Proteome Profiling for Assessing Diversity: Analysis of Individual Heads of *Drosophila melanogaster* Using LC−Ion Mobility−MS

John A. Taraszka,[†,§] Xinfeng Gao,[†] Stephen J. Valentine,[†] Renã A. Sowell,[†] Stormy L. Koeniger,[†] 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

The proteomes of three heads of individual *Drosophila melanogaster* organisms have been analyzed and compared by a combination of liquid chromatography, ion mobility spectrometry, and mass spectrometry approaches. In total, 197 proteins are identified among all three individuals (an average of 120 ± 20 proteins per individual), of which at least 101 proteins are present in all three individuals. Within all three datasets, more than 25 000 molecular ions (an average of 9000 ± 2000 per individual) corresponding to protonated precursor ions of individual peptides have been observed. A comparison of peaks among the datasets reveals that peaks corresponding to protonated peptides that are found in all heads are more intense than those features that appear between pairs of or within only one of the individuals. Moreover, there is little variability in the relative intensities of the peaks common among all individuals. It appears that it is the lower abundance components of the proteome that play the most significant role in determining unique features of individuals.

Keywords: *Drosophila melanogaster* • ion mobility • mass spectrometry • proteomics

Introduction

A cornerstone of evolution is associated with the diversity of individuals within a population. This diversity is generally understood to arise at the genetic level and leads to characteristics that may be advantageous or disadvantageous within the context of the environment.[1,2] Although the relationships of genes and evolution are documented,[1,2] this information alone is incomplete because of issues related to when, where, and how gene products are expressed. It is anticipated that diversity among individuals should increase within the products of the genome.[3] The emerging field of proteomics,[4,5] in which large mixtures of proteins are characterized in a single experimental sequence, may allow the assessment of variability or similarity within individuals at the level of the proteome.

Many cellular processes affect the diversity at the proteome level. For example, in eukaryotic organisms, such as *Drosophila melanogaster* (the fruit fly, hereafter referred to as *Drosophila*), factors such as alternative splicing,[6−8] DNA recombination,[9,10] transcription start sites,[11,12] RNA editing,[13,14] polyadenylation,[15,16] and post-translational modifications[17,18] cause divergence at the level of individual proteomes. Moreover, each of these processes is complex. For example, there are at least eleven alternative splicing patterns that can cause divergence in transcripts[6−8,19] and the level of diversity that is introduced can be substantial (e.g., the single gene *Dscam*, found in *Drosophila*, is highly expressed in the embryonic nervous system and may generate over 38 000 protein isoforms).[20]

From the increased complexity introduced by these pathways, it is estimated that the number of proteins in a proteome can readily exceed the number of genes in a genome by orders of magnitude.[7] A number of new technologies have been developed to study complex mixtures of proteins directly.[4,5] One of the most influential involves the combination of multiple dimensions of condensed-phase separations with mass spectrometry (MS).[4] In this approach, mass-to-charge (m/z) measurements are compared against information in databases in order to identify proteins.[4]

In the last several years, our group has worked to develop ion mobility spectrometry (IMS) as a high-speed, gas-phase separation for incorporation into liquid chromatography (LC)−MS platforms.[21−24] Because the time scale required for the ion mobility separation (on the order of ms) is between that required for LC separations (on the order of seconds to minutes) and time-of-flight (TOF) mass detection (μs), it is possible to include an IMS separation at no cost in the total experimental time. Inclusion of a high-resolution IMS separation allows some peptide isomers that cannot be resolved by LC−MS analysis to be distinguished.[21,22] Additionally, the IMS separation reduces spectral congestion and can reduce effects of chemical noise.[25] This combination of LC−IMS−MS methodologies, where peptides are dispersed according to their hydrophobicity, ionic average collision cross-section, and ionic m/z ratio, makes it possible to generate collision-induced

Scheme 1. Flowchart Illustrating the Overall Experimental Procedure Used in the Analysis of the Individual *Drosophila* Head Proteome
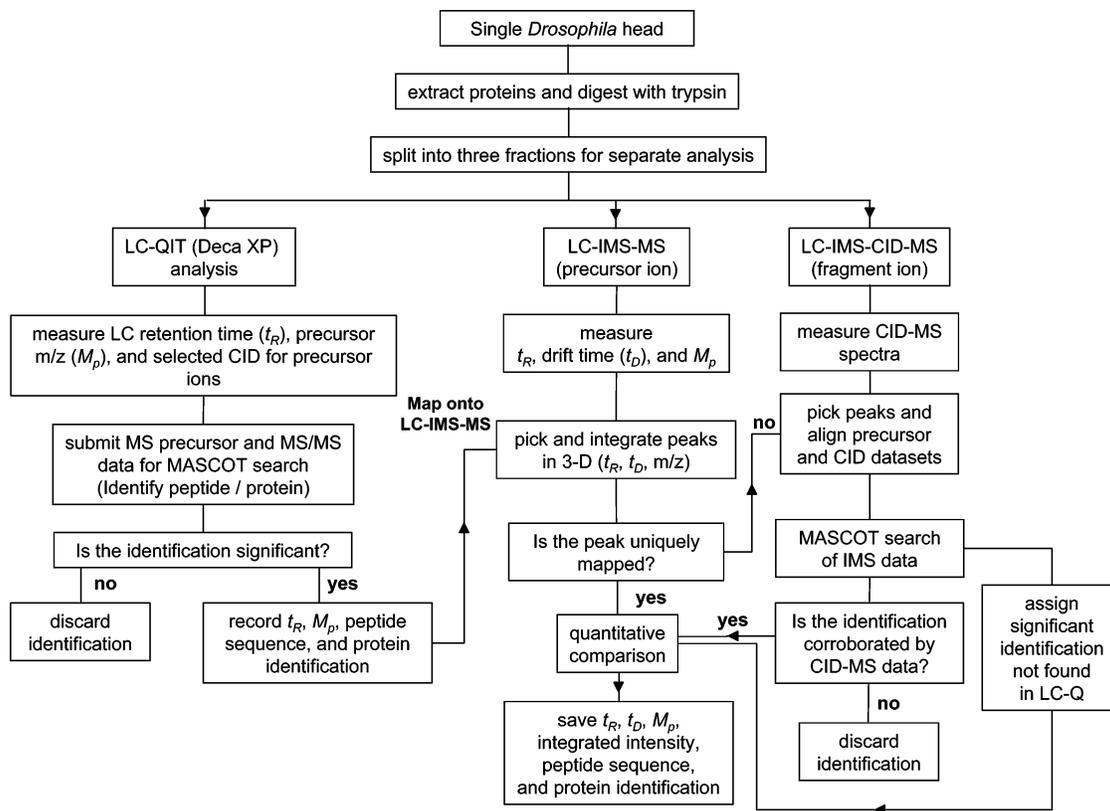
dissociation (CID) patterns for ions in parallel.[26] The present experiments employed a low-resolution drift tube with a resolving power $[t_D(\text{total})/\Delta t_D(\text{fwhm})]$ ranging from ~17 to 35 for different ions across the spectrum. Although it is often not possible to resolve isobaric species within a given charge state, the current experiments do allow for separation of ions based on their charge state. Hence, two ions with a nominal m/z equal to 800 in two different charge states are readily resolved. In addition, parent ions of different m/z ratios within a given charge state are resolved in the IMS (drift time) dimension. Resolution of parent ions allows the corresponding daughter ions (generated in CID mode) to be resolved and readily matched to their parent ion. Furthermore, the resolution of the daughter ions, greatly increases the peak capacity of an LC–IMS–CID–MS experiment relative to that of a typical LC–CID–MS experiment. In the IMS approach CID spectra are collected without parent ion preselection and many series of daughter ions are resolved over the drift time space in a given LC window. In contrast, in a LC–CID–MS experiment CID spectra are collected sequentially due to the use of parent ion preselection. Last, we point out that IMS experiments are extremely sensitive; previously we have reported detection limits ranging from 10 to 100 amol.[24]

