is a greater representation of the transcriptome than the proteome. It is also interesting that the fractional overlap between transcripts that are expressed in the embryo and head states is less than the proteome overlap between these states.

Comparison of the Genome, Transcriptome, and Proteome in Adult Heads and Whole Embryos. Figures 6 and 7 show Venn diagram representations of the genome, transcriptome, and proteome data for the adult head and whole embryo, respectively. From these figures, one sees that the overall numbers of components that are detected as transcripts or proteins in specific GO components vary substantially depending on the component and the organismal state—raising some interesting questions. For example: why is there an overall disparity between mRNA and protein expression? One may expect that an abundant protein may also be highly expressed at the transcriptome level, and this appears to be the case in prokaryotic systems.⁶¹ In eukaryotic systems, the situation appears to be more complex. For example, measurements of mRNA and protein abundances in *Saccharomyces cerevisiae* have shown 20- to 30-fold differences between mRNA and protein levels.⁶² Clearly, we are only at the beginning of understanding how the genome is expressed and much work remains to be done.

Summary and Conclusions

A new approach that integrates LC and MS techniques with a gas-phase separation based on IMS has created a multidimensional analytical map of peptides from proteins for two states of the *Drosophila* proteome: the embryo and the adult

head. This approach can be modulated between LC–IMS–MS and LC–IMS–(CID)–MS modes. With this approach precursor ions are not isolated for fragmentation; thus, at least in concept no ions are discriminated against during the time that one ion is selected for MS/MS analysis (as is the case in the commercial LC–QIT) approach. The present paper describes the first use of this approach for a comparative proteomics study.

Although a significant amount of work has been done to get these techniques to the stage that they can be used for comparative proteomics, the progress that is described represents only a first step in the inclusion of IMS technologies for use in proteomics platforms. In particular, we are at the earliest stages of interpreting the large data arrays that are generated. The present approach utilizes a calibration to overlay peptides that were identified by commercial LC–QIT techniques onto the LC–IMS–MS datasets. With this method it is possible to unambiguously identify and compare datasets; however, only ~50% of identified LC–QIT peaks can be uniquely mapped to specific peaks in the SCX–LC–IMS–MS dataset.

In the map provided in Table S-1 ~40% of peptides identified by LC–QIT analysis are not unambiguously assigned in the IMS dimension. This may be rationalized by one of several explanations (or a combination thereof). As discussed above, in some cases it appears that ions having the correct $t_R(m/z)$ values are present in the IMS data; however, they do not have a unique or in some cases well-defined mobility. This highlights a weakness of a dispersive approach in which ions are not selected by their m/z ratios for CID analysis. Although several high-resolution IMS instruments have been developed in the past few years (having resolving powers in excess of 200 in some cases),⁶³ the one used in the present studies has only a limited resolving power (~17 to 35); thus many types of ions remain unresolved even after the combined SCX–LC–IMS separation. Due to this lower resolution, a given CID spectra may contain fragment ions from two or more parent ions. We have chosen to begin mapping proteomes using a low-resolution drift tube because of the large signals associated with this approach. Current efforts are underway to incorporate a higher-resolution gas-phase separation; this should allow substantially more peaks to be unambiguously mapped.

To identify peaks from the LC–IMS–MS and LC–IMS–(CID)–MS datasets, we have used peak-picking algorithms that were developed in house to determine the positions of peaks within multidimensional datasets. This approach generates MS and CID–MS data that is resolved in LC and IMS dimensions. The MS and CID–MS positions that are obtained can be used in combination with database searching techniques to identify some peptides and proteins that were not identified based on the LC–QIT analysis. In some cases, the LC–IMS–MS approach offers significant advantages. For example, Figures 2 and 3 show one case where a small peak that was observed in both the LC–QIT and LC–IMS–MS datasets was identified based on the CID–MS analysis (using the IMS approach) but was not identified during the LC–QIT analysis (in this case, because it was not selected for MS/MS analysis). In the LC–IMS–(CID)–MS dataset that was shown (Figure 3) this approach allowed 54 (16%) additional peptides that were not identified in a single LC–QIT experiment to be identified. Across the map, ~10% of ions are identified only by LC–IMS–MS analysis. It appears likely that upon further refinement and development of the LC–IMS–MS techniques that this type of advantage will complement scanning techniques (eventually offering a substantially greater coverage of the proteome).

