

1 **DGRPpool: A web tool leveraging harmonized *Drosophila* Genetic** 2 **Reference Panel phenotyping data for the study of complex traits**

3 Vincent Gardeux^{1,2}, Roel P.J. Bevers^{1,‡}, Fabrice P.A. David^{1,2,3}, Emily Rosschaert^{1,4}, Romain
4 Rochepeau¹, Bart Deplancke¹

5 ¹Laboratory of Systems Biology and Genetics, Institute of Bioengineering, School of Life Sciences, Ecole Polytechnique
6 Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland.

7 ²Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland.

8 ³Bioinformatics Competence Center, EPFL, Switzerland.

9 ⁴KU Leuven, Belgium

10 [‡]Present address: Genomics England Ltd., E14 5AB London, England.

11 **Abstract**

12 Genome-wide association studies have advanced our understanding of complex traits, but
13 studying how a GWAS variant can affect a specific trait in the human population remains
14 challenging due to environmental variability. *Drosophila melanogaster* is in this regard an
15 excellent model organism for studying the relationship between genetic and phenotypic
16 variation due to its simple handling, standardized growth conditions, low cost, and short
17 lifespan. The *Drosophila* Genetic Reference Panel (DGRP) in particular has been a valuable
18 tool for studying complex traits, but proper harmonization and indexing of DGRP
19 phenotyping data is necessary to fully capitalize on this resource. To address this, we
20 created a web tool called **DGRPpool** (dgrpools.epfl.ch), which aggregates phenotyping data of
21 935 phenotypes across 125 DGRP studies in a common environment. DGRPpool enables
22 users to download data and run various tools such as genome-wide association analyses
23 (GWAS) and Phenome-WAS analyses. As a proof-of-concept, DGRPpool was used to study
24 the longevity phenotype and uncovered both established and unexpected correlations with
25 other phenotypes such as locomotor activity, sleep duration, and oxidative stress resistance.
26 DGRPpool has the potential to facilitate new genetic and molecular insights of complex traits
27 in *Drosophila* and serve as a valuable, interactive tool for the scientific community.

28

29

30

31

32

33 Introduction

34 *Drosophila melanogaster* is an excellent model organism for studying genotype-to-
35 phenotype relationships. It is a short-living species and is very easy to maintain in similar
36 laboratory conditions, which limits confounding factors such as the environment. The
37 *Drosophila* Genetic Reference Panel (DGRP) was created in the early 2010s and now
38 consists of 205 inbred lines that are fully sequenced, of which 192 are still available in the
39 Bloomington *Drosophila* Stock Center (<https://bdsc.indiana.edu/>)^{1,2}. The DGRP has proven
40 highly valuable to study the genetic basis of complex traits, as illustrated by the many
41 studies that have used GWAS principles to identify variants that contribute to traits related to
42 morphology, metabolism, behavior, aging, disease susceptibility etc. (**Figure 1A**).
43 Furthermore, since the DGRP lines were inbred for many generations, they are almost fully
44 homozygous, which simplifies the identification of putatively causal alleles and elucidation of
45 implicated molecular mechanisms³. Moreover, the fact that the same lines can be studied by
46 various researchers for diverse traits should leverage these data generation efforts to
47 uncover unexpected correlations between phenotypes or relationships between genetic
48 variants and a wide range of traits.

49 However, there is currently only one major data resource that aims to compile DGRP
50 information, the DGRP2 website (<http://dgrp2.gnets.ncsu.edu/>)^{1,2}. This website hosts the
51 genotyping data, its annotation, and potential covariates, as well as 31 phenotypes from 12
52 studies (**Table 1**). The data is primarily hosted as static files, downloadable from the website,
53 along with limited RNA expression data. In addition, a very important tool, used by the
54 DGRP community, is the possibility for any user to submit their own phenotype files for
55 running a GWAS analysis (corrected with known covariates). This is particularly useful,
56 especially for researchers that do not have the bioinformatics knowledge or capacity to
57 perform these tasks internally. However, the DGRP2 website has not been updated for an
58 extended period as the last referenced paper dates back to 2015, and, except for the GWAS
59 computation, remains thus static. This means that any meta-study, which would aim to
60 aggregate datasets across available phenotypes, would require hours (if not days) of work to
61 transform the data into an appropriate and common format. Moreover, the result of such
62 effort would unlikely become available to the rest of the community, and thus any other
63 group would need to redo this work in order to gather similar information, while the data of
64 other phenotyping studies beyond the 12 available would not be easily accessible.

65 For all these reasons, we decided to create a web application, DGRPpool (dgrp.pool.epfl.ch),
66 that would both act as a repository of DGRP phenotyping datasets and also as an online tool
67 for assisting researchers with some basic systems genetics-inspired analyses. Our goal was

68 to index all existing literature about DGRP phenotyping data —where possible— in order for
69 users to quickly search through the website using simple keywords. We manually associated
70 each study with broad and tailored categories such as “ageing”, “metabolism”, or “olfactory”.
71 We specifically spent important time curating the datasets to avoid any errors or
72 misrepresentations of datasets. To avoid the “maintenance issue” that is common to online
73 tools, and keep the data up to date, we implemented specific curators tools, to help maintain
74 the web application in the future. These tools allow any user to submit a novel dataset, which
75 is then attributed to a curator, in order to manually format and validate all phenotyping data
76 and metadata associated with the study. Importantly, any user can become a curator, as
77 advertised on the main page of the resource, since we strongly believe that such a
78 community-run resource architecture is most optimal to keep a web tool state-of-the-art and
79 allow crowd-based curation work⁴.

80 In addition, we set out to build important tools for the DGRP community such that DGRPpool
81 would not only be a static repository for downloading phenotyping data but could also be
82 used as an interactive data analysis tool. For example, users can correlate phenotypes
83 together, from the same study or across studies. We also implemented an automated GWAS
84 analysis (using PLINK2, and known covariates) which we pre-calculated on all the
85 phenotyping data that are currently available. Using this data, users can simply browse
86 through their genes or variants of interest and directly find related phenotypes. A PheWAS
87 page also allows exploration of each variant’s impact across multiple phenotypes. Moreover,
88 these tools are applicable to user-submitted phenotypes, so that anyone can upload their
89 own phenotypes to search the DGRPpool database for correlated phenotypes or to run
90 GWAS analyses.

91 Our goal is to ensure that DGRP phenotyping data is findable, accessible, interoperable, and
92 reusable (FAIR)⁵ to fully leverage the opportunities that stem from this unique genotyping-
93 phenotyping resource. To this end, we made user access our priority, both for removing the
94 bottleneck of data harmonization, and also to allow for better, more reproducible research.

95 To showcase the potential of our tool in facilitating new biological discoveries, we conducted
96 a proof-of-concept study focusing on the longevity phenotype, a well-studied trait in
97 *Drosophila* research with clear relevance to human longevity⁶. By leveraging the data
98 harmonization and curation efforts in DGRPpool, we identified multiple phenotypes that are
99 significantly associated with longevity across 18 different studies, such as oxidative stress
100 resistance⁷, sleep duration^{8,9}, desiccation survival^{10,11}, and starvation resistance^{10,12,13}.
101 Interestingly, we also observed correlations between shorter lifespan and certain
102 phenotypes, such as locomotor activity¹⁴ and food intake^{15,16}. These results validate prior

103 knowledge and illustrate how our tool can provide novel biological insights with just a few
104 clicks. Therefore, we firmly believe that tools such as DGRPool —which ultimately could
105 become entirely community-driven— are essential not only for catalyzing novel research, but
106 also for leveraging the diversity and richness of existing datasets.

107 **Results**

108 **A thousand phenotypes across 125 studies**

109 To start our data collection, we searched for DGRP studies that reference any phenotyping
110 data and in parallel implemented diverse tools to automatically aggregate these data and
111 their associated metadata from the journals hosting the datasets. However, we quickly
112 realized that it was difficult to automate the entire process. Specifically, the import of
113 phenotyping data proved challenging since i) datasets tended to be hosted in very different
114 formats such as Excel files or PDF, ii) data was stored within the journal's supplementary
115 section, or in external repositories such as Figshare; and iii) the format of the phenotyping
116 data differed from one publication to another. Because of these challenges, we implemented
117 a curation page to manually review, edit, and correct datasets that were automatically
118 aggregated, aiming to prevent errors in the imported datasets. In addition, this allows the
119 curator to add relevant remarks or comments on the study under review, thus providing
120 enhanced context for future analyses of these datasets.

121 In line with the community-resourcing philosophy of DGRPool, we created a specific
122 “curator” role that any logged-in user can claim, again with the underlying rationale of
123 assuring long-term sustainability of our web application. With this role, the user has access
124 to additional functionalities on the DGRPool website, including the modification of any
125 metadata attached to a study (title, authors, description, categories), and the submission or
126 modification of attached phenotypes (see **Supp. Figure S1**). Although this may entail a
127 considerable amount of time, we assert that this approach is the most effective means of
128 furnishing high-quality data. Consistent with this philosophy, we have incorporated a
129 functionality on the homepage which empowers any user to submit a DOI as a
130 recommendation for a study that could be absent from the DGRPool repository. If the DOI is
131 not in the database, it triggers the same automated scripts that were originally used to
132 incorporate the 125 studies. The corresponding study is then created on DGRPool, and its
133 metadata (authors, links, ...) are automatically imported. Once a study has been created,
134 one of three possible labels can be assigned to describe the state of curation of a study: 1)
135 **Submitted** (default), when no curator is yet assigned to the study, 2) **Under curation**, when
136 a curator is assigned, and 3) **Curated** when all phenotyping data and metadata have been

137 curated, and the study received final approval by the curator. At this time, DGRPool hosts
138 125 studies, including 41 that have already been fully curated, 81 still under curation, and 3
139 under a submitted status given that the latter were used to test DGRPool's DOI feature. All
140 metadata of these three studies were correctly imported into DGRPool, but not the
141 associated phenotypes, which is also the case for a portion of the other 122 studies. Indeed,
142 in total, 74 studies have attached phenotyping data; 100% of the curated ones, and only
143 40% of the non-curated ones. Altogether, the total number of studies in DGRPool is currently
144 125, and we expect that this number will continue to grow upon its public release, along with
145 the number of curated studies.

146 Since the curation process is still ongoing, we will be referring to two different datasets in the
147 manuscript: 1) The **full dataset**, comprising **125** studies (independent of "curation" status),
148 and 2) the **curated dataset**, comprising **41** studies that already underwent tedious curation
149 and contributed about 500 phenotypes (see below). Of note, for all tools available on the
150 website, it is possible to run these on either all studies or (as is currently the default), only on
151 the curated studies.

152 For all of the curated studies, we carefully separated the data by sex when information on
153 sex-specific phenotypes was available, or we assigned it as *NA* when flies were sex-mixed,
154 when there was no information on sex, or when the phenotype is inherent to a population
155 (e.g. in the case of non-sexual chromosomal traits, like inversions). We also extracted this
156 information from the phenotyping data itself for the non-curated studies, when available, but
157 when not findable, it was set to *NA*, waiting for a more in-depth curation and careful reading
158 of the paper method's section. Therefore, across all 125 studies, this led to an overall
159 equilibrium between all represented sexes, with slightly more data for females and slightly
160 less unannotated data (**Figure 1B**). However, when focusing only on the 41 curated
161 datasets, the proportion of phenotypes without assigned sex (*NA*) dropped drastically to
162 ~15%. This effect highlights the importance of tedious curation, which typically requires the
163 curator to read through the entire manuscript to understand the utilized experimental
164 protocols to select the appropriate sex, even if this information is not explicitly indicated in
165 the phenotyping data itself.

