

1 **Novel subgroups of type 2 diabetes based on multi-Omics profiling: an IMI-**
2 **RHAPSODY Study**

3
4 Shiyong Li¹, Iulian Dragan⁴, Chun Ho Fung², Dmitry Kuznetsov⁴,
5 Michael K. Hansen¹², Joline W.J. Beulens^{5,6}, Leen M. 't Hart^{5,6,7,8},
6 Roderick C. Slieker⁷, Louise A. Donnelly⁹, Mathias J. Gerl¹⁰, Christian Klose¹⁰, Florence
7 Mehl⁴, Kai Simons¹⁰, Petra JM Elders¹¹, Ewan R. Pearson⁹,
8 Guy A. Rutter^{1,2,3*} and Mark Ibberson^{4*}

9
10 ¹Centre de Recherche du CHUM, and Faculty of Medicine, University of Montreal, QC,
11 Canada

12 ²Section of Cell Biology and Functional Genomics, Department of Metabolism, Diabetes
13 and Reproduction, Imperial College of London, du Cane Road, London W120NN, United
14 Kingdom

15 ³Lee Kong Chian School of Medicine, Nan Yang Technological University, Singapore.

16 ⁴Vital-IT Group, SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

17 ⁵Department of Epidemiology and Data Sciences, Amsterdam University Medical Center,
18 location VUmc, Amsterdam, The Netherlands

19 ⁶Amsterdam Public Health, Amsterdam, The Netherlands

20 ⁷Department of Cell and Chemical Biology, Leiden University Medical Center, Leiden,
21 The Netherlands

22 ⁸Department of Biomedical Data Sciences, Section Molecular Epidemiology, Leiden
23 University Medical Center, Leiden, The Netherlands

24 ⁹Division of Population Health & Genomics, School of Medicine, University of Dundee, UK

25 ¹⁰Lipotype GmbH Dresden Germany

26 ¹¹Department of General Practice and Elderly Care Medicine, Amsterdam Public Health
27 Research Institute, Amsterdam UMC–location VUmc, Amsterdam, the Netherlands

28 ¹²Janssen Research and Development, Philadelphia, PA

29
30 Address correspondence to

31 Guy A. Rutter guy.rutter@umontreal.ca, 514 890-8000, ext. 27081

32 Mark Ibberson mark.ibberson@sib.swiss, (41) 21 692 4084

33
34 Word count: 3823 Number of Figures: 3 Number of Tables: 1

35

36

37 **Abstract**

38 Type 2 diabetes is a complex, multifactorial disease with varying presentation and
39 underlying pathophysiology. Recent studies using data-driven cluster analysis have led
40 to a stratification of type 2 diabetes into novel subgroups based on six clinical
41 measurements. Whether these subgroups truly correspond to the underlying
42 phenotypic differences is nevertheless unclear. Here, we apply an unsupervised, data-
43 driven clustering method (Similarity Network Fusion) to characterize type 2 diabetes in
44 two independent cohorts involving 1,134 subjects in total based on integrated plasma
45 lipidomics and peptidomics data without pre-selection. Logistic regression was then
46 used to explore clustering based on ≥ 180 circulating lipids and 1,195 protein
47 biomarkers, alongside clinical signatures. Two subgroups were identified, one of which
48 associated with elevated C-peptide levels, diabetic complications and more severe
49 insulin resistance compared to the other. GWAS analysis against 403 type 2 diabetes
50 risk variants revealed associations of several SNPs with clusters and altered molecular
51 profiles. We thus demonstrate that heterogeneity in type 2 diabetes can be captured by
52 circulating omics alone using an unsupervised bottom-up approach. Such multi-omics
53 signatures could reflect pathological mechanisms underlying type 2 diabetes and thus
54 may help inform on precision medicine approaches to disease management.

55

56 **Introduction**

57 Type 2 diabetes is global problem of increasing proportions and a substantial threat to
58 human health (1). Recently, Ahlqvist et al. (2) proposed an approach to sub-classifying
59 patients with type 2 diabetes into four different subgroups by *K-means* clustering using
60 six clinical measurements: age at diagnosis, Body Mass Index (BMI), Glutamic acid

61 decarboxylase (GAD) autoantibodies (GADA), HOMA2-B (beta-cell function), HOMA2-IR
62 (insulin resistance) and HbA1c. Since then, we (3) and others (4,5) have applied a
63 similar clustering method with independent cohorts and obtained similar results. These
64 studies suggest the new classification system facilitates meaningful sample subgroup
65 discovery across different populations. However, the value of this stratification
66 approach has been questioned by some researchers (6,7) in terms of its stability and
67 utility. Importantly, it is still debatable whether type 2 diabetes patients with different
68 underlying aetiologies are represented faithfully, by placing them into subsets, based
69 solely on a limited number of clinical features. Indeed, others (8) have proposed that
70 positioning individuals within a multi-dimensional, and potentially fluid, continuum of
71 variables may provide a more useful prognostic strategy.

72

73 An alternative clustering approach, Similarity Network Fusion (SNF) (9) combines
74 multiple levels of data from the same patients into a similarity network, enabling
75 exploratory multidimensional data-driven clustering. First, patient similarity networks
76 are generated for each data level independently then merged into a single global
77 network through an iterative process. Clustering can then be performed on this network
78 using a graph-based algorithm. As opposed to clustering patients based on a limited
79 number of pre-selected clinical features, this unsupervised bottom-up approach can
80 leverage much larger biological data sets (“omics”), and holds the promise of
81 uncovering new insights, which may be missed by models based on single data types,
82 limited clinical features or supervised approaches.

83

84 In the present study, we performed distributed analysis of two independent patient
85 cohorts to explore the stratification of type 2 diabetes based on combined plasma

86 lipidomics and peptidomics using a federated database system. In addition, we explore
87 the genetic factors predisposing to membership in these groups and the potential
88 correlations with patient molecular profiles.

89

90 **Methods**

91 **Study populations**

92 We used data from two type 2 diabetes cohorts: Hoorn Diabetes Care System (DCS) and
93 Genetics of Diabetes Audit and Research in Tayside Scotland (GoDARTS) .

94 The DCS cohort recruits almost all type 2 diabetes patients from 103 GPs in the West-
95 Friesland region of the Netherlands. This prospective, regional cohort study started in
96 1998 and by the time of 2017, it holds 12,673 type 2 diabetes patients with a median of
97 0.7 years (IQR 0.2-3.7) after diagnosis (10).

98

99 Between 1996 and 2015, GoDARTS recruited 10149 type 2 diabetes patients from the
100 Tayside region of Scotland. Patients in the GoDARTS cohort were not necessarily
101 recruited at the time of diagnosis (11).

