Comprehensive patient-level classification and quantification of driver events in TCGA PanCanAtlas cohorts

Short title: Patient-level analysis of driver events in TCGA PanCanAtlas cohorts

Alexey D. Vyatkin[§], Danila V. Otnyukov, Sergey V. Leonov and Aleksey V. Belikov*

Laboratory of Innovative Medicine, School of Biological and Medical Physics, Moscow Institute of Physics and Technology, 141701 Dolgoprudny, Moscow Region, Russia

[§]In memory of Alexey D. Vyatkin, 13.07.1997-14.02.2021

*Corresponding author: belikov.research@gmail.com

Abstract

There is a growing need to develop novel therapeutics for targeted treatment of cancer. The prerequisite to success is the knowledge about which types of molecular alterations are predominantly driving tumorigenesis. To shed light onto this subject, we have utilized the largest database of human cancer mutations – TCGA PanCanAtlas, multiple established algorithms for cancer driver prediction (2020plus, CHASMplus, CompositeDriver, dNdScv, DriverNet, HotMAPS, OncodriveCLUSTL, OncodriveFML) and developed four novel computational pipelines: SNADRIF (Single Nucleotide Alteration DRIver Finder), GECNAV (Gene Expression-based Copy Number Alteration Validator), ANDRIF (ANeuploidy DRIver Finder) and PALDRIC (PAtient-Level DRIver Classifier). A unified workflow integrating all these pipelines, algorithms and datasets at cohort and patient levels was created.

We have found that there are on average 12 driver events per tumour, of which 0.6 are single nucleotide alterations (SNAs) in oncogenes, 1.5 are amplifications of oncogenes, 1.2 are SNAs in tumour suppressors, 2.1 are deletions of tumour suppressors, 1.5 are driver chromosome losses, 1 is a driver chromosome gain, 2 are driver chromosome arm losses, and 1.5 are driver chromosome arm gains. The average number of driver events per tumour increases with age (from 7 to 15) and cancer stage (from 10 to 15) and varies strongly between cancer types (from 1 to 24). Patients with 1 and 7 driver events per tumour are the most frequent, and there are very few patients with more than 40 events. In tumours having only one driver event, this event is most often an SNA in an oncogene. However, with increasing number of driver events per tumour, the contribution of SNAs decreases, whereas the contribution of copy-number alterations and aneuploidy events increases.

Author Summary

By analysing genomic and transcriptomic data from 10000 cancer patients through our custom-built computational pipelines and previously established third-party algorithms, we have found that half of all driver events in a patient's tumour appear to be gains and losses of chromosomal arms and whole chromosomes. We therefore suggest that future therapeutics development efforts should be focused on targeting aneuploidy. We have also found that approximately a third of driver events in a patient are whole gene amplifications and deletions. Thus, therapies aimed at copy-number alterations also appear very promising. On the other hand, drugs aiming at point mutations are predicted to be less successful, as these alterations are responsible for just a couple of drivers per tumour. One notable exception are patients having only one driver event in their tumours, as this event is almost always a single nucleotide alteration in an oncogene.

Introduction

Driver events are the molecular and cellular events that drive cancer progression, often called driver mutations [1]. The knowledge about the spectrum and quantity of driver events in individual patients and groups of patients is crucial to inform design and selection of targeted therapeutics. Historically, most attention has been devoted to point mutations or single nucleotide alterations (SNAs), as most driver prediction algorithms work only with this class of driver events. As the average numbers of driver SNAs per tumour in various cancer types have been estimated before and shown to vary within 1-10 range, with the average across all cancer types of 4 [2], this defined the current thinking that the number of driver events per tumour is relatively low.

However, SNA drivers represent only the tip of the iceberg, and are likely not the most crucial contribution to cancer progression. SNAs are very well tolerated by cancer cells [2]. It is known that cancer cells contain large numbers of deletions and amplifications (often called copy number alterations, or CNAs), translocations, inversions, full chromosome and chromosomal arm gains and losses (aneuploidy), as well as epigenetic modifications [3–5], but the driver potential of these alterations has been left almost unexplored due to the scarcity of driver prediction algorithms for these classes of events. Here, we set the goal to predict, classify and quantify as many different classes of driver events as possible, using clear and straightforward principles, and developed custom computational pipelines for this purpose.

Another shortcoming of the majority of existing driver prediction algorithms is that they work at the cohort level, i.e. they predict driver mutations for large groups of patients, usually having a particular cancer type. This does not allow to look at the composition of driver events in individual patients. We wrote specific scripts to convert cohort-level predictions into patient-

72 level events, which also allowed seamless integration of the results from various third-party
73 algorithms, including 2020plus [6], CHASMplus [7], CompositeDriver [8], dNdScv [2], DriverNet
74 [9], HotMAPS [10], OncodriveCLUSTL [11], and OncodriveFML [12].. This is useful, as each
75 individual driver prediction algorithm has its own strengths and shortcomings, and combining
76 results from multiple algorithms allows to obtain more complete and balanced picture,
77 ensuring that less driver mutations have been missed. Having patient-level data is also
78 convenient for analysing driver event composition in various demographic and clinical groups
79 (e.g. patients grouped by age, sex or cancer stage) without the need to rerun the driver
80 prediction algorithms on subgroups of patients, which would otherwise not only consume
81 additional time and resources but, most importantly, reduce the predictive power. Patient-level
82 data also allow to study heterogeneity of cancer patient cohorts, e.g. by using distribution
83 histograms.
84
85 In addition to these existing driver prediction algorithms, we decided to create our own, using
86 clear and simple rules to have an internal reference standard. We called this algorithm SNADRIF
87 – SNA DRIver Finder. It is a Python 3.7 software package that predicts cancer driver genes from
88 the TCGA PanCanAtlas SNA data and classifies them into oncogenes and tumour suppressors.
89 Driver prediction is based on calculating the ratio of nonsynonymous SNAs to silent SNAs [2],
90 whereas driver classification is based on calculating the ratio of hyperactivating SNAs to
91 inactivating SNAs [13]. Bootstrapping is used to calculate statistical significance and Benjamini–
92 Hochberg procedure is used to keep false discovery rate under 5%.
93
94 Copy-number alterations (CNA) usually involve large chunks of DNA containing tens or
95 hundreds of genes, which makes CNA data not very useful for uncovering individual driver
96 genes. Nevertheless, it is an important source of information about amplifications and

97 deletions of driver genes predicted from SNA data. However, due to CNA data coarseness, we
98 wanted to clarify the actual copy number status of individual genes using mRNA and miRNA
99 expression data available at TCGA PanCanAtlas. For this purpose, we created another Python
100 3.7 software package called GECNAV - Gene Expression-based CNA Validator. CNA validation is
101 based on comparing the CNA status of a given gene in a given patient to expression of this gene
102 in this patient relative to the median expression of this gene across all patients.
103
104 Aneuploidy – chromosome arm and full chromosome gains and losses – makes a substantial
105 contribution to the number of driver alterations per tumour, however, there are no existing
106 algorithms to differentiate driver aneuploidies from passenger ones. Therefore, we built our
107 own pipeline called ANDRIF - ANeuploidy DRIver Finder. It is a Python 3.7 software package that
108 predicts driver chromosomal arm or full chromosome gains or losses from the TCGA
109 PanCanAtlas aneuploidy data. Driver prediction is based on calculating the average alteration
110 status for each arm or chromosome in each cancer type. Bootstrapping is used to obtain the
111 realistic distribution of the average alteration statuses under the null hypothesis. Benjamini–
112 Hochberg procedure is performed to keep the false discovery rate under 5%.
113
114 Finally, we needed an algorithm to integrate all data on driver alterations from different
115 algorithms - our own and third-party. We called this algorithm PALDRIC - PAtient-Level DRIver
116 Classifier. It is a Python 3.7 software package that translates cohort-level lists of driver genes or
117 mutations to the patient level, classifies driver events according to the molecular causes and
118 functional consequences, and presents comprehensive statistics on various kinds of driver
119 events in various demographic and clinical groups of patients. PALDRIC allows to natively
120 combine outputs of various third-party algorithms to investigate optimal combinations.
121

Our overall workflow can be seen in **Fig 1**.

Fig 1. Workflow. Actual names of the files generated and used in the workflow are shown.

125 **Results**

126

127 To get two different perspectives on the number and composition of driver events, we
128 performed two different analyses. In the first one, we used the combination of results from
129 SNADRIF, several third-party algorithms - 2020plus, CHASMplus, CompositeDriver, HotMAPS,
130 OncodriveFML, and a consensus driver gene list from 26 algorithms [8], each of them applied to
131 the whole TCGA PanCanAtlas dataset. We also used a list of COSMIC Cancer Gene Census Tier 1
132 genes affected by somatic SNAs and CNAs, as it represents the current gold standard of verified
133 cancer drivers [14]. To minimize false positives, we used only genes that were predicted as
134 drivers by at least two of our sources, including CGC and the list from Bailey et al. [8].
135 Benchmarking (see **S1 Files)** showed that such combinatory approach allows to identify more
136 high-confidence driver genes than CGC or Bailey et al. lists contain (244 vs 182 vs 164, **Fig 2A**)
137 and to recover more CGC genes (62%) than any individual algorithm or even Bailey et al list (**Fig**
138 **2B**). The overlap of three sources recovered fewer CGC genes (45%, **Fig 2B**), therefore we
139 concluded that the overlap of two sources is optimal. SNADRIF sensitivity was comparable to
140 CompositeDriver (**Fig 2B**), and it contributed 16 driver genes (6.5%) to the overall consensus,
141 i.e. it validated 16 genes that were found only in one other source, including two genes (*MYCL*
142 and *TBL1XR1*) that were present only in CGC and not predicted by any third-party algorithm
143 (**Table 1**). Unfortunately, application of driver gene lists discovered through pancancer analysis
144 equally to every cancer type results in unrealistically high numbers of driver events per patient,
145 which is to be expected as this approach ignores tissue specificity of driver genes. Therefore,
146 we present this analysis only as **S2 Files**.