One other feature of the IMS dimension that is likely to be useful comes about because mobilities are a measure of the cross section-to-charge ratio of the ions. As shown previously, cross section and mass are not entirely independent parameters.^{42–44,64} That is, ions sizes and masses are correlated; however, the measured mobility still provides an additional means of characterizing the shapes of peptides; and, because different sequences may have identical m/z values but different shapes characterizing the mobility should help refine assignments. Over the last several years a number of methods for predicting the mobilities for different peptide sequences and charge states have been developed. The most rigorous method for determining a mobility involves the use of molecular modeling approaches combined with cross section calculation algorithms.⁶⁵ We (and others) have developed size parameters for individual amino acids that can be rapidly combined to roughly predict cross sections.^{64,66,67} Of course, once the mobility of an ion has been measured this value can be used to predict the drift time of the ion in any other system. We are currently working to incorporate a mobility parameter that will complement the scores obtained from MS/MS database searches.⁶⁸ In this case, two sequences which give identical significance scores (e.g., from a MASCOT search) may be further delineated based on differences in their cross sections (thus, reducing the number of false-positive assignments).

The discussions provided above have attempted to put these early experimental findings into a biological context. In these studies 1133 unique proteins were characterized in two states of *Drosophila*: the adult head and the embryo. Of these, 307 are observed in both states, indicating that there are significant changes in the proteome. Further investigation reveals that the number of proteins within a given GO cellular component can vary substantially between states. For example, nuclear and ribosomal proteins (and mRNAs) are more numerous in the embryonic state, which is reasonable given the flux of protein synthesis during early development. In contrast, proteins belonging to cellular components associated with visual and neuronal pathways are more numerous in the adult head. We also observed an increase in the number of mitochondrial proteins in adult heads; this was rationalized by noting that mitochondria are found in higher densities in neurons. Similar to other eukaryotic organisms,^{69–71} a comparison of mRNAs detected in cDNA libraries to proteins observed in these studies suggests a low correlation between the detected transcriptome and measured proteome. Overall, the overlap between detected proteins and detected mRNAs varied from 47 to 57% between the embryo and head states. A global comparison to the genome further reveals that overall only a small fraction (30 to 37%) of the predicted genome is sampled with current transcriptome and proteome technologies.

In closing, we are currently working on including other states of *Drosophila* as a part of this analytical map. We have recently obtained preliminary data for the earliest (0 to 2 h) and latest (20 to 22 h) stages of embryogenesis. This project will utilize the map to study developmental factors related to the proteome. We have also recorded data for the adult head state as a function of organism age. These studies are aimed at understanding neurological changes that occur as the brain ages. These types of systems, which are intractable in humans, are possible in model organisms and may provide clues about general biological mechanisms associated with development.

Acknowledgment. The authors acknowledge partial funding of this work from the National Institutes of Health (1R01GM-59145-03), the National Science Foundation (CHE-0078737) and from the Indiana Genomics initiative (INGEN). We are grateful for many helpful technical discussions with our colleague Dr. Xinfeng (Frank) Gao (from the molecular visualization facility). This work also benefited from many discussions with Drs. Steven Naylor and Eric Neuman (from Beyond Genomics Inc., Boston, MA) and Professor Fred Regneir (Purdue University).