In this paper, we utilize the LC–IMS–(CID)–MS combination for a comparative proteome analysis of three individuals. In summary, we find evidence for 197 proteins across three individuals; of these, at least 101 proteins are present in all three of the individuals. The intensities of peaks that are in common to all three individuals are also consistent. Below, we provide evidence that proteins expressed in all individuals are expressed at relatively high levels; unique features within an individual appear to arise from lower abundance peaks in the proteome.

Experimental and Methods

Protein Isolation and Tryptic Digestion. In these experiments wild-type Oregon-R *Drosophila* (that are nearly 100% genetically identical) are grown under identical conditions described previously.[27] Briefly, heads were obtained from adult female flies that were one week old, and proteins from each head were extracted using a mortar and pestle into 100 μL of a phosphate buffered saline solution containing 4 M urea and 0.1 mM α-toluenesulfonyl flouride. A Bradford assay indicated that ~8 μg of protein is obtained from a single head, which has a dry mass of ~20 μg. Reduction, alkylation, and trypsin digestion of the extracted proteins were carried out using standard protocols.[27] For analysis, the sample is dissolved in 20 μL of water and 2 uL is injected onto the LC system.

Overview of Analysis of *Drosophila* Head Proteomes. In this study, two aspects of the analysis are considered: identification of peptides and proteins within each individual and a quantitative comparison of peaks between individuals (even if the peaks are not identified). Scheme 1 provides an overview of the experimental procedures. A mixture of tryptic peptides is split into three fractions. In one analysis (the left side of Scheme 1) tryptic peptides are analyzed with a commercial LC–QIT to measure the retention times (t_R), precursor ion mass-to-charge ratios (M_p), and selected MS/MS spectra. The two other fractions are subjected to LC–IMS–MS and LC–IMS–(CID)–MS analysis (the right side of Scheme 1) to obtain precursor ion and CID datasets, respectively. From IMS experiments ion information obtained includes t_R, drift time (t_D), M_p, integrated peak intensities, precursor ion and CID–MS information.

Nanoflow LC Conditions. An Agilent 1100 CapPump (Agilent Technologies, Palo Alto, CA) was used for the LC separations. The setup of the nanoflow system is described elsewhere.[23,27]

Briefly, peptides are eluted from a pulled-tip nanocolumn at a flowrate of 250 nL·min⁻¹ using a gradient consisting 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).

Overview of IMS−MS and IMS−(CID)−MS Techniques. Ion mobility techniques have been used for gas-phase separations,[28] and for studying gas-phase ion structure.[29] A number of authoritative reviews are available.[30] The ion mobility instrument used in these experiments has been described previously.[27,31] Only a brief overview is presented here. Peptides eluting from the pulled-tip nanocolumn are electrosprayed into an octopole linear ion trap where ions are stored between drift tube experiments (duty cycle = 167 Hz). Nested drift(flight) time measurements are initiated by ejecting a 100 μs pulse of ions out of the trap into a ∼20 cm-long low-field (∼5 V·cm⁻¹) region of the drift tube filled with ∼1.6 Torr of 300 K He buffer gas. As ions exit the low-field region they enter a shorter (∼1.2 cm) second drift region that is modulated between low-field conditions (to transmit precursor ions) and high-field conditions (to induce fragmentation).[31] Ions exit the drift tube and enter the source region of an orthogonal-reflectron TOF mass spectrometer. Flight times (t_F) for the highest m/z ions of interest are about 2 orders of magnitude shorter than the drift times associated with the lowest mobility ions of interest, and thus, hundreds of TOF spectra can be obtained in a single drift experiment.[32]

Nomenclature for Nested LC−IMS−MS Measurements. We report peak positions associated with the different separation dimensions (e.g., t_R, t_D, and t_F) using a nomenclature that has been described previously.[23] In a three-dimensional LC−IMS−MS measurement, the position of a single peak is indicated by the following: $t_R[t_D(m/z)]$ in units of min[ms(μs)]. In a similar fashion, the positions of peaks in two-dimensional plots are represented by $t_R[t_D]$ for drift time versus retention time plots or by $t_R(m/z)$ for LC−MS plots.

LC-QIT Measurements. LC−MS experiments were performed on a LCQ Deca XP ion quadrupole trap (QIT) mass spectrometer (ThermoElectron Inc., Waltham, MA) coupled to a nanoflow LC system (Dionex Inc., Sunnyvale CA) as described previously.[27] The LC gradient used in LC-QIT experiments is identical to the gradient used in LC−IMS−MS experiments.

Assignment of Peptide Sequences from LC−QIT and LC−IMS−(CID)−MS Data. The methods employed to assign peptides have been discussed previously.[27] Briefly, the MASCOT (Matrix Science Ltd., London, UK) program is used to search the National Center for Biotechnology Information *Drosophila* protein database.[33,34] In these experiments the same MASCOT parameters are used as previously described.[27] A protein is considered identified only if at least one peptide unique to the protein obtains a significant score; a significant score indicates that the match has a less than 5% chance of occurring at random.[33] If the score is not significant (Scheme 1), then the identification is discarded. In some cases, peptides are identified from the LC−QIT data but are not identified in the LC−IMS−(CID)−MS analysis. In such cases, the LC−QIT assignments can be mapped onto the LC−IMS−MS data in order to assign peaks.[27]

Comparison of LC−IMS−MS Datasets. The precursor ion datasets from LC−IMS−MS experiments are calibrated and superimposed to determine the peaks that are in common among the different datasets. In these experiments, the cali-

brated precursor ion data from three different individuals are analyzed and grouped into three categories of peaks: (1) common among all three individuals; (2) found exclusively between pairs of individuals; and, (3) unique to individuals. Here, peaks present in different datasets are considered identical if their positions are within $t_R[t_D(t_F)]$ tolerances of ±0.24-[0.125(0.008)] min[ms(μs)].

Estimation of the Relative Abundances of Peptides and Proteins among Individuals using LC−IMS−MS Data. One advantage of the IMS technique is the ability to correlate peak intensities and peptide abundance. To compare intensities of peaks between datasets, peaks within a LC−IMS−MS dataset are normalized to the total ion current (TIC). In this approach, normalized intensities can be compared among the three datasets (individuals). Here, three methods examine the abundances of peptides and proteins among the three individuals.