147

148 **Fig 2. Benchmarking of driver prediction algorithms.** (**A**) Number of driver genes predicted by
149 each algorithm applied to pancancer dataset. CGC **–** COSMIC Cancer Gene Census Tier 1 genes

150 affected by somatic SNAs and CNAs; Overlap 2 and Overlap 3 – genes predicted by at least 2 or
151 3 of the listed algorithms, respectively. (**B**) Sensitivity of algorithms applied to pancancer
152 dataset, estimated as the percentage of genes from positive control lists identified. Positive
153 control lists include CGC (blue), Overlap 2 (orange) and Overlap 3 (yellow), see legend for (A).
154 (**C**) Number of driver gene-cohort pairs predicted by each algorithm applied individually to each
155 cancer type (cohort). CGC - COSMIC Cancer Gene Census Tier 1 gene-cohort pairs affected by
156 somatic SNAs and CNAs; Overlap 2 and Overlap 3 – gene-cohort pairs predicted by at least 2 or
157 3 of the listed algorithms, respectively. (**D**) Sensitivity of algorithms applied individually to each
158 cancer type (cohort), estimated as the percentage of gene-cohort pairs from positive control
159 lists identified. Positive control lists include CGC (blue), Overlap 2 (orange) and Overlap 3
160 (yellow), see legend for (C).
161
162 **Table 1. Driver genes uniquely predicted by SNADRIF and one other source.**

Gene	Entrez Gene ID	Also predicted by
AMD1	262	HotMAPS
BMPR1A	657	2020plus
CMTR2	55783	OncodriveFML
ELL2	22936	HotMAPS
HNRNPDL	9987	2020plus
LYRM4	57128	OncodriveFML
METTL14	57721	HotMAPS
MYCL	4610	CGC
PPP3R1	5534	HotMAPS
PTMA	5757	Bailey et al, 2018
RPL22	6146	Bailey et al, 2018
SAMHD1	25939	HotMAPS
TBL1XR1	79718	CGC
TDG	6996	HotMAPS
TNNI3K	51086	HotMAPS
TTK	7272	HotMAPS

163
164 In the second analysis we used the combination of results from 2020plus, CHASMplus,
165 CompositeDriver, dNdScv, DriverNet, HotMAPS, OncodriveCLUSTL, OncodriveFML, the

166 consensus list from [8] and a list of Cancer Gene Census Tier 1 genes affected by somatic SNAs
167 and CNAs, applied separately to each cancer cohort of TCGA PanCanAtlas. Applying algorithms
168 to individual cohorts allows to discover cancer type-specific drivers and avoid contamination by
169 false positives, i.e. driver genes discovered during pancancer analysis that do not in reality play
170 any role in a given cancer type. On the other hand, much fewer patients are available for
171 cohort-specific analysis, and this decreases statistical power to discover new driver genes. Of
172 note, our SNADRIF algorithm works best for pancancer analysis and struggles with small
173 cohorts, due to scarcity of point mutations. However, when a combination of driver prediction
174 algorithms is used, there are lower chances of missing an important driver gene even in a
175 cohort-specific analysis, as algorithms based on differing principles complement each other. To
176 minimize false positives, we used only genes that were predicted as drivers in the same cancer
177 type by at least two of our sources, including CGC and the list from Bailey et al [8]. CGC
178 attribution of driver genes to various cancer types is not very precise or comprehensive,
179 therefore benchmarking of cohort-specific algorithms on CGC likely underestimates their true
180 performance. Anyhow, benchmarking (see **S1 Files)** showed that our combinatory approach
181 allows to recover more CGC gene-cohort pairs (32%) than any individual algorithm or even
182 Bailey et al list (**Fig 2D**). Similarly to pancan analysis, the overlap of three sources recovered
183 fewer CGC gene-cohort pairs (23%, **Fig 2D**), therefore we chose the overlap of two sources as
184 optimal. Benchmarking also showed that our combinatory approach allows to identify more
185 high-confidence driver gene-cohort pairs than are present on the Bailey et al. list (705 vs 537,
186 **Fig 2C**). The results of this analysis would be presented in the following paragraphs, whereas
187 the underlying data and additional graphs could be found in **S3 Files.**
188
189 **Fig 3** shows the average number of driver events of different classes in males and females. It
190 could be seen that half of all driver events are chromosome- and arm-level gains and losses.

191 Moreover, a third of all driver events are amplifications of oncogenes and deletions of tumour
192 suppressors. Only one sixth of all driver events are SNAs. Although the overall difference
193 between males and females in the average number of driver events is not significant (P=0.462,
194 two-tailed heteroscedastic t-test, see **S4 Files**), there are significant differences in particular
195 types of driver events. Males have significantly less SNAs in oncogenes ($P=3.62*10^{-15}$) and in
196 tumour suppressors ($P=1.06*10^{-16}$), but significantly more amplifications of oncogenes
197 ($P=1.97*10^{-6}$), simultaneous occurrences of SNAs in one allele and deletions of the other allele
198 in tumour suppressors ($P=1.57*10^{-5}$), driver chromosome losses (P=0.01) and driver
199 chromosome gains ($P=3.5*10^{-10}$). High number of SNAs in females might be at least partially
200 explained by the contribution from UCEC cohort (see **Fig 7**).
201
202 **Fig 3. Driver event distribution by patient's gender.** Two-tailed heteroscedastic t-test was
203 performed. p-values indicate a significant increase in the average number of driver events of a
204 given type compared to the opposite gender. n.s. - difference not significant.
205
206 **Fig 4** demonstrates an increase in the number of driver events with age in males and females. It
207 can be seen that in female patients the number of driver events increases dramatically from the
208 earliest age until approximately 50 y.o., i.e. the age of menopause. After this, the growth of the
209 number of driver events slows down. In males the picture is different. There is a high number of
210 driver events in the earliest age group, due to the frequent occurrence of testicular cancer
211 (TGCT) in this age group and a relatively high number of driver events in this cancer type (see
212 **Fig 7**). The next age group has much lower number of drivers (P=0.041, one-tailed
213 heteroscedastic t-test, see **S4 Files**), and like in females, their number increases with age, but
214 until an older age, approximately 70 y.o., when their growth slows down. Surprisingly, the
215 number of SNAs in oncogenes does not increase with age, at least since 30 y.o.

216

217 **Fig 4. Driver event distribution by patient's age.** Colour coding as in Fig 3. One-tailed
218 heteroscedastic t-test was performed. p-values indicate a significant difference in the average
219 number of driver events between indicated age groups. n.s. - difference not significant.

220

221 **Fig 5** demonstrates that in females, the number of driver events increases dramatically from
222 Stage I to Stage II ($P=9*10^{-15}$, two-tailed heteroscedastic t-test, see **S4 Files**), after which stays
223 more or less constant. It is interesting that this increase is due to CNAs and aneuploidy events
224 but not due to point mutations. The number of SNAs actually decreases. In males, the increase
225 in driver events with progression from Stage I to Stage IV is more gradual and chaotic. As in
226 females, there is a prominent increase from Stage I to Stage II ($P=4*10^{-5}$) but, for unclear
227 reason, there appears to be less driver events in Stage III than in Stage II tumours (P=0.003).
228 However, Stage IV tumours have more events than Stage II ($P=3.3*10^{-9}$) and Stage III ($P=4.9*10^{-}$
229 ^{17}) tumours. Unlike in females, the increase is mediated also by SNAs, in addition to CNAs and
230 aneuploidy events. In males, there are no significant changes in the number of driver
231 chromosome losses between stages. This complicated picture might be explained by unequal
232 representation of patients with different cancer stages amongst cancer types. Indeed, a closer
233 look at the data (see **S3 Files**) shows that there are twice as many PRAD patients with Stage III
234 than Stage II cancer, and because PRAD patients typically have a lower than average number of
235 driver events (see **Fig 7**), this leads to the counterintuitive result of Stage III tumours having less
236 driver events than Stage II tumours in the pooled analysis of male patients. Likewise, there are
237 six times as many UCEC patients with Stage I than Stage II cancer (see **S3 Files**), and because
238 UCEC patients typically have a higher than average number of driver SNAs (see **Fig 7**), this leads
239 to the counterintuitive result of Stage II tumours having less SNA driver events than Stage I
240 tumours in the pooled analysis of female patients.

241

242 **Fig 5. Driver event distribution by cancer stage.** Colour coding as in Fig 3. Two-tailed
243 heteroscedastic t-test was performed. p-values indicate a significant difference in the average
244 number of driver events of a given type between the indicated stages. n.s. - difference not
245 significant.

246

247 In **Fig 6** we aimed to show how different classes of driver events increase proportionally to the
248 total number of driver events in a patient. It is striking that the number of SNAs in oncogenes
249 does not increase, considering that if a tumour has only one driver event it is almost certain to
250 be an SNA in an oncogene. There are very few patients with more than 30 driver events (see **Fig**
251 **8**), therefore the results for those groups become unreliable.

252

253 **Fig 6. Driver event distribution by total number of driver events per patient.** Colour coding as
254 in Fig 3.

255

256 **Fig 7** shows the average number of driver events of different classes in different cancer cohorts.
257 It can be immediately seen that there is a dramatic difference in the number and distribution of
258 driver events per patient between cancer types. Some cancer types (ACC, CHOL, KICH, KIRP,
259 MESO, PCPG, SARC, TGCT) are dominated by aneuploidy events, whereas in others aneuploidy
260 events contribute half or less drivers. In ACC, KICH, KIRP and TGCT, the absolute majority of
261 driver events are chromosome gains and losses. ACC, CHOL, KICH, KIRC, MESO and PCPG have
262 almost no driver alterations in oncogenes. Additionally, BRCA, ESCA, GBM, HNSC, KIRP, SARC
263 and TGCT have almost no SNAs in oncogenes, whereas THCA and THYM have no amplifications
264 of oncogenes. Moreover, TGCT, THCA, THYM and UVM have almost no driver alterations in
265 tumour suppressors, whereas KICH and PCPG are missing only SNAs in tumour suppressors.

266 Tumour suppressors in DLBC and UCEC and oncogenes in DLBC, LGG, PAAD, THCA, THYM and
267 UVM are predominantly affected by SNAs, which is not seen in other cancer types.
268
269 **Fig 7. Driver event distribution by cancer type.** Colour coding as in Fig 3.
270
271 **Fig. 8** shows the actual distributions of driver event quantities amongst patients of all cancer
272 types combined. It can be seen that these distributions have a sharp peak at one driver event
273 per tumour, which is the dominant peak in females, and a smoother peak at 7 events, which is
274 the dominant peak in males. THCA is the major contributor to the one-driver-event peak, but
275 also PRAD and THYM. There are almost no patients with more than 40 driver events per
276 tumour.
277
278 **Fig 8. Patient distribution by total number of driver events per patient.**

Discussion

Our results show that half of all driver events are chromosome- and arm-level gains and losses, a third are amplifications of oncogenes and deletions of tumour suppressors and only one sixth are SNAs. Hopefully, this finding will help to shift the current perception that there are very few driver events per tumour and they are mostly SNAs to the greater appreciation of copy-number alterations and especially aneuploidy events. This may stimulate efforts in development of novel therapeutics targeting these types of driver alterations.

