

1 Kawasaki Disease patient stratification and pathway 2 analysis based on host transcriptomic and proteomic 3 profiles

4 Heather Jackson*¹, Stephanie Menikou*¹, Shea Hamilton¹, Andrew McArdle¹, Chisato Shimizu²,
5 Rachel Galassini¹, Honglei Huang³, Jihoon Kim², Adriana Tremoulet², Marien de Jonge⁴, Taco
6 Kuijpers⁵, Victoria Wright¹, Jane Burns², Climent Casals-Pascual⁶, Jethro Herberg¹, Mike Levin^{#1},
7 Myrsini Kaforou^{#1} and on behalf of the PERFORM consortium

8 ¹ Faculty of Medicine, Imperial College London, London, SW7 2AZ

9 ² Health Sciences, University of California San Diego, La Jolla, CA 92093

10 ³ Oxford Biomedica (UK) Ltd, Oxford, OX4 6LT

11 ⁴ Radboud University Medical Center, Nijmegen, 6525

12 ⁵ Amsterdam University Medical Center (AMC), Amsterdam, 1105

13 ⁶ Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN

14 Correspondence: m.kaforou@imperial.ac.uk

15 **Abstract:**

16 The aetiology of Kawasaki Disease (KD), an acute inflammatory disorder of childhood, remains
17 unknown despite various triggers of KD having been proposed. Host 'omic profiles offer insights
18 into the host response to infection and inflammation, with the interrogation of multiple 'omic
19 levels in parallel providing a more comprehensive picture. We used differential abundance
20 analysis, pathway analysis, clustering and classification techniques to explore whether the host
21 response in KD is more similar to the response to bacterial or viral infection at the transcriptomic
22 and proteomic levels through comparison of 'omic profiles from children with KD to those with
23 bacterial and viral infections. Pathways activated in patients with KD included those involved in
24 anti-viral and anti-bacterial responses. Unsupervised clustering showed that the majority of KD
25 patients clustered with bacterial patients on both 'omic levels, whilst application of diagnostic
26 signatures specific for bacterial and viral infections revealed that many transcriptomic KD samples
27 had low probabilities of having bacterial or viral infections, suggesting that KD may be triggered
28 by a different process not typical of either common bacterial or viral infections. Clustering based
29 on the transcriptomic and proteomic responses during KD revealed three clusters of KD patients
30 on both 'omic levels, suggesting heterogeneity within the inflammatory response during KD. The
31 observed heterogeneity may reflect differences in the host response to a common trigger, or
32 variation dependent on different triggers of the condition.

33 **Keywords:** infectious diseases; paediatrics; transcriptomics; proteomics; Kawasaki Disease; host
34 'omics; systems biology; pathway analysis; clustering; classification

36 **1. Introduction**

37 Kawasaki disease (KD) is an acute inflammatory disorder first described in Japan over 50 years
38 ago [1]. KD occurs most frequently in children under five years of age [2]. Untreated KD leads to the
39 formation of coronary artery aneurysms (CAAs) in 10-30% of children [3-5], causing it to be the most
40 common cause of acquired heart disease in children in Europe, Japan and North America [6].

41 The aetiology of KD remains unknown, however the seasonality and epidemicity seen in areas
42 of high incidence, including Japan, suggest that it is caused by an infectious trigger [7]. The current
43 consensus is that, in some genetically predisposed children, an infectious trigger initiates an
44 abnormal immune response [8,9]. Multiple viral and bacterial pathogens have been suggested as
45 candidates for the trigger, in addition to airborne environmental and fungal triggers [8,10]. Despite

46 the many theories postulated, none have been independently confirmed, and some studies have
47 concluded that KD is likely to be caused by multiple environmental triggers [11].

48 As the coronavirus disease 2019 (COVID-19) pandemic evolved in early to mid-2020, an increase
49 in cases of children with unusual febrile illnesses, some with features resembling KD, was observed
50 [12]. This new condition, which was later termed “Paediatric Inflammatory Multisystem Syndrome
51 Temporally associated with SARS-CoV-2”, or “Multisystem Inflammatory Syndrome in Children”
52 (PIMS-TS or MIS-C) [12–15], tends to arise several weeks after SARS-CoV-2 infection [14]. The finding
53 of increased KD-like cases after the emergence of a novel viral pathogen raises questions about
54 whether more than one type of trigger might cause KD, and whether KD might represent a
55 constellation of overlapping inflammatory syndromes.

56 Study of host transcriptomic and proteomic profiles can reveal perturbations caused by infection
57 or inflammation. Comparison of the transcriptional response in different diseases has revealed
58 different host responses to individual pathogens such as TB, bacterial and viral infections [16,17].
59 Previous studies of host ‘omics in the context of KD have characterised the pathways involved in the
60 disease and have established biomarker signatures with diagnostic potential [18,19]. Interrogating
61 multiple ‘omic datasets in parallel provides more accurate insights into the molecular dynamics of
62 infection as information captured in one ‘omic layer might not necessarily be captured in another
63 ‘omic layer.

64 We explored the host transcriptomic and proteomic profiles of children with KD and those with
65 viral and bacterial infections, aiming to elucidate whether the inflammatory response in KD is more
66 similar to that of a bacterial or viral infection, or indeed neither. We also used the approach to explore
67 the heterogeneity within the transcriptional and translational response of patients with KD.

68 2. Results

69 2.1. Description of datasets

70 Whole blood transcriptomic profiles generated from 414 children were included in the
71 analysis, obtained from children with Kawasaki Disease (KD; $n = 178$), confirmed (definite) bacterial
72 infection (DB; $n = 54$), confirmed (definite) viral infection (DV; $n = 120$), and healthy controls (HC; $n =$
73 62). Two transcriptomic datasets were used. The ‘discovery’ transcriptomic dataset, which was
74 generated by HumanHT-12 version 4.0 BeadChips, was used for all steps of the analysis. The
75 ‘validation’ transcriptomic dataset, which was created by merging two datasets generated by
76 HumanHT-12 version 3.0 and 4.0 BeadChips, was used to test the classifiers trained on the
77 discovery dataset.

78 In addition, proteomic profiles from the plasma or serum of 329 children in the same groups
79 were studied: from children with KD ($n = 52$), DB ($n = 121$) and DV ($n = 106$) infections, and HC ($n =$
80 50). Liquid chromatography with tandem mass spectrometry (LC-MS/MS) and the SomaScan [20]
81 platform were used to generate the proteomic ‘discovery’ and ‘validation’ datasets, respectively.
82 The ‘discovery’ proteomic dataset, generated from plasma samples using LC-MS/MS, was used for
83 all steps of the analysis. The ‘validation’ proteomic dataset, generated from serum samples using
84 the SomaScan platform [20], was used to test the classifiers trained on the discovery dataset.

85 On both ‘omic levels, the datasets that were used as ‘discovery’ datasets were selected due to
86 their higher number of bacterial and viral samples. There was no overlap between the patients
87 included in the proteomic datasets and those included in the transcriptomic datasets.

88 KD patients were defined according to AHA guidelines [21]. DB patients had a bacterial
89 pathogen identified in a sample from a sterile site. DV patients had a virus identified that was
90 consistent with the presenting syndrome; had no bacteria identified in blood or relevant culture
91 sites; and had a CRP of $<60\text{mg/L}$. Further details on the clinical definitions used to define the DV
92 and DB groups can be found in the Supplementary Text.

93 The median ages (months) of KD patients in the transcriptomic discovery and validation
 94 datasets were 26 (IQR: 29) and 37 (IQR: 34), respectively. The proportions of male KD patients were
 95 55% and 60% for the transcriptomic discovery and validation datasets, respectively. For the
 96 proteomic KD group, the median ages (months) were 30 (IQR: 36) and 16 (IQR: 39) for the discovery
 97 and validation datasets, respectively. The proportion of males was 69% for both the discovery and
 98 validation datasets (Table S1). Table S2 contains clinical information for the KD patients included in
 99 the four datasets analysed. The causative pathogens for the patients with bacterial and viral
 100 infections from all datasets are shown in Table S3. The median duration of fever when the blood
 101 sample was taken for transcriptomic analysis from KD patients was 5 (range of 2-7 days) and 6 days
 102 (range of 2-10 days) for the discovery and validation datasets, respectively. For the proteomic KD
 103 samples, the median duration of fever when the sample was taken was 7 (range of 3-20 days) and
 104 6.5 days (range of 4-22 days) for the discovery and validation dataset, respectively.

105 **Table 1.** The datasets used in the analysis. KD, DB, DV and HC are abbreviations for Kawasaki
 106 Disease, definite bacterial, definite viral, and healthy control, respectively. LC-MS/MS is an
 107 abbreviation for liquid chromatography with tandem mass spectrometry. * = not used in analysis.

Dataset name	GEO accession(s)	Platform(s) used for generation	KD	DB	DV	HC	Citation(s)
Transcriptomic discovery	GSE73461	Microarray: HumanHT-12 version 4.0	77	31	92	62	[18]
Transcriptomic validation	GSE73462 GSE73463	Microarrays: 1x HumanHT-12 version 3.0 1x HumanHT-12 version 4.0	101	23	28	16*	[19,22]
Proteomic discovery	NA	LC-MS/MS	26	73	75	25	<i>unpublished</i>
Proteomic validation	NA	SomaScan [20]	26	48	31	25	<i>unpublished</i>

108

109 2.2. Comparison of Kawasaki Disease to Bacterial and Viral infection

110 We explored whether the host response during KD is more similar to the host response during
 111 bacterial or viral infections using transcriptomic (gene-level) and proteomic data. We first assessed
 112 the variance in the discovery datasets using Principal Component Analysis (PCA; Fig. S1-S2). In the
 113 transcriptomic dataset, PC1 (29.24%) appears to be capturing lymphocyte number and disease
 114 group, with the KD patients located between the bacterial and viral groups. In the proteomic
 115 dataset, PC1 (29.18%) appears to be capturing variation caused by age differences, while PC2
 116 (13.39%) and PC3 (10.56%) strongly capture the disease group effects, with the KD patients grouped
 117 together between the clearly separated bacterial and viral groups.

118 2.2.1. Differential abundance analysis

119 Limma [23] was used to identify genes and proteins differentially abundant between each
 120 disease group (KD, DB, DV) and healthy controls (HC), whilst accounting for age, sex and, for the
 121 transcriptomic dataset, immune cell proportions. Features were considered significantly
 122 differentially abundant (SDA) at a FDR of 5%. Differential abundance analysis was applied to 13035
 123 genes and 344 proteins. For the transcriptomics, 3,218, 3,124, and 4,663 genes were SDA between

124 KD vs HC, DB vs HC, and DV vs HC, respectively. For the proteomics, 113, 125, and 78 proteins
125 were SDA between KD vs HC, DB vs HC, and DV vs HC, respectively. Genes and proteins SDA
126 between KD vs HC are listed in the Supplementary File 1.