In the first approach, we compare the intensities of peaks that are grouped into the three categories discussed in the previous section. Here, integrated peak intensities are binned in increments of 50 units using a minimum value of 384 and a maximum equal to the most intense peak in a given dataset. Then, the normalized fraction of peaks that fall within each intensity bin is calculated. Using this method one can ascertain if peaks that fall within the three categories have different overall intensities.

A second approach determines if there are any changes in the relative abundances of peptide peaks that are common among individuals. This is accomplished by calculating the intensity ratios between pairs of individuals for these peaks. For this analysis, a normalized intensity ratio (NIR) is defined as shown in eq 1

$$\text{NIR} = \frac{I_x}{I_y} \tag{1}$$

where I_x is the normalized intensity of the peak in individual x and I_y is the normalized intensity of the same peak in individual y. A NIR close to unity indicates that a peptide is not changing in abundance between two individuals.

The third approach estimates protein abundances by examining the peak intensities for peaks assigned to a specific protein. This approach is divided into two parts. The first part ascertains if there are any proteins that change in relative abundance between two individuals. In this part of the analysis a protein intensity ratio (PIR) is defined as shown in eq 2

$$\text{PIR} = \frac{\sum_{i=1}^{n_x} I_x}{\sum_{i=1}^{n_y} I_y} \tag{2}$$

where n_x is the total number of peptides observed for a specific protein in individual x, I_x is the normalized intensity of each peptide, n_y is the total number of peptides observed for the same protein in individual y, and I_y is the normalized intensity of each peptide. Similar to the NIR calculated for peptides, a PIR close to unity indicates that a specific protein is expressed similarly between two individuals. In the second part of the analysis, the relative protein abundances are estimated. The average total intensity ($\langle I \rangle$) is defined as the average of the total peak intensities (for a given protein) among all three individu-
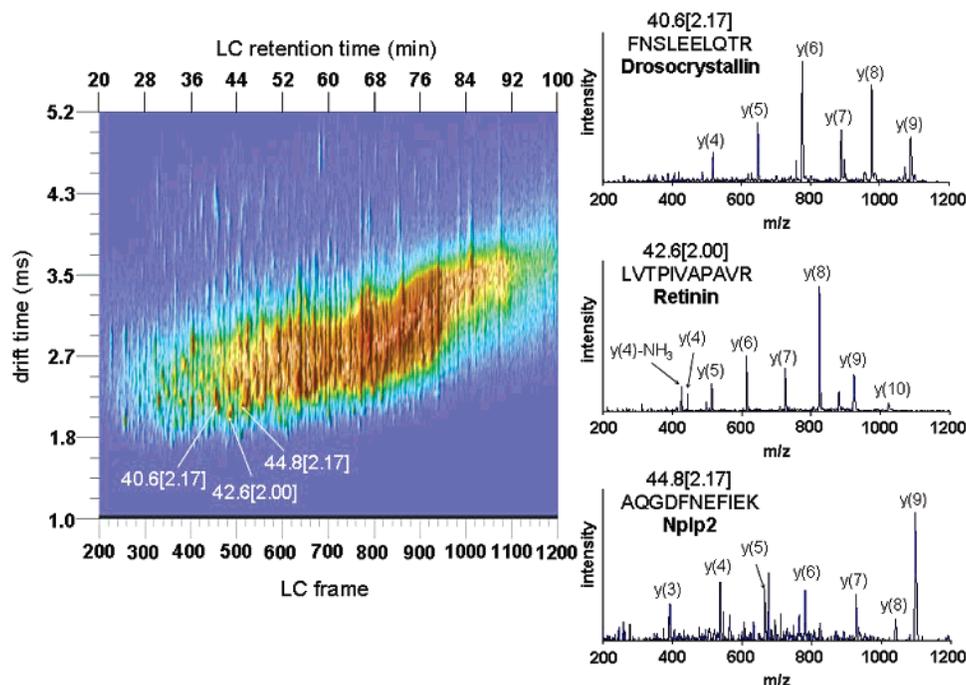
Figure 1. Left portion shows a two-dimensional representation of a three-dimensional LC−IMS−(CID)−MS dataset obtained from an individual *Drosophila* head. The two-dimensional representation is obtained by integrating all time-of-flight bins at given drift and retention times. Each individual spot of the graph contains a separate CID mass spectrum. Shown on the right are three CID mass spectra extracted from the positions indicated on the plot shown on the left. Each of these spectra corresponds to a peptide that is unique to the protein specified in the figure.

als. A large $\langle I \rangle$ value indicates that a particular protein is present at a relatively high concentration (in terms of mass·volume^{-1}).

Results and Discussion

Typical LC−IMS−MS Data for an Individual *Drosophila* Head. To simplify visualization of multidimensional data, it is useful to collapse dimensions and utilize two-dimensional representations. Figure 1 shows a typical plot of drift times as a function of retention times for a LC−IMS−(CID)−MS dataset from a single *Drosophila* head. In this representation, each spot generally corresponds to a peptide (or group of peptides) that is resolved from the complex mixture based on the LC−IMS separation. In the plot shown here peptides are separated over a range of $t_R[t_D] \approx 20[2.0]$ to $90[5.0]$. In the present data, the m/z axis contains mass spectra that correspond to conditions in which the peptides may fragment in the second drift region.

Figure 1 also provides three examples of mass spectra obtained from the three-dimensional dataset. Under the employed conditions, the mass spectra appear similar to MS/MS fragmentation patterns generated by other methods [e.g., CID data obtained in LC−QIT experiment (data not shown)]. The peak at $t_R[t_D] = 40.6[2.17]$ contains a series of intense fragment ions at 40.6[2.17(518.52, 647.78, 776.47, 889.79, 976.59, and 1090.89)]. When this information is combined with the precursor ion m/z value, MASCOT identifies the peak as the peptide FNSLEELQTR—a peptide unique to the protein drosocrystallin. Similar analysis of the peaks at 42.6[2.00] and 44.8[2.17] indicates that the peaks correspond to the peptides LVTPIVA-PAVR and AQGDFNEFIEK, respectively. The peptides LVTPIVA-PAVR and AQGDFNEFIEK are unique to the proteins retinin and neuropeptide-like precursor 2 (Nplp2), respectively. In all three spectra, an intense series of y-ion fragments are present.

It is also useful to view a subset of a three-dimensional

dataset to gain an understanding of the peak features. Figure 2 shows a $t_R[t_D(m/z)]$ precursor ion dataset for an individual *Drosophila* head that ranges from 40[1.83 (637)] to 48[2.67-(692)]. The subset contains over 20 intense peaks (and many other lower intensity features) and corresponds to less than 1% of the total separation space. In this figure three peaks are labeled as representative examples from the dataset. The peak at 43.2[2.17(638.40)] corresponds to the [TGEELQAAED-KINHLNK + 3H]$^{3+}$ ion, which is unique to myosin heavy chain. This peak has an integrated intensity of 4668 at the full width at half-maximum (fwhm) in all dimensions. Another example of a well-resolved intense peak is observed at 41.4[2.04(655.90)]; its integrated intensity is 2422. Finally, we point out a peak at 43.9[2.46(674.60)] that is ∼10 times less intense than the peak at 43.2[2.17(638.40)]. This peak also has not been assigned. Only a small fraction of the precursor ions are assigned to specific peptides; however, intensity analysis has been done for all peptide peaks.