It is interesting to speculate why the number of driver events steadily increases in women until 50 y.o. and in men until 70 y.o. As 50 y.o. is the average age of menopause whereas males are fertile until much older ages, it is logical to suggest that driver accumulation is somehow promoted by high levels of sex hormones. If we concur with the theory that driver events accumulate in stem cells in large part due to errors in DNA replication and chromosome segregation [15], and combine it with the knowledge that sex hormones increase the rate of stem cell division [16–18], the link becomes clear. It is known that individuals with higher levels of sex hormones have higher rates of cancer [19].

In tumours with one driver event, this event is almost always an SNA in an oncogene, and with increasing number of total driver events the number of CNAs and aneuploidy events increases dramatically, whereas the number of SNAs in tumour suppressors increases only slightly. Most strikingly, the number of SNAs in oncogenes does not increase at all. Thus, it is tempting to suggest that a hyperactivating SNA in an oncogene is the seed which initiates cancer and enables later appearance of CNAs and aneuploidy events, the latter two driving the progression

303 of cancer. However, it cannot be excluded that in tumours having more than one driver event
304 the initiating event is CNA or aneuploidy, as these are cross-sectional and not longitudinal data.
305
306 Our results on the number and composition of driver events amongst various cancer types and
307 individual patients pose several interesting questions. First, why there is such a great difference
308 in the number and composition of driver events between cancer types? Why only one driver
309 event is sufficient to initiate thyroid cancer but two dozens are observed in bladder
310 carcinomas? Why some cancer types are dominated by aneuploidy drivers and others by SNA
311 drivers? Why some cancers have no alterations in oncogenes whereas others have none in
312 tumour suppressors? Are these differences explained by different tissue microenvironments to
313 which these tumours have to adapt? Why than for some patients within the same cancer type
314 one driver event is sufficient to develop a detectable tumour whereas the others are not
315 diagnosed until dozens of events are accumulated?
316
317 One possible explanation is that in patients with low number of predicted driver events some
318 actually important drivers have not been discovered due to technical reasons, from sequencing
319 to bioinformatic analysis. Some of them might even be epigenetic [20], a class of drivers we
320 have not analysed in this study. Another possible explanation is that in patients with high
321 numbers of predicted drivers some of those drivers are actually false positives due to
322 imperfections of driver prediction algorithms. Although both of these explanations are likely
323 true to some degree, their contribution does not seem to be so large as to dramatically affect
324 the results. The most probable explanation is the inherent stochasticity of carcinogenesis
325 processes [21,22] combined with unequal strength of different driver events [23]. For example,
326 patients having tumours with only one driver event were unlucky that this event happened (by
327 chance) in one of the most crucial genes such as *BRAF*. On the other hand, other patients did

328 not develop a detectable tumour until they have accumulated 30 events because those events
329 involved only weak drivers. It would be interesting to develop a bioinformatic approach to
330 determine the strength of driver genes, e.g. based on how often a given gene is found amongst
331 a few sufficient ones.
332
333 Overall, our study brought some clarity to the distributions of driver events of various classes in
334 various demographic and clinical groups of cancer patients and highlighted the importance of
335 aneuploidy events and copy-number alterations. Many crucial questions, such as the reason
336 behind the dramatic variability in the number and composition of driver events between cancer
337 types and patients, remain to be answered.

338 **Methods**

339 Source files and initial filtering

340 TCGA PanCanAtlas data were used. Files "Analyte level annotations -

341 merged_sample_quality_annotations.tsv", "ABSOLUTE purity/ploidy file -

342 TCGA_mastercalls.abs_tables_JSedit.fixed.txt", "Aneuploidy scores and arm calls file -

343 PANCAN_ArmCallsAndAneuploidyScore_092817.txt", "Public mutation annotation file -

344 mc3.v0.2.8.PUBLIC.maf.gz", "gzipped ISAR-corrected GISTIC2.0 all_thresholded.by_genes file -

345 ISAR_GISTIC.all_thresholded.by_genes.txt", "RNA batch corrected matrix -

346 EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2.geneExp.tsv", "miRNA batch corrected

347 matrix - pancanMiRs_EBadjOnProtocolPlatformWithoutRepsWithUnCorrectMiRs_08_04_16.csv",

348 were downloaded from https://gdc.cancer.gov/about-data/publications/PanCan-CellOfOrigin.

349

350 Using TCGA barcodes (see https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/

351 and https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/sample-type-codes), all

352 samples except primary tumours (barcoded 01, 03, 09) were removed from all files. Based on

353 the information in the column "Do_not_use" in the file "Analyte level annotations -

354 merged_sample_quality_annotations.tsv", all samples with "True" value were removed from

355 all files. All samples with "Cancer DNA fraction" <0.5 or unknown or with "Subclonal genome

356 fraction" >0.5 or unknown in the file "TCGA_mastercalls.abs_tables_JSedit.fixed.txt" were

357 removed from the file "PANCAN_ArmCallsAndAneuploidyScore_092817.txt". Moreover, all

358 samples without "PASS" value in the column "FILTER" were removed from the file

359 "mc3.v0.2.8.PUBLIC.maf.gz" and zeros in the column "Entrez_Gene_Id" were replaced with

360 actual Entrez gene IDs, determined from the corresponding ENSEMBL gene IDs in the column

361 "Gene" and external database

362 ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz. Filtered

363 files were saved as “Primary_whitelisted_arms.tsv”,

364 “mc3.v0.2.8.PUBLIC_primary_whitelisted_Entrez.tsv”,

365 “ISAR_GISTIC.all_thresholded.by_genes_primary_whitelisted.tsv”,

366 ”EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2-v2.geneExp_primary_whitelisted.tsv”,

367 “pancanMiRs_EBadjOnProtocolPlatformWithoutRepsWithUnCorrectMiRs_08_04_16_primary_whit

368 elisted.tsv”.

369

370 <u>RNA filtering of CNAs</u>

371 Using the file “EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2-

372 v2.geneExp_primary_whitelisted.tsv”, the median expression level for each gene across

373 patients was determined. If the expression for a given gene in a given patient was below 0.05x

374 median value, it was encoded as “-2”, if between 0.05x and 0.75x median value, it was encoded

375 as “-1”, if between 1.25x and 1.75x median value, it was encoded as “1”, if above 1.75x median

376 value, it was encoded as “2”, otherwise it was encoded as “0”. The file was saved as

377 “EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2-

378 v2.geneExp_primary_whitelisted_median.tsv.” The same operations were performed with the

379 file

380 “pancanMiRs_EBadjOnProtocolPlatformWithoutRepsWithUnCorrectMiRs_08_04_16_primary_

381 whitelisted.tsv”, which was saved as

382 “pancanMiRs_EBadjOnProtocolPlatformWithoutRepsWithUnCorrectMiRs_08_04_16_primary_

383 whitelisted_median.tsv”

384 Next, the file “ISAR_GISTIC.all_thresholded.by_genes_primary_whitelisted.tsv” was processed

385 according to the following rules: if the gene CNA status in a given patient was not zero and had

386 the same sign as the gene expression status in the same patient (file

387 “EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2-

388 v2.geneExp_primary_whitelisted_median.tsv" or

389 "pancanMiRs_EBadjOnProtocolPlatformWithoutRepsWithUnCorrectMiRs_08_04_16_primary_

390 whitelisted_median.tsv" for miRNA genes), then the CNA status value was replaced with the

391 gene expression status value, otherwise it was replaced by zero. If the corresponding

392 expression status for a given gene was not found then its CNA status was not changed. The

393 resulting file was saved as

394 "ISAR_GISTIC.all_thresholded.by_genes_primary_whitelisted_RNAfiltered.tsv"

395

396 We named this algorithm GECNAV (Gene Expression-based CNA Validator) and created a

397 Github repository: https://github.com/belikov-av/GECNAV. The package used to generate data

398 in this article is available as **S5 Files**.

399

400 Aneuploidy driver prediction

401 Using the file "Primary_whitelisted_arms.tsv", the average alteration status of each

402 chromosomal arm was calculated for each cancer type and saved as a matrix file

403 "Arm_averages.tsv". By drawing statuses randomly with replacement (bootstrapping) from any

404 cell of "Primary_whitelisted_arms.tsv", for each cancer type the number of statuses

405 corresponding to the number of patients in that cancer type were generated and their average

406 was calculated. The procedure was repeated 10000 times, the median for each cancer type was

407 calculated and the results were saved as a matrix file "Bootstrapped_arm_averages.tsv".

408

409 P-value for each arm alteration status was calculated for each cancer type. To do this, first the

410 alteration status for a given cancer type and a given arm in "Arm_averages.tsv" was compared

411 to the median bootstrapped arm alteration status for this cancer type in

412 "Bootstrapped_arm_averages.tsv". If the status in "Arm_averages.tsv" was higher than zero

413 and the median in “Bootstrapped_arm_averages.tsv”, the number of statuses for this cancer
414 type in “Bootstrapped_arm_averages.tsv” that are higher than the status in
415 “Arm_averages.tsv” was counted and divided by 5000. If the status in “Arm_averages.tsv” was
416 lower than zero and the median in “Bootstrapped_arm_averages.tsv”, the number of statuses
417 for this cancer type in “Bootstrapped_arm_averages.tsv” that are lower than the status in
418 “Arm_averages.tsv” was counted and divided by 5000, and marked with minus to indicate arm
419 loss. Other values were ignored (cells left empty). The results were saved as a matrix file
420 “Arm_Pvalues_cohorts.tsv”.
421
422 For each cancer type, Benjamini–Hochberg procedure with FDR=5% was applied to P-values in
423 “Arm_Pvalues_cohorts.tsv” and passing P-values were encoded as “DAG” (Driver arm gain) or
424 “DAL” (Driver arm loss) if marked with minus. The other cells were made empty and the results
425 were saved as a matrix file “Arm_drivers_FDR5_cohorts.tsv”.
426
427 Alterations were classified according to the following rules: if the arm status in a given patient
428 (file “Primary_whitelisted_arms.tsv”) was “-1” and the average alteration status of a given arm
429 in the same cancer type (file “Arm_drivers_FDR5_cohorts.tsv”) was “DAL”, then the alteration
430 in the patient was classified as “DAL”. If the arm status in a given patient was “1” and the
431 average alteration status of a given arm in the same cancer type was “DAG”, then the alteration
432 in the patient was classified as “DAG”. In all other cases an empty cell was written. The total
433 number of DALs and DAGs was calculated, patients with zero drivers were removed, and the
434 results were saved as a matrix file “Arm_drivers_FDR5.tsv”.
435
436 Using the file “Primary_whitelisted_arms.tsv”, the values for the whole chromosomes were
437 calculated using the following rules: if both p- and q-arm statuses were “1” then the

