RESEARCH ARTICLE

Individualized discovery of rare cancer drivers in global network context

Iurii Petrov[1,2], **Andrey Alexeyenko**[1,2,3]*

[1]Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden; [2]Science for Life Laboratory, Solna, Sweden; [3]Evi-networks, enskild konsultföretag, Huddinge, Sweden

Abstract Late advances in genome sequencing expanded the space of known cancer driver genes several-fold. However, most of this surge was based on computational analysis of somatic mutation frequencies and/or their impact on the protein function. On the contrary, experimental research necessarily accounted for functional context of mutations interacting with other genes and conferring cancer phenotypes. Eventually, just such results become 'hard currency' of cancer biology. The new method, NEAdriver employs knowledge accumulated thus far in the form of global interaction network and functionally annotated pathways in order to recover known and predict novel driver genes. The driver discovery was individualized by accounting for mutations' co-occurrence in each tumour genome – as an alternative to summarizing information over the whole cancer patient cohorts. For each somatic genome change, probabilistic estimates from two lanes of network analysis were combined into joint likelihoods of being a driver. Thus, ability to detect previously unnoticed candidate driver events emerged from combining individual genomic context with network perspective. The procedure was applied to 10 largest cancer cohorts followed by evaluating error rates against previous cancer gene sets. The discovered driver combinations were shown to be informative on cancer outcome. This revealed driver genes with individually sparse mutation patterns that would not be detectable by other computational methods and related to cancer biology domains poorly covered by previous analyses. In particular, recurrent mutations of collagen, laminin, and integrin genes were observed in the adenocarcinoma and glioblastoma cancers. Considering constellation patterns of candidate drivers in individual cancer genomes opens a novel avenue for personalized cancer medicine.

***For correspondence:** andrej.alekseenko@scilifelab.se

Competing interest: The authors declare that no competing interests exist.

Funding: See page 23

Received: 17 September 2021
Preprinted: 05 October 2021
Accepted: 20 May 2022
Published: 20 May 2022

Reviewing Editor: C Daniela Robles-Espinoza, International Laboratory for Human Genome Research, Mexico

© Copyright Petrov and Alexeyenko. This article is distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use and redistribution provided that the original author and source are credited.

Editor's evaluation

In this work, Petrov and Alexeyenko present a novel network-based method to infer cancer driver genes that is not based on frequency of mutations, NEAdriver, and evaluate its performance across a large dataset. This manuscript addresses a topic of high interest in the cancer genomics community and is a welcome addition to the literature.

Introduction

Carcinogenesis is a complex, multi-step process, during which cellular genomes accumulate new, somatic alterations which might also interact with e.g. germline variants. Mutations that cause or facilitate cancer initiation and progression are called drivers. On the other hand, many mutations occur spuriously due to impairment of chromosome maintenance, replication errors etc. Therefore, a cancer cell genome usually represents a mixture of driver and passenger mutations (***Torkamani et al., 2009***). Apart from the boosting genome instability that generates passengers as well as additional drivers, cancer cells should acquire selective advantages, such as apoptosis evasion, unconstrained

proliferation, or survival in low-oxygen environment which correspond to the hallmarks of cancer (**Hanahan and Weinberg, 2011**). Given the avalanche of new data from cancer genome sequencing, it became possible to complement earlier known, 'core' cancer gene sets with multitudes of computationally inferred drivers.

The existing computational approaches to cancer driver discovery can be classified into three major method groups, possibly combined within a certain implementation:

1. Mutation frequency analyses, based on the idea that driver genes appear mutated more often than expected by chance (**Lawrence et al., 2014**);(**Mermel et al., 2011**). In order to keep discovering novel drivers, frequency methods should capture increasingly more rare events (**Vogelstein et al., 2013**) – which is limited by practically achievable genomics dataset sizes. As an example, MuSiC driver analysis included only point mutations (PM) that occurred in more than 5% of tumours (**Dees et al., 2012**). A close-to-comprehensive frequency analysis might require 600–5000 samples per tumour type, depending on background mutation frequency (**Lawrence et al., 2015**). Thus, despite all the advancements, cancer sequencing often fails to identify any driver events in a certain cancer genome.
2. Evaluation of functional impact of sequence alterations using protein structural information, physicochemical features, evolutionary conservation etc. (**Reva et al., 2011**);(**Sim et al., 2012**);(**Adzhubei et al., 2010**) Such methods might also include frequency analyses and were often trained on smaller sets of best known cancer genes (**Martelotto et al., 2014**) which might lead to overfitting. Although some positive correlation with higher mutation frequency has been demonstrated (**Gnad et al., 2013**), predictions by different methods often disagreed even for most studied genes (**Tamborero et al., 2013**).
3. Commonality of protein function to disease genes established via expert judgement or computational analysis of literature associations (**Jimenez-Sanchez et al., 2001**) and global gene network context (**Torkamani and Schork, 2009**; **Tranchevent et al., 2011**; **Doncheva et al., 2012**). In contrast to the approaches described above, this 'guilt-by-association' (GBA) methodology (**Oliver, 2000**) did not require information on mutations per se and could thus be applied to all known genes. Most commonly, likelihood of a general function such as cancer 'driverness' was assigned by a GBA algorithm alone which, when applied to all the genes, generated thousands of predictions with prohibitively high false positive rates (FPR).

A particular challenge would be to identify drivers among gene copy number alterations (CNA), which may encompass longer chromosomal regions with multiple genes were gained or lost at once, 'competing' for a driver role assignment. Therefore, CNA genes were often excluded from the analyses described above.

