

Defining the characteristics of interferon-alpha-stimulated human genes: insight from expression data and machine-learning --Manuscript Draft--

Manuscript Number:	GIGA-D-22-00042R2	
Full Title:	Defining the characteristics of interferon-alpha-stimulated human genes: insight from expression data and machine-learning	
Article Type:	Research	
Funding Information:	Medical Research Council (MC_UU_1201412)	Prof David L. Robertson
	China Scholarship Council (201706620069)	Mr Haiting Chai
Abstract:	<p>Background: A virus-infected cell triggers a signalling cascade resulting in the secretion of interferons (IFNs), which in turn induces the up-regulation of the IFN stimulated genes (ISGs) that play a role in anti-pathogen host defence. Here, we conducted analyses on large-scale data relating to evolutionary, gene expression, sequence composition, and network properties to elucidate factors associated with the stimulation of human genes in response to IFN-α.</p> <p>Results: We find that ISGs are less evolutionary conserved than genes that are not significantly stimulated in IFN experiments (non-ISGs). ISGs show obvious depletion of GC-content in the coding region. This influences the representation of some compositions following the translation process. IFN repressed human genes (IRGs), down-regulated genes in IFN experiments, can have similar properties to the ISGs. Additionally, we design a machine-learning framework integrating the support vector machine and novel feature selection algorithm that achieves an area under the receiver operating characteristic curve (AUC) of 0.7455 for ISG prediction. Its application in other IFN-systems suggests the similarity between the ISGs triggered by type I and III IFNs.</p> <p>Conclusions: ISGs have some unique properties that make them different from the non-ISGs. The representation of some properties have strong correlations with genes' expression following IFN-α stimulation, which can be used as predictive features in machine learning. Our model predicts several genes as putative ISGs that so far have shown no significant differential expression when stimulated with IFN-α in the cell/tissue types in the available databases. A webserver implementing our method is accessible at http://isgpre.cvr.gla.ac.uk/. The docker image at https://hub.docker.com/r/hchai01/isgpre can be downloaded to reproduce the prediction.</p>	
Corresponding Author:	Joseph Hughes University of Glasgow Centre for Virus Research Glasgow, Glasgow UNITED KINGDOM	
Corresponding Author Secondary Information:		
Corresponding Author's Institution:	University of Glasgow Centre for Virus Research	
Corresponding Author's Secondary Institution:		
First Author:	Haiting Chai	
First Author Secondary Information:		
Order of Authors:	Haiting Chai	
	Quan Gu	
	David L. Robertson	
	Joseph Hughes	
Order of Authors Secondary Information:		

Response to Reviewers:	We believe these revisions will address all the concerns raised by you and the reviewers. We apologize for the misunderstanding with respect to the ROC curve and AUC values. We have added these in the legend for each dataset. We have also added ISGPRES to biotools and registered ISGPRES with scicrunch.com and received the RRID:SCR_022730, however it does not yet appear when browsing the scicrunch dashboard. We have also made a few additional clarifications to the text, in particular, after much discussion, we decided to change the 'noisy' feature terminology to poorly performing feature (this was originally called disruptive feature).
Additional Information:	
Question	Response
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics	Yes
<p>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends.</p> <p>Have you included all the information requested in your manuscript?</p>	
Resources	Yes
<p>A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?</p>	
Availability of data and materials	Yes
All datasets and code on which the conclusions of the paper rely must be either included in your submission or	

deposited in [publicly available repositories](#) (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.

Have you have met the above requirement as detailed in our [Minimum Standards Reporting Checklist](#)?

1 **Defining the characteristics of interferon-alpha-stimulated human genes:**
2 **insight from expression data and machine-learning**

3

4 Haiting Chai¹, Quan Gu¹, David L. Robertson^{1,*}, Joseph Hughes^{1,*}

5

6 ¹MRC-University of Glasgow Centre for Virus Research, Glasgow, United Kingdom

7

8 *david.l.robertson@glasgow.ac.uk, joseph.hughes@glasgow.ac.uk

9 **ORCID:**

10 Haiting Chai [0000-0003-0558-3965]; Quan Gu [0000-0002-1201-6734];

11 Joseph Hughes [0000-0003-2556-2563]; David L Robertson [0000-0001-6338-
12 0221]

13

14 **Abstract**

15 **Background:** A virus-infected cell triggers a signalling cascade resulting in the secretion of
16 interferons (IFNs), which in turn induces the up-regulation of the IFN stimulated genes (ISGs)
17 that play a role in anti-pathogen host defence. Here, we conducted analyses on large-scale data
18 relating to evolutionary, gene expression, sequence composition, and network properties to
19 elucidate factors associated with the stimulation of human genes in response to IFN- α .

20 **Results:** We find that ISGs are less evolutionary conserved than genes that are not significantly
21 stimulated in IFN experiments (non-ISGs). ISGs show obvious depletion of GC-content in the
22 coding region. This influences the representation of some compositions following the
23 translation process. IFN repressed human genes (IRGs), down-regulated genes in IFN

24 experiments, can have similar properties to the ISGs. Additionally, we design a machine-
25 learning framework integrating the support vector machine and novel feature selection
26 algorithm that achieves an area under the receiver operating characteristic curve (AUC) of
27 0.7455 for ISG prediction. Its application in other IFN-systems suggests the similarity between
28 the ISGs triggered by type I and III IFNs.

29 **Conclusions:** ISGs have some unique properties that make them different from the non-ISGs.
30 The representation of some properties have strong correlations with genes' expression
31 following IFN- α stimulation, which can be used as predictive features in machine learning. Our
32 model predicts several genes as putative ISGs that so far have shown no significant differential
33 expression when stimulated with IFN- α in the cell/tissue types in the available databases. A
34 webserver implementing our method is accessible at <http://isgpre.cvr.gla.ac.uk/>. The docker
35 image at <https://hub.docker.com/r/hchai01/isgpre> can be downloaded to reproduce the
36 prediction.

37
38 **Key words:** anti-viral response, interferon, interferon stimulated genes, omics data analyses,
39 machine-learning.

42 **Introduction**

43 Interferons (IFNs) are a family of cytokines defined for their capacity to interfere with viral
44 replication. They are secreted from host cells after an infection by pathogens such as bacteria
45 or viruses to trigger the innate immune response with the aim of inhibiting viral spread by
46 'warning' uninfected cells [1]. The response induced by IFNs is rapid and feedforward, to
47 synthesize new IFNs, which guarantees a full response even if the initial activation is limited
48 [2]. In humans, several IFNs have been discovered (e.g., IFN- $\alpha/\beta/\epsilon/\kappa/\omega/\gamma/\lambda$ [3-8]). IFN- α , IFN-

49 β , IFN- ϵ , IFN- κ , IFN- ω are grouped into type I IFNs for signalling through the common IFN-
50 α receptor (IFNAR) complex present on target cells [3-6] (**Figure 1A**). IFN- α comprises 13
51 subtypes in humans while the remaining type I IFNs are encoded by a specific gene [9]. IFN-
52 λ targets IFN- λ receptor 1 (IFNLR1)/interleukin-10 receptor 2 (IL-10R2) and was classified as
53 type III IFN following its discovery in 2003 [8] (**Figure 1C**). Similar to type I IFNs, IFN- λ
54 also exert antiviral properties but functions less intensely [10-12]. IFN- γ is classified as type II
55 IFN and manifests its biological effects by interacting with IFN- γ receptor (IFNGR) [7]
56 (**Figure 1B**). In contrast to type I and III IFNs, IFN- γ is also anti-pathogen, immunomodulatory,
57 and proinflammatory but more focused on establishing cell immunity [3,7,11,13].

58 All three types of IFNs are capable of activating the Janus kinase/signal transducer and
59 activator of transcription (JAK-STAT) pathway and inducing the transcriptional up-regulation
60 of approximately 10% of human genes that prime cells for stronger pathogen detections and
61 defence [9,14,15]. These up-regulated human genes are referred to as IFN-stimulated genes
62 (ISGs). They play an important role in the establishment of the cellular antiviral state, inhibition
63 of viral infection and return to cellular homeostasis [3,9,14,16]. For example, the ectopic
64 expression of heparinase (HPSE) can inhibit the attachment of multiple viruses [17,18];
65 interferon induced transmembrane proteins (IFITM) can impair the entry of multiple viruses
66 and traffic viral particles to degradative lysosomes [19,20]; MX dynamin like GTPase proteins
67 (MX) can effectively block early steps of multiple viral replication cycles [21]. Abnormality
68 in the IFN-signalling cascade, for example, the absence of signal transducer and activator of
69 transcription 1 (STAT1), will lead to the failure of activating ISGs, making the host cell highly
70 susceptible to virus infections [22].

71

72 **Figure 1. Illustration of signalling cascade triggered by different IFNs.** In (A), type I IFN
73 signals through IFNAR, Janus kinase 1(JAK1), tyrosine kinase 2 (TYK2), STAT, and IFN

74 regulatory factor 9 (IRF9) to form IFN stimulated gene factor 3 complex (ISGF3), and binds
75 to IFN stimulated response elements (ISRE) to induce the expression of type I ISGs. In (B),
76 type II IFN signals through IFNGR, JAK1 and JAK2 to form IFN- γ activation factor (GAF)
77 and binds to gamma-activated sequence promoter elements (GAS) to induce the expression of
78 type II ISGs. In (C), type III IFN signals through IFNLR1, IL-10R2, JAK1, TYK2, STAT, and
79 IRF9 to form ISGF3, and then bind to ISRE to induce the expression of type III ISGs. Figure
80 created using the BioRender (<https://biorender.com/>).

81

82 Most research on ISGs has focused on elucidating their role in antiviral activities or
83 discovering new ISGs within or across species [3,9,14,19,23,24]. The identification of ISGs
84 can be achieved via various approaches. Associating gene expression with suppression of viral
85 infection is a reasonable strategy to identify ISGs with obvious antiviral performance,
86 exemplified by the influenza inhibitor, MX dynamin like GTPase 1 (MX1), and the human
87 immunodeficiency virus 1 inhibitor, MX dynamin like GTPase 2 (MX2) [21]. CRISPR
88 screening is a loss-of-function experimental approach to identify ISGs required for IFN-
89 mediated inhibition to viruses. It enabled the discovery of tripartite motif containing 5 (TRIM5),
90 MX2 and bone marrow stromal cell antigen 2 (BST2) [25]. Monitoring the ectopic expression
91 of ISGs is another instrumental way to identify ISGs that are individually sufficient for viral
92 suppression [26], for example, interferon stimulated exonuclease gene 20 (ISG20) and ISG15
93 ubiquitin like modifier (ISG15). Using RNA-sequencing [27] and fold change-based criteria
94 to measure whether a target human gene is induced by IFN signalling is routinely used. In
95 most cases, a gene is defined as IFN stimulated (up-regulated) when its expression value is
96 increased in the presence of IFNs (fold change > 2).

97 There are several online databases to support IFN- or ISG-related research. For example,
98 Interferome (<http://www.interferome.org>) provides an excellent resource by compiling *in vivo*

99 and *in vitro* gene expression profiles in the context of IFN stimulation [24]. The Orthologous
100 Clusters of Interferon-stimulated Genes (OCISG, <http://isg.data.cvr.ac.uk>) demonstrates an
101 evolutionary comparative approach of genes differentially expressed in the type I IFN system
102 for ten different species [3].

103 Experimental data in the Interferome database indicate that a human gene may show
104 differential responses to different IFNs in different tissues or cells [24]. Despite some well-
105 investigated ISGs, the majority of classified ISGs have limited expression following IFN
106 stimulation [3,24]. This means that the difference between ISGs and those human genes not
107 significantly up-regulated in the presence of IFNs (non-ISGs) may not be obvious especially
108 when being assessed more generally. It should also be noted that, within non-ISGs, there are a
109 group of genes down-regulated during IFN stimulations. We refer to them as interferon-
110 repressed human genes (IRGs) and they constitute another major part of the IFN regulation
111 system [3,31]. Collectively, the complex nature of the IFN-stimulated system results in
112 knowledge that is far from comprehensive.

113 In this study, we try to associate the inherent properties of human genes with their
114 expression following IFN- α stimulation. We show that it is feasible to make ISG predictions
115 on human genes with a model only compiled from the knowledge of IFN- α responses in the
116 human fibroblast cells. To achieve this, we first constructed a refined high-confidence dataset
117 consisting of 620 ISGs and 874 non-ISGs by checking the genes across multiple databases
118 including OCISG [3], Interferome [24], and Reference Sequence (RefSeq) [32]. The analyses
119 were conducted primarily on our refined data using genome- and proteome-based features that
120 were likely to influence the expression of human genes in the presence of IFN- α (**Figure 2**).
121 Based on the calculated features, we designed a machine learning framework with an optimised
122 feature selection strategy for the prediction of putative ISGs in different IFN systems. Finally,

123 we also developed an online web server and Docker application to implement our machine
124 learning method.

125

126 **Figure 2. Diagrammatic representation of the project pipeline.** Human genes used in
127 analyses and machine learning modelling are classified based on their clinical representations
128 following IFN- α treatment in human fibroblast cells. ISGs (pink block) and non-ISGs (green
129 block) in other IFN systems are only used for testing. The figure is created using images from
130 Wikimedia Commons, <https://commons.wikimedia.org>.

131

132

133 **Results**

134 **Evolutionary characteristics of ISGs**

135 In this study, we constructed dataset S2 from 10836 well-annotated human genes (dataset S1).
136 It consists of 620 ISGs and 874 non-ISGs with high confidence based on their records in both
137 the OCISG [3] and Interferome [24]. Dataset S1 was used as the background set. Human genes
138 in this set were evolutionarily unrelated to each other as they were retrieved from the OCISG
139 [3]. Detailed information about our compiled datasets is provided in **Table 5** and
140 **Supplementary Data S1**.

141 Here, we explored features relating to alternative splicing [33], duplication [34] and
142 mutation [35]. We found that more highly upregulated human genes tended to have fewer open
143 reading frames (ORFs) (Pearson's correlation coefficient (PCC) = -0.287, **Figure 3A**),
144 transcripts (PCC = -0.407, **Figure 3B**), and protein-coding exons (PCC = -0.441, **Figure 3C**).
145 These results illustrate that alternative splicing may be linked to IFN- α up-regulation.
146 Particularly, the data points of IRGs are generally placed below those of non-ISGs, suggesting
147 these three features (number of ORFs, number of transcripts and the usage of protein-coding

148 exons) are all differentially represented in some IRGs compared to the remaining non-ISGs.
149 This distribution also indicates that some IRGs have similar feature properties to ISGs,
150 especially to those highly up-regulated in the presence of IFN- α (right part of the scatter plots
151 in **Figure 3A/B/C**).

152

153 **Figure 3. The average representation of alternative splicing features associated with IFN-**
154 **α stimulations in experiments.** (A) The numbers of ORFs and (B) transcripts are used as
155 measurements of the diversity of the alternative splicing process. (C) The count of exons used
156 for coding is used as a measurement of the complexity of alternative splicing process. These
157 three plots are drawn based on the expression data of 8619 human genes with valid fold change
158 in the IFN- α experiments (**Supplementary Data S1**). The 0.1-length sliding-window is
159 adopted to divide the data into 126 bins with different $\text{Log}_2(\text{Fold Change})$. Vertical dashed
160 lines $x=-0.871$ and $x=0.686$ are used to divide the plot into three regions. Data points in the left
161 and right regions are produced by IRG and ISGs, respectively. Data points in the middle region
162 come from ISGs or non-ISGs (including IRGs). 2217 human genes are not shown in these
163 figures as they had insufficient read coverage to determine a fold change in the experiments
164 (**Table 5**). Points in the scatter plot are located based on the average feature representation of
165 genes with similar expression performance in experiments.

166

167 To determine whether ISGs tend to originate from duplication events, we firstly
168 counted the number of human paralogs of each gene (**Figure 4A**). We found that there were
169 around 22% of singletons in our main dataset, whilst ISGs had 15% and non-ISGs had 26%.
170 The result of a Mann-Whitney U test [36] indicated that the number of human paralogs was
171 significantly under-represented in the ISGs compared to the background human genes ($M_1 =$
172 10.5 , $M_2 = 11.5$, $p = 8.8\text{E-}03$). . Next, we used the number of non-synonymous substitutions

173 (dN) and synonymous substitutions (dS) within human paralogues to measure the type and
174 strength of selection pressure acting on human genes [37]. As shown in **Figure 4B**, non-
175 synonymous substitutions are more frequently observed in the ISGs than in the background
176 human genes ($M_1 = 0.62$, $M_2 = 0.55$, $p = 4.0E-03$). On the other hand, the ISGs tend to have a
177 higher frequency of synonymous substitutions than the background human genes ($M_1 = 37.7$,
178 $M_2 = 34.6$, $p = 1.1E-02$) (**Figure 4C**) but the difference is not as obvious as for non-
179 synonymous substitutions. In **Figure 4D**, the distribution of dN/dS ratios for human paralogues
180 indicates that most human genes, including ISGs and non-ISGS, are constrained by natural
181 selection but the ISGs, in general, tend to be moderately less constrained ($M_1 = 0.036$, $M_2 =$
182 0.045 , $p = 8.3E-03$). When eliminating the influence of duplication events, the ISGs still
183 receive less selection pressure than the non-ISGs but the difference in the dN/dS ratio is not
184 significant ($M_1 = 0.053$, $M_2 = 0.031$, $p > 0.05$).

185

186 **Figure 4. Differences in the evolutionary constraints of human genes.** (A) Paralogues
187 within *Homo sapiens*. (B) Non-synonymous substitutions within human paralogues. (C)
188 Synonymous substitutions within human paralogues. (D) dN/dS ratios within human
189 paralogues. Here, the ISGs and non-ISGs are taken from dataset S2 while the background
190 human genes are from dataset S1 (**Table 5**). Mann-Whitney U tests are applied for the
191 hypothesis testing between the feature distribution of different classes. Boxes in the plot
192 represent the major distribution of values (from the first to the third quartile); outliers are added
193 for values higher than two-fold of the third quartile; cross symbol marks the position of the
194 average value including the outliers; upper and lower whiskers show the maximum and
195 minimum values excluding the outliers.

196

197 **Differences in the coding region of the canonical transcripts**

198 Compared to general profile features (e.g., number of ORFs), the sequences themselves provide
199 more direct mapping to the protein function and structure [38]. Here, we encoded 344 discrete
200 features and 7026 categorical features from complementary DNA (cDNA) of the canonical
201 transcript to explore features specific to ISGs. We divided the discrete features into four
202 categories (nucleotide compositions/dinucleotide compositions/codon usages/nucleotide 4-
203 mer compositions) and compared their representations among three different groups of human
204 genes including recompiled ISGs from dataset S2, recompiled non-ISGs from dataset S2, and
205 the background human genes from dataset S1 (**Figure 5**).

206 Firstly, guanine and cytosine were both more depleted in ISGs than non-ISGs, leading
207 to an under-representation of GC-content in the ISGs (Mann-Whitney U test: $M_1 = 52\%$, $M_2 =$
208 55% , $p = 2.3E-11$). This attribute is the opposite to the GC-biased gene conversion (gBGC)
209 process, and would result in ISGs being less stable with weak evolutionary conservation
210 (**Figure 4**) [39]. Additionally, the under-representation of GC-content also influenced the
211 representation of other dinucleotide features. Among all dinucleotide depletions in ISGs, CpG
212 depletion was ranked first followed by GpG and GpC depletions ($p = 2.9E-14$, $4.9E-13$ and
213 $1.2E-10$, respectively). In turn, adenine and thymine-related dinucleotide compositions,
214 exemplified by ApT and TpA were more enriched in ISGs than non-ISGs ($p = 8.0E-10$ and
215 $8.5E-10$, respectively).

216 We compared the usage of 64 different codons in the third category as their frequencies
217 influence transcription efficiency [40]. Differences between the ISGs and background human
218 genes were observed in codons for 11 amino acids including leucine (L), isoleucine (I), valine
219 (V), serine (S), threonine (T), alanine (A), glutamine (Q), lysine (K), glutamic acid (E),
220 arginine (R), and glycine (G). The most significant difference was observed in the usage of
221 codon 'AGA'. Among all arginine-targeted alternative codons, codon 'AGA' was usually

222 favoured, and its presence reached an estimated 25% in the ISGs but reduced to 22% in the
223 background human genes ($p = 1.4E-05$). It was significantly lower in the non-ISGs, at 18% (p
224 $= 1.9E-13$). On the other hand, compared to the background human genes, the codon ‘CAG’
225 coding for amino acid ‘Q’ was the most under-represented in the ISGs. It was less favoured by
226 the ISGs than non-ISGs ($M_1 = 72\%$, $M_2 = 78\%$, $p = 7.3E-13$) although it dominated in coding
227 patterns. As for the three stop codons, compared with the background human genes, the usage
228 of the TAA stop codon was over-represented in the ISGs ($M_1 = 28\%$, $M_2 = 33\%$, $p = 9.7E-03$).
229 In this category of codon usage, the features with different frequencies between the ISGs and
230 background human genes became more discriminating when comparing the ISGs with non-
231 ISGs. Significant differences in codon usages between the ISGs and non-ISGs were widely
232 observed except for methionine (M) and tryptophan (W). Hence, despite the limited differences
233 of codon usages between the ISGs and background human genes, these features were useful
234 for discriminating the ISGs from non-ISGs.

235 In the last category, we calculated the occurrence frequency of 256 nucleotide 4-mers
236 to add some positional resolution for finding and comparing interesting organisational
237 structures [41]. Among the 256 4-mers, 46 of them were differentially represented between the
238 ISGs and background human genes (**Supplementary Data S2**). Most of these 4-mers were
239 over-represented by the ISGs except two with the pattern ‘TAAA’ and ‘CGCG’. Interestingly,
240 the feature of ‘TAAA’ composition became a positive factor when comparing ISGs and non-
241 ISGs ($M_1 = 4.1\%$, $M_2 = 3.7\%$, $p = 4.1E-06$), suggesting it might be a suitable feature to discern
242 potential or incorrectly labelled ISGs. We found six nucleotide 4-mers: ‘ACCC’, ‘AGTC’,
243 ‘AGTG’, ‘TGCT’, ‘GACC’, and ‘GTGC’ were over-represented in the ISGs when compared
244 to the background human genes. However, they were not differentially represented when
245 comparing the ISGs with non-ISGs. These six features might be inherently biased for some
246 reason and were not powerful enough to contribute to distinguishing the ISGs from non-ISGs.

247 In addition to the aforementioned 40 features (except 4-mer ‘ACCC’, ‘AGTC’, ‘AGTG’,
248 ‘TGCT’, ‘GACC’, and ‘GTGC’) that were differentially represented in ISGs compared to
249 background human genes, we found a further 39 features nucleotide 4-mers differentially
250 represented between ISGs and non-ISGs (**Supplementary Data S2**).