438 chromosome status was written as “1”; if both p- and q-arm statuses were “-1” then the
439 chromosome status was written as “-1”; if at least one arm status was not known (empty cell)
440 then the chromosome status was written as empty cell; in all other cases the chromosome
441 status was written as “0”. For one-arm chromosomes (13, 14, 15, 21, 22), their status equals
442 the status of the arm. The resulting file was saved as “Primary_whitelisted_chromosomes.tsv”.
443
444 The same procedures as described above for chromosomal arms were repeated for the whole
445 chromosomes, with the resulting file “Chromosome_drivers_FDR5.tsv”. Chromosome drivers
446 were considered to override arm drivers, so if a chromosome had “DCL” (Driver chromosome
447 loss) or “DCG” (Driver chromosome gain), no alterations were counted on the arm level, to
448 prevent triple counting of the same event.
449
450 We named this algorithm ANDRIF (ANeuploidy DRIver Finder) and created a Github repository:
451 https://github.com/belikov-av/ANDRIF. The package used to generate data in this article is
452 available as **S6 Files**.
453
454 SNA driver prediction
455 Using the file “mc3.v0.2.8.PUBLIC_primary_whitelisted_Entrez.tsv” all SNAs were classified
456 according to the column “Variant_Classification”. “Frame_Shift_Del”, “Frame_Shift_Ins”,
457 “Nonsense_Mutation”, “Nonstop_Mutation” and “Translation_Start_Site” were considered
458 potentially inactivating; “De_novo_Start_InFrame”, “In_Frame_Del”, “In_Frame_Ins” and
459 “Missense_Mutation” were considered potentially hyperactivating;
460 “De_novo_Start_OutOfFrame” and “Silent” were considered passengers; the rest were
461 considered unclear. The classification results were saved as the file
462 “SNA_classification_patients.tsv”, with columns “Tumor_Sample_Barcode”, “Hugo_Symbol”,

463 “Entrez_Gene_Id”, “Gene”, “Number of hyperactivating SNAs”, “Number of inactivating SNAs”,

464 “Number of SNAs with unclear role”, “Number of passenger SNAs”.

465

466 Using this file, the sum of all alterations in all patients was calculated for each gene. Genes

467 containing only SNAs with unclear role (likely, noncoding genes) were removed, also from

468 “SNA_classification_patients.tsv”. Next, the Nonsynonymous SNA Enrichment Index (NSEI) was

469 calculated for each gene as

$$\boldsymbol{NSEI} = \frac{\textbf{Number of hyperactivating SNAs} + \textbf{Number of inactivating SNAs} + \mathbf{1}}{\textbf{Number of passenger SNAs} + \mathbf{1}}$$

470 and the Hyperactivating to Inactivating SNA Ratio (HISR) was calculated for each gene as

$$\boldsymbol{HISR} = \frac{\textbf{Number of hyperactivating SNAs} + \mathbf{1}}{\textbf{Number of inactivating SNAs} + \mathbf{1}}$$

471 Genes for which the sum of hyperactivating, inactivating and passenger SNAs was less than 10

472 were removed to ensure sufficient precision of NSEI and HISR calculation, and the results were

473 saved as “SNA_classification_genes_NSEI_HISR.tsv”.

474

475 Using the file “SNA_classification_patients.tsv”, the gene-patient matrix “SNA_matrix.tsv” was

476 constructed, encoding the “Number of hyperactivating SNAs”, “Number of inactivating SNAs”,

477 “Number of SNAs with unclear role” and “Number of passenger SNAs” as one number

478 separated by dots (e.g. “2.0.1.1”). If data for a given gene were absent in a given patient, it was

479 encoded as “0.0.0.0”. By drawing statuses randomly with replacement (bootstrapping) from

480 any cell of “SNA_matrix.tsv” 10000 times for each patient, the matrix file

481 “SNA_matrix_bootstrapped.tsv” was created. The sums of statuses in

482 “SNA_matrix_bootstrapped.tsv” were calculated for each iteration separately, and then the

483 corresponding NSEI and HISR indices were calculated and the results were saved as

484 “SNA_bootstrapped_NSEI_HISR.tsv”. Null hypothesis P-values were calculated for each

485 iteration as the number of NSEI values higher than a given iteration's NSEI value and divided by
486 10000. The histogram of bootstrapped p-values was plotted to check for the uniformity of the
487 null hypothesis p-value distribution.
488
489 P-value for each gene was calculated as the number of NSEI values in
490 "SNA_bootstrapped_NSEI_HISR.tsv" more extreme than its NSEI value in
491 "SNA_classification_genes_NSEI_HISR.tsv" and divided by 10000. The results were saved as
492 "SNA_classification_genes_NSEI_HISR_Pvalues.tsv". Benjamini–Hochberg procedure with
493 FDR(Q)=5% was applied to P-values in "SNA_classification_genes_NSEI_HISR_Pvalues.tsv", and
494 genes that pass were saved as "SNA_driver_gene_list_FDR5.tsv".
495
496 We named this algorithm SNADRIF (SNA DRIver Finder) and created a Github repository:
497 https://github.com/belikov-av/SNADRIF. The package used to generate data in this article is
498 available as **S7 Files**.
499
500 Driver prediction algorithms sources and benchmarking
501 Lists of driver genes and mutations predicted by various algorithms (**Table 2**) applied to
502 PanCanAtlas data were downloaded from https://gdc.cancer.gov/about-
503 data/publications/pancan-driver (2020plus, CompositeDriver, DriverNet, HotMAPS,
504 OncodriveFML), https://karchinlab.github.io/CHASMplus (CHASMplus), as well as received by
505 personal communication from Francisco Martínez-Jiménez,
506 Institute for Research in Biomedicine, Barcelona, francisco.martinez@irbbarcelona.org
507 (dNdScv, OncodriveCLUSTL, OncodriveFML). All genes and mutations with q-value > 0.05 were
508 removed. Additionally, a consensus driver gene list from 26 algorithms applied to PanCanAtlas
509 data [8] was downloaded from https://www.cell.com/cell/fulltext/S0092-8674(18)30237-X and

510 a COSMIC Cancer Gene Census (CGC) Tier 1 gene list [14] was downloaded from

511 https://cancer.sanger.ac.uk/cosmic/census?tier=1. Only genes affected by somatic SNAs and

512 CNAs present in the TCGA cancer types were used for further analyses from the CGC list. Cancer

513 type names in the CGC list were manually converted to the closest possible TCGA cancer type

514 abbreviation. Entrez Gene IDs were identified for each gene using HUGO Symbol and external

515 database

516 ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz.

517

518 **Table 2. Driver prediction algorithms.**

Name	Ref.	Repository	Level	Principles
20/20plus	[6]	https://github.com/KarchinLab/2020plus	gene	Machine learning, trained on Cancer Genome Landscapes (20/20 rule); Nonsynonymous/Synonymous, clustering, conservation (uses UCSC's 46-way vertebrate alignment and SNVBox), impact (uses VEST), network (uses BioGrid), expression, chromatin, replication (uses MutSigCV)
ANDRIF	This paper	https://github.com/belikov-av/ANDRIF	Chromosomal arm, chromosome	Recurrence
CHASMplus	[7]	https://github.com/KarchinLab/CHASMplus	mutation	Machine learning, trained on TCGA; clustering (uses HotMAPS 1D), conservation (uses UCSC Multiz-100-way and SNV box), network (uses Interactome Insider)
CompositeDriver	[8]	https://github.com/mil2041/CompositeDriver	gene	Recurrence, impact (uses FunSeq2)
dNdScv	[2]	https://github.com/im3sanger/dndscv	gene	Nonsynonymous/Synonymous
DriverNet	[9]	https://github.com/shahcompbio/drivernet https://bioconductor.org/packages/release/bioc/html/DriverNet.html	gene	Network (uses MGSA and a human functional protein interaction network), impact (uses gene expression outliers)
HotMAPS	[10]	https://github.com/karchinlab/HotMAPS	mutation	3D clustering (uses Protein Data Bank and ModPipe)
OncodriveCLUSTL	[11]	http://bbglab.irbbarcelona.org/oncodriveclustl/analysis https://bitbucket.org/bbglab/oncodriveclustl/src/master/	gene	Clustering
OncodriveFML	[12]	http://bbglab.irbbarcelona.org/oncodrivefml/analysis https://bitbucket.org/bbglab/o	gene	Recurrence, Impact (uses CADD and RNAsnp)

		ncodrivefml/src/master/		
SNADRIF	This paper	https://github.com/belikov-av/SNADRIF	gene	Nonsynonymous/Synonymous
Bailey et al, 2018	[8]	https://www.cell.com/cell/fulltext/S0092-8674(18)30237-X	gene	Consensus driver gene list from 26 algorithms applied to PanCanAtlas data
COSMIC Cancer Gene Census	[14]	https://cancer.sanger.ac.uk/cosmic/census?tier=1	gene	Manually curated list of cancer driver genes, current "gold standard"

The sensitivity of algorithms was assessed as the percentage of genes in a positive control list that were predicted as drivers by an algorithm, because Sensitivity=True positives/(True positives + False negatives). Three positive control lists were used – CGC Tier 1 genes affected by somatic SNAs or CNAs in TCGA cancer types, a list of genes identified by at least two of all our sources (including CGC and Bailey), and a list of genes identified by at least three of all our sources (including CGC and Bailey). Sensitivity was assessed separately for algorithms applied to individual cancer types as the percentage of gene-cohort pairs in a positive control list that were matched by gene-cohort pairs predicted by an algorithm.

Conversion of population-level data to patient-level data

For lists of driver *genes*, all entries from the file "mc3.v0.2.8.PUBLIC_primary_whitelisted_Entrez.tsv" were removed except those that satisfied the following conditions simultaneously: "Entrez Gene ID" matches the one in the driver list; cancer type (identified by matching "Tumor_Sample_Barcode" with "bcr_patient_barcode" and "acronym" in "clinical_PANCAN_patient_with_followup.tsv") matches "cohort" in the driver list or the driver list is for pancancer analysis; "Variant_Classification" column contains one of the following values: "De_novo_Start_InFrame", "Frame_Shift_Del", "Frame_Shift_Ins", "In_Frame_Del", "In_Frame_Ins", "Missense_Mutation", "Nonsense_Mutation", "Nonstop_Mutation", "Translation_Start_Site".