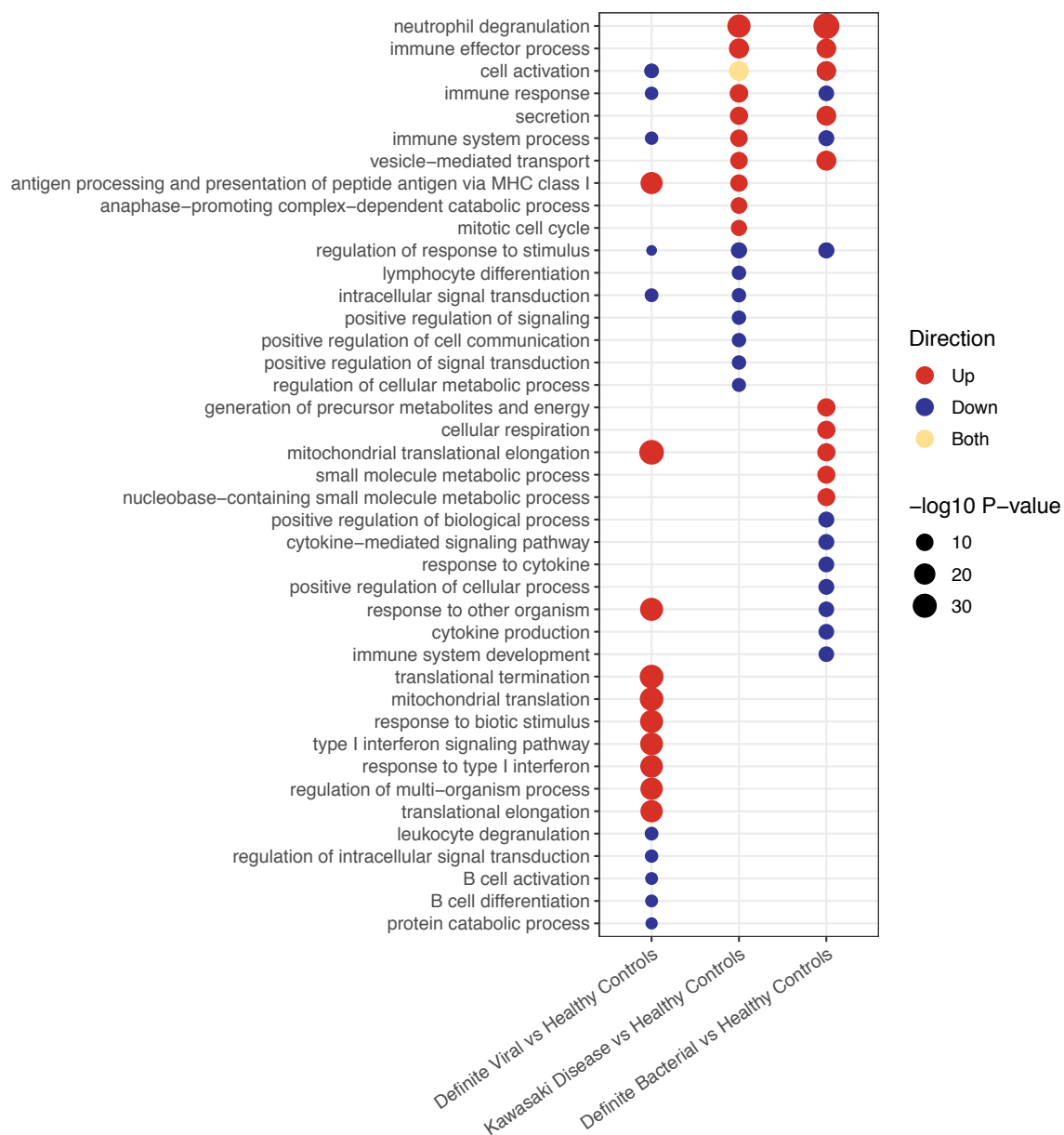
127 2.2.2. Pathway analysis

128 The lists of SDA features identified in 2.2.1 were subjected to pathway analysis using
129 g:Profiler2 [24] to determine which pathways were upregulated and downregulated in the three
130 disease groups in the discovery datasets for the transcriptomic (Fig.1a) and proteomic (Fig. 1b)
131 datasets. The full lists of pathways are provided in Supplementary File 2.

132 In the transcriptomic pathway analysis, some pathways were found to be enriched across 2 or
133 3 of the disease conditions, whereas others were found in a single condition (Fig. 1a). For example,
134 neutrophil degranulation, which was the top pathway in both KD and bacterial infections, and
135 vesicle-mediated transport were both upregulated in KD and bacterial infections, whereas antigen
136 presentation via MHC class I was upregulated in KD and viral infections. Of the top 17 pathways
137 enriched in KD, 6 pathways were also present with concordant directions in the top bacterial
138 pathways and 3 were present with concordant directions in the top viral pathways. Seven were
139 unique to KD.

140 In the proteomics data, all pathways overlapping between KD and either bacterial or viral
141 infections had concordant directions of regulation (Fig. 1b). Of the top 19 pathways enriched in KD,
142 13 were also enriched in bacterial samples, 10 in viral samples, and 4 were unique to KD. Eight of
143 the top 19 KD pathways were enriched in both bacterial and viral samples. All of these are involved
144 in the immune response. The higher frequency of overlapping concordant pathways makes it
145 harder to identify differences between the pathways enriched in the proteomic dataset than the
146 transcriptomic dataset. Overall there was a much lower number of proteins SDA between KD vs
147 HC ($n = 113$) than genes SDA between KD vs HC ($n = 3,218$). Furthermore, the total number of
148 proteins remaining following quality control and filtering for missingness ($n = 344$) was much lower
149 than the total number of genes remaining following quality control ($n = 13,035$), which could justify
150 why the differences in pathways enriched between the disease groups are more apparent in the
151 gene expression data.

152 Three pathways were enriched on both 'omic levels. These were: immune effector process
153 pathway (upregulated in KD and bacterial patients); immune response (upregulated in KD
154 patients); and vesicle-mediated transport (upregulated in KD and bacterial patients).

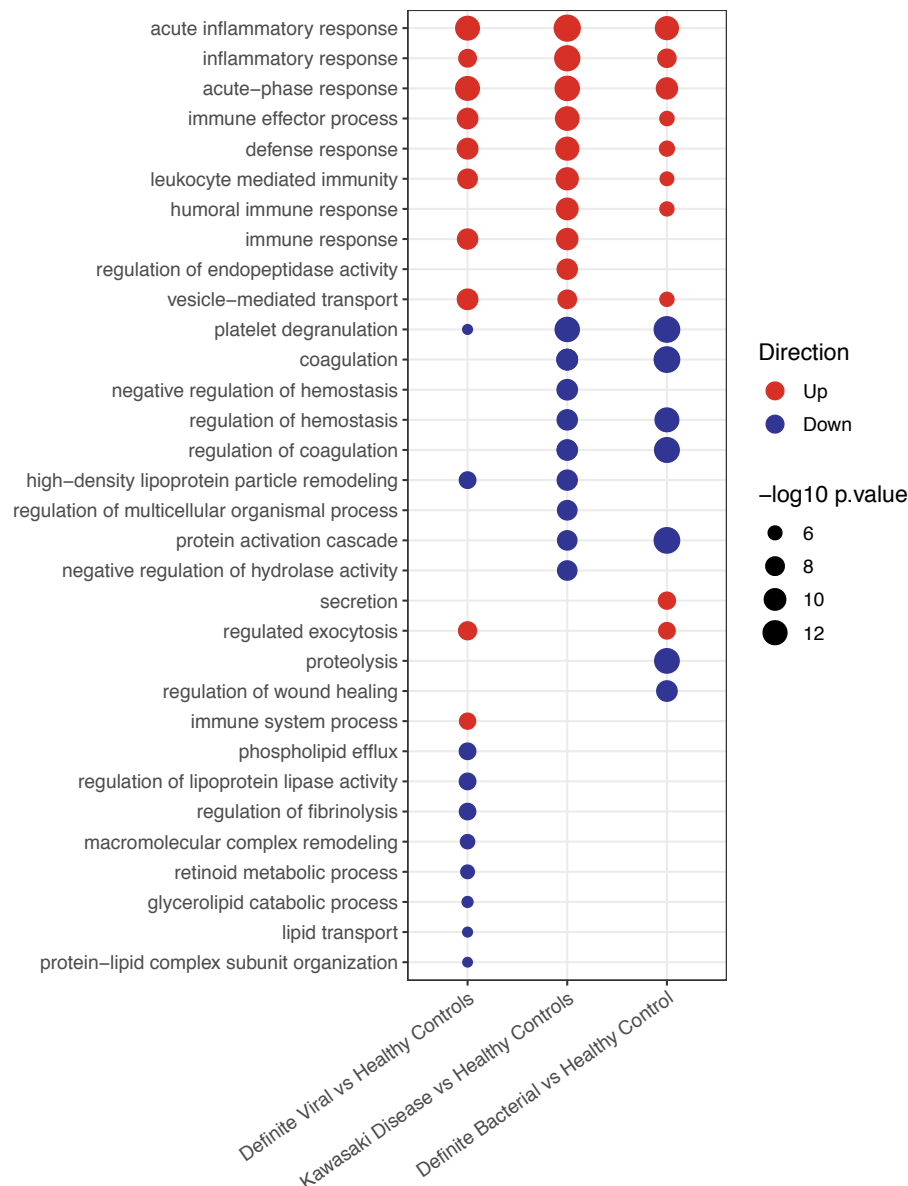


155

156

157

Figure 1a. Pathways upregulated and downregulated in bacterial, KD and viral patients compared to healthy controls in the transcriptomic dataset.



158

159 **Figure 1b.** Pathways upregulated and downregulated in bacterial, KD and viral patients compared
 160 to healthy controls in the proteomic dataset.

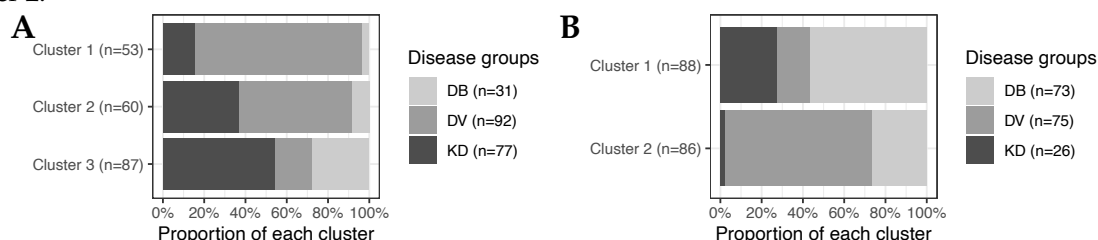
161 2.2.3. Clustering

162 K-Means clustering was used to determine whether the KD patients were more likely to cluster
 163 with bacterial or viral patients in the discovery datasets. Prior to clustering analysis, gene expression
 164 values were corrected for age, sex and immune cell proportions by taking the residual gene
 165 expression values after removing the contributions of these variables. Immune cell proportions were
 166 estimated using CIBERSORTx [25], an online tool for estimating immune cell proportions from gene
 167 expression data. The same process was performed to remove the contribution of age and sex from
 168 the protein abundance values. NbClust [26] was used to determine the optimal number of clusters
 169 (k). The value of k most frequently selected across the 12 indices measured by NbClust was selected
 170 as the optimal number of clusters for downstream analyses. In the transcriptomic analysis, 3 clusters
 171 were identified as optimal, whereas on the protein level, 2 clusters were identified as optimal.

172

173 We assessed the proportion of KD, bacterial and viral patients in each of the clusters for the
 174 transcriptomic (A) and proteomic (B) datasets (Fig. 2). In the transcriptomic analysis, an over-
 175 representation of viral patients was observed in cluster 1 and, to a lesser extent, cluster 2. An over-

176 representation of bacterial patients was observed in cluster 3 (Fig. 2), resulting in two viral-like
177 clusters and one bacterial-like cluster. Of the 77 transcriptomic KD samples, 47 (61%) belonged to
178 cluster 3, 22 (29%) belonged to cluster 2, and 8 (10%) belonged to cluster 1. In the proteomic
179 analysis, an over-representation of bacterial patients was found in cluster 1, whereas an over-
180 representation of viral patients was observed in cluster 2, leading to one viral-like and one bacterial-
181 like cluster. Of the 26 proteomic KD samples, 24 (92%) belonged to cluster 1 and 2 (8%) belonged to
182 cluster 2.



183 **Figure 2.** The proportion of patients from each disease group in each cluster for transcriptomics (A) and
184 proteomics (B). DB, DV and KD represent definite bacterial, definite viral, and Kawasaki Disease.
185

186 The association between KD patient cluster membership and various clinical variables was
187 tested. CRP levels (p-value: 0.048), lymph node swelling (p-value: 0.044), and peeling (p-value:
188 0.050) were significantly associated with cluster membership of KD transcriptomic samples. Higher
189 levels of CRP were found in transcriptomic KD samples in cluster 3 which had the highest
190 proportion of bacterial samples. Out of the 55 patients displaying peeling, 38 were found in cluster
191 3, as were 17 of the 21 patients with lymph node swelling. No clinical variables were associated
192 with cluster membership in the proteomic dataset. On both 'omic levels, CRP levels were highest in
193 the clusters in which the majority of bacterial samples were found (Fig. S5-S6), as expected since a
194 CRP cut-off of <60mg/L was required for patients in the DV groups. This pattern was also observed
195 for the WBC counts in the transcriptomic dataset (Fig. S5).