102

103 Lipidomics and peptidomics are available for a subset of type 2 diabetes patients in
104 both DCS and GoDARTS cohorts. These data were generated as part of the IMI2
105 RHAPSODY project (3). The sample availability for omics was within 2 years of
106 diagnosis. Of note, individuals were selected without taking into consideration pre-
107 cluster assignments.

108

109

110

111 **Measurements**

112 Data from both cohorts were used with informed consent obtained through the relevant
113 local ethical committees. In DCS, all blood measurements are performed in fasted
114 individuals. In contrast, all measurements in GoDARTS are performed in the non-fasted
115 state. In DCS, Haemoglobin A1c was measured based on the turbidimetric inhibition
116 immunoassay for haemolysed whole EDTA blood (Cobas c501, Roche Diagnostics,
117 Mannheim, Germany). The levels of triglycerides, total cholesterol and HDL cholesterol
118 were measured enzymatically (Cobas c501, Roche Diagnostics) (10). In DCS and
119 GoDARTS, C-peptide was measured on a DiaSorin Liaison (DiaSorin, Saluggia, Italy).
120 Plasma lipids were determined using a QExactive mass spectrometer (Thermo
121 Scientific). Plasma protein levels were measured on the SomaLogic SOMAscan platform
122 (Boulder, Colorado, USA). For more details refer to (3).

123

124 **Cluster analysis**

125 A federated database of type 2 diabetes cohorts including DCS and GoDARTS has
126 previously been set up as part of the RHAPSODY project. This system enables statistical
127 and machine learning analysis to be performed on cohort data at a distance without any
128 disclosure of sensitive data. The federated database system was interrogated using the
129 R statistical programming language (version 4.0.4). In both DCS and GoDARTS cohorts,
130 patients with complete lipidomics and peptidomics (589 and 545 patients, DCS and
131 GoDARTS, respectively) were used for clustering and subsequent statistical analysis.
132 Lipidomics and peptidomics values were centered to a mean value of 0 and an SD of 1 in
133 each cohort using the *scale* function in R. Euclidean distances between each pair of
134 patients were then calculated based on the normalized lipidomics or peptidomics data
135 by using *dist2* function from *dssSNF* (*dsSwissKnifeClient* package) (12). Patient similarity

136 matrices were generated from the Euclidean distance matrices with parameters of K
137 (the number of nearest neighbours) equal to 20 and hyperparameter alpha equal to 5.
138 The patient similarity matrices for lipidomics and peptidomics were then fused using
139 the *SNF* function within *SNFtool* package (13) with T (number of iterations) equal to 20.
140 In both DCS and GoDARTS, men and women were clustered together without pre-
141 separation. Two-step clustering was then performed, in which the optimal number of
142 clusters is estimated in the first step using the silhouette method and patient
143 assignments made in the second step. Silhouette widths of fused cluster patients were
144 calculated from the similarity matrices using the *Silhouette_SimilarityMatrix* function
145 (*CancerSubtypes* package) (14).

146

147 SNF models were validated by bootstrapping tests (n=1000 iterations) that compared
148 model classification to models using randomized Euclidean distance matrices. Euclidean
149 distance matrices were randomized using the function *randomizeMatrix* within the
150 *picante* package (15). For each simulated model, the same parameters were used and
151 the number of clusters is set to correspond to the number of clusters in the study model.
152 The significance of each cluster (P-value ≤ 0.05) was calculated by assessing the
153 frequency of achieving a model with an equal or greater mean local cluster coefficient
154 for randomized data. Local cluster coefficient calculation was done with unweighted,
155 undirected adjacency matrices using the *transitivity* function in the *igraph* package (16).
156 The adjacency matrices were generated from the similarity matrices with the top 5%
157 similarity values set as 1 and the rest as 0.

158

159 The possibility of using clinical measurements to replicate our multi-omics clustering
160 results was assessed in a logistic regression model with pseudo R square as the

161 recorded results. One of the common interpretations of pseudo R square is the
162 indication of the improvement to which the model parameters improve upon the
163 prediction of the null model. In a logistic regression model, we treated the patient's
164 assignment results as the dependent variable (cluster 1:1; cluster 2:0) with six clinical
165 measurements: age, BMI, C-peptide, LDL cholesterol, HDL cholesterol, Triglycerides act
166 as univariate or together as covariates. Pseudo R square=1-(deviation/ null deviation).

167

168 **Statistical analysis**

169 Logistic regression modelling was performed by using the *ds.glm* function within the
170 *dsBaseClient* package to assess the association level of each molecule (lipidomics and
171 peptidomics), clinical measurements and type 2 diabetes SNPs (17) with the multi-
172 omics clusters. The cluster classification was treated as a dependent categorical variable
173 with age, gender and BMI acting as covariates. Subsequently, for lipids, peptides and
174 clinical measurements, each cluster was subset for the strongly associated (P-value
175 ≤ 0.05) features and the mean values of each group of features in each cluster were
176 calculated. Mean values were used since, to protect the patient's identity, individual-
177 level data cannot be downloaded from the federated database system. However, it is
178 possible to display a mean value if it is derived from 5 or more patients. Several mean
179 values for each feature from 5 or more patients were calculated based on the default
180 patients' order in the remote server for each cluster. The feature values were
181 normalized. Hierarchical clustering was then performed, and the results were visualized
182 as heatmaps using the R package *gplots* (18).

183

184 For each patient in the database, the clinical cluster assignments were defined
185 previously, similar to (2), but with five clinical variables: age, BMI, HbA1c, high-density

186 lipoprotein (HDL) and C-peptide (3). Our clusters were compared with these clinical
187 clusters by calculating the hypergeometric overlap p-value (*phyper* function in R).

188

189 Cox regression was performed using the *dssCoxph* function in the *dssSwissKnifeClient*
190 package. Time to insulin requirement was defined as the length of time from diagnosis
191 until an individual started insulin treatment for a period of more than six months, or
192 alternatively as more than two independent HbA1c measurements greater than 69
193 mmol/mol (8.5%) at least three months apart and when ≥ 2 non-insulin glucose-
194 lowering drugs were taken. Hazard ratios for time to insulin treatment requirement
195 within the clusters were calculated using Cox regression with age, gender and BMI as
196 covariates.

197

198 The associations between type 2 diabetes SNPs (17) and multi-omics clusters or
199 molecules with altered expression profiles were assessed using a linear regression
200 model with age, BMI and gender as co-independent variables. The routinely used
201 genome-wide common variant association threshold $p \leq 5 \times 10^{-8}$ was applied in
202 this study (19) to assess significance.