540 For lists of driver *mutations*, the procedures were the same, except that Ensembl Transcript ID
541 and nucleotide/amino acid substitution were used for matching instead of Entrez Gene ID.
542 These data (only columns "TCGA Barcode", "HUGO Symbol", "Entrez Gene ID") were saved as
543 "AlgorithmName_output_SNA.tsv".
544
545 Additionally, all entries from the file
546 "ISAR_GISTIC.all_thresholded.by_genes_primary_whitelisted.tsv" were removed except those
547 that satisfied the following conditions simultaneously: "Locus ID" matches "Entrez Gene ID" in
548 the driver list; cancer type (identified by matching Tumor Sample Barcode with
549 "bcr_patient_barcode" and "acronym" in "clinical_PANCAN_patient_with_followup.tsv")
550 matches "cohort" in the driver list or the driver list is for pancancer analysis; CNA values are
551 "2", "1", "-1" or "-2". These data were converted from the matrix to a list format (with columns
552 "TCGA Barcode", "HUGO Symbol", "Entrez Gene ID") and saved as
553 "AlgorithmName_output_CNA.tsv".
554
555 Finally, the files "AlgorithmName_output_SNA.tsv" and "AlgorithmName_output_CNA.tsv"
556 were combined, duplicate TCGA Barcode-Entrez Gene ID pairs were removed, and the results
557 saved as "AlgorithmName_output.tsv".
558
559 Driver event classification and analysis
560 The file "Clinical with Follow-up - clinical_PANCAN_patient_with_followup.tsv" was
561 downloaded from https://gdc.cancer.gov/node/905/. All patients with "icd_o_3_histology"
562 different from XXXX/3 (primary malignant neoplasm) were removed, as well as all patients not
563 simultaneously present in the following three files:
564 "mc3.v0.2.8.PUBLIC_primary_whitelisted_Entrez.tsv",

565 “ISAR_GISTIC.all_thresholded.by_genes_primary_whitelisted.tsv” and

566 “Primary_whitelisted_arms.tsv”. The resulting file was saved as

567 “clinical_PANCAN_patient_with_followup_primary_whitelisted.tsv”.

568

569 Several chosen “AlgorithmName_output.tsv” files were combined and all TCGA Barcode-Entrez

570 Gene ID pairs not present in at least two output files were removed. Entries with TCGA

571 Barcodes not present in “clinical_PANCAN_patient_with_followup_primary_whitelisted.tsv”

572 were removed as well. Matching “Number of hyperactivating SNAs” and “Number of

573 inactivating SNAs” for each TCGA Barcode-Entrez Gene ID pair were taken from the

574 “SNA_classification_patients.tsv” file, in case of no match zeros were written. Matching HISR

575 value was taken from “SNA_classification_genes_NSEI_HISR.tsv” for each Entrez Gene ID, in

576 case of no match empty cell was left. Matching CNA status was taken from

577 “ISAR_GISTIC.all_thresholded.by_genes_primary_whitelisted_RNAfiltered.tsv” for each TCGA

578 Barcode-Entrez Gene ID pair, in case of no match zero was written.

579

580 Each TCGA Barcode-Entrez Gene ID pair was classified according to the **Table 3**:

581 **Table 3. Driver event classification rules.**

Driver type	Number of nonsynonymous SNAs	Number of inactivating SNAs	HISR	CNA status	Count as … driver event(s)
SNA-based oncogene	≥1	0	>5	0	1
CNA-based oncogene	0	0	>5	1 or 2	1
Mixed oncogene	≥1	0	>5	1 or 2	1
SNA-based tumour suppressor	≥1	≥0	≤5	0	1
CNA-based tumour suppressor	0	0	≤5	-1 or -2	1
Mixed tumour suppressor	≥1	≥0	≤5	-1 or -2	1
Passenger	0	0		0	0
Low-probability driver	All the rest				0

582

583 Results of this classification were saved as “AnalysisName_genes_level2.tsv”.

584

585 Using this file, the number of driver events of each type was counted for each patient.
586 Information on the number of driver chromosome and arm losses and gains for each patient
587 was taken from the files “Chromosome_drivers_FDR5.tsv” and “Arm_drivers_FDR5.tsv”. All
588 patients not present in the files “AnalysisName_genes_level2.tsv”,
589 “Chromosome_drivers_FDR5.tsv” and “Arm_drivers_FDR5.tsv”, but present in the file
590 “clinical_PANCAN_patient_with_followup_primary_whitelisted.tsv”, were added with zero
591 values for the numbers of driver events. Information on the cancer type (“acronym”), gender
592 (“gender”), age (“age_at_initial_pathologic_diagnosis”) and tumor stage (“pathologic_stage”, if
593 no data then “clinical_stage”, if no data then “pathologic_T”, if no data then “clinical_T”) was
594 taken from the file “clinical_PANCAN_patient_with_followup_primary_whitelisted.tsv”. The
595 results were saved as “AnalysisName_patients.tsv”.

596

597 Using the file “AnalysisName_patients.tsv”, the number of patients with each integer total
598 number of driver events from 0 to 100 was counted for each cancer type, also for males and
599 females separately, and histograms were plotted for each cancer type-gender combination.
600 Using the same file “AnalysisName_patients.tsv”, the average number of various types of driver
601 events was calculated for each cancer type, tumour stage, age group, as well as for patients
602 with each integer total number of driver events from 1 to 100. Analyses were performed for
603 total population and for males and females separately, and cumulative histograms were plotted
604 for each file.

605

We named this algorithm PALDRIC (PAtient-Level DRIver Classifier) and created a Github repository: https://github.com/belikov-av/PALDRIC. The package used to generate data in this article is available as **S8 Files**.

Acknowledgements

AVB acknowledges MIPT 5-100 program support for early career researchers.

612 **References**

613 1. Pon JR, Marra MA. Driver and Passenger Mutations in Cancer. Annual Review of
614 Pathology: Mechanisms of Disease. 2015;10: 25–50. doi:10.1146/annurev-pathol-012414-
615 040312

616 2. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal
617 Patterns of Selection in Cancer and Somatic Tissues. Cell. 2017;171: 1029-1041.e21.
618 doi:10.1016/j.cell.2017.09.042

619 3. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer
620 patterns of somatic copy number alteration. Nature Genetics. 2013;45: 1134–1140.
621 doi:10.1038/ng.2760

622 4. Ben-David U, Amon A. Context is everything: aneuploidy in cancer. Nature Reviews
623 Genetics. 2020;21: 44–62. doi:10.1038/s41576-019-0171-x

624 5. Dawson MA, Kouzarides T. Cancer Epigenetics: From Mechanism to Therapy. Cell.
625 2012;150: 12–27. doi:10.1016/j.cell.2012.06.013

626 6. Tokheim CJ, Papadopoulos N, Kinzler KW, Vogelstein B, Karchin R. Evaluating the
627 evaluation of cancer driver genes. Proceedings of the National Academy of Sciences. 2016;113:
628 14330 LP – 14335. doi:10.1073/pnas.1616440113

629 7. Tokheim C, Karchin R. CHASMplus Reveals the Scope of Somatic Missense Mutations
630 Driving Human Cancers. Cell Systems. 2019;9: 9-23.e8. doi:10.1016/j.cels.2019.05.005

631 8. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al.
632 Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. 2018;173: 371-
633 385.e18. doi:10.1016/j.cell.2018.02.060

634 9. Bashashati A, Haffari G, Ding J, Ha G, Lui K, Rosner J, et al. DriverNet: uncovering the
635 impact of somatic driver mutations on transcriptional networks in cancer. Genome Biology.
636 2012;13: R124. doi:10.1186/gb-2012-13-12-r124

637 10. Tokheim C, Bhattacharya R, Niknafs N, Gygax DM, Kim R, Ryan M, et al. Exome-Scale
638 Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. Cancer
639 Research. 2016;76: 3719 LP – 3731. doi:10.1158/0008-5472.CAN-15-3190
640 11. Arnedo-Pac C, Mularoni L, Muiños F, Gonzalez-Perez A, Lopez-Bigas N.
641 OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers.
642 Bioinformatics. 2019;35: 4788–4790. doi:10.1093/bioinformatics/btz501
643 12. Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, López-Bigas N.
644 OncodriveFML: a general framework to identify coding and non-coding regions with cancer
645 driver mutations. Genome Biology. 2016;17: 128. doi:10.1186/s13059-016-0994-0
646 13. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer
647 Genome Landscapes. Science. 2013;339: 1546 LP – 1558. doi:10.1126/science.1235122
648 14. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene
649 Census: describing genetic dysfunction across all human cancers. Nature Reviews Cancer.
650 2018;18: 696–705. doi:10.1038/s41568-018-0060-1
651 15. Tomasetti C, Li L, Vogelstein B. Stem cell divisions, somatic mutations, cancer etiology,
652 and cancer prevention. Science. 2017;355: 1330 LP – 1334. doi:10.1126/science.aaf9011
653 16. Nakada D, Oguro H, Levi BP, Ryan N, Kitano A, Saitoh Y, et al. Oestrogen increases
654 haematopoietic stem-cell self-renewal in females and during pregnancy. Nature. 2014;505:
655 555–558. doi:10.1038/nature12932
656 17. Asselin-Labat M-L, Vaillant F, Sheridan JM, Pal B, Wu D, Simpson ER, et al. Control of
657 mammary stem cell function by steroid hormone signalling. Nature. 2010;465: 798–802.
658 doi:10.1038/nature09027
659 18. Joshi PA, Jackson HW, Beristain AG, Di Grappa MA, Mote PA, Clarke CL, et al.
660 Progesterone induces adult mammary stem cell expansion. Nature. 2010;465: 803–807.
661 doi:10.1038/nature09091

662 19. Folkerd EJ, Dowsett M. Influence of Sex Hormones on Cancer Progression. Journal of
663 Clinical Oncology. 2010;28: 4038–4044. doi:10.1200/JCO.2009.27.4290
664 20. Lyu J, Li JJ, Su J, Peng F, Chen YE, Ge X, et al. DORGE: Discovery of Oncogenes and tumoR
665 suppressor genes using Genetic and Epigenetic features. Science Advances. 2020;6: eaba6784.
666 doi:10.1126/sciadv.aba6784
667 21. Belikov AV. The number of key carcinogenic events can be predicted from cancer
668 incidence. Scientific Reports. 2017;7. doi:10.1038/s41598-017-12448-7
669 22. Belikov AV, Vyatkin A, Leonov SV. The Erlang distribution approximates the age
670 distribution of incidence of childhood and young adulthood cancers. Coates P, editor. PeerJ.
671 2021;9: e11976. doi:10.7717/peerj.11976
672 23. Zhang M, Jang H, Nussinov R. PI3K Driver Mutations: A Biophysical Membrane-Centric
673 Perspective. Cancer Research. 2021;81: 237 LP – 247. doi:10.1158/0008-5472.CAN-20-0911
674

