

1 **Systems-Level Annotation of Metabolomics Data Reduces 25,000**
2 **Features to Fewer than 1,000 Unique Metabolites**

3

4 Nathaniel G. Mahieu and Gary J. Patti*

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19 Departments of Chemistry and Medicine, Washington University, St. Louis, Missouri
20 63130, United States

21

22

23 *Contact: gjpattij@wustl.edu, 314-935-3512

24

25 **SUMMARY**

26 When using liquid chromatography/mass spectrometry (LC/MS) to perform untargeted
27 metabolomics, it is now routine to detect tens of thousands of features from biological samples.
28 Poor understanding of the data, however, has complicated interpretation and masked the
29 number of unique metabolites actually being measured in an experiment. Here we place an
30 upper bound on the number of unique metabolites detected in *Escherichia coli* samples
31 analyzed with one untargeted metabolomic method. We first group multiple features arising from
32 the same analyte, which we call “degenerate features”, using a context-driven annotation
33 approach. Surprisingly, this analysis revealed thousands of previously unreported degeneracies
34 that reduced the number of unique analytes to ~2,961. We then applied an orthogonal approach
35 to remove non-biological features from the data by using the ¹³C-based credentialing
36 technology. This further reduced the number of unique analytes to less than 1,000.

37

38 INTRODUCTION

39 It has become increasingly popular to perform untargeted metabolomics by using liquid
40 chromatography/mass spectrometry (LC/MS). This is at least in part due to the large number of
41 signals or features that are typically detected from most biological samples.¹⁻³ While it is often
42 assumed that these tens of thousands of detected signals provide "global" coverage of the
43 metabolome, the number of metabolites being measured in an experiment has not been
44 rigorously assessed. The major barrier preventing this type of analysis has been the challenge
45 of identifying metabolites.⁴ To date, the overwhelming majority of the detected signals in any
46 one untargeted metabolomics experiment have not been named. Even comprehensive efforts to
47 identify as many metabolites as possible in a data set by using the most advanced informatic
48 resources currently available have resulted in relatively small percentages of the total number of
49 signals being identified.⁵⁻⁷ Thus, the basic question of how many unique metabolites are being
50 profiled in an untargeted metabolomics experiment has remained outstanding.

51 It is important to note that uncertainties related to experimental coverage have not
52 prevented the widespread application of the untargeted metabolomics technology.
53 Improvements in instrumentation and software have made performing untargeted metabolomics
54 with LC/MS relatively routine.⁸ Accordingly, the number of research cores offering LC/MS
55 untargeted metabolomics services has increased dramatically over the last decade.⁹ The
56 conventional workflows used by most research facilities, however, essentially sidestep the issue
57 of experimental coverage.¹⁰ Their experimental output is a long list of signals or features,
58 without thorough annotation. The data sets are either mined in a targeted fashion for specific
59 metabolites with known retention times and fragmentation patterns, or only the small subset of
60 signals that have a statistically significant difference between sample classes are further
61 investigated.¹¹ For many of these signals altered between sample classes, further investigation
62 does not lead to identification because their accurate mass and fragmentation patterns do not

63 match the accurate mass and fragmentation patterns of any known reference standard in
64 metabolomic databases.¹² Although it is common to refer to these unmatched signals as
65 "unknown metabolites", rarely is such a designation justified. Signals associated with
66 contaminants, artifacts, and many adducts also do not return matches from metabolomic
67 databases. These possibilities and others must be ruled out before gaining confidence that a
68 signal is a bona fide, unique metabolite with an unknown structure.

69 The goal of the current study was to accurately enumerate contaminants, artifacts, and
70 degeneracies at the systems level within the same data set to get an upper estimate of the
71 number of unique metabolites detected in a representative LC/MS-based metabolomics
72 experiment. For the purposes of this work, contaminant refers to a detected signal that does not
73 originate from the biological sample being measured (e.g., solvent impurities and plastic
74 leechables). Artifacts refer to features detected due to informatic error. As an example, artifacts
75 can be caused by baseline fluctuations and poorly resolved components.^{23,24} Finally,
76 degeneracy refers to multiple signals arising from a single analyte. There are many causes of
77 degeneracy including: fragmentation, analyte adduction with various charge carriers (e.g., a
78 proton, sodium, potassium, etc.), and the detection of naturally occurring isotopes (e.g., ¹³C,
79 ¹⁵N, etc.).^{16,17,25} A final, largely under-annotated source of degeneracy is the adduction of an
80 analyte with other species present, including other analytes or the chemical background.

81 Although some degenerate relationships are well known and commonly annotated with
82 the approaches described above, the prevalence of many degenerate relationships has not
83 been previously estimated.¹⁴ Here we introduce and apply an approach that recovers
84 relationships implied by the experimental data, rather than relying on a hypothetical
85 predetermined list as is typically done in metabolomics. The approach allows for more
86 comprehensive annotation, especially in the case of under-annotated adducts that may be
87 specific to a single laboratory or experiment.

88 RESULTS AND DISCUSSION

89 *Generating a representative untargeted metabolomic data set*

90 In untargeted metabolomics, signals are often referred to as features, a convention we will
91 follow here. A feature is a detected ion with a peak shape, unique m/z , and retention time. To
92 estimate the number of unique analytes detected in a representative untargeted metabolomic
93 data set, we set out to annotate three types of features: (i) degenerate features, (ii) contaminant
94 features, and (iii) artifactual features. This work attempts to accurately enumerate each of these
95 types of features simultaneously for the same data set. We annotated degenerate features by
96 using mz.unity and a new contextual approach to find degeneracies implied by the data. We
97 annotated contaminant features and artifactual features by using the credentialing approach.²⁶ A
98 requirement of the credentialing approach is uniform ^{13}C -labeling. Given that there are
99 convenient and well-established methods to culture *E. coli* on a uniformly labeled carbon
100 source, we chose to focus our work on *E. coli*. Although mz.unity and credentialing were
101 previously introduced for the targeted analysis of features, here we have developed these
102 resources for systems-level studies so that they can be utilized simultaneously to assess
103 analyte content in untargeted metabolomic data sets.

104 Metabolites from *E. coli* cells were extracted and analyzed with an LC/MS-based
105 untargeted metabolomics platform, as detailed in Methods. In brief, metabolite extraction was
106 achieved by using a combination of methanol, acetonitrile, and water. Extracted metabolites
107 were separated with reversed-phase chromatography prior to being analyzed in positive polarity
108 by a Q Exactive Plus mass spectrometer. These experimental methods (or variations thereof)
109 are commonly applied in untargeted metabolomics.^{27–29} To process the resulting LC/MS data,
110 we employed a custom informatic workflow (Figure 1). The workflow used an iterative, two-
111 phase peak detection process. An in-house model-based feature detection algorithm was run on

112 each of five individual replicates. Many of the resulting features are inconsistent between
113 replicates due to subtle differences in the chromatograms from each file. It is common for some
114 peaks to go undetected, or some peaks to be integrated differently between runs.³⁰ These
115 errors make further analysis challenging because a one-to-one feature grouping cannot be
116 specified between replicates, and the established groups contain artificial variation in feature
117 areas. To refine the features detected in the five replicates, we utilized the Warpgroup
118 algorithm.³⁰ Warpgroup considers all files in concert to identify “consensus features”, a set of
119 feature integrations supported by all replicates. The result is a near one-to-one matching of
120 features between samples (Figure 2A-B) and decreased variation introduced by informatic
121 processing (Figure 2C-D). The Warpgroup refined feature detection is highly sensitive, allowing
122 the recovery of features that, when processed in isolation, would be challenging to detect
123 (Figure 2E). Here, we retained only features with a signal-to-noise ratio >5 and a coefficient of
124 variation <0.5 after Warpgrouping. This resulted in 25,230 high-quality features in our
125 representative data set. It is worth mentioning that analysis of our data set with the standard
126 XCMS software resulted in the detection of more features compared to Warpgroup (see
127 Supplementary Table 1). These data show that our informatic methods are not contributing to
128 atypically high feature counts.