A Brief Overview of Proteins Identified in Individual *Drosophila* Heads. Before discussing the trends in the proteome data among the individuals, it is useful to summarize the proteins that are observed in these studies. Table 1 lists all proteins (197) identified in this study and contains information about the cellular component obtained from Gene Ontology (GO) databases[35,36] as well as a description that illustrates the assigned function of these proteins. An examination of GO databases reveals that 78 out of 197 (40%) of the identified proteins have unspecified cellular components. Of the proteins with specified cellular components, 52 out of 112 (46%) are associated with the mitochondrion. Additional information on specific proteins can be obtained from the *Drosophila* genome database, FlyBase.[36] Over 95% of proteins observed in these studies were previously detected in other studies that examined
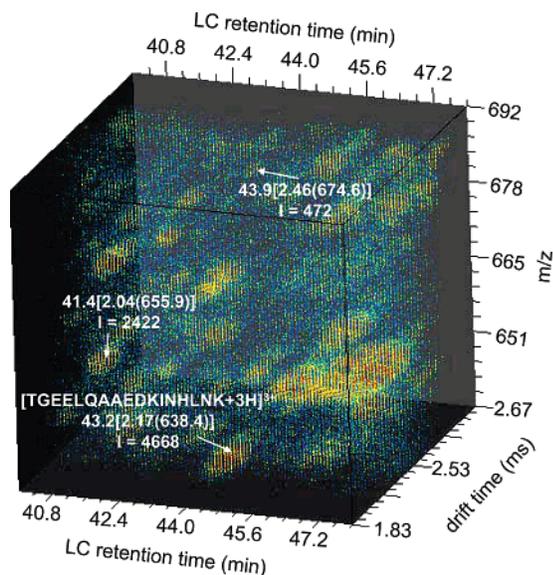
Figure 2. Three-dimensional $t_R[t_D(m/z)]$ precursor ion LC−IMS−MS dataset for an individual *Drosophila* head. Peaks in this plot appear as clusters of data points. In this plot peaks are represented using a false color scheme, where blue features have the lowest number of counts (3) and red features have the highest number of counts (10 or greater). An intensity cutoff equal to three was used to minimize background. Positions of peaks are denoted using the $t_R[t_D(m/z)]$ nomenclature, and integrated peak intensities (I) are also provided. The integrated peak intensities are calculated in three dimensions over the full width at half-maximum (fwhm) of the peaks.

a population of *Drosophila* heads.[27] It is important to note that many of the listed proteins have single peptide identification. If we only count proteins with two or more peptides only 62 are identified between the three individuals. However, one needs to be very careful in using this approach. Many of the proteins that have single peptide identifications are found in two or more samples. For example, the protein FBgn0052029 is identified by the peptide AQQQQGYVAPSVR in all three individuals (i.e., three different samples).

Comparison of Three Individuals. Figure 3 shows another representation of the LC−IMS−MS datasets recorded for the three different *Drosophila* heads, in this case a $t_R(m/z)$ representation. Thousands of peaks are observed in each plot, and many features are common among all three individuals. We show examples of three peaks that are common to all individuals [i.e., $t_R(m/z)$ = 24.8(1125.56), 36.7(1003.74), and 44.1-(891.36)], two peaks that are in common to two of the three individuals [i.e., $t_R(m/z)$ = 25.8(1192.87) and 27.8(1105.18)], and one peak that is unique [i.e., $t_R(m/z)$ = 50.9(1629.88)]. Although a visual inspection of the LC−MS plots provides a general overview, this analysis is limited by intensity cutoffs and the collapse of the ion mobility dimension. Any peaks that are resolved in the t_D dimension [but not in $t_R(m/z)$ dimensions] are not distinguishable in the LC−MS plots.

A more detailed analysis of the individual proteomes examines the overlap of precursor ion (peptide) peaks among the three LC−IMS−MS datasets. This is performed by superimposing LC−IMS−MS datasets in all three dimensions of the analytical space. In these experiments, we observed 8846, 10452, and 7358 precursor ion peaks in datasets corresponding to individual 1, 2, and 3, respectively. These peaks have integrated peak intensities of 350 or greater and are well above
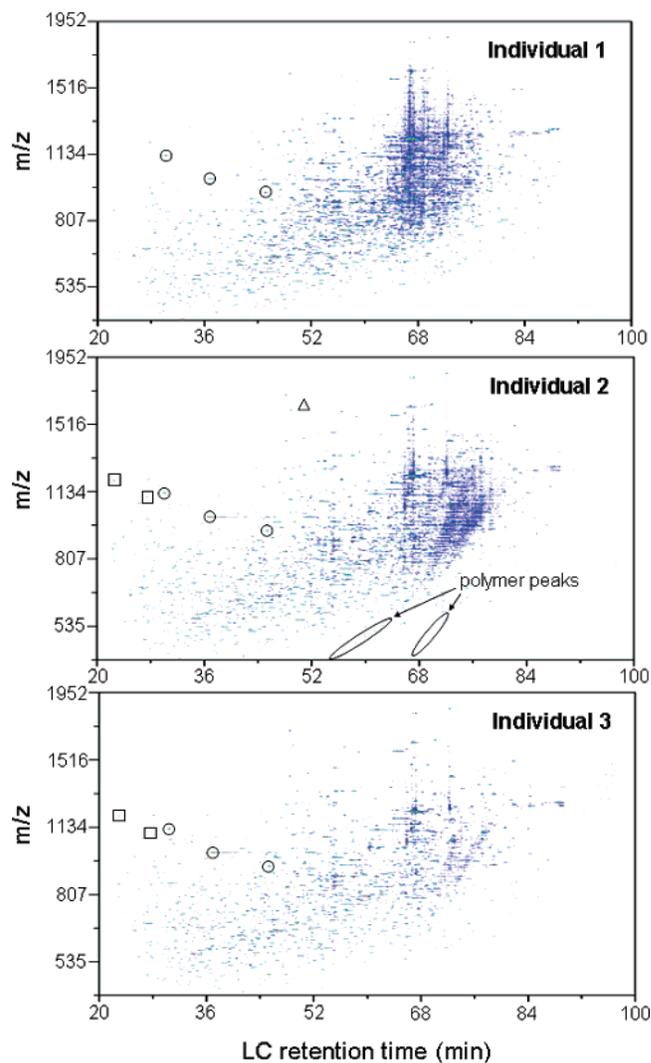
Figure 3. Two-dimensional LC−MS plots of three-dimensional LC−IMS−MS datasets. These plots are generated by integrating the three-dimensional data arrays over all drift times at given retention and flight times. The t_R dimension has been empirically calibrated to correct for slight shifts in retention times between the experiments. Circled features indicate peptide peaks that were found in all three individuals; boxed peaks indicate two peaks that are found in individuals 2 and 3, but not in individual 1. The peak in individual 2 that is enclosed by a triangle is unique to that individual. In the plot of individual 2 we observe polymer peaks that result from column bleed. Because these peaks are not related to the proteome, they are excluded from the analysis.