Network analysis is an important tool for cancer driver gene discovery: it not only implements the GBA principle, but also assesses genomic events by employing the network-defined entities, such as modules and pathways. Non-biological algorithms of network analysis, such as PageRank (**Page et al., 1999**) and Random Walk with Restart (RWR), exist since long ago and were adopted, usually with minimal or no changes, by bioinformatics frameworks (**Erten et al., 2011**; **Fang and Gough, 2014**; **Köhler et al., 2008**; **Ozturk et al., 2018**; **Winter et al., 2012**). For a combination of natural and historical reasons, interpretation of these algorithms tend to focus on network hubs, which could miss novel disease genes with lower node degrees (**Barabási et al., 2011**). Furthermore, GBA methods might be work regardless of predefined pathways, for example consider expression correlates (**Torkamani and Schork, 2009**), physical interactions (**Ciriello et al., 2012**), or shared annotations (**Freudenberg and Propping, 2002**). Such methods though, when applied to either all or to frequently mutated genes, would either suffer from the high false positive rate or miss rare drivers. Another solution was offered by the method of network enrichment analysis (NEA)(**Alexeyenko et al., 2012**), where network connectivity is normalized by gene node degrees, which allowed studying genes poorly covered with experimental data. The other advantage of NEA is its high sensitivity and robustness due to considering the multitude of edges available in the global network (**Jeggari and Alexeyenko, 2017**; **Franco et al., 2019**). The concept of enrichment, that is detection of signal that prevails over noise is implemented in NEA via counting network edges that connect gene nodes. Significant excess of actual number of edges over expected by chance can distinguish functionally relevant genes, that is drivers differ from passengers by relevant fragments of network connectivity.

The above-mentioned problem of high FPR can be efficiently addressed by combining probabilities from multiple evidence channels. In the presented analysis we did that in a three-pronged way.

First, we reduced FPR by considering only genes altered in a given tumor genome, whereas genes mutated elsewhere were ignored. Second, we employed the idea that driver mutations of mutated gene sets (MGS) in individual samples should be mutually related and identified such cases by network enrichment against each other, that is within MGS. Third, we detected driver roles by summarizing network connectivity of MGS to diverse potentially informative pathways. Since the full set of such pathways was not known in advance, we started from hundreds pathway profiles, followed by feature selection and creation of predictive cohort-specific sparse models. Combining the two predictors decreased FPR even further.

We applied the analysis pipeline to nine largest TCGA cohorts as well as to a newly compiled meta-cohort of medulloblastoma (MB). We evaluate agreement between our and earlier published driver sets, relative contributions of the driver score components, significance, prediction error rates, and prove the method robustness across over a broad range of mutation rates (from very low in MB to very high in skin melanoma) and variable disease aggressiveness. We demonstrate functional relations between driver mutation patterns, gene expression in affected pathways, and patient survival. Finally, the analysis exposed so far underestimated protein categories with individually rare genomic alterations in their members which appeared essential for several cancer hallmarks.

Results

Algorithm outline: two evidence channels for driver prediction

The procedure evaluated likelihood of each genomic alteration reported in MAF or CNA files after level 3 analysis being a driver in the given tumor genome. This was done by considering functional network context in two parallel, independent analysis channels (***Figure 1***):

MutSet channel

Evaluated network enrichment between each altered gene *m* (*m*∈ MGS) and the constellation of all other altered genes *n* (*n*∈ MGS; $n \neq m$). The resulting NEA scores $Z_{m \leftrightarrow MGS}$ accounted for network degrees of the interacting MGS genes and expressed strength of the cumulative interaction compared to a value expected by chance, that is when *m* would be functionally unrelated to the rest of MGS.