129 We note that there is no universally accepted experimental platform for untargeted
130 metabolomics at this time. The extraction techniques, chromatography, mass spectrometers,
131 and peak detection algorithms used vary between laboratories and are often multiplexed.^{31,32}
132 However, it is routine to detect tens of thousands of signals from a biological sample in most
133 LC/MS experiments.^{33,34} Our detection of 25,230 consensus features from five replicates
134 resulted in a data set with complexity that is typical of an untargeted metabolomics experiment.

135 ***Simple annotations***

136 As a first step to place an upper bound on the number of unique metabolites detected in our
137 experiment, we performed a background subtraction. Specifically, we filtered features that were
138 not at least two-fold higher than the signal detected in extraction blanks. These features
139 represent contaminants or artifacts that are introduced during the sample extraction or data-
140 processing steps. This reduced our list of 25,230 features to 12,797 (Figure 3A). Notably, this
141 approach was developed to limit the number of artifact features returned by other processing
142 software.

143 Next, we set out to annotate degenerate features (i.e., those features arising from the
144 same analyte). We started our analysis by identifying simple relationships that are already
145 commonly annotated in untargeted metabolomics.^{14,35–37} This included degeneracy due to
146 carbon and other isotopes as well as common adducts and neutral losses. Annotations were
147 made by using mz.unity, and degenerate features were grouped together.¹⁷ Because features
148 within the same group arise from the same analyte, the number of “feature groups” provides a
149 much better estimate of the maximum number of unique analytes detected in an experiment
150 than the number of total features (Table 1 and Figure 3 B-C). In our subsequent descriptions,
151 we will therefore transition from counting features to counting feature groups. A feature for
152 which no degeneracy has been identified constitutes its own feature group, which we refer to as
153 a singlet. Figure 3B shows the progressive decrease in the number of feature groups as
154 isotopes, common charge carriers, and common neutral losses are annotated.

155 When isotopes, common charge carriers, and neutral losses are annotated, the number
156 of feature groups decreases from 12,797 to 7,318. We note that currently employed annotation
157 approaches end here with the identification of simple relationships (see vertical line in Figure
158 3B). These results might suggest that there are as many as 7,318 unique analytes detected in
159 the sample, but two observations suggested that much degeneracy still remained unannotated
160 in our *E. coli* data set. First, about 50% of our feature groups still contain only a single feature

161 (i.e., singlets with no detected relationships). Although in some cases singlets result from low-
162 abundance analytes with no natural isotopes detected above noise level, the prevalence of
163 singlets suggested that additional relationships remained unannotated. Second, we also know
164 that the set of relationships annotated thus far are only a small subset of the possible
165 degeneracies. A recent targeted study of glutamate demonstrated that many additional,
166 complex sources of degeneracy can exist in LC/MS-based metabolomics that are not currently
167 annotated with existing informatic resources.¹⁷ Glutamate was found to produce over 100
168 spectral peaks and exhibited complex adduct formation. Our objective was to comprehensively
169 characterize these additional sources of degeneracy within a data set (*E. coli*) for the first time.

170 ***Homo and hetero multimers***

171 We then expanded our search for degenerate relationships to complex adducts (i.e., two
172 or more species non-covalently bound to one another, such as dimers, trimers, etc.). Our search
173 included analytes adducted with themselves (homo-relationships), as well as analytes adducted
174 with different analytes (hetero-relationships). We considered all coeluting features as potential
175 multimer partners evaluating all [m, z] values as possible adduct formers. The charge state was
176 specified based on observed isotopes, or assumed to be a charge state of 1. As our conditions
177 generally form ions with a single charge, we balance the +2 charge from the observed ions with
178 the loss of a proton [1.00783, +1] for each multimer. Thus, a complex hetero-relationship
179 between three detected features will satisfy: $[m_1, z_1] + [m_2, z_2] - [1.00783, 1] = [m_3, z_3]$. Grouping
180 these detected complex adducts reduced the number of feature groups in our data set to 3,400
181 (see “multimers” bar in Figure 3B-C).

182 ***Frequent intrinsic relationships show previously unannotated degeneracy***

183 Current annotation approaches in untargeted metabolomics face the major challenge of
184 determining the specific relationships to search for. While some relationships are well known

185 and occur ubiquitously (such as the commonly annotated sodium or potassium adducts),
186 constraining annotation to only these is significantly limiting. Other degenerate relationships are
187 specific to experimental methodologies or the materials and reagents used during the analysis.
188 Since there is no way to determine these relationships *a priori*, they have gone unannotated to
189 date. Here we introduce an informatic approach to find data set wide, experimentally unique
190 relationships that are implied by their context in the data. We then estimate their prevalence
191 within our *E. coli* data set.

192 Common adducts and fragments will always coelute with the original analyte and will
193 occur multiple times throughout the run.¹⁷ We leverage this fact and recover “frequent intrinsic
194 relationships” by performing a frequency analysis of mass differences between all pairs of
195 features eluting within one second of each other.¹⁶ Unrelated but coeluting analytes will exhibit
196 mass spacing that is random and, as such, will not be enriched in the frequency distribution.
197 Thus, frequently occurring mass differences represent probable degenerate relationships. Mass
198 differences were calculated assuming a charge state of 1, a simplification that limits the analysis
199 to relationships that do not include a charge-state conversion. A Gaussian kernel density
200 estimation was performed on the observed mass differences with a bandwidth of 0.00001 Da
201 (our observed scan-to-scan mass error) (Figure 4A). The heights of the local maxima represent
202 the frequency and mass dispersion of each mass difference. Mass differences that are frequent
203 and similar in mass will have large density estimates. The 24 most frequently observed mass
204 differences are listed in Table 2.

205 The effectiveness of the approach was confirmed by the recovery of two commonly
206 known relationships as the most frequent relationships in our data set: the exchange of H⁺ and
207 Na⁺ and the exchange of Na⁺ and NH₄⁺. This result indicated that the analysis of frequent
208 intrinsic relationships offers novel insight into the nature of features detected in metabolomic
209 data sets. Notably, the approach returned a multitude of relationships that had not been

210 included in our prior searches. These commonly occurring relationships are likely adducts or
211 fragments, and may be specific to our sample or experimental equipment/materials. Figure 4B
212 shows the peak pairs observed with mass difference [23.0760, 0] throughout the data set.

213 We recognize that the recovery of frequent intrinsic relationships can also return
214 relationships between commonly coeluting, non-degenerate analyte pairs. Fully saturated and
215 partially unsaturated lipids, for example, commonly coelute and have a mass difference of
216 [2.0156, 0] (H_2).³⁸ We observed 176 occurrences of such a mass difference in our experiment.
217 To minimize the risk of grouping unrelated features, we removed relationships with mass
218 differences smaller than 15 Da and we applied two frequency cutoffs to illustrate the possible
219 range of degeneracy. The conservative cutoff annotated and grouped frequent intrinsic
220 relationships occurring more than 200 times (see bar labeled “commons n>200” in Figure 3B-
221 C), while the aggressive cutoff annotated and grouped frequent intrinsic relationships occurring
222 more than 50 times (see bar labeled “commons n>50” in Figure 3B-C). The inclusion of frequent
223 intrinsic relationships in our data set annotation reduced the number of feature groups to 5,281
224 or 3,769, depending on the cutoff.

225 ***Situational adducts due to background ions contribute significantly to degeneracy***

226 To further expand the scope of our annotation, we considered a source of adduct ions that are
227 present throughout the run: the chemical background. These ions lack a chromatographic peak
228 shape, but they are detected throughout the experiment due to the ionization of solvents, their
229 additives, or any contaminants present. Because the background ions coelute with every
230 feature, it is reasonable to expect that they will produce many adducts. We refer to adducts
231 between analytes and other presently observed species (such as background ions) as
232 “situational adducts”.

233 A low-mass spectrum was collected, deisotoped, and background ions appearing at
234 intensities higher than 200,000 were used as potential participants in situational adduct
235 formation (Figure 5). Annotation of the identified situational adducts reduced our number of
236 feature groups to approximately 3,000 (see bar labeled “background” in Figure 3B-C). This
237 significant reduction in feature groups indicates that background ions are indeed a major source
238 of feature inflation in our experiment. We also note that annotation of situational adducts
239 reduced the number of feature groups containing only a single feature (i.e., singlets) to 1,288.