the detection limit. Among the three individuals, 2124 precursor ion peaks are found in common. Examination of the peaks in common exclusively between individual pairs (i.e., excluding the 2124 peaks found in common among all three) reveals that there are 1606 peaks in common between individuals 1 and 2, 1030 peaks in common between individuals 1 and 3, and 2196 peaks in common between individuals 2 and 3. Last, we note that each individual has thousands of peaks that are unique. Specifically, individual 1, 2, and 3 have 4086, 4526, and 2008 unique peaks, respectively. When a peak is not found this means that we have readily detected it in other samples, and it should be easily observed above the detection limit. So far, 273 tryptic peptides (corresponding to 101 proteins) are assigned to 316 of the 2124 peaks found in common among all

Table 1. List of *Drosophila* Proteins Identified from Three Heads

FBgn no.[a]	description[b]	cellular component[c]	FBgn no.	description	cellular component
0000055	Alcohol dehydrogenase		0000064	Aldolase	
0000116	Arginine kinase		0000121	Arrestin 2	Rhabdomere
0000253	Calmodulin	Cytoplasm, Rhabdomere	0000409	Cytochrome *c* proximal	Mitochondrion
0000551	Ecdysone-dependent gene 78E		0000556	Elongation factor 1α 48D	Cytoplasm
0000559	Elongation factor 2b	Cytoplasm	0000579	Enolase	
0000592	Esterase 6		0000667	A actinin	PM
0001090	Bangles and beads		0001091	Glyceraldehyde 3 phosphate dehydrogenase 1	Cytoplasm
0001092	Glyceraldehyde 3 phosphate dehydrogenase 2	Cytoplasm	0001128	Glycerol 3 phosphate dehhydrogenase	Cytoplasm
0001145	Glutamine synthetase 2	Cytoplasm	0001197	Histone H2A variant	Nucleus, Chromosome
0001208	Henna		0001218	Heat shock protein cognate 3	ER
0001219	Heat shock protein cognate 4	Mitochondrion, Nucleus	0001224	Heat shock protein 28	
0001233	Heat shock protein 83	Centrosome, Cytoplasm	0002593	Ribosomal protein P2	Cytosolic ribosome
0002611	Ribosomal Protein L12	Cytosolic ribosome	0002719	Malic enzyme	
0002741	Myosin heavy chain	Muscle fiber	0002772	Myosin alkali light chain 1	Muscle fiber
0002773	Myosin light chain 2	Myosin	0002855	Accessory gland-specific peptide 26Aa	Extracellular
0002921	Na pump α subunit	PM, Na/K-ATPase complex	0002938	ninaC	Cytoplasm, Rhabdomere
0002940	ninaE	Rhabdomere; ER	0003074	Phosophoglucose isomerase	Cytoplasm
0003149	Paramyosin	Muscle fiber	0003178	Pyruvate kinase	
0003360	Stress-sensitive B	Mitochondrion	0003462	Superoxide dismutase	Cytoplasm
0003721	Tropomyosin 1	Muscle fiber	0003738	Triose phosphate isomerase	
0003861	Transient receptor potential	Rhabdomere	0003884	α-tubulin at 84B	Microtubule
0003885	α-tubulin at 84D	Microtubule	0003887	B-tubulin at 56D	Microtubule
0004028	Wings up A	Muscle fiber	0004045	Yolk protein 1	
0004047	Yolk protein 3		0004117	Tropomyosin 2	Muscle fiber
0004169	Upheld	Muscle fiber	0004362	High mobility group protein D	
0004363	Porin	Mitochondrion	0004432	Cyclophilin 1	Cytosol
0004435	G-protein β49B	PM, Rhabdomere	0004507	Glycogen phosphorylase	
0004516	Glutamic acid decarboxylase 1		0004551	Calcium ATPase at 60A	ER
0004623	G-protein β76C	Cytoplasm, PM, Rhabdomere	0004907	14−3−3	Ring canal
0005391	Yolk protein 2	Nucleus	0005664	Drosocrystallin	
0005671	Vacuolar H+ ATPase 55 kD β subunit	Mitochondrion	0010100	Aconitase	Mitochondrion
0010213	Superoxide dismutase 2 (Mn)	Mitochondrion	0010217	ATP sythase β subunit	Mitochondrion
0010228	HMG protein Z	Nucleus	0010387	Diazepam-binding inhibitor	
0010397	Lamin C	Nucleus	0010516	Walrus	Mitochondrion
0010531	CCS		0010612	l(2)06225	Mitochondrion
0010808	l(3)03670		0011211	Bellwether	Mitochondrion
0011280	Pheromone-binding protein-related protein 2	Extracellular	0011361	Mitochondrial acyl carrier protein 1	Mitochondrion
0011643	Muscle LIM protein at 60A	Nucleus	0011693	Photoreceptor dehydrogenase	
0011695	Ejaculatory bulb protein III	Extracellular	0011726	Twinstar	
0013334	Synapse-associated protein 47kD	Synaptic junction	0013733	Short stop	Microtubule
0013954	FK506-binding protein 2	Cytoplasm	0014002	Protein disulfide isoemerase	ER
0014391	Stunted	Mitochondrion	0014869	Phosphoglyceromutase	
0015031	Cyclope	Mitochondrion	0015221	Ferritin 2 light chain	Ferritin complex
0015222	Ferritin 1 heavy chain	Ferritin complex	0015245	Heat shock protein 60	Mitochondrion
0015288	Ribosomal protein L22	Cytosolic ribosome	0015324	Vacuolar H+ ATPase 2 6kD E subunit	Mitochondrion
0015390	futsch	Microtubule cytoskeleton	0016119	ATPase coupling factor 6	Mitochondrion
0016120	ATP synthase subunit d	Mitochondrion	0016685	Nucleoplasmin	Nucleus
0016691	Oligomycin sensitivity-conferring protein	Mitochondrion	0016724	Retinoid- and fatty-acid binding protein	
0017539	Succinyl conenzyme A synthetase flavoprotein subunit	Mitochondrion	0017565	Nac	NPAC
0017567	NADH:ubiquinone reductase 23 kD	Mitochondrion	0019624	Cytochrome *c* oxidase subunit Va	Mitochondrion
0020235	ATP synthase γ chain	Mitochondrion	0020238	14−3−3	Ring canal
0020439	Fau		0020907	Sacroplasmic calcium-binding protein 2	Sacroplasmic reticulum
0021765	Scully	Mitochondrion	0022355	Transferrin 1	
0024289	Sorbitol dehydrogenase 1		0025839	CG3621	Mitochondrion
0026170	Smt3		0026409	Mitochondrial phosphate carrier protein	Mitochondrion

Table 1. (Continued)