Supporting Information Available: The complete list of the positions of peaks for specific peptides and proteins that have been identified (as discussed below) as well as relevant genomic information (Table S-1). This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Hemming, B. S. *Insect Development and Evolution*; Cornell University Press: Ithaca, New York, 2003.
- (2) Helfand, S. L.; Rogina, B. *BioEssays* **2003**, *25*, 134–141.
- (3) The roles of steroid and peptide hormones in *Drosophila* development have been previously discussed. See for example: Riddiford, L. M. *Adv. Insect Physiol.* **1993**, *24*, 213–274. Henrich, V. C.; Rybczynski, R.; Gilbert, L. I. *Vitamins and Hormones* **1999**, *55*, 73–125. Riddiford, L. M.; Cherbas, P.; Truman, J. W. *Vitam. Horm.* **2001**, *60*, 1–73.
- (4) Zhu, S.; Chiang, A.-S.; Lee, T. *Development* **2003**, *130*, 2603–2610.
- (5) Roberson, K.; Mergliano, J.; Minden, J. S. *Dev. Biol.* **2003**, *260*, 124–127.
- (6) Kumar, J. P. *Nat. Gen. Rev.* **2001**, *2*, 846–857.
- (7) Lee, S. B.; Cho, K. S.; Kim, E.; Chung, J. *Development* **2003**, *130*, 4001–4010.
- (8) Estrada, B.; Casares, F.; Sánchez-Herrero, E. *Differentiation* **2003**, *71*, 299–310.
- (9) Aravin, A. A.; Lagos-Quntana, M.; Yalcin, A.; Zavolan, M.; Marks, D.; Snyder, B.; Gaasterland, T.; Meyer, J.; Tuschli, T. *Dev. Cell* **2003**, *5*, 337–350.
- (10) Lee, L. A.; Orr-Weaver, T. L. *Annu. Rev. Genet.* **2003**, *37*, 545–578.
- (11) Arbeitman, M. N.; Furlong, E. E. M.; Imam, F.; Johnson, E.; Null, B. H.; Baker, B. S.; Kransnow, M. A.; Scott, M. P.; Davis, R. W.; White, K. P. *Science* **2002**, *297*, 2270–2275.
- (12) Several groups have examined the proteins in *Drosophila*. See for example: Prokopenko, S. N.; He, Y.; Lu, Y.; Bellen, H. J. *Genetics* **2000**, *156*, 1691–1715. Shieh, B.; Parker, L.; Popescu, D. *J. Biochem.* **2002**, *156*, 523–527. Ou, C.; Pi, H.; Chien, C. *Trends in Genetics* **2003**, *19*, 382–389.
- (13) Vierstraete, E.; Cerstiaens, A.; Baggerman, B.; Van den Bergh, G.; De Loof, A.; Schoofs, L. *Biochem. Biophys. Res. Comm.* **2003**, *304*, 831–838.
- (14) Busby, S. A.; Steele, H. A.; Karr, T. L.; Shabanowitz, J.; Hunt, D. F. *Proc. 51st ASMS Conference on Mass Spectrometry and Allied Topics*, Montreal, Quebec, CA (June 8–12, **2003**).
- (15) Krijgsfeld, J.; Ketting, R. F.; Mahmoudi, T.; Johansen, J.; Artal-Sanz, M.; Verrijzer, C. P.; Plasterk, R. H. A.; Heck, A. J. R. *Nat. Biotechnol.* **2003**, *21*, 927–931.
- (16) Taraszka, J. A.; Gao, X.; Valentine, S. J.; Koeniger, S. L.; Miller, D. E.; Kaufman, T. C.; Clemmer, D. E. *J. Proteome Res.* **2005**, *4*, 1238–1247.
- (17) Baggerman, G.; Cerstiaens, A.; De Loof, A.; Schoofs, L. *J. Biol. Chem.* **2002**, *277*, 40368–40374.
- (18) Johnson, E. C.; Bohn, L. M.; Barak, L. S.; Birse, R. T.; Nässel, D. R.; Caron, M. G.; Taghert, P. H. *J. Biol. Chem.* **2003**, *278*, 52172–52178.
- (19) Verleyen, P.; Baggerman, G.; Wiehart, U.; Schoeters, E.; Van Lommel, A.; De Loof, A.; Schoofs, L. *J. Neurochem.* **2004**, *88*, 311–319.
- (20) Uttenweiler-Joseph, S.; Moniatte, M.; Lagueus, M.; Van Dorsse-laer, A.; Hoffmann, J. A.; Bulet, P. *Proc. Nat'l. Acad. Sci. U.S.A.* **1998**, *95*, 11342–11347.
- (21) Giot, L.; Bader, J. S.; Brouwer, C.; Chaudhuri, A.; Kuang, B.; Li, Y.; Hao, Y. L.; Ooi, C. E.; Godwin, B.; Vitols, E.; Vijayadamar, G.; Pochart, P.; Machineni, H.; Welsh, M.; Kong, Y.; Zerhusen, G.; Malcolm, R.; Varrone, Z.; Collis, A.; Minto, A.; Burgess, S.; McDaniel, L.; Stimpson, E.; Spriggs, F.; Williams, J.; Neurath, K.