251 To check the effect of these aforementioned 343 features on the level of stimulation in
252 the IFN- α system ($\text{Log}_2(\text{Fold Change}) > 0$), we calculated the PCC for the normalised features
253 (**Equation 2**) and found 106 features were positively related to the increase of fold change, and
254 34 features were suppressed when human gene were more up-regulated after IFN- α treatments
255 (Student t-test: $p < 0.05$) (**Supplementary Data S3**). ApA composition showed the most
256 obvious positive correlation with stimulation level (PCC = 0.464, $p = 8.8\text{E-}06$), while negative
257 association between the representation of 4-mer ‘CGCG’ and IFN- α -induced up-regulation was
258 the most significant (PCC = -0.593, $p = 3.2\text{E-}09$). Human genes with higher up-regulation in
259 the presence of IFN- α contained more codons ‘CAA’, rather than ‘CAG’ for coding amino acid
260 ‘Q’. The depletion of GC-content, especially cytosine content, promotes the suppression of
261 many nucleotide compositions in the cDNA, e.g., CpG composition.

262

263 **Figure 5. Differences in the representation of discrete features encoded from coding**
264 **regions (canonical).** Mann-Whitney U tests are applied for hypothesis testing on the whole
265 comparing data without sampling and the results are provided in **Supplementary Data S2**.
266 Here, the ISGs and non-ISGs are taken from dataset S2 (No. = 620 and 874) while the
267 background human genes are from dataset S1 (No. = 10836) (**Table 5**).

268

269 To find conserved sequence patterns relating to gene regulation [42], we checked the existence
270 of 2940, 44100 and 661500 short linear nucleotide patterns (SLNPs) consisting of three to five
271 consecutive nucleobases in the group of the ISGs and non-ISGs. By using a positive 5%

272 difference in the occurrence frequency as the cut-off threshold, we found 7884 SLNPs with a
273 maximum difference in representation of around 15%. After using Pearson's chi-squared tests
274 and Benjamini-Hochberg correction to avoid type I error in multiple hypotheses [43], 7025
275 SLNPs remained with an adjusted p-value lower than 0.01 (**Supplementary Data S4**), hereon
276 referred to as “flagged” SLNPs. The differentially represented 7025 SLNPs were ranked
277 according to the adjusted p-value. As shown in **Figure 6A**, dinucleotide ‘TpA’ dominates in
278 the top 10, top 100, top 1000, and all differentially represented SLNPs even if TpA
279 representation is suppressed in the cDNA of genes’ canonical transcripts compared to other
280 dinucleotides. Dinucleotide ‘ApT’ and ‘ApA’ are also frequently observed in the flagged
281 SLNPs but their occurrences do not show significant differences in the top 100 SLNPs
282 (Pearson's chi-squared test: $p > 0.05$). GC-related dinucleotides, e.g., ‘CpC’, ‘GpC’ and ‘GpG’
283 are rarely observed in the flagged SLNPs especially in the top 10 or top 100. In view of this,
284 we hypothesize that the differential representation of nucleotide compositions influences and
285 reflects on the pattern of SLNPs in the ISGs. By checking the co-occurrence status of the
286 flagged SLNPs, we found that these sequence patterns had a cumulative effect in distinguishing
287 the ISGs from non-ISGs, especially when the number of cooccurring SLNPs reached around
288 5320 (Pearson's chi-squared test: $p = 7.9E-13$, **Figure 6B**). There were eight (~1.3%) ISGs in
289 dataset S2 containing all the flagged 7025 SLNPs. Their up-regulations after IFN- α treatment
290 were generally low with a fold change fluctuating around 2.2. However, some of these eight
291 genes such as desmoplakin (DSP) were clearly highly up-regulated in endothelial cells isolated
292 from human umbilical cord veins after not only IFN- α treatments (fold change = 11.1) but also
293 IFN- β treatments (fold change = 13.7). We also found some non-ISGs (e.g., hemicentin 1
294 (HMCN1)) and human genes with limited expression in the IFN- α experiments (ELGs) (e.g.,
295 tudor domain containing 6 (TDRD6)) containing the flagged SLNPs, but their frequencies were
296 lower than that in the ISGs.

297 **Figure 6. SLNPs in the coding regions (canonical).** (A) Influence of dinucleotide
298 compositions on the flagged SLNPs. (B) The co-occurrence status of SLNPs in different human
299 genes. Ranks in (A) are generated based on the adjusted p-value given by Pearson's chi-squared
300 tests after the Benjamini-Hochberg correction procedure. Detailed results of the hypothesis
301 tests are provided in **Supplementary Data S4**. Here, the ISGs and non-ISGs are taken from
302 dataset S2 while the background human genes are from dataset S1 (**Table 5**).

303

304 **Differences in the protein amino acid sequence**

305 We used the amino acid sequences generated by the canonical transcript to extract features at
306 the proteomic level. In addition to the basic composition of 20 standard amino acids, we
307 considered 17 additional features related to physicochemical (e.g., hydrophathy and polarity) or
308 geometric properties (e.g., volume) [44,45]. We found several amino acids that were either
309 enriched or depleted in the ISG products compared to the background human proteins, which
310 were produced by genes in dataset S1 (**Figure 7**). The differences were even more marked
311 between protein products of the ISGs and non-ISGs, highlighting some differences that were
312 not observed when comparing the ISG products to the background human proteins (e.g.,
313 isoleucine composition). The differences observed in the amino acid compositions were at least
314 in part associated with the patterns previously observed in features encoded from genetic
315 coding regions. For example, asparagine (N) showed significant over-representation in the ISG
316 products compared to the non-ISG products or background human proteins (Mann-Whitney U
317 test: $p = 2.8E-12$ and $1.2E-03$, respectively). This was expected as there are only two codons,
318 i.e., 'AAT' and 'AAC' coding for amino acid 'N', and dinucleotide 'ApA' showed a
319 remarkable enrichment in the coding region of ISGs. A similar explanation could be given for
320 the relationship between the deficiency of GpG content and amino acid 'G'. The translation of
321 amino acid 'K' was also influenced by ApA composition but was not significant due to the

322 mild representation of dinucleotide 'ApG' in the genetic coding region. Additionally, as
323 previously mentioned, the ISGs showed a significant depletion in the CpG content, and
324 consequently, the amino acid 'A' and 'R' in the ISG products were significantly under-
325 represented. Cysteine (C) was not frequently observed in human proteins but still showed a
326 relatively significant enrichment in the ISG products ($M_1 = 2.3\%$, $M_2 = 2.5\%$, $p = 1.8E-03$).

327 When focusing on the composition of amino acid sequences grouped by
328 physicochemical or geometric properties, we found some features differentially represented
329 between the ISG products and background human proteins. The result showed that hydroxyl
330 (amino acid 'S' and 'T'), amide (amino acid 'N' and 'Q'), or sulfur amino acids (amino acid
331 'C' and 'M') were more abundant in the ISG products compared to the background human
332 proteins (Mann-Whitney U test: $p = 0.04$, $1.0E-03$ and 0.02 , respectively). Small amino acids
333 (amino acid 'N', 'C', 'T', aspartic acid (D) and proline (P), the volume ranging from 108.5 to
334 116.1 cubic angstroms) were more frequently observed in the ISG products than in background
335 human proteins ($M_1 = 22.1\%$, $M_2 = 21.7\%$, $p = 0.02$). These differences became more marked
336 when comparing the representation of these features between the ISG and non-ISG products.
337 For example, features relating to chemical properties of the side chain (e.g., aliphatic), charge
338 status and geometric volume showed differences between proteins produced by the ISGs and
339 non-ISGs. Some features such as neutral amino acids that include amino acid 'G', 'P', 'S', 'T',
340 histidine (H) and tyrosine (Y) were not differentially represented between the ISG and non-
341 ISG products, but they indicated an obvious association with the change of IFN- α -triggered
342 stimulations (PCC = -0.556 , $p = 4.1E-08$) (**Supplementary Data S3**).

343

344 **Figure 7. Differences in the representation of discrete features encoded from protein**
345 **sequences.** Mann-Whitney U tests are applied for hypothesis testing on the whole data without
346 sampling and the results are provided in **Supplementary Data S2**. Here, the ISGs and non-

347 ISGs are taken from dataset S2 (No. = 620 and 874) while the background human genes are
348 from dataset S1 (No. = 10836) (**Table 5**). Aliphatic group: amino acid 'A', 'G', 'I', 'L', 'P'
349 and 'V'; aromatic/huge group: amino acid 'F', 'W' and 'Y' (volume > 180 cubic angstroms);
350 sulfur group: amino acid 'C' and 'M'; hydroxyl group: amino acid 'S' and 'T';
351 acidic/negative_charged group: amino acid 'D' and 'E'; amide group: amino acid 'N' and 'Q';
352 positive_charged group: amino acid 'R', 'H' and 'K'; hydrophobic group: amino acid 'A', 'C',
353 'I', 'L', 'M', 'F', 'V', and 'W' that participates to the hydrophobic core of the structural
354 domains [46]; neutral group: amino acid 'G', 'H', 'P', 'S', 'T' and 'Y'; hydrophilic group:
355 amino acid 'R', 'N', 'D', 'Q', 'E' and 'K'; Tiny group: amino acid 'G', 'A' and 'S' (volume <
356 90 cubic angstroms); small group: amino acid 'N', 'D', 'C', 'P' and 'T' (volume ranged from
357 109 to 116 cubic angstroms); medium group: amino acid 'Q', 'E', 'H' and 'V' (volume ranged
358 within 138 to 153 cubic angstroms); large group: amino acid 'R', 'I', 'L', 'K' and 'M' (volume
359 ranged within 163 to 173 cubic angstroms); uncharged group: the remaining 15 amino acids
360 except electrically charged ones; polar group: amino acid 'R', 'H', 'K', 'D', 'E', 'N', 'Q', 'S',
361 'T' and 'Y'; nonpolar group: the remaining 10 amino acids except polar ones.

362

363 Next, we searched the sequence of the ISG products against that of the non-ISG
364 products to find conserved short linear amino acid patterns (SLAAPs), which might be
365 constrained by strong purifying selection [47]. As opposed to the analysis of the genetic
366 sequence, we only obtained 19 enriched sequence patterns with a Pearson's chi-squared p-value
367 ranging from 1.5E-04 to 0.02 (**Table 1**), hereon referred to as flagged SLAAPs. They were
368 greatly influenced by four polar amino acids: 'K', 'N', 'E' and 'S', and one nonpolar amino
369 acid: 'L'. Some of these flagged SLAAPs, for example, SLAAP 'NVT' and 'S-N-E', were
370 clearly over-represented in the ISG products compared to the background human proteins and
371 could be used as features to differentiate the ISGs from background human genes. The third

372 column in **Table 1** indicates a number of patterns that are lacking in the non-ISG products and
373 hence may be the reason for the lack of up-regulation in the presence of IFN- α . Particularly,
374 we noticed that SLAAP 'KEN' was a destruction motif that could be recognised or targeted by
375 anaphase promoting complex (APC) for polyubiquitination and proteasome-mediated
376 degradation [48,49]. Results shown in **Figure 8A** illustrate that the co-occurrence of
377 differentially represented SLAAPs (flagged) has a cumulative effect in distinguishing the ISGs
378 from non-ISGs. This cumulative effect can even be achieved with only two random SLAAPs
379 (Pearson's chi-squared test: $p = 4.6E-10$). The bias in the co-occurring SLAAPs (flagged) in
380 the background human proteins towards a pattern similar to the non-ISG products further
381 proves the importance of these 19 SLAAPs. However, their co-occurrence is not associated
382 with the level of IFN-triggered stimulations (PCC = 0.015, $p > 0.05$) (**Figure 8B**).

383 Regions that lacked stable structures under normal physiological conditions within
384 proteins are termed intrinsically disordered regions (IDRs). They play an important role in cell
385 signalling [50]. Compared with ordered regions, IDRs are usually more accessible and have
386 multiple binding motifs, which can potentially bind to multiple partners [51]. According to the
387 results calculated by IUPred [52], we found 6721, 10510, and 119071 IDRs (IUPred score no
388 less than 0.5) in proteins produced by the ISGs, non-ISGs and background human genes
389 respectively. We hypothesize that enriched SLAAPs widely detected in the IDRs may be
390 important for human protein-protein interactions or potentially virus mimicry [53]. For instance,
391 in the ISG products, about 40.8% of SLAAP 'SxNxT' were observed in the IDRs, 14.9% higher
392 than that in non-ISG products (**Table 1**). This difference reflected the importance of SLAAP
393 'SxNxT' for target specificity of IFN- α -induced protein-protein interactions (PPIs) [9] even if
394 it was not statistically significant. By contrast, the conditional frequency of SLAAP 'SxNxE'
395 in the IDRs of the ISG and non-ISG products were almost the same, indicating that SLAAP
396 'SxNxE' might have an association with some inherent attributes of the ISGs but was less likely

397 to be involved in the IFN- α -induced PPIs. SLAAP ‘KEN’ in the IDRs also showed some
398 interesting differences: in the non-ISG products, 41.9% of SLAAP ‘KEN’ were observed in
399 the IDRs, 14.6% higher than that in the ISG products, which provided an effective approach to
400 distinguish the ISGs from non-ISGs. When SLAAP ‘KEN’ is discovered in the ordered
401 globular region of a protein sequence, statistically, the protein is more likely to be produced by
402 an ISG, but this assumption is reversed if the SLAAP is located in an IDR (Pearson's chi-
403 squared tests: $p = 0.03$). Despite the relatively low conditional frequency of SLAAP ‘KEN’ in
404 the IDRs of the ISG products, these SLAAPs in the IDR are more likely to be functionally
405 active than those falling within ordered globular regions [54].

406

407 **Table 1. Representation of SLAAPs in protein sequences and their IDRs.**

SLAAP ^a	Frequency in ISG/non-ISG products ^b	Bias based on the frequency in human proteins	P value ^c	Conditional frequency in the IDRs of ISG/non-ISG products/background human proteins ^{c,d}	P value ^e
SxNxExE	15.2%/8.8%	+47.6%/-14.2%	1.5E-04	39.4%/40.3%/33.4%	0.90
ENE	15.0%/8.8%	+20.9%/-29.0%	2.1E-04	37.6%/42.9%/40.9%	0.49
SxNxTxT	11.5%/6.2%	+21.9%/-34.2%	2.9E-04	40.8%/25.9%/27.3%	0.08
SVI	15.2%/9.2%	+37.6%/-16.9%	3.6E-04	18.1%/11.3%/15.2%	0.21
LxNL	23.7%/16.4%	+13.2%/-21.9%	4.0E-04	10.2%/11.9%/9.4%	0.65
LxKL	30.8%/22.8%	+18.0%/-12.8%	4.9E-04	12.6%/10.1%/8.7%	0.43
NVT	13.7%/8.5%	+52.1%/-6.1%	1.2E-03	18.8%/21.6%/15.4%	0.66
ISS	20.5%/14.3%	+20.7%/-15.7%	1.7E-03	29.9%/25.6%/23.8%	0.44
LKxK	24.4%/17.7%	+24.5%/-9.3%	1.8E-03	14.6%/20.6%/20.0%	0.16
IKxE	14.2%/9.0%	+34.2%/-14.5%	1.8E-03	26.1%/16.5%/25.8%	0.13
EKxI	15.8%/10.4%	+31.0%/-13.7%	2.0E-03	15.3%/20.9%/16.0%	0.32
KxExS	16.9%/11.4%	+21.9%/-17.7%	2.4E-03	36.2%/36.0%/39.2%	0.98
LNS	17.7%/12.1%	+21.2%/-17.1%	2.4E-03	20.0%/25.5%/20.5%	0.34
KEN	16.0%/10.6%	+33.5%/-11.0%	2.4E-03	27.3%/41.9%/34.8%	0.03
LxNxL	22.6%/17.5%	+14.3%/-11.4%	1.5E-02	10.7%/11.8%/9.5%	0.78
KxExL	25.8%/20.5%	+25.7%/-0.3%	1.5E-02	18.8%/17.9%/18.7%	0.84
KLL	27.1%/21.9%	+9.9%/-11.4%	1.9E-02	11.3%/8.4%/9.9%	0.35
LKE	29.8%/24.5%	+18.2%/-3.0%	2.1E-02	19.5%/24.8%/20.1%	0.20
LKxL	33.2%/27.7%	+15.0%/-4.2%	2.1E-02	7.8%/12.4%/10.0%	0.11

408 *a: ‘x’ in SLAAPs indicates one position occupied by a standard amino acid;*

409 *b: here, the ISGs and non-ISGs are taken from dataset S2 while the background human genes use samples from*
410 *dataset S1 (Table 5);*

411 *c: p values in this column use Pearson's chi-squared tests to measure the difference of SLAAPs occurrences in*
412 *the ISG and non-ISG products;*

413 *d: frequencies in this column are calculated based on a condition that corresponding SLAAPs are observed in*
414 *the protein sequence;*

415 *e: p values in this column use Pearson's chi-squared tests to measure the difference of SLAAPs occurrences in*
416 *the IDRs of the ISG and non-ISG products.*

417

418 **Figure 8. Representation of co-occurring SLAAPs (flagged) in our main dataset.** (A) The
419 co-occurrence status of SLAAPs in different classes. (B) Relationship between co-occurrence
420 of the marked SLAAPs and $\text{Log}_2(\text{Fold Change})$ after IFN- α treatments. Here, the ISGs and
421 non-ISGs are taken from dataset S2 while the background human genes are from dataset S1
422 (Table 5). Points in (B) are located based on the average feature representation of genes with
423 similar expression performance in IFN- α experiments.

424

425 Differences in network profiles

426 We constructed a network with 332,698 experimentally verified interactions among 17603
427 human proteins (confidence score > 0.63) from the Human Integrated Protein-Protein
428 Interaction rEference (HIPPIE) database [55] to investigate if the connectivity among human
429 proteins has an association with genes' expression in the IFN- α experiments. 10169 out of
430 10836 human proteins produced by genes in our background dataset S1 were included in the
431 network. Nodes and edges of this network can be downloaded from our webserver at
432 <http://isgpre.cvr.gla.ac.uk/>. Based on this network, we calculated eight features as defined in
433 the methods including the average shortest path, closeness, betweenness, stress, degree,
434 neighbourhood connectivity, clustering coefficient, and topological coefficient.

435 As illustrated in Figure 9B/G, ISG products tend to have higher values of betweenness
436 and stress than background human proteins (Mann-Whitney U test: $p = 0.01$, and 0.03 ,
437 respectively), which means they are more likely to locate at key paths connecting different
438 nodes of the PPI network. Some ISG products with high values of betweenness and stress, e.g.,
439 tripartite motif containing 25 (TRIM25), can be considered as the shortcut or bottleneck of the
440 network and play important roles in many PPIs including those related to the IFN- α -triggered

441 immune activities [56,57]. However, such differential representation of betweenness does not
442 mean ISG products are more likely to be or even be close to bottlenecks of the network
443 compared to the background human proteins. Some examples shown in **Table 2** indicate that
444 ISG products are less-connected by top-ranked bottlenecks and hubs of the network than non-
445 ISG products or the background human proteins. This conclusion is not influenced by the
446 hub/bottleneck protein's performance in the IFN- α experiments. Comparing proteins produced
447 by the ISGs and non-ISGs, we found the former tends to have lower values of clustering
448 coefficient and neighbourhood connectivity (Mann-Whitney U test: $p = 0.04$ and $7.9E-03$,
449 **Figure 9D/F**). This discovery indicates that the ISG products and some of their interacting
450 proteins are less likely to be targeted by lots of proteins. It also supports the finding that the
451 ISG products are involved in many shortest paths for nodes but are away from hubs or
452 bottlenecks in the network. To some extent, this location also increases the length of the
453 average shortest paths through ISG products in the network (**Figure 9A**).

454 When investigating the association between IFN- α -induced gene stimulation and
455 network attributes of gene products, we only found the feature of neighbourhood connectivity
456 was under-represented as the level of differential expression in the presence of IFN increases
457 (PCC = -0.392, $p = 2.2E-04$). This suggests that proteins produced by genes that are highly up-
458 regulated in response to IFN- α are further away from hubs in the PPI networks.

459

460 **Figure 9. Differences in network preferences.** The included features are (A) average shortest
461 path (B) betweenness, (C) closeness, (D) clustering coefficient, (E) degree, (F) neighbourhood
462 connectivity, (G) stress, and (H) topological coefficient. Mann-Whitney U tests are applied for
463 hypothesis testing on the whole comparing data without sampling and the results were provided
464 in **Supplementary Data S2**. Here, the ISGs and non-ISGs are taken from dataset S2 (No. =

465 620 and 874) while the background human genes use samples from dataset S1 (No. = 10836)
 466 (Table 5).

467

468 **Table 2. Interaction profiles of human proteins connecting top hubs/bottlenecks of the**
 469 **HIPPIE network.**

Human protein	TRIM25	ELAVL1	ESR2	NTRK1
Gene class	ISG	IRG	Not included in S1 ^a	
Degree (hub rank)	2295 (2nd)	1787 (4th)	2500 (1st)	1976 (3rd)
Betweenness (bottleneck rank)	0.067 (1st)	0.048 (4th)	0.051 (3rd)	0.026 (5th)
Difference in interacting partners (ISG products versus non-ISG products) ^b	Depleted P = 0.01	P > 0.05	Depleted P = 1.1E-4	Depleted P = 5.5E-3
Difference in interacting partners (ISG products versus the background human proteins) ^b	P > 0.05	P > 0.05	Depleted P = 8.1E-3	Depleted P = 0.03

470 *a: ESR2 and NTRK1 were not included in dataset S1 as their expression data were not compiled in OCISG;*

471 *b: differences here are measured via Pearson's chi-squared tests on human proteins interacting with the*
 472 *corresponding hub/bottleneck protein.*

473

474 **Features highly associated with the level of IFN stimulations**

475 In this study, we encoded a total of 397 discrete and 7046 categorical features covering the
 476 aspects of evolutionary conservation, nucleotide composition, transcription, amino acid
 477 composition, and network profiles. In order to find out some key factors that may enhance or
 478 suppress the stimulation of human genes in the IFN- α system, we compared the representation
 479 of discrete features of human genes with different but positive Log₂(Fold Change). Two
 480 features on the co-occurrence of SLNPs and SLAAPs were not taken into consideration here
 481 as they were more subjective than the other discrete features and were greatly influenced by
 482 the number of sequence patterns. Upon the calculation of PCC and the result of hypothesis
 483 tests, we found 168 features highly associated with the level of IFN- α -triggered stimulations
 484 (Student t-tests: $p < 0.05$) (**Supplementary Data S3**). Among them, 118 features showed a
 485 positive correlation (**Figure 10**) while the remaining 50 features showed a negative correlation
 486 (**Figure 11**) with the change of up-regulation in IFN- α experiments. Among these 168 features,
 487 the number of ORFs, alternative splicing results, and counts of exons used for coding were

488 encoded from characteristics of the gene. Average dN/dS and average dS within human
489 paralogues were encoded based on the sequence alignment results from Ensembl [58]. 140
490 and 22 features were encoded from the genetic sequence and proteomic sequence respectively.
491 The last one, neighbourhood connectivity, was obtained from the network profile of a human
492 interactome constructed based on experimentally verified data in the HIPPIE database [55].