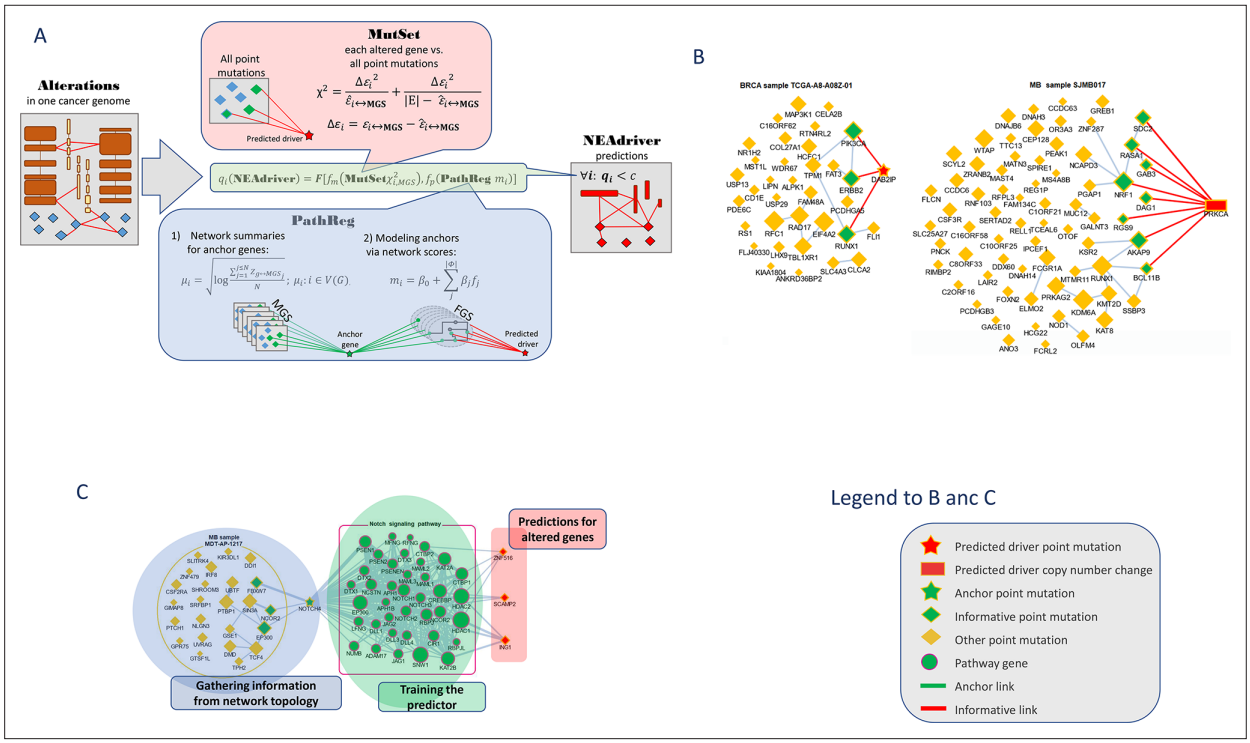

Figure 1. Visualization of NEAdriver analysis. (**A**) Workflow according to the algorithm described in Methods. (**B**) Examples of MutSet analysis to mutation gene sets in two cohorts.(**C**) Example PathReg analysis in MB cohort. Legend to nodes and edges.

PathReg channel

Evaluated likelihood of being a cancer driver in a two-step procedure. First, each of the *N* genes present in the network (N=19035) were characterized – in the same way as in MutSet – by network enrichment scores against each individual MGS. These 'anchor' scores $Z_{i\leftrightarrow MGS}$ were summarized per gene *i* and within each of the 10 cohorts *c*. The reasoning behind this was that if MGSs included some actual (but not explicitly declared) drivers, then potential cancer involvement of a gene could be expressed as a summary of its network interactions over the MGS collection. For a highly scoring gene, this should become evidence of being a driver when it was altered in a given genomic sample. The cohort-specific gene vectors called 'anchor.summary' should already contain all information gathered from genes' interactions with all the MGSs. However, given that both MGSs and the available network were likely incomplete, some genes in anchor.summary vectors would not be fully evaluated. Therefore, we introduced a second step: boosting via pathway enrichment scores. Cohort-specific predictive models were created using anchor.summary vectors as independent variables (being split into training and testing halves, *N'*=*N*/2 and N"=*N*/2). Dependent variables for the models were provided by pre-calculated NEA scores for the same *N* genes against a collection of *P*=320 pathways, generally called functional gene sets (FGS), forming an N×*P* matrix. Then sparse multivariate models (k=31…83 pathways with non-zero coefficients) were obtained via the lasso training procedure under cross-validation and by controlling performance and robustness with error terms and information criteria. Due to big sample sizes (*N'*~10^4), the models reproduced well on the test sets (Spearman rank R between observed and predicted vectors of anchor.summary were 0.62…0.79; Suppl.File 1). The sparse models produced cohort-specific PathReg scores for each of the *N* genes. Using published driver sets as references, performance of the scores was compared to the original anchor.summary, which demonstrated clear superiority of PathReg and thus gain in driver-related information via pathway enrichment. Thus, MutSet estimated driverness strictly in the context of individual cancer genomes, whereas PathReg models were originally derived from individual MGSs and then presented as universal, cohort-specific values. Importantly, neither of the two employed information on mutation frequencies. Enrichment heatmaps for genes that scored highest in PathReg versus pathways included in the models are presented in ***Supplementary file 1***.

Similarly to NetSig5000 method (***Horn et al., 2018***), we also tested the approach of network evaluation of mutations against sets of most frequently mutated genes, but this channel did not yield any further advantage and was not used. The MutSet and PathReg scores were calibrated and converted into *p* and respective *q* (false discovery rate) values. Evidence from MutSet and PathReg was combined under OR condition, that is the resulting product q(MutSet&PathReg)=q(MutSet)*q(PathReg) reported the probability of NOT being a driver despite positive evidence obtained from either channel. In this way, MGSs were reduced to driver gene sets, DGS (***Figure 1A***). For the purposes of further testing, the DGSs were defined at two significance thresholds q(MutSet&PathReg)<0.05 and q(MutSet&PathReg)<0.01. Under the former cutoff the fractions of drivers were 14…67% larger than

Table 1. Fractions of individual alterations and unique genes predicted by NEAdriver.

		BRCA	COAD	GBM	LUAD	LUSC	OV	PAAD	PRAD	SKCM	MB
No. of	genes	17,216	16,553	11,484	17,028	14,853	14,055	14,766	11,660	17,253	18,244
	samples	989	269	284	519	178	461	185	300	346	564
	alterations (PM&CNA)	158,982	115,250	36,230	193,285	71,238	81,719	64,027	33,985	230,159	96,263
Fraction of cases when received q<0.05	PathReg	1.55%	2.23%	3.81%	3.02%	2.94%	3.81%	0.16%	3.44%	4.42%	0.06%
	MutSet	2.95%	3.52%	3.75%	4.88%	3.68%	3.53%	1.21%	2.62%	9.24%	4.48%
	PathReg & MutSet	8.52%	6.81%	10.63%	9.67%	7.72%	9.15%	2.79%	7.74%	13.42%	9.14%
Fraction of cases when received q<0.01	PathReg	0.55%	1.37%	1.24%	2.41%	2.3%	2.16%	0.03%	2.80%	3.17%	0.02%
	MutSet	2.17%	2.77%	2.65%	4.01%	2.57%	2.68%	0.70%	1.80%	8.00%	3.65%
	PathReg & MutSet	7.11%	5.57%	7.81%	8.21%	6.16%	7.17%	1.65%	6%	11.81%	6.98%
No. of genes which received q(PathReg&MutSet)<0.05 in>90% samples		221	343	226	498	334	270	14	180	766	5

under the latter (***Table 1***). The full lists of genes for the ten cohorts are presented in the summary tables (***Supplementary file 8***).