- Ioime, N.; Agee, M.; Voss, E.; Furtak, K. Renzulli, R.; Aanensen, N.; Carrolla, S.; Bickelhaupt, E.; Lazovatsky, Y.; DaSilva, A.; Zhong, J.; Stanyon, C. A.; Finley, R. L.; White, K. P.; Braveman, M.; Jarvie, T.; Gold, S.; Leach, M.; Kinght, J.; Shimkets, R. A.; McKenna, M. P.; Chant, J.; Rothberg, J. M. *Science* **2003**, *302*, 1727–1736.
- (22) The FlyBase Consortium *Nucl. Acids Res.* **2003**, *31*, 172–175. <http://flybase.org/>.
- (23) Adams, M. D.; et al. *Science* **2000**, *287*, 2185–2195.
- (24) The Gene Ontology Consortium *Nat. Genet.* **2000**, *25*, 25–29.
- (25) Several reviews have addressed proteomic technologies. See for example: Rabilloud, T. *Proteomics* **2002**, *2*, 3–10. Templin, M. F.; Stoll, D.; Schwenk, J. M.; Pötz, O.; Kramer, S.; Joos, T. O. *Proteomics* **2003**, *3*, 2155–2166. Lion, N.; Rohner, T. C.; Dayon, L.; Arnaud, I. L.; Damoc, E.; Youhnovski, N.; Wu, Z.; Roussel, C.; Jossennad, J.; Jensen, H.; Rossier, J. S.; Przybylski, M.; Girault, H. H. *Electrophor.* **2003**, *24*, 3533–3562.
- (26) Several recent reviews have discussed the advances in mass spectrometry based proteomics. See for example: Smith, R. D. *Trends Biotech.* **2002**, *20*, S3–S7. Wu, C. C.; Yates, J. R., III *Nat. Biotech.* **2003**, *21*, 262–267. Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198–207. Standing, R. *Curr. Opin. Struct. Biol.* **2003**, *13*, 595–601.
- (27) Ficarro, S. B.; McClelland, M. L.; Stukenberg, P. T.; Burke, D. J.; Ross, M. M.; Shabanowitz, J.; Hunt, D. F.; White, F. M. *Nat. Biotechnol.* **2002**, *20*, 301–305.
- (28) Zhang, H.; Li, X. J.; Martin, D. B.; Aebersold, R. *Nat. Biotechnol.* **2003**, *21*, 660–666.
- (29) Mann, M.; Jensen, O. N. *Nat. Biotechnol.* **2003**, *21*, 255–261.
- (30) Posewitz, M. C.; Tempst, P. *Anal. Chem.* **1999**, *71*, 2883–2892.
- (31) Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. *Nat. Biotechnol.* **1999**, *17*, 994–999.
- (32) Geng, M.; Ji, J.; Regnier, F. E. *J. Chromatogr. A* **2000**, *870*, 295–313.
- (33) Yao, X.; Freas, A.; Ramirez, J.; Demirev, P. A.; Fenselau, C. *Anal. Chem.* **2001**, *73*, 2836–2842.
- (34) Gerber, S. A.; Rush, J.; Stemman, O.; Kirschner, M. W.; Gygi, S. P. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 6940–6945.
- (35) Regnier, F. E.; Riggs, L.; Zhang, R.; Xiong, L.; Liu, P.; Chakraborty, A.; Seeley, E.; Sioma, C.; Thompson, R. A. *J. Mass Spectrom.* **2002**, *37*, 133–145.
- (36) Brunet, S.; Thibault, P.; Gagnon, E.; Kearney, P.; Bergeron, J. J. M.; Desjardins, M. *Trends Cell Biol.* **2003**, *13*, 629–638.
- (37) Srebalus, C. A.; Hilderbrand, A. E.; Valentine, S. J.; Clemmer, D. E. *Anal. Chem.* **2002**, *74*, 26–36.
- (38) Hilderbrand, A. E.; Myung, S.; Srebalus-Barnes, C. A.; Clemmer, D. E. *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 1424–1436.
- (39) Moon, M. H.; Myung, S.; Plasencia, S.; Hilderbrand, A. E.; Clemmer, D. E. *J. Proteome. Res.* **2003**, *2*, 589–597.
- (40) Myung, S.; Lee, Y. J.; Moon, M. H.; Taraszka, J.; Sowell, R.; Koeniger, S.; Hilderbrand, A. E.; Valentine, S. J.; Cherbas, L.; Cherbas, P.; Kaufmann, T. C.; Miller, D. F.; Mechref, Y.; Novontny, M. V.; Ewing, M. A.; Sporleder, C. R.; Clemmer, D. E. *Anal. Chem.* **2003**, *75*, 5137–5145.