203

204 All analyses were performed using the R statistical programming language (version
205 4.0.4). The Benjamini-Hochberg procedure was used to determine the false discovery
206 rate correcting for multiple tests. Figures were produced using *gplots* (version 3.1.1),
207 *ggplot2* (version 3.3.4) and *igraph* (version 1.2.6). Figure 1a was generated using
208 *Biorender*.

209

210

211 **Data availability statement**

212 Information about accessing the Rhapsody federated database can be found at:

213 <https://imi-rhapsody.eu/>. Summary statistics of lipidomic and proteomic data will be

214 available from an interactive Shiny dashboard, available upon publication.

215

216 **Results**

217 **Unsupervised multi-Omics clustering defines two subgroups in independent**

218 **cohorts**

219 Similarity Network Fusion (SNF) was performed using plasma lipidomics and

220 peptidomics data from a total of 1022 type 2 diabetes patients in two independent

221 cohorts. 589 and 545 type 2 diabetes patients, respectively, from the DCS and GoDARTS

222 cohorts with complete cases of plasma lipidomics and peptidomics measurements were

223 selected. The characteristics of these two cohorts were generally comparable in terms

224 of average age and BMI, with a majority of males (3). Individuals were clustered using

225 SNF as described in Methods. Through silhouette analysis, we were able to demonstrate

226 that clustering integrating both lipidomics and peptidomics outperforms either of the

227 single -omics in terms of cluster assignment quality (Supplemental Figure 1). Based on a

228 two-step clustering procedure, the optimal number of clusters was found to be two

229 (Supplemental Figure 1). The significance of the clustering results was validated by

230 bootstrapping (n=1000 iterations) against simulated datasets (Supplemental Figure 2).

231 In the DCS cohort, cluster 1 comprised 46.5% of the patients and was characterized by a

232 relatively high BMI value compared to cluster 2 (53.5%). In both GoDARTS (cluster

233 1:40.7%; cluster 2:59.3%) and DCS cohorts, patients have similar age, HbA1c values.

234 Furthermore, in the DCS cohort, triglycerides, C-peptide and homeostatic model

235 assessment (HOMA) 2, HOMA2-B and HOMA2-IR, which were calculated based on

236 fasting blood glucose and fasting C-peptide concentration, showed a higher
237 concentration in cluster 1 compared to cluster 2. In contrast, LDL cholesterol, HDL
238 cholesterol and HOMA2-S show a higher concentration in cluster 2 (Supplemental
239 Figure 3). Similarly, a higher concentration of LDL cholesterol and HDL cholesterol, and
240 a lower concentration of triglycerides and C-peptide, were also observed in cluster 2
241 compared to cluster 1 in the GoDARTS cohort (Figure 2a; Supplemental Figure 3).

242 We compared type 2 diabetes progression rates between clusters in both cohorts using
243 Cox Proportional hazard models . In both cohorts, type 2 diabetes progression rate
244 showed a nominally significant difference in DCS and GoDARTS, respectively, between
245 multi-omics cluster 1 and cluster 2 ($p=0.0156$ (unadjusted), DCS; $p=0.0308$
246 (unadjusted), GoDARTS; corrected for gender). Cluster 2 reduces the hazard ratio by
247 44% or 27% in DCS or GoDARTS, respectively (Figure 2d, left). Results were in the same
248 direction but were less significant when additionally correcting for age and BMI (figure
249 2d, right).

250 We then investigated whether the multi-omics clusters could be replicated by using
251 clinical measurements alone. Pseudo R square was used to assess the possibility of
252 using six clinical features: age, BMI, LDL cholesterol, HDL cholesterol, C-peptide and
253 triglycerides to predict the multi-omics clustering results in a logistic model for both
254 cohorts (Supplemental Table 1). The highest pseudo R^2 scores are shown when all
255 seven clinical measurements act as covariates with 29.9% and 46.9% improvement
256 over random in DCS and GoDARTS, respectively. These results clearly indicate that the
257 molecular clusters we observe are not merely reflecting differences in routine clinical
258 measurements.

259 In summary, by combining plasma lipidomics and peptidomics data in a data-driven and
260 unsupervised analysis, we were able to separate type 2 diabetes patients from two

261 independent cohorts into two subgroups that appear to differ in diabetes-related
262 clinical characteristics.

263

264 **Similar molecular signatures separate subgroups in both cohorts**

265 We investigated the correlation of each plasma lipidomics or peptidomics feature with
266 the clustering results using logistic regression models adjusting for gender, BMI and
267 age. In DCS, out of a total of 1195 different peptides and 180 different lipids that were
268 used for clustering, 123 proteins and 137 lipids showed significant differences between
269 clusters. In GoDARTS, out of a total of 1195 different peptides and 199 different lipids
270 that were used for clustering, 130 proteins and 149 lipids showed significant
271 differences between clusters. 50 significant common peptides and 109 significant
272 common lipids are shared between the DCS and GoDARTS cohorts (Figure 2b).

273

274 A hierarchical clustering of the relative concentration patterns of the cluster-
275 associated features is shown as a heatmap in DCS and GoDARTS (Figure 2b;
276 Supplemental Figure 5). In DCS, a number of discriminative omics features were
277 observed when comparing clusters. Phosphatidylcholine, Triacylglycerol, Diacylglycerol
278 and Ceramide were relatively higher in cluster 1; Phosphatidylcholine O and
279 Sphingomyelin were relatively higher in cluster 2. These pronounced expression
280 patterns were also apparent in GoDARTS. For proteins, the difference in the molecular
281 profiles between cluster 1 and 2 can be observed in both cohorts, but the differences
282 were more modest compared to lipids.

283

284

285 **GWAS analysis reveals putative associations between type 2 diabetes molecular**
286 **profiles and genetic factors**

287 Finally, in order to investigate possible genetic differences between molecular clusters,
288 we analysed genetic loci previously associated with type 2 diabetes (17). Multi-omics
289 clusters in each cohort were compared with type 2 diabetes SNPs using logistic
290 regression correcting for age, BMI and gender. We did not detect any significant
291 associations of the type 2 diabetes variants with our clusters in either cohort.

292 Since we had identified around 110 proteins and 130 lipids that showed altered levels
293 between clusters, we then tested the association between the altered molecular profiles
294 and type 2 diabetes variants. The SNP and molecular profile association was assessed in
295 a linear regression model with SNP dosages, age, BMI and gender as independent
296 variables. In DCS, a novel protective type 2 diabetes variant (17), rs146886108, showed
297 putative association with several molecules such as heat shock protein 90 beta (p
298 =1.08e-13) and glucose 6 phosphate isomerase ($p = 2.04e-15$). A consistent association
299 result between rs505922 and E-selection can also be observed in both DCS ($p = 1.23e-$
300 15) and GoDARTS ($p = 1.40e-16$) cohorts.