166 Upon data curation, the assigned curator(s) has to specify a few phenotypic categories for
167 each study, for example, "Metabolism", "Nutrition", or "Ageing" (**Figure 1C**). Since these
168 categories are browsable, it facilitates searching for a set of specific studies or linking the
169 studies together. Interestingly, the top annotated categories are either "Behaviour", "Life
170 History Traits", or "Resistance", which is consistent with historical behavioral and immune
171 studies conducted for *Drosophila* as a model organism¹⁷⁻²¹. The number of phenotypes per

172 study ranges from 1 to 89 (**Figure 1D, Supp Figure S2**), with a median of 5, and a mean of
173 11, revealing that while a low number of phenotypes (usually less than 10) tends to be the
174 norm, some studies aggregate lots of (often similar) phenotypes. An example of the latter is
175 Chaston et al., 2016²² which investigated the impact of microbiota on nutritional traits. The
176 authors studied 76 different microbial taxa, whose effect was quantified independently,
177 generating a high number of phenotypes. Similarly, Dembeck et al., 2015²³ studied cuticular
178 hydrocarbon composition, considering 66 different cuticular components, while Vonesch et
179 al., 2016²⁴ studied organismal size traits, regrouping 28 morphological phenotypes such as
180 wing length or intraocular distance. In total, the 41 curated studies aggregate 312 M + 220 F
181 + 132 NA = 664 sex-specific phenotypes, for a total of 500 unique phenotypes (~60%), while
182 the remaining non-curated studies provide another 57 M + 34 F + 267 NA = 358 sex-specific
183 phenotypes, for a total of 329 unique phenotypes (~40%).

184 **Harmonization and formatting of phenotyping data**

185 DGRP phenotyping data are often available as a supplemental data table, published along
186 with the main paper on the journal's website. Such data can also be stored on external
187 websites such as Figshare and, as already indicated, the corresponding file can be in
188 varying formats (i.e. Excel, text, or PDF), so it is challenging to entirely automate extraction
189 algorithms. Usually, the data are presented in the form of a matrix, with DGRP lines in rows
190 and phenotypes in columns. But sometimes, they can be in a more "exotic" format²⁵,
191 requiring a hands-on approach to format it appropriately. Also, the provided phenotyping
192 data are often not sufficiently self-informative and thus require in-depth reading of the
193 original manuscript to grasp abbreviations or identify the correct measurement units. These
194 are important, in particular, to assure reproducibility, but especially when aggregating
195 multiple studies together such that the scale of the values is similar. In DGRPpool, we
196 therefore created a common matrix format to represent all studies, and we implemented a
197 "Unit" metadata for each phenotype. Then, for each study, we mapped all phenotypes to
198 their appropriate format and units (**Supplement Figure S3**). This part is fully accessible to
199 the curator, who can update or add any phenotype that would be missing, with their
200 corresponding units and meta-data description.

201 Another issue that we faced is that phenotypes are often averaged across multiple individual
202 flies and that the authors only provide these "Summary datasets". This can be problematic in
203 terms of reproducibility, since some figures may show boxplots or distributions of values for
204 each DGRP line, but these plots are not reproducible when only summary data is available
205 (i.e. means or medians). Fortunately, some studies do provide "raw datasets" which contain
206 multiple phenotypic values per DGRP line, often corresponding to replicate flies of the same

207 genotype. These values tend to be of much greater interest since they enable statistical
208 analyses and/or the computation of further summary statistics (not only mean or median, but
209 also the standard error of means or other often non-provided summary values).

210 Finally, for some studies, phenotyping data were not or no longer available from the journal's
211 website²⁶⁻²⁸, which is often the journal's responsibility. However, in all cases, we were able
212 to contact the authors directly to recover the missing datasets.

213 To avoid such issues in the future, we have formulated a couple of good practice guidelines
214 for authors to facilitate and improve upon our and future datasets with the aim of enabling
215 harmonized and reproducible research. These guidelines are detailed in the Discussion
216 section of this manuscript. All curated datasets in DGRPpool are formatted following these
217 guidelines (where possible), and phenotypes can now be easily downloaded in a standard
218 TSV format from a particular study, or from a phenotype page.

219 **How to leverage these datasets by correlating phenotypes**

220 Our formatting and harmonizing of all datasets now enables interesting cross-phenotype
221 analyses to generate new biological insights. One strategy to perform such analyses is to
222 download a summary table that contains all the phenotypes in a common format and that is
223 available from DGRPpool's front page. However, we deemed this still insufficient as a
224 catalyzing resource, which is why we implemented tools to correlate existing and user-
225 submitted phenotypes with all the other phenotypes in DGRPpool (**Supp. Figure S4**).

226 To better understand the structure of these phenotypes, and how they relate together, we
227 also computed a global visualization of the phenotype correlations across all curated studies
228 (**Figure 2A, Supp. Figure S5**). This revealed a clear trend, with phenotypes belonging to the
229 same study (within-study) correlating in general stronger than those from different studies
230 (**Figure 2B, Supp. Figure S6**). This is expected since a given study will typically contain
231 phenotypes that have been acquired for a given research topic, thus they will share
232 similarities. Another potential factor that could explain this similarity is the well-known "batch
233 effect". Indeed, phenotypes acquired in the same environment (same lab, technician,
234 reagents etc.) may sometimes show greater similarity than those acquired across different
235 labs and conditions²⁹. The longevity phenotype however, assessed in at least six of the
236 studies in DGRPpool^{27,30-34} across different laboratories, illustrates that phenotype and its
237 measurements not only exhibits strong correlation across sexes (**Figure 2C**), but are also
238 sufficiently robust between laboratories (**Figure 2D**). This example illustrates both the high
239 robustness of results acquired in the context of DGRP studies (stable genotype, stable

240 environment) and the robustness of the phenotype itself, which highlights its potential high
241 heritability.

242 **Cross-study correlations highlight phenotype relationships**

243 **Figure 2A** also highlights interesting cross-study correlations. For example, we can see a
244 strong correlation between (Vonesch et al, 2016)²⁴ and (Grubbs et al, 2013)³⁵ which is
245 perhaps expected since both studies examine fly morphology traits. The first one measures
246 different organismal size traits such as eye interocular distance, or wing length, while the
247 second studies leg and antenna development from imaginal discs, resulting in measuring
248 phenotypes such as leg and bone length (**Figure 3A**). Similarly, three studies: (MacKay et
249 al, 2012)¹, (Richardson et al, 2012)³⁶ and (Huang et al, 2014)² are expectedly correlated
250 since all three investigate the influence of the *Wolbachia* endosymbiont. Another interesting
251 correlation is between (Chow et al., 2013)³⁷ and (Durham et al., 2014)²⁷ which both studied
252 fecundity and yield a cross-study correlation between remating proportion (Chow et al.,
253 2013)³⁷ vs. mean fecundity (Durham et al., 2014)²⁷ (**Figure 3B**). While potentially
254 conceptually obvious, this correlation suggests that females that are more likely to mate with
255 multiple males tend to also produce a greater number of eggs.

256 These examples were all generated using DGRPool phenotype correlation tools, supporting
257 our notion that it can leverage cross-study comparisons of multiple phenotypes to unveil
258 potentially new interesting phenotype interaction/associations. As a further proof of concept
259 and given society's strong interest in defining "healthy aging" determinants³⁸, we continued
260 investigating the "mean longevity" phenotype from (Arya et al, 2010)³⁰ and we selected 50
261 phenotypes that were significantly correlated with it at 25% FDR threshold (**Figure 3C**). The
262 hierarchical clustering clearly separated the phenotypes into three clusters: longevity-like
263 phenotypes (strongly correlated together), other longevity-associated phenotypes (correlated
264 with longevity), and phenotypes that seem antagonistic to longevity (anti-correlated
265 phenotypes). Among the phenotypes that positively correlated with longevity, some may be
266 expected such as starvation resistance^{10,12,13} and oxidative stress resistance⁷ but some are
267 less intuitive such as desiccation survival^{10,11}, certain cuticular components of the
268 epicuticle³⁹, and sleep duration^{8,9}, whose relationship to longevity is complex and still not
269 fully understood⁴⁰. Although we cannot exclude spurious correlations, some of these more
270 surprising correlations appear biologically highly interesting, illustrating the capacity of
271 DGRPool to unveil new research avenues that seem worth exploring in greater molecular
272 detail. Also of interest is the group of often unexpected phenotypes that significantly anti-
273 correlates with longevity. These include locomotor activity¹⁴, some other cuticular
274 components of the epicuticle⁴¹, and food intake^{15,16}, suggesting that higher locomotor activity

275 or food intake is linked to reduced longevity. Whether these are direct or indirect links
276 remains unanswered, but appears worthy for a more in-depth scrutiny that is beyond the
277 scope of this paper.

278 Inversely, our analyses also revealed that some expected phenotype correlations could not
279 be detected. For example, in the context of metabolic energy expenditure⁴², it might seem
280 intuitive that higher activity⁴³ would lead to greater food intake⁴⁴. However, we did not
281 observe such a correlation. Similarly, higher activity levels may reflect increased mating
282 behaviour³⁷, but this was also not observed. These are just a few examples of several cases
283 where expected correlations did not materialize, collectively signifying that the genetic
284 architecture underlying such traits appears inherently complex.

285 These proof-of-concept examples demonstrate in our opinion the utility of the DGRP lines
286 and by extension DGRPpool to serve as powerful tools that will facilitate the identification of
287 non-intuitive phenotype correlations and their underlying molecular basis as well as the
288 discovery of putative genotype to phenotype relationships, as detailed below.

289 **From phenotypes to associated genotypes**

290 The goal of most DGRP phenotyping studies is to eventually be able to link the phenotypes
291 to potentially causal variants or sets of variants⁴⁵. In response, tools like DGRP2 GWAS
292 (<http://dgrp2.gnets.ncsu.edu/>)^{1,2} have been put in place to accommodate geno-phenotype
293 relationship analyses.

294 With the goal of providing an integrative analytical environment, we therefore also
295 implemented GWAS tools within DGRPpool, aiming to assist researchers with performing
296 GWAS analyses and interpreting the respective output. Specifically, we precalculated GWAS
297 analyses using PLINK2 on every existing phenotype in DGRPpool (see **Methods**), thereby
298 considering all ~4M available DGRP variants while correcting for six main covariates
299 (*Wolbachia* status, and five major insertions)². Consequently, users can browse the GWAS
300 results from any phenotype page on DGRPpool (**Supp. Figure S7**). These comprise a
301 QQplot, for assessing the validity of the results, or potentially over-estimated p-values, and a
302 Manhattan plot, for visualizing the significant loci across the *D. melanogaster* genome. It also
303 displays a table with the top 1000 associated variants and allows the user to download the
304 table of all significant hits, at a p-value<0.01 threshold. The tool further runs an ANOVA
305 between the phenotype and the six main covariates to uncover potential confounder effects
306 (prior correction), which is displayed as a “warning” table to inform the user about potential
307 associations of the phenotype and any of the covariates. The interface also allows plotting
308 an independent boxplot for each variant to visualize the effect of each allele on the

309 phenotype. Importantly, for each variant, we also implemented a PheWAS button to visualize
310 the effect of a particular variant over all existing phenotypes in DGRP. We also annotated
311 all the variants for impact (non-synonymous effects, stop-codon gain, etc.) and for potential
312 regulatory effect (transcription factor binding motif disruption), which should assist
313 researchers with prioritizing the variants in terms of potential consequences. For all of these
314 variants, we also provide links to their description in Flybase⁴.

315 As mentioned, these GWAS results are available for each existing phenotype in DGRP,
316 directly from the phenotype's page. But users can also submit their own phenotype files
317 (through the 'Tool' menu in the header), and visualize the same information for their own
318 phenotypes. The GWAS analysis runs in the backend and takes about 1-2 minutes before
319 displaying the results. This is implemented using a queuing system which prevents
320 overloading the server in case of a peak of users or requests.