- (41) Wild-type Oregon-R *Drosophila* were obtained from the *Drosophila* Stock Center at Indiana University, Bloomington.
- (42) Several groups have applied ion mobility techniques to separations. See for example: Hagen, D. F. *Anal. Chem.* **1979**, *51*, 870–874. Leasure, C. S.; Eiceman, G. A. *Anal. Chem.* **1985**, *57*, 1890–1894. Lee, D. S.; Wu, C.; Hill, H. H., Jr. *J. Chromatogr. A* **1998**, *822*, 1–9. Asbury, G. A.; Wu, C.; Siems, W. F.; Hill, H. H., Jr. *Anal. Chim. Acta* **2000**, *404*, 273–283. Matz, L. M.; Hill, H. H., Jr. *Anal. Chem.* **2001**, *73*, 1664–1669. Bluhm, B. K.; North, S. W.; Russell, D. H. *J. Chem. Phys.* **2001**, *114*, 1709–1715. Ruotolo, B. T.; Gillig, K. J.; Stone, E. G.; Russell, D. H.; Fuhrer, K.; Gonin, M.; Schultz, J. A. *Int. J. Mass Spectrom. Ion Processes* **2001**, *219*, 253–267.
- (43) Ion mobility spectrometry techniques have been used extensively to probe gas-phase structures. See for example: Hunter, J.; Fye, J.; Jarrold, M. F. *Science* **1993**, *260*, 784–786. Bowers, M. T.; Kemper, P. R.; vonHelden, G.; Bowers, M. T. *Science* **1993**, *260*, 1446–1451. Hunter, J. M.; Jarrold, M. F. *J. Am. Chem. Soc.* **1995**, *117*, 10317–10324. Wyttenbach, T.; von Helden, G.; Bowers, M. T. *J. Am. Chem. Soc.* **1996**, *118*, 8355–8364. Shelimov, K. B.; Jarrold, M. F. *J. Am. Chem. Soc.* **1997**, *119*, 2987–2994. Lee, S.; Wyttenbach, T.; Bowers, M. T. *Int. J. Mass Spectrom. Ion Proc.* **1997**, *167*, 605–614. Gidden, J.; Wyttenbach, T.; Jackson, A. T.; Scrivens, J. H.; Bowers, M. T. *J. Am. Chem. Soc.* **2000**, *122*, 4692–4699.
- (44) Several reviews have discussed ion mobility techniques. See for example: St. Louis, R. H.; Hill, H. H. *Crit. Rev. Anal. Chem.* **1990**, *21*, 321–355. Clemmer, D. E.; Jarrold, M. F. *J. Mass Spectrom.* **1997**, *32*, 577–592. Hoaglund Hyzer, C. S.; Counterman, A. E.; Clemmer, D. E. *Chem. Rev.* **1999**, *99*, 3037–3079. Shvartsburg, A. A.; Hudgins, R. R.; Dugourd, P.; Jarrold, M. F. *Chem. Soc. Rev.* **2001**, *30*, 26–35. Collins, D. C.; Lee, M. L. *Anal. Bioanal. Chem.* **2002**, *372*, 66–73. Wyttenbach, T.; Bowers, M. T. *Modern Mass Spectrom. Topics Curr. Chem.* **2003**, *225*, 207–232.
- (45) Valentine, S. J.; Koeniger, S. L.; Clemmer, D. E. *Anal. Chem.* **2003**, *75*, 6202–6208.
- (46) Valentine, S. J.; Counterman, A. E.; Hoaglund, C. S.; Reilly, J. P.; Clemmer, D. E. *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 1213–1216.
- (47) Taraszka, J. A.; Counterman, A. E.; Clemmer, D. E. *Fresenius J. Anal. Chem.* **2001**, *369*, 234–245.
- (48) Mason, E. A.; McDaniel, E. W. *Transport Properties of Ions in Gases*; Wiley: New York, 1988.
- (49) Hoaglund, C. S.; Valentine, S. J.; Sporleder, C. R.; Reilly, J. P.; Clemmer, D. E. *Anal. Chem.* **1998**, *70*, 2236–2242.
- (50) Srebalus, C. A.; Li, J.; Marshall, W. S.; Clemmer, D. E. *Anal. Chem.* **1999**, *71*, 3918–3927.
- (51) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophor.* **1999**, *20*, 3551–3567.