301

302 **Multi-Omics clusters show similarity to known clinical subgroups**

303 We next compared the multi-omics clusters with clusters based on clinical
304 measurements (3). Similar to a previous study (2), these clusters can be classified as
305 severe insulin-deficient diabetes (SIDD); severe insulin-resistant diabetes (SIRD); mild
306 obesity-related diabetes (MOD); mild diabetes (MD) and mild diabetes with high HDL
307 (MDH). The overlap of clinical clusters with molecular clusters is shown in figure 3a. In
308 DCS, multi-omics cluster 1 showed a significant overlap with clinical cluster SIRD and
309 MOD. On the other hand, multi-omics cluster 2 showed a significant overlap with clinical

310 cluster MDH. In GoDARTS, a similar trend was observed, as multi-omics cluster 1
311 showed a larger overlap with SIRD and MOD, and multi-omics cluster 2 showed an
312 overlap with MDH (figure 3a).

313 Box plots were used to further visualize the differences between multi-omics and
314 clinical clusters (figure 3b, Supplemental figure 4). In DCS, we observed that multi-
315 omics cluster 1, SIRD and MOD showed similar levels of C-peptide, HOMA-2B, HOMA-
316 2IR, and triglycerides levels. On the other hand, multi-omics cluster 2 and MDH both
317 showed a low C-peptide and a high HDL cholesterol level (Supplemental figure 4a). A
318 similar pattern was observed in GoDARTS (Supplemental figure 4b).

319 In summary, we conclude that multi-omics cluster 1 shares similar clinical features to
320 SIRD and MOD clinical subgroups whereas multi-omics cluster 2 shows higher
321 similarity to the MDH clinical subgroup.

322

323 **Discussion**

324 Using an unsupervised, bottom-up approach, we were able to cluster individuals with
325 diabetes from two large European cohorts into two robust clusters based solely on
326 plasma-measured multi-omics data. Clinical features such as age and BMI were
327 excluded as confounding variables driving SNF clustering as they were not significantly
328 different between cluster 1 and cluster 2. Moreover, using six clinical features to
329 replicate the multi-omics clustering results was unsuccessful for both cohorts. These
330 findings suggest that patient stratification using molecular features can reveal novel
331 insights that are complementary to clinical data.

332 In both the DCS and GoDARTS cohorts, patients assigned to multi-omics clusters 1 and
333 2 showed conserved differences in disease severity markers as measured by HOMA2-B,
334 -IR, C-peptide and triglyceride levels, which is known to be associated with insulin

335 resistance (20). Individuals with type 2 diabetes are at risk of many complications such
336 as kidney disease and cardiovascular disease. It has been demonstrated that insulin
337 resistance is associated with hallmarks of kidney damage such as glomerular
338 hypertension, hyperfiltration and proteinuria (21, 22). We show evidence from Cox
339 proportional hazards modelling that patients in different clusters have slower disease
340 progression rates in both cohorts. Therefore, these multi-omics clusters could
341 potentially define molecular signatures related to disease severity, progression and
342 complications risk which could be relevant for more precision-based therapy of type 2
343 diabetes.

344 We tested the associations of type 2 diabetes genetic loci with multi-omics clusters
345 but found no clear evidence to suggest that the multi-omics clusters were associated
346 with specific type 2 diabetes genetic variants in either cohort. This suggests that known
347 type 2 diabetes genes are not drivers of molecular differences between clusters. A
348 similar clustering result can be observed in both cohorts and cluster 1 and cluster 2
349 show similar HbA1c levels suggesting that our clusters are stable and at least partially
350 mechanistically distinct. It would be intriguing to see whether similar results can be
351 obtained from both newly diagnosed patients and long-term diabetes patients could be
352 performed in the future to support this view.

353 In both cohorts, we have identified ~100 proteins and ~130 lipids that showed an
354 altered expression pattern between multi-omics cluster 1 and cluster 2. Currently, we
355 cannot fully identify which molecules likely to be the driver molecules, nor we can say
356 these molecules play an important role in type 2 diabetes development. Nevertheless,
357 we assessed the association between these molecules and the type 2 diabetes variants.
358 A novel type 2 diabetes SNP, rs146886108, a missense variant that encodes
359 p.Arg187Gln in ANKH, which was first identified by Mahajan et al. (17), and showed a

360 strong correlation with several measured proteins, including heat shock protein 90
361 beta, glucose 6 phosphate isomerase. The UK Biobank Mendelian trait GWAS study also
362 showed a strong association between rs146886108 and random glucose. The fact that
363 we found a strong association between rs146886108 and glucose 6 phosphate
364 isomerase could potentially identify a mechanism behind the action of rs146886108 on
365 glucose regulation. Further investigations may help to evaluate the causal roles of these
366 type 2 diabetes variants, with their associated molecules and possible mechanisms.

367 Using five clinical measurements, age, BMI, HbA1c, c-peptide and HDL, Slieker and
368 Donnelly et al. (3) have shown that people with type 2 diabetes could be grouped into
369 five distinct clusters: insulin-deficient (SIDD), insulin-resistant (SIRD), an obesity
370 cluster (MOD), mild diabetes (MD) and MD with a high HDL (MDH). Comparing with
371 these clinical clusters, we were able to demonstrate that the multi-omics clusters
372 significantly overlap with the clinically defined clusters. Multi-omics cluster 1,
373 characterized by a relatively high C-peptide and triglycerides level, showed a significant
374 overlap with SIRD and MOD. Multi-omics cluster 2, a cluster with a relatively high HDL
375 and low C-peptide value, showed a significant overlap with MDH. These results further
376 underline the complementarity of the multi-omics and clinical clustering approaches
377 and suggests that multi-omics signatures can in theory be linked to clinical
378 measurements and outcomes to explore molecular mechanisms linked to the underlying
379 heterogeneity of disease.

380 The present approach offers the prospect of multi-omics personalized medicine, in
381 which an extended set of molecular characteristics is used to assess disease
382 development and progression. Through multi-omics health management, each patient
383 might be positioned with respect to the underlying dysregulated pathways, with
384 therapies selected accordingly. Whether this approach will indeed provide better

385 outcomes than current practice, based on clinical features and more limited blood
386 biochemistry, will need to be assessed in the future.

387

388 **Limitations of study**

389 From the patient network (Figure 1) it is clear that the multi-omics clusters are not
390 completely distinct. The two clusters therefore represent different molecular
391 signatures, where a patient shows more similarity to one cluster compared to the other
392 on a continuous scale. Nevertheless, we do not recommend using such clusters to assign
393 patients, but rather the use of the clustering results to explore the underlying molecular
394 pathology and as an alternative route to identifying novel biomarkers for risk of type 2
395 diabetes and its complications.