FBgn no.[a]	description[b]	cellular component[c]	FBgn no.	description	cellular component
0026415	Imaginal disk growth factor 4		0027571	CG3523	
0027580	CG1616	Mitochondrion	0027779	Vacuolar H+ ATPase SFD subunit	Mitochondrion
0028479	CG4389	Mitochondrion	0028737	Elongation factor 1	Cytosol
0029704	CG2982		0029721	CG7010	Mitochondrion
0029869	CG3861	Mitochondrion	0029889	CG4094	Cytoplasm; Mitochondrion
0030136	CG2998	Cytosolic ribosome	0030184	CG2968	Mitochondrion
0030362	Regucalcin		0030733	CG3560	Mitochondrion
0031024	CG12233	Mitochondrion	0031037	CG14207	
0031066	CG14235	Mitochondrion	0031208	CG11023	
0031408	CG10882		0031453	CG9894	
0031692	CG6514		0031800	CG9497	
0031830	CG11015	Mitochondrion	0031912	CG5261	Mitochondrion
0032114	CG3752	Mitochondrion	0032237	CG5362	Cytsol
0032286	CG7300		0032299	CG17127	
0032686	Phosphodiesterase 11		0032820	Fructose-1,6-bisphosphatase	
0032833	CG10664	Mitochondrion	0032946	Nervana 3	PM, Na/K-ATPase complex
0033029	l(2)NC136		0033446	CG1648	
0033597	CG9079		0033663	Erp60	
0033728	CG8505		0033730	CG8511	
0034471	Odorant-binding protein 56e		0034643	CG10321	
0034877	CG17280	Mitochondrion	0035281	CG1919	
0035499	Chd64		0035817	CG7409	
0035917	CG6416		0036106	CG6409	ER
0036182	CG6084		0036334	CG11267	Mitochondrion
0036619	CG4784		0036642	CG4169	Mitochondrion
0036762	CG7430	TCA cycle enzyme complex	0036824	CG3902	
0036927	CG7433	Mitochondrion	0037138	CG7145	Mitochondrion
0037643	CG11963	TCA cycle enzyme complex	0037874	CG4800	Cytoplasm
0037891	CG5214	Mitochondrion	0038224	CG3321	Mitochondrion
0038271	CG3731	Mitochondrion	0038294	CG6803	
0038569	CG7218		0038587	CG7998	Mitochondrion
0038840	CG5621	PM	0039512	CG14264	
0039682	Odorant-binding protein 99c		0039697	CG7834	Mitochondrion
0039713	CG7808	Cytosolic ribosome	0039737	CG7920	
0039802	dj-1		0040064	Yippee interacting protein 2	Mitochondrion
0040066	Will die slowly		0040074	Retinin	
0040282	CG11956		0040309	Thioredoxin peroxidase 1	Cytosol
0040361	CG14627		0040660	CG13551	
0040813	Neuropeptide-like precursor 2	Extracellular	0041605	cpx	
0042119	CG18778		0050045	CG30045	
0050115	CG30115		0051305	CG31305	Mitochondrion
0051618	CG31618		0051878	CG31878	
0052039	CG32039		0052920	CG32920	
0061209	CG17949				

[a] The FlyBase gene number is provided. This number can be used to search the FlyBase database. [b] The name of the protein or gene is given as a description of the identified protein. In circumstances where there is no specified name, the computed gene (CG) number is provided as a cross reference. [c] Cellular components are obtained from gene ontology databases accessible from Flybase.[35,36] In some cases, the most specific cellular component is not given for conciseness. For example, mitochondrial proteins are simply listed as mitochondrion and not into other subcategories, such as mitochondrial inner membrane. No entry indicates that the cellular component is not specified. The following abbreviations are used throughout the table: plasma membrane (PM), endoplasmic reticulum (ER), and nascent polypeptide associated complex (NPAC).

individuals. Because ~1800 peaks in common are still not assigned to peptides, the remaining 96 proteins may still be found in all three individuals. However, at least 101 proteins are common among all individuals. At this stage, it is unknown what factors are influencing the observed differences. Although we have grown identical *Drosophila* strains under identical conditions, physiological differences may still play a role. For example, one individual may ingest food at a different time than another individual.

Comparisons of Precursor Ion Intensity Distributions in LC–IMS–MS Datasets. Figure 4 shows a plot of the normalized fraction of peaks as a function of integrated peak intensities for those features that are common to all three organisms, two of the three organisms, and unique to one fruit fly. The distribution corresponding to unique peaks is dominated by lower intensity peaks; 93% of unique peaks have peak intensities less than 900. For peaks found between pairs the results are similar; 81% have peak intensities less than 900. In contrast, only 47% of peaks that are in common to all these organisms have peak intensities less than 900. The shapes of these profiles are also interesting. The features that are unique to a single animal dominate the distribution as the peak intensity is lowered. That is, low abundance features appear to be favored for a single individual. In all of these data, the peak intensity
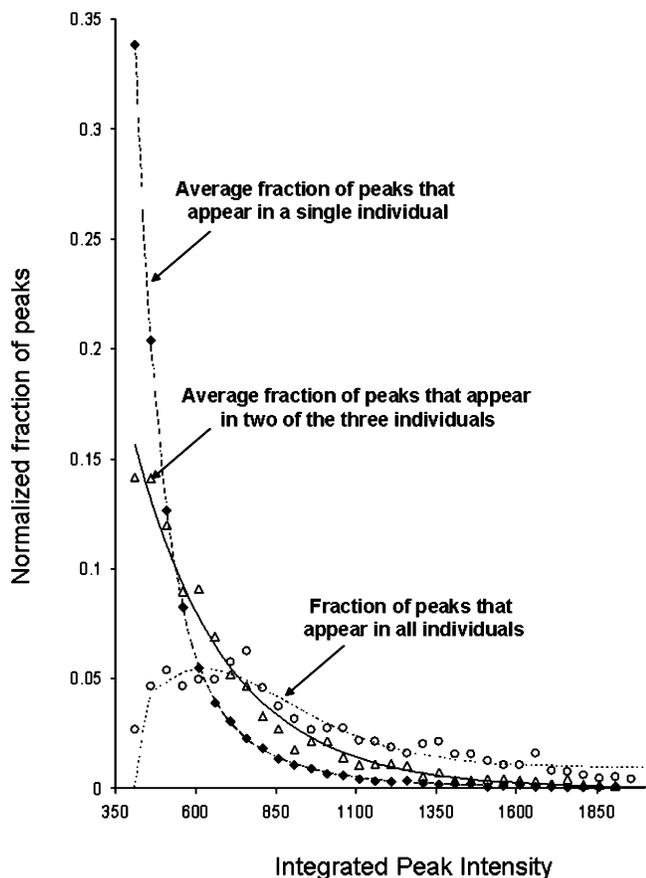
Figure 4. Plot of the normalized fraction of peaks detected as a function of peak intensity for peaks common to all, found between pairs, and unique to individuals. For the common peaks (open circles, small dashed line) the average normalized intensities are used. For the peaks found in pairs (open triangles, solid line) and unique to individuals (solid diamonds, large dashed line), an average normalized fraction is reported at a given peak intensity. In this figure, the sum of the fraction of peaks in each category equals one. The distributions of peaks found between pairs and individuals were empirically fitted to a general exponential decay function of the form $y = y_0 + Ae^{-x/b}$. The distribution of common peaks was empirically fitted to a Giddings peak function in Origin 7.0 (OriginLab Inc., Northampton, MA).

is well above the detection limit. Thus, these appear to be components that are unique to one individual. Those peaks that are in common to two animals often display peaks that are slightly larger. These features dominate this plot for a narrow region of peaks having intensities (I) of ∼650 to 750. Interestingly, larger peaks ($I > ∼850$) are associated most frequently with all three heads. These trends indicate that the peptides found in common among individuals are more abundant than those found between pairs or only in individuals. While it remains to be corroborated these data suggest that unique features are overall lowest in intensity indicating that lower abundance species may play a role in determining individual characteristics.