Comparison with alternative gene sets

We first evaluated performance of the method [q(MutSet&PathReg)<0.05] by ability to detect gene members of 11 alternative reference sets, which were either derived from curated resources or published as computational analysis results (***Figure 2A***). Overlaps with the reference sets were mostly significant: 90 out of 110 pairwise comparisons by Fisher's exact test and 61 out 110 by Mann-Whitney test received a Bonferroni-adjusted p-value below 0.05.

In order to see differences between functional landscapes of NEAdriver (the sets of predicted drivers at (q(MutSet&PathReg)<0.05) vs. the alternative sets), we calculated network enrichment scores of each gene set as a whole versus each of 50 hallmark gene sets (***Liberzon et al., 2015***; ***Figure 2B***). The heatmap revealed that NEAdriver detected genes from a different hallmark subspace than most of the computational methods. On the other hand, the NEAdriver predictions often clustered together with the curated sets KEGG05200 'Pathways in cancer' and the union of five 'general' (not cohort-specific) cancer-relevant KEGG gene sets. Similar patterns were observed while looking at the outputs from PathReg and MutSet channels separately, although the former was somewhat closer to the curated sets (Figure Supplements to ***Figure 2B***). Compared to the computational methods, the NEAdriver sets and the curated sets showed higher enrichment in EMT, angiogenesis, and suppressed KRAS signaling, glycolysis, inflammatory response, and hypoxia while depleted in cell cycle, DNA replication/repair, peroxisome as well as MYC and mTOR signaling. For comparison, the computational sets were much more similar to the original, full MGSs (Figure Supplement to ***Figure 2B***), which confirmed that the NEAdriver pattern was functionally specific and distinct rather than reflected the initial mutation composition. We also noted that nearly all the alternative cohorts (except MutSig) abounded in genes with higher network degree, which were likely better known and studied than the genes predicted with NEAdriver. Node degrees of the latter were closer to an average level, as illustrated by comparisons against random gene samples (***Supplementary file 2***). Remarkably, the network-based method NetSig5000 also prioritized genes with higher node degree.

Estimation of discovery rates

The same reference gene sets were used for a more detailed evaluation of NEAdriver error rates. The best combination of true positive and true negative rates was found against the gold standard cohort-specific sets, which were either literature-based or available as KEGG pathways (***Figure 3A and B*** and upper left plots in Figure Supplements to ***Figure 3***), except BRCA cohort, where better results were found for NetSig5000 set which was derived from just this cohort (***Horn et al., 2018***).

Precision was first estimated using the common definition as fraction of true positives among all positives:

$$Precision = \frac{TP}{TP+FP}$$

If applied to the full set of *n* genes, regardless of their mutation status in specific cancer genomes – which corresponded to GBA approach – then these estimates appeared very low and never exceeded 20% at TPR = 10% (upper right plots in Figure Supplements to ***Figure 3***).

A more advanced estimate could be provided by using the hybrid positive predictive value (PPV) formula by John Ioannidis (***Ioannidis, 2005***), where the frequentist terms – error rates of types I and II – were combined with odds (i.e. the ratio of actual drivers versus non-drivers among the mutated genes), which represented a Bayesian component:

$$PPV = \frac{(1-\beta)*R}{(1-\beta)*R+\alpha}$$

The odds could be estimated from the union of all test results on the cohort's MGSs as

$$R = \frac{TP+FN}{FP+TN}$$

By expanding the Bayesian approach, the type I and type II errors would be, respectively: $\alpha = \frac{FP}{FP+TN}$ and $\beta = \frac{FN}{TP+FN}$.