321 After having run GWAS on all phenotypes in DGRP, we observed the distribution of the
322 number of significant variants per phenotype at $p \leq 1 \times 10^{-5}$ threshold, which is an often used
323 arbitrary threshold for GWAS analyses in DGRP studies (**Figure 4A**). This threshold yields
324 on average 382 significant hits per tested phenotype, which is skewed due to some
325 phenotypes leveraging lots of results (median = 38). Conveniently, this threshold seems
326 sufficient for avoiding an over-abundant number of false positives, as is clearly visible from
327 other, less stringent, thresholds (**Supp. Figure S8**). Another very often used threshold, is the
328 Bonferroni one, which is much more stringent and varies from $p \leq 1.126 \times 10^{-8}$ (if considering
329 all 4M variants) to $p \leq 2.64 \times 10^{-8}$ (if removing variants with low MAF or high number of
330 missing values). In our results, the Bonferroni threshold ($p \leq 2.64 \times 10^{-8}$) yielded 73
331 significant hits on average (median = 0, **Supp. Figure S8**) which could be limiting for many
332 studies as it may mask potentially interesting variants that, while minimally contributing on an
333 individual basis, may collectively point to implicated pathways or biological processes⁴⁶.
334 Thus, while choosing an optimal threshold is in general challenging, our results indicate that
335 any threshold below 1×10^{-5} is reasonable given that at this threshold, the p -values appear
336 not over-estimated, as observed on the respective QQplots. We also verified if any variant is
337 over-selected across all phenotypes to uncover a possible bias in our studies (**Figure 4B**),
338 but we did not find such variants, even at different thresholding values (data not shown).

339 As a proof-of-concept and a validation of our approach, we compared our results with a
340 randomly selected study that identified several variants associated with survival to azinphos-

341 methyl at different doses (0.25, 0.5, 1, and 2 $\mu\text{g/ml}$)²⁶. Of note, this study is available in
342 DGRP Pool under <https://dgrpools.epfl.ch/studies/3>. In particular, this study showed that
343 survival to azinphos-methyl is highly variable among DGRP lines, even at a “low” 0.25 $\mu\text{g/ml}$
344 dose. Importantly, the results of this study are reproduced in DGRP Pool as can be observed
345 on the respective phenotype’s page (<https://dgrpools.epfl.ch/phenotypes/20>, **Figure 4C**). For
346 example, DGRP Pool’s GWAS results are very similar to those of the study
347 (https://dgrpools.epfl.ch/phenotypes/20/gwas_analysis, **Figure 4D**) and show a strong
348 association at a 2R locus. Interestingly, the top variant we found, 2R:8072884 ($p = 1.966 \times$
349 10^{-26}), a 509bp insertion polymorphism, is the *Accord* LTR insertion. It is annotated as
350 located upstream of the *Cyp6g1* gene and has a high likelihood to be the main causal
351 gene^{47,48}. As described in the author’s Ph.D. thesis⁴⁹, the minor allele at this variant—which
352 corresponds to NOT having the insertion—correctly genotypes eight out of nine susceptible
353 DGRP lines that are homozygous for the ancestral *Cyp6g1*^M arrangement at this locus
354 (DGRP lines 091, 486, 642, 776, 802, 821, **843**, 852, and 857). The presence of the *Accord*
355 LTR insertion is associated with increased resistance to organophosphates, suggesting that
356 derived alleles of *Cyp6g1* confer organophosphate resistance in the DGRP (**Figure 4E**).

357 These results show that DGRP Pool is able to accurately reproduce results from existing
358 studies, and that new biological findings can be leveraged from its interactive results and
359 plots. Revisiting the same organophosphate study²⁶, the PheWAS page present in the
360 GWAS results shows that this top variant is not only significant at other doses, but that it is
361 also significant in the context of other studies, in particular one study on cuticular
362 hydrocarbon composition²³, and another study investigating *Drosophila* microbiota²². This
363 could help with fine-tuning putative causal variants, but also with uncovering potential
364 associations between certain phenotypes that in turn could enable studies aimed at
365 providing underlying genetic and molecular mechanisms.

366 **Extreme phenotypes**

367 After having collected and harmonized thousands of DGRP phenotypes, we investigated if
368 we could identify outliers amongst DGRP lines that would potentially bias phenotypic
369 associations. Indeed, if a particular DGRP line is repeatedly ranked in the extreme of all
370 phenotypes, it could be that there are unknown cofactors that make the line “weaker” in
371 general, or inversely. Although it is difficult to judge what phenotype is particularly
372 advantageous or disadvantageous due to the presence of potential trade-offs^{50,51}, we can
373 determine how often a DGRP line is in the top or bottom 15% of a given phenotype. By
374 focusing on phenotypes that are likely impacting overall viability, we ranked DGRP lines for
375 each associated phenotype. Upon ranking the DGRP lines, we calculated whether the rank

376 falls within the top or bottom 15% performers of the phenotype. We then assessed for each
377 DGRP line how often they are ‘extreme’ and divided this by the total number of phenotypes
378 in which the DGRP line has been included to obtain a “fraction of extremeness” (FoE).
379 Finally, we filtered for lines which had at least 50 phenotypic measures available to ensure
380 that our values were not driven by a low number of observations (**Figure 5A**). Looking
381 broadly, we observed a mild correlation of fraction of extremeness (FoE) across the sexes
382 (**Figure 5B**, Spearman’s $\rho = 0.3514$, $p < 1.57 \times 10^{-5}$). While this may indicate that
383 extremeness is a population-wide feature, it is not sufficiently profound to conclude that
384 DGRP lines are generally extreme in both sexes, which may only be the case for specific
385 DGRP lines.

386 Upon considering individual DGRP lines, we can observe to what extent they are extreme for
387 each individual phenotype. In **Figure 5C**, we show the most extreme and “moderate” (i.e.
388 least distinctive) DGRP lines for each sex using an adjusted fraction of extremeness for
389 plotting purposes in which lower scores represent DGRP lines with a high fraction of
390 extremeness. While females of DGRP_879 and males of DGRP_783 tend to be extreme in
391 some cases, for the majority of phenotypes they are considered moderate. Conversely,
392 females of DGRP_757 and males of DGRP_352 are more likely to be labeled as extreme.

393 These examples only represent extremeness for individual DGRP lines of a given sex,
394 however, their counterpart may not be as extreme or moderate. We therefore also looked for
395 DGRP lines which can be considered extreme in both females and males, and are
396 potentially more extreme on a population-wide basis. **Figure 5D** describes such populations
397 where the overall fraction of extremeness between males and females differed on average at
398 most 0.05. In these cases, DGRP_852 and DGRP_042 are more likely to be extreme across
399 sexes, which may be attributed to at least two factors. First, this may indicate that the
400 population is generally not healthy if they consistently display a low lifespan, or second, and
401 conversely, well-documented trade-offs of life history traits such as lifespan vs fecundity may
402 be strongly at play here. The former does not however seem to be the case, as shown in
403 **Figure 5E**. Both DGRP_852 and DGRP_042 generally display lifespan values around the
404 mean lifespan of all DGRP lines, suggesting that they are more likely extreme for other
405 phenotypes and are thus not by definition weak lines. However, DGRP_757 and DGRP_765
406 consistently display lower longevity in lifespan studies. These lines may therefore on the one
407 hand be of particular interest for those studying life history traits in an evolutionary context,
408 even though we did not observe strong lifespan and fecundity trade-offs across our
409 phenotype dataset. On the other hand though, it may be advisable not to include DGRP_757

410 and DGRP_765 when studying the genetic basis of these complex traits as their outlier
411 status may not reflect common genetic principles.

412 **Discussion**

413 There are many studies across organisms where collated phenotyping data has led to novel
414 insights^{52,53}. Even though the *Drosophila* Genetic Reference Panel was formally released
415 more than ten years ago, the resulting phenotype data of over 100 studies has so far not
416 been combined into a single accessible resource. We anticipate that providing wider access
417 to this data, as driven by FAIR principles⁵, will therefore facilitate our general understanding
418 of the relationship between genotypes and phenotypes.

419 We have previously shown that using a subset of this resource effectively enabled us to
420 establish a relationship between mitochondrial haplotypes and feeding behavior which we
421 experimentally validated⁵⁴. Next to our own study, other studies have used a similar
422 approach and compared their results to already published phenotypes. For example, Wang
423 et al.⁵⁵ studied the resistance and tolerance of DGRP flies to the fungal pathogen
424 *Metarhizium anisopliae* (Ma549) and found that the host's defense to Ma549 was correlated
425 with its defense to the bacterium *Pseudomonas aeruginosa* (Pa14). But they also compared
426 this result to several previously published DGRP phenotypes including oxidative stress
427 sensitivity⁵⁶, aggression⁵⁷, nutritional scores⁵⁸, sleep indices⁴³, and others. Similarly, Zwarts
428 et al.⁵⁹ studied the size of the cerebral cortex and the mushroom bodies (MB). They showed
429 that these phenotypes were correlated with phenotypes from other studies like aggression⁶⁰
430 and sleep⁴³. Therefore, we believe that DGRPpool will either aid with validating the findings of
431 a given study (i.e. higher bacterial resistance linked to overall resistance phenotypes) or by
432 placing a study's phenotype data into a wider context (for example, linking brain size to
433 behavioral phenotypes).

434 Moreover, having access to multiple studies studying similar phenotypes can also be of help
435 for meta-analyses and increased statistical power. In the case of longevity for example, there
436 are multiple studies that aggregated this phenotype, across similar or complementary DGRP
437 lines. Therefore, one could conduct a meta-GWAS analysis⁶¹ by leveraging the replicates or
438 combining the different lines into a single dataset. This tends to be a challenging process
439 given the need for data harmonization and curation, which is exactly what we aimed to
440 address by establishing with DGRPpool. Of course, since similar DGRP lines across
441 laboratories still have the same genotype, they should not be treated as biological replicates,
442 but phenotypes could be averaged across similar lines, which would reduce hidden
443 covariates such as laboratory adaptation or batch effects. Moreover, complementary lines

444 can be used to enhance power and potentially find more small-effect associations. Indeed,
445 researchers are increasingly advocating for collaboration and joining efforts to combine
446 resources⁶² to enable more accurate, and reproducible results.

447 Our data collection and harmonization efforts have already enabled us to conduct some
448 interesting cross-study analyses, including an investigation into the presence of biases
449 stemming from outlier DGRP lines. Our "extremeness" analysis revealed that caution is
450 warranted when selecting DGRP lines for specific studies, because, while some DGRP lines
451 may be situated at the outer edge of the phenotypic spectrum by chance, DGRP_757 and
452 DGRP_765 generally display lower lifespans in longevity studies. It is important to note that
453 a shorter lifespan does not necessarily imply lower viability, as populations can still be
454 propagated healthily. However, a shorter lifespan may also result from an impaired
455 development⁶³ or developmental environment, which may confound the study of healthy
456 aging⁶⁴. Consequently, researchers should consider excluding these extreme lines from their
457 experimental designs to prevent loss of power or potential covariate biases.

458 Furthermore, and beyond our current focus on DGRP lines, we may in the future also
459 consider adding standard *D. melanogaster* lines such as w1118, YWB, YWN or ORB to
460 DGRP. This is because such lines have often been included as controls in DGRP
461 studies³⁴, and for most of these, genomic information is also available.

462 Finally, in order to sustain the value of the DGRP as a resource and to promote more
463 findings, we provide the following guidelines for future DGRP phenotyping studies:

- 464 ● When available, report the raw datasets with values per fly. Optionally, but only in
465 addition, the summary datasets can be provided, with values averaged across flies.
- 466 ● Provide the data as a separate Excel or text file (TSV/CSV) in the form of a matrix, with
467 DGRP lines in rows and phenotypes in columns. Avoid reporting the values in the form of
468 a PDF or an image, because it complicates data extraction afterward.
- 469 ● Clearly define the abbreviations in the tables and the units used for all phenotypes, so
470 that the phenotyping dataset is self-explanatory and does not require an extended search
471 in the main manuscript.
- 472 ● Report all DGRP lines in the first column of the phenotyping file, and the corresponding
473 sex in the second column (M, F, or NA), before all phenotypes. Be careful to use the
474 same format for all DGRP lines (e.g. DGRP_XXX).
- 475 ● Pick a common format for all NA values. Whether reporting NA, or as an empty cell. But
476 avoid mixing different formats.