396 In both DCS and GoDARTS, patients were not necessarily recruited after diagnosis, and
397 this heterogeneity in the disease duration could potentially have an effect on the blood
398 measurements obtained. Moreover, as we mentioned above, in DCS, blood samples are
399 taken in a fasted state, in contrast to GoDARTS, where blood measurements are done in
400 a non-fasted state. However, it has been demonstrated that a normal food intake only
401 has slight effects on cholesterol levels, and lipid profiles at most change minimally in
402 response to normal food intake in general populations (23, 24).

403 One of the common questions regarding patients' clustering results is whether patients
404 move between clusters over time. In a recent 5-year follow-up study (25), 367 diabetes
405 patients were assigned based on the nearest centroid approaches at both baseline and
406 5-year follow-up. 23% of patients switched their cluster allocations after 5 years. This
407 can be expected as the variables that were used for clustering such as age, BMI and
408 HOMA2 can change over time, following the time increase, lifestyle choice changes and
409 disease progress. For our study, roughly 1375 plasma molecules were used for

410 clustering. Some molecule concentrations such as triacylglycerol and diacylglycerol can
411 change following lifestyle changes or medical treatment. Other “housekeeping”
412 molecules are more likely to remain constant over time. The stability of molecular
413 clustering results needs to be assessed in further prospective studies.

414 We cannot at this stage assert that the different clusters represent different aetiologies
415 of type 2 diabetes, nor that the clustering method represents an optimal classification
416 for type 2 diabetes subgroups. The multi-omics clusters in our study were derived
417 primarily from patients from Europe. Consequently, the applicability of these results to
418 other ethnic groups needs to be assessed.

419

420 **Acknowledgments**

421 The authors acknowledge the support of the Health Informatics Centre, University of
422 Dundee, for managing and supplying the anonymized data.

423 **Funding.** G.R. was supported by a Wellcome Trust Investigator Award
424 (212625/Z/18/Z), MRC Programme grant (MR/R022259/1) Diabetes UK
425 (BDA/15/0005275, BDA 16/0005485) grants and a start-up grant from the CR-CHUM,
426 Université de Montréal. This project has received funding from the Innovative
427 Medicines Initiative 2 Joint Undertaking, under grant agreement no. 115881
428 (RHAPSODY). This Joint Undertaking receives support from the European Union’s
429 Horizon 2020 research and innovation programme and EFPIA. This work is supported
430 by the Swiss State Secretariat for Education, Research and Innovation (SERI), under
431 contract no. 16.0097.

432 **Duality of Interest.** G.A.R. has received grant funding from, and is a consultant for, Sun
433 Pharmaceuticals Inc. K.S. is CEO of Lipotype. K.S. and C.K. are shareholders of Lipotype.
434 M.J.G. is an employee of Lipotype. M.K.H. is an employee of Janssen Research &

435 Development. All other authors declare that there are no relationships or activities that
436 might bias, or be perceived to bias, their work.

437 **Contribution Statement.** S.L., M.I. and G.A.R. designed the research. S.L. wrote the
438 manuscript. S.L. performed all data analysis with technical support from I.D., I.D., D.K.,
439 and M.I. set up a federated node system for data analysis. R.C.S., L.A.D., M.K.H., J.W.J.B.,
440 L.M't.H., F.M, P.J.M.E. and E.R.P. were involved in generating the peptidomics data,
441 generating the clinical cluster assignment results, managing and handling the data in
442 databases. M.J.G., C.K. and K.S. generated the Lipotype data. The research was designed
443 partly based on the pre-work results by C.H.F. All authors read and approved the
444 manuscript and contributed according to the ICMJE criteria (www.icmje.org/). G.A.R.
445 and M.I. co-supervised the work. M.I. is the guarantor of this work and, as such, had full
446 access to all the data in the study and takes responsibility for the integrity of the data
447 and the accuracy of the data analysis.

448

449 **Table 1. Genetic associations with altered molecular profiles between multi-**
 450 **omics clusters reaching significance.** The correlations between molecules and SNPs
 451 were estimated using linear regression models with age, BMI and gender adjustments.
 452 The commonly accepted threshold $P \leq 5 \times 10^{-8}$ for genome-wide association studies is
 453 applied in this present study.

DCS						
Locus	rs ID	Chr	Pos	Alleles	association molecule	P-value (adjusted by BMI, age and sex)
ANKH	rs146886108	5	14,751,305	C/T	Integrin alpha I beta 1 complex	5.88E-10
ANKH	rs146886108	5	14,751,305	C/T	Heat shock protein HSP 90 alpha beta	1.13E-21
ANKH	rs146886108	5	14,751,305	C/T	Heat shock protein HSP 90 beta	1.08E-13
ANKH	rs146886108	5	14,751,305	C/T	Glucose 6 phosphate isomerase	2.04E-15
ANKH	rs146886108	5	14,751,305	C/T	Alcohol dehydrogenase NADP	4.04E-32
ABO	rs505922	9	133,273,813	C/T	E selectin	1.23E-15

ABO	rs505922	9	133,273,813	C/T	CD209 antigen	2.09E-14
GoDARTS						
Locus	rs ID	Chr	Pos	Alleles	association molecule	P-value (adjusted by BMI, age and sex)
ABO	rs505922	9	133,273,813	C/T	E selectin	1.40E-16

454

455

456

457 **References**

- 458 1. International Diabetes Federation. IDF Diabetes Atlas, 10th edn. Brussels,
459 Belgium: International Diabetes Federation, 2021.
- 460 2. Ahlqvist E, Storm P, Käräjämäki A, Martinell M, Dorkhan M, Carlsson A, et al.
461 Novel subgroups of adult-onset diabetes and their association with outcomes: a
462 data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol*. 2018
463 May;6(5):361–9.
- 464 3. Slieker RC, Donnelly LA, Fitipaldi H, Bouland GA, Giordano GN, Åkerlund M, et al.
465 Distinct Molecular Signatures of Clinical Clusters in People With Type 2 Diabetes:
466 An IMI-RHAPSODY Study. *Diabetes*. 2021 Nov;70(11):2683–93.
- 467 4. Tanabe H, Saito H, Kudo A, Machii N, Hirai H, Maimaituxun G, et al. Factors
468 Associated with Risk of Diabetic Complications in Novel Cluster-Based Diabetes
469 Subgroups: A Japanese Retrospective Cohort Study. *J Clin Med Res [Internet]*.
470 2020 Jul 2;9(7).
- 471 5. Safai N, Ali A, Rossing P, Ridderstråle M. Stratification of type 2 diabetes based on
472 routine clinical markers. *Diabetes Res Clin Pract*. 2018 Jul;141:275–83.
- 473 6. Kahkoska AR, Geybels MS, Klein KR, Kreiner FF, Marx N, Nauck MA, et al.
474 Validation of distinct type 2 diabetes clusters and their association with diabetes
475 complications in the DEVOTE, LEADER and SUSTAIN-6 cardiovascular outcomes
476 trials. *Diabetes Obes Metab*. 2020 Sep;22(9):1537–47.
- 477 7. Lugner M, Gudbjörnsdóttir S, Sattar N, Svensson AM, Miftaraj M, Eeg-Olofsson K,
478 et al. Comparison between data-driven clusters and models based on clinical
479 features to predict outcomes in type 2 diabetes: nationwide observational study.
480 *Diabetologia*. 2021 Sep;64(9):1973–81.