675 **S1 Files. Benchmarking of driver prediction algorithms.**

676 **S2 Files. Output from PALDRIC algorithm applied to pancancer dataset.**

677 **S3 Files. Output from PALDRIC algorithm applied to individual cohorts.**

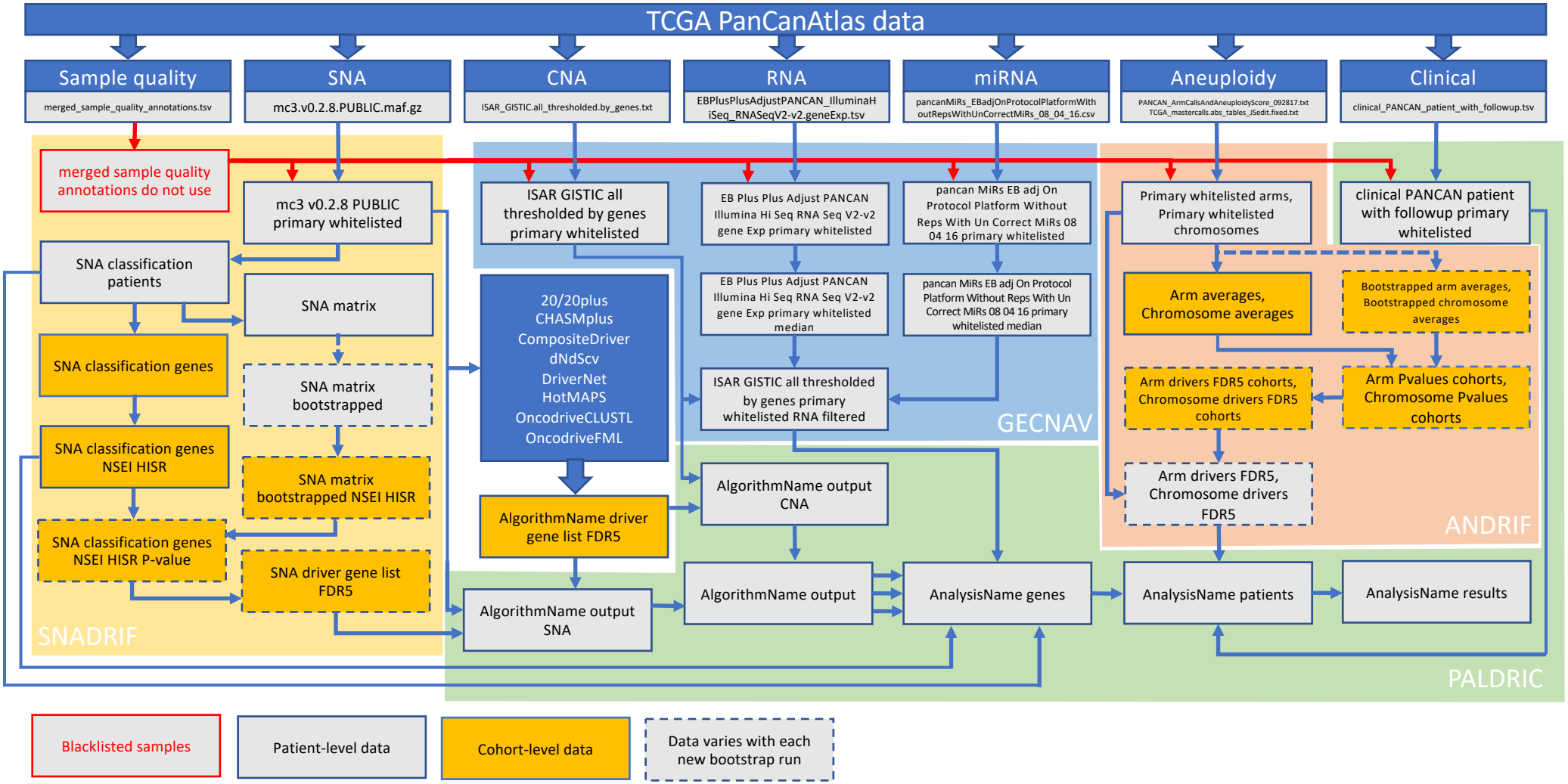
678 **S4 Files. Statistical significance testing for Figures 3, 4, 5.**