240 ***Background ions give rise to some frequent intrinsic relationships***

241 Some frequent intrinsic relationships that we detected are indicative of novel adduction or
242 fragmentation phenomena in our untargeted metabolomic data set, and we were interested in
243 the origin of these unknown relationships. We speculated that some of the frequent intrinsic
244 relationships that we discovered were the result of analyte adduction with the chemical
245 background described above. In the simplest of cases, we found that some frequently occurring
246 mass-to-charge differences between features corresponded to the mass-to-charge values of
247 background ions. However, we also noted that mass differences in the background ions were
248 found in features. This indicated that a single analyte formed adducts with multiple background
249 ions (Figure 5 and Figure 6) and therefore multiple situational adducts were detected for the
250 same analyte. As the spacings between the background ions determine the spacings in the
251 situational adduct features, we expect these repeatedly occurring spacings to be returned as
252 frequent intrinsic relationships. Inspecting the returned frequent intrinsic relationships, we found
253 several mass differences that also appear in the chemical background. This result is an
254 additional confirmation of the effectiveness of frequent intrinsic relationship discovery and
255 suggests that chemical background is a large source of feature inflation.

256 We also performed formula decomposition on the frequent intrinsic relationships to
257 further elucidate their origins. Interestingly, chemical formula CH_2 , C_2H_4 , and C_3H_6 were found in
258 the frequent intrinsic relationships exhibited by the chemical background. Additional analysis of
259 the background ions indicated that they were an alkyl amine series. These alkyl amine species
260 are known to form strong adducts and are commonly found as contaminants in alcohol
261 solvents.³⁹ We note that our laboratory has never performed ion-pairing experiments and the
262 source of these reagents was solvent impurity as indicated by the series rather than sole
263 presence of triethylamine. In developing our methods, we attempted to find solvents with the
264 lowest possible levels of chemical background (Burdick & Jackson brand purchased from
265 Honeywell). Unfortunately, alkyl amines seem to be ubiquitous in methanol and isopropanol
266 LC/MS solvents.

267 ***Removing artifacts and contaminants by credentialing***

268 The degenerate relationships that we annotated above led to a striking reduction in the number
269 of feature groups, indicating that fewer than 15% of the total 25,230 features that were detected
270 in *E. coli* correspond to unique analytes. Even after this extensive annotation process, however,
271 two sources of feature inflation remained in artifacts and contaminants. We applied an
272 alternative experimental approach called credentialing to filter these features associated with
273 artifacts and contaminants. The credentialing process introduces an isotopic signature into
274 biological analytes during *E. coli* growth.²⁶ Features in our data set displaying this isotopic
275 signature are deemed “credentialed”, as they are known to be of *E. coli* origin. In contrast,
276 features that do not display this isotopic signature are annotated as artifacts or contaminants.
277 Credentialing does not rely on any of the relationship annotation approaches that we described
278 above, and is thus an orthogonal and highly complementary approach to data analysis.

279 We first filtered non-credentialed features from the raw data set on the basis of isotopic
280 signatures. The resulting set of features is free of artifacts, noise, and contaminants. This
281 process returned 2,462 high-quality, credentialed features. We then took these credentialed
282 features through the same annotation process as the full data set to remove degeneracy.
283 Annotation of degeneracy reduced the estimated number of unique *E. coli* analytes being
284 measured to 832 (Figure 3C).

285 **CONCLUSION**

286 Detecting tens of thousands of LC/MS features from biological samples is typical in untargeted
287 metabolomics, however, to date it has been unclear how many unique metabolites are actually
288 being profiled. Our work here evaluated one representative untargeted metabolomics data set
289 from *E. coli* to set an upper bound on the number of unique metabolites being measured. By
290 using a new context-driven approach to identify degenerate features arising from the same
291 metabolite, we determined that the ~25,000 features detected in our experiment corresponded
292 to fewer than 2,961 unique analytes. An orthogonal and complimentary approach using
293 credentialing isotope signatures to identify artifacts and contaminants similarly reduced the
294 number of unique analytes detected. Out of the total ~25,000 features detected, only 832
295 passed both our degeneracy and credentialing filters.

296 We wish to emphasize that our work is unrelated to the size of the *E. coli* metabolome
297 and should not be interpreted as an indication of the total number of intracellular metabolites
298 present. There are certainly more than 832 *E. coli* metabolites.⁴⁴ The purpose of our work is
299 only to assess how many unique metabolites are being measured in a representative
300 untargeted metabolomics experiment. Additionally, we note that our context-driven analysis of
301 degeneracy is not exhaustive. Relationships that are uncommon and not indicated by
302 background ions remain unannotated and may further reduce the number of unique analytes

303 detected. Notwithstanding, our results suggest that there are an order of magnitude more
304 features than unique metabolites in untargeted metabolomics experiments. This has important
305 implications for designing untargeted metabolomics experiments and influences strategies for
306 interpreting the data produced before establishing metabolite identifications.

307 **METHODS**

308 ***Materials***

309 U-¹³C-D-glucose was purchased from Cambridge Isotope Laboratories Inc. (Andover, MA). *E.*
310 *coli* strain K12, MG1655 was purchased from ATCC (Manassas, VA). Lennox LB broth powder
311 and 5x M9 salts were purchased from Sigma-Aldrich (St. Louis, MO). Cell culture was
312 performed with ultrapure water provided by a Milli-Q system (Millipore). LC/MS grade, Burdick &
313 Jackson brand water, acetonitrile, methanol, and isopropanol were purchased from Honeywell
314 (Morris Plains, NJ). Cortecs T3 reversed phase UPLC columns and column guards were
315 purchased from Waters Corporation (Milford, MA).

316 ***Generating credentialed samples***

317 *E. coli* was grown in a rotary shaker at 37 °C and 300 rpm as previously described (Mahieu et
318 al., 2014). A 100 mL volume of M9 minimal media was used with a glucose concentration of 2
319 g/L. Two cultures were grown in parallel, one using natural abundance glucose and a second
320 using U-¹³C-glucose as the only carbon source. Cultures were grown to OD₆₀₀ = 0.7, at which
321 point they were harvested.

322 For harvest, flasks were removed from the shaker and placed on ice. The contents of
323 each flask were pipetted into 50 mL conical tubes and centrifuged at 3200g and 4°C for 10
324 minutes. The supernatant was decanted and remaining media was gently rinsed off the top of
325 the pellet with 0.5 mL of water. Conical tubes were then placed in liquid nitrogen and lyophilized

326 for 24 hours, or until dry. This powdered, credentialed *E. coli* standard was then extracted to
327 generate samples for untargeted metabolomic analysis.

328 Several replicate extractions were performed in parallel by using a previously described
329 method.²⁶ Briefly, five 2.5 mg samples of each ¹²C and ¹³C material were weighed out, while two
330 empty tubes were included as extraction blanks. To these, 1,000 μ L of 2:2:1
331 methanol:acetonitrile:water was added, followed by three freeze-thaw cycles with sonication and
332 vortexing. After centrifugation, the supernatant was vacuum concentrated and reconstituted in
333 100 μ L of 1:1 acetonitrile:water with internal standards. From these extracts, three samples
334 were aliquoted for LC/MS analysis: natural abundance extract, a mix of 1:1 natural abundance
335 extract and ¹³C extract, and the blank extract.