- 481 8. McCarthy MI. Painting a new picture of personalised medicine for diabetes.
482 Diabetologia. 2017 May;60(5):793–9.
- 483 9. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network
484 fusion for aggregating data types on a genomic scale. Nat Methods. 2014
485 Mar;11(3):333–7.
- 486 10. van der Heijden AA, Rauh SP, Dekker JM, Beulens JW, Elders P, 't Hart LM, et al.
487 The Hoorn Diabetes Care System (DCS) cohort. A prospective cohort of persons
488 with type 2 diabetes treated in primary care in the Netherlands. BMJ Open. 2017
489 Jun 6;7(5):e015599.
- 490 11. Hébert HL, Shepherd B, Milburn K, Veluchamy A, Meng W, Carr F, et al. Cohort
491 Profile: Genetics of Diabetes Audit and Research in Tayside Scotland (GoDARTS).
492 Int J Epidemiol. 2018 Apr 1;47(2):380–1j.
- 493 12. Iulian Dragan (2021). dsSwissKnifeClient: DataSHIELD Tools and Utilities - client
494 side. R package version 0.2.0
- 495 13. Bo Wang, Aziz Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno,
496 Benjamin Haibe-Kains and Anna Goldenberg (2021). SNFtool: Similarity Network
497 Fusion. R package version 2.3.1.
- 498 14. Xu T, Le TD, Liu L, Su N, Wang R, Sun B, et al. CancerSubtypes: an R/Bioconductor
499 package for molecular cancer subtype identification, validation and visualization.
500 Bioinformatics. 2017 Oct 1;33(19):3131–3.
- 501 15. Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, et al.
502 Picante: R tools for integrating phylogenies and ecology. Bioinformatics. 2010
503 Jun 1;26(11):1463–4.

- 504 16. Csardi G, Nepusz T: The igraph software package for complex network research,
505 InterJournal, Complex Systems 1695. 2006.
- 506 17. Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, et al.
507 Fine-mapping type 2 diabetes loci to single-variant resolution using high-density
508 imputation and islet-specific epigenome maps. *Nat Genet.* 2018 Nov;50(11):1505
509 –13.
- 510 18. Gregory R. Warnes, Ben Bolker, Lodewijk Bonebakker, Robert Gentleman,
511 Wolfgang Huber, Andy Liaw, Thomas Lumley, Martin Maechler, Arni Magnusson,
512 Steffen Moeller, Marc Schwartz and Bill Venables (2020). *gplots: Various R*
513 *Programming Tools for Plotting Data.* R package version 3.1.1.
- 514 19. Panagiotou OA, Ioannidis JPA, Genome-Wide Significance Project. What should
515 the genome-wide significance threshold be? Empirical replication of borderline
516 genetic associations. *Int J Epidemiol.* 2012 Feb;41(1):273–86.
- 517 20. Glueck CJ, Khan NA, Umar M, Uppal MS, Ahmed W, Morrison JA, et al. Insulin
518 resistance and triglycerides. *J Investig Med.* 2009 Dec;57(8):874–81.
- 519 21. Groop L, Ekstrand A, Forsblom C, Widén E, Groop PH, Teppo AM, et al. Insulin
520 resistance, hypertension and microalbuminuria in patients with type 2 (non-
521 insulin-dependent) diabetes mellitus. *Diabetologia.* 1993 Jul;36(7):642–7.
- 522 22. Gnudi L, Coward RJM, Long DA. Diabetic Nephropathy: Perspective on Novel
523 Molecular Mechanisms. *Trends Endocrinol Metab.* 2016 Nov;27(11):820–30.
- 524 23. Langsted A, Freiberg JJ, Nordestgaard BG. Fasting and nonfasting lipid levels:
525 influence of normal food intake on lipids, lipoproteins, apolipoproteins, and
526 cardiovascular risk prediction. *Circulation.* 2008 Nov 11;118(20):2047–56.

- 527 24. Liu MM, Peng J, Cao YX, Guo YL, Wu NQ, Zhu CG, et al. The difference between
528 fasting and non-fasting lipid measurements is not related to statin treatment.
529 Ann Transl Med. 2021 Mar;9(5):386.
- 530 25. Zaharia OP, Strassburger K, Strom A, Bönhof GJ, Karusheva Y, Antoniou S, et al.
531 Risk of diabetes-associated diseases in subgroups of patients with recent-onset
532 diabetes: a 5-year follow-up study. Lancet Diabetes Endocrinol. 2019
533 Sep;7(9):684–94.
- 534
- 535
- 536
- 537
- 538

539 **Figure legends**

540 **Figure 1. Experiment design and the molecular clusters displayed as both**
541 **similarity matrices and patient network. (b,c)** Molecular clusters were generated by
542 integrated multi-omics (lipidomics and peptidomics) data in both DCS (b) and GoDARTS
543 (c), respectively. The similarity matrices were calculated based on Euclidean distances
544 with parameters of K=20 (the number of shared neighbours of datapoint), alpha=0.5
545 (hyperparameter) and T=20 (number of iterations). Each point in the similarity
546 matrices represents patient-to-patient similarity, the higher the colour intensity, the
547 more similar the patients. Boxed regions represent the potential clusters. The patient
548 networks were generated based on unweighted adjacency matrices. The nodes
549 represent the patients, edge thickness reflects the strength of similarity, the size of the
550 node represents the betweenness. Nodes were coloured based on their cluster
551 assignments. DCS= Hoorn Diabetes Care System. GoDARTS=Genetics of Diabetes Audit
552 and Research in Tayside Scotland.