493 In the positive group, the feature of 'large' amino acid compositions that includes the
494 composition of five amino acids with geometric volume ranging from 163 to 173 cubic
495 angstroms was ranked first for having the highest PCC at 0.593 (Student t-test: $p = 2.8E-09$).
496 This feature was not highlighted previously as it did not have a strong signal for discriminating
497 the ISGs from non-ISGs (Mann-Whitney U test: $p > 0.05$). Similar phenomena were found on
498 87 features (64 positive correlations and 23 negative correlations) such as AG-content, ApG
499 content and previously mentioned neutral amino acid composition. The strongest negative
500 correlation between feature representation and IFN- α -triggered stimulations was found on the
501 feature of 4-mer 'CGCG' (PCC = -0.593, $p = 3.2E-09$). This feature also showed a differential
502 distribution between the ISGs and non-ISGs, providing useful information to distinguish the
503 ISGs from non-ISGs. Similar phenomena were found on 81 features (54 positive correlations
504 and 27 negative correlations) such as previously mentioned GC-content, CpG content and the
505 usage of codon 'GCG' coding for amino acid 'A'.

506 Collectively, the biased effect on the basic composition of nucleotide sequences
507 influences the correlation between the representation of sequence-based features and IFN- α -
508 triggered stimulations. Human genes that show over-representation in more features listed in
509 **Figure 10** are expected to be more up-regulated after IFN- α treatments at least in the human
510 fibroblast cells. Meanwhile, the under-representation of features listed in **Figure 11** also
511 contributes to the level of up-regulation in the IFN- α experiments.

512

513 **Figure 10. 118 features positively associated with higher up-regulation after IFN- α**
514 **treatments.** Features here are screened based on the PCC and results of Student t-tests ($p <$
515 0.05). Features with higher PCC indicate stronger positive correlations. Detailed results about
516 PCC and hypothesis tests are provided in **Supplementary Data S3**.

517

518 **Figure 11. 50 features negatively associated with higher up-regulation after IFN- α**
519 **treatments.** Features here are screened based on the PCC and results of Student t-tests ($p <$
520 0.05). Features with lower PCC indicate stronger negative correlations. Detailed results about
521 PCC and hypothesis tests are provided in **Supplementary Data S3**.

522

523 **Difference in feature representation of interferon-repressed genes and genes with low**
524 **levels of expression**

525 We grouped human genes into two classes based on their response to the IFN- α in the human
526 fibroblast cells. Genes significantly up-regulated in IFN- α experiments were included in the
527 ISG class, while those that did not were put into the non-ISG class. However, there is also
528 another group of human genes down-regulated in the presence of IFN- α , i.e., the IRGs. They
529 were labelled as the non-ISGs, but contain unique patterns that constitute an important aspect
530 of the IFN response [3]. Some of these IRGs were not up-regulated in any known type I IFN
531 systems, thus they have been placed in a refined non-ISG class for analyses and predictions.
532 Additionally, there are a number of genes that have insufficient levels of expression in the
533 experiments to determine a fold change, i.e., ELGs. Here, we used the previously defined
534 features to compare the ISGs from dataset S2 with the IRGs and ELGs divided from the
535 background dataset S1 (**Table 5**).

536 As shown in **Figure 12**, the IRGs are differentially represented to a lower extent in the
537 majority of nucleotide 4-mer compositions than the ISGs, indicating the deficiency of some

538 nucleotide sequence patterns in the coding region of IRGs. Note that, many nucleotide 4-mer
539 composition features are more suppressed in the ISGs than non-ISGs although the differences
540 are small. The biased representation of these features in the IRGs suggests that the IRGs have
541 characteristics similar to the ISGs rather than non-ISGs. Additionally, there are a very limited
542 number of features relating to evolutionary conservation, nucleotide sequence compositions or
543 codon usage showing obvious differences between the ISGs and IRGs, but many of them are
544 differentially represented when comparing the ISGs with non-ISGs. Therefore, involving the
545 IRGs in the class of the non-ISGs will increase the risk for machine learning models to produce
546 more false positives. However, there are some informative features differentiating the IRGs
547 from ISGs. For example, compared to the ISGs, the IRGs are more enriched in CpGs (Mann-
548 Whitney U test: $p = 5.6E-03$), which is also mentioned in [59]. The IRGs tend to have higher
549 closeness centrality and neighbourhood connectivity than the ISGs (Mann-Whitney U test: $p =$
550 0.04 and $6.4E-06$ respectively), suggesting that the IRGs are more central in the human PPI
551 network and connected to key proteins with many interaction partners. Differences in some
552 amino acid composition features between the ISGs and IRGs can also be observed in **Figure**
553 **12**. Therefore, good predictability is still expected when using features extracted from protein
554 sequences.

555 **Figure 12** illustrates 161 features showing significant differences (Mann-Whitney U
556 tests: $p < 0.05$) in the representation of the ISGs and ELGs. An estimated 82% of these features
557 were also differentially represented between the ISGs and non-ISGs. 79% of these significant
558 features displayed similar over-representation or under-representation in two comparisons, i.e.,
559 ISGs versus ELGs and ISGs versus non-ISGs. These ratios indicate that the majority of the
560 ELGs are less likely to be ISGs based on their feature profile as well as their low expression
561 levels in cells induced with IFN- α . Network analyses showed that the ELG products tended to
562 have lower values of all calculated network features than ISG products with the exception of

563 topological coefficient. This means that the ELG products are less connected to other human
564 proteins in the human PPI network. Particularly, their abnormal representation on the feature
565 of average shortest paths indicates that some ELGs (e.g., vascular cell adhesion molecule 1
566 (VCAM1) and ubiquitin D (UBD)) may still have high connectivity in the human PPI network.
567

568 **Figure 12. Differential expressions of discrete features between different genes and their**
569 **coded proteins.** Mann-Whitney U tests are applied for hypothesis testing on the whole
570 comparing data without sampling and the results were provided in **Supplementary Data S2**.
571 Here, the ISGs and non-ISGs are taken from dataset S2 (No. = 620 and 874); the IRGs and
572 ELGs are taken from dataset S4 (No. = 1006) and dataset S8 (No. = 2217); the background
573 human genes are from dataset S1 (No. = 10836) (**Table 5**).
574

575 **Implementation with machine learning framework**

576 In this study, we encoded 397 discrete and 7046 categorical features for the analyses. As excess
577 of features will greatly increase the dimension of feature spaces and complicate the
578 classification task for the classifier, we limited the number of SLNPs to the top 100 based on
579 the adjusted p-value and we expected these to be sufficient to provide a picture of short linear
580 sequence patterns in the coding region of the canonical transcript. Accordingly, features
581 measuring the co-occurrence status of multiple SLNPs were recalculated based on the selected
582 100 SLNPs. As a result, we prepared 518 features (**Supplementary Data S5**) for our machine
583 learning framework. To reduce the impact of noisy data on classifications, we only used the
584 refined ISGs and non-ISGs from dataset S2 for training and modelling.

585 **Table 3** firstly shows the comparisons of prediction performance among different
586 machine learning methods. The threshold is determined by maximising the value of MCC. As
587 the random forest (RF) classifier was built based on randomly selected samples and features

588 [60], we repeated its modelling procedure ten times. These initial comparisons showed that the
589 support vector machine (SVM) [61] is superior to k-nearest neighbors (KNN) and RF [60]. The
590 poor prediction performance of the best base classifier (SVM, AUC = 0.6509) indicates that
591 there are a number of poorly performing features hidden in the set. As some genes respond to
592 IFNs in a cell-specific manner [2], it is hard to produce predictions unless we detect key
593 discriminating features, which are robust to the change of biological environment.

594 Here, we considered two iterative strategies for selecting predictive features. The first
595 one is the forward feature selection (FFS) [62] wherein features are added sequentially based
596 on their individual performance. This strategy did not work well (**Table 3**) as the prediction
597 performances were all poor when the feature was used individually (**Supplementary Data S5**).
598 The second strategy is developed based on the backward feature elimination scheme but uses
599 fewer iterations to achieve the end result, namely AUC-driven subtractive iteration algorithm
600 (ASI) (**Figure 15**).

601 Pre-processing using the ASI algorithm showed that there were at least 28% of features
602 disrupting the prediction model. The loss of some of the individual nucleotide 4-mer feature
603 seemed not to influence the performance of the classifier at this stage, but the similarities
604 between IRGs and ISGs (**Figure 12**) particularly in the 4-mer features was a cause for concern
605 when the model was used to predict new data, especially unknown IRGs.

606 When using the ASI algorithm, the number of disrupting features did not stabilise until
607 the algorithm reached the 11-th iteration. The remaining 74 features constituted our optimum
608 feature set for predicting the ISGs (**Table 4**). Among them, 14 and 9 features displayed positive
609 and negative correlations with the level of up-regulation in IFN- α experiments ($p < 0.05$).
610 During the procedure, the AUC kept increasing steadily and reached 0.7479 at the end (**Table**
611 **3**). The Matthews correlation coefficient (MCC) also showed an overall improvement although
612 it fluctuated slightly during the last few iterations. By ranking the scores calculated by the

613 prediction model, we found 68.1% of the 496 genes (equal to the number of ISGs in the training
614 dataset) were successfully predicted as the ISGs. **Figure 13B** illustrates the distribution of
615 prediction scores generated by the ASI-optimised model for human genes with different
616 expressions in IFN- α experiments. Human genes with higher up-regulation in IFN- α
617 experiments tend to obtain higher prediction scores from our optimised machine learning
618 model (PCC = 0.243, $p = 4.2E-10$).

619 However, there were also some ISGs incorrectly predicted by our model even though
620 they were highly up-regulated, for example, basic leucine zipper ATF-like transcription factor
621 2 (BATF2, prediction score = 0.34). The model produced 33 ISGs with a prediction score
622 higher than 0.8 but this number for the non-ISGs reduced to six, including one IRG (tripartite
623 motif containing 59 (TRIM59)). The highest prediction score within the non-ISGs was found
624 on ubiquitin conjugating enzyme E2 R2 (UBE2R2, prediction score = 0.88). It contains many
625 features similar to the ISGs but was not differentially expressed in the presence of IFN- α in the
626 human fibroblast cells [3]. The lowest prediction score within ISGs was found on cap
627 methyltransferase 1 (CMTR1, prediction score = 0.12) due to the weak signal from its features.
628 For instance, CMTR1 protein does not contain any ISG-favoured SLAAPs listed in **Table 1**.
629 The influence of the IRGs on the prediction was reflected in the training dataset but was not
630 significant. Compared with human genes not differentially expressed in the IFN- α experiments
631 (non-ISGs but not IRGs), there were slightly more IRGs unsuccessfully classified when using
632 a threshold of 0.549 (Pearson's chi-squared tests: $M_1 = 27\%$, $M_2 = 24\%$, $p > 0.05$).

633

634 **Table 3. Performance of different machine learning classifiers on the training dataset S2'**
635 **via five-fold cross-validation.**

Classifier	Method	Features	Threshold-dependent					Threshold-independent	
			Score range	Threshold ^a	Sensitivity	Specificity	MCC	SN ₄₉₆ ^b	AUC
Basic	KNN ^c	518	0.100~0.900	0.500~0.550	0.593	0.621	0.214	0.607±0.014	0.6305
	RF ^d	Random	0.080~0.900	0.380~0.579	0.590±0.168	0.617±0.183	0.219±0.019	0.600±0.007	0.6413±0.0082

	SVM	518	0.328–0.743	0.542	0.567	0.681	0.250	0.615	0.6509
Optimised	SVM+FFS	78 ^c	0.170–0.836	0.561	0.518	0.760	0.287	0.621	0.6768
	SVM+ASI	74 ^c	0.098–0.918	0.549	0.623	0.750	0.376	0.681	0.7479

636 *a: this threshold is provided by maximising the value of MCC;*

637 *b: this sensitivity is measured among tested genes with the top 496 prediction probabilities;*

638 *c: k-value here is set as the square root of the size of the training samples in five-fold cross validation, i.e., k =*
639 *20 [63];*

640 *d: this random forest algorithm uses 50 random grown trees and the modelling and validation procedures are*
641 *repeated ten times;*

642 *e: these features constitute the best/optimum feature set for the current machine learning method.*

643

644 **Figure 13. The optimisation of the machine learning model with the ASI algorithm. (A)**

645 shows the change of the prediction models based on the one generated with all 518 features

646 (poorly performing feature vector = 144, best MCC = 0.250, SN₄₉₆ = 0.615, and AUC =

647 0.6509). (B) shows the distribution of prediction scores generated by the ASI-optimised model

648 for human genes with different expression levels in the IFN- α system. The ISGs and non-ISGs

649 shown in (B) are randomly selected through an undersampling strategy [64] on dataset S2. The

650 list of gene names can be found in **Supplementary Data S1**.

651

652 **Table 4. The optimum 74 features contributing to predicting the ISGs.**

Evolutionary features (2)		
Number of human paralogues, average dS within human paralogues ^N .		
Codon usage features (10)		
Codon usage: CTA (L) ^P	Codon usage: ATT (I)	Codon usage: TAT (Y)
Codon usage: GCG (A) ^N	Codon usage: CAC (H) ^N	Codon usage: TGC (C)
Codon usage: CGT (R)	Codon usage: CGA (R)	Codon usage: CGG (R) ^N
Codon usage: AGA (R) ^P		
Genetic composition features (40)		
DNA AC content	Dinucleotide CpT composition	DNA 4-mer CGCG composition ^N
DNA 4-mer AATC composition ^P	DNA 4-mer TCGT composition	DNA 4-mer GATG composition ^P
DNA 4-mer AACA composition	DNA 4-mer TGAG composition ^P	DNA 4-mer GACC composition
DNA 4-mer ATAT composition	DNA 4-mer TGTA composition	DNA 4-mer GACG composition
DNA 4-mer ATGT composition ^P	DNA 4-mer CACG composition	DNA 4-mer GAGT composition ^P
DNA 4-mer ACAC composition	DNA 4-mer CTCC composition	DNA 4-mer GTAC composition
DNA 4-mer ACTA composition	DNA 4-mer CCAC composition	DNA 4-mer GTGT composition
DNA 4-mer ACTC composition	DNA 4-mer CCTA composition	DNA 4-mer GTGC composition
DNA 4-mer ACCG composition	DNA 4-mer CCTC composition ^P	DNA 4-mer GTGG composition
DNA 4-mer TATG composition	DNA 4-mer CCGT composition	DNA 4-mer GCAA composition ^P

DNA 4-mer TTCT composition	DNA 4-mer CGAG composition	DNA 4-mer GCTC composition
DNA 4-mer TTCG composition	DNA 4-mer CGTG composition	DNA 4-mer GCCT composition
DNA 4-mer TTGA composition	DNA 4-mer CGCA composition	DNA 4-mer GGGG composition
DNA 4-mer TCAT composition		

Proteomic composition features (9)

Arginine composition, cysteine composition^P, methionine composition;
 Basic amino acid composition (R/H/K)^P Sulfur amino acid composition (C&M)^P
 Hydroxyl amino acid composition (S&T)^N Small amino acid composition (N/D/C/P/T)^N
 Large amino acid composition (R/I/L/K/M)^P
 Uncharged amino acid composition (A/N/C/Q/G/I/L/M/F/P/S/T/W/Y/V)^N

Features about human interactome network (3)

Average shortest paths^P, betweenness, neighborhood connectivity^N.

Sequence pattern features (8)

SLNP: ATA[AG][TG]	SLNP: TAT[AT]T	SLNP: T[AT]AAA
SLNP: [ATG]TGTA	SLAAP: SxNxExE	SLAAP: ENE
SLAAP: SVI	Co-occurrence of SLAAPs (count)	

653 *P*: features are positively associated with the level of up-regulation in IFN- α experiments ($p < 0.05$);

654 *N*: features are negatively associated with the level of up-regulation in IFN- α experiments ($p < 0.05$).

655

656 **Review of different testing datasets**

657 In this study, we trained and optimised a SVM model from our training dataset S2', and
 658 prepared seven testing datasets (dataset S2''/S3/S4/S5/S6/S7/S8) to assess the generalisation
 659 capability of our model under different conditions (**Table 5**). The S2'' testing dataset was a
 660 subset of dataset S2. The prediction performance on this testing dataset was close to that in the
 661 training stage with an AUC of 0.7455 (**Figure 14A**). The best MCC value (0.345) was achieved
 662 when setting the judgement threshold to 0.438, which meant that the prediction model was
 663 sensitive to signals related to ISGs. In this case, it performed predictions with high sensitivity
 664 but inevitably produced many false positives, especially within IRGs.

665 In the S3 testing dataset, we used 695 ISGs with low confidence. The overall accuracy
 666 (equals to sensitivity as there were no negatives) only reached 44.0% when using a judgement
 667 threshold of 0.549, about 0.18 lower than SN under the same threshold in the training dataset
 668 S2' (**Table 3**). It is expected as some of their inherent attributes make them slightly up-
 669 regulated, silent or even repressed (e.g., become non-ISGs in other IFN systems) in response
 670 to some IFN-triggered signalling. On this testing dataset, our machine learning model produced
 671 38 (5.5%) ISGs with a prediction score higher than 0.8. This number was also lower than that

672 on the training dataset S2'. It further indicates the relatively low confidence for the ISGs
673 included in dataset S3.

674 The S4 testing dataset was constructed to illustrate our hypothesis that there are some
675 patterns shared among the ISGs and IRGs at least in the IFN- α system in the human fibroblast
676 cells. On this testing dataset, the prediction accuracy (equals to SP as there were no positives)
677 was 60.2% under the judgement threshold of 0.549, about 0.15 lower than the SP under the
678 same threshold in the training dataset S2' (**Table 3**). Leucine rich repeat containing 2 (LRRC2),
679 carbohydrate sulfotransferase 10 (CHST10) and eukaryotic translation elongation factor 1
680 epsilon 1 (EEF1E1) showed strong signals of being ISGs (probability score > 0.9). In total,
681 there were 56 (5.6%) IRGs being incorrectly predicted as ISGs with prediction scores higher
682 than 0.8. This high score was found in an estimated 8.1% of the ISGs but was only observed
683 in 1.2% of human genes not differentially expressed in the IFN- α experiments (**Figure 13B**).
684 These results indicate that there is a considerable number of IRGs incorrectly predicted as ISGs
685 in the S4 testing dataset due to their close distance to the ISGs in the high-dimensional feature
686 space. This may be the case for many other datasets including dataset S2'', S5, S6, S7, and S8.
687 It also supports our hypothesis about the shared patterns from the machine learning aspect and
688 is consistent with the results shown in **Figure 12**.

689 The next three testing datasets (S5, S6, and S7) were collected from the Interferome
690 database [24] to test the applicability of the machine learning model across different IFN types.
691 The ISGs in these testing datasets were all highly up-regulated ($\text{Log}_2(\text{Fold Change}) > 1.0$) in
692 the corresponding IFN systems while all the non-ISGs were not up-regulated after
693 corresponding IFN treatments ($\text{Log}_2(\text{Fold Change}) < 0$). The results shown in **Figure 14** reveal
694 that the ISGs triggered by type I or III IFN signalling can still be predicted by our machine
695 learning model, but the performance is limited to some extent (AUC = 0.6677 and 0.6754

696 respectively). However, it is almost impossible to make normal predictions with the current
697 feature space for human genes up-regulated by type II IFNs (AUC = 0.5532).

698

699 **Figure 14. The performance of our optimised model on different datasets.** (A) and (B)
700 illustrate the AUC and best MCC. S2' is the training dataset used in this study. It randomly
701 includes 496 ISGs and an equal number of non-ISGs from dataset S2 that contains ISGs/non-
702 ISGs with high confidence (**Table 5**). Evaluation on this dataset in (A) is processed via five-
703 fold cross validation. S2'' is the testing dataset constructed with the remaining human genes in
704 dataset S2. S5, S6, and S7 are collected from the Interferome database [24], including human
705 genes with different responses to the type I, II and III IFNs, respectively. The label and usage
706 of these human genes are provided in **Supplementary Data S1**.

707

708 The S8 testing dataset consisted of 2217 human genes that were insufficiently expressed
709 in IFN- α experiments in the human fibroblast cells [3]. The results showed that there were
710 around 41.2% ELGs being predicted as the ISGs when using a judgement threshold of 0.549.
711 This was approximately 0.21 lower than the SN under the same threshold in the training dataset
712 S2' (**Table 3**). It suggests that there are more non-ISGs than ISGs in this dataset, which is
713 consistent with the results shown in **Figure 12**. Particularly, we found ten ELGs with prediction
714 scores higher than 0.9: CD48 molecule, CD53 molecule, lipocalin 2 (LCN2), uncoupling
715 protein 1 (UCP1), coiled-coil domain containing 68 (CCDC68), potassium calcium-activated
716 channel subfamily M regulatory beta subunit 2 (KCNMB2), potassium voltage-gated channel
717 interacting protein 4 (KCNIP4), zinc finger HIT-type containing 3 (ZNHIT3), serpin family B
718 member 4 (SERPINB4), and fibrinogen silencer binding protein (FSBP). By retrieving data
719 from the Genotype-Tissue Expression project [65], we found that the expression of these ELGs
720 was generally limited with the exception of CD53 and ZNHIT3 (**Figure 15**). The expression

721 data of CD53 were not included in the OCISG database [3] and were also limited in the
722 Interferome database [24]. It only showed slight up-regulation after type I IFN treatments in
723 blood, liver, and brain but there is currently no record of its expression level in the presence of
724 IFN- α in the human fibroblast cells. ZNHIT3 is another well-expressed gene lacking
725 information in the OCISG. In the Interferome database [24], we found that ZNHIT3 could be
726 up-regulated after IFN treatments in some fibroblast cells on the skin. As for the remaining
727 eight ELGs, despite their limited expression in the human fibroblast cells, their features suggest
728 that they are very likely to be IFN- α stimulated in a currently untested cell type.

729

730 **Figure 15. Expression of the ELGs in different tissues.** Expression data for ten ELGs are
731 collected from the Genotype-Tissue Expression project (<https://gtexportal.org/>) [65]. The
732 tissues in red are not included in the Interferome database [24]. White boxes in the heatmap
733 indicate that there is no data available for genes in the corresponding tissues. The overall
734 expression level of these ten ELGs are reflected via human perspective photo retrieved from
735 Expression Atlas (<https://www.ebi.ac.uk/gxa>) [66].