336 ***Data set generation***

337 Each sample was analyzed five times as analytical replicates. The untargeted LC/MS data set
338 was generated in positive polarity on a Q Exactive Plus mass spectrometer with a HESI II
339 source coupled to a Dionex 3000RSLC. The data set was collected with the following settings:
340 aux gas, 5; sheath gas, 35; sweep gas, 2; capillary temperature, 300 °C; aux gas temperature,
341 200 °C; spray voltage, 3.5 kV; needle diameter, 34 ga; s-lens, 75 V; mass range, 100–1500 Da;
342 resolution 70,000; micro scans, 1; max injection time; 100 ms; automatic gain control target:
343 1e6. Reversed-phase chromatography was performed with the Waters Cortecs T3 (2.1mm x
344 50mm, 1.6 μ m) column at a flow rate of 300 μ L/min and a column temperature of 50 °C. Solvents
345 were: A, water + 5mM ammonium acetate + 5 μ M ammonium phosphate; B, 9:1
346 isopropanol:methanol + 5mM ammonium acetate + 5 μ M ammonium phosphate. An injection
347 volume of 2 μ L was used with a linear gradient of (minutes, %A): 0, 100; 28, 0; 30, 0; 30, 100;
348 35, 100.

349 Chromatographic features were detected by using a set of in-house algorithms. Mass
350 traces were retained if they were longer than 10 scans, excluding missing peaks. Baselines for
351 each mass trace were calculated by using the iterative restricted least squares method from the
352 baseline R package. Model based peak detection was performed by using the skew normal
353 distribution as a model peak distribution. This process resulted in a set of features detected in
354 each replicate run. Features were grouped by mass and retention time using a density based
355 method. Retention time drift and mass drift were corrected by fitting a loess curve of degree 2 to
356 the distance from the mean value of each group against the mean retention time of each group.

357 Subtle variations from run to run cause many features to be integrated differently and
358 sometimes not integrated in each file. Further, closely eluting peaks often lead to incorrectly
359 grouped features. To resolve these missing values, refine the individual datasets and get a set
360 of detected peaks consistent with all replicate runs, we applied the Warpgroup algorithm.³⁰
361 Warpgroup is available at <https://github.com/nathaniel-mahieu/warpgroup>. Warpgroup takes as
362 input the raw data and each file's detected features combining them to output a set of
363 consensus features. Parameters: `sc.aligned.lim`, 9; `pct.pad`, 0.1; `min.peaks`, 3. Of the detected
364 peaks we retained only features with a signal-to-noise ratio >5 and a coefficient of variation <0.5
365 after Warpgrouping. This resulted in 25,230 "high-quality" features in our representative data
366 set.

367 This consensus data set set is the standard output of an untargeted metabolomics
368 experiment. As such, it was taken as a representative dataset for annotation of detected
369 signals.

370 ***Mz.unity based annotation***

371 Mz.unity was applied to the dataset to detect mass and charge ($[m, z]$) relationships between
372 eluting signals derived from a single analyte.¹⁷ We use $[m, z]$ to denote the mass and charge of

373 a species, where both are specified as opposed to m/z where the two are convolved. These
374 searches find sets of features that have $[m, z]$ s differing by a specific amount. Differences are
375 specific to relationships, for example, loss of ^{12}C and gain of ^{13}C ($[+1.003355, 0]$), or loss of
376 water ($[-18.01057, 0]$).

377 Searches were first performed for the following relationships: isotopes, common charge
378 carriers, common neutral losses, and common adducts. We then searched for dimers between
379 coeluting features. The dimer search posits each eluting $[m, z]$ as a possible adduct former.
380 The charge state was specified based on observed isotopes, or assumed to be a charge of 1.
381 As dimers are normally formed with a charge from only one constituent, we also assumed the
382 loss of a proton $[1.00783, +1]$ for each pair.

383 Mz.unity is available at <https://github.com/nathaniel-mahieu/mz.unity>.

384 ***Frequent intrinsic relationships***

385 Groups of features eluting within 1 second of each other were taken, and their pairwise $[m, z]$
386 differences were calculated after assuming a charge state of 1. A Gaussian kernel density
387 estimation was performed on the mass differences with a bandwidth of 0.00001 Da (our
388 observed scan-to-scan mass error). Local maxima of the density estimate were detected along
389 with the estimated density at those locations. The heights of the local maxima represent the
390 frequency and mass dispersion of each mass difference. Mass differences that are more
391 frequent and more similar in mass will have larger density estimates.

392 We took enriched mass differences larger than 15 Da and occurring more than 50 times
393 throughout the dataset into the mz.unity search.

394 ***Situational adducts***

395 Background ions that lack a chromatographic peak shape are an ever-present set of species
396 that often form adducts with eluting analytes. These situational adducts are then detected as
397 features having a chromatographic peak shape. A low mass background spectrum was
398 collected, containing detected ions above 50 Da. This spectrum was deisotoped and
399 background species appearing at higher than 200,000 intensity were used to seed possible
400 adduct relationships. The [m, z]s of each background peak were included in the dimer search,
401 as above after specifying the charge state based on observed isotopes or assuming a charge of
402 1.

403 ***Credentialing***

404 A high-confidence set of features were recovered from the $^{12+13}\text{C}$ dataset by applying version
405 3.0 of the credentialing algorithm, which is available at <https://github.com/pattilab/credential>.
406 Credentialing searches for pairs of peaks that have precise isotopic spacing expected from U-
407 ^{12}C and U- ^{13}C analytes.²⁶ This provides a filter against many forms of noise, contaminants, and
408 artifact features. Credentialing was run with the parameters: ppmwid, 8; rtwid, 1.2; cd, 1.00335;
409 mpc, c(12, 120); ratio, 1; ratio.lim, 0.1; maxnmer, 4. Credentialed features from the $^{12+13}\text{C}$ data
410 set were then matched to the ^{12}C dataset by applying retention time and mass correction as
411 above before grouping.

412

413 **FIGURE LEGENDS**

414 Figure 1: Our informatic workflow. Raw data were processed with in-house algorithms to first
415 identify high-quality, consensus features (i.e., recurring features between replicates) and
416 discriminate against processing artifacts. This consensus data set was further characterized by
417 mz.unity (to estimate signal degeneracy) and credentialing (to estimate contaminants and
418 artifacts).

419 Figure 2: An overview of the consensus data set. (A) The base peak chromatogram of a
420 representative run. The number of features detected during each second is overlaid. (B) The
421 number of features detected in each group before (pink) and after (green) Warpgroup.
422 Inconsistencies are resolved by Warpgroup. (C) The within group CVs of peak areas is
423 decreased by Warpgroup. (D) The within group CVs of peak width are decreased by
424 Warpgroup. (E) Several representative features detected by the informatic workflow. The
425 estimated baseline is plotted in red.

426 Figure 3: Plotting the maximum number of unique analytes detected throughout the steps of our
427 annotation process. (A) Removal of features occurring in the blank. (B) Features are grouped as
428 additional relationships are annotated. This reduces the maximum number of unique analytes.
429 When a feature group contains multiple features, it is shown in green. When a feature group
430 contains only a single feature (i.e., is a singlet), then it is shown in pink. Relationships from left
431 to right: no relationships; isotopes; charge carriers; neutral losses; complex dimers (homo and
432 hetero); frequent intrinsic relationships; situational adducts (background). (C) Similar annotation
433 of features that were credentialed.

434 Figure 4: Detection of frequent intrinsic relationships. (A) The Gaussian kernel density of all
435 pairwise peak relationships in the data set. Inset is a zoomed-in section around 14 Da. Known
436 relationships are labeled with a formula. Unknown relationships are labeled with mass and

437 charge transitions [m, z]. (B) Peak pairs of the recovered frequent intrinsic relationship [23.0760,
438 0] plotted in mass/charge and retention time (points). Line segments connect pairs with the
439 specified spacing.

440 Figure 5: Situational adducts. (A) The persistent background spectrum observed in this
441 experiment. The three indicated background peaks have mass spacings that correspond to a
442 methylene group. These are likely an alkyl amine series with carbon numbers 5, 6, and 7. When
443 these background species adduct with an analyte, situational adducts are formed. (B) An
444 example of a situational adduct forming between background ion 102.1280 (a six carbon alkyl
445 amine) and an eluting analyte. This process likely occurs with all three alkyl amine species
446 throughout the run, giving rise to the frequent intrinsic relationships of mass 14.0157 (see Table
447 2, Row 8).

448 Figure 6: Schematic showing how background ions give rise to frequent intrinsic relationships.
449 Analyte A is detected as an adduct of each background ion (B_1 and B_2). The spacing between
450 the adducts ($A+B_1-H$ and $A+B_2-H$) is equal to the spacing between the background ions.