553

554 **Figure 2. (a,b,c) Clinical characteristics and molecular profiles of multi-omics**
555 **clusters in both DCS and GoDARTS clusters. (d) The hazard ratio for time to**
556 **insulin requirement across clusters.** Multi-omics clustering was done without
557 separating the women and men. In each cohort, the statistical difference between
558 cluster 1 and cluster 2 was determined by logistic regression with age, BMI and gender
559 as covariates. Multi-analysed P-values were adjusted by the Benjamini-Hochberg
560 procedure, and a false discovery rate (FDR)-adjusted P-value ≤ 0.05 was considered
561 significant. **(a)** Distributions of nine clinical measurements at baseline in both DCS and
562 GoDARTS cohorts for each cluster. *:P-value ≤ 0.05 . **:P-value ≤ 0.01 . ***:P-
563 value ≤ 0.001 . **(b)** Venn diagrams showing overlaps of significant (significantly different

564 between clusters) proteins (upper) or lipids (lower) across DCS and GoDARTS cohorts.
565 (c) The relative concentration heatmaps of common significant proteins (upper) or
566 lipids (lower) between DCS and GoDARTS cohorts. Concentrations of each feature were
567 first converted into the log(concentration) and then normalized in order to investigate
568 the relative expression level of each feature between different clusters (Row Z-score).
569 Heatmap cells were coloured in red/green to represent the relative expression level
570 (red=low, green=high). The federated database does not allow to download data from
571 individual samples in order to protect the patient's identity. Each vertical line in the
572 heatmap represents 5 different patients' mean values for each feature. y-labels were
573 hidden for a better presentation. Fully labelled versions are available in Supplemental
574 Figure 5. (d) X-axis, hazard ratio, y-axis, database. Cox model with correction of gender
575 (left) or with corrections for gender, age and BMI (right). The number in the panel
576 represents the hazards with 95% confidence intervals. Multi-omics cluster 1:1 and
577 Multi-omics cluster 2:2. Cox model gives the hazard ratio for the second group relative
578 to the first group. The hazard ratio is not significant if the 95% confidence interval
579 includes 1. DBP= Diastolic Blood Pressure. SBP= Systolic Blood Pressure. HbA1C=
580 Haemoglobin A1C. DCS= Hoorn Diabetes Care System. GoDARTS=Genetics of Diabetes
581 Audit and Research in Tayside Scotland.

582

583 **Figure 3 (a)** Comparison of clustering in Slieker and Donnelly et al., 2021 versus
584 clustering in the current study. The first number in each cell represents the number of
585 individuals overlapping, the second number represents the hypergeometric overlap p-
586 value. (b) Distributions of nine clinical measurements at baseline in both DCS and
587 GoDARTS cohorts for multi-omics clusters and clinical clusters. CL1= multi-omics
588 cluster #1. CL2= multi-omics cluster #2. MD= mild diabetes. MDH= mild diabetes with

589 high HDL. MOD= mild obesity-related diabetes. SIDDD= severely insulin deficient

590 diabetes. SIRD= severely insulin resistant diabetes. SBP=systolic blood pressure.

591 DBP=diastolic blood pressure. HbA1C=Hemoglobin A1C.

592

593

594

595

596

597

598

599

600

601

602

603

604

605

606

607

608

609

610

611

612

613

Figure 1

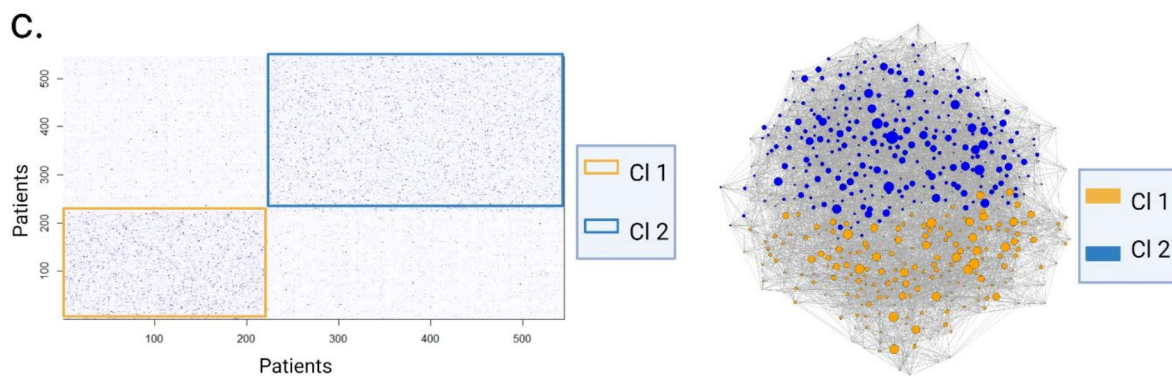
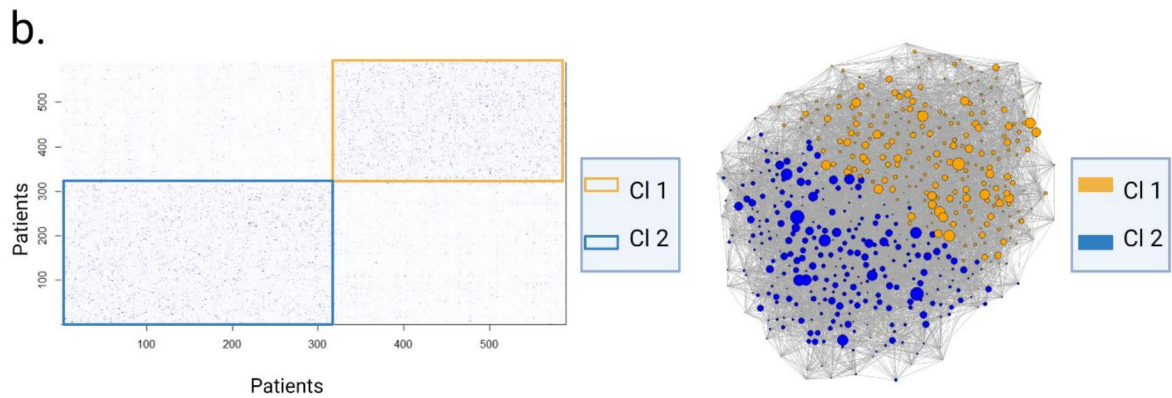
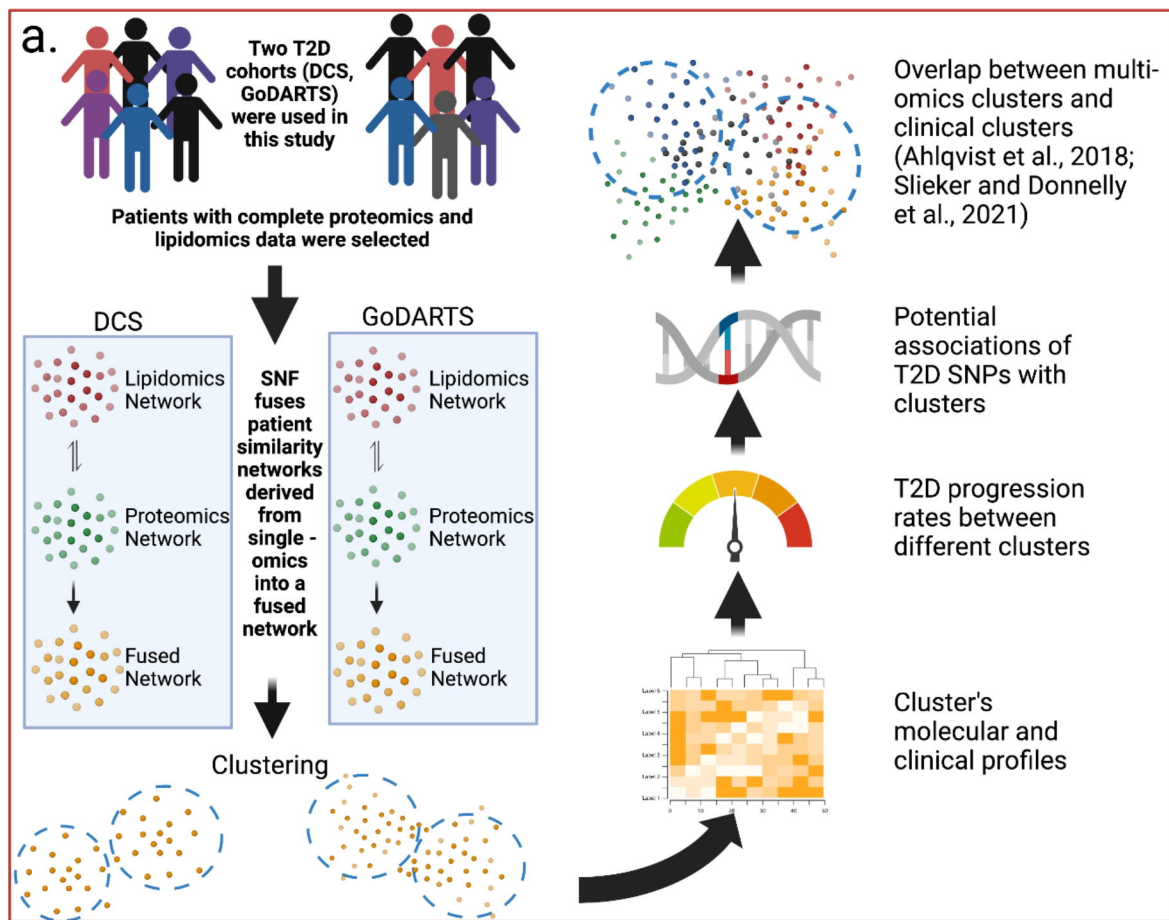