477 In conclusion, we propose that DGRPool has two primary purposes within the *Drosophila*
478 community and beyond. First, it can be used to evaluate potential associations between
479 phenotypes and contribute to understanding the genetic architecture underlying complex
480 traits. Second, it can serve as a catalyst for further research and inform broader validation
481 experiments, as exemplified in our previous work⁵⁴. In the latter study, the validation of our
482 hypothesis would not have been feasible without a harmonized dataset of phenotype data,
483 as the connection between mitochondrial haplotypes and food intake would have remained
484 theoretical. To maximize the benefits of DGRPool, it should therefore remain subject to all
485 FAIR principles, which unfortunately are still too often only implemented in terms of "open"
486 and "sharing." In other words, when large amounts of data are made publicly available
487 without systematic curation or homogenization, data interoperability and reproducibility can
488 be highly problematic. DGRPool is in this regard a crucial initial step towards making DGRP
489 phenotyping data widely accessible and usable for the entire *Drosophila* research
490 community.

491 **Methods**

492 **Data availability**

493 All phenotyping data aggregated in DGRPool can be downloaded in a common format on
494 each phenotype page. In the "Download" section on the front page, we also provide four .tsv
495 files containing 1) All studies and their metadata (authors, citation, ...), 2) All phenotypes and
496 their metadata (name, description, unit, ...), 3) All DGRP lines and their metadata (name,
497 bloomington accession, ...), and 4) a global file with all numerical phenotypes across all
498 studies, formatted following our recommendations.

499

500 All codes used to produce the figures of this manuscript are also available on our GitHub:

501 <https://github.com/DeplanckeLab/DGRPool>

502 **Web application**

503 The DGRPool web application is hosted on a virtual machine at EPFL. All compute-intensive
504 calculations (i.e. GWAS) are performed on an HPC within EPFL and results are then moved
505 to the virtual machine's local storage. The back-end is implemented with Ruby-on-Rails
506 (RoR) 7 and all data is stored in a PostgreSQL relational database. The front-end uses
507 different JavaScript libraries and is set to enable interactive usage. For instance, the
508 application implements bootstrap tooltips to display HTML texts within tooltips, plotly.js
509 v.2.16.1²⁹ to generate the scatter plots, bar plots and box plots, using *scattergl*, *bar* and *box*

510 modes respectively, or JQuery autocomplete for phenotype search combined with a SOLR
511 search engine running on the server side (used for the phenotype comparison tool).

512 **Semi-automated referencing of studies and/or phenotypes**

513 To submit a new study, any user can submit a DOI from the front page. Then, all metadata
514 associated with this study (authors, journal, date, ...) are automatically imported from the
515 Crossref⁶⁵ API. When the study is created, it acquires the “Submitted” state, and
516 administrators are notified. Then, a curator is assigned to the study and needs to manually
517 verify all information. A specific curator page allows him/her to 1) edit the metadata, 2) edit
518 the categories associated with the study, or 3) add/remove/modify the phenotyping data and
519 edit their names/types/units.

520 Identifiers from GEO⁶⁶, ArrayExpress⁶⁷, or the Sequence Read Archive (SRA)⁶⁸ can be
521 associated manually with any study, for example for referencing additional gene expression
522 data that would be published along with the phenotyping data.

523 **Phenotypes correlated with longevity**

524 We computed the correlation of the “mean longevity” phenotype from (Arya et al, 2010)³⁰
525 and selected 50 phenotypes that were significantly correlated with it using a 25% FDR
526 threshold. For this, we used the phenotype correlation tools available in DGRPpool (result list
527 available at https://dgrpool.epfl.ch/phenotypes/1315/compute_correlation) which makes our
528 results reproducible and freely accessible, following the FAIR principles.

529 **GWAS**

530 GWAS analyses (whether pre-calculated, or using the web tool) use Plink2 v2.00a3LM (1 Jul
531 2021). It runs on all available variants in the DGRP database which is using the dm3
532 assembly (4'438'427 variants: 3'963'420 SNPs, 293'363 deletions, 169'053 insertions and
533 12'591 MNPs) with options “--glm --geno 0.2 --maf 0.05”. We corrected the model for six
534 main covariates (*Wolbachia* status, and 5 major insertions) that were described in ² and also
535 used on the DGRP2 website. Of note, these covariates are phenotypes, and thus are also
536 available as a separate, browsable study on DGRPpool (<https://dgrpool.epfl.ch/studies/17>).

537 **Extremeness**

538 Fraction of extremeness was calculated for each phenotypic spectrum separately by ranking
539 the values with ties being assigned the minimum rank. We then calculated a cut-off to assign
540 ranks in the bottom or upper 15% of a phenotypic range. This rank cut-off was further
541 rounded up to be more inclusive on either end (i.e. if the cut-off was 1.2 or 1.8, the cut-off
542 would become 2). Phenotypes equal or lower than the cut-off were assigned -1, whereas
543 phenotypes equal to the max rank minus the cutoff or higher were assigned 1. Remaining
544 phenotypic values were assigned 0. DGRP lines with phenotypic values of either -1 or 1
545 were then considered extreme for a given phenotype.

546 To calculate the overall fraction of extremeness for each DGRP line, we counted the number
547 of times a DGRP line was assigned -1 or 1 and divided this by the total number of
548 phenotypes available for that particular DGRP line. For most of our analyses, we only
549 included DGRP lines for which at least 50 phenotypes were available unless stated
550 otherwise.

551

552 The adjusted fraction of extremeness was calculated by dividing the phenotypic ranking by
553 the max rank of a given phenotype. Values were adjusted with 1 minus the value if the value
554 was above 0.5 (e.g. if $x = 0.91$, the adjusted value is $1 - 0.91 = 0.09$). Only adjusted fraction of
555 extremeness values below 0.15 are therefore considered extreme. As no rounding was
556 performed in this case, it is possible for DGRPs to be assigned -1 and labeled as extreme,
557 even though the DGRP line may have a value of 0.167. Further analysis shows that this
558 'violation' only takes place for 1.1% (417 out 36,753) of the observations. At a *per* DGRP
559 view, this would amount to less than 1 per 50 phenotypes, the cut-off for the number of
560 phenotypes which a line needs to adhere to in order to be included in our analysis.

561 **Acknowledgments**

562 The authors gratefully acknowledge the help and suggestions from Nathan M. Fiorellino and
563 Jasper Deplancke in the early stages of the development of this web tool. This work was
564 funded by the Ecole Polytechnique Fédérale de Lausanne (EPFL) and SNSF Project Grant
565 (#310030_197082) to BD.

566 Author contributions

567 RB and BD initiated the project. VG, RB and BD wrote the article. RR implemented the
568 automatic pipeline to retrieve phenotypic data from articles. VG and ER curated the studies.
569 FPAD designed and implemented the web application and its database. FPAD designed and
570 set up the unified format to represent phenotype data. FPAD and VG implemented the
571 different tools (GWAS, PheWAS, Correlation). VG tested the web application extensively.
572 VG and RB performed supporting analyses (e.g. GWAS, extremeness analysis).

573 Competing interests

574 The authors declare that they have no conflict of interest.

575 References

- 576 1. Mackay, T. F. C. *et al.* The *Drosophila melanogaster* Genetic Reference Panel. *Nature*
577 **482**, 173–178 (2012).
- 578 2. Huang, W. *et al.* Natural variation in genome architecture among 205 *Drosophila*
579 *melanogaster* Genetic Reference Panel lines. *Genome Res.* **24**, 1193–1208 (2014).
- 580 3. Bou Sleiman, M. S. *et al.* Genetic, molecular and physiological basis of variation in
581 *Drosophila* gut immunocompetence. *Nat. Commun.* **6**, 7829 (2015).
- 582 4. Gramates, L. S. *et al.* FlyBase: a guided tour of highlighted features. *Genetics* **220**,
583 iyac035 (2022).
- 584 5. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and
585 stewardship. *Sci. Data* **3**, 160018 (2016).
- 586 6. Piper, M. D. W. & Partridge, L. *Drosophila* as a model for ageing. *Biochim. Biophys. Acta*
587 *BBA - Mol. Basis Dis.* **1864**, 2707–2717 (2018).
- 588 7. Finkel, T. & Holbrook, N. J. Oxidants, oxidative stress and the biology of ageing. *Nature*
589 **408**, 239–247 (2000).
- 590 8. Bushey, D., Hughes, K. A., Tononi, G. & Cirelli, C. Sleep, aging, and lifespan in
591 *Drosophila*. *BMC Neurosci.* **11**, 56 (2010).
- 592 9. Thompson, J. B., Su, O. O., Yang, N. & Bauer, J. H. Sleep-length differences are
593 associated with altered longevity in the fruit fly *Drosophila melanogaster*. *Biol. Open* **9**,
594 bio054361 (2020).
- 595 10. Rion, S. & Kawecki, T. J. Evolutionary biology of starvation resistance: what we have
596 learned from *Drosophila*: Starvation resistance in *Drosophila*. *J. Evol. Biol.* **20**, 1655–1664
597 (2007).
- 598 11. Hoffmann, A. A. & Harshman, L. G. Desiccation and starvation resistance in
599 *Drosophila*: patterns of variation at the species, population and intrapopulation levels.
600 *Heredity* **83**, 637–643 (1999).
- 601 12. Chippindale, A. K., Chu, T. J. F. & Rose, M. R. Complex Trade-Offs and the
602 Evolution of Starvation Resistance in *Drosophila melanogaster*. *Evolution* **50**, 753 (1996).
- 603 13. Jang, T. & Lee, K. P. The genetic basis for mating-induced sex differences in
604 starvation resistance in *Drosophila melanogaster*. *J. Insect Physiol.* **82**, 56–65 (2015).
- 605 14. Magwere, T. *et al.* Flight Activity, Mortality Rates, and Lipoxidative Damage in
606 *Drosophila*. *J. Gerontol. Ser. A* **61**, 136–145 (2006).
- 607 15. Lee, K. P. *et al.* Lifespan and reproduction in *Drosophila*: New insights from
608 nutritional geometry. *Proc. Natl. Acad. Sci.* **105**, 2498–2503 (2008).
- 609 16. Piper, M. D. W. & Partridge, L. Dietary Restriction in *Drosophila*: Delayed Aging or
610 Experimental Artefact? *PLoS Genet.* **3**, e57 (2007).