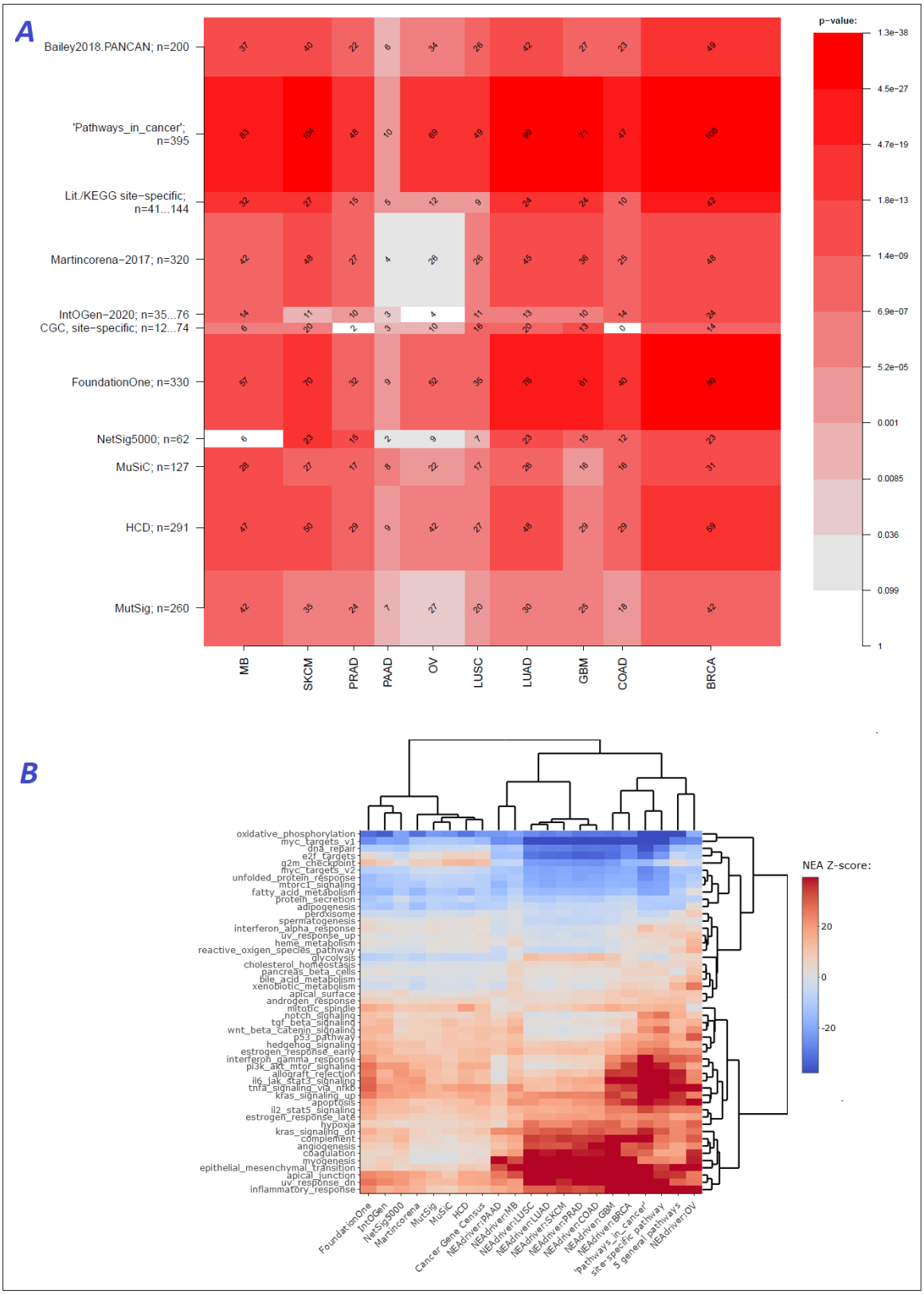

Figure 2. Agreement between NEAdriver and reference gene sets. (**A**) The heatmap matrix elements represent overlap between the cohort-specific sets of predicted NEAdriver gene sets at q(MutSet&PathReg)<0.05 and gene sets from curated resources and alternative methods. Row and column widths are proportional to gene sets sizes. All the reference gene sets had fixed, 'pan-cancer' member sets independent of cancer site, except Cancer Gene Census, IntOGen, and literature/KEGG site-specific sets, for which size ranges are given. B. Network enrichment of the cancer gene sets with

Figure 2 continued on next page

Figure 2 continued

regard to 50 hallmarks (***Liberzon et al., 2015***). NEAdriver sets defined at q(MutSet&PathReg)<0.05 are represented by 165 genes for each cohort, most frequent across its samples(n=165 was chosen for being half of the size of FoundationOne set).

The online version of this article includes the following figure supplement(s) for figure 2:

Figure supplement 1. Network enrichment of the cancer gene sets with regard to 50 hallmarks.

Figure supplement 2. NEAdriver sets defined at q(MutSet)<0.05 and represented by 165 genes for each cohort.

Figure supplement 3. NEAdriver sets defined at q(PathReg)<0.05 and represented by 165 genes for each cohort.

Figure supplement 4. Correlations between from PathReg and MutSet channels.

The value of FP + TN could be approximated as the total number of genes, since the number of drivers (TP and FN) should be negligible compared to that.

However, estimating the TP and FN via reference (gold standard) sets was more challenging, since the source publications and databases never claimed that their gene sets are truly complete. Thus, PPV estimates were particularly sensitive to biases in TP and FN and we therefore tried each of the nine sets. PPV ranged from 30% to 70% at TPR = 10%, but even at TPR = 100% almost never dropped below 20% (***Figure 3C and D*** and all cohorts in Figure Supplements to ***Figure 3***). Again, the best performance was achieved using the literature/KEGG sets (PPV = 44…68% at TPR = 10%).

Since this approach considered any genes not listed in each given set as false findings, the PPV estimates must have been excessively conservative. Therefore, we next investigated the potential of discovering novel drivers using genes collected from site-specific literature or respective KEGG pathways. An alternative, pan-cancer estimate was made with a set of 369 'known cancer genes' (***Martincorena et al., 2017***). Applying the PPV adjustment to these sets under assumption that they were just 50% complete increased the PPV estimate by 20…30% (dotted curves at ***Figure 3E and F***). Recalling that (1 – PPV) is essentially synonymous to false discovery rate (i.e. q-value) allowed us also to compare error rate estimates from the two independent approaches: the gold-standard based PPV versus the continuous NEAdriver q(MutSet&PathReg). Although the relation was not linear over the range PPV = 0…100%, at PPV = 30…70% the cancer site-specific estimates were remarkably close in each of the 10 cohorts. On the other hand, the pan-cancer benchmark of 369 known (mostly computationally) cancer genes estimated NEAdriver q as inflated by 20…40% while assuming that the gene set was complete (solid lines) but well matching the PPV while assuming 50% incompleteness (dotted lines) (***Figure 3E and F*** and Figure Supplements to ***Figure 3***). Therefore, we used NEAdriver q(MutSet&PathReg) for reporting confidence of driver predictions in this work.