Figure 2

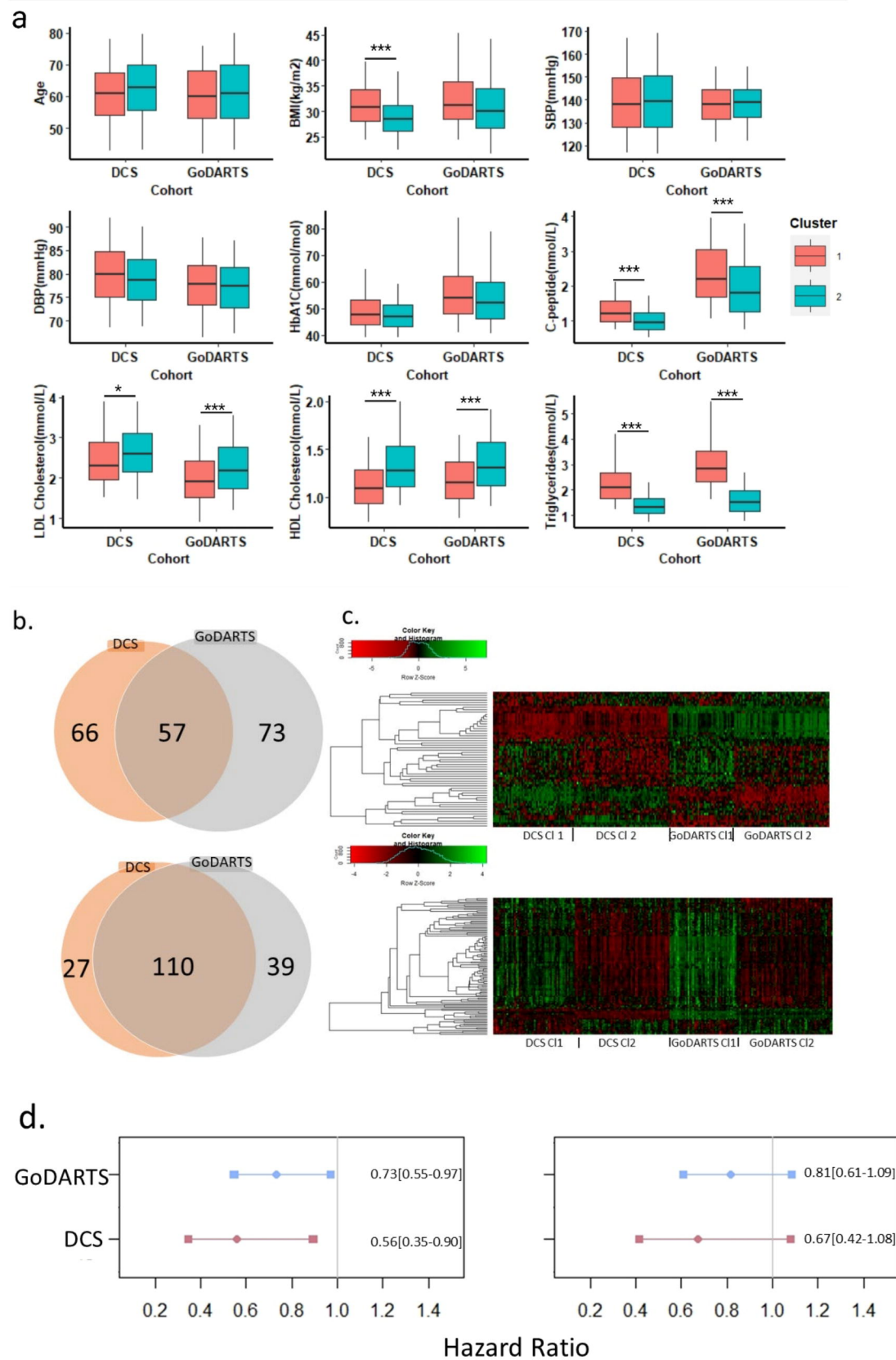


Figure 3