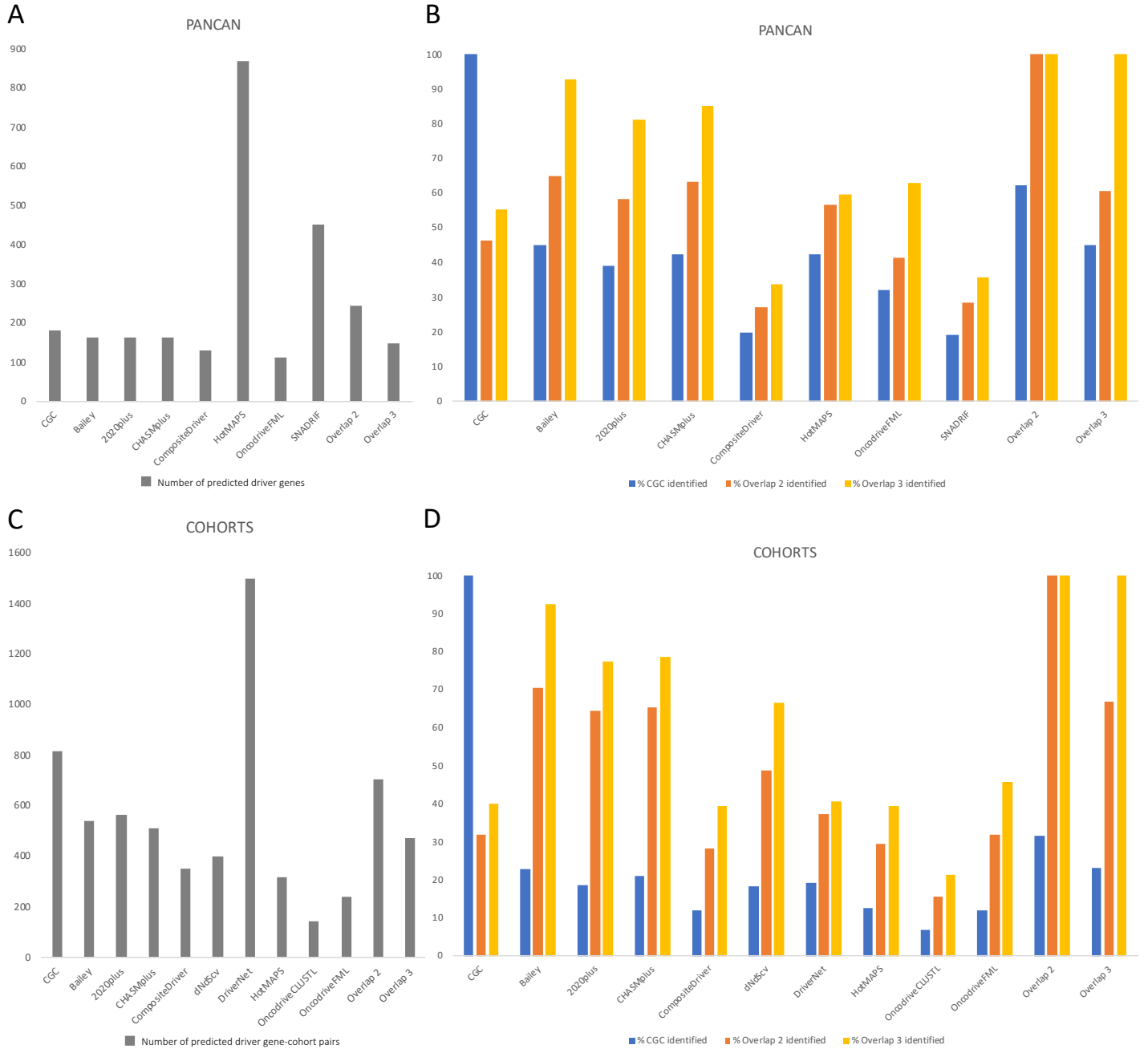
679 **S5 Files. GECNAV package.**

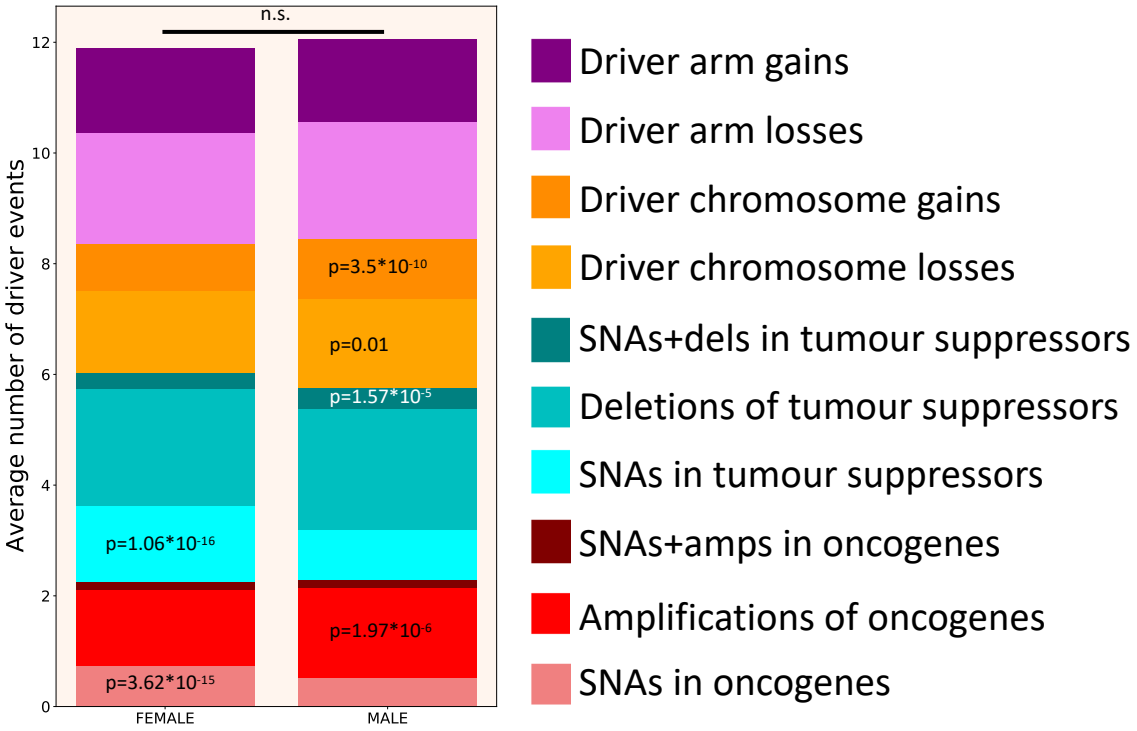
680 **S6 Files. ANDRIF package.**

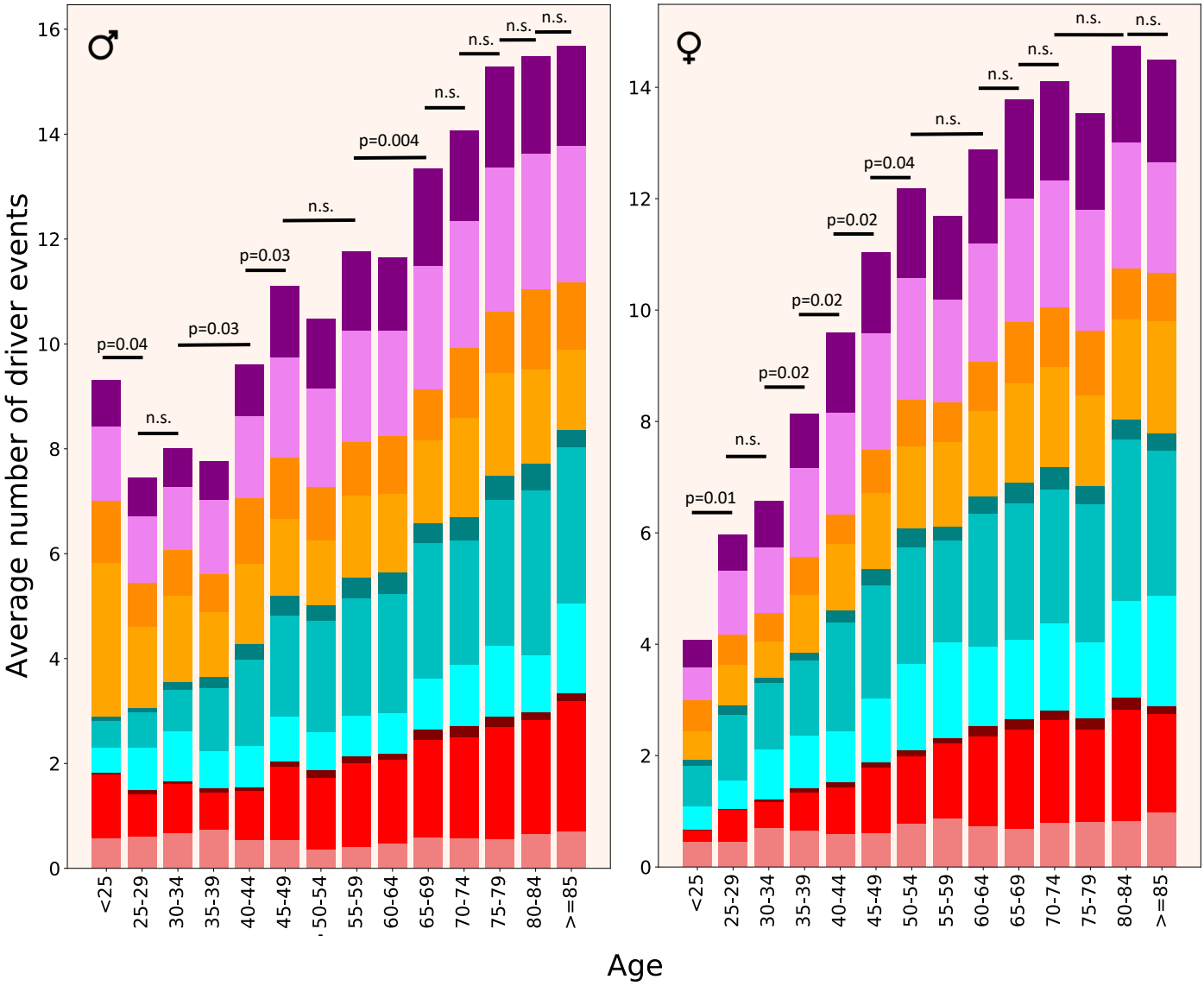
681 **S7 Files. SNADRIF package.**

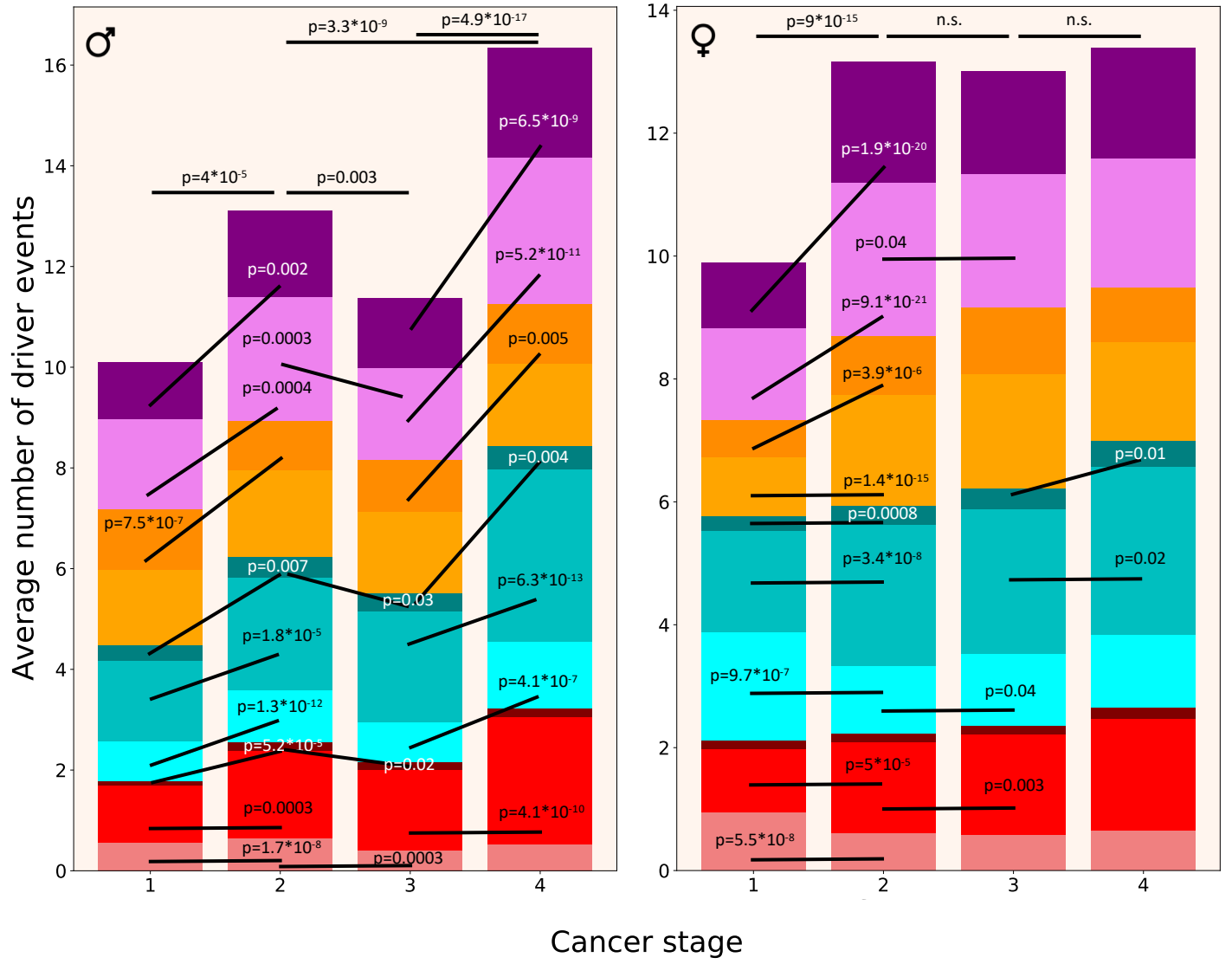
682 **S8 Files. PALDRIC package.**

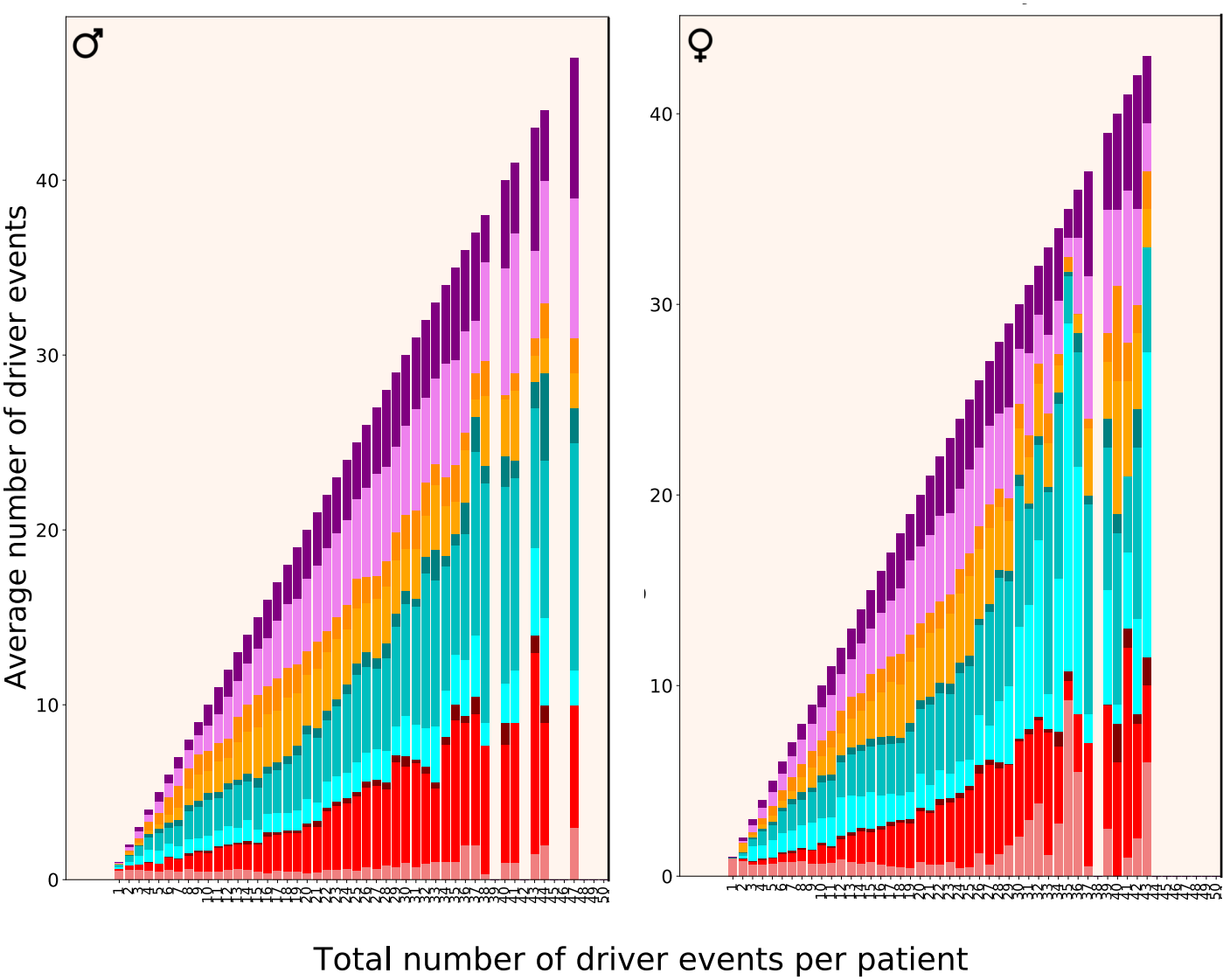

TCGA PanCanAtlas data
Sample quality
merged_sample_quality_annotations.tsv
SNA
mc3.v0.2.8.PUBLIC.maf.gz
CNA
ISAR_GISTIC.all_thresholded.by_genes.txt
RNA
EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2-v2.geneExp.tsv
miRNA
pancanMiRs_EBadjOnProtocolPlatformWithoutRepsWithUnCorrectMiRs_08_04_16.csv
Aneuploidy
PANCAN_ArmCallsAndAneuploidyScore_092817.txt
TCGA_mastercalls.abs_tables_JSedit.fixed.txt
Clinical
clinical_PANCAN_patient_with_followup.tsv
merged sample quality annotations do not use
mc3 v0.2.8 PUBLIC primary whitelisted
ISAR GISTIC all thresholded by genes primary whitelisted
EB Plus Plus Adjust PANCAN Illumina Hi Seq RNA Seq V2-v2 gene Exp primary whitelisted
pancan MiRs EB adj On Protocol Platform Without Reps With Un Correct MiRs 08 04 16 primary whitelisted
Primary whitelisted arms, Primary whitelisted chromosomes
clinical PANCAN patient with followup primary whitelisted
SNA classification patients
SNA matrix
20/20plus
CHASMplus
CompositeDriver
dNdScv
DriverNet
HotMAPS
OncodriveCLUSTL
OncodriveFML
EB Plus Plus Adjust PANCAN Illumina Hi Seq RNA Seq V2-v2 gene Exp primary whitelisted median
pancan MiRs EB adj On Protocol Platform Without Reps With Un Correct MiRs 08 04 16 primary whitelisted median
Arm averages, Chromosome averages
Bootstrapped arm averages, Bootstrapped chromosome averages
SNA classification genes
SNA matrix bootstrapped
ISAR GISTIC all thresholded by genes primary whitelisted RNA filtered
GECNAV
Arm drivers FDR5 cohorts, Chromosome drivers FDR5 cohorts
Arm Pvalues cohorts, Chromosome Pvalues cohorts
SNA classification genes NSEI HISR
SNA matrix bootstrapped NSEI HISR
AlgorithmName driver gene list FDR5
AlgorithmName output CNA
Arm drivers FDR5, Chromosome drivers FDR5
ANDRIF
SNA classification genes NSEI HISR P-value
SNA driver gene list FDR5
AlgorithmName output SNA
AlgorithmName output
AnalysisName genes
AnalysisName patients
AnalysisName results
SNADRIF
PALDRIC
Blacklisted samples
Patient-level data
Cohort-level data
Data varies with each new bootstrap run


A
PANCAN
900
800
700
600
500
400
300
200
100