451

452

Stage	Groups with more than one feature		Singlets	
	All Features	Credentialed Features	All Features	Credentialed Features
Blank Subtracted	0	0	12797	2462
Isotopes	3986	1066	5071	1326
Charge Carriers	3620	1137	4384	992
Neutral Losses	3640	1174	3678	790
Multimers	3400	1117	3381	712
Commons n>200	2809	1063	2472	495
Commons n>50	2149	864	1620	353
Background	1673	659	1288	233

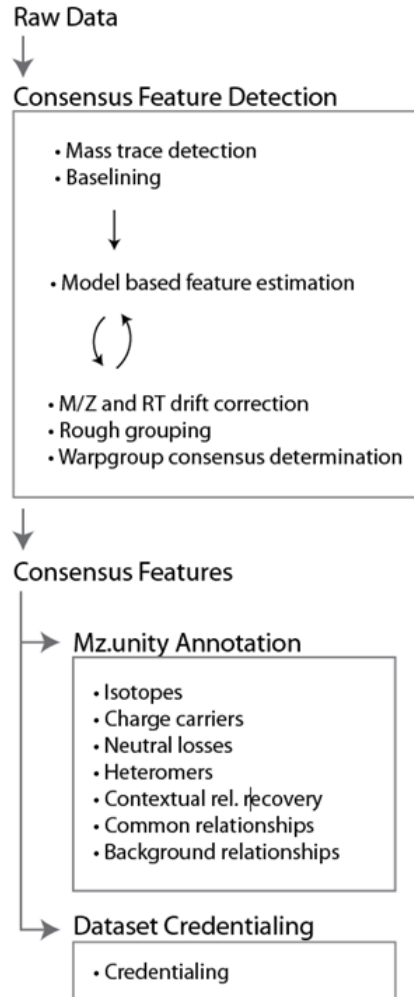
453 **Table 1:** A breakdown of the analyte number observed after each annotation step.

454

<u>Δ Mass</u>	<u>Δ Charge</u>	<u>Density</u>	<u>Known Species</u>
21.9820	0	60.4	$H^+ \leftrightarrow Na^+$
4.9554	0	55.2	$NH_4^+ \leftrightarrow Na^+$
23.0760	0	33.6	
18.0107	0	32.5	H_2O
17.0266	0	30.5	NH_3
28.0314	0	26.7	C_2H_4
45.0580	0	23.4	C_2H_7N
14.0157	0	23.2	CH_2
65.1230	0	19.6	
87.1046	0	18.2	$C_5H_{13}N$
42.0470	0	16.6	C_3H_6
44.0262	0	15.3	C_2H_4O
39.9926	0	13.3	C_2O
7.1020	0	13.1	
15.9740	0	13.0	$K^+ \leftrightarrow Na^+$
70.0783	0	12.5	
29.0518	0	11.6	
36.0713	0	11.3	
15.9949	0	10.1	
1.9967	0	9.3	$^{41}K \leftrightarrow ^{39}K$
56.0627	0	9.3	
12.9952	0	8.7	
35.0373	0	8.7	
20.9292	0	8.5	$NH_4^+ \leftrightarrow K^+$

455

456 **Table 2:** Recovered frequent intrinsic relationships. Not all recovered relationships shown
 457 were used in the annotation. The local maxima of the density are ordered by the number of
 458 occurrences. These frequently occurring differences are good candidates for peak
 459 relationships. Several well-known relationships are present, including alternative charge
 460 carriers at the top of the list.



461

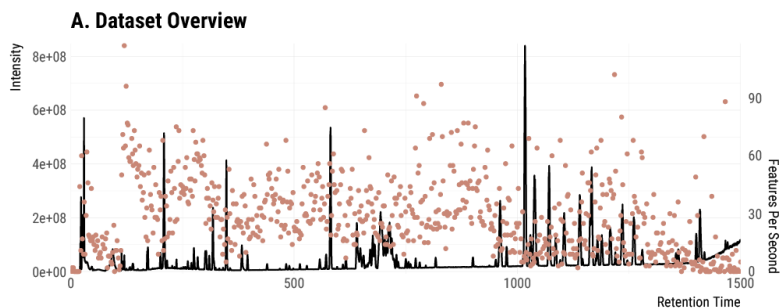
462 Figure 1: Our informatic workflow. Raw data were processed with in-house algorithms to first
463 identify high-quality, consensus features (i.e., recurring features between replicates) and
464 discriminate against processing artifacts. This consensus data set was further characterized by
465 mz.unity (to estimate signal degeneracy) and credentialing (to estimate contaminants and
466 artifacts).