We also compared the results to a number of previously suggested network-based methods that considered impact of somatic alterations on the transcriptome: DriverNet (***Bashashati et al., 2012***) and HotNet2 (***Leiserson et al., 2015***) which implemented the cohort level approach as well as SCS (***Guo et al., 2018***), OncoIMPACT (***Bertrand et al., 2015***), and DawnRank (***Hou and Ma, 2014***) which worked at the individualized, single-patient level. The comparison also included naïve frequency-based estimates as provided by Guo and co-authors (***Supplementary file 3***). These publications presented short candidate driver lists combined over all samples. Agreement of the integrated ranks with NEAdriver confidence q(MutSet&PathReg) in the four TCGA cohorts proved to be significant albeit rather weak (Spearman R=0.22…0.30). Further, we focused on lists of top 50 genes from each of the methods. In three cohorts (OV was the exception), a good agreement was found between the methods and NEAdriver (***Supplementary file 4***). Out of top 50 driver lists, between 9 and 41 genes received NEAdriver q(MutSet&PathReg)<0.05 (the overlaps were significant after Bonferroni-adjusted Fisher's exact test p<0.001).

Rates of driver discovery versus mutation frequency and possible confounders

Driver discovery from large-scale genome sequencing might produce false positives due to various biasing factors, such as gene length, transcriptional activity, or DNA replication rate (***Lawrence et al., 2013***). Although NEAdriver was designed to be independent from alteration frequency – in order to be thus more sensitive to rare events – we still performed in-depth analysis of NEAdriver output in relation to such factors and in comparison with alternative methods.

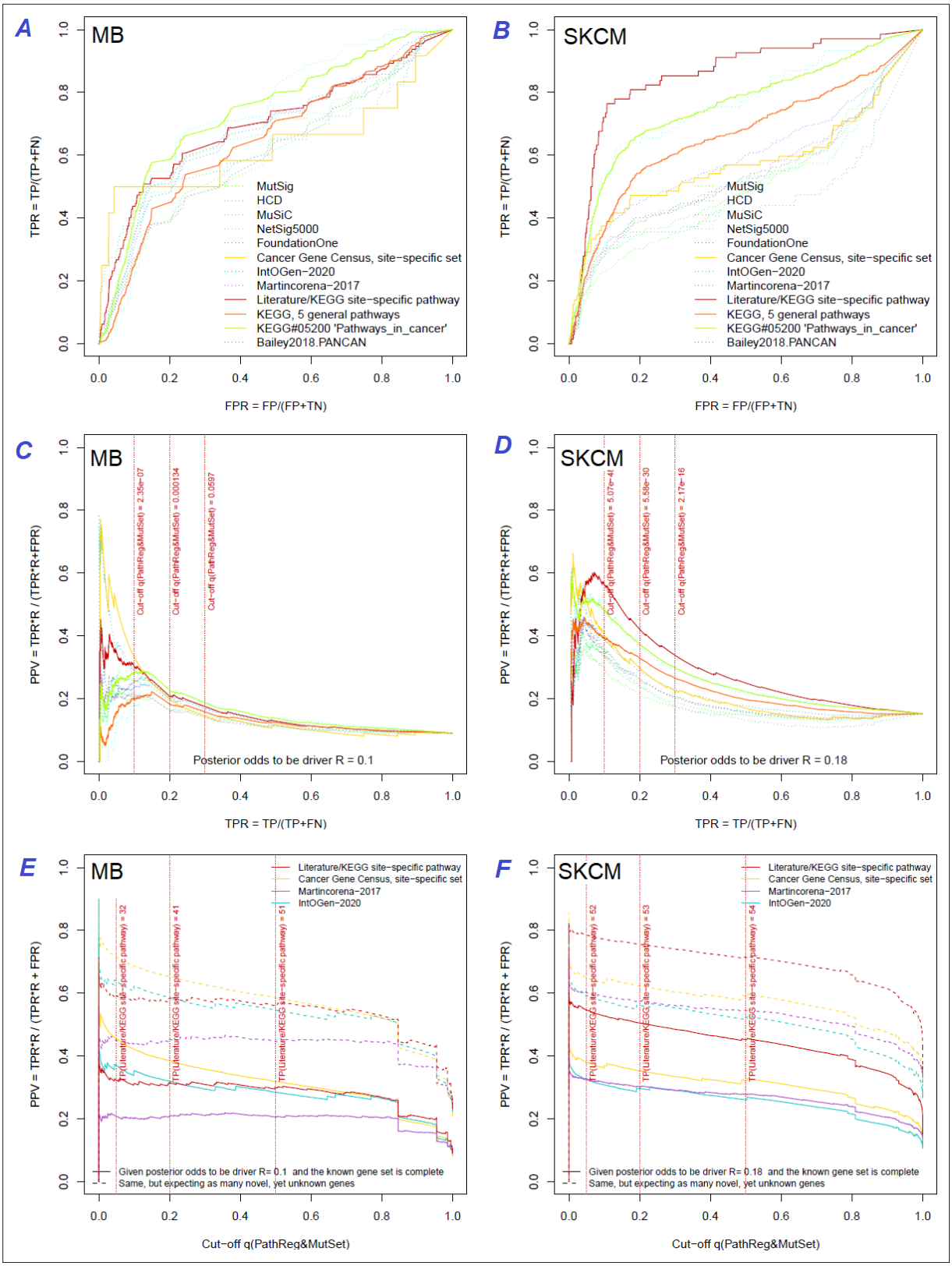

Figure 3. Performance of the new driver prediction evaluated on different benchmarks. Two cohorts with very low versus high passenger mutation load, medulloblastoma (MB: **A,C,E**) and skin cutaneous melanoma (SKCM: **B,D,F**), represent contrast conditions for computational driver discovery. The NEAdriver predictions were quantified by the cumulative statistic q(MutSet&PathReg)<0.05 and matched to reference sets. (**A**) and (**B**): ROC curves in the space of true positive versus false positive rates in the classical definition of 'precision'. (**C**) and (**D**): Precision-recall curves where precision was