736

737

738 **Discussion**

739 In this study, we investigated the characteristics that influence the expression of human genes
740 in IFN- α experiments. We compared the ISGs and non-ISGs through multiple procedures to
741 guarantee strong signals for the ISGs and to avoid cell-specific influences that resulted in the
742 lack of ISG expression in certain cell types [2]. Even some highly up-regulated ISGs can
743 become down-regulated when the biological conditions change, exemplified by the
744 performance of C-X-C motif chemokine ligand 10 (CXCL10) on liver biopsies after IFN- α
745 treatment. This refinement is necessary as the representation of features between the ISGs and

746 background human genes shows that many non-ISGs especially IRGs have similar feature
747 patterns to the ISGs (**Figure 12**).

748 Generally, the ISGs are less evolutionarily conserved and include more human
749 paralogues than the non-ISGs. They have specific nucleotide patterns exemplified by the
750 depletion of GC-content and have a unique codon usage preference in coding proteins. There
751 are a number of SLNPs widely observed in the cDNA of the ISGs which are relatively rare in
752 the non-ISGs (**Supplementary Data S4**). Likewise, there are also many SLAAPs highlighted
753 in the sequences of ISG products that are absent or rare in the non-ISG products (**Table 1**). In
754 the human PPI network, the ISG products tend to have higher betweenness than the background
755 human protein. Abnormal expression or knockout of these proteins will increase the diameter
756 of the network and may lead to some lethal consequences that are not tolerated in signalling
757 pathways [67-69]. These ISG-specific patterns may be the result of the evolution of the innate
758 immune system in vertebrates and could be adaptations to the cellular environment induced by
759 interferon following a pathogenic infection [70]. It is also possible that some of the particular
760 SLNPs and SLAAPs may be functionally important as the cell changes from non-infected to
761 infected. Experimental evidence will be necessary to investigate this.

762 Some inherent properties of the ISGs facilitate or elevate their expression after IFN- α
763 treatments but may also be used by viruses to escape from IFN- α -mediated antiviral response
764 [22]. For instance, we found that a higher dN/dS ratio was positively correlated with gene up-
765 regulation following IFN- α treatments (**Figure 10**). This suggests ISGs are on average under
766 stronger adaptive evolutionary selection pressure than the non-ISGs possibly linked to their
767 evolution as anti-viral molecules. , It will also facilitate the virus to interfere with IFN- α
768 signalling through the JAK-STAT pathway and inactivate downstream cellular factors
769 involved in IFN- α signal transductions [22]. We found arginine was under-represented in the
770 ISG products compared to the non-ISG products. As arginine is essential for the normal

771 proliferation and maturation of human T cells [72], such depletion in the ISG products may
772 leave a risk of inhibiting T- cell function and potentially increase susceptibility to infections
773 [73]. Furthermore, the special pattern of the ISGs also promotes the representation of some
774 features even if they are not well represented in nature, for example, the higher cysteine
775 composition in the ISGs. We hypothesize that it may be helpful to activate T-cell to regulate
776 protein synthesis, proliferation and secretion of immunoregulatory cytokines [74,75]. There
777 are also some features (e.g., methionine composition) not differentially represented between
778 the ISGs and non-ISGs but play important roles in IFN- α -mediated immune responses. For
779 example, there is evidence for the methionine content playing a role in the biosynthesis of S-
780 Adenosylmethionine (SAM), which can improve interferon signalling in cell culture [76,77].

781 As previously mentioned, there were similar patterns between the feature representation
782 of the ISGs and IRGs, which led to the unclear boundary for the ISGs and non-ISGs in the
783 feature space. We found significant differences in the representation of features on
784 evolutionary conservation (**Figure 4**) between the ISGs and non-ISGs, but they became non-
785 significant when comparing the ISGs with IRGs. Similar phenomena were observed on many
786 features deciphered from the canonical transcript, e.g., dinucleotide composition and codon
787 usage features. We hypothesis that IRGs are former ISGs that have evolved to be down
788 regulated to avoid any unintended harmful consequences. Furthermore, despite so many
789 similarities between the ISGs and IRGs, the separate classification of these genes is still
790 possible. 4-mer compositions can be considered as the key features as most of them are
791 differentially represented between ISGs and IRGs (**Figure 12**). Using proteomic features can
792 also help to differentiate the ISGs from IRGs but is not as predictive as using 4-mer features.

793 In the machine learning framework, we developed the ASI algorithm to remove poorly
794 performing features but kept features that do not influence prediction performance when
795 removed individually from iterations. Features may have synergistic effects on the prediction

796 performance. The elimination of some specific features may ruin such improvement even when
797 they were individually uninformative for the improvement of the classifier. In this case,
798 keeping as many useful features as possible seems to be a reasonable option but will greatly
799 increase the dimension of the feature space and increase the risk of overfitting [78]. By contrast,
800 our ASI algorithm avoided such a risk and kept the synergistic effect of different features
801 through iterations.

802 In the prediction task, we found some previously labelled non-ISGs with very high
803 prediction scores, suggesting that they had some inherent properties consistent with them being
804 stimulated after IFN- α treatments. Some, for example, UBE2R2 has been shown to be
805 significantly up-regulated after IFN- α treatment [79]. The non-ISG label had been assigned
806 because the relevant expression data in the presence of IFN- α were not included in the OCISG
807 [3] and Interferome databases [24]. We also found ten ELGs with very high prediction scores
808 (> 0.9). Literature searches on these genes indicate that they are likely to be involved in the
809 innate immune response [80,81]. Their responses may be limited to certain tissues or cell types
810 for which there is limited expression data in the Interferome database [24]. For example, LCN2
811 has been shown to mediate an innate immune response to bacterial infections by sequestering
812 iron [80] and is induced in the central nervous system of mice infected with West Nile virus
813 encephalitis [82]. CD48 was shown to increase in levels in the context of human IFN- $\alpha/\beta/\gamma$
814 stimulation [81]. Interestingly, CD48 is also the target of immune evasion by viruses [83] and
815 has been captured in the genome of cytomegalovirus and undergone duplication [84]. Evidence
816 for other ELGs is harder to assess, particularly those for which expression is absent in a range
817 of tissues (e.g., UCP1 in **Figure 15**). UCP1 is a mitochondrial carrier protein expressed in
818 brown adipose tissue (BAT) responsible for non-shivering thermogenesis [85]. It is possible
819 that UCP1 is stimulated directly or indirectly by IFN- α in BAT, resulting in the defended
820 elevation of body temperature in response to infection.

821 We developed the machine learning model based on experimental data from the human
822 fibroblast cells stimulated by IFN- α . It can be generalised to type I or III IFN systems,
823 presumably because activations of type I and III ISGs are both controlled by ISRE [9] and aim
824 to regulate host immune response [10-12]. However, our model cannot be used for predictions
825 in the type II IFN system (AUC = 0.5532, best MCC = 0.083, **Figure 14**). It may be caused by
826 the different control elements and the different roles in human immune activities [14]. One
827 feasible strategy is to reclassify the ISGs/non-ISGs based on the IFN experiments in the type
828 II IFN system. Using only the overlapping ISGs and non-ISGs in both type I and type II IFN
829 system for modelling could be another solution. In summary, our analyses highlight some key
830 sequence-based features that are helpful to distinguish the ISGs from non-ISGs, or IRGs. While
831 reliable ISG prediction remains a difficult challenge, our machine learning model is able to
832 produce a list of putative ISGs to support IFN-related research. As knowledge of the ISG
833 functions continues to be elucidated by experimentalists, the *in-silico* approach applied here
834 can in future be extended to classify the different functions of ISGs. The ‘important’ features
835 mentioned in this study may become a focus for investigating the interferon antagonists
836 expressed by different viruses [86].

837

838

839 **Methods**

840 **Dataset curation**

841 In this study, we retrieved 2054 ISGs (up-regulated), 12379 non-ISGs (down-regulated or not
842 differentially expressed), and 3944 unlabelled human genes (ELGs with less than one count
843 per million reads mapping across the three biological replicates [87,88]) from the OCISG
844 database (<http://isg.data.cvr.ac.uk/>) [3]. Gene clusters in the OCISG database were built
845 through Ensembl Compara [89], which provided a thorough account of gene orthology based

846 on whole genomes available in Ensembl [58]. Labels of these human genes were defined based
847 on the fold change and a false discovery rate (FDR) following the IFN- α treatments in the
848 human fibroblast cells. We searched the collected 18377 entries against the RefSeq database
849 (<https://www.ncbi.nlm.nih.gov/refseq/>) [32] to decipher features based on appropriate
850 transcripts (canonical) [90] coding for the main functional isoforms of these human genes. It
851 produced 1315, 7304, and 2217 results for the ISGs, non-ISGs and ELGs, respectively. These
852 10836 human genes were well-annotated by multiple online databases and were used as the
853 background dataset S1 in the analyses.

854 For the purpose of generating a set of human genes with high confidence of being up-
855 regulated and non-up-regulated in response to the IFN- α , we searched the recompiled 8619
856 human genes (ISGs or non-ISGs) against Interferome (<http://www.interferome.org/>) [24]. We
857 filtered out the ISGs without high up-regulation ($\text{Log}_2(\text{Fold Change}) > 1.0$) or with obvious
858 down-regulation ($\text{Log}_2(\text{Fold Change}) < -1.0$) in the presence of type I IFNs. This procedure
859 guaranteed a refined ISG dataset with strong levels of stimulation induced by any type I IFNs
860 and reduced biases driven by the IRGs for the analyses and predictions. We filtered out the
861 non-ISGs showing enhanced expression after type I IFN treatments ($\text{Log}_2(\text{Fold Change}) > 0$).
862 The exclusion of these non-ISGs could effectively reduce the risk of involving false negatives
863 in analyses and producing false positives in predictions. As a result, the refined dataset S2
864 contains 620 ISGs and 874 non-ISGs with relatively high confidence.

865 The training procedure in the machine learning framework was conducted on the
866 balanced dataset S2'. It consisted of 992 randomly selected ISGs and non-ISGs from dataset
867 S2. The remaining human genes in S2 were used for independent testing. Additionally, we also
868 constructed another six testing datasets for the purpose of review and assessment. Dataset S3
869 contained 695 ISGs with low confidence compared to those ISGs in dataset S2. Some of them
870 could be non-ISGs or even IRGs in the type I IFN system. Dataset S4 contained 1006 IRGs

871 from the human fibroblast cell experiments. Dataset S5, S6, and S7 were constructed based on
872 records for experiments in type I, II, and III IFN systems from Interferome
873 (RRID:SCR_007743) [24]. The criterion for an ISG in the latter three datasets was a high level
874 of up-regulation ($\text{Log}_2(\text{Fold Change}) > 1.0$) while that for non-ISGs was no up-regulation after
875 IFN treatments ($\text{Log}_2(\text{Fold Change}) < 0$). The last testing dataset S8 was derived from our
876 background dataset S1, containing 2217 ELGs. A breakdown of the aforementioned eight
877 datasets is shown in **Table 5**. Detailed information of the human genes used in this study is
878 provided in **Supplementary Data S1**. The cDNA and protein sequences are accessible at
879 <http://isgpre.cvr.gla.ac.uk/>.

880

881 **Table 5. A breakdown of datasets used in this study.**

Dataset	Brief description	IFN system	ISGs	Non-ISGs	ELGs	Usage
S1	Background human genes	IFN- α in fibroblast cells	1315	7304	2217	Analyses
S2	Dataset with high confidence	IFN- α in fibroblast cells	620	874	0	Analyses
S2'	Training subset of S2	IFN- α in fibroblast cells	496	496	0	Training
S2''	Testing subset of S2	IFN- α in fibroblast cells	124	378	0	Testing
S3	ISGs with low confidence in S1	IFN- α in fibroblast cells	695	0	0	Testing
S4	IRGs divided from S1	IFN- α in fibroblast cells	0	1006	0	Analyses/ testing
S5	ISGs from Interferome [24]	Type I IFNs in all cells	1259	872	0	Testing
S6	ISGs from Interferome [24]	Type II IFN in all cells	2229	755	0	Testing
S7	ISGs from Interferome [24]	Type III IFN in all cells	33	1683	0	Testing
S8	ELGs divided from S1	IFN- α in fibroblast cells	0	0	2217	Testing

882

883 **Generation of discrete features**

884 We encoded 397 discrete features from aspects of evolution, nucleotide composition,
885 transcription, amino acid composition, and network preference. Original values of these
886 features for our compiled 10836 human genes are accessible at <http://isgpre.cvr.gla.ac.uk/>.

887 From the perspective of evolution, we used the number of transcripts, open reading
888 frames (ORFs) and count of exons used for coding to quantify the alternative splicing process.
889 Genes with more transcripts and ORFs have higher alternative splicing diversity to produce
890 proteins with similar or different biological functions [33,91,92]. Frequent use of protein-

891 coding exons indicates more complex alternative splicing products [93]. Here, duplication and
892 mutation features were measured by the number of within-species paralogues and substitutions
893 [34,35]. These data were collected from BioMart (RRID:SCR_002987)[58] to assess the
894 selection on protein sequences and mutational processes affecting the human genome [94].

895 From the perspective of nucleotide composition, we calculated the percent of adenine,
896 thymine, cytosine, guanine, and their four-category combinations in the coding region of the
897 canonical transcript. The first category measured the proportion of two different nitrogenous
898 bases out of the implied four bases, e.g., GC-content. The second category also focused on the
899 combination of two nucleotides but added the impact of phosphodiester bonds along the 5' to
900 3' direction, e.g., CpG-content [95]. The third category calculated the occurrence frequency of
901 4-mers, e.g., 'CGCG' composition to involve some positional resolution [41]. The last category
902 considered the co-occurrence of SLNPs. From the perspective of transcription, we calculated
903 the usage of 61 coding codons and three stop codons in the coding region of the canonical
904 transcripts. Codon usage biases are observed when there are multiple codons available for
905 coding one specific amino acid. They can affect the dynamics of translation thus regulate the
906 efficiency of translation and even the folding of the proteins [40,96].

907 From the perspective of amino acid composition, we calculated the percentage of 20
908 standard amino acids and their combinations based on their physicochemical properties [46].
909 Patterns in the amino acid level are considered to have a direct impact on the establishment of
910 biological functions or to reflect the result of strong purifying selection [47]. Based on the
911 chemical properties of the side chain, we grouped amino acids into seven classes including
912 aliphatic, aromatic, sulfur, hydroxyl, acidic, amide, and basic amino acids. We also grouped
913 amino acids based on geometric volume, hydrophathy, charge status, and polarity, but found
914 some overlaps among these features. For instance, amino acids with basic side chains are all
915 positively charged. Aromatic amino acids all have large geometric volumes (volume > 180

916 cubic angstroms). Likewise, we also considered the co-occurrence of short linear sequence
917 patterns at the protein level. These co-occurring SLAAPs may relate to potential mechanisms
918 regulating the expression of the ISGs [97].

919 When trying to measure the network preference for the gene products, we constructed
920 a human PPI network based on 332,698 experimentally verified interactions (confidence score >
921 0.63) from HIPPIE (RRID:SCR_014651)[55]. Nodes and edges of this network are provided
922 at <http://isgpre.cvr.gla.ac.uk/>. Eight network-based features including the average shortest path,
923 closeness, betweenness, stress, degree, neighbourhood connectivity, clustering coefficient, and
924 topological coefficient were calculated from this network. Isolated nodes or proteins were not
925 included in our network and were assigned zero values for all these eight features. The shortest
926 path measures the average length of the shortest path between a focused node and others in the
927 network. Closeness of a node is defined as the reciprocal of the length of the average shortest
928 path. Proteins with a low value of the shortest paths or closeness are close to the centre of the
929 network. Betweenness reflects the degree of control that one node exerted over the interactions
930 of other nodes in the network [98]. Stress of a node measures the number of shortest paths
931 passing through it. Proteins with a high value of betweenness or stress are close to the
932 bottleneck of the network. Degree of a node counts the number of edges linked to it while
933 neighbourhood connectivity reflected the average degree of its neighbours. Proteins with high
934 values of degree or neighbourhood connectivity are close to the hub of the network. They are
935 considered to play an important role in the establishment of the stable structure of the human
936 interactome [99]. Clustering and topological coefficient measure the possibility of a node to
937 form clusters or topological structures with shared neighbours. The former coefficient can be
938 used to identify the modular organisation of metabolic networks [100] while the latter one may
939 be helpful to find out virus mimicry targets [53].

940

941 **Generation of categorical features**

942 In this study, categorical features were used to check the occurrence of short linear sequence
943 patterns in the genome and proteome. SLNPs constructed in this study contained three to five
944 random nucleotides, producing 708,540 alternative choices. SLNPs with no restrictions on their
945 first or last position were not taken into consideration as their patterns could be expressed in a
946 more concise way. A SLNP was picked out to encode a binary feature when its occurrence
947 level in the coding region of the canonical ISG transcripts was significantly higher than that
948 for the non-ISGs (Pearson's chi-squared test: $p < 0.05$). SLAAPs were constructed with three
949 to four fixed amino acids separated by putative gaps. The gap could be occupied by at most
950 one random amino acid, producing 1,312,000 alternative choices. Likewise, binary features
951 were prepared for SLAAPs showing significant enrichment in the ISG products than in the
952 non-ISG products (Pearson's chi-squared test: $P < 0.05$). Since there were lots of results
953 rejecting the null-hypothesis, we adopted the Benjamini-Hochberg correction procedure to
954 avoid type I error [43]. Additionally, we also encoded two features to check the co-occurrence
955 or absence of multiple SLNPs and SLAAPs. This co-occurrence status might be a better
956 representation of functional sites composed of short stretches of adjacent nucleobases or amino
957 acids surrounding SLNPs or SLAAPs [47].

958

959 **Assessment of associations between feature representation and IFN-triggered** 960 **stimulations**

961 We obtained 8619 human genes with expression data from the OCISG database [3]. 4111 of
962 them were annotated with a positive $\text{Log}_2(\text{Fold Change})$ ranging from 0 to 12.6, which meant
963 they were up-regulated after IFN- α treatments in the human fibroblast cells. In order to measure
964 the average level of feature representation (AREP) for genes with similar expression during
965 IFN stimulations, we introduced a 0.1-length sliding-window to divide the data into 126 bins

966 with different $\text{Log}_2(\text{Fold Change})$. Here, PCC was introduced to test the association between
 967 the representation of discrete features and IFN- α -triggered stimulation ($\text{Log}_2(\text{Fold Change}) >$
 968 0). It can be formulated as:

$$PCC(f) = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{LFC_i - M_0}{SD_0} \right) \times \left(\frac{AREP_i - M_f}{SD_f} \right) \quad (1)$$

969 where n is the number of divided parts that equals to 126 in this study; LFC_i and $AREP_i$ are
 970 the value of $\text{Log}_2(\text{Fold Change})$ and AREP in the i -th part; M_0 and SD_0 are the mean and
 971 standard deviation of $\text{Log}_2(\text{Fold Change})$, which is set as 6.4 and 3.7 respectively in this study;
 972 M_f and SD_f are the mean and standard deviation of 126 AREP that reflect the representation
 973 of the considered feature. To make fair comparisons among features with different scales, we
 974 normalised them based on the major value of their representations:

$$Norm(f) = \begin{cases} 1, f > UB(f) \\ \frac{f - LB(f)}{UB(f) - LB(f)}, LB(f) < f < UB(f) \\ 0, f < LB(f) \end{cases} \quad (2)$$

975 where $LB(f)$ and $UB(f)$ are the lower and upper bound representing the 5th and 95th
 976 percentile within representation values for the target feature. The representation of feature was
 977 considered to have a stronger positive/negative association with IFN- α -triggered stimulations
 978 if the PCC calculated from the normalised features was closer to 1.0/-1.0 and the p-value
 979 calculated by the Student t-test was lower than 0.05.

980

981 **Machine learning and optimisation**

982 We designed a machine learning framework for the prediction of ISGs. Firstly, all features
 983 were encoded and normalised based on their major representations (**Equation 2**). Then we
 984 used an under-sampling procedure [64] to generate a balanced dataset from dataset S2 for
 985 training and modelling. The SVM with radial basis function [61] was used as the basic classifier.
 986 It maps the normalised feature space to a higher dimension to generate a space plane to better

987 classify the majority of positive and negative samples. In order to avoid overfitting [78] and
988 made it easier for the SVM model to generate an appropriate classification plane that involved
989 fewer false positives and false negatives, here, we propose a subtractive iteration algorithm
990 driven by the change of AUC. This algorithm is developed based on the traditional backward
991 feature elimination method [62] but uses fewer iterations to filter out poorly performing
992 features (**Figure 16**). In each iteration, we traversed the features and removed those that did
993 not improve the AUC of the prediction results. In the testing procedure, we encoded the
994 optimum features for testing samples and placed them in the optimised feature space. Samples
995 with longer distance to the optimised classification plane indicated a stronger signal of being
996 the ISGs or non-ISGs. They were more likely to get higher prediction scores (close to 0 or 1)
997 from the SVM model.

998

999 **Figure 16. The pseudo-code of the AUC-driven subtractive iteration algorithm.**

1000

1001 **Performance evaluation**

1002 In this study, the prediction results were evaluated with three threshold-dependent criteria
1003 including sensitivity, specificity, and MCC [101] and two threshold-independent criteria: SN_n
1004 and AUC. Sensitivity and specificity were used to assess the quality of the machine learning
1005 model in recognising ISGs and non-ISGs respectively while MCC provided a comprehensive
1006 evaluation for both positives and negatives. The number of 'n' in the SN_n criterion was
1007 determined based on the number of ISGs used for testing. It was used to measure the upper
1008 limit of the prediction model as well as to check the existence of important false positives close
1009 to the class of ISGs from the perspective of data expression. Finally, AUC was a widely used
1010 criterion to evaluate the prediction ability of a binary classifier system. The group of interest

1011 was almost unpredictable in a specific binary classifier system if the AUC of the classifier was
1012 close to 0.5.

1013

1014

1015 **Availability of source code and requirements**

1016 • Project name: ISGPRES

1017 • Project home page: <http://isgpre.cvr.gla.ac.uk/>

1018 • Operating system: Platform independent

1019 • Programming language: Java

1020 • Other requirements: Docker or JDK 8+

1021 • Docker image: <https://hub.docker.com/repository/docker/hchai01/isgpre>

1022 • Biotoools repository: <https://bio.tools/isgpre>

1023 • Research Resource Identification Initiative ID: SCR_022730

1024 • Documentation and tutorials: <https://github.com/HChai01/ISGPRES>

1025 License: GNU GPL v3.0

1026

1027 **Data Availability**

1028 The implemented web server and all reproduceable data are freely accessible at
1029 <http://isgpre.cvr.gla.ac.uk/> and <https://github.com/HChai01/ISGPRES>. Code snapshots and
1030 other supplementary data are also available in the *GigaScience* GigaDB repository [102].