- 611 17. Arch, M., Vidal, M., Koiffman, R., Melkie, S. T. & Cardona, P.-J. *Drosophila*
612 melanogaster as a model to study innate immune memory. *Front. Microbiol.* **13**, 991678
613 (2022).
- 614 18. Dissel, S. *Drosophila* as a Model to Study the Relationship Between Sleep, Plasticity,
615 and Memory. *Front. Physiol.* **11**, 533 (2020).
- 616 19. Flatt, T. Life-History Evolution and the Genetics of Fitness Components in *Drosophila*
617 melanogaster. *Genetics* **214**, 3–48 (2020).
- 618 20. Harnish, J. M., Link, N. & Yamamoto, S. *Drosophila* as a Model for Infectious
619 Diseases. *Int. J. Mol. Sci.* **22**, 2724 (2021).
- 620 21. O’Kane, C. J. *Drosophila* as a Model Organism for the Study of Neuropsychiatric
621 Disorders. in *Molecular and Functional Models in Neuropsychiatry* (ed. Hagan, J. J.) vol. 7
622 37–60 (Springer Berlin Heidelberg, 2011).
- 623 22. Chaston, J. M., Dobson, A. J., Newell, P. D. & Douglas, A. E. Host Genetic Control of
624 the Microbiota Mediates the *Drosophila* Nutritional Phenotype. *Appl. Environ. Microbiol.*
625 **82**, 671–679 (2016).
- 626 23. Dembeck, L. M. *et al.* Genetic architecture of natural variation in cuticular
627 hydrocarbon composition in *Drosophila melanogaster*. *eLife* **4**, e09861 (2015).
- 628 24. Vonesch, S. C., Lamparter, D., Mackay, T. F. C., Bergmann, S. & Hafen, E. Genome-
629 Wide Analysis Reveals Novel Regulators of Growth in *Drosophila melanogaster*. *PLoS*
630 *Genet.* **12**, e1005616 (2016).
- 631 25. Hope, K. A. *et al.* The *Drosophila* Gene Sulfateless Modulates Autism-Like
632 Behaviors. *Front. Genet.* **10**, 574 (2019).
- 633 26. Battlay, P., Schmidt, J. M., Fournier-Level, A. & Robin, C. Genomic and
634 Transcriptomic Associations Identify a New Insecticide Resistance Phenotype for the
635 Selective Sweep at the *Cyp6g1* Locus of *Drosophila melanogaster*. *G3*
636 *GenesGenomesGenetics* **6**, 2573–2581 (2016).
- 637 27. Durham, M. F., Magwire, M. M., Stone, E. A. & Leips, J. Genome-wide analysis in
638 *Drosophila* reveals age-specific effects of SNPs on fitness traits. *Nat. Commun.* **5**, 4338
639 (2014).
- 640 28. Najarro, M. A., Hackett, J. L. & Macdonald, S. J. Loci Contributing to Boric Acid
641 Toxicity in Two Reference Populations of *Drosophila melanogaster*. *G3*
642 *GenesGenomesGenetics* **7**, 1631–1641 (2017).
- 643 29. Ackermann, M. *et al.* Effects of assay conditions in life history experiments with
644 *Drosophila melanogaster*: Assay environment in life history experiments. *J. Evol. Biol.* **14**,
645 199–209 (2001).
- 646 30. Arya, G. H. *et al.* Natural Variation, Functional Pleiotropy and Transcriptional
647 Contexts of *Odorant Binding Protein* Genes in *Drosophila melanogaster*. *Genetics* **186**,
648 1475–1485 (2010).
- 649 31. Huang, W. *et al.* Context-dependent genetic architecture of *Drosophila* life span.
650 *PLoS Biol.* **18**, e3000645 (2020).
- 651 32. Ivanov, D. K. *et al.* Longevity GWAS Using the *Drosophila* Genetic Reference Panel.
652 *J. Gerontol. A. Biol. Sci. Med. Sci.* **70**, 1470–1478 (2015).
- 653 33. Zhao, X. *et al.* The metabolome as a biomarker of aging in *Drosophila melanogaster*.
654 *Aging Cell* **21**, (2022).
- 655 34. Hoffman, J. M., Dudeck, S. K., Patterson, H. K. & Austad, S. N. Sex, mating and
656 repeatability of *Drosophila melanogaster* longevity. *R. Soc. Open Sci.* **8**, 210273 (2021).
- 657 35. Grubbs, N. *et al.* New Components of *Drosophila* Leg Development Identified
658 through Genome Wide Association Studies. *PLoS ONE* **8**, e60261 (2013).
- 659 36. Richardson, M. F. *et al.* Population Genomics of the *Wolbachia* Endosymbiont in
660 *Drosophila melanogaster*. *PLoS Genet.* **8**, e1003129 (2012).
- 661 37. Chow, C. Y., Wolfner, M. F. & Clark, A. G. Large Neurological Component to Genetic
662 Differences Underlying Biased Sperm Use in *Drosophila*. *Genetics* **193**, 177–185 (2013).
- 663 38. Friedman, S. M. Lifestyle (Medicine) and Healthy Aging. *Clin. Geriatr. Med.* **36**, 645–
664 653 (2020).

- 665 39. Wang, Z. *et al.* Desiccation resistance differences in *Drosophila* species can be
666 largely explained by variations in cuticular hydrocarbons. *eLife* **11**, e80859 (2022).
- 667 40. Consensus Conference Panel *et al.* Joint Consensus Statement of the American
668 Academy of Sleep Medicine and Sleep Research Society on the Recommended Amount
669 of Sleep for a Healthy Adult: Methodology and Discussion. *Sleep* **38**, 1161–1183 (2015).
- 670 41. Nghiem, D., Gibbs, A. G., Rose, M. R. & Bradley, T. J. Postponed aging and
671 desiccation resistance in *Drosophila melanogaster*. *Exp. Gerontol.* **35**, 957–969 (2000).
- 672 42. Chatterjee, N. & Perrimon, N. What fuels the fly: Energy metabolism in *Drosophila*
673 and its application to the study of obesity and diabetes. *Sci. Adv.* **7**, eabg4336 (2021).
- 674 43. Harbison, S. T., McCoy, L. J. & Mackay, T. F. Genome-wide association study of
675 sleep in *Drosophila melanogaster*. *BMC Genomics* **14**, 281 (2013).
- 676 44. Garlapow, M. E., Huang, W., Yarboro, M. T., Peterson, K. R. & Mackay, T. F. C.
677 Quantitative Genetics of Food Intake in *Drosophila melanogaster*. *PLOS ONE* **10**,
678 e0138129 (2015).
- 679 45. Mackay, T. F. C. & Huang, W. Charting the genotype–phenotype map: lessons from
680 the *Drosophila melanogaster* Genetic Reference Panel. *WIREs Dev. Biol.* **7**, (2018).
- 681 46. Uffelmann, E. *et al.* Genome-wide association studies. *Nat. Rev. Methods Primer* **1**,
682 59 (2021).
- 683 47. Daborn, P. J. *et al.* A single P450 allele associated with insecticide resistance in
684 *Drosophila*. *Science* **297**, 2253–2256 (2002).
- 685 48. Schmidt, J. M. *et al.* Copy number variation and transposable elements feature in
686 recent, ongoing adaptation at the *Cyp6g1* locus. *PLoS Genet.* **6**, e1000998 (2010).
- 687 49. Battlay, P. The quantitative genetics of insecticide resistance in *Drosophila*
688 *melanogaster*. (University of Melbourne, 2019).
- 689 50. Zwaan, B., Bijlsma, R. & Hoekstra, R. F. Direct selection on life span in *Drosophila*
690 *Melanogaster*. *Evolution* **49**, 649–659 (1995).
- 691 51. Rose, M. & Charlesworth, B. A test of evolutionary theories of senescence. *Nature*
692 **287**, 141–142 (1980).
- 693 52. Greene, D. *et al.* Genetic association analysis of 77,539 genomes reveals rare
694 disease etiologies. *Nat. Med.* **29**, 679–688 (2023).
- 695 53. Doust, C. *et al.* Discovery of 42 genome-wide significant loci associated with
696 dyslexia. *Nat. Genet.* **54**, 1621–1629 (2022).
- 697 54. Bevers, R. P. J. *et al.* Mitochondrial haplotypes affect metabolic phenotypes in the
698 *Drosophila* Genetic Reference Panel. *Nat. Metab.* **1**, 1226–1242 (2019).
- 699 55. Wang, J. B., Lu, H.-L. & St. Leger, R. J. The genetic basis for variation in resistance
700 to infection in the *Drosophila melanogaster* genetic reference panel. *PLOS Pathog.* **13**,
701 e1006260 (2017).
- 702 56. Jordan, K. W. *et al.* Genome-Wide Association for Sensitivity to Chronic Oxidative
703 Stress in *Drosophila melanogaster*. *PLoS ONE* **7**, e38722 (2012).
- 704 57. Shorter, J. *et al.* Genetic architecture of natural variation in *Drosophila melanogaster*
705 aggressive behavior. *Proc. Natl. Acad. Sci.* **112**, (2015).
- 706 58. Unckless, R. L., Rottschaefer, S. M. & Lazzaro, B. P. A Genome-Wide Association
707 Study for Nutritional Indices in *Drosophila*. *G3 GenesGenomesGenetics* **5**, 417–425
708 (2015).
- 709 59. Zwarts, L. *et al.* The genetic basis of natural variation in mushroom body size in
710 *Drosophila melanogaster*. *Nat. Commun.* **6**, 10115 (2015).
- 711 60. Zwarts, L. *et al.* Complex genetic architecture of *Drosophila* aggressive behavior.
712 *Proc. Natl. Acad. Sci.* **108**, 17070–17075 (2011).
- 713 61. Zeggini, E. & Ioannidis, J. P. Meta-analysis in genome-wide association studies.
714 *Pharmacogenomics* **10**, 191–201 (2009).
- 715 62. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits:
716 consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**, 356–369 (2008).
- 717 63. May, C. M., Doroszuk, A. & Zwaan, B. J. The effect of developmental nutrition on life
718 span and fecundity depends on the adult reproductive environment in *D. rosophila*

- 719 *melanogaster*. *Ecol. Evol.* **5**, 1156–1168 (2015).
- 720 64. Iliadi, K. G., Knight, D. & Boulianne, G. L. Healthy Aging – Insights from *Drosophila*.
721 *Front. Physiol.* **3**, (2012).
- 722 65. Hendricks, G., Tkaczyk, D., Lin, J. & Feeney, P. Crossref: The sustainable source of
723 community-owned scholarly metadata. *Quant. Sci. Stud.* **1**, 414–427 (2020).
- 724 66. Barrett, T. *et al.* NCBI GEO: archive for high-throughput functional genomic data.
725 *Nucleic Acids Res.* **37**, D885–D890 (2009).
- 726 67. Brazma, A. ArrayExpress--a public repository for microarray gene expression data at
727 the EBI. *Nucleic Acids Res.* **31**, 68–71 (2003).
- 728 68. Kodama, Y., Shumway, M., Leinonen, R., & on behalf of the International Nucleotide
729 Sequence Database Collaboration. The sequence read archive: explosive growth of
730 sequencing data. *Nucleic Acids Res.* **40**, D54–D56 (2012).
- 731

732 Figures

733 **Figure 1. General content of the DGRP pool web tool.** **A.** Pubmed search on “*Drosophila*
734 DGRP” terms unveiled 131 results from 2012 to 2023 (search made on March 2023). **B.** Sex
735 of the DGRP lines used across all 125 studies (left) and 41 curated studies (right), for each
736 phenotype. Studies have only been curated up to study 41 at the time of writing. **C.** Number
737 of studies per phenotype category. Studies can be assigned to multiple categories. **D.**
738 Number of phenotypes per study and per sex. Studies without attached phenotypes were not
739 plotted. Of note, a given phenotype can be measured for different sexes and thus counted
740 multiple times.

741 **Figure 2. Within- and cross-study phenotype correlations.** **A.** Spearman’s correlation of
742 all phenotypes available in the 41 curated studies. Of note, we separately computed the
743 phenotype correlations when data per sex were available (M, F or NA), and we restricted the
744 computation to quantitative (non-categorical) phenotypes. Phenotypes are grouped by study
745 (colored box at the bottom of the plot). **B.** Absolute value of the Spearman’s correlation of
746 pairs of phenotypes that originated from the same study (within-study) and those that
747 originated from two different studies (cross-study). Of note, displayed values are median.
748 Mean values are 0.099 for cross-study, and 0.260 for within-study. Again, we restricted the
749 calculation to the 41 curated studies. **C.** Correlation of two longevity phenotypes from the
750 same study (Arya et al, 2010)³⁰, revealing a strong correlation between Female (F) and Male
751 (M) longevity. **D.** Correlation of two phenotypes from different studies: mean lifespan
752 (Durham et al, 2014)²⁷ and mean longevity (Arya et al, 2010)³⁰. Of note, both the C and D
753 plots were generated using the “phenotype correlation” tool in DGRP pool.