Figure 3 continued on next page

Figure 3 continued

calculated via inclusion of odds 'driver/non-driver'.(**E** and **F**) calibration of positive predictive value, PPV against false discovery rate (q~1 – PPV; solid lines) and modeling of PPV in presence of true, but yet unknown drivers (dot-dashed lines) using site-specific and pan-cancer benchmarks. The dotted vertical cutoff lines referto cancer site specific pathway sets, taken from either to the literature or respective KEGG pathway. Cutoffs in (**C**) and (**D**) display q(MutSet&PathReg) values, whereas TP counts in (**E**) and (**F**) are numbers of unique site-specific genes discovered under variable q(MutSet&PathReg) threshold shown at X-axis.

The online version of this article includes the following figure supplement(s) for figure 3:

Figure supplement 1. Performance of the new driver prediction evaluated on different benchmarks.

Figure supplement 2. Performance of the new driver prediction evaluated on different benchmarks, all cohorts.

Figure supplement 3. Performance of the new driver prediction evaluated on different benchmarks, PRAD.

Figure supplement 4. Performance of the new driver prediction evaluated on different benchmarks, PAAD.

Figure supplement 5. Performance of the new driver prediction evaluated on different benchmarks, OV.

Figure supplement 6. Performance of the new driver prediction evaluated on different benchmarks, LUSC.

Figure supplement 7. Performance of the new driver prediction evaluated on different benchmarks, LUAD.

Figure supplement 8. Performance of the new driver prediction evaluated on different benchmarks, GBM.

Figure supplement 9. Performance of the new driver prediction evaluated on different benchmarks, COAD.

Figure supplement 10. Performance of the new driver prediction evaluated on different benchmarks, BRCA.

The genes listed by FoundationOne, MutSig, HCD, MuSiC and in particular by Cancer Gene Census and IntOGen had generally more mutation events per cohort than drivers predicted at q(MutSet&PathReg)<0.05 (***Figure 4***, left panes). The same tendency was observed when comparing copy number altered genes predicted by NEAdriver (***Figure 4—figure supplements 1–10*** to ***Figure 4***). Exceptions could only be found in SKCM cohort (which was not analysed in most of these projects) and in a few cohorts for MutSig (which implemented advanced normalization approaches). Sensitivity of NetSig5000 to rare mutations was comparable to our method – likely due to its network-based approach – but again mostly to genes with higher network degree (see details in ***Figure 4—figure supplements 1–20*** to ***Figure 4***). On the contrary, when mutation frequencies were normalized by coding sequence length, the differences between the methods became less pronounced (***Figure 4***, right panes). This was not surprising, since shorter genes manifest mutations less frequently.

Lawrence and co-authors (***Lawrence et al., 2013***) demonstrated that simply considering mutation frequency per gene without accounting for genomic factors results in multiple false positive associations. Similarly to their Figure 3 , we evaluated influence of suggested confounders, namely gene length, replication rate, expression level, and total mutation burden per sample on the NEAdriver predictions. Each of the linear models included one of these confounders together with number of mutations per cohort per gene as well as being a known driver or a known artifact. While the mentioned ***Figure 3*** by Lawrence and co-authors showed strong correlations of mutation rate versus replication or expression, respective correlations of NEAdriver score were not strong at all (albeit formally significant; absolute values of Kendall tau <0.05; ***Supplementary file 5***). NEAdriver predictions were stronger associated with gene length (absolute values of Kendall tau = 0.07…0.15) – while we also noticed such association in all the alternative gene sets (***Figure 4—figure supplement 21*** to ***Figure 4***). Also, genes of the latter sets contained point mutations more frequently and had higher mutation frequency per b.p. length, both absolute and normalized by genome-specific mutation load. This also characterized the driver set by ***Lawrence et al., 2014***, which would supposedly be least affected by these factors due to inclusion of these covariates in their models. Otherwise, probability of being predicted by NEAdriver after adjustment for gene length and TMB proved to be weak (Kendall tau <0.05 in seven out of ten cohorts; ***Supplementary file 5***). We also note that the replication and expression analyses must vary between tissues and datasets, while being expensive to measure (e.g. the analyses by Lawrence and colleagues used data from less than 100 cell lines).

Finally, we specifically considered 'artefactual' or 'false positives' genes that had cropped up in earlier cancer genome studies and were explicitly listed in later literature (34)(40)(41). We also included olfactory receptor genes as a whole category, supposedly prone to artifacts (although a majority these lacked any network edges and could not produce non-zero NEA scores). Among NEAdriver predictions, the fractions of any artefactual genes were much smaller than the overall false discovery rate