Determination of the Relative Abundances of Common Peaks among Individuals. Figure 5 shows a representation of the NIRs for the 2124 peaks in common among all three individuals. It is observed that NIRs cluster around a value of 1.0; the average NIR values range from 0.9 ± 0.4 (individuals 2:1) to 1.1 ± 0.3 (individuals 3:2). However, some values (those greater than 2.8 and less than 0.35) indicate a change in peak
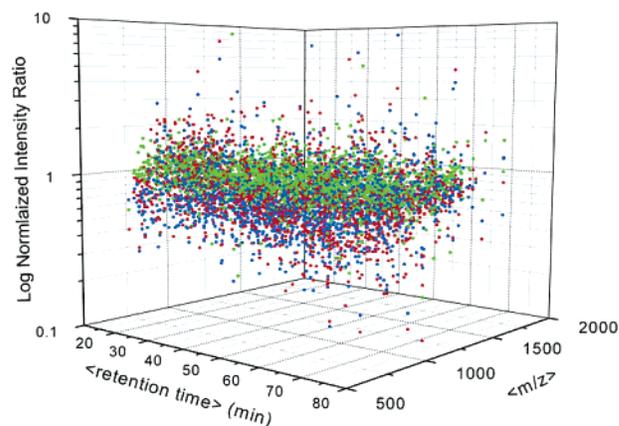
Figure 5. In this plot, NIRs (see eq 1) are calculated for pairs of individuals (2:1, 3:1, and 3:2). The NIRs shown correspond to individuals 2:1 (blue dots), 3:1 (red dots), and 3:2 (green dots). The average and standard deviations (@ ± σ) of the NIRs for individuals 2:1, 3:1, and 3:2 are 0.9 ± 0.4, 0.9 ± 0.5, and 1.1 ± 0.3, respectively. In control experiments, involving three back-to-back runs of the same sample, the normalized intensities agree to within 24% (relative uncertainty). With this uncertainty, the relative uncertainty in a NIR is 48%. Given the 48% uncertainty, we estimate that NIRs greater than unity are only significant [i.e., indicative of a change in peak intensity (and peptide abundance)] if they are greater than 2.8, and NIRs less than unity are only significant if they are less than 0.35.

intensity (and peptide abundance). For NIRs of individuals 2:1, 3:1, and 3:2 we find that there are 46, 51, and 6 NIRs that indicate a change in peak intensity, respectively. Individuals 2 and 3 have the fewest peaks that change in abundance (i.e., only 6 *NIR*s indicate a significant change). In contrast, ∼46 to 51 peaks have different intensities in head 1 relative to heads 2 or 3. Although these small changes exist, over 95% of the common peaks have the same intensities among the three studied individuals.

Abundances of Proteins Common to All Individuals. One of the simplest means for estimating the relative protein abundances between samples is to tabulate the average number of peptides identified in each sample.[37] Here, we use a similar method that includes the normalized intensities of the peaks corresponding to the identified peptides. Table 2 lists the average number of peptides observed for a given protein, PIR values and ⟨*I*⟩ values for the 33 proteins that have three or more common peptides identified. Proteins containing fewer than three common peptides are avoided to minimize sampling errors. PIR values estimate the relative changes in protein abundances among individuals, and ⟨*I*⟩ values estimate the relative abundances of proteins by using the intensity of identified peptide ion peaks. Values of PIRs range from 0.543 (for alcohol dehydrogenase between individuals 3:1) to 1.99 (for retinin between individuals 3:1). Within the estimated relative uncertainty of 48% all PIRs are within one sigma of unity; that is, the relative abundances of these proteins do not change among the individuals.

Yet, the abundances of common proteins vary. For example, myosin heavy chain has an average of 39 peptides identified in an individual and has the largest ⟨*I*⟩ value of 77602. ninaC, which is a calcium-dependent calmodulin binding protein, has an average of 3 peptides identified and has an ⟨*I*⟩ value of 3864. Utilizing the method described by Opiteck and co-workers, the estimated concentration of myosin heavy chain is ∼13 times

Table 2. Average Number of Peptides for a Given Protein, the Protein Intensity Ratios (PIR), and Average Total Intensity (⟨*I*⟩) for Identified Proteins Containing Three or More Peptides Found in All Three Individuals

FBgn No.[a]	description[b]	no.[c]	PIR 2:1[d]	PIR 3:1[d]	PIR 3:2[d]	⟨*I*⟩[e]
0000055	Alcohol dehydrogenase	5	0.609	0.543	0.893	11284
0000064	Aldolase	10	0.864	1.03	1.19	18788
0000116	Arginine Kinase	9	0.555	0.630	1.14	14206
0000253	Calmodulin	4	1.17	1.25	1.07	17142
0000579	Enolase	4	0.649	0.601	0.926	16203
0002741	Myosin Heavy chain	39	1.50	1.56	1.04	77602
0002772	Myosin Alkali light chain 1	4	1.30	0.894	0.687	7485
0002773	Myosin light chain 2	3	1.37	1.25	0.909	5653
0002921	Na pump α subunit	11	1.33	1.47	1.10	17832
0002938	ninaC	3	1.21	1.32	1.10	3864
0003149	Paramyosin	11	0.942	1.05	1.11	20567
0003721	Tropomyosin 1	8	1.06	1.14	1.08	9159
0003887	β-Tubulin at 56D	5	1.26	1.46	1.16	9022
0004117	Tropomyosin 2	8	1.98	1.711	0.863	15651
0004169	Upheld	4	0.817	0.799	0.978	5410
0004363	Porin	3	0.685	0.653	0.954	4331
0004432	Cyclophilin 1	4	1.05	0.816	0.774	4428
0004551	Calcium ATPase at 60A	4	1.07	1.32	1.23	6425
0005664	Drosocrystallin	6	1.27	1.27	1.00	14972
0010217	ATP synthase β subunit	9	0.838	0.831	0.991	19694
0011211	Bellwether	13	0.649	0.782	1.21	27451
0011693	Photoreceptor dehydrogenase	4	0.893	1.09	1.22	6832
0016120	ATP synthase subunit d	5	1.50	1.63	1.09	3225
0016724	Retinoid- and fatty-acid binding protein	3	0.791	0.825	1.04	4715
0029869	CG3861	3	1.27	1.66	1.30	6088
0032114	CG3752	4	1.00	0.809	0.806	5337
0036619	CG4784	6	1.25	1.36	1.09	16371
0037138	CG7145	4	0.737	0.785	1.07	7546
0038587	CG7998	4	0.813	0.990	1.22	8159
0040074	Retinin	6	1.77	1.99	1.14	33959
0040813	Neuropeptide-like precursor 2	3	1.09	1.20	1.10	21541
0051305	CG31305	3	0.858	0.626	0.730	8965
0051878	CG31878	3	0.587	0.710	1.21	3447

[a] The FlyBase gene number is provided. This number can be used to search the FlyBase database. [b] The name of the protein or gene is given as a description of the identified protein. In cases where there is no specified name, the computed gene (CG) number is provided as a cross reference. [c] The average number of peptides observed for a given protein is reported. [d] The protein intensity ratios (PIR) are provided for the individuals studied. The ratios are given between individuals 2:1, 3:1, and 3:2. [e] The average total intensity (of three measurements) for peptides belonging to a specified protein (see text).