754 **Figure 3. Phenotype correlations contribute new biological insights.** **A.** Correlation of
755 mean femur length (Grubbs et al., 2013)³⁵ vs. mean head width (Vonesch et al., 2016)²⁴
756 showing the significant cross-study association of organismal size traits. **B.** Correlation of
757 remating proportion (Chow et al., 2013)³⁷ vs. mean fecundity (Durham et al., 2014)²⁷. **C.** 50
758 phenotypes correlated with longevity (Arya et al, 2010)³⁰ at a 25% FDR threshold, revealing
759 three main groups of phenotypes: lifespan phenotypes (middle rows), other correlated
760 phenotypes (bottom rows) and anti-correlated phenotypes (top rows). Of note, both the A
761 and B plots were generated using the “phenotype correlation” tool in DGRP pool.

762 **Figure 4. Overview of GWAS results across phenotypes and one case study.** **A.**
763 Distribution of the number of significant variants after a GWAS, for each phenotype available
764 in DGRP pool. Of note, all values >1000 have been set to 1000, for easier visualization. **B.** For
765 each variant, we plotted the number of times it was significantly associated with a phenotype
766 (y-axis = number of occurrences). It is worth noting that we chose a Manhattan plot for
767 representing this information, but this is not a “real” GWAS Manhattan plot. **C.** Case study on
768 survival to azinphos-methyl exposure (Battlay et al., 2016)²⁶, here to a 0.25 µg/ml dose. This
769 plot was extracted from the phenotype’s page on DGRP pool at
770 <https://dgrpools.epfl.ch/phenotypes/20>. **D.** Manhattan plot (taken from DGRP pool’s result page
771 https://dgrpools.epfl.ch/phenotypes/20/gwas_analysis) showing the association of variants to
772 “survival at 0.25 µg/ml dose” phenotype. **E.** Boxplot (taken from DGRP pool’s result page
773 https://dgrpools.epfl.ch/phenotypes/20/gwas_analysis), showing the effect of the top variant,
774 2R:8072884, which is a long insertion.

775 **Figure 5. Analysis of extremeness among DGRP lines across 40 phenotypes. A.**
776 Fraction of extremeness of a given DGRP line. DGRP lines are assigned as 'extreme' in a
777 phenotype when they are in the top or bottom 15% of the phenotypic spectrum. Phenotypes
778 were selected based on the curated studies which had the following categories assigned to
779 them: Life history traits, Immunity, Toxicity, Resistance, Fecundity, Aging. DGRP lines were
780 included if they had at least 50 phenotypic measures. **B.** Scatter plot for the fraction of
781 extremeness of DGRP lines. On the x-axis, the fraction of extremeness is plotted for
782 females, whereas males are plotted on the y-axis. **C.** Most extreme and moderate DGRP
783 lines per sex. On the x-axis, the adjusted fraction of extremeness is provided. Individual
784 fractions of extremeness per phenotype were retrieved for each DGRP line. The fraction was
785 adjusted by 1 minus the fraction of extremeness if the fraction of extremeness was above
786 0.5. Because extremeness can range from 0 to 0.15 or 0.85 to 1, we adjusted the fraction of
787 extremeness for plotting purposes. DGRP lines with a low adjusted fraction of extremeness
788 are therefore more extreme, whereas a high adjusted fraction of extremeness is
789 representative of more moderate DGRP lines. **D.** Extreme and moderate DGRP line
790 pairings. On the x-axis, the adjusted fraction of extremeness is provided. Extreme and
791 moderate line pairings were retrieved by searching for DGRP lines for which the fraction of
792 extremeness between females and males was not greater than 0.05 while still having the
793 highest and lowest average fraction of extremeness (across sex). **E.** Looking at phenotypes
794 from Figure 2D marked as longevity/lifespan, for DGRP lines which are in the top 5 of
795 fraction of extremeness for each respective sex, including DGRP_852 and DGRP_042 (red
796 shades) from **5D**. We specifically highlight DGRP_757, DGRP_765 in blue shades to show
797 that they are across multiple studies in the lower end of the lifespan as is expected given
798 that the lifespan trait is robust across studies. Similarly, DGRP_320 shows a trend in which it
799 displays above average lifespan. Other extreme DGRP lines which were in each respective
800 top 5 are displayed in gray.

801

802 Supplement Figures

803 **Supplemental Figure S1. Screenshot from the curator's view for a given study - Metadata**
804 **section.** This screenshot shows the metadata section of the editing page for a study, where the
805 curator can edit any of the fields. We expect the curator to set a description (short abstract) for the
806 study, and associate some categories. The curator can also deactivate a phenotype if he/she
807 considers that it is not a proper phenotype (like the number of replicates). Once the curation is done,
808 the "Status" field can be changed to "Validated", which signifies that the curation process is finished,
809 allowing the study to be widely visible to the users.

810 **Supplemental Figure S2. Number of phenotypes per study.** Studies have only been curated up to
811 study 41 at the time of writing. Studies without attached phenotypes were not plotted. We here
812 disregard the sex and thus count the unique phenotypes irrespective of the available sex associated
813 with them. The 41 curated studies have 500 different phenotypes (~60%), while the remaining (S42-
814 S125) studies provide another 329 phenotypes (~40%).

815 **Supplemental Figure S3. Screenshot from the curator's view for a given study - Phenotype**
816 **section.** This screenshot shows the phenotype section of the editing page for a study, where the
817 curator can create or update the phenotyping data associated with the study. Here, the data is from
818 (Huang et al, 2020)³¹, taken as an example study. It is divided into 4 columns (from left to right): 1)
819 dataset type (raw or summary), 2) phenotypes, 3) DGRP lines, and 4) actions. If the curator submits
820 or updates a phenotype, a parsing script is then run to check the data format, and then the data is
821 updated in the DGRP database. For each study, the curator can submit, update or delete a unique
822 summary dataset, containing summary data for each DGRP line (for e.g. mean or median values).
823 The curator can also submit multiple raw datasets, if the raw data is available for this study. Raw data
824 means that the phenotyping data is not summarized, i.e. there are multiple values for the same DGRP
825 line (e.g. because of replicate flies). **Note:** Gray phenotypes are deactivated phenotypes, i.e. not
826 treated as real phenotypes (here, it is a block number for each fly).

827 **Supplemental Figure S4. Screenshot from the phenotype correlation tool result page.** This
828 screenshot shows the results obtained after running the phenotype vs phenotype correlation tool,
829 available directly from a phenotype page, by clicking the "Compute Correlation" button. Of note, there
830 is also the possibility to run this tool from the "Tool" section displayed on the banner of the DGRP
831 website on any user-submitted phenotype file.

832 **Supplemental Figure S5. Spearman's correlation of all phenotypes available in the 41 curated**
833 **studies.** Here, we applied a binary coloring using a fixed threshold to better visualize the correlations.
834 All correlations above $\text{abs}(\text{Spearman's } \rho) > 0.3$ are shown in black (therefore anti-correlated
835 phenotypes are also in black), the others are in white.

836 **Supplemental Figure S6. Comparison of correlation within and cross-study.** We calculated the
837 absolute value of the Spearman's correlation of pairs of phenotypes that originated from the same
838 study (within-study) and those that originated from two different studies (cross-study). Of note,
839 displayed values are median. Mean values are 0.170 for cross-study, and 0.287 for within-study.
840 These values are calculated across all phenotypes (125 studies).

841 **Supplemental Figure S7. Screenshot from the GWAS result page.** This screenshot shows the
842 results obtained after running the GWAS analysis, available directly from a phenotype page, by
843 clicking the "GWAS" button. Of note, there is also the possibility to run this tool from the "Tool"
844 section displayed on the banner of the DGRP website on any user-submitted phenotype file.

845 **Supplemental Figure S8. Distribution of the number of GWAS hits per phenotype depending**
 846 **on the significance threshold.**

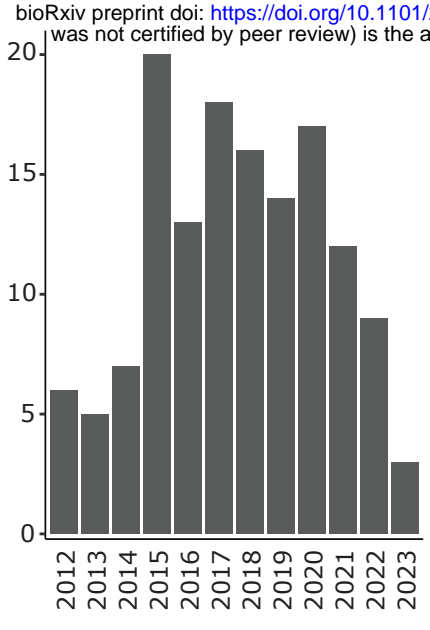
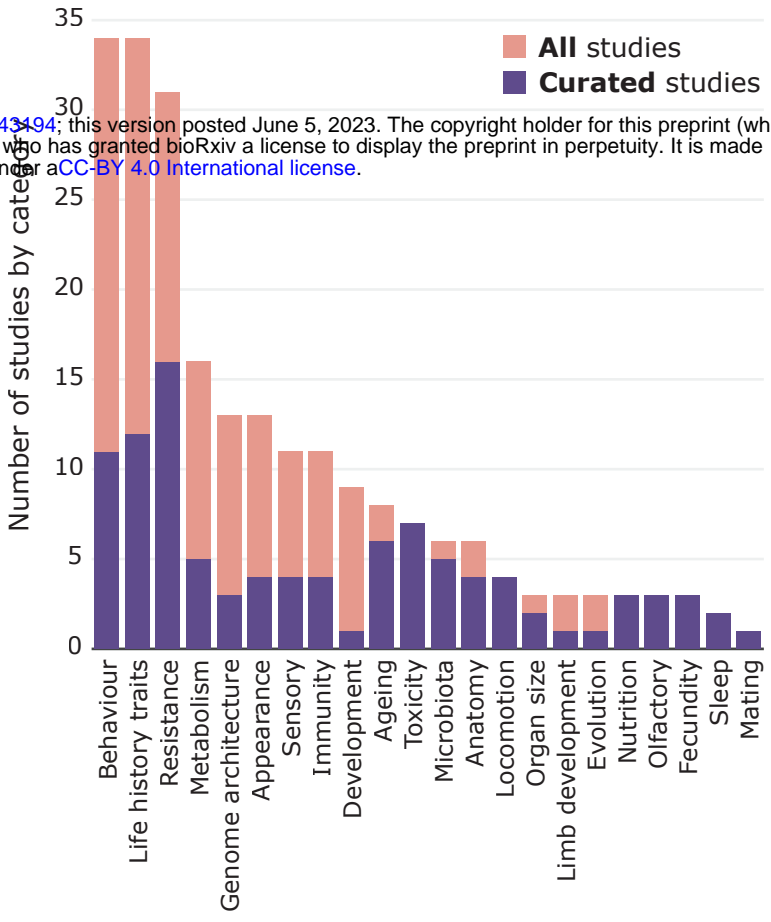
847 **Tables**

REFERENCE		DGRPpool	DGRP2 (Mackay et al., 2012) (Huang, Massouras, et al., 2014)	
DATA	DGRP lines	341	205	
	DGRP studies	125 (41 fully curated)	12	
	Phenotypes	935	31	
	Gene Expression data	External links	✓	
TOOLS	GWAS	Calculated on all phenotypes	✓	
		User upload	✓	
		Method	Plink2	FastLMM
		Covariates	Wolbachia + 5 Insertions	Wolbachia + 5 Insertions
		Boxplot of REF vs ALT	✓	
		PheWAS of top variants	✓	
	Phenotype correlation	Calculated on all phenotypes	✓	
	User upload	✓		
WEB	URL	https://dgrpools.epfl.ch/	http://dgrp2.gnets.ncsu.edu/	
	Backend	Ruby-on-rails + PostgreSQL	NA	
	Frontend	Javascript, Plotly	NA	
FEAT.	Curation system & tools	✓		
	Publish new studies	✓		
	Interactive plots	✓		

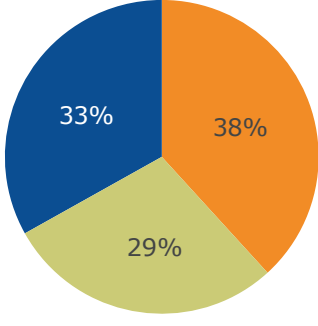
848
 849 **Table 1. Comparison of the two currently available web portals organizing DGRP phenotyping**
 850 **data.** This table compares different features available in DGRPpool, with DGRP2, the current main
 851 resource for DGRP data. It separates the features into 1) **Data**, which summarizes the available
 852 phenotyping data, 2) **Tools**, which lists the available tools and options, mainly GWAS, PheWAS and
 853 phenotype correlation, 3) **Web**, which describes the website itself, and 4) **Additional features**, that
 854 are available in DGRPpool, such as the curation system, the possibility to publish new studies and the
 855 interactive plots.