467

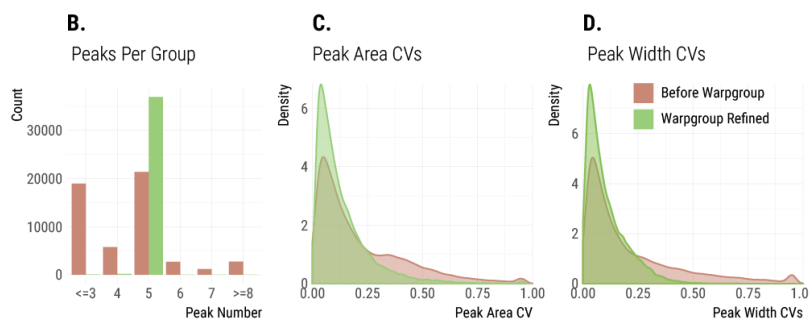
468

469

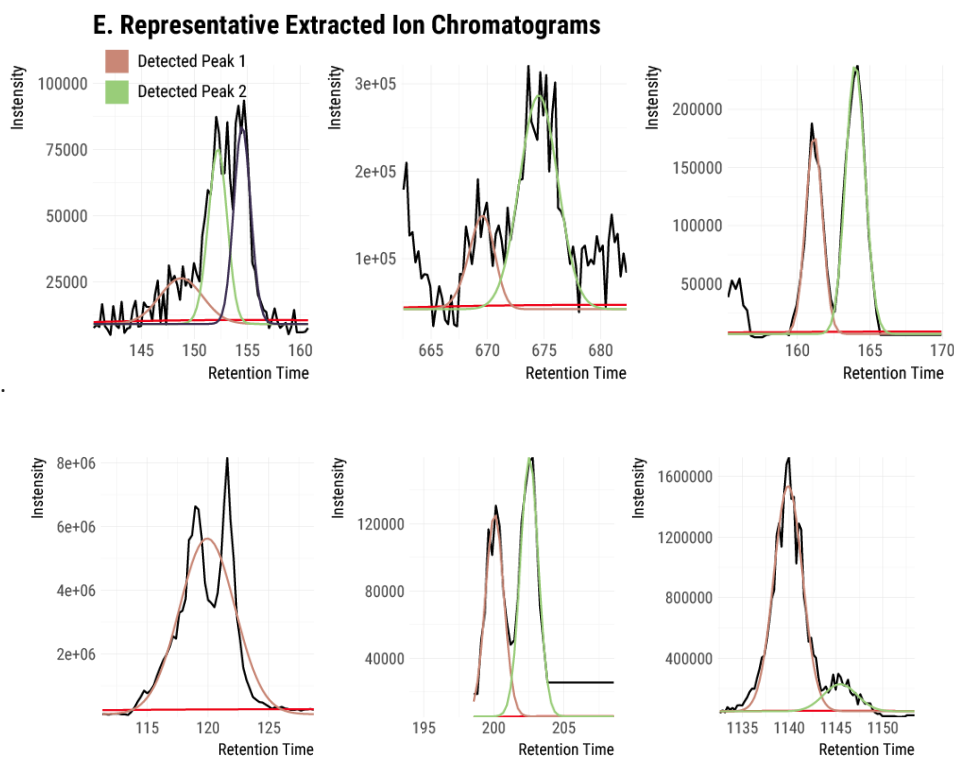
470



471

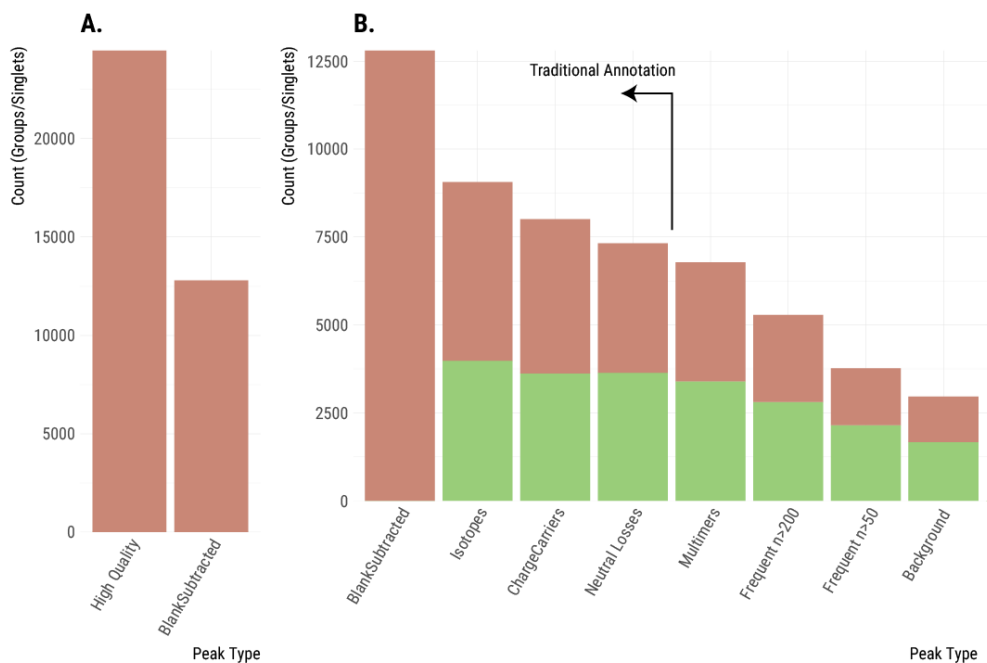


472



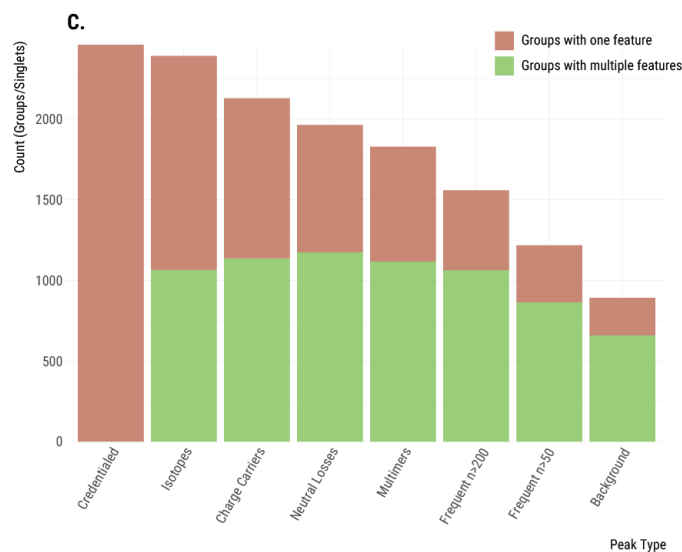
473

474 Figure 2: An overview of the consensus data set. (A) The base peak chromatogram of a
475 representative run. The number of features detected during each second is overlaid. (B) The
476 number of features detected in each group before (pink) and after (green) Warpgroup.
477 Inconsistencies are resolved by Warpgroup. (C) The within group CVs of peak areas is
478 decreased by Warpgroup. (D) The within group CVs of peak width are decreased by
479 Warpgroup. (E) Several representative features detected by the informatic workflow. The
480 estimated baseline is plotted in red.