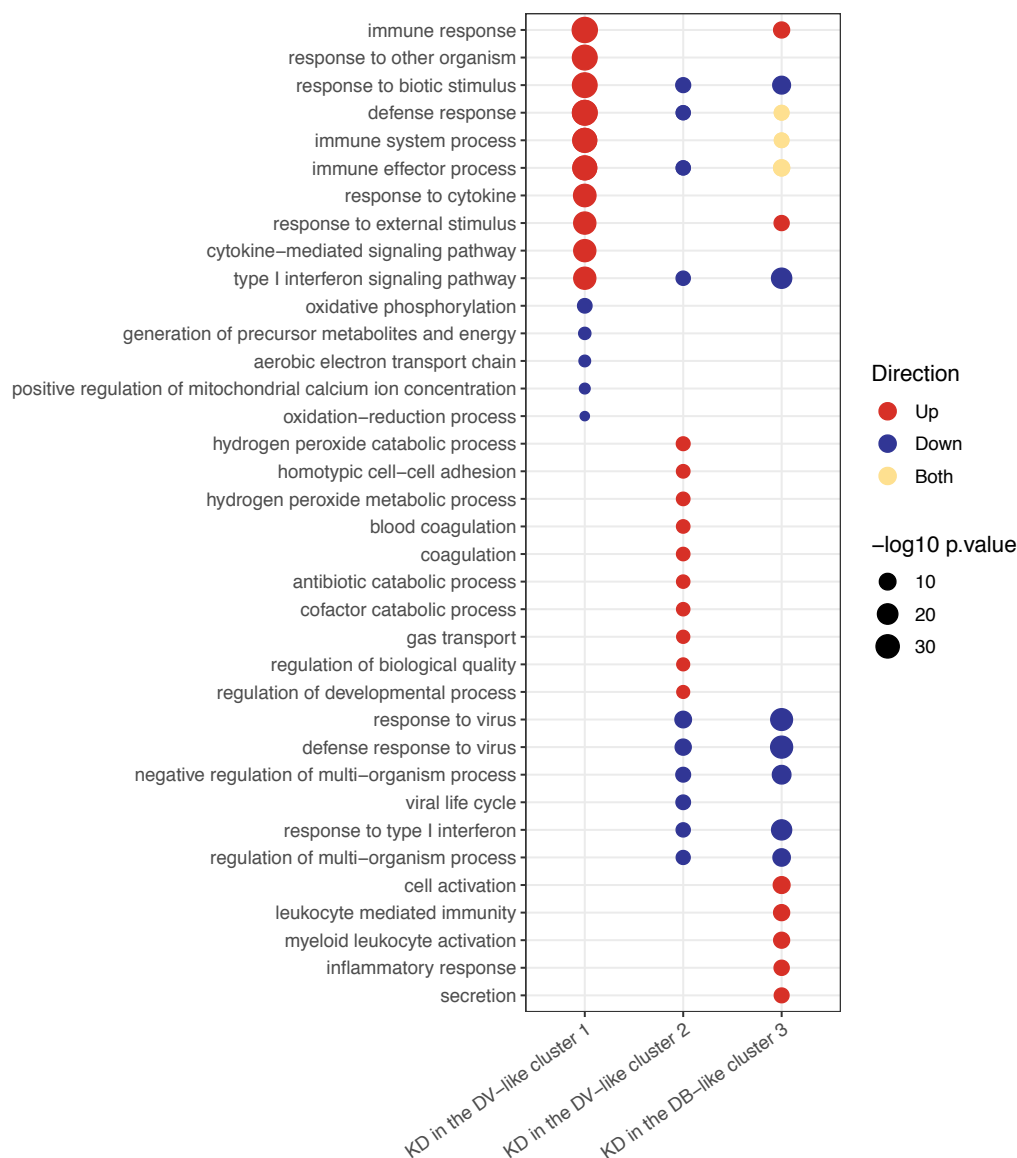
196 Differential abundance analysis was performed to compare feature abundance in the KD
197 samples that fell into different clusters. There were 503 genes SDA between transcriptomic KD
198 samples in cluster 1 vs clusters 2 and 3, 454 genes SDA between KD samples in cluster 2 vs clusters
199 1 and 3, and 651 genes SDA between KD samples in cluster 3 vs clusters 1 and 2. These lists of SDA
200 genes were subjected to pathway analysis using g:Profiler2 [24] to identify pathways upregulated
201 and downregulated within the clusters (Fig. 3). Complete lists of pathways are in Supplementary
202 File 3.

203 For the transcriptomics, cluster 1 had the highest proportion of viral patients compared to the
204 other clusters (Fig. 2a). The majority of the Adenovirus (19/23) and Influenza (16/23) patients were
205 in cluster 1. Cluster 1 KD patients were characterised by upregulation of anti-viral response
206 pathways, such as interferon and cytokine signalling (Fig. 3). In cluster 2, although the majority of
207 patients were viral, their proportion was not quite as high as it was in cluster 1 (Fig. 2). The majority
208 of the RSV (15/27) patients were in cluster 2. In the KD patients in cluster 2, various pathways
209 associated with the anti-viral response were downregulated (Fig. 3). Cluster 3 had the highest
210 proportion of bacterial patients and KD patients (Fig. 2). Similarly to cluster 2, the top pathways
211 downregulated for KD patients in cluster 3 were associated with the anti-viral response, while the
212 inflammatory response pathway was strongly upregulated suggesting that the KD patients in this
213 cluster were different to those in cluster 1 and that their response was not as viral-like as those in
214 cluster 1 (Fig. 3).

215 Three pathways - response to biotic stimulus (i.e. a stimulus caused or produced by a living
216 organism), response to other organism and type I interferon signalling - were upregulated in viral
217 transcriptomic samples (Fig. 1a) and also in the KD samples in the viral-like cluster 1 (Fig. 3).
218 Furthermore, four pathways, including two associated with interferon signaling, were upregulated
219 in viral transcriptomic samples (Fig. 1a) and downregulated in the KD samples in clusters 2 and 3

220 (Fig. 3). There were five pathways downregulated in bacterial transcriptomic samples (Fig. 1a) and
 221 upregulated in KD transcriptomic samples in cluster 1 (Fig. 3), including two related to cytokine
 222 signaling.

223 For the proteomic dataset, two proteins were SDA between cluster 1 and 2: serum amyloid A1
 224 (SAA1) and retinol binding protein 4 (RBP4). Both of these proteins have been identified previously
 225 as Kawasaki markers, with RBP4 abundance being lower in active KD [27] and SAA1 being
 226 elevated in KD [28]. The two KD patients in cluster 2 displayed the opposite pattern, with higher
 227 RBP4 and lower SAA1 abundance than the other KD patients.



228

229 **Figure 3.** Pathways upregulated and downregulated in the KD patients in clusters 1, 2 and 3 for the
 230 transcriptomic dataset. Clusters were identified using *K*-Means applied to KD, DB and DV patients.
 231 There were 151, 52 and 137 pathways upregulated in clusters 1, 2 and 3, respectively, and 5, 66 and 137
 232 pathways downregulated in clusters 1, 2 and 3, respectively.

233 2.2.4. Classification using Disease Risk Scores

234 To further assess if the KD patients elicited more bacterial-like or more viral-like responses, we
 235 built two classifiers that returned the probabilities that a patient is bacterial or viral through two
 236 separate disease risk scores (DRS). A DRS translates the abundance of features in a discriminatory

237 signature, selected by Lasso [29], into a single value that can be assigned to each individual [16].
238 Through using two independent classifiers, the possibility of a patient being neither bacterial nor
239 viral was allowed. The classifiers were trained using the 'omic data that was corrected for age, sex
240 and, for the transcriptomic dataset, immune cell proportions.

241 The Lasso model selected 38 genes for the bacterial classifier, of which 26 had increased
242 abundance and 12 had decreased abundance in bacterial patients compared to viral patients and
243 healthy controls (Table S4). The viral classifier included 32 genes, of which 13 had increased
244 abundance and 19 had decreased abundance in viral patients compared to bacterial patients and
245 healthy controls (Table S5). The classifiers trained in the transcriptomic discovery dataset were
246 tested on bacterial and viral patients from the transcriptomic validation dataset. The bacterial
247 classifier achieved an area under the ROC curve (AUC) of 0.935 (95% CI: 0.869-1) and the viral
248 classifier achieved an AUC of 0.935 (95% CI: 0.856-1).

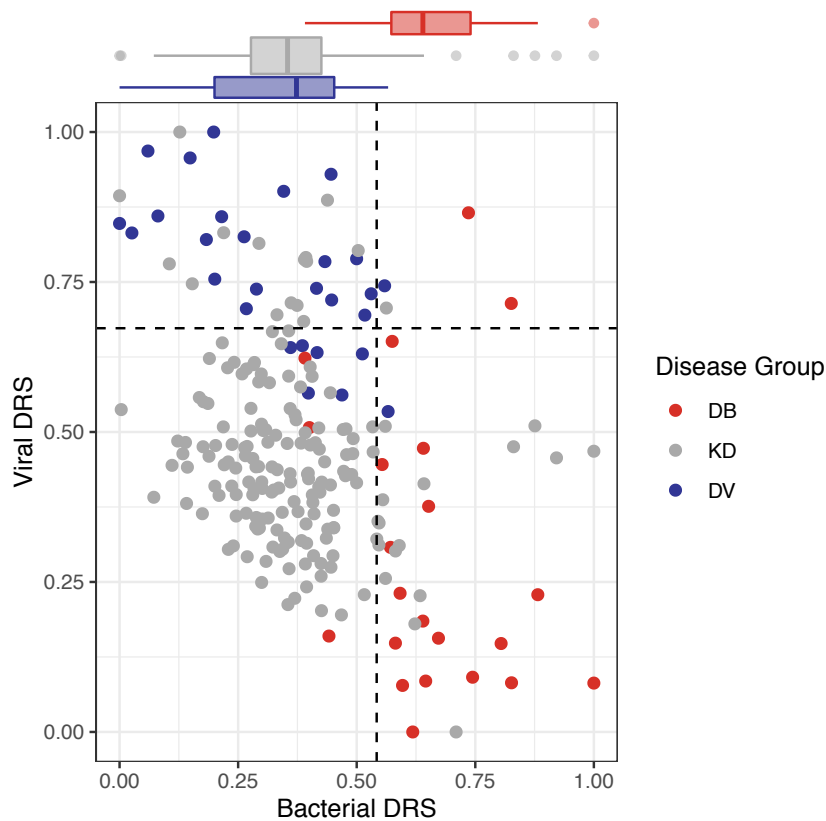
249 The Lasso model selected 26 proteins for the bacterial classifier, of which 12 had increased
250 abundance and 14 had decreased abundance in bacterial patients compared to viral patients and
251 healthy controls (Table S6). The viral classifier included 20 proteins, of which 11 had increased
252 abundance and 9 had decreased abundance in viral patients compared to bacterial patients and
253 healthy controls (Table S7). When testing the classifiers trained in the proteomic discovery dataset
254 on bacterial and viral patients from the validation dataset, the bacterial classifier achieved an AUC
255 of 0.925 (95% CI: 0.867-0.984) and the viral classifier achieved an AUC of 0.891 (95% CI: 0.821-0.962).
256 For both 'omic levels, the 90% sensitivity of the classifiers in classifying these samples was used to
257 determine the DRS threshold above which a sample would be classified as bacterial or viral.

258 The classifiers were applied to KD patients from the discovery and validation datasets for both
259 'omic levels, resulting in bacterial DRS (DB-DRS) and viral DRS (DV-DRS) for each KD patient (Fig.
260 4-5). Classification labels (DB or DV) were assigned to the KD patients using the DB-DRS and DV-
261 DRS thresholds calculated from applying the classifiers to the bacterial and viral patients in the
262 validation datasets (Fig. S9). Of the 178 transcriptomic KD samples, 18 (10%) samples had DB-DRS
263 high enough to be classified as bacterial and 16 (9%) samples had DV-DRS high enough to be
264 classified as viral. 145 (81%) samples did not achieve DB-DRS nor DV-DRS sufficiently high to lead
265 to bacterial or viral classification, and 1 sample was classified as both bacterial and viral (Fig. 4). Of
266 the 52 proteomic KD samples, 40 (78%) achieved DB-DRS high enough to be classified as bacterial
267 and 18 (35%) achieved DV-DRS high enough to be classified as viral. 10 (19%) proteomic KD
268 samples achieved DB-DRS and DV-DRS high enough for them to be classified as both bacterial and
269 viral, and 4 (7.7%) were classified as neither (Fig. 5).

270 To further examine the 'omic profiles of the KD patients with DB-DRS and DV-DRS too low for
271 them to be classified as either bacterial or viral, we performed pathway analysis on the genes or
272 proteins SDA between these KD patients and healthy controls. Amongst the pathways upregulated
273 on the transcriptomic level, were 'defense response to fungus' (p-value: 6.6e-08) and 'response
274 to fungus' (p-value: 6.7e-07).