(39/3) that of ninaC.[37] In our method, ⟨*I*⟩ values, which incorporate integrated peak intensities, indicate that myosin heavy chain is ~20 times (77602/3864) more abundant than ninaC. These two methods agree within a factor of 2 and appear to complement each other. These calculations are best viewed as estimations of the relative amount of proteins present in the individuals. Several experimental factors limit the accuracy of these calculations; these factors include: (1) the solubility of proteins and tryptic peptides; (2) the retention of peptides on the LC trapping and analytical columns; (3) the efficiency of trypsin cleavage; and, (4) the ionization efficiencies of peptides. In addition, it is difficult if not impossible to perform such estimations if the proteins being compared have few peptide identifications.

In summary, although there are many pathways for divergence in an individual organism, specific differences are not apparent in this study. This can be due to several reasons. First, we are only characterizing a very small percentage of the total proteome (<1% by current estimates). The identified proteins clearly represent the most abundant proteins in the individuals; it is likely that differences between individuals are influenced by lower abundance proteins (in some cases believed to be only 1 copy per cell). Second, differences between individuals at the proteome level may not become apparent until later in life; in

this study one-week old flies were used, and the lifespan of a wild-type *Drosophila* is typically 60 days. Third, differences between individuals may be due to alternative splicing and post-translational modifications (PTMs). In the studies presented here, we do not discern any alternatively spliced proteins and do not identify any PTMs. Fourth, environmental factors can substantially increase divergence in organisms; however, the *Drosophila* studied here were grown under identical conditions.

Conclusions

The proteomes of three individual *Drosophila* heads using a LC−IMS−MS approach have been examined. The results suggest that features unique to individuals may arise from lower abundant species. The more abundant proteins are expressed similarly among individuals suggesting that large variations in the abundances of these proteins may affect the viability of the organism.

References

(1) Dobzhansky, T. *Genetics of the Evolutionary Process*; Columbia University Press: New York, 1970.

(2) Masatoshi, N. *Molecular Evolutionary Genetics*; Columbia University Press: New York, 1987.

(3) Maniatis, T.; Tasic, B. *Nature* **2002**, *418*, 236−243.

(4) 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; and, Standing, R. *Curr. Opin. Struct. Biol.* **2003**, *13*, 595−601.

(5) 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; and, Lion, N.; Rohner, T. C.; Dayon, L.; Arnaud, I. L.; Damoc, E.; Youhnovski, N.; Wu, Z.; Roussel, C.; Jossernad, J.; Jensen, H.; Rossier, J. S.; Przybylski, M.; Girault, H. H. *Electrophor.* **2003**, *24*, 3533−3562.

(6) Black, D. L. *Cell*, **2000**, *103*, 367−370.

(7) Graveley, B. R. *Nature* **2001**, *17*, 100−107.

(8) Black, D. L. *Annu. Rev. Biochem.* **2003**, *72*, 291−336.

(9) Bénédicte, M. *Trends Biochem. Sci.* **2000**, *25*, 173−178.

(10) Barbour, L.; Xiao, W. *Mutat. Res.* **2003**, *532*, 137−155.

(11) Quelle, D. E.; Zindy, F.; Ashmun, R. A.; Sherra, C. J. *Cell* **1995**, *83*, 993−1000.

(12) Xue, S. A.; Jones, M. D.; Lu, Q. L.; Middledorp, J. M.; Griffin, B. E. *Mol. Cell Biol.* **2003**, *23*, 2192−2201.

(13) Keegan, L. P.; Gallo, A.; O'Connell, M. A. *Nat. Rev. Genet.* **2001**, *2*, 869−878.

(14) Blanc, V.; Davidson, N. O. *J. Biol. Chem.* **2003**, *278*, 1395−1398.

(15) Gautheret, D.; Poiroit, O.; Lopez, F.; Audic, S.; Claverie, J. M.; *Genome Res.* **1998**, *8*, 524−530.

(16) Macdonald, P. *Curr. Opin. Cell Biol.* **2001**, *13*, 326−331.

(17) Mann, M.; Jensen, O. N. *Nat. Biotech.* **2003**, *21*, 255−261.

(18) Cohen, P. *Trends Biochem. Sci.* **2000**, *25*, 596−601.

(19) Chow, L. T.; Gelinas, R. E.; Broker, T. R.; Roberts, R. J. *Cell* **1977**, *12*, 1−8.

(20) Schmucker, D.; Clemens, J. C.; Shu, H.; Worby, C. A.; Xiao, J.; Muda, M.; Dixon, J. E.; Zipursky, L. S. *Cell* **2000**, *101*, 617−684.

(21) Srebalus, C. A.; Hilderbrand, A. E.; Valentine, S. J.; Clemmer, D. E. *Anal. Chem.* **2002**, *74*, 26−36.

(22) Hilderbrand, A. E.; Myung, S.; Srebalus-Barnes, C. A.; Clemmer, D. E. *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 1424−1436.

(23) Moon, M. H.; Myung, S.; Plasencia, M.; Hilderbrand, A. E.; Clemmer, D. E. *J. Proteome. Res.* **2003**, *2*, 589−597.

(24) 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.

(25) Taraszka, J. A.; Counterman, A. E.; Clemmer, D. E. *Fren. J. Anal. Chem.* **2001**, *369*, 234−245.

(26) Hoaglund-Hyzer, C. S.; Jianwei, L.; Clemmer, D. E. *Anal. Chem.* **2000**, *72*, 7237−2740.

(27) Taraszka, J. A.; Kurulugama, R.; Sowell, R.; Valentine, S. J.; Koeniger, S. L.; Miller, D. E.; Kaufman, T. C.; Clemmer, D. E. *J. Proteome Res.* **2005**, *4*, 1223−1237.

(28) 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.

(29) 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 Processes* **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.

(30) 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.

(31) Valentine, S. J.; Koeniger, S. L.; Clemmer, D. E. *Anal. Chem.* **2003**, *75*, 6202−6208.

(32) Hoaglund, C. S.; Valentine, S. J.; Sporleder, C. R.; Reilly, J. P.; Clemmer, D. E. *Anal. Chem.* **1998**, *70*, 2236−2242.

(33) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophor.* **1999**, *20*, 3551−3567.

(34) The *Drosophila* protein database used in these studies obtained from http://www.ncbi.nlm.nih.gov/

(35) The Gene Ontology Consortium *Nature Genet.* **2000**, *25*, 25−29.

(36) The FlyBase Consortium *Nucleic Acids Res.* **2003**, *31*, 172−175. http://flybase.org/

(37) Pang, J. X.; Ginanni, N.; Dongre, A. R.; Hefta, S. A.; Opiteck, G. J. *J. Proteome Res.* **2002**, *1*, 161−169.