1031

1032

1033 **Additional Files**

1034 **Supplementary Data S1. Basic information and usage of our compiled 10836 human**
1035 **genes.**

1036 **Supplementary Data S2. The result of Mann-Whitney U tests for discrete features.**
1037 **Supplementary Data S3. Association between feature representations and IFN- α**
1038 **stimulations.**
1039 **Supplementary Data S4. The result of Pearson's chi-squared tests for sequence motifs.**
1040 **Supplementary Data S5. Features and their individual performance in machine learning.**

1041

1042 **Abbreviations**

1043 APC: anaphase promoting complex; AREP: average level of feature representation; ASI:
1044 AUC-driven subtractive iteration algorithm; AUC: area under the receiver operating
1045 characteristic curve; cDNA: complementary DNA; dN: non-synonymous substitutions; dS:
1046 synonymous substitutions; ELGs: human genes with limited expression in the IFN- α
1047 experiments; FDR: false discovery rate; FFS: forward feature selection; GAF: IFN- γ activation
1048 factor; GAS: gamma-activated sequence promoter elements; gBGC: GC-biased gene
1049 conversion; HIPPIE: Human Integrated Protein-Protein Interaction rEference; IDRs:
1050 intrinsically disordered regions; IFNAR: interferon- α receptor; IFNGR: IFN- γ receptor;
1051 IFNLR1: IFN- λ receptor 1; IFNs: interferons; IL-10R2: interleukin-10 receptor 2; IRF9:
1052 interferon regulatory factor 9; IRG: interferon repressed (down-regulated) human genes;
1053 ISGF3: interferon stimulated gene factor 3 complex; ISGs: interferon stimulated (up-regulated)
1054 human genes; ISRE: interferon stimulated response elements; JAK1: Janus kinase 1; KNN: k-
1055 nearest neighbours; MCC: Matthews correlation coefficient; non-ISGs, human genes not
1056 significantly up-regulated by interferons; OCISG: Orthologous Clusters of Interferon-
1057 stimulated Genes; ORF: open reading frame; PCC: Pearson's correlation coefficient; PPI:
1058 protein-protein interaction; RefSeq: Reference Sequence; RF: random forest; SLAAP: short
1059 linear amino acid pattern; SLNP: short linear nucleotide pattern; SN₄₉₆: sensitivity of

1060 samples with the top 496 prediction scores; STAT: signal transducer and activator of
1061 transcription; SVM: support vector machine.

1062

1063

1064 **Competing Interests**

1065 The authors have declared that no competing interests exist.

1066

1067

1068 **Funding**

1069 HC is supported by the China Scholarship Council (201706620069). JH, QG and DLR are
1070 supported by the Medical Research Council (MC_UU_1201412). The funders had no role in
1071 study design, data collection and analysis, decision to publish, or preparation of the manuscript.

1072

1073

1074 **Authors' Contributions**

1075 Conceptualization: all authors; data curation: H. C.; formal analysis: H. C.; funding acquisition:
1076 D. L. R.; investigation: H. C.; methodology: H. C.; project administration: D. L. R., J. H.;
1077 resources: Q. G., J. H., D. L. R.; web server: H. C.; software: H. C.; supervision: Q. G., J. H.,
1078 D. L. R.; validation: all authors; visualization: H. C.; writing original draft: H. C.; writing
1079 review & editing: all authors.

1080

1081

1082 **Acknowledgments**

1083 The authors wish to thank Drs Andrew Davison, Suzannah Rihn and Sam Wilson for helpful
1084 discussions and recommendations, and Scott Arkison for help setting up the website.

1085

1086

1087 **References**

- 1088 1. Rönnblom L. The type I interferon system in the etiopathogenesis of autoimmune
1089 diseases. *Ups J Med Sci.* 2011;116(4):227-37.
- 1090 2. Mostafavi S, Yoshida H, Moodley D, LeBoité H, Rothamel K, Raj T, et al. Parsing the
1091 interferon transcriptional network and its disease associations. *Cell.* 2016;164(3):564-
1092 78.
- 1093 3. Shaw AE, Hughes J, Gu Q, Behdenna A, Singer JB, Dennis T, et al. Fundamental
1094 properties of the mammalian innate immune system revealed by multispecies
1095 comparison of type I interferon responses. *PLoS Biol.* 2017;15(12):e2004086.
- 1096 4. Shalhoub S. Interferon beta-1b for COVID-19. *The Lancet.* 2020;395(10238):1670-1.
- 1097 5. Harris BD, Schreiter J, Chevrier M, Jordan JL and Walter MR. Human interferon- ϵ and
1098 interferon- κ exhibit low potency and low affinity for cell-surface IFNAR and the
1099 poxvirus antagonist B18R. *J Biol Chem.* 2018;293(41):16057-68.
- 1100 6. Li S-f, Zhao F-r, Shao J-j, Xie Y-l, Chang H-y and Zhang Y-g. Interferon-omega:
1101 Current status in clinical applications. *Int Immunopharmacol.* 2017;52):253-60.
- 1102 7. Kak G, Raza M and Tiwari BK. Interferon-gamma (IFN- γ): exploring its implications
1103 in infectious diseases. *Biomol Concepts.* 2018;9(1):64-79.
- 1104 8. Hemann EA, Gale Jr M and Savan R. Interferon lambda genetics and biology in
1105 regulation of viral control. *Front Immunol.* 2017;8):1707.
- 1106 9. Schneider WM, Chevillotte MD and Rice CM. Interferon-stimulated genes: a complex
1107 web of host defenses. *Annu Rev Immunol.* 2014;32):513-45.
- 1108 10. Kotenko SV and Durbin JE. Contribution of type III interferons to antiviral immunity:
1109 location, location, location. *J Biol Chem.* 2017;292(18):7295-303.
- 1110 11. Fensterl V and Sen GC. Interferons and viral infections. *Biofactors.* 2009;35(1):14-20.
- 1111 12. Lazear HM, Schoggins JW and Diamond MS. Shared and distinct functions of type I
1112 and type III interferons. *Immunity.* 2019;50(4):907-23.
- 1113 13. Takaoka A and Yanai H. Interferon signalling network in innate defence. *Cell*
1114 *Microbiol.* 2006;8(6):907-22.
- 1115 14. Stark GR and Darnell Jr JE. The JAK-STAT pathway at twenty. *Immunity.*
1116 2012;36(4):503-14.
- 1117 15. Schoggins JW. Interferon-stimulated genes: what do they all do? *Annu Rev Virol.*
1118 2019;6):567-84.
- 1119 16. Aso H, Ito J, Koyanagi Y and Sato K. Comparative description of the expression profile
1120 of interferon-stimulated genes in multiple cell lineages targeted by HIV-1 infection.
1121 *Front Microbiol.* 2019;10):429.
- 1122 17. Dang W, Xu L, Yin Y, Chen S, Wang W, Hakim MS, et al. IRF-1, RIG-I and MDA5
1123 display potent antiviral activities against norovirus coordinately induced by different
1124 types of interferons. *Antiviral Res.* 2018;155):48-59.
- 1125 18. Masola V, Bellin G, Gambaro G and Onisto M. Heparanase: A multitasking protein
1126 involved in extracellular matrix (ECM) remodeling and intracellular events. *Cells.*
1127 2018;7(12):236.
- 1128 19. Schoggins JW. Recent advances in antiviral interferon-stimulated gene biology.
1129 *F1000Research.* 2018;7

- 1130 20. Spence JS, He R, Hoffmann H-H, Das T, Thinon E, Rice CM, et al. IFITM3 directly
1131 engages and shuttles incoming virus particles to lysosomes. *Nat Chem Biol.*
1132 2019;15(3):259-68.
- 1133 21. Haller O, Staeheli P, Schwemmler M and Kochs G. Mx GTPases: dynamin-like antiviral
1134 machines of innate immunity. *Trends Microbiol.* 2015;23(3):154-63.
- 1135 22. García-Sastre A. Ten strategies of interferon evasion by viruses. *Cell Host Microbe.*
1136 2017;22(2):176-84.
- 1137 23. Giotis ES, Robey RC, Skinner NG, Tomlinson CD, Goodbourn S and Skinner MA.
1138 Chicken interferome: avian interferon-stimulated genes identified by microarray and
1139 RNA-seq of primary chick embryo fibroblasts treated with a chicken type I interferon
1140 (IFN- α). *Vet Res.* 2016;47(1):1-12.
- 1141 24. Rusinova I, Forster S, Yu S, Kannan A, Masse M, Cumming H, et al. Interferome v2.
1142 0: an updated database of annotated interferon-regulated genes. *Nucleic Acids Res.*
1143 2012;41(D1):D1040-D6.
- 1144 25. OhAinle M, Helms L, Vermeire J, Roesch F, Humes D, Basom R, et al. A virus-
1145 packageable CRISPR screen identifies host factors mediating interferon inhibition of
1146 HIV. *Elife.* 2018;7):e39823.
- 1147 26. Zhang Y, Burke CW, Ryman KD and Klimstra WB. Identification and characterization
1148 of interferon-induced proteins that inhibit alphavirus replication. *J Virol.*
1149 2007;81(20):11246-55.
- 1150 27. Stark R, Grzelak M and Hadfield J. RNA sequencing: the teenage years. *Nature*
1151 *Reviews Genetics.* 2019;20(11):631-56.
- 1152 28. Pamela C, Kanchwala M, Liang H, Kumar A, Wang L-F, Xing C, et al. The IFN
1153 response in bats displays distinctive IFN-stimulated gene expression kinetics with
1154 atypical RNASEL induction. *The Journal of Immunology.* 2018;200(1):209-17.
- 1155 29. Feld JJ, Nanda S, Huang Y, Chen W, Cam M, Pusek SN, et al. Hepatic gene expression
1156 during treatment with peginterferon and ribavirin: Identifying molecular pathways for
1157 treatment response. *Hepatology.* 2007;46(5):1548-63.
- 1158 30. Dalman MR, Deeter A, Nimishakavi G and Duan Z-H. Fold change and p-value cutoffs
1159 significantly alter microarray interpretations. In: *BMC Bioinformatics* 2012, pp.1-4.
1160 BioMed Central.
- 1161 31. Trilling M, Bellora N, Rutkowski AJ, de Graaf M, Dickinson P, Robertson K, et al.
1162 Deciphering the modulation of gene expression by type I and II interferons combining
1163 4sU-tagging, translational arrest and in silico promoter analysis. *Nucleic Acids Res.*
1164 2013;41(17):8107-25.
- 1165 32. O'Leary NA, Wright MW, Brister JR, Ciuffo S, Haddad D, McVeigh R, et al. Reference
1166 sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and
1167 functional annotation. *Nucleic Acids Res.* 2016;44(D1):D733-D45.
- 1168 33. Bragg JG, Potter S, Bi K and Moritz C. Exon capture phylogenomics: efficacy across
1169 scales of divergence. *Mol Ecol Resour.* 2016;16(5):1059-68.
- 1170 34. Kondrashov FA, Rogozin IB, Wolf YI and Koonin EV. Selection in the evolution of
1171 gene duplications. *Genome Biol.* 2002;3(2):1-9.
- 1172 35. Esposito M and Moreno-Hagelsieb G. Non-synonymous to synonymous substitutions
1173 suggest that orthologs tend to keep their functions, while paralogs are a source of
1174 functional novelty. *bioRxiv.* (2018):354704.
- 1175 36. MacFarland TW and Yates JM. Mann–whitney u test. *Introduction to nonparametric*
1176 *statistics for the biological sciences using R.* Springer; 2016. p. 103-32.
- 1177 37. Van den Eynden J and Larsson E. Mutational signatures are critical for proper
1178 estimation of purifying selection pressures in cancer somatic mutation data when using
1179 the dN/dS metric. *Front Genet.* 2017;8):74.

- 1180 38. Song H, Bremer BJ, Hinds EC, Raskutti G and Romero PA. Inferring protein sequence-
1181 function relationships with large-scale positive-unlabeled learning. *Cell Syst.* 2020;
- 1182 39. Pessia E, Popa A, Mousset S, Rezvoy C, Duret L and Marais GA. Evidence for
1183 widespread GC-biased gene conversion in eukaryotes. *Genome Biol Evol.*
1184 2012;4(7):675-82.
- 1185 40. Zhou Z, Dang Y, Zhou M, Li L, Yu C-h, Fu J, et al. Codon usage is an important
1186 determinant of gene expression levels largely through its effects on transcription.
1187 *Proceedings of the National Academy of Sciences.* 2016;113(41):E6117-E25.
- 1188 41. Sievers A, Bosiek K, Bisch M, Dreessen C, Riedel J, Froß P, et al. K-mer content,
1189 correlation, and position analysis of genome DNA sequences for the identification of
1190 function and evolutionary features. *Genes.* 2017;8(4):122.
- 1191 42. Lee NK, Li X and Wang D. A comprehensive survey on genetic algorithms for DNA
1192 motif prediction. *Inf Sci.* 2018;466):25-43.
- 1193 43. Noble WS. How does multiple testing correction work? *Nat Biotechnol.*
1194 2009;27(12):1135-7.
- 1195 44. Di Rienzo L, Miotto M, Bò L, Ruocco G, Raimondo D and Milanetti E. Characterizing
1196 hydropathy of amino acid side chain in a protein environment by investigating the
1197 structural changes of water molecules network. *Front Mol Biosci.* 2021;8
- 1198 45. Bhadra P, Yan J, Li J, Fong S and Siu SW. AmPEP: Sequence-based prediction of
1199 antimicrobial peptides using distribution patterns of amino acid properties and random
1200 forest. *Sci Rep.* 2018;8(1):1-10.
- 1201 46. Pommié C, Levadoux S, Sabatier R, Lefranc G and Lefranc MP. IMGT standardized
1202 criteria for statistical analysis of immunoglobulin V- REGION amino acid properties.
1203 *J Mol Recognit.* 2004;17(1):17-32.
- 1204 47. Kumar M, Gouw M, Michael S, Sámano-Sánchez H, Pancsa R, Glavina J, et al. ELM—
1205 the eukaryotic linear motif resource in 2020. *Nucleic Acids Res.* 2020;48(D1):D296-
1206 D306.
- 1207 48. Pflieger CM and Kirschner MW. The KEN box: an APC recognition signal distinct from
1208 the D box targeted by Cdh1. *Genes Dev.* 2000;14(6):655-65.
- 1209 49. Fehr AR and Yu D. Control the host cell cycle: viral regulation of the anaphase-
1210 promoting complex. *J Virol.* 2013;87(16):8818-25.
- 1211 50. Bösl K, Ianevski A, Than TT, Andersen PI, Kuivanen S, Teppor M, et al. Common
1212 nodes of virus–host interaction revealed through an integrated network analysis. *Front*
1213 *Immunol.* 2019;10):2186.
- 1214 51. Wright PE and Dyson HJ. Intrinsically disordered proteins in cellular signalling and
1215 regulation. *Nat Rev Mol Cell Biol.* 2015;16(1):18-29.
- 1216 52. Mészáros B, Erdős G and Dosztányi Z. IUPred2A: context-dependent prediction of
1217 protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.*
1218 2018;46(W1):W329-W37.
- 1219 53. Hagai T, Azia A, Babu MM and Andino R. Use of host-like peptide motifs in viral
1220 proteins is a prevalent strategy in host-virus interactions. *Cell Rep.* 2014;7(5):1729-39.
- 1221 54. Michael S, Travé G, Ramu C, Chica C and Gibson TJ. Discovery of candidate KEN-
1222 box motifs using cell cycle keyword enrichment combined with native disorder
1223 prediction and motif conservation. *Bioinformatics.* 2008;24(4):453-7.
- 1224 55. Alanis-Lobato G, Andrade-Navarro MA and Schaefer MH. HIPPIE v2.0: enhancing
1225 meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids*
1226 *Res.* 2016):gkw985.
- 1227 56. Abedi M and Gheisari Y. Nodes with high centrality in protein interaction networks are
1228 responsible for driving signaling pathways in diabetic nephropathy. *PeerJ.*
1229 2015;3):e1284.

- 1230 57. Ozato K, Shin D-M, Chang T-H and Morse HC. TRIM family proteins and their
1231 emerging roles in innate immunity. *Nat Rev Immunol.* 2008;8(11):849-60.
- 1232 58. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl
1233 2020. *Nucleic Acids Res.* 2020;48(D1):D682-D8.
- 1234 59. Shaw AE, Rihn SJ, Mollentze N, Wickenhagen A, Stewart DG, Orton RJ, et al. The
1235 antiviral state has shaped the CpG composition of the vertebrate interferome to avoid
1236 self-targeting. *PLoS Biol.* 2021;19(9):e3001352.
- 1237 60. Zhang M-L and Zhou Z-H. ML-KNN: A lazy learning approach to multi-label learning.
1238 *Pattern recognition.* 2007;40(7):2038-48.
- 1239 61. Chang C-C and Lin C-J. LIBSVM: A library for support vector machines. *ACM Trans*
1240 *Intell Syst Technol.* 2011;2(3):1-27.
- 1241 62. Sivaranjani S, Ananya S, Aravinth J and Karthika R. Diabetes prediction using machine
1242 learning algorithms with feature selection and dimensionality reduction. In: *2021 7th*
1243 *International Conference on Advanced Computing and Communication Systems*
1244 *(ICACCS) 2021*, pp.141-6. IEEE.
- 1245 63. Cheng D, Zhang S, Deng Z, Zhu Y and Zong M. kNN algorithm with data-driven k
1246 value. In: *International Conference on Advanced Data Mining and Applications 2014*,
1247 pp.499-512. Springer.
- 1248 64. Liu X-Y, Wu J and Zhou Z-H. Exploratory undersampling for class-imbalance learning.
1249 *IEEE Trans Syst Man Cybern.* 2008;39(2):539-50.
- 1250 65. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The genotype-tissue
1251 expression (GTEx) project. *Nat Genet.* 2013;45(6):580-5.
- 1252 66. Papatheodorou I, Moreno P, Manning J, Fuentes AM-P, George N, Fexova S, et al.
1253 Expression Atlas update: from tissues to single cells. *Nucleic Acids Res.*
1254 2020;48(D1):D77-D83.
- 1255 67. Jeong H, Mason SP, Barabási A-L and Oltvai ZN. Lethality and centrality in protein
1256 networks. *Nature.* 2001;411(6833):41-2.
- 1257 68. Hahn MW and Kern AD. Comparative genomics of centrality and essentiality in three
1258 eukaryotic protein-interaction networks. *Mol Biol Evol.* 2005;22(4):803-6.
- 1259 69. Batada NN, Hurst LD and Tyers M. Evolutionary and physiological importance of hub
1260 proteins. *PLoS Comput Biol.* 2006;2(7):e88.
- 1261 70. Pérez-Martínez D. Innate immunity in vertebrates: an overview. *Immunology.*
1262 2016;148(2):125-39.
- 1263 71. Jopling CL. Mutations: Stop that nonsense! *Elife.* 2014;3):e04300.
- 1264 72. Zhu X, Pribis JP, Rodriguez PC, Morris Jr SM, Vodovotz Y, Billiar TR, et al. The
1265 central role of arginine catabolism in T-cell dysfunction and increased susceptibility to
1266 infection after physical injury. *Ann Surg.* 2014;259(1):171-8.
- 1267 73. Morris CR, Hamilton- Reeves J, Martindale RG, Sarav M and Ochoa Gautier JB.
1268 Acquired amino acid deficiencies: a focus on arginine and glutamine. *Nutr Clin Pract.*
1269 2017;32):30S-47S.
- 1270 74. Levring TB, Hansen AK, Nielsen BL, Kongsbak M, Von Essen MR, Woetmann A, et
1271 al. Activated human CD4+ T cells express transporters for both cysteine and cystine.
1272 *Sci Rep.* 2012;2(1):1-6.
- 1273 75. Sikalidis AK. Amino acids and immune response: a role for cysteine, glutamine,
1274 phenylalanine, tryptophan and arginine in T-cell function and cancer? *Pathol Oncol Res.*
1275 2015;21(1):9-17.
- 1276 76. Yin C, Zheng T and Chang X. Biosynthesis of S-Adenosylmethionine by magnetically
1277 immobilized *Escherichia coli* cells highly expressing a methionine adenosyltransferase
1278 variant. *Molecules.* 2017;22(8):1365.

- 1279 77. Feld JJ, Modi AA, El-Diwany R, Rotman Y, Thomas E, Ahlenstiel G, et al. S-adenosyl
1280 methionine improves early viral responses and interferon-stimulated gene induction in
1281 hepatitis C nonresponders. *Gastroenterology*. 2011;140(3):830-9.
- 1282 78. Yeom S, Giacomelli I, Fredrikson M and Jha S. Privacy risk in machine learning:
1283 Analyzing the connection to overfitting. In: *2018 IEEE 31st Computer Security*
1284 *Foundations Symposium (CSF) 2018*, pp.268-82. IEEE.
- 1285 79. Li S-W, Lai C-C, Ping J-F, Tsai F-J, Wan L, Lin Y-J, et al. Severe acute respiratory
1286 syndrome coronavirus papain-like protease suppressed alpha interferon-induced
1287 responses through downregulation of extracellular signal-regulated kinase 1-mediated
1288 signalling pathways. *J Gen Virol*. 2011;92(5):1127-40.
- 1289 80. Flo TH, Smith KD, Sato S, Rodriguez DJ, Holmes MA, Strong RK, et al. Lipocalin 2
1290 mediates an innate immune response to bacterial infection by sequestering iron. *Nature*.
1291 2004;432(7019):917-21.
- 1292 81. Tissot C, Rebouissou C, Klein B and Mechti N. Both human α/β and γ interferons
1293 upregulate the expression of CD48 cell surface molecules. *J Interferon Cytokine Res*.
1294 1997;17(1):17-26.
- 1295 82. Noçon AL, Ip JP, Terry R, Lim SL, Getts DR, Müller M, et al. The bacteriostatic protein
1296 lipocalin 2 is induced in the central nervous system of mice with West Nile virus
1297 encephalitis. *J Virol*. 2014;88(1):679-89.
- 1298 83. Zarama A, Perez-Carmona N, Farre D, Tomic A, Borst EM, Messerle M, et al.
1299 Cytomegalovirus m154 hinders CD48 cell-surface expression and promotes viral
1300 escape from host natural killer cell control. *PLoS Pathog*. 2014;10(3):e1004000.
- 1301 84. Martínez-Vicente P, Farré D, Engel P and Angulo A. Divergent Traits and Ligand-
1302 Binding Properties of the Cytomegalovirus CD48 Gene Family. *Viruses*.
1303 2020;12(8):813.
- 1304 85. Ricquier D. UCP1, the mitochondrial uncoupling protein of brown adipocyte: a
1305 personal contribution and a historical perspective. *Biochimie*. 2017;134):3-8.
- 1306 86. Hossain MA, Larrous F, Rawlinson SM, Zhan J, Sethi A, Ibrahim Y, et al. Structural
1307 elucidation of viral antagonism of innate immunity at the STAT1 interface. *Cell Rep*.
1308 2019;29(7):1934-45. e8.
- 1309 87. Yu X, Liu H, Hamel KA, Morvan MG, Yu S, Leff J, et al. Dorsal root ganglion
1310 macrophages contribute to both the initiation and persistence of neuropathic pain. *Nat*
1311 *Commun*. 2020;11(1):1-12.
- 1312 88. Chen Y, Lun AT and Smyth GK. From reads to genes to pathways: differential
1313 expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-
1314 likelihood pipeline. *F1000Research*. 2016;5:1438. doi: 10.12688/f1000research.8987.2.
- 1315 89. Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, et al. Ensembl
1316 comparative genomics resources. *Database*. 2016;2016):bav096.
- 1317 90. Li HD, Menon R, Omenn GS and Guan Y. Revisiting the identification of canonical
1318 splice isoforms through integration of functional genomics and proteomics evidence.
1319 *Proteomics*. 2014;14(23-24):2709-18.
- 1320 91. Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, Mayr C, et al. Alternative
1321 isoform regulation in human tissue transcriptomes. *Nature*. 2008;456(7221):470-6.
- 1322 92. Sieber P, Platzer M and Schuster S. The definition of open reading frame revisited.
1323 *Trends Genet*. 2018;34(3):167-70.
- 1324 93. Pan Q, Shai O, Lee LJ, Frey BJ and Blencowe BJ. Deep surveying of alternative
1325 splicing complexity in the human transcriptome by high-throughput sequencing. *Nat*
1326 *Genet*. 2008;40(12):1413-5.