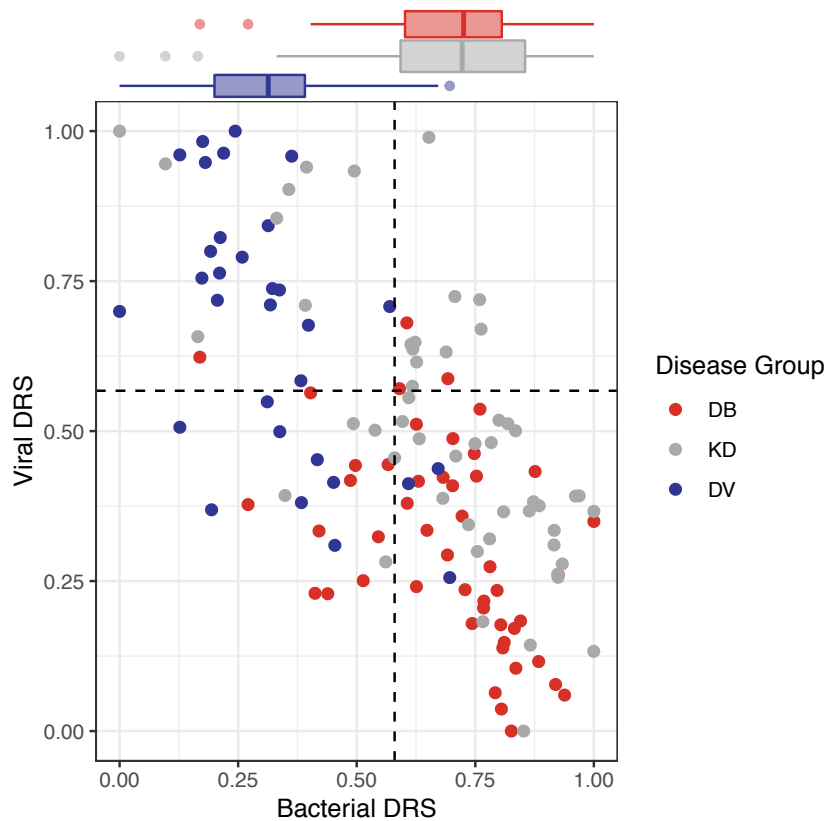
275 The associations of DB-DRS, DV-DRS, bacterial classification as predicted from the DB-DRS,
276 and viral classification as predicted from the DV-DRS, with various clinical variables were tested
277 for KD samples from both 'omic levels. In the transcriptomic KD samples, clinical measurements of
278 CRP were positively associated with DB-DRS (p-value: 0.002) and bacterial classification (p-value:
279 0.0001), and negatively associated with DV-DRS (p-value: 0.002) and viral classification (p-value:
280 0.023). In the proteomic KD samples, CRP levels were significantly positively associated with DB-
281 DRS (p-value: 0.013) and bacterial classification (p-value: 0.007). Peeling was significantly
282 associated with higher DB-DRS on both 'omic levels (transcriptomic p-value: 0.041, proteomic p-
283 value: 0.007). Strawberry tongue was significantly associated with a low score on the transcriptomic
284 DV-DRS (p-value: 0.045).

285 For the KD patients from the discovery datasets, the associations between DB-DRS or DV-DRS
286 and the cluster membership of patients was tested. Transcriptomic KD sample cluster membership
287 was significantly associated with DB-DRS (p-value: 0.005) and DV-DRS (p-value: 0.0006), with a
288 stepwise increase in DB-DRS and decrease in DV-DRS from clusters 1 to 3, where cluster 1 was the
289 most viral-like cluster, and cluster 3 was the most bacterial-like cluster. Proteomic KD sample
290 cluster membership was significantly associated with DB-DRS (p-value: 0.002) and DV-DRS (p-
291 value: 0.023), with higher DB-DRS and lower DV-DRS in KD patients in cluster 1, where cluster 1
292 was the more bacterial-like cluster and cluster 2 was the more viral-like cluster.



293

294 **Figure 4.** Bacterial DRS (DB-DRS) plotted against viral DRS (DV-DRS) for KD (discovery and validation),
295 definite bacterial (DB; validation) and definite viral (DV; validation) patients from the transcriptomic datasets.
296 Boxplots are shown for each disease group.



297

298

299

300

Figure 5. Bacterial DRS (DB-DRS) plotted against viral DRS (DV-DRS) for KD (discovery and validation), definite bacterial (DB; validation) and definite viral (DV; validation) patients from the proteomic datasets. Boxplots are shown for each disease group.

301

2.3. Clustering of Kawasaki Disease patients alone

302

303

304

305

306

307

308

309

We performed unsupervised clustering for the KD patients from the discovery datasets to explore the natural patient stratification formed in the absence of bacterial and viral comparator patients. For both 'omic levels, 3 clusters were optimal, as determined by NbClust [26]. The clusters were identified using the 'omic data that was corrected for age, sex and, for the transcriptomic dataset, immune cell proportions. Of the 77 transcriptomic KD samples, 32 (41%) were in cluster 1 (cluster KD1-T), 23 (30%) were in cluster 2 (cluster KD2-T), and 22 (29%) were in cluster 3 (cluster KD3-T). Of the 26 proteomic KD samples, 4 (15%) were in cluster 1 (cluster KD1-P), 7 (27%) were in cluster 2 (cluster KD2-P), and 15 (58%) were in cluster 3 (cluster KD3-P).

310

311

312

313

314

315

There was high overlap between the samples in cluster KD1-T and those in the transcriptomic bacterial-like cluster 3 described in 2.2.3 (Fig. S10). All except one of the samples found previously in the transcriptomic viral-like cluster 1 were found in cluster KD2-T. The majority ($n = 14$; 64%) of the samples in KD3-T were also found in transcriptomic cluster 2. On the proteome level, in 2.2.3, all KD samples except two clustered together in cluster 1, however the two remaining samples that were previously in cluster 2, were not assigned to the same cluster.

316

317

318

319

320

321

322

323

The association between cluster membership and various clinical variables was tested. CRP levels were significantly associated with cluster membership for both 'omic layers (transcriptomics p-value: 0.041, proteomics p-value: 0.010). Furthermore, coronary artery aneurysm (CAA) formation was significantly associated with cluster membership in the proteomic dataset (p-value: 0.020) with 13 of the 21 patients known to not have CAAs being in cluster KD3-P. On the transcriptomic level, the highest WBC counts and CRP levels were in cluster KD1-T, and on the proteomic level, WBC counts and CRP levels were highest in clusters KD2-P and KD1-P, respectively (Fig. S7-S8).

324 The associations between DB-DRS or DV-DRS and cluster membership of KD patients when
325 clustered alone was tested. The transcriptomic KD samples' cluster membership was significantly
326 associated with DB-DRS (p-value: 0.006) with the highest DB-DRS in cluster KD1-T. Although the
327 association between transcriptomic KD samples' cluster membership and DV-DRS was not
328 significant, the highest DV-DRS values were observed in cluster KD2-T. There were no significant
329 associations between the proteomic KD samples' cluster membership and DB-DRS or DV-DRS.

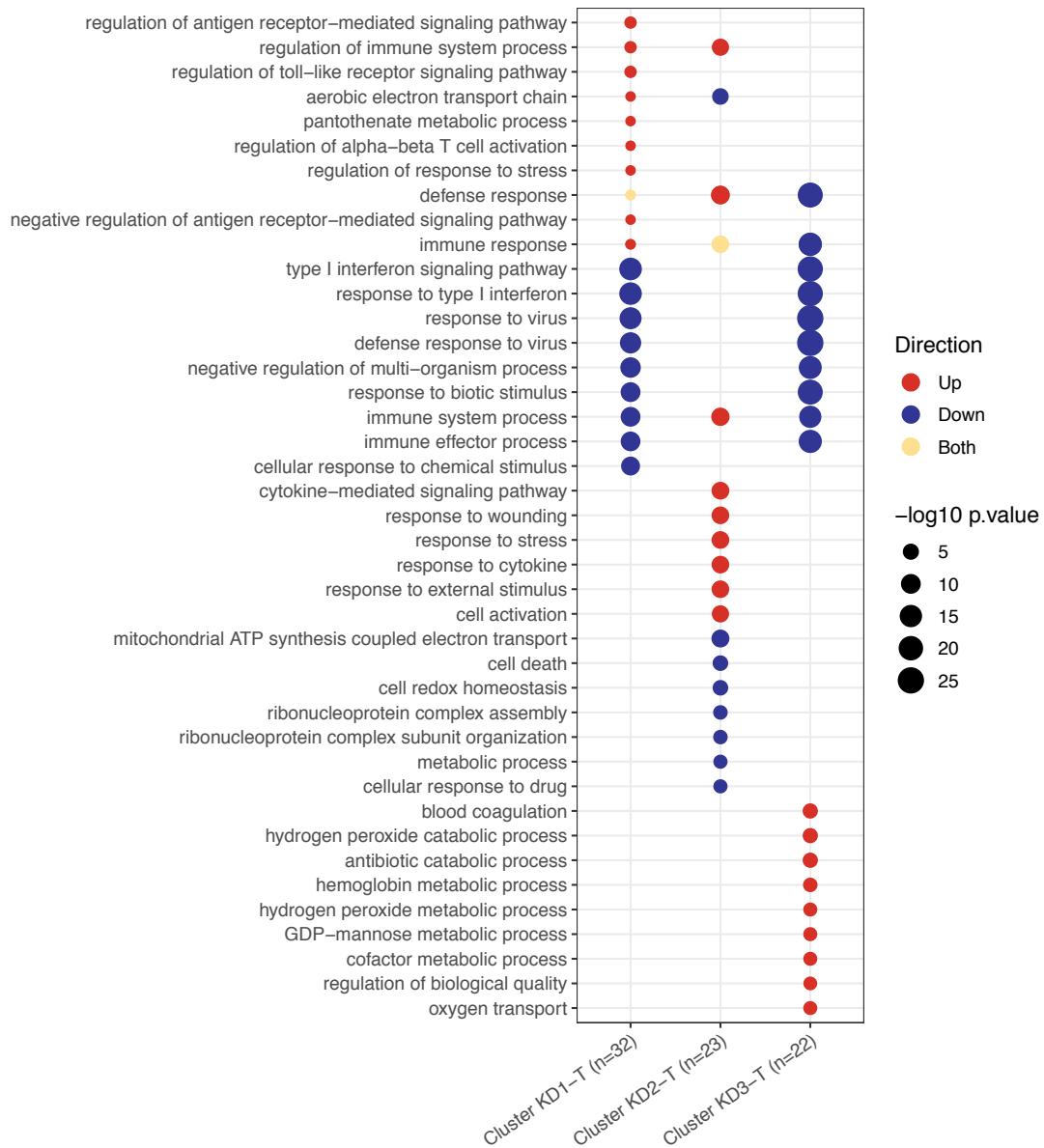
330 Differential abundance analysis was performed on the patients that fell into different clusters.
331 For the transcriptomics, there were 494 genes SDA between cluster KD1-T vs clusters KD2-T and
332 KD3-T, 461 genes SDA between cluster KD2-T vs clusters KD1-T and KD3-T, and 320 genes SDA
333 between cluster KD3-T vs clusters KD1-T and KD2-T. For the proteomics, 42 proteins were SDA
334 between cluster KD1-P vs clusters KD2-P and KD3-P, 25 proteins were SDA between cluster KD2-P
335 vs clusters KD1-P and KD3-P, and 38 proteins were SDA between cluster KD3-P vs clusters KD1-P
336 and KD2-P. These lists of SDA features were subjected to pathway analysis using g:Profiler2 [24] to
337 identify pathways upregulated and downregulated within the clusters (Fig. 6). Complete lists of
338 pathways are found in Supplementary Files 4-5.

339 In the transcriptomic analysis (Fig. 6a), cluster KD2-T had features in common with an anti-
340 viral response, whilst the others did not. Many pathways associated with the anti-viral response
341 were downregulated in clusters KD1-T and KD3-T whilst patients in cluster KD2-T were
342 characterised by the upregulation of viral pathways, including those associated with cytokine
343 signaling.

344 The response to biotic stimulus and type I interferon signaling pathways were previously
345 identified as being upregulated in viral transcriptomic samples (Fig. 1a) and in KD samples in the
346 viral-like cluster 1 when *K*-Means was applied to KD, DB and DV (Fig. 3). These pathways were
347 downregulated in clusters KD1-T and KD3-T (Fig. 6a), indicating that the transcriptomic response
348 in these samples was less viral-like than samples in cluster KD2-T.