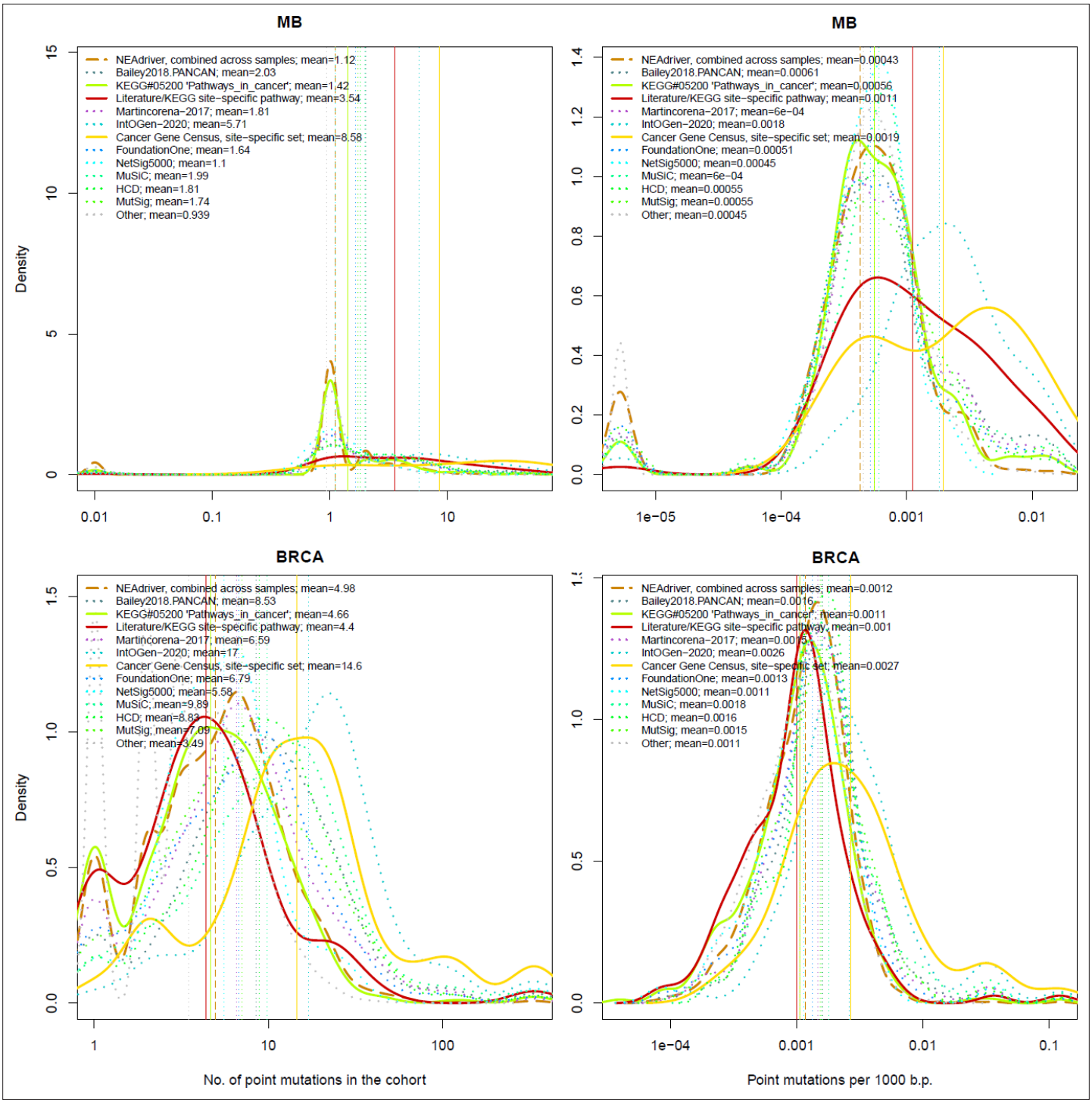

Figure 4. Comparative analysis of point mutation frequency among genes included in cancer gene sets. Density plots shape the distributions in each of the alternative sets, predictions by NEAdriver (q(MutSet&PathReg)<0.05; brown dashed line), and genes not included in any of the above ('other'; gray dotted line). Vertical lines correspond to mean values provided in the legend.

The online version of this article includes the following figure supplement(s) for figure 4:

Figure supplement 1. Comparative analysis of point mutation frequency among genes included in cancer gene sets.

Figure supplement 2. Performance of the new driver prediction evaluated on different benchmarks, SKCM.

Figure supplement 3. Comparative analysis of point mutation frequency among genes included in cancer gene sets, PRAD.

Figure supplement 4. Comparative analysis of point mutation frequency among genes included in cancer gene sets, PAAD.

Figure 4 continued on next page

Figure 4 continued

Figure supplement 5. Comparative analysis of point mutation frequency among genes included in cancer gene sets, OV.

Figure supplement 6. Comparative analysis of point mutation frequency among genes included in cancer gene sets, LUSC.

Figure supplement 7. Comparative analysis of point mutation frequency among genes included in cancer gene sets, LUSC.

Figure supplement 8. Comparative analysis of point mutation frequency among genes included in cancer gene sets, GBM.

Figure supplement 9. Comparative analysis of point mutation frequency among genes included in cancer gene sets, COAD.

Figure supplement 10. Comparative analysis of point mutation frequency among genes included in cancer gene sets, BRCA.

Figure supplement 11. Ten cohort files with density plots shaping mutation frequency distributions for genes in each of the alternative sets, predictions by NEAdriver (brown dashed line), and genes not included in any of the above ('other'; gray dotted line).

Figure supplement 12. Comparative analysis of point mutation frequency among genes included in cancer gene sets, SKCM.

Figure supplement 13. Comparative analysis of point mutation frequency among genes included in cancer gene sets, point mutations, PRAD.

Figure supplement 14. Comparative analysis of point mutation frequency among genes included in cancer gene sets, point mutations, PAAD.

Figure supplement 15. Comparative analysis of point mutation frequency among genes included in cancer gene sets, point mutations, OV.

Figure supplement 16. Comparative analysis of point mutation frequency among genes included in cancer gene sets, point mutations, LUSC.

Figure supplement 17. Comparative analysis of point mutation frequency among genes included in cancer gene sets, point mutations, LUAD.

Figure