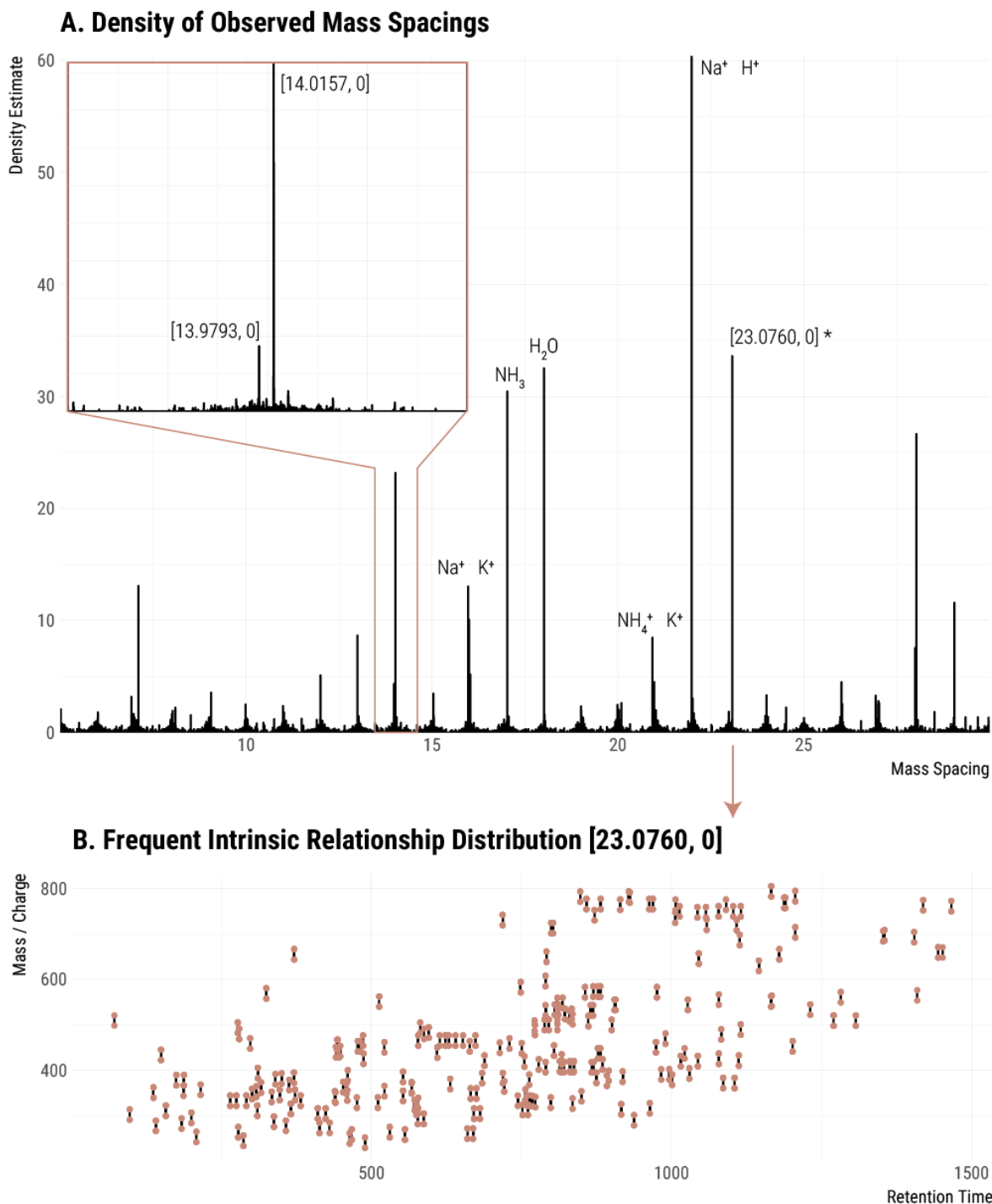
481

482



483

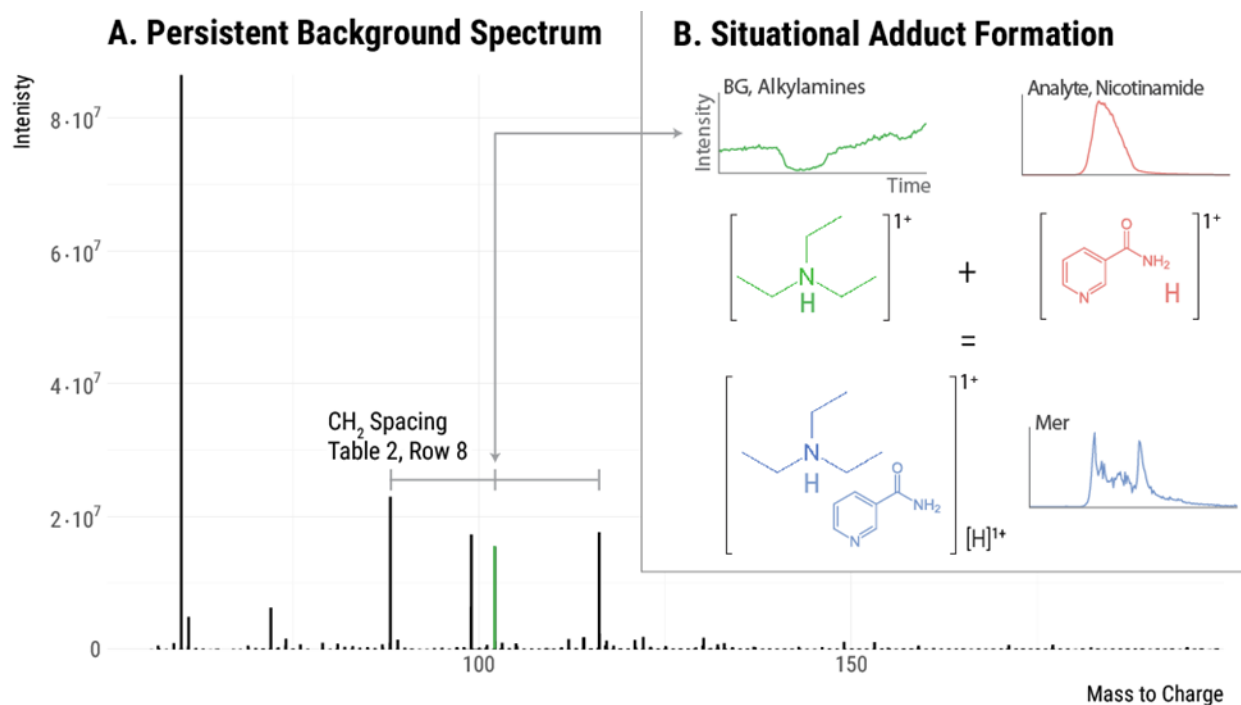
484 Figure 3: Plotting the maximum number of unique analytes detected throughout the steps of our
485 annotation process. (A) Removal of features occurring in the blank. (B) Features are grouped as
486 additional relationships are annotated. This reduces the maximum number of unique analytes.
487 When a feature group contains multiple features, it is shown in green. When a feature group
488 contains only a single feature (i.e., is a singlet), then it is shown in pink. Relationships from left
489 to right: no relationships; isotopes; charge carriers; neutral losses; complex dimers (homo and
490 hetero); frequent intrinsic relationships; situational adducts (background). (C) Similar annotation
491 of features that were credentialed.



492

493 Figure 4: Detection of frequent intrinsic relationships. (A) The Gaussian kernel density of all
494 pairwise peak relationships in the data set. Inset is a zoomed-in section around 14 Da. Known
495 relationships are labeled with a formula. Unknown relationships are labeled with mass and
496 charge transitions [m, z]. (B) Peak pairs of the recovered frequent intrinsic relationship [23.0760,
497 0] plotted in mass/charge and retention time (points). Line segments connect pairs with the
498 specified spacing.

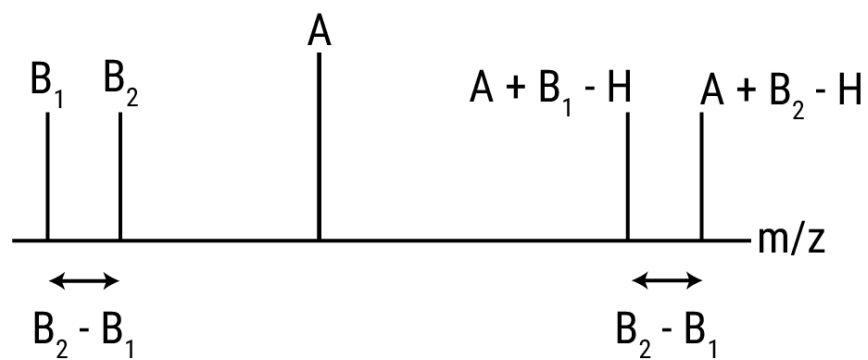
499



500

501 Figure 5: Situational adducts. (A) The persistent background spectrum observed in this
502 experiment. The three indicated background peaks have mass spacings that correspond to a
503 methylene group. These are likely an alkyl amine series with carbon numbers 5, 6, and 7. When
504 these background species adduct with an analyte, situational adducts are formed. (B) An
505 example of a situational adduct forming between background ion 102.1280 (a six carbon alkyl
506 amine) and an eluting analyte. This process likely occurs with all three alkyl amine species
507 throughout the run, giving rise to the frequent intrinsic relationships of mass 14.0157 (see Table
508 2, Row 8).

509



510

511 Figure 6: Schematic showing how background ions give rise to frequent intrinsic relationships.
512 Analyte A is detected as an adduct of each background ion (B₁ and B₂). The spacing between
513 the adducts (A+B₁-H and A+B₂-H) is equal to the spacing between the background ions.

514

515 References

- 516
- 517 1. Broeckling, C. D., Afsar, F. A., Neumann, S., Ben-Hur, A. & Prenni, J. E. RAMClust: A
518 Novel Feature Clustering Method Enables Spectral-Matching-Based Annotation for
519 Metabolomics Data. *Anal. Chem.* **86**, 6812–6817 (2014).
 - 520 2. Uppal, K. *et al.* Computational Metabolomics: A Framework for the Million Metabolome.
521 *Chem. Res. Toxicol.* **29**, 1956–1975 (2016).
 - 522 3. Xia, J., Sinelnikov, I. V., Han, B. & Wishart, D. S. MetaboAnalyst 3.0—making
523 metabolomics more meaningful. *Nucleic Acids Res.* **43**, W251–W257 (2015).
 - 524 4. Dunn, W. B. *et al.* Mass appeal: metabolite identification in mass spectrometry-focused
525 untargeted metabolomics. *Metabolomics* **9**, 44–66 (2013).
 - 526 5. Benton, H. P. *et al.* Autonomous metabolomics for rapid metabolite identification in global
527 profiling. *Anal. Chem.* **87**, 884–891 (2014).
 - 528 6. Stanstrup, J., Gerlich, M., Dragsted, L. O. & Neumann, S. Metabolite profiling and
529 beyond: approaches for the rapid processing and annotation of human blood serum mass
530 spectrometry data. *Anal. Bioanal. Chem.* **405**, 5037–5048 (2013).
 - 531 7. Tautenhahn, R. *et al.* An accelerated workflow for untargeted metabolomics using the
532 METLIN database. *Nat. Biotechnol.* **30**, 826–828 (2012).
 - 533 8. Mahieu, N. G., Genenbacher, J. L. & Patti, G. J. A roadmap for the XCMS family of
534 software solutions in metabolomics. *Curr. Opin. Chem. Biol.* **30**, 87–93 (2016).
 - 535 9. Sud, M. *et al.* Metabolomics Workbench: An international repository for metabolomics
536 data and metadata, metabolite standards, protocols, tutorials and training, and analysis
537 tools. *Nucleic Acids Res.* **44**, D463–D470 (2016).
 - 538 10. Cho, K., Mahieu, N. G., Johnson, S. L. & Patti, G. J. After the feature presentation:
539 technologies bridging untargeted metabolomics and biology. *Curr. Opin. Biotechnol.* **28**,
540 143–148 (2014).
 - 541 11. Patti, G. J., Yanes, O. & Siuzdak, G. Innovation: Metabolomics: the apogee of the omics
542 trilogy. *Nat. Rev. Mol. Cell Biol.* **13**, 263–269 (2012).
 - 543 12. Nikolskiy, I., Mahieu, N. G., Chen, Y.-J., Tautenhahn, R. & Patti, G. J. An untargeted
544 metabolomic workflow to improve structural characterization of metabolites. *Anal. Chem.*
545 **85**, 7713–7719 (2013).
 - 546 13. Lee, T. S., Ho, Y. S., Yeo, H. C., Lin, J. P. Y. & Lee, D.-Y. Precursor mass prediction by
547 clustering ionization products in LC-MS-based metabolomics. *Metabolomics* **9**, 1301–
548 1310 (2013).
 - 549 14. Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R. & Neumann, S. CAMERA: An
550 Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid
551 Chromatography/Mass Spectrometry Data Sets. *Anal. Chem.* **84**, 283–289 (2012).
 - 552 15. Scheltema, R. *et al.* Simple data-reduction method for high-resolution LC-MS data in