349 Four pathways previously identified as being downregulated in bacterial transcriptomic
350 samples (Fig. 1a), including two pathways associated with cytokine signaling, were upregulated in
351 cluster KD2-T. In addition, six pathways upregulated in cluster KD2-T had already been identified
352 as being upregulated in the viral-like cluster 1 identified previously (Fig. 3). Five pathways of the 9
353 top upregulated pathways in cluster KD3-T (Fig. 6a) were also upregulated in cluster 2 when *K*-
354 Means was applied to KD, DB and DV (Fig. 3). These were blood coagulation, hydrogen peroxide
355 catabolic process, antibiotic catabolic processes, hydrogen peroxide metabolic process and
356 regulation of biological quality.

357 In the proteomic analysis (Fig. 6b), one of the KD clusters had features in common with the
358 anti-viral response, whilst another KD cluster was more bacterial-like. Amongst the top pathways
359 enriched in cluster KD1-P and KD2-P were pathways involved in inflammation. The top pathways
360 enriched in cluster KD3-P were associated with lipids. Of the 37 pathways enriched in the
361 proteomic KD samples (Fig. 6b), 21 were previously identified as enriched in proteomic samples
362 (Fig. 1b). Of these, 6 were enriched in proteomic viral samples (Fig. 1b) and samples in cluster KD2-
363 P (Fig. 6b) with concordant directions. Furthermore, 7 pathways enriched in cluster KD1-P (Fig. 6b)
364 were also enriched in bacterial proteomic samples (Fig. 1b) with concordant directions. These
365 results suggest that cluster KD1-P is a more bacterial-like cluster, whereas cluster KD2-P is a more
366 viral-like cluster. Some pathways were enriched on both 'omic levels, including those associated
367 with blood coagulation, the response to stress and immune effector processes.



368

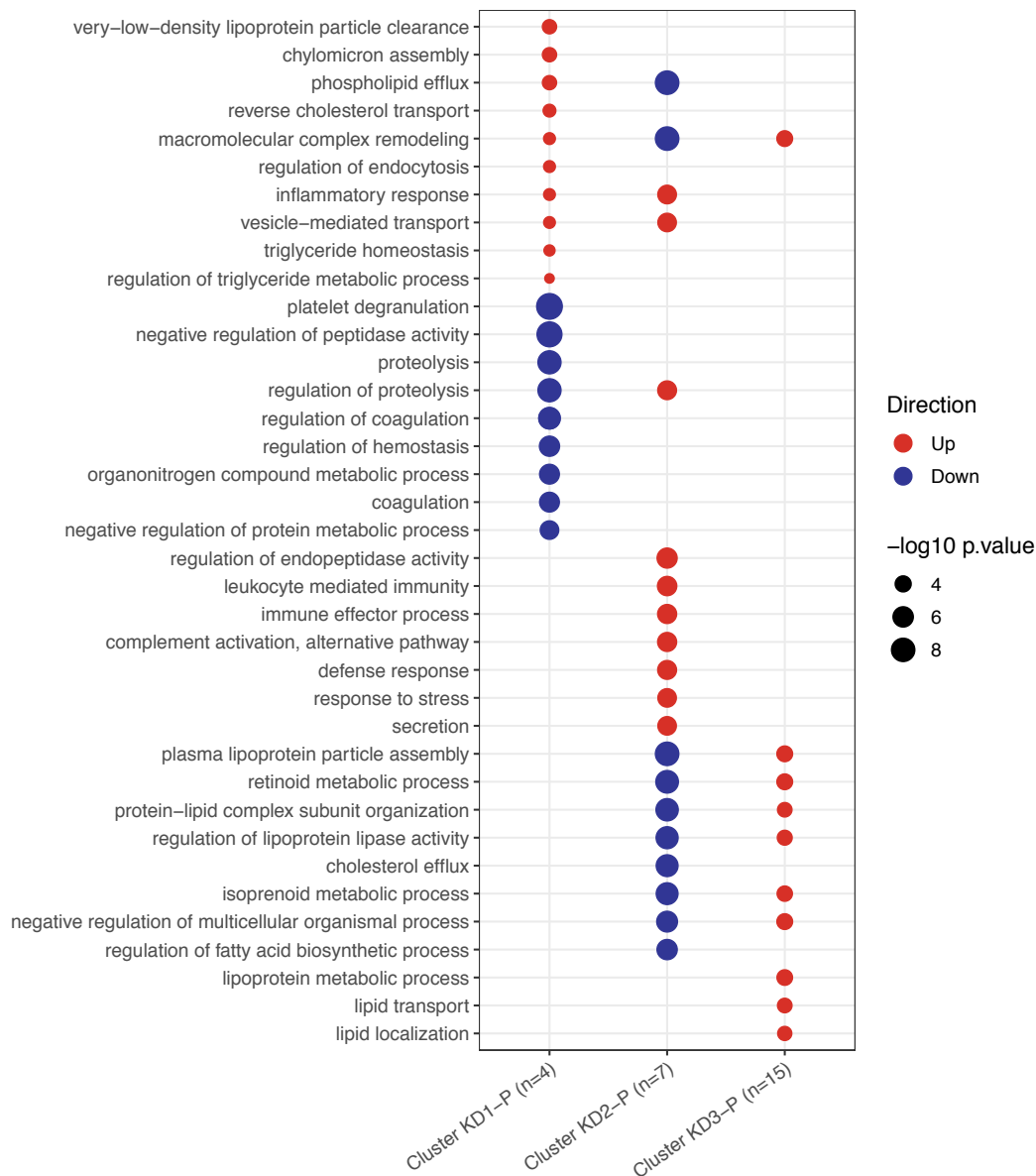
369

370

371

372

Figure 6a. Pathways upregulated and downregulated in transcriptomic KD patients between clusters. Clusters were identified by K-Means ran on KD patients alone. There were 24, 118 and 24 pathways upregulated in clusters KD1-T, KD2-T and KD3-T, respectively, and 94, 68 and 75 pathways downregulated in clusters KD1-T, KD2-T and KD3-T, respectively.



373

374 **Figure 6b.** Pathways upregulated and downregulated in proteomic KD patients between clusters. Clusters
 375 were identified by *K*-Means ran on KD patients alone. There were 77, 94 and 61 pathways upregulated in
 376 clusters KD1-P, KD2-P and KD3-P, respectively, and 64, 104 and 53 pathways downregulated in clusters KD1-
 377 P, KD2-P and KD3-P, respectively.

378 3. Discussion

379 Although the cause of Kawasaki Disease has not been identified, there is growing clinical,
 380 epidemiological and immunological evidence that it may be caused by different infectious triggers,
 381 with data pointing to bacteria, viruses or fungi. We explored the transcriptomes and proteomes of
 382 children with KD and definite bacterial and viral infections, using multiple approaches to compare
 383 the host response to these diseases to the response during KD. We found that there was a diversity
 384 of responses in the proteomic and transcriptomic profiles of KD patients, suggesting that KD is not
 385 a homogenous condition, and that whilst some patients had a more viral- or bacterial-like profile,
 386 the majority were defined as neither bacterial nor viral when their transcriptomic response was
 387 mapped onto viral and bacterial disease risk scores (DRS).

388 Within the host response profiles, some elements of KD appeared more viral-like and some
 389 elements appeared more bacterial-like. This is shown through overlapping pathways that were

390 enriched in KD and either bacterial or viral infections. For example, the antigen presentation via
391 MHC class I pathway was upregulated in KD and viral infections on the transcriptomic level. Major
392 histocompatibility complex (MHC) molecules are expressed on the cell surface to present antigenic
393 peptides to T cells, and their expression is increased by a broad range of immune activators
394 including interferons [30,31]. The finding of upregulated MHC class I expression in KD and viral
395 patients may reflect interferon-induced activation in these groups. Also, on the transcriptomic level,
396 KD and bacterial infections share neutrophil degranulation as their most upregulated pathway.
397 Neutrophils are the first responders to infection and inflammation, and the expansion and
398 activation of the neutrophil population is a characteristic feature of acute KD. In the initial days of
399 KD illness there is an intense inflammatory response with neutrophil leucocytosis [32]. Studies have
400 found elevated levels of human neutrophil elastase and IL-8, a C-X-C chemokine that activates
401 neutrophils [33,34].

402 The host response during KD is highly heterogenous, as demonstrated through the enrichment
403 of certain pathways in the KD patients in different clusters when *K*-Means was applied to KD,
404 bacterial and viral transcriptomic samples. For example, anti-viral response pathways were
405 upregulated in KD patients in the majority viral cluster 1 and downregulated in KD patients in
406 clusters with decreasing numbers of viral samples (cluster 2, 3), relative to each other. In the
407 majority bacterial cluster 3, pathways associated with the inflammatory response were upregulated.
408 The heterogeneity of the host response during KD was also apparent when *K*-Means was applied to
409 KD patients alone. Three distinct clusters were identified on both 'omic levels, and, in each cluster,
410 a distinct set of pathways was enriched. The range of pathways enriched in the different clusters
411 further demonstrate the heterogeneity in the host response during KD, with some clusters enriched
412 for viral response pathways and some clusters enriched for bacterial response pathways.
413 Unsurprisingly, amongst the patients clustering in the more bacterial-like clusters, their DB-DRS
414 tended to be higher, and amongst the patients clustering in the more viral-like clusters, their DV-
415 DRS tended to be higher.

416 The two different approaches to clustering (with and without bacterial and viral comparator
417 samples) produced similar clusters of KD patients, providing reassurance that the clusters
418 described here are biologically meaningful. Despite the similarities, however, the clusters identified
419 in 2.2.3 and 2.3 were not completely identical, indicating that the inclusion of well characterized
420 bacterial and viral patients adds further insights to the solely data-driven KD-based analysis.

421 Although there are shared features between the response to KD and both bacterial and viral
422 infections, the distinct pathways enriched in each disease group demonstrate the variation in the
423 molecular host response; the distinctiveness of the responses is also supported by the ability of
424 RNA and protein signatures to discriminate KD from bacterial and viral infections [18,19,35]. These
425 differences between the response during KD and the responses to bacterial and viral infections
426 suggest that KD may be triggered by a novel process not typical of either common bacterial or viral
427 infections. Despite the host omics profiles' heterogeneity observed in KD, commonalities are also
428 shown.

429 A two-way classifier approach highlighted that it is not a simple dichotomous question as to
430 whether the response during KD more closely resembles the responses to bacterial or viral
431 infections, when focusing on key discriminatory features. We found that 145 of the 178
432 transcriptomic KD samples were not assigned DRS high enough for them to be classified as either
433 bacterial or viral, and amongst pathways upregulated in these KD patients compared to healthy
434 controls were two pathways associated with the fungal response. This finding is intriguing, given
435 the evidence suggesting that KD could be caused by a fungal trigger that has been reported
436 elsewhere [36,37].

437 The heterogeneity and the different clusters of responses to KD which have elements shared
438 with bacterial, viral or fungal responses, could indicate multiple microbial triggers of KD, as has
439 been suggested by Rypdal et al. [11]. An alternative explanation for the heterogeneity observed here
440 in the response during KD could be that a single pathogen that causes KD leads to heterogeneous
441 responses in different hosts, as has been observed in children infected with SARS-CoV-2, where
442 many children remain asymptomatic, some experience severe inflammation [38,39], and some
443 develop PIMS-TS/MIS-C [13,14,40]. Variations in the host condition, such as epigenetic differences
444 and differences in prior pathogen exposure, could cause the spectrum of host responses to KD
445 observed here. Differences in host genetics could also be responsible for the heterogeneity in host
446 response during KD as the severity of KD, including the formation of CAAs, is already known to be
447 impacted by the host's genetic background [41].

448 This study has certain limitations. The proteomic discovery dataset was a lower dimensional
449 dataset (n= 867) than the transcriptomic discovery dataset (n=47,323) with high rates of missingness,
450 as is common in quantitative proteomics. Only proteins with no missingness were used for the
451 clustering and classification, so key proteins for distinguishing KD could be absent from the
452 analysis. On the proteomic level, many pathways were enriched in multiple disease groups (Fig.
453 1b), making it difficult to identify a disease-specific pathway signature. This could be caused by
454 plasma samples, which were used in this dataset, capturing a noisy signal due to the release of
455 substances from various tissues into the bloodstream. The proteomic response during KD shared
456 more similarities with the proteomic response to bacterial infection, with more pathways
457 overlapping between KD and bacterial infections (Fig. 1b) and all but two KD proteomic samples
458 clustering with bacterial proteomic samples (Fig. 2). This follows observations of striking clinical
459 similarities between KD and bacterial streptococcal and staphylococcal toxic shock syndromes
460 [42,43], and could reflect the hypothesis that the proteome is closer to the observed phenotype than
461 the transcriptome [44].