A

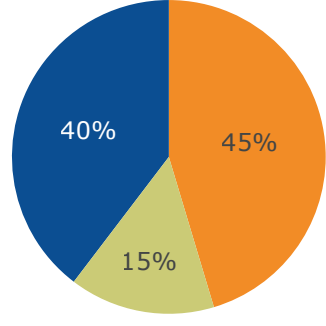
Number of "DGRP" & "Drosophila" publications per year

**C****B**

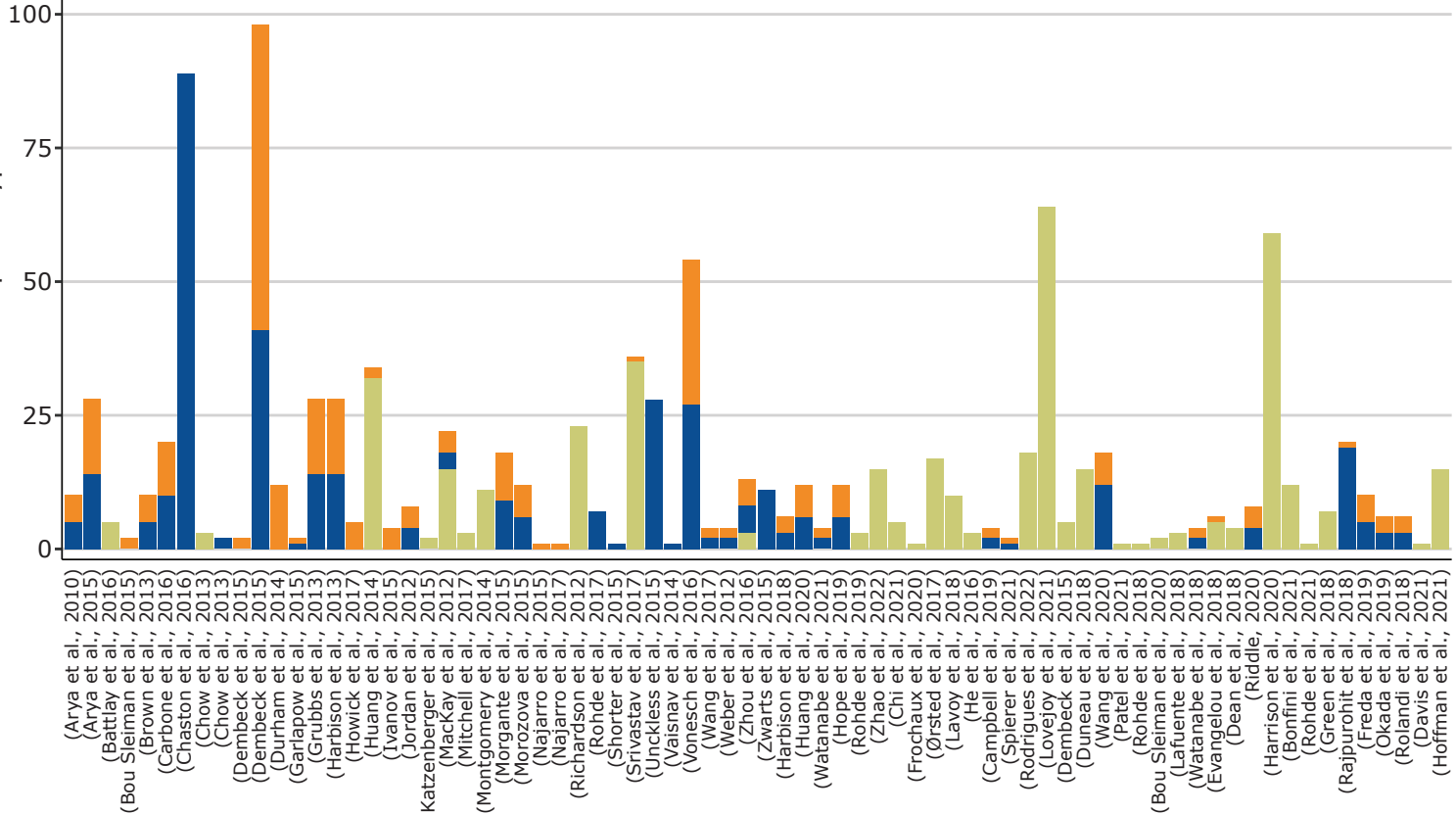
Sex of DGRP lines across all phenotypes

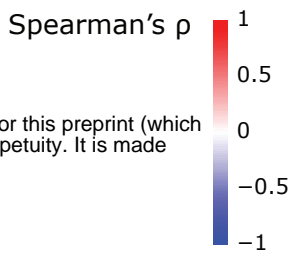


Sex of DGRP lines across curated phenotypes

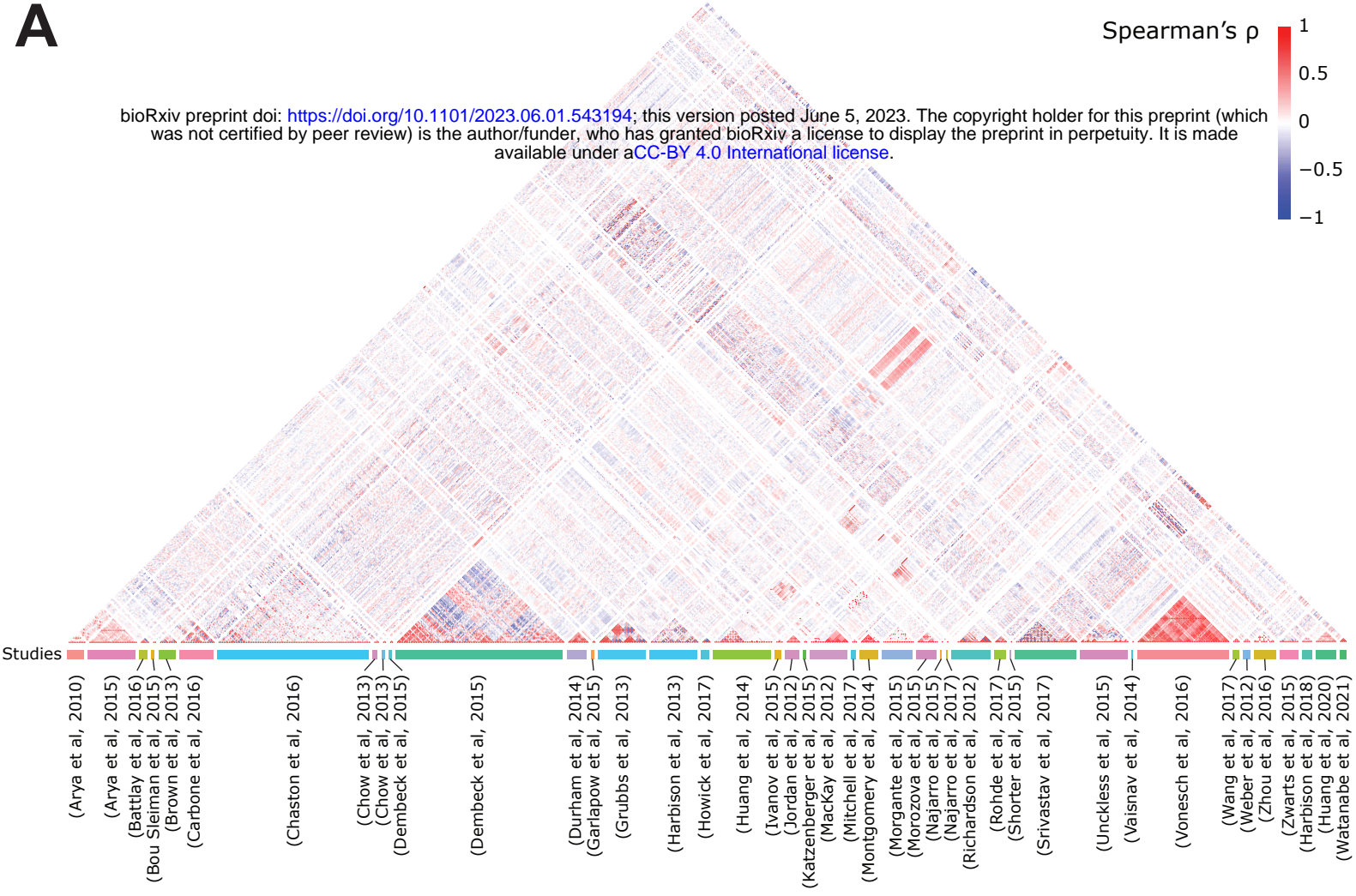
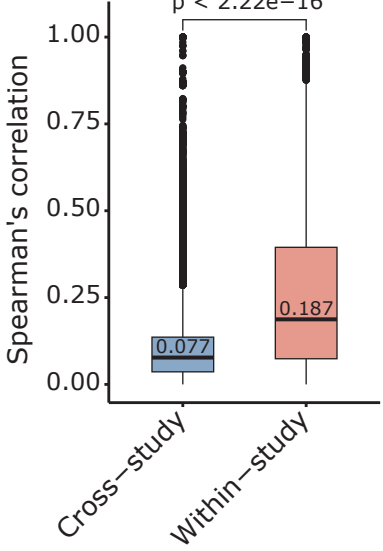
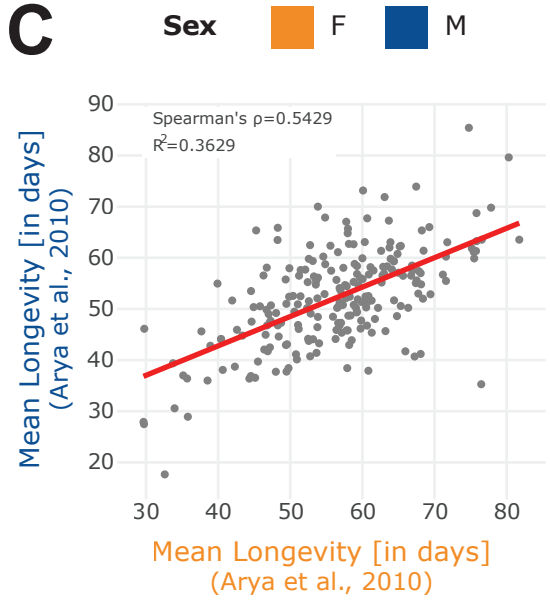
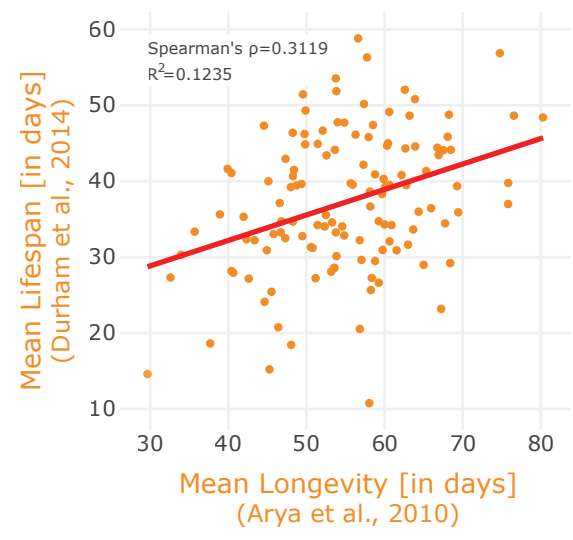
**Sex****D**