- 553 metabolomics. *Bioanalysis* **1**, 1551–1557 (2009).
- 554 16. Brown, M. *et al.* Mass spectrometry tools and metabolite-specific databases for molecular
555 identification in metabolomics. **134**, (2009).
- 556 17. Mahieu, N. G., Spalding, J. L., Gelman, S. J. & Patti, G. J. Defining and Detecting
557 Complex Peak Relationships in Mass Spectral Data: The Mz.unity Algorithm. *Anal.*
558 *Chem.* **88**, 9037–9046 (2016).
- 559 18. Chokkathukalam, A. *et al.* mzMatch-ISO: an R tool for the annotation and relative
560 quantification of isotope-labelled mass spectrometry data. *Bioinformatics* **29**, 281–283
561 (2013).
- 562 19. Zhou, R., Tseng, C.-L., Huan, T. & Li, L. IsoMS: Automated Processing of LC-MS Data
563 Generated by a Chemical Isotope Labeling Metabolomics Platform. *Anal. Chem.* **86**,
564 4675–4679 (2014).
- 565 20. Stupp, G. S. *et al.* Isotopic ratio outlier analysis global metabolomics of *Caenorhabditis*
566 *elegans*. *Anal. Chem.* **85**, 11858–11865 (2013).
- 567 21. de Jong, F. A. & Beecher, C. Addressing the current bottlenecks of metabolomics:
568 Isotopic Ratio Outlier AnalysisTM, an isotopic-labeling technique for accurate biochemical
569 profiling. *Bioanalysis* **4**, 2303–2314 (2012).
- 570 22. Bueschl, C. *et al.* A novel stable isotope labelling assisted workflow for improved
571 untargeted LC–HRMS based metabolomics research. *Metabolomics* **10**, 754–769 (2014).
- 572 23. Tong, H., Bell, D., Tabei, K. & Siegel, M. M. Automated data massaging, interpretation,
573 and e-mailing modules for high throughput open access mass spectrometry. *J. Am. Soc.*
574 *Mass Spectrom.* **10**, 1174–1187 (1999).
- 575 24. Zhu, J. & Cole, R. B. Formation and decompositions of chloride adduct ions, [M + Cl]⁻, in
576 negative ion electrospray ionization mass spectrometry. *J. Am. Soc. Mass Spectrom.* **11**,
577 932–941 (2000).
- 578 25. Brown, M. *et al.* Automated workflows for accurate mass-based putative metabolite
579 identification in LC/MS-derived metabolomic datasets. *Bioinformatics* **27**, 1108–12
580 (2011).
- 581 26. Mahieu, N. G., Huang, X., Chen, Y.-J. & Patti, G. J. Credentialing features: a platform to
582 benchmark and optimize untargeted metabolomic methods. *Anal. Chem.* **86**, 9583–9
583 (2014).
- 584 27. Cajka, T. & Fiehn, O. Toward Merging Untargeted and Targeted Methods in Mass
585 Spectrometry-Based Metabolomics and Lipidomics. *Anal. Chem.* **88**, 524–545 (2016).
- 586 28. Contrepois, K., Jiang, L. & Snyder, M. Optimized Analytical Procedures for the
587 Untargeted Metabolomic Profiling of Human Urine and Plasma by Combining Hydrophilic
588 Interaction (HILIC) and Reverse-Phase Liquid Chromatography (RPLC)–Mass
589 Spectrometry. *Mol. Cell. Proteomics* **14**, 1684–1695 (2015).
- 590 29. Ivanisevic, J. *et al.* Toward 'omic scale metabolite profiling: a dual separation-mass
591 spectrometry approach for coverage of lipid and central carbon metabolism. *Anal. Chem.*
592 **85**, 6876–6884 (2013).

- 593 30. Mahieu, N. G., Spalding, J. L. & Patti, G. J. Warpgroup: increased precision of
594 metabolomic data processing by consensus integration bound analysis. *Bioinformatics*
595 *btv564* (2015).
- 596 31. Vinayavekhin, N. & Saghatelian, A. in *Current Protocols in Molecular Biology Chapter 30*,
597 Unit 30.1.1-24 (John Wiley & Sons, Inc., 2010).
- 598 32. Wishart, D. S. Emerging applications of metabolomics in drug discovery and precision
599 medicine. *Nat. Rev. Drug Discov.* **15**, 473–484 (2016).
- 600 33. Melamud, E., Vastag, L. & Rabinowitz, J. D. Metabolomic Analysis and Visualization
601 Engine for LC–MS Data. *Anal. Chem.* **82**, 9818–9826 (2010).
- 602 34. Milne, S. B., Mathews, T. P., Myers, D. S., Ivanova, P. T. & Brown, H. A. Sum of the
603 Parts: Mass Spectrometry-Based Metabolomics. *Biochemistry* **52**, 3829–3840 (2013).
- 604 35. Daly, R. *et al.* MetAssign: probabilistic annotation of metabolites from LC-MS data using a
605 Bayesian clustering approach. *Bioinformatics* **30**, 2764–71 (2014).
- 606 36. Kessler, N. *et al.* ALLocator: An Interactive Web Platform for the Analysis of Metabolomic
607 LC-ESI-MS Datasets, Enabling Semi-Automated, User-Revised Compound Annotation
608 and Mass Isotopomer Ratio Analysis. *PLoS One* **9**, e113909 (2014).
- 609 37. Zeng, Z. *et al.* Ion Fusion of High-Resolution LC–MS-Based Metabolomics Data to
610 Discover More Reliable Biomarkers. *Anal. Chem.* **86**, 3793–3800 (2014).
- 611 38. Han, X., Yang, K. & Gross, R. W. Multi-dimensional mass spectrometry-based shotgun
612 lipidomics and novel strategies for lipidomic analyses. *Mass Spectrom. Rev.* **31**, 134–178
613 (2012).
- 614 39. Keller, B. O., Sui, J., Young, A. B. & Whittall, R. M. Interferences and contaminants
615 encountered in modern mass spectrometry. *Anal. Chim. Acta* **627**, 71–81 (2008).
- 616 40. Barbazuk, W. B. *et al.* The syntenic relationship of the zebrafish and human genomes.
617 *Genome Res.* **10**, 1351–8 (2000).
- 618 41. Hillier, L. D. *et al.* Generation and analysis of 280,000 human expressed sequence tags.
619 *Genome Res.* **6**, 807–28 (1996).
- 620 42. Masson, P., Alves, A. C., Ebbels, T. M. D., Nicholson, J. K. & Want, E. J. Optimization
621 and Evaluation of Metabolite Extraction Protocols for Untargeted Metabolic Profiling of
622 Liver Samples by UPLC-MS. *Anal. Chem.* **82**, 7779–7786 (2010).
- 623 43. Yanes, O., Tautenhahn, R., Patti, G. J. & Siuzdak, G. Expanding Coverage of the
624 Metabolome for Global Metabolite Profiling. *Anal. Chem.* **83**, 2152–2161 (2011).
- 625 44. Sajed, T. *et al.* ECMDDB 2.0: A richer resource for understanding the biochemistry of *E.*
626 *coli*. *Nucleic Acids Res.* **44**, D495–D501 (2016).
- 627