462 Although bacterial patients with known viral coinfections have been removed from the
463 analysis, it is impossible to say with confidence that an individual does not have a coinfection, the
464 presence of which could falsely increase heterogeneity in the host response in a given disease
465 group. Coinfection is common in KD, with one study identifying confirmed infections in a third of
466 KD patients [45]. Despite being unable to rule-out that some KD patients included had co-incident
467 viral or bacterial infections, we found that most KD transcriptomic samples were neither classified
468 as viral nor bacterial when the respective DRS scores were applied. Amongst the patients classified
469 as bacterial or viral, it is possible that some patients could be suffering from an intercurrent
470 infection in addition to KD.

471 There are variations in the range of bacterial and viral pathogens and the severity of illness
472 represented in the two 'omic datasets. The bacterial and viral patients included in the
473 transcriptomic datasets and the proteomic validation dataset were more severely unwell than those
474 included in the proteomic discovery dataset (Table S3) due to the inclusion criteria of the studies to
475 which they were recruited. The KD patients included in the transcriptomic dataset were collected
476 from San Diego, USA, whereas the KD patients included in the proteomic dataset were collected
477 from London, UK, although the same case definition was used. There remains no diagnostic test for
478 KD, thus some KD patients presented here may have unrecognized alternative diagnoses. The 2 KD
479 samples in the proteomic dataset that cluster separately (Fig. 2b) and are distinguished from the
480 other KD samples by their levels of SAA1 and RBP4, two previously identified KD markers [27,28],
481 are a possible example of this.

482

483 4. Materials and Methods

484 4.1. Patient recruitment

485 All samples were obtained from patients with written parental informed consent. Case
486 definitions can be found in the Supplementary Text. The definite bacterial (DB), definite viral (DV),
487 healthy control (HC) and Kawasaki Disease (KD) samples used in the transcriptomic discovery and
488 validation datasets were recruited in the United Kingdom and Spain as part of the IRIS
489 (Immunopathology of Respiratory, Inflammatory and Infectious Disease; NIHR ID 8209) and
490 GENDRES (Genetic, Vitamin D, and Respiratory Infections Research Network;
491 <http://www.gendres.org>) studies [17,22] and in the United States through the US-Based Kawasaki
492 Disease Research Center Program
493 (<https://medschool.ucsd.edu/som/pediatrics/research/centers/kawasaki-disease/pages/default.aspx>).

494 The DB, DV and HC samples used in the proteomic discovery and validation datasets were
495 enrolled in the EUCLIDS (European Union Childhood Life-Threatening Infectious Disease Study;
496 11/LO/1982) study [46] and the PERFORM (Personalised Risk assessment in Febrile illness to
497 Optimise Real-life Management across the European Union) study (<https://www.perform2020.org/>;
498 16/LO/1684). KD samples used in the proteomic datasets were recruited from the ongoing UK
499 Kawasaki study “Genetic determinants of Kawasaki Disease for susceptibility and outcome”
500 (13/LO/0026). This study recruits acutely unwell children with KD during hospital admission in
501 participating hospitals around the UK.

502 4.2. Data generation

503 4.2.1. Transcriptomic datasets

504 The transcriptomic discovery dataset was generated from whole blood samples obtained from
505 KD patients, healthy controls, and patients with bacterial and viral infections using the HumanHT-
506 12 version 4.0 (Illumina) microarray[18]. In order to have a transcriptomic validation dataset
507 containing the same disease groups as the transcriptomic discovery dataset, two datasets were
508 merged. One dataset contained gene expression values (HumanHT-12 version 4.0 Illumina
509 microarray) from whole blood samples obtained from acute and convalescent KD samples [19]. The
510 other dataset consisted of gene expression values (HumanHT-12 version 3.0 Illumina microarray)
511 from whole blood samples obtained from patients with bacterial and viral infections [22]. For all
512 three independent microarray experiments, one batch of samples was processed, and samples were
513 randomly positioned across the arrays.

514 4.2.1. Proteomic datasets

515 The proteomic discovery dataset was generated from plasma samples using LC-MS/MS. Full
516 details of the experimental protocol are in the Supplementary Text. The proteomic validation
517 dataset was generated from serum samples using the SomaScan aptamer-based platform [20]. Prior
518 to pre-processing, 867 proteins were measured in the discovery dataset (LC-MS/MS) and 1,300 in
519 the validation dataset (SomaScan). Samples in the proteomic validation dataset were split across
520 three plates with KD, DB, DV and HC samples present on each plate in relative proportions.

521 4.3. Statistical methods

522 All analysis was conducted using the statistical software R (R version 3.6.1, [49]). Code used for the
523 analytical pipeline described here is found at [https://github.com/heather-](https://github.com/heather-jackson/KawasakiDisease_IJMS)
524 [jackson/KawasakiDisease_IJMS](https://github.com/heather-jackson/KawasakiDisease_IJMS). Note, the code is signposted for the transcriptomic datasets but can
525 be modified for other ‘omic levels.

526 4.3.1. Pre-processing of gene expression data

527 Background correction, robust spline normalisation (RSN), and log₂-transformation were
528 applied to the raw discovery gene expression dataset using the R package lumi [47]. Probes were

529 retained if at least 80% of samples in each comparator group had a detection p-value <0.01. Low
530 variance probes and those significantly associated with the UCSD recruitment site were removed.
531 Bacterial samples with known viral coinfections were removed from the analysis at this stage to
532 ensure that the signal from the bacterial samples was not diluted. KD samples that had been
533 administered IVIG treatment were also removed at this stage, but their inclusion was irrespective of
534 coincident viral or bacterial detection, for which data was not available. A KD sample previously
535 identified as an outlier [18] was removed.

536 As mentioned, two microarray gene-expression datasets were merged to form the validation
537 dataset. Background subtraction and RSN normalisation were applied to these two datasets
538 independently, using the R package lumi [47], prior to using ComBat [48] to remove the batch effects
539 in the merged dataset [18].

540

541 4.3.2. Pre-processing of protein abundance data

542 The raw discovery dataset files generated by LC-MS/MS were processed using MaxQuant
543 (1.6.10.43) [50], with matching between runs activated. Relative quantification was performed using
544 the MaxLFQ algorithm [51]. The resulting LFQ values were log₂-transformed. Bacterial samples with
545 known viral coinfections were removed from the analysis at this stage to ensure that the signal from
546 the bacterial samples was not diluted. Protein groups were removed if they were identified as
547 contaminants, or if they were missing in over 90% of samples in each disease group.

548 The proteomic validation dataset was generated from the SomaScan platform [20]. Quality
549 control steps used scale factors returned from SomaScan to correct for variations in aptamer
550 hybridisation efficiency, inter- and intra-assay variability, variability in the starting quantities of
551 proteins, and plate effects. Further batch effect corrections were carried out using COCONUT
552 normalisation [52].

553

554 4.3.3. Comparison of Kawasaki Disease to bacterial and viral infections

555 4.3.3.1. Differential abundance analysis

556 Differential abundance analysis was carried out to compare the overall transcriptomic and
557 proteomic responses to KD, definite bacterial (DB) and definite viral (DV) infections. The degree to
558 which genes and proteins were differentially abundant between KD and healthy controls (HC), DB
559 and HC, and DV and HC was quantified using Limma [23] on the transcriptomic and proteomic
560 discovery datasets separately. Age and sex were included as covariates for both datasets. Immune
561 cell proportions, calculated using the online CIBERSORTx portal [25], were used as additional
562 covariates for the transcriptomic dataset. The immune cell proportions included were lymphocytes,
563 neutrophils, monocytes, mast cells and eosinophils. Features were considered significantly
564 differentially abundant (SDA) at a false discovery rate (FDR) [53] of 5%.

565

566 4.3.3.2. Pathway analysis

567 The pathways upregulated and downregulated in KD, DB, and DV samples were identified
568 from the lists of SDA features identified for each disease group in as outlined in 4.3.3.1. Pathways
569 were identified using g:Profiler2 [24] and redundancy in the pathways identified was removed
570 using REVIGO [54].

571 4.3.3.3. Clustering analysis

572 K-Means clustering [55] was applied separately to transcriptomic and proteomic discovery
573 datasets. Healthy controls were excluded as we were only interested in the clustering of KD with
574 pathological patients. For the proteomic dataset, only proteins with no missing data points were
575 used (n = 106).

576 To explore the effects of sex and age on clustering in the proteomic dataset, the contribution of
577 these variables was removed by regressing out their effects on every protein and taking the residual

578 values as the 'corrected' abundance. This process was also followed in the transcriptomic dataset
579 but the contributions of the immune cell proportions listed in 4.3.3.1 were also removed. Prior to
580 clustering, and after correction, features were removed if their variance was lower than 0.25. To
581 determine the optimal number of clusters (k) for each corrected and non-corrected dataset, the R
582 package NbClust [26] was used, with 12 indices tested. The indices tested were: KL [56], CH [57],
583 Hartigan [58], McClain [59], Dunn [60], SDIndex [61], SDbw [62], C-Index [63], Silhouette [64], Ball
584 [65], Ptbiserial [66,67] and Ratkowsky [68]. The number of clusters tested by NbClust ranged
585 between 2-10 clusters. The most frequently selected k by the 12 indices was used for downstream
586 analyses. The lowest k selected the most frequently was taken in cases where there were multiple
587 values of k selected the most frequently.

588 Once clusters were identified, features that were SDA (5% FDR) between KD samples in the
589 different clusters were identified. Pathway analysis was done using these lists of SDA features to
590 determine the pathways upregulated and downregulated in KD samples in the different clusters.
591 The R package g:Profiler2 [24] was used for pathway analysis, with pathways with p-values < 0.01
592 considered significant. Redundancy in the pathways identified was removed using REVIGO [54].

593 The association between cluster membership and various clinical variables was tested. For
594 categorical variables, Fisher's Exact test was used. For continuous variables, One-Way ANOVA was
595 used. P-values < 0.05 were considered significant. The categorical variables tested in both datasets
596 were: strawberry tongue (yes/no/unknown); lymph node swelling (yes/no/unknown); and peeling
597 (yes/no). Continuous variables tested in both datasets were: levels of C-reactive protein (CRP);
598 month of year; and the duration of fever at sampling. Coronary artery aneurysms (CAA) was
599 available only as a dichotomous variable for the patients submitted for proteomic analysis. For the
600 patients submitted for transcriptomic analysis, maximal coronary artery Z-scores were available
601 and were used instead.

602 4.3.3.4. Classification

603 Two independent classifiers were built for each 'omic dataset. One classifier was for classifying
604 DB patients (DB classifier), and the other for classifying DV patients (DV classifier). The DB
605 classifiers and DV classifiers were trained on features SDA between DB vs DV and HC patients
606 combined, and DV vs DB and HC patients combined, respectively. Only features present in both
607 datasets (discovery and validation) were used for training the classifiers. The discovery datasets
608 that were corrected for age, sex, and, for transcriptomics, immune cell proportions were used to
609 train the classifiers. The validation datasets were also corrected for age, sex and, for the
610 transcriptomic validation dataset, immune cell proportions as determined by CIBERSORTx [25].
611 The proteomic discovery and validation datasets were generated using different platforms.
612 Therefore, each dataset was scaled so that all abundance values were between 0-1, and then the two
613 datasets were quantile normalised together. Proteins with no missing values that were also found in
614 the proteomic validation dataset were used to train the proteomic classifier.

615 The DB classifiers were trained to identify DB patients from DV and HC patients, whereas the
616 DV classifiers were trained to identify DV patients from DB and HC patients. Lasso regularised
617 regression [29] was used to identify the discriminatory features and their weights for each classifier.
618 For each sample, a disease risk score (DRS) was calculated using the abundance of the features
619 selected by Lasso, as described by Kafrou et al. in [16]. The DRS was calculated by totalling the
620 abundance of features with positive Lasso weights and subtracting from this total the abundance of
621 features with negative Lasso weights. Features were only included in the DRS if their Lasso weight
622 direction and log-fold change direction were concordant. DRS were scaled between 0-1.