- 1327 94. Guéguen L and Duret L. Unbiased estimate of synonymous and nonsynonymous
1328 substitution rates with nonstationary base composition. *Mol Biol Evol.* 2018;35(3):734-
1329 42.
- 1330 95. Takata MA, Gonçalves-Carneiro D, Zang TM, Soll SJ, York A, Blanco-Melo D, et al.
1331 CG dinucleotide suppression enables antiviral defence targeting non-self RNA. *Nature.*
1332 2017;550(7674):124-7.
- 1333 96. Yu C-H, Dang Y, Zhou Z, Wu C, Zhao F, Sachs MS, et al. Codon usage influences the
1334 local rate of translation elongation to regulate co-translational protein folding. *Mol Cell.*
1335 2015;59(5):744-54.
- 1336 97. Tufarelli C, Ahmad A, Strohbuecker S, Scotti C and Sottile V. In Silico Identification
1337 of SOX1 Post-Translational Modifications Highlights a Shared Protein Motif. 2020;
- 1338 98. Yoon J, Blumer A and Lee K. An algorithm for modularity analysis of directed and
1339 weighted biological networks based on edge-betweenness centrality. *Bioinformatics.*
1340 2006;22(24):3106-8.
- 1341 99. Friedel CC and Zimmer R. Influence of degree correlations on network structure and
1342 stability in protein-protein interaction networks. *BMC Bioinformatics.* 2007;8(1):1-10.
- 1343 100. Ravasz E, Somera AL, Mongru DA, Oltvai ZN and Barabási A-L. Hierarchical
1344 organization of modularity in metabolic networks. *Science.* 2002;297(5586):1551-5.
- 1345 101. Chicco D and Jurman G. The advantages of the Matthews correlation coefficient (MCC)
1346 over F1 score and accuracy in binary classification evaluation. *BMC Genomics.*
1347 2020;21(1):1-13.
- 1348 102. Chai H; Gu Q; Robertson DL; Hughes J (2022): Supporting data for "Defining the
1349 characteristics of interferon-alpha-stimulated human genes: insight from expression
1350 data and machine-learning" GigaScience Database. <http://dx.doi.org/10.5524/102322>
1351

Figure 1

[Click here to access/download;Figure;Figure_1.eps](#)

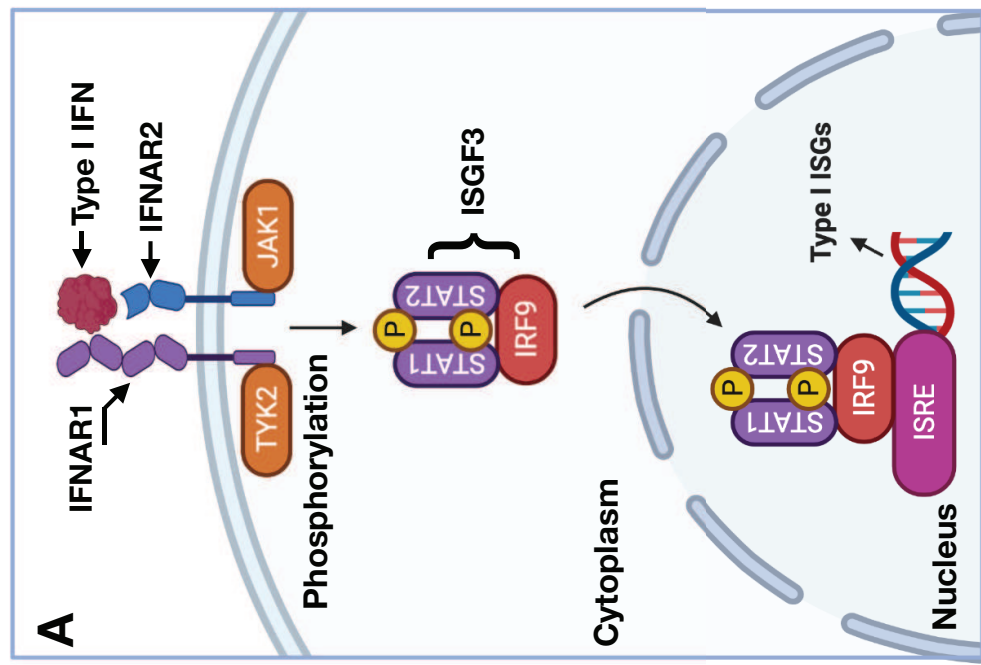
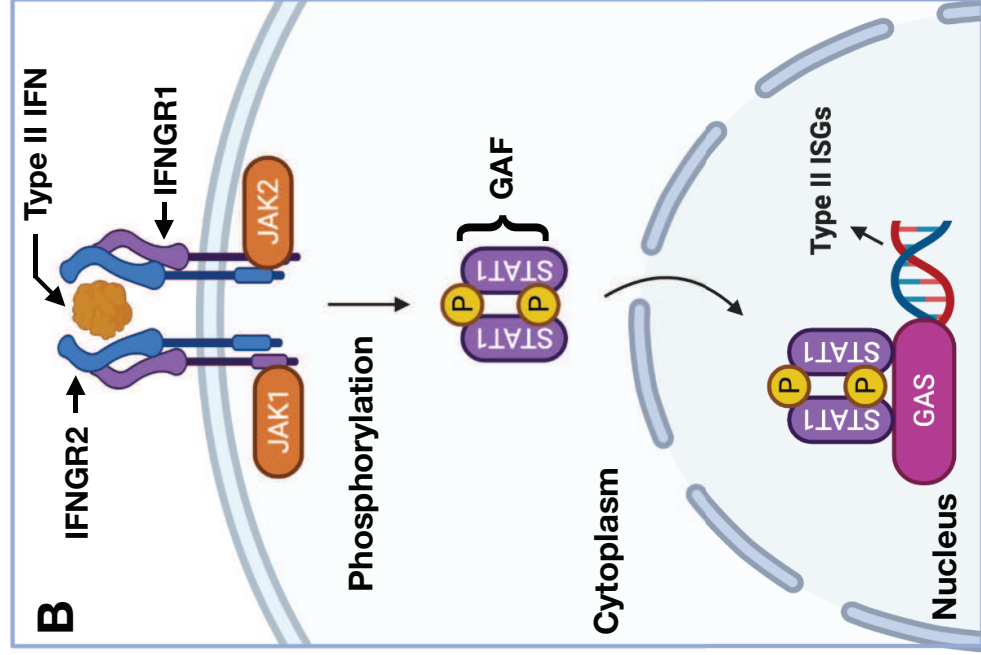
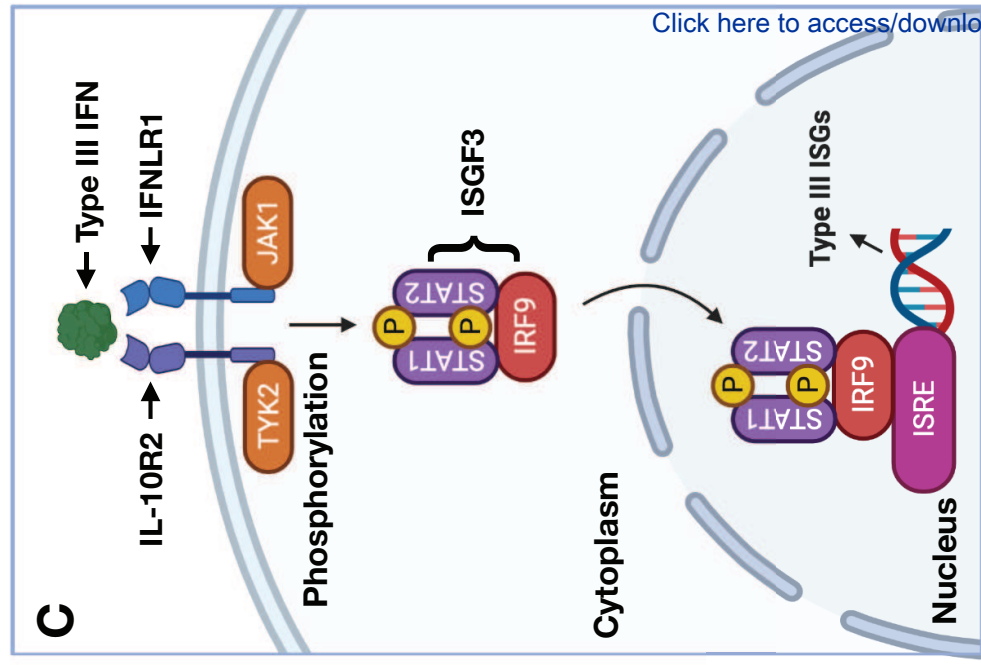
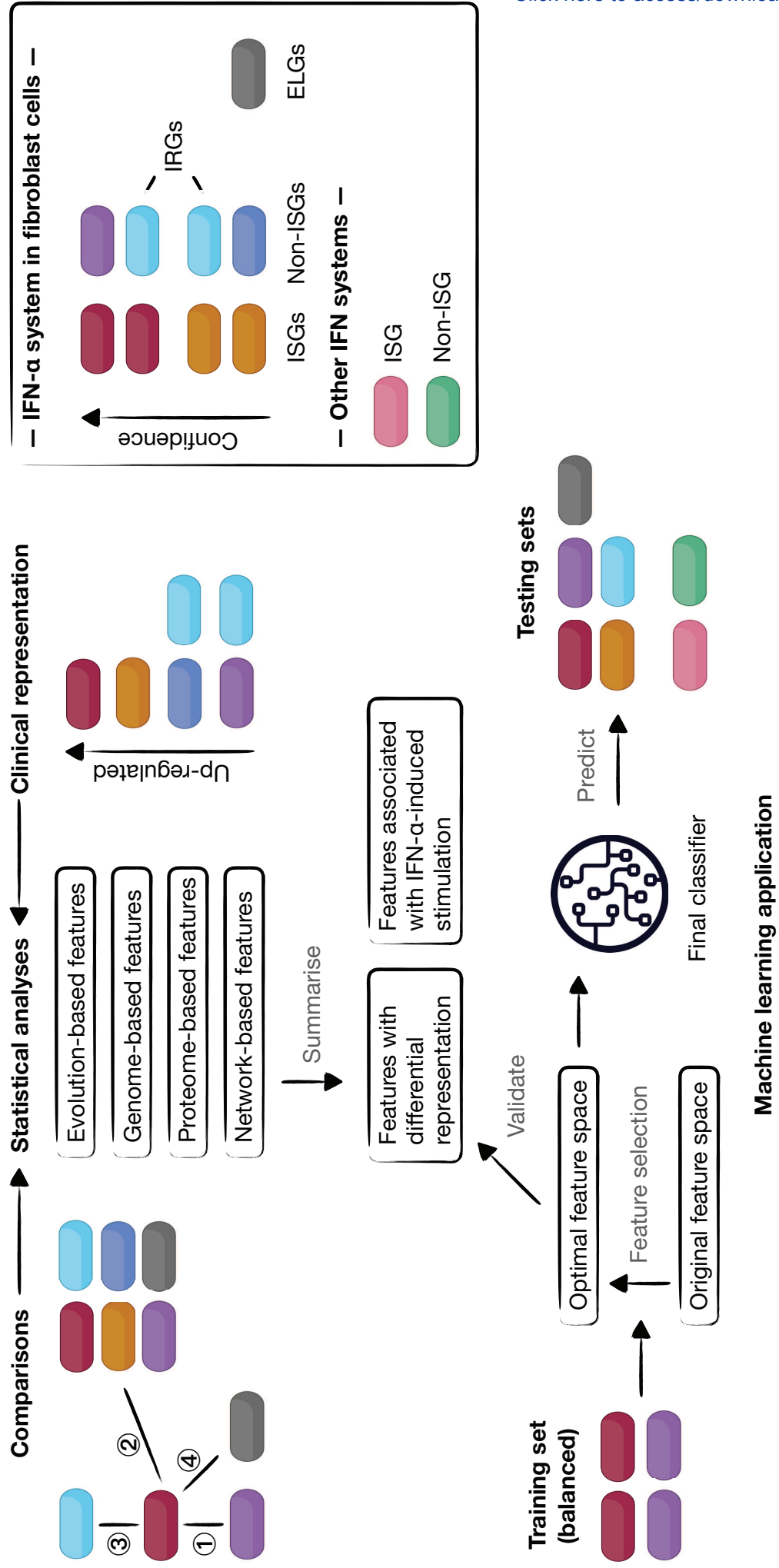


Figure 2

[Click here to access/download;Figure;Figure_2.eps](#)