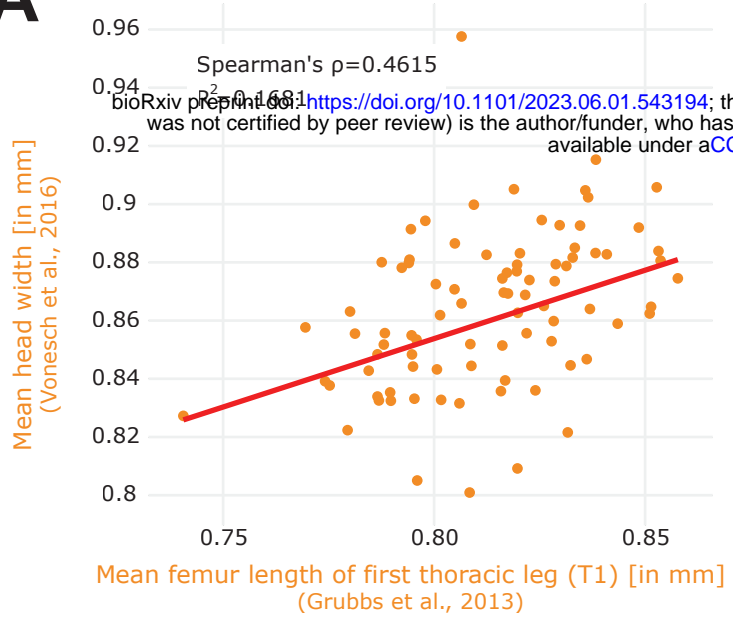
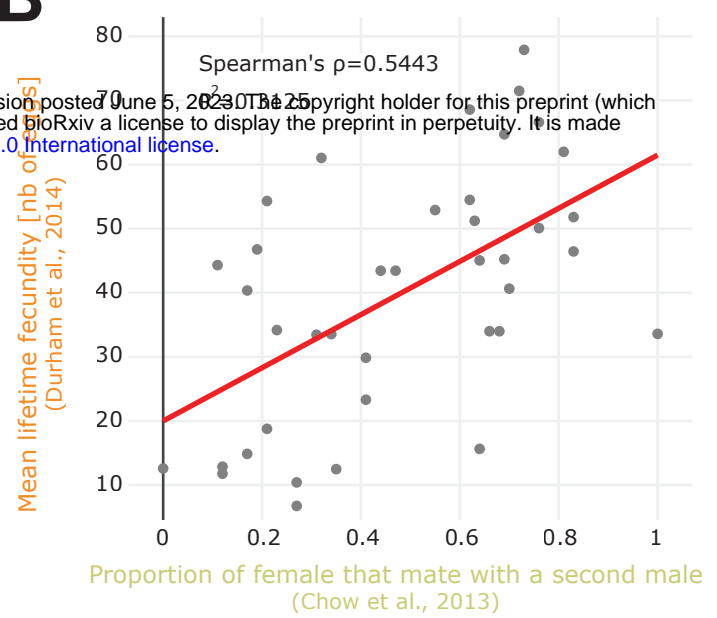
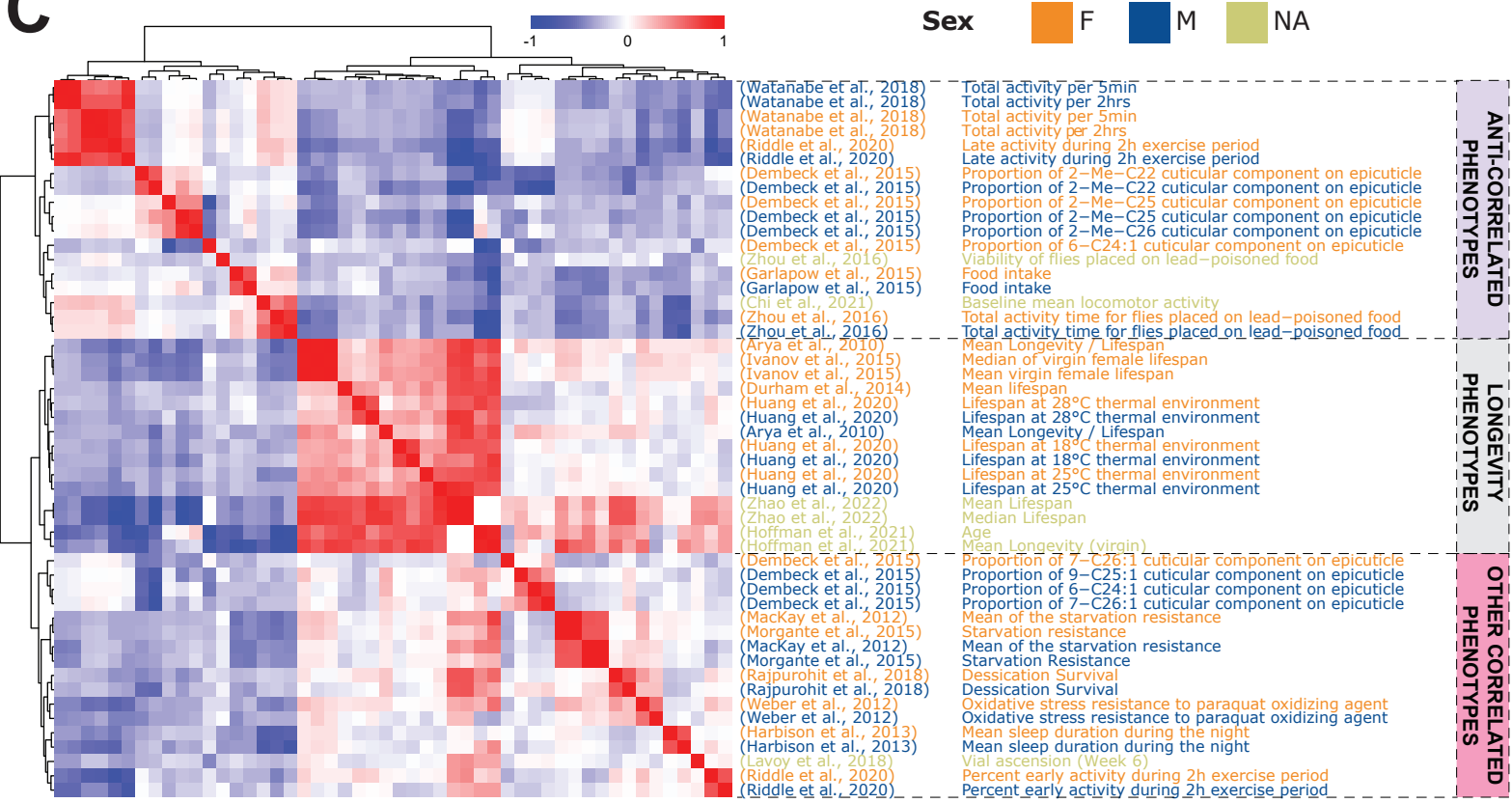
Number of phenotypes



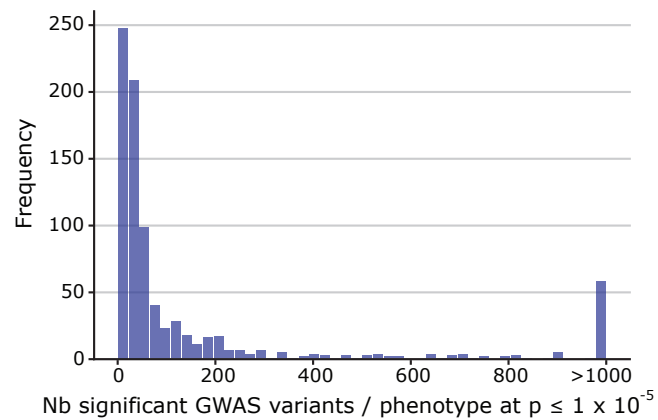
A

bioRxiv preprint doi: <https://doi.org/10.1101/2023.06.01.543194>; this version posted June 5, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

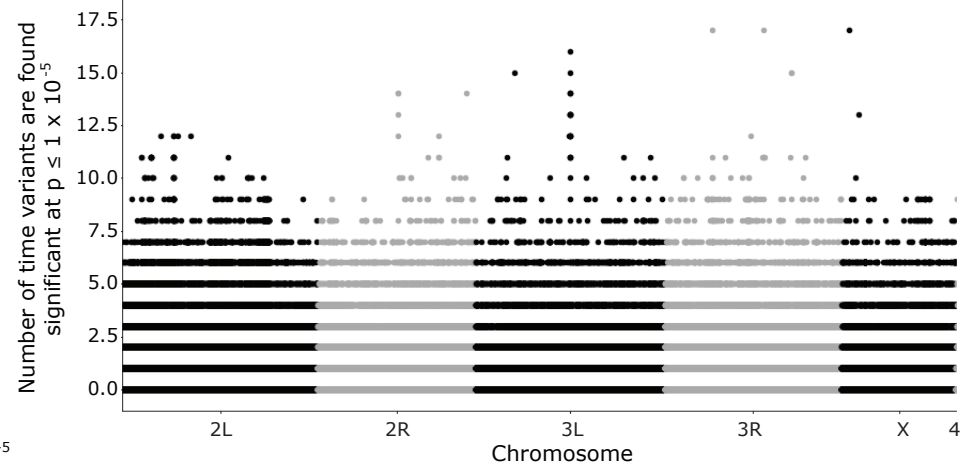
**B****C****D**

A**B****C**

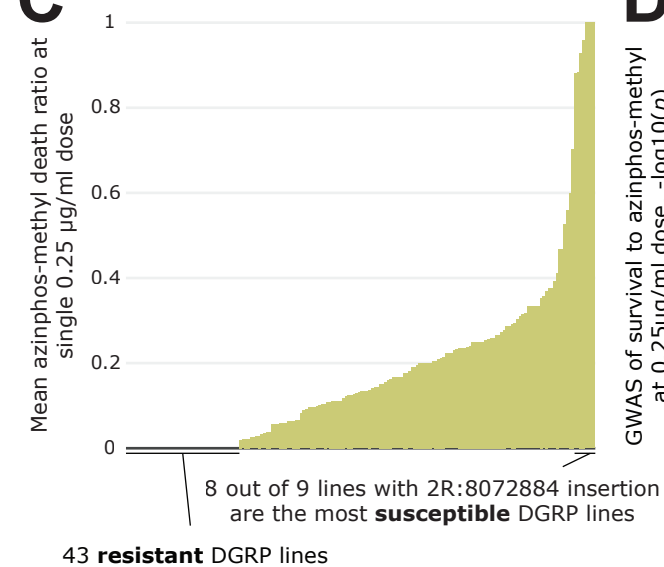
A



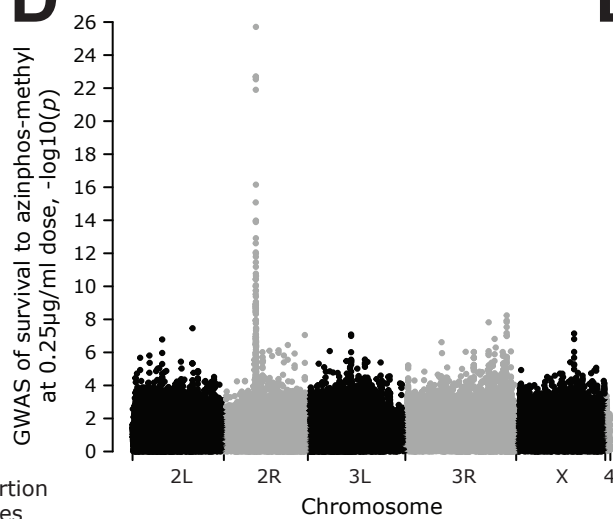
B



C



D



E