623 The classifiers were tested on the DB and DV patients of their respective validation dataset.
624 The cut-off threshold above which a sample was classified as DB or DV was calculated using the
625 coords function in the R package pROC [69] using a sensitivity cut-off of 90%. The classifiers were
626 then tested on the KD patients from the discovery and validation datasets and the thresholds

627 identified by pROC were used to determine if the KD patients were classified as DB or DV. If
628 patients were classified as neither bacterial nor viral according to their DRS, differential abundance
629 analysis followed by pathway analysis (as described in 4.3.3.2) was done to identify the pathways
630 enriched in these patients compared to healthy controls.

631 4.3.4. Exploration of Kawasaki Disease samples alone

632 In order to identify the natural clusters formed by KD patients in the absence of bacterial or
633 viral patients, K-Means clustering was done separately on the KD patients. The process followed
634 was the same as outlined in 4.3.3.3. The association between cluster membership and clinical
635 variables was tested. The clinical variables and the statistical tests used were the same as outlined in
636 4.3.3.3.

637 5. Conclusions

638 Taken together, the results from differential abundance analysis, pathway analysis, clustering
639 and classification suggest that the host transcriptomic and proteomic responses during KD are highly
640 heterogenous. Different clusters of host responses during KD were identified, some of which
641 resemble elements of host responses to bacterial, viral and fungal infections. These differences in the
642 host responses could imply that KD is triggered either by several different pathogens, or by a single
643 pathogen that has different manifestations according to the underlying genetic and environmental
644 situation of the host. Whilst there are similarities between the host response during KD and the host
645 response to bacterial infections and viral infections, there are also many differences in the responses,
646 suggesting that KD may be triggered by a novel process not typical of either common bacterial or
647 viral infections. This was demonstrated by the majority of the KD transcriptomic samples falling into
648 a non-bacterial, non-viral group following classification, raising the possibility that the minority of
649 KD transcriptomic samples with bacterial or viral profiles were possibly suffering from intercurrent
650 infection in addition to a separate KD trigger. Our data further suggest that research into the
651 etiologies of KD should be focused on cohorts of KD patients who share similar clinical characteristics
652 in order to identify shared molecular responses.

653 **Author Contributions:** Conceptualization: MK and ML; methodology: SH, HJ, AM, SM, CCP, MK; formal
654 analysis: HJ, AM, MK; investigation: SH, MJ, SM, CCP, CS, VW; resources: JCB, RG, SH, MJ, SM, CS, AHT, VW;
655 data curation: JCB, SH, JH, HJ, AM, SM, CS, VW; writing—original draft preparation: HJ; writing—review and
656 editing: ALL; validation: HJ; visualization: HJ; supervision: JH, ML, MK; project administration: HJ, CS, VW,
657 MK; funding acquisition: HJ, SM, SH, RG, AHT, MJ, TK, VW, JCB, CP, JH, ML, MK. All authors have read and
658 agreed to the published version of the manuscript.

659 **Funding:** HJ receives support from the Wellcome Trust (4-Year PhD programme, grant number 215214/Z/19/Z).
660 MK receives support from the Wellcome Trust (Sir Henry Wellcome Fellowship grant number 206508/Z/17/Z).
661 RG, JH, SH, MJ, MK, TK, ML, SM, CCP, VW acknowledge funding for the EUCLIDS and PERFORM studies,
662 funded by the European Union, grant numbers 668303 and 279185. JCB and AHT acknowledge the Marilyn and
663 Gordon Macklin Foundation and the National Institutes of Health, Heart, Lung Blood Institute RO1HL140898.
664 JH, ML and MK acknowledge support from the NIHR Imperial College BRC.

665 **Acknowledgments:** Many thanks to the patients and their families who took part in the studies that the data
666 presented here originated from.

667 **Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the
668 study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to
669 publish the results.

670 Abbreviations

CAA	Coronary artery aneurysm
DB	Definite bacterial
DRS	Disease risk score
DV	Definite viral
FDR	False discovery rate

HC	Healthy control
KD	Kawasaki Disease
LFC	Log-fold change
SDA	Significantly differentially abundant

671 References

- 672 1. Kawasaki, T.; Kosaki, F.; Okawa, S.; Shigematsu, I.; Yanagawa, H. A New Infantile Acute Febrile
673 Mucocutaneous Lymph Node Syndrome (MLNS) Prevailing in Japan. *Pediatrics* **1974**, *54*.
- 674 2. Ramphul, K.; Mejias, S.G. Kawasaki disease: a comprehensive review. *Arch. Med. Sci. - Atheroscler. Dis.*
675 **2018**, *3*, 41–45, doi:10.5114/amsad.2018.74522.
- 676 3. Ogata, S.; Shimizu, C.; Franco, A.; Touma, R.; Kanegaye, J.T.; Choudhury, B.P.; Naidu, N.N.; Kanda, Y.;
677 Hoang, L.T.; Hibberd, M.L.; et al. Treatment Response in Kawasaki Disease Is Associated with
678 Sialylation Levels of Endogenous but Not Therapeutic Intravenous Immunoglobulin G. *PLoS One* **2013**,
679 *8*, e81448, doi:10.1371/journal.pone.0081448.
- 680 4. Skochko, S.M.; Jain, S.; Sun, X.; Sivilyay, N.; Kanegaye, J.T.; Pancheri, J.; Shimizu, C.; Sheets, R.;
681 Tremoulet, A.H.; Burns, J.C. Kawasaki Disease Outcomes and Response to Therapy in a Multiethnic
682 Community: A 10-Year Experience. *J. Pediatr.* **2018**, *203*, 408–415.e3, doi:10.1016/j.jpeds.2018.07.090.
- 683 5. Brogan, P.; Burns, J.C.; Cornish, J.; Diwakar, V.; Eleftheriou, D.; Gordon, J.B.; Gray, H.H.; Johnson,
684 T.W.; Levin, M.; Malik, I.; et al. Lifetime cardiovascular management of patients with previous
685 Kawasaki disease. *Heart* **2020**, *106*, 411–420.
- 686 6. Singh, S.; Vignesh, P.; Burgner, D. The epidemiology of Kawasaki disease: A global update. *Arch. Dis.*
687 *Child.* **2015**, *100*, 1084–1088.
- 688 7. Nagata, S. Causes of Kawasaki Disease—From Past to Present. *Front. Pediatr.* **2019**, *7*, 18,
689 doi:10.3389/fped.2019.00018.
- 690 8. Dietz, S.M.; van Stijn, D.; Burgner, D.; Levin, M.; Kuipers, I.M.; Hutten, B.A.; Kuijpers, T.W. Dissecting
691 Kawasaki disease: a state-of-the-art review. *Eur. J. Pediatr.* **2017**, *176*, 995–1009.
- 692 9. Nakamura, A.; Ikeda, K.; Hamaoka, K. Aetiological significance of infectious stimuli in Kawasaki
693 disease. *Front. Pediatr.* **2019**, *7*, 244.
- 694 10. Rodó, X.; Ballester, J.; Cayan, D.; Melish, M.E.; Nakamura, Y.; Uehara, R.; Burns, J.C. Association of
695 Kawasaki disease with tropospheric wind patterns. *Sci. Rep.* **2011**, *1*, doi:10.1038/srep00152.
- 696 11. Rypdal, M.; Rypdal, V.; Burney, J.A.; Cayan, D.; Bainto, E.; Skochko, S.; Tremoulet, A.H.; Creamean, J.;
697 Shimizu, C.; Kim, J.; et al. Clustering and climate associations of Kawasaki Disease in San Diego
698 County suggest environmental triggers. *Sci. Rep.* **2018**, *8*, 1–9, doi:10.1038/s41598-018-33124-4.
- 699 12. Levin, M. Childhood Multisystem Inflammatory Syndrome — A New Challenge in the Pandemic. *N.*
700 *Engl. J. Med.* **2020**, *383*, 393–395, doi:10.1056/NEJMe2023158.
- 701 13. Whittaker, E.; Bamford, A.; Kenny, J.; Kaforou, M.; Jones, C.E.; Shah, P.; Ramnarayan, P.; Fraisse, A.;
702 Miller, O.; Davies, P.; et al. Clinical Characteristics of 58 Children with a Pediatric Inflammatory
703 Multisystem Syndrome Temporally Associated with SARS-CoV-2. *JAMA - J. Am. Med. Assoc.* **2020**, *324*,
704 259–269, doi:10.1001/jama.2020.10369.
- 705 14. Dufort, E.M.; Koumans, E.H.; Chow, E.J.; Rosenthal, E.M.; Muse, A.; Rowlands, J.; Barranco, M.A.;
706 Maxted, A.M.; Rosenberg, E.S.; Easton, D.; et al. Multisystem Inflammatory Syndrome in Children in
707 New York State. *N. Engl. J. Med.* **2020**, *383*, 347–358, doi:10.1056/NEJMoa2021756.
- 708 15. McCrindle, B.W.; Manlhiot, C. SARS-CoV-2-Related Inflammatory Multisystem Syndrome in Children:
709 Different or Shared Etiology and Pathophysiology as Kawasaki Disease? *JAMA - J. Am. Med. Assoc.*

- 710 2020, 324, 246–248.
- 711 16. Kaforou, M.; Wright, V.J.; Oni, T.; French, N.; Anderson, S.T.; Bangani, N.; Banwell, C.M.; Brent, A.J.;
712 Crampin, A.C.; Dockrell, H.M.; et al. Detection of Tuberculosis in HIV-Infected and -Uninfected
713 African Adults Using Whole Blood RNA Expression Signatures: A Case-Control Study. *PLoS Med.*
714 **2013**, *10*, e1001538, doi:10.1371/journal.pmed.1001538.
- 715 17. Kaforou, M.; Herberg, J.A.; Wright, V.J.; Coin, L.J.M.; Levin, M. Diagnosis of Bacterial Infection Using a
716 2-Transcript Host RNA Signature in Febrile Infants 60 Days or Younger. *JAMA* **2017**, *317*,
717 doi:10.1001/jama.2017.1365.
- 718 18. Wright, V.J.; Herberg, J.A.; Kaforou, M.; Shimizu, C.; Eleftherohorinou, H.; Shailes, H.; Barendregt,
719 A.M.; Menikou, S.; Gormley, S.; Berk, M.; et al. Diagnosis of Kawasaki Disease Using a Minimal
720 Whole-Blood Gene Expression Signature. *JAMA Pediatr.* **2018**, *172*, e182293,
721 doi:10.1001/jamapediatrics.2018.2293.
- 722 19. Hoang, L.T.; Shimizu, C.; Ling, L.; Naim, A.N.M.; Khor, C.C.; Tremoulet, A.H.; Wright, V.; Levin, M.;
723 Hibberd, M.L.; Burns, J.C. Global gene expression profiling identifies new therapeutic targets in acute
724 Kawasaki disease. *Genome Med.* **2014**, doi:10.1186/s13073-014-0102-6.
- 725 20. Gold, L.; Ayers, D.; Bertino, J.; Bock, C.; Bock, A.; Brody, E.N.; Carter, J.; Dalby, A.B.; Eaton, B.E.;
726 Fitzwater, T.; et al. Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS*
727 *One* **2010**, doi:10.1371/journal.pone.0015004.
- 728 21. McCrindle, B.W.; Rowley, A.H.; Newburger, J.W.; Burns, J.C.; Bolger, A.F.; Gewitz, M.; Baker, A.L.;
729 Jackson, M.A.; Takahashi, M.; Shah, P.B.; et al. Diagnosis, treatment, and long-term management of
730 Kawasaki disease: A scientific statement for health professionals from the American Heart Association.
731 *Circulation* **2017**, *135*, e927–e999, doi:10.1161/CIR.0000000000000484.
- 732 22. Herberg, J.A.; Kaforou, M.; Gormley, S.; Sumner, E.R.; Patel, S.; Jones, K.D.J.; Paulus, S.; Fink, C.;
733 Martinon-Torres, F.; Montana, G.; et al. Transcriptomic profiling in childhood H1N1/09 influenza
734 reveals reduced expression of protein synthesis genes. *J. Infect. Dis.* **2013**, *208*, 1664–1668,
735 doi:10.1093/infdis/jit348.
- 736 23. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. limma powers differential
737 expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **2015**, *43*, e47–e47,
738 doi:10.1093/nar/gkv007.
- 739 24. Raudvere, U.; Kolberg, L.; Kuzmin, I.; Arak, T.; Adler, P.; Peterson, H.; Vilo, J. g:Profiler: a web server
740 for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **2019**,
741 *47*, W191–W198, doi:10.1093/nar/gkz369.
- 742 25. Newman, A.M.; Steen, C.B.; Liu, C.L.; Gentles, A.J.; Chaudhuri, A.A.; Scherer, F.; Khodadoust, M.S.;
743 Esfahani, M.S.; Luca, B.A.; Steiner, D.; et al. Determining cell type abundance and expression from
744 bulk tissues with digital cytometry. *Nat. Biotechnol.* **2019**, *37*, 773–782, doi:10.1038/s41587-019-0114-2.
- 745 26. Charrad, M.; Ghazzali, N.; Boiteau, V.; Niknafs, A. Nbclust: An R package for determining the relevant
746 number of clusters in a data set. *J. Stat. Softw.* **2014**, doi:10.18637/jss.v061.i06.
- 747 27. Kimura, Y.; Yanagimachi, M.; Ino, Y.; Aketagawa, M.; Matsuo, M.; Okayama, A.; Shimizu, H.; Oba, K.;
748 Morioka, I.; Imagawa, T.; et al. Identification of candidate diagnostic serum biomarkers for Kawasaki
749 disease using proteomic analysis. *Sci. Rep.* **2017**, doi:10.1038/srep43732.
- 750 28. Whitin, J.C.; Yu, T.T.S.; Ling, X.B.; Kanegaye, J.T.; Burns, J.C.; Cohen, H.J. A novel truncated form of
751 serum amyloid a in kawasaki disease. *PLoS One* **2016**, *11*, doi:10.1371/journal.pone.0157024.
- 752 29. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288.