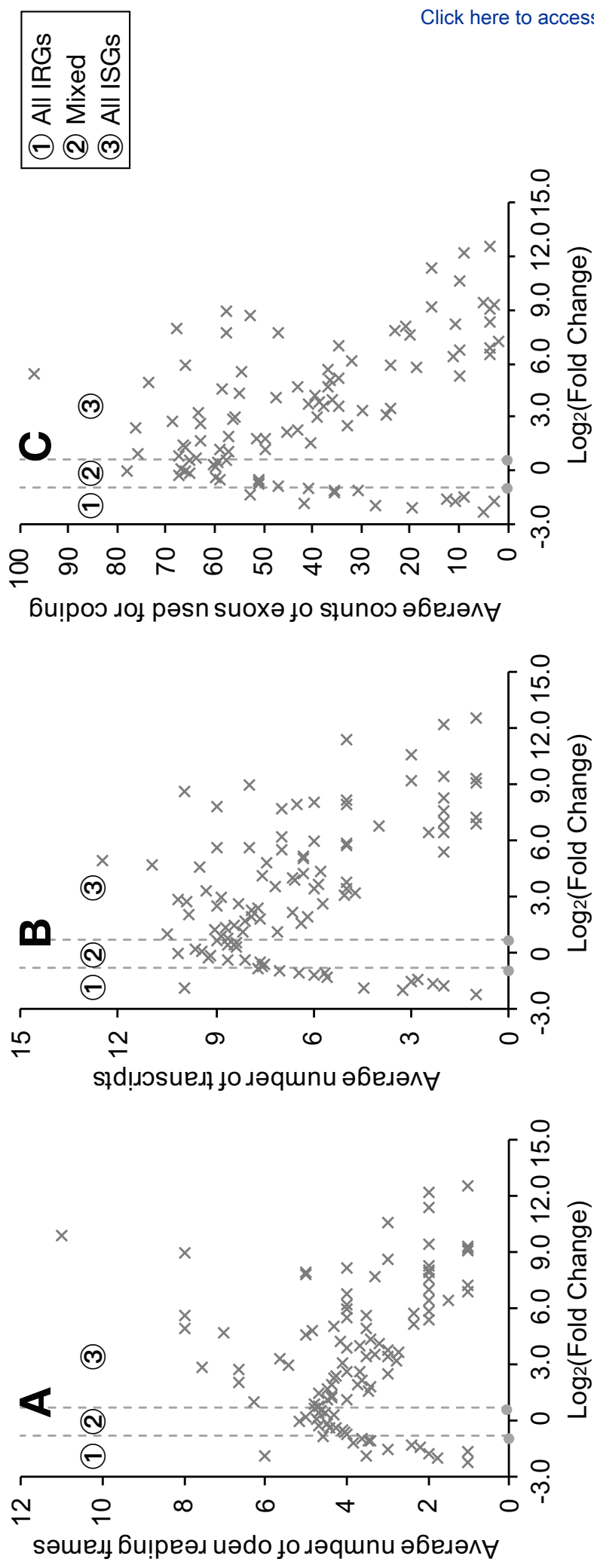


Figure 4

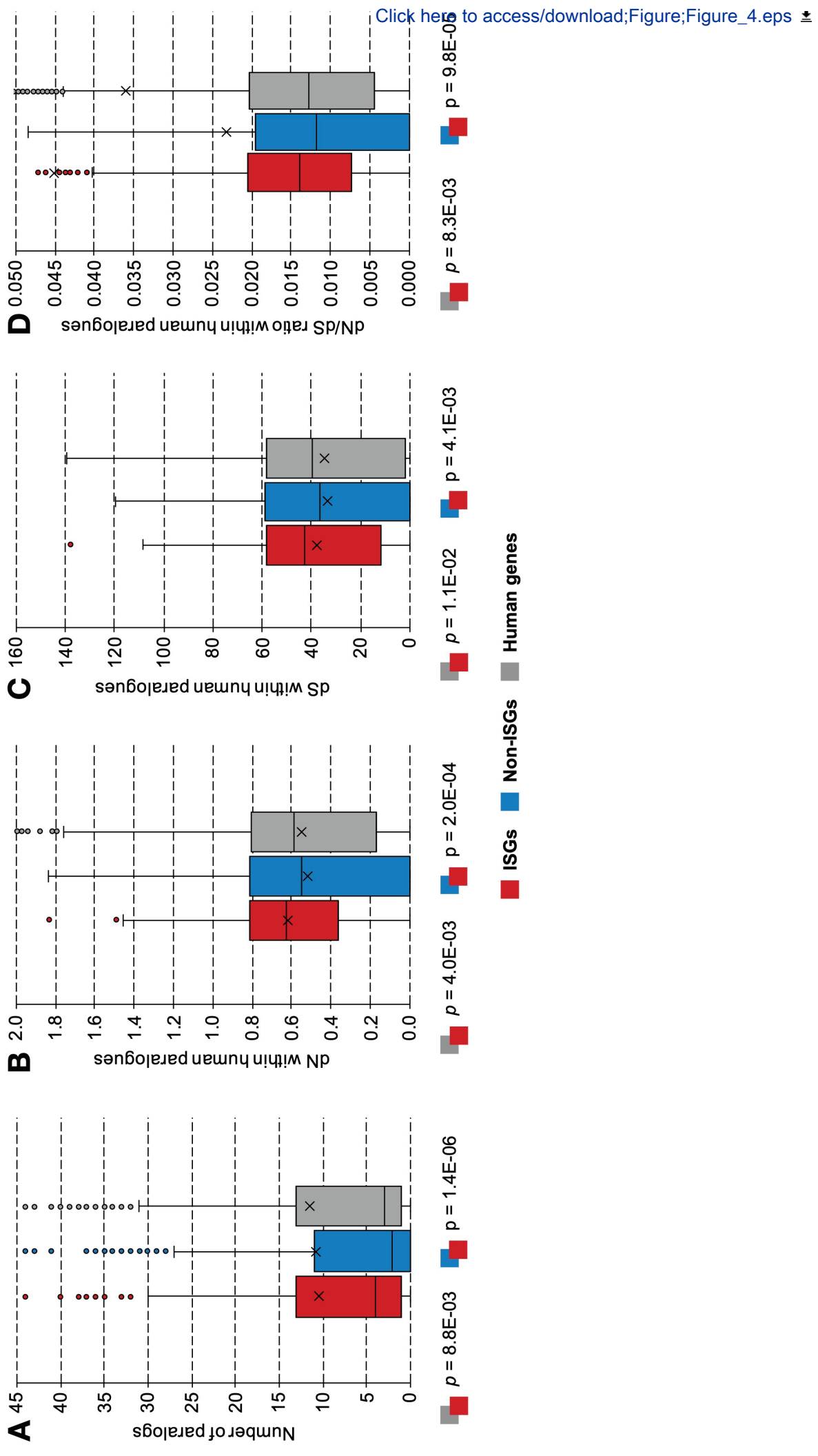
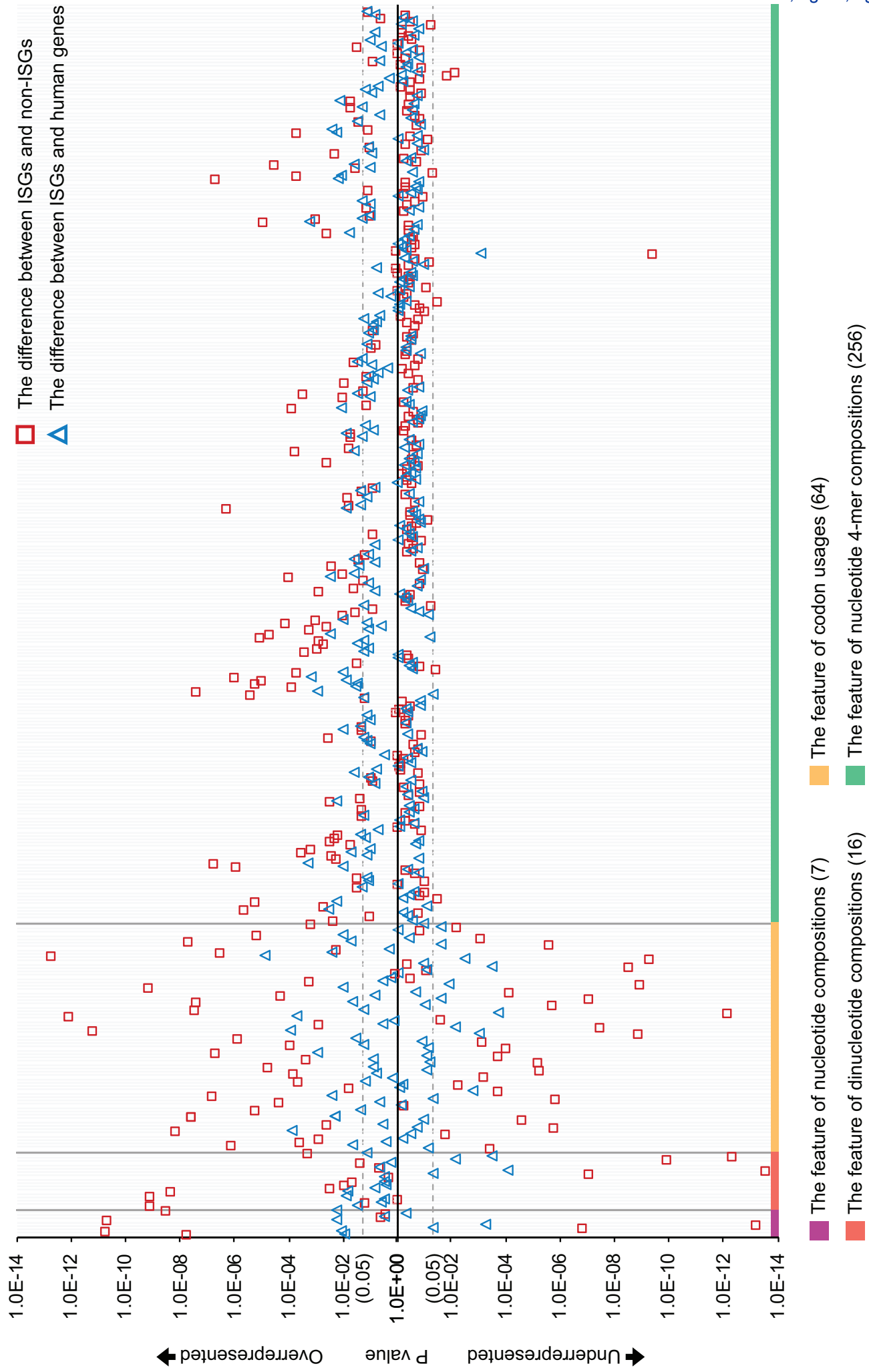
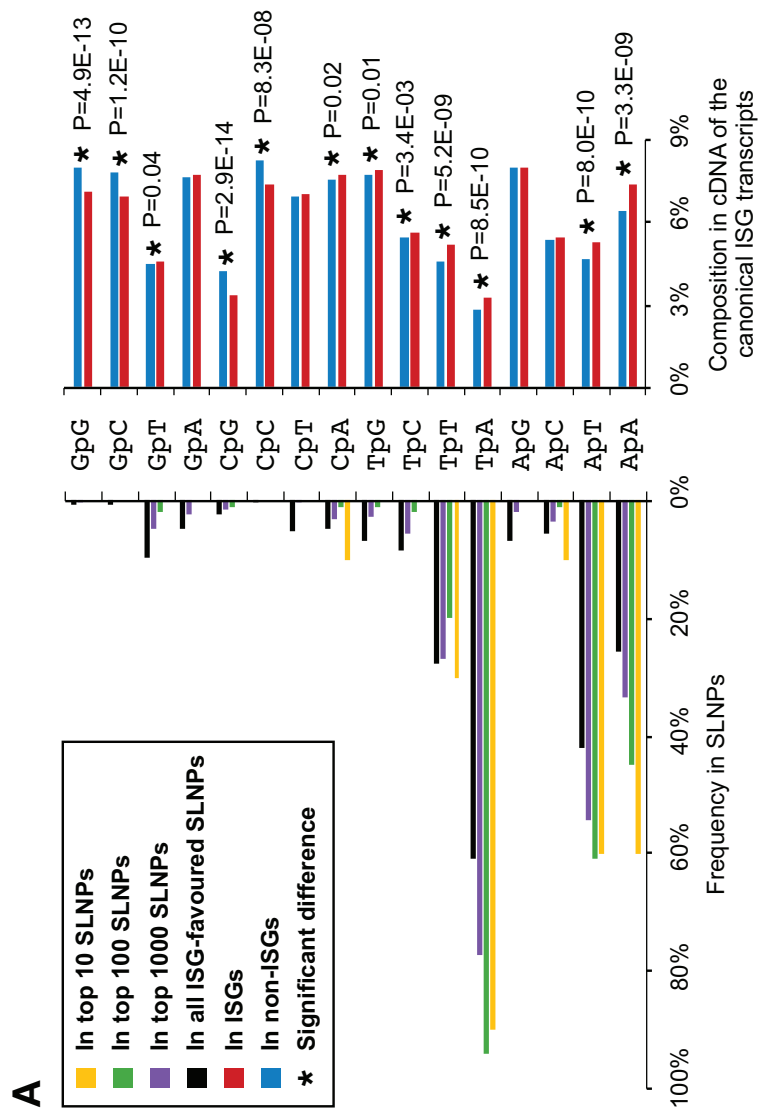
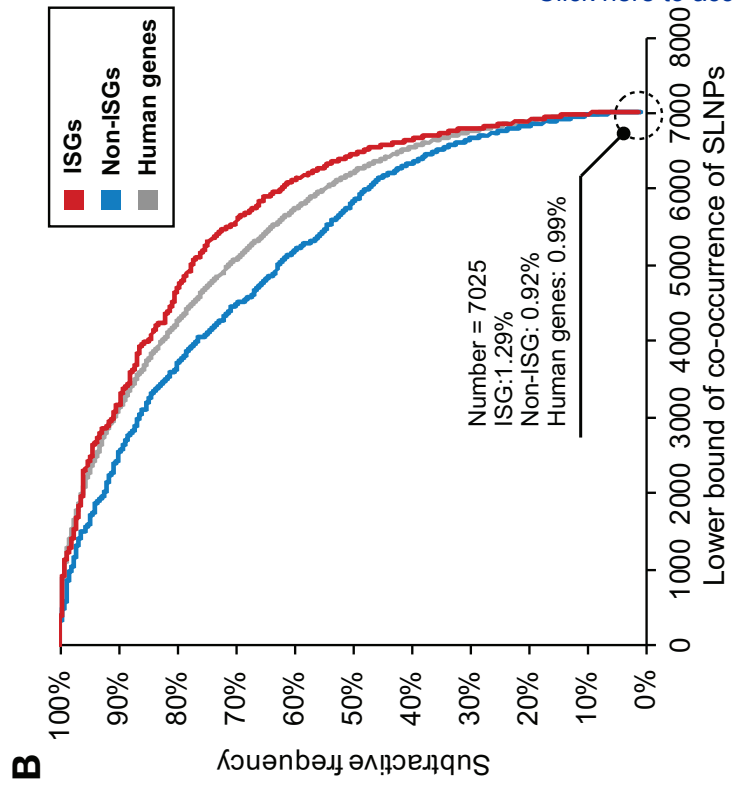
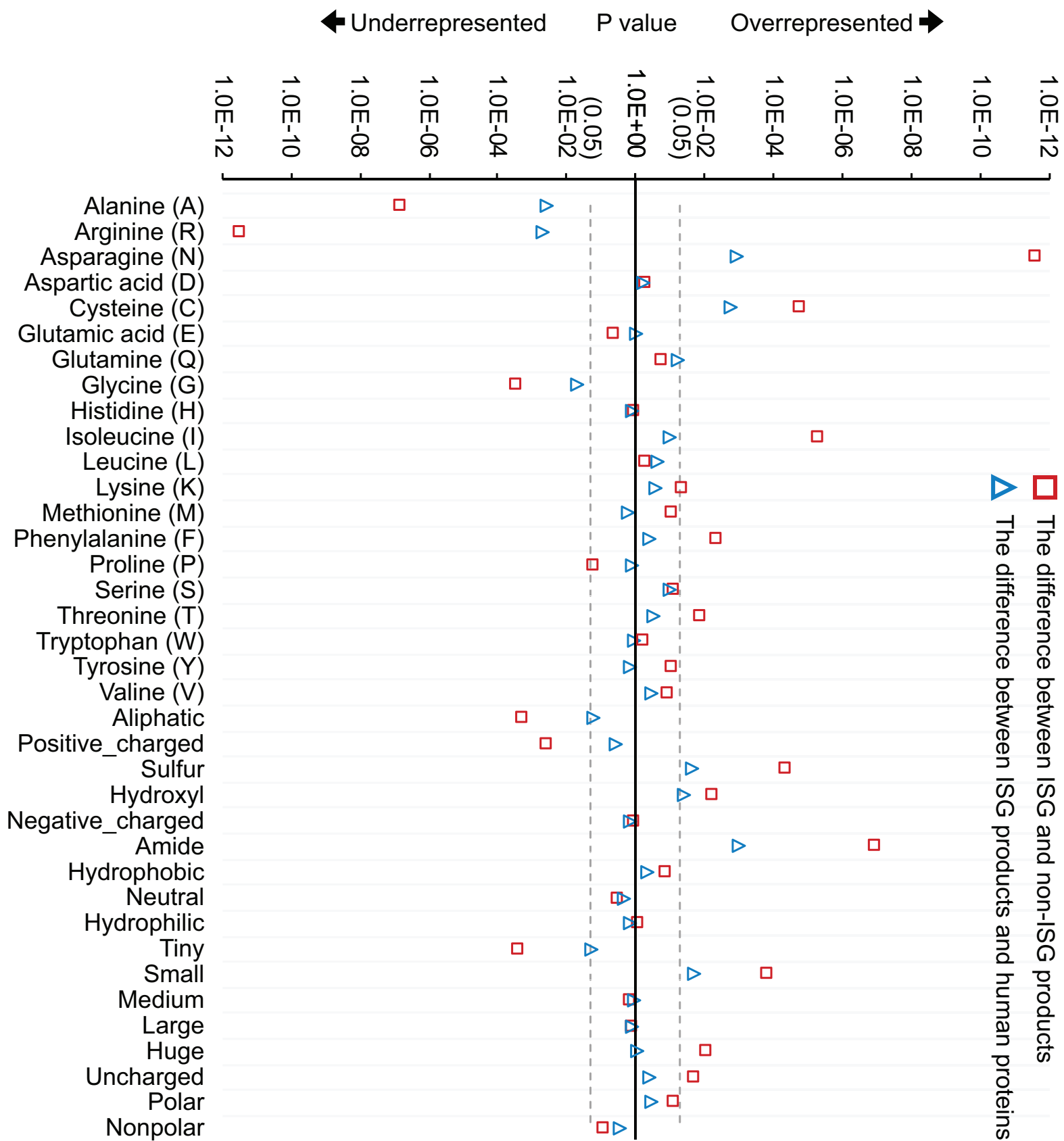


Figure 5

[Click here to access/download;Figure;Figure_5.eps](#)







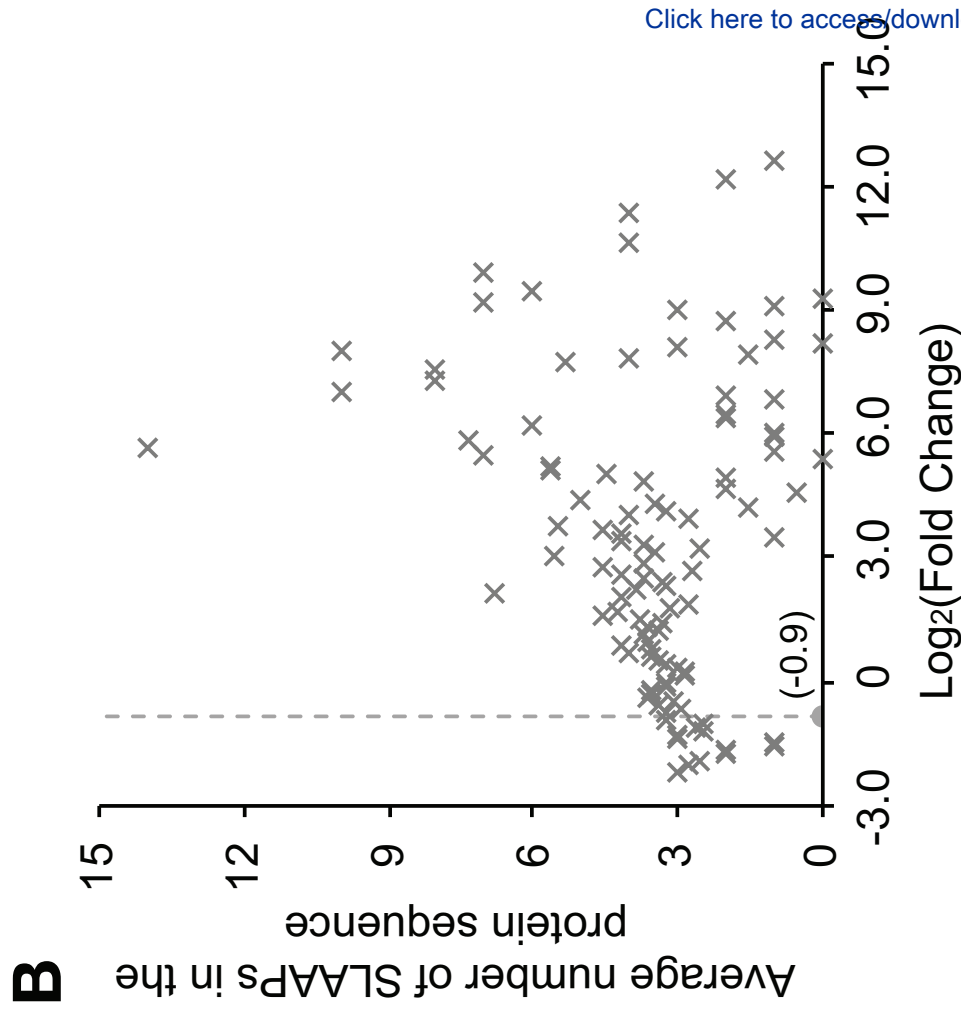
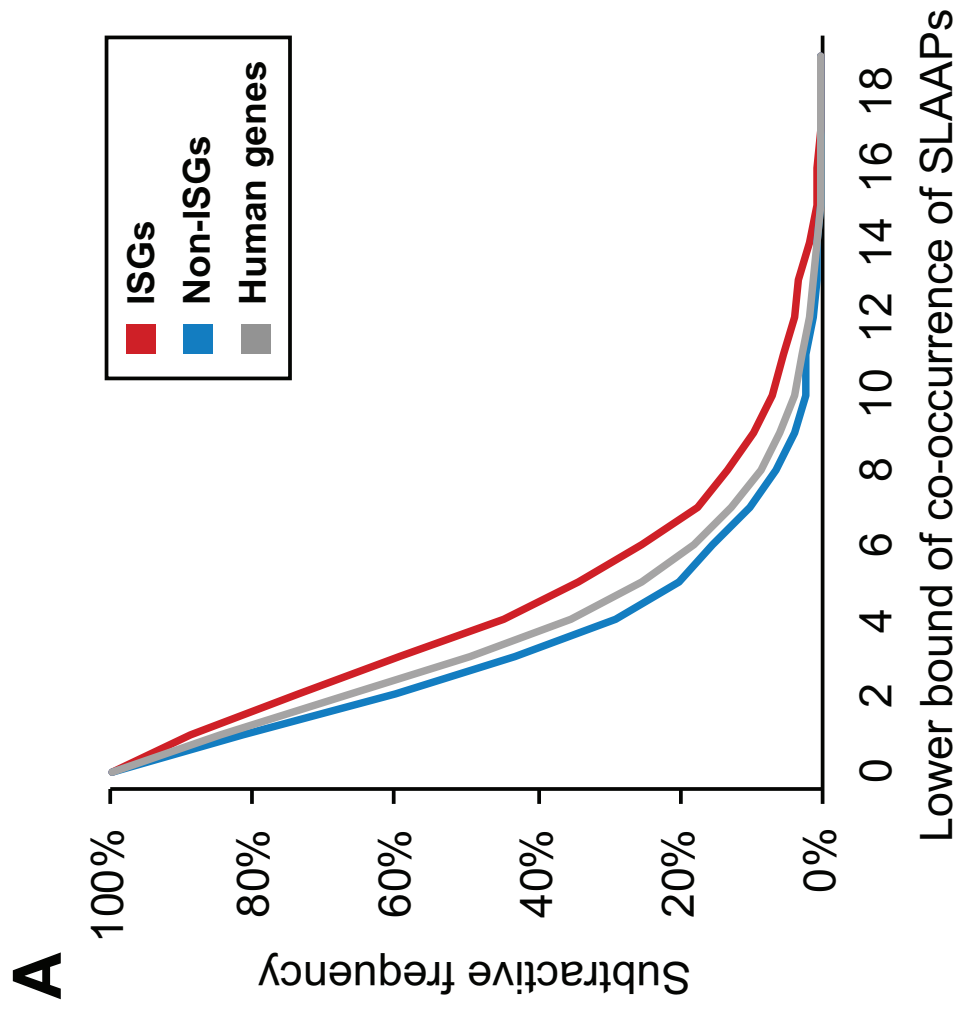


Figure 9

[Click here to access/download;Figure;Figure_9.eps](#)

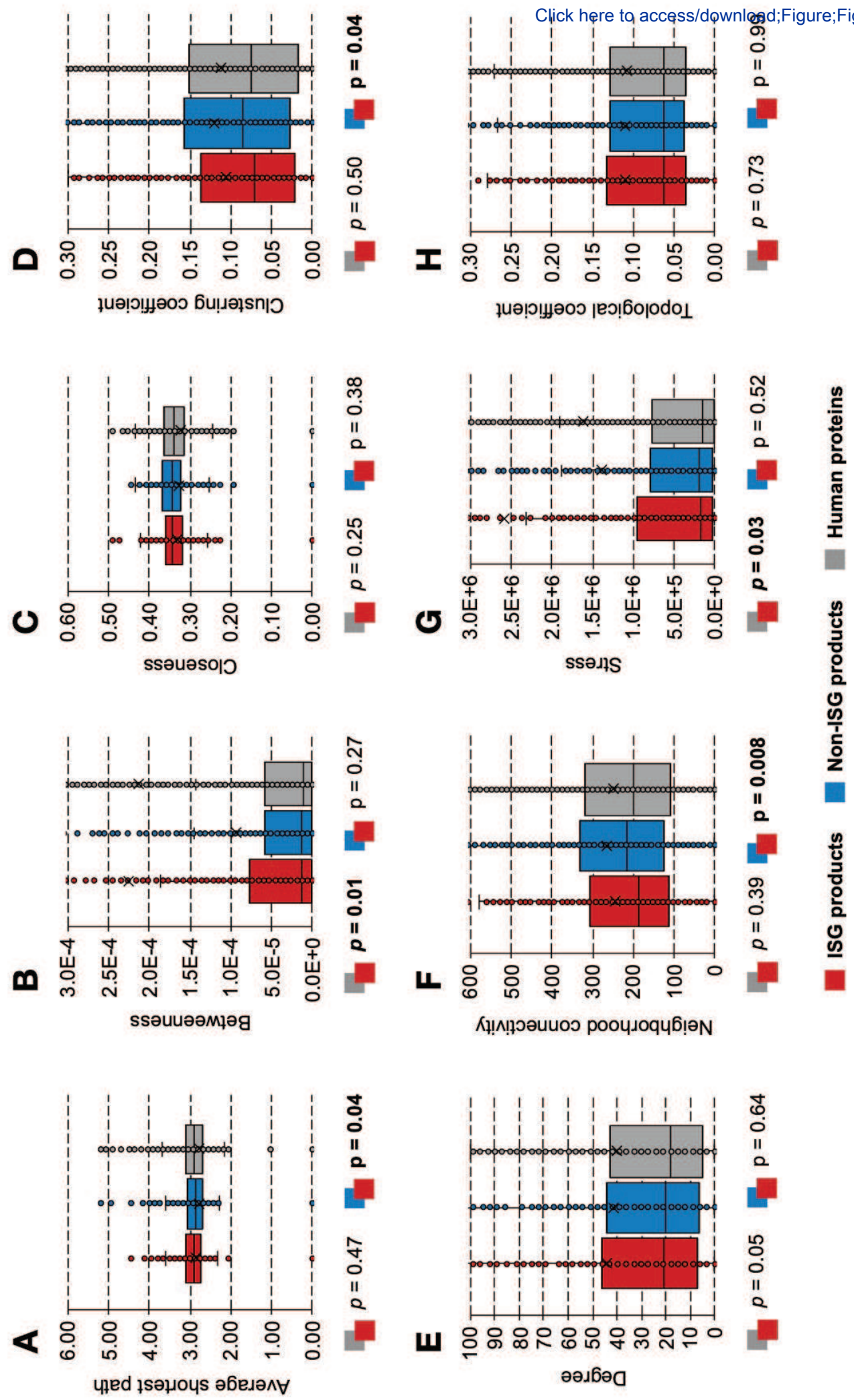
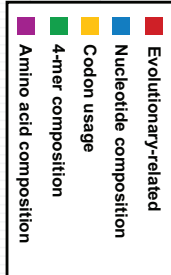
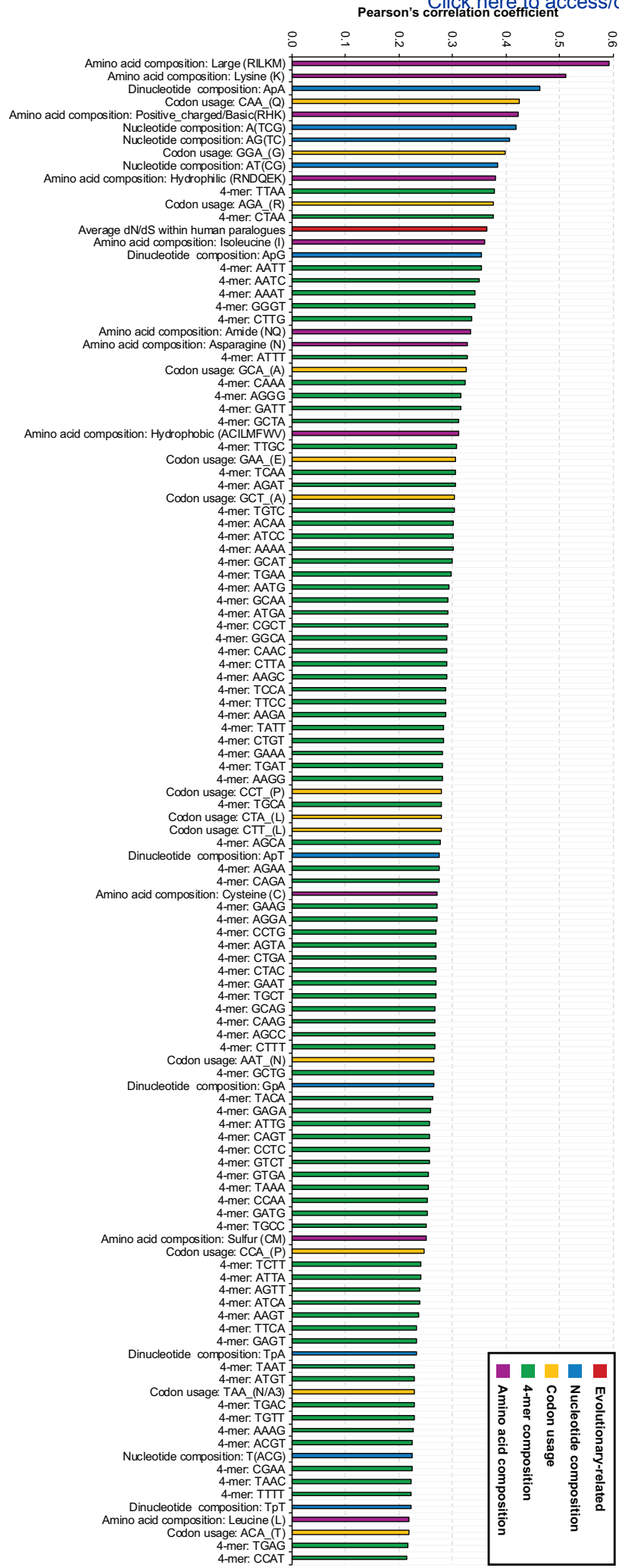