- (52) The *Drosophila* protein database used in these studies obtained from www.ncbi.nlm.nih.gov. It contains 59518 protein sequences.
- (53) <http://www.chem.indiana.edu/facilities/proteomics/parser/main.htm>.
- (54) Stapleton, M.; Carlson, J.; Brokstein, P.; Yu, C.; Champe, M.; George, R.; Guarin, H.; Kronmiller, B.; Pacleb, J.; Park, S.; Wan, K.; Rubin, G. M.; Celniker, S. E. *Genome Biol.* **2002**, *3*, 1–8.
- (55) Tabb, D. L.; Huang, Y.; Wysocki, V. H.; Yates, J. R., III *Anal. Chem.* **2004**, *76*, 1243–1246.
- (56) Klassen, J. S.; Blades, A. T.; Kebarle, P. J. *Phys. Chem.* **1995**, *99*, 15508–15517.
- (57) Levitan, I. B.; Kaczmarek, L. K. *The Neuron: Cell and Molecular Biology*; Oxford University Press: New York, 1991.
- (58) Hardie, R. C.; Raghu, P. *Nature* **2001**, *413*, 186–193.
- (59) Michal, G. *Biochemical Pathways*, Wall chart 1982 Boehringer Mannheim GMBH, Germany.
- (60) Marie, M. L.; Champeil, P.; Møller, J. V. *Biochim. Biophys. Acta* **2000**, *1508*, 86–111.
- (61) Corbin, R. W.; Paily, O.; Yang, F.; Shabanowitz, J.; Platt, M.; Lyons, C. E.; Root, K.; McAuliffe, J.; Jordan, M. I.; Kustu, S.; Soupeine, E.; Hunt, D. E. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 9232–9237.
- (62) Gygi, S. P.; Rochon, Y.; Franza, B. R. Aebersold, R. *Mol. Cell Biol.* **1999**, *19*, 1720–1730.
- (63) Several researchers have utilized high-resolution ion mobility methods. See for example: Hudgins R. R.; Motoharu, I.; Jarrold, M. F.; Dugourd, P. *J. Chem. Phys.* **1999**, *111*, 7865–7870. Wu, C.; Siems, W. F.; Klasmeier, J.; Hill, H. H., Jr. *Anal. Chem.* **2000**, *72*, 391–395. Collins, D. C.; Lee, M. L. *Fresenius J. Anal. Chem.* **2001**, *369*, 225–233. Valentine, S. J.; Kulchania, M.; Barnes, C. A. S.; Clemmer, D. E. *Int. J. Mass Spectrom. Ion Processes* **2001**, *212*, 97–109.
- (64) Valentine, S. J.; Counterman, A. E.; Hoaglund-Hyzer, C. S.; Clemmer, D. E. *J. Phys. Chem. B* **1999**, *103*, 1203–1207.
- (65) Several references discuss algorithms for calculating theoretical cross sections. See for example: Mesleh, M. F.; Hunter, J. M.; Shvartsburg, A. A.; Schatz, G. C.; Jarrold, M. F. *J. Phys. Chem.* **1996**, *100*, 16082–16086. Shvartsburg, A. A.; Jarrold, M. F. *Chem. Phys. Lett.* **1996**, *261*, 86–91. Wyttenbach, T.; von Helden, G.; Batka, J. J., Jr.; Carlat, D.; Bowers, M. T. *J. Am. Soc. Mass Spectrom.* **1997**, *8*, 275–282.
- (66) Shvartsburg, A. A.; Sui, K. W. M.; Clemmer, D. E. *J. Am. Soc. Mass Spectrom.* **2001**, *12*, 885–888.
- (67) Mosier, P. D.; Counterman, A. E.; Jurs, P. C.; Clemmer, D. E. *Anal. Chem.* **2002**, *74*, 1360–1370.
- (68) Valentine, S. J.; Clemmer, D. E., work in progress.
- (69) Griffin, T. J.; Gygi, S. P.; Ideker, T.; Rist, B.; Eng, J.; Hood, L.; Aebersold, R. *Mol. Cell Proteomics* **2002**, *1*, 323–333.
- (70) Washburn, M. P.; Koller, A.; Oshiro, G.; Ulaszek, R. R.; Plouffe, D.; Cosmin, D.; Winzeler, E.; Yates, J. R. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 2107–3112.
- (71) Fletcher, B.; Latter, G. I.; Monardo, P.; McLaughlin, C. S.; Garrels, J. I. *Mol. Cell Biol.* **1999**, *19*, 7357–7368.

PR050038G