- 753 30. Danese, S. Nonimmune cells in inflammatory bowel disease: from victim to villain. *Trends Immunol.*
754 2008, 29, 555–564.
- 755 31. Tanaka, K.; Yoshioka, T.; Bieberich, C.; Jay, G. Role of the Major Histocompatibility Complex Class I
756 Antigens in Tumor Growth and Metastasis. *Annu. Rev. Immunol.* **1988**, 6, 359–380,
757 doi:10.1146/annurev.iy.06.040188.002043.
- 758 32. Tremoulet, A.H.; Jain, S.; Chandrasekar, D.; Sun, X.; Sato, Y.; Burns, J.C. Evolution of laboratory values
759 in patients with Kawasaki disease. *Pediatr. Infect. Dis. J.* **2011**, 30, 1022–1026,
760 doi:10.1097/INF.0b013e31822d4f56.
- 761 33. Biezeveld, M.H.; van Mierlo, G.; Lutter, R.; Kuipers, I.M.; Dekker, T.; Hack, C.E.; Newburger, J.W.;
762 Kuijpers, T.W. Sustained activation of neutrophils in the course of Kawasaki disease: an association
763 with matrix metalloproteinases. *Clin. Exp. Immunol.* **2005**, 141, 183–188, doi:10.1111/j.1365-
764 2249.2005.02829.x.
- 765 34. Asano, T.; Ogawa, S. Expression of IL-8 in Kawasaki disease. *Clin. Exp. Immunol.* **2000**, 122, 514–519,
766 doi:10.1046/j.1365-2249.2000.01395.x.
- 767 35. Zandstra, J.; van de Geer, A.; Tanck, M.W.T.; van Stijn-Bringas Dimitriades, D.; Aarts, C.E.M.; Dietz,
768 S.M.; van Bruggen, R.; Schweintzger, N.A.; Zenz, W.; Emonts, M.; et al. Biomarkers for the
769 Discrimination of Acute Kawasaki Disease From Infections in Childhood. *Front. Pediatr.* **2020**, 8, 355,
770 doi:10.3389/fped.2020.00355.
- 771 36. Manlhiot, C.; Mueller, B.; O’Shea, S.; Majeed, H.; Bernknopf, B.; Labelle, M.; Westcott, K. V.; Bai, H.;
772 Chahal, N.; Birken, C.S.; et al. Environmental epidemiology of Kawasaki disease: Linking disease
773 etiology, pathogenesis and global distribution. *PLoS One* **2018**, 13, doi:10.1371/journal.pone.0191087.
- 774 37. Rodó, X.; Curcoll, R.; Robinson, M.; Ballester, J.; Burns, J.C.; Cayan, D.R.; Lipkin, W.I.; Williams, B.L.;
775 Couto-Rodriguez, M.; Nakamura, Y.; et al. Tropospheric winds from northeastern China carry the
776 etiologic agent of Kawasaki disease from its source to Japan. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, 111,
777 7952–7957, doi:10.1073/pnas.1400380111.
- 778 38. Lu, X.; Zhang, L.; Du, H.; Zhang, J.; Li, Y.Y.; Qu, J.; Zhang, W.; Wang, Y.; Bao, S.; Li, Y.; et al. SARS-
779 CoV-2 Infection in Children. *N. Engl. J. Med.* **2020**, 382, 1663–1665, doi:10.1056/NEJMc2005073.
- 780 39. Göttinger, F.; Santiago-García, B.; Noguera-Julián, A.; Lanaspa, M.; Lancellata, L.; Calò Carducci, F.I.;
781 Gabrovská, N.; Velizarova, S.; Prunk, P.; Osterman, V.; et al. COVID-19 in children and adolescents in
782 Europe: a multinational, multicentre cohort study. *Lancet Child Adolesc. Heal.* **2020**, 0, doi:10.1016/S2352-
783 4642(20)30177-2.
- 784 40. Davies, P.; Evans, C.; Kanthimathinathan, H.K.; Lillie, J.; Brierley, J.; Waters, G.; Johnson, M.; Griffiths,
785 B.; du Pré, P.; Mohammad, Z.; et al. Intensive care admissions of children with paediatric
786 inflammatory multisystem syndrome temporally associated with SARS-CoV-2 (PIMS-TS) in the UK: a
787 multicentre observational study. *Lancet Child Adolesc. Heal.* **2020**, 0, doi:10.1016/S2352-4642(20)30215-7.
- 788 41. Burgner, D.; Davila, S.; Breunis, W.B.; Ng, S.B.; Li, Y.; Bonnard, C.; Ling, L.; Wright, V.J.; Thalamuthu,
789 A.; Odam, M.; et al. A genome-wide association study identifies novel and functionally related
790 susceptibility Loci for Kawasaki disease. *PLoS Genet* **2009**, doi:10.1371/journal.pgen.1000319.
- 791 42. Curtis, N.; Zheng, R.; Lamb, J.R.; Levin, M. Evidence for a superantigen mediated process in Kawasaki
792 disease. *Arch. Dis. Child.* **1995**, 72, 308–311, doi:10.1136/adc.72.4.308.
- 793 43. Han, S.B.; Lee, S.Y. Antibiotic use in children with Kawasaki disease. *World J. Pediatr.* **2018**, 14, 621–622.
- 794 44. Diz, A.P.; Martínez-Fernández, M.; Rolán-Alvarez, E. Proteomics in evolutionary ecology: Linking the
795 genotype with the phenotype. *Mol. Ecol.* **2012**, 21, 1060–1080.

- 796 45. Benseler, S.M.; McCrindle, B.W.; Silverman, E.D.; Tyrrell, P.N.; Wong, J.; Yeung, R.S.M. Infections and
797 Kawasaki disease: Implications for coronary artery outcome. *Pediatrics* **2005**, *116*, e760–e766,
798 doi:10.1542/peds.2005-0559.
- 799 46. Martínón-Torres, F.; Salas, A.; Rivero-Calle, I.; Cebey-López, M.; Pardo-Seco, J.; Herberg, J.A.;
800 Boeddha, N.P.; Klobassa, D.S.; Secka, F.; Paulus, S.; et al. Life-threatening infections in children in
801 Europe (the EUCLIDS Project): a prospective cohort study. *Lancet Child Adolesc. Heal.* **2018**, *2*, 404–414,
802 doi:10.1016/S2352-4642(18)30113-5.
- 803 47. Du, P.; Kibbe, W.A.; Lin, S.M. lumi: A pipeline for processing Illumina microarray. *Bioinformatics* **2008**,
804 doi:10.1093/bioinformatics/btn224.
- 805 48. Leek, J.T.; Johnson, W.E.; Parker, H.S.; Jaffe, A.E.; Storey, J.D. The SVA package for removing batch
806 effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **2012**, *28*, 882–883,
807 doi:10.1093/bioinformatics/bts034.
- 808 49. R Foundation for Statistical Computing R: A language and environment for statistical computing. *R A*
809 *Lang. Environ. Stat. Comput.* **3.3.1** 2016.
- 810 50. Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range
811 mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367–1372,
812 doi:10.1038/nbt.1511.
- 813 51. Cox, J.; Hein, M.Y.; Lubner, C.A.; Paron, I.; Nagaraj, N.; Mann, M. Accurate proteome-wide label-free
814 quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol.*
815 *Cell. Proteomics* **2014**, *13*, 2513–2526, doi:10.1074/mcp.M113.031591.
- 816 52. Sweeney, T.E. COCONUT: COmbat CO-Normalization Using conTrols (COCONUT). R package
817 version 1.0.2. **2017**.
- 818 53. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach
819 to Multiple Testing. *J. R. Stat. Soc. Ser. B* **1995**, *57*, 289–300.
- 820 54. Supek, F.; Bošnjak, M.; Škunca, N.; Šmuc, T. REVIGO Summarizes and Visualizes Long Lists of Gene
821 Ontology Terms. *PLoS One* **2011**, *6*, e21800, doi:10.1371/journal.pone.0021800.
- 822 55. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A K-Means Clustering Algorithm. *Appl. Stat.* **1979**, *28*,
823 100, doi:10.2307/2346830.
- 824 56. Krzanowski, W.J.; Lai, Y.T. A Criterion for Determining the Number of Groups in a Data Set Using
825 Sum-of-Squares Clustering. *Biometrics* **1988**, *44*, 23, doi:10.2307/2531893.
- 826 57. Caliński, T.; Harabasz, J. A Dendrite Method For Cluster Analysis. *Commun. Stat.* **1974**, *3*, 1–27,
827 doi:10.1080/03610927408827101.
- 828 58. Gordon, A.D.; Hartigan, J.A. Clustering Algorithms. *J. Am. Stat. Assoc.* **1976**, doi:10.2307/2286880.
- 829 59. McClain, J.O.; Rao, V.R. CLUSTISZ: A Program to Test for the Quality of Clustering of a Set of Objects.
830 *J. Mark. Res.* **12**, 456–460.
- 831 60. Dunn, J.C. Well-separated clusters and optimal fuzzy partitions. *J. Cybern.* **1974**, *4*, 95–104,
832 doi:10.1080/01969727408546059.
- 833 61. Halkidi, M.; Vazirgiannis, M.; Balisalakos, V. Quality scheme assessment in the clustering process. In
834 Proceedings of the Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial
835 Intelligence and Lecture Notes in Bioinformatics); Springer Verlag, 2000; Vol. 1910, pp. 265–276.
- 836 62. Halkidi, M.; Vazirgiannis, M. Clustering validity assessment: Finding the optimal partitioning of a data
837 set. In Proceedings of the Proceedings - IEEE International Conference on Data Mining, ICDM; 2001;
838 pp. 187–194.

- 839 63. Hubert, L.J.; Levin, J.R. A general statistical framework for assessing categorical clustering in free
840 recall. *Psychol. Bull.* **1976**, *83*, 1072–1080, doi:10.1037/0033-2909.83.6.1072.
- 841 64. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J.*
842 *Comput. Appl. Math.* **1987**, *20*, 53–65, doi:10.1016/0377-0427(87)90125-7.
- 843 65. Ball, G.H.; Hall, D.J. ISODATA, A NOVEL METHOD OF DATA ANALYSIS AND PATTERN
844 CLASSIFICATION. *undefined* **1965**.
- 845 66. Milligan, G.W. An examination of the effect of six types of error perturbation on fifteen clustering
846 algorithms. *Psychometrika* **1980**, *45*, 325–342, doi:10.1007/BF02293907.
- 847 67. Milligan, G.W. A monte carlo study of thirty internal criterion measures for cluster analysis.
848 *Psychometrika* **1981**, *46*, 187–199, doi:10.1007/BF02293899.
- 849 68. Ratkowsky, D.; Lance, G. A Criterion for Determining the Number of Groups in a Classification. *Aust.*
850 *Comput. J.* **1978**, *10*, 115–117.
- 851 69. Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.C.; Müller, M. pROC: An open-
852 source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **2011**, *12*, 77,
853 doi:10.1186/1471-2105-12-77.
- 854



© 2020 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

855