Figure 10

[Click here to access/download;Figure;Figure_10.eps](#)



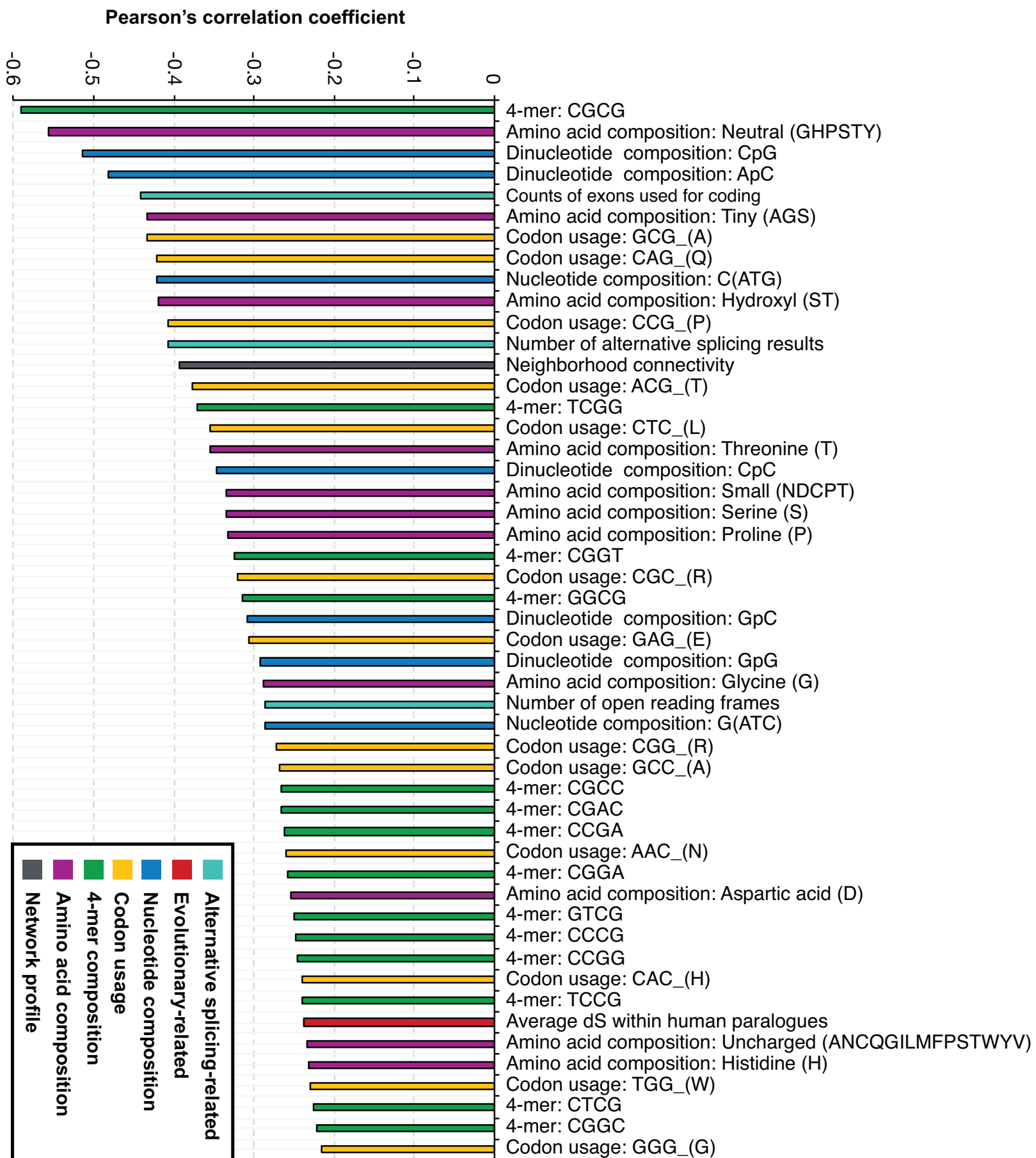
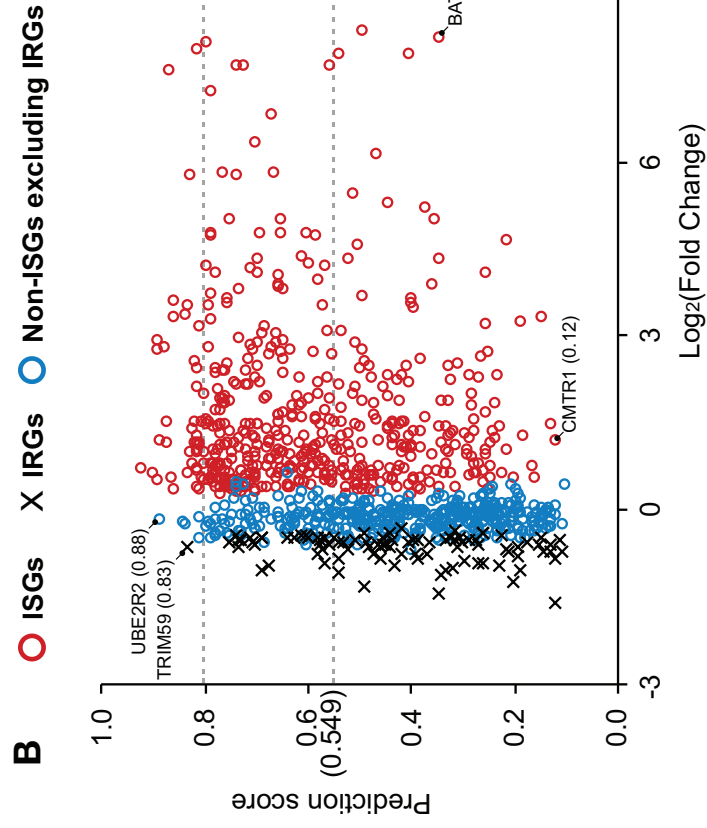
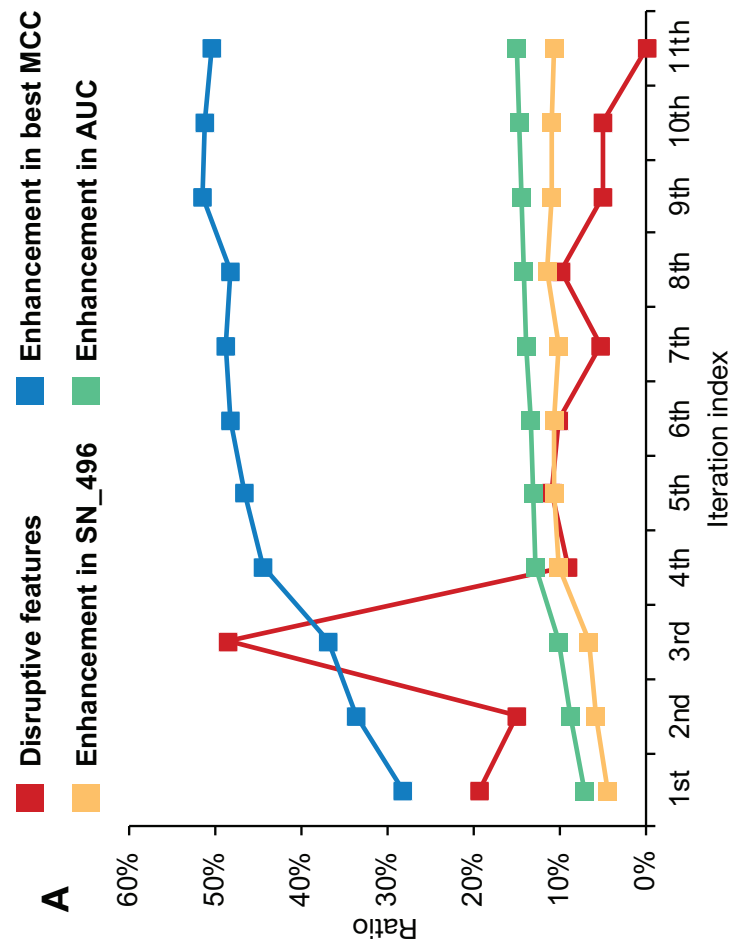


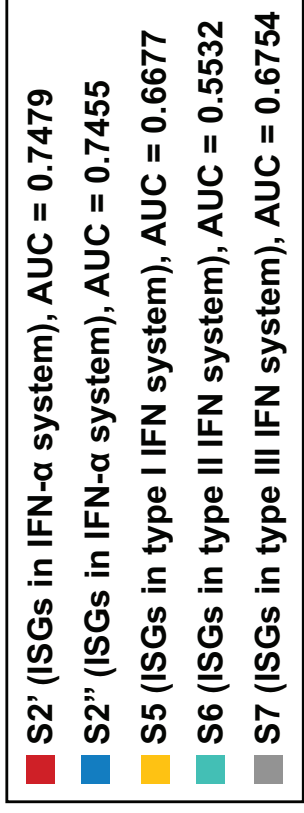
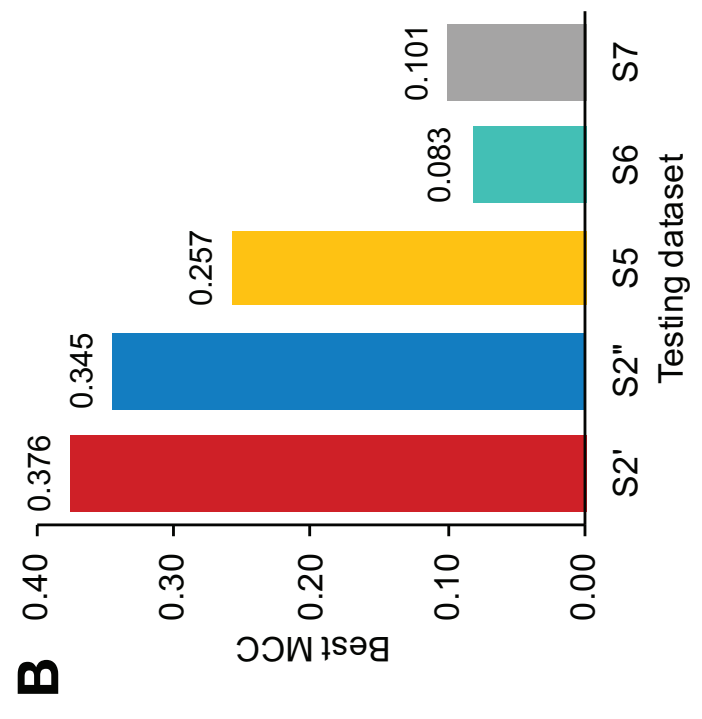
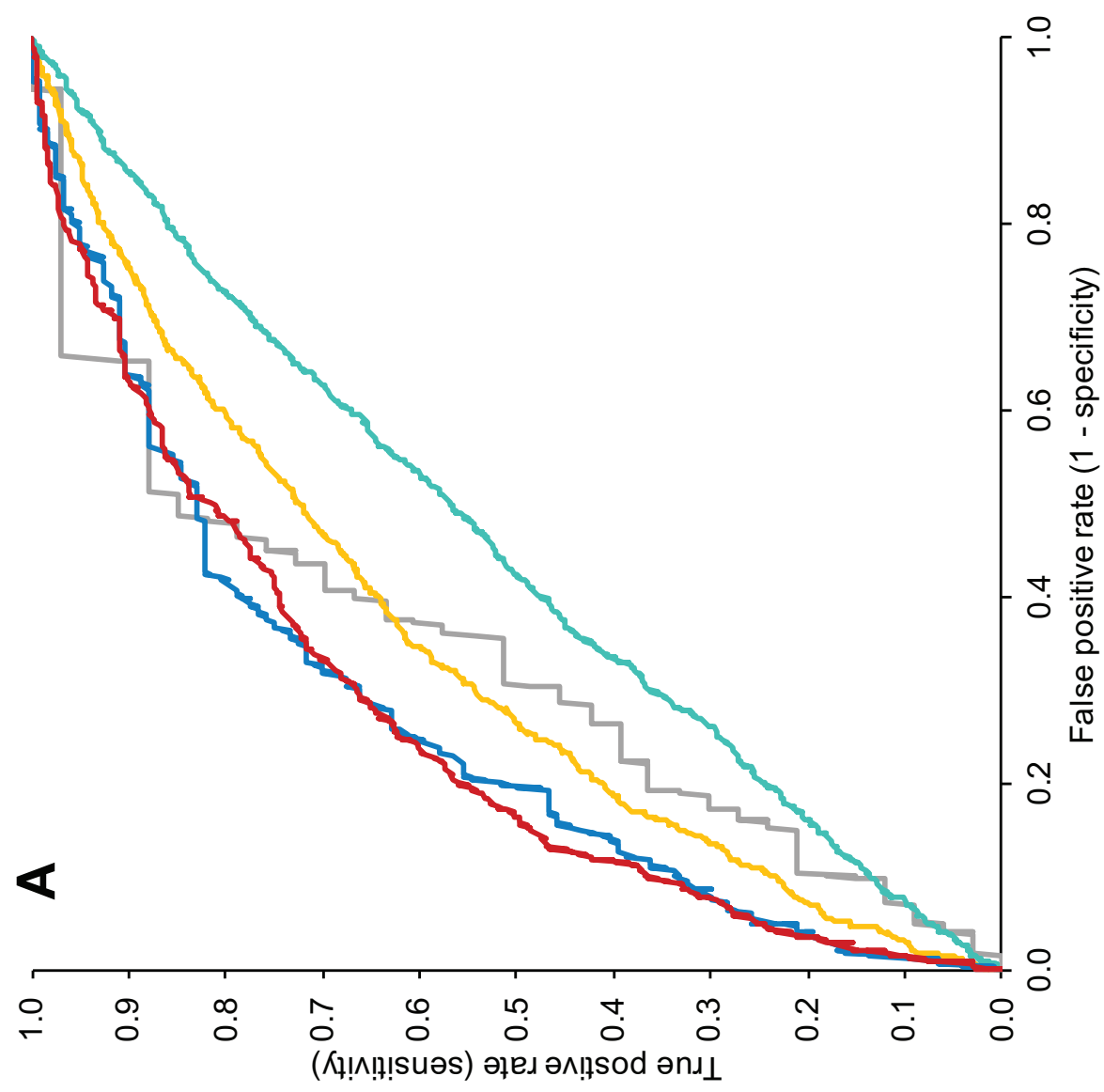
Figure 12

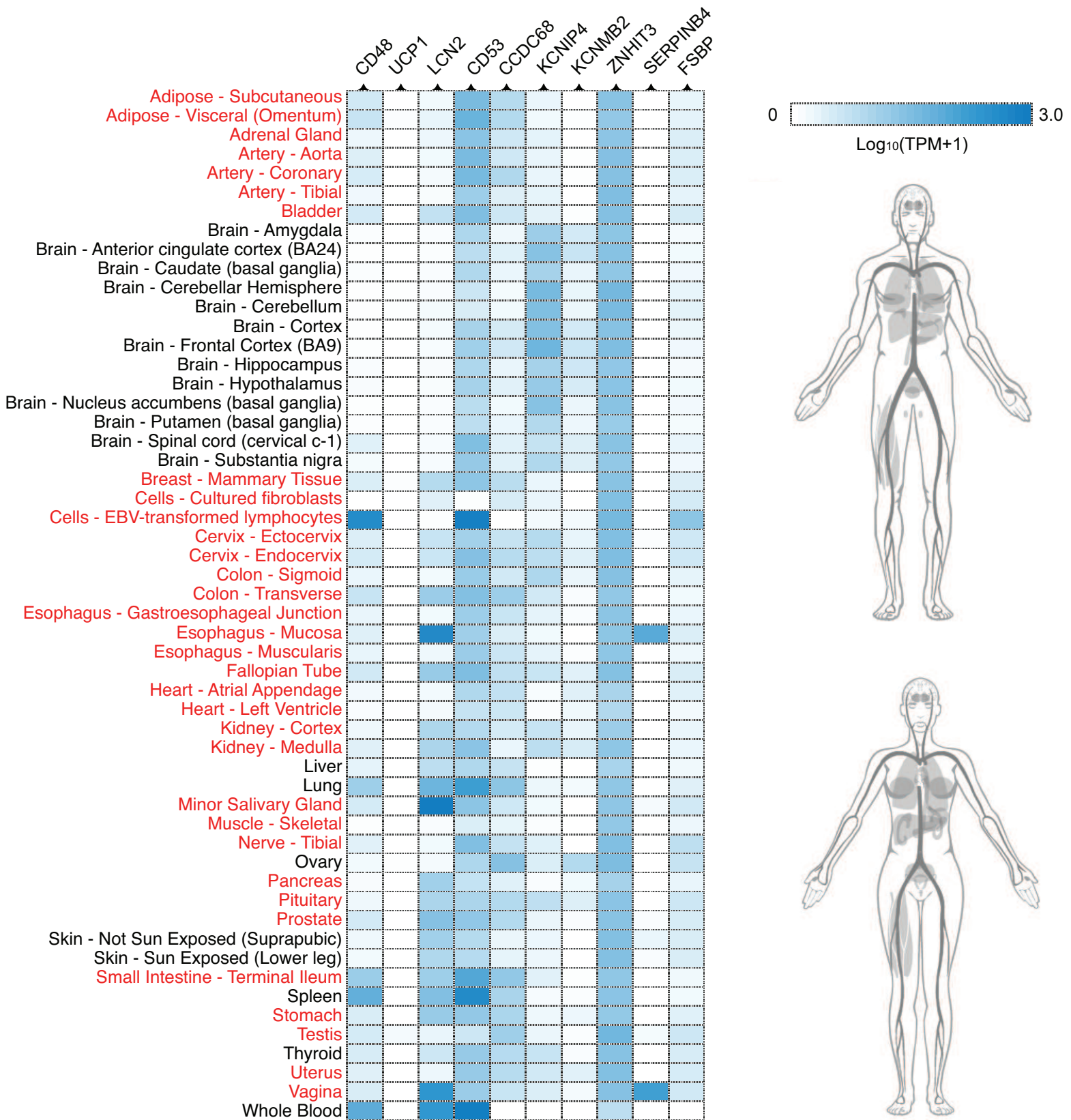


Figure 13

[Click here to access/download;Figure;Figure_13.eps](#)







BEGIN

Initialisation: Balanced dataset $S_0 = \{(1, v_1^0), \dots, (1, v_n^0), (0, v_{n+1}^0) \dots (0, v_{2n}^0)\}$, dimension of the feature vector D_0 , machine learning algorithm A , number of disruptive feature $d_0 = D_0$, and iteration round $i = 0$.

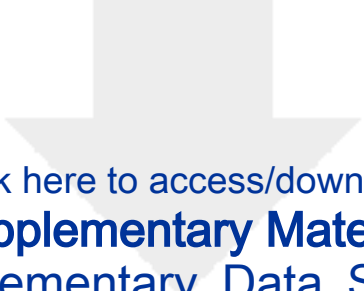
While $d_0 > 0$ (i^{th} iteration):

- 1) Use five-fold cross validation on dataset S_i , prediction $P_i = A(S_i)$;
- 2) Evaluate the P_i with the criterion of AUC;
- 3) Remove one feature from feature vector v^i and generate a temporary dataset T_i ;
- 4) Use five-fold cross validation on dataset T_i , prediction $P'_i = A(T_i)$;
- 5) Evaluate the P'_i with the criterion of AUC;
- 6) Repeat 4) and 5) for the traversal of D_i features;
- 7) Traverse v^i and remove m features helpful to improve AUC of P'_i , $d_i = m$;
- 8) Update dataset $S_{i+1} = \{(1, v_1^{i+1}), \dots, (1, v_n^{i+1}), (0, v_{n+1}^{i+1}) \dots (0, v_{2n}^{i+1})\}$, $D_{i+1} = D_i - m$.

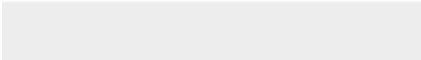

End

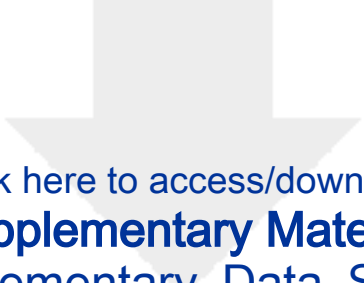
Output: dataset S_{i-1} encoded by D_{i-1} features.

END

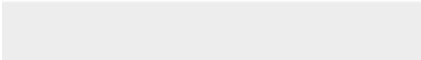



Click here to access/download
Supplementary Material
Supplementary_Data_S1.csv





Click here to access/download
Supplementary Material
Supplementary_Data_S2.csv






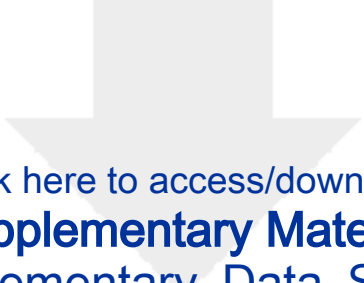
Click here to access/download
Supplementary Material
Supplementary_Data_S3.csv



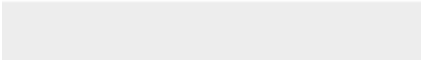



Click here to access/download
Supplementary Material
Supplementary_Data_S4.csv





Click here to access/download
Supplementary Material
Supplementary_Data_S5.csv

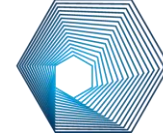




Medical
Research
Council



University
of Glasgow



CVR
Centre for
Virus Research

Editors

GigaScience

16th September 2022

Dear Editors

On behalf of my co-authors please consider the resubmission of our research article entitled ‘Defining the characteristics of interferon-alpha-stimulated human genes: insight from expression data and machine learning’ for consideration in your journal. We believe these revisions will address all the concerns raised by you and the reviewers. We apologise for the misunderstanding with respect to the ROC curve and AUC values. We have added these in the legend for each dataset. We have also added ISGPRES to biotools and registered ISGPRES with scicrunch.com and received the RRID:SCR_022730, however it does not yet appear when browsing the scicrunch dashboard. We have also made a few additional clarifications to the text, in particular, after much discussion, we decided to change the ‘noisy’ feature terminology to poorly performing feature (this was originally called disruptive feature).

We confirm that all authors have read and approved this version of the manuscript.

Yours Sincerely,

A handwritten signature in black ink, appearing to read 'J Hughes'.

Joseph Hughes