0
CGC
Bailey
2020plus
CHASMplus
CompositeDriver
HotMAPS
OncodriveFML
SNADRIF
Overlap 2
Overlap 3
Number of predicted driver genes
B
PANCAN
100
90
80
70
60
50
40
30
20
10
0
CGC
Bailey
2020plus
CHASMplus
CompositeDriver
HotMAPS
OncodriveFML
SNADRIF
Overlap 2
Overlap 3
% CGC identified
% Overlap 2 identified
% Overlap 3 identified
C
COHORTS
1600
1400
1200
1000
800
600
400
200
0
CGC
Bailey
2020plus
CHASMplus
CompositeDriver
dNdScv
DriverNet
HotMAPS
OncodriveCLUSTL
OncodriveFML
Overlap 2
Overlap 3
Number of predicted driver gene-cohort pairs
D
COHORTS
100
90
80
70
60
50
40
30
20
10
0
CGC
Bailey
2020plus
CHASMplus
CompositeDriver
dNdScv
DriverNet
HotMAPS
OncodriveCLUSTL
OncodriveFML
Overlap 2
Overlap 3
% CGC identified
% Overlap 2 identified
% Overlap 3 identified


n.s.
Average number of driver events
0
2
4
6
8
10
12
p=3.5*10-10
p=0.01
p=1.57*10-5
p=1.06*10-16
p=1.97*10-6
p=3.62*10-15
FEMALE
MALE
Driver arm gains
Driver arm losses
Driver chromosome gains
Driver chromosome losses
SNAs+dels in tumour suppressors
Deletions of tumour suppressors
SNAs in tumour suppressors
SNAs+amps in oncogenes
Amplifications of oncogenes
SNAs in oncogenes

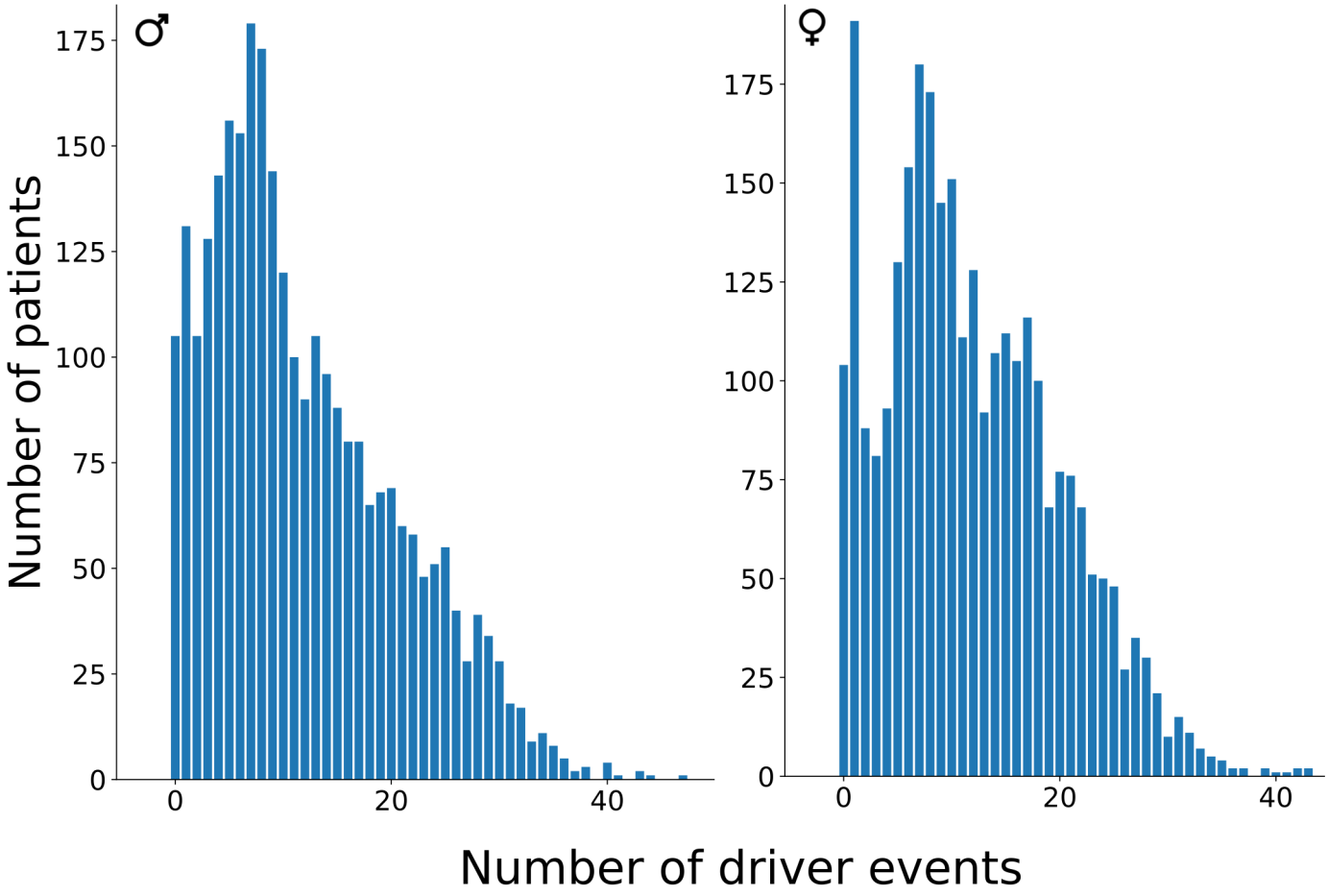

♂
♀
Average number of driver events
Age
<25
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80-84
>=85
p=0.04
n.s.
p=0.03
p=0.03
n.s.
p=0.004
n.s.
n.s.
n.s.
n.s.
p=0.01
n.s.
p=0.02
p=0.02
p=0.04
n.s.
n.s.
n.s.
n.s.
n.s.


♂
♀
Average number of driver events
Cancer stage
p=4*10-5
p=0.003
p=3.3*10-9
p=4.9*10-17
p=0.002
p=0.0003
p=0.0004
p=7.5*10-7
p=0.007
p=1.8*10-5
p=1.3*10-12
p=5.2*10-5
p=0.0003
p=1.7*10-8
p=0.03
p=0.02
p=0.0003
p=6.5*10-9
p=5.2*10-11
p=0.005
p=0.004
p=6.3*10-13
p=4.1*10-7
p=4.1*10-10
p=9*10-15
n.s.
n.s.
p=1.9*10-20
p=0.04
p=9.1*10-21
p=3.9*10-6
p=1.4*10-15
p=0.0008
p=3.4*10-8
p=9.7*10-7
p=5*10-5
p=5.5*10-8
p=0.04
p=0.003
p=0.01
p=0.02


♂
♀
Average number of driver events
Total number of driver events per patient
0
10
20
30
40

♂
Average number of driver events
0
5
10
15
20
ACC
BLCA
BRCA
CHOL
COAD
DLBC
ESCA
GBM
HNSC
KICH
KIRC
KIRP
LGG
LIHC
LUAD
LUSC
MESO
PAAD
PCPG
PRAD
READ
SARC
SKCM
STAD
TGCT
THCA
THYM
UVM
PANCAN

♀
Average number of driver events
0
5
10
15
20
25
ACC
BLCA
BRCA
CESC
CHOL
COAD
DLBC
ESCA
GBM
HNSC
KICH
KIRC
KIRP
LGG
LIHC
LUAD
LUSC
MESO
OV
PAAD
PCPG
READ
SARC
SKCM
STAD
THCA
THYM
UCEC
UCS
UVM
PANCAN


♂
♀
Number of patients
Number of driver events
0
25
50
75
100
125
150
175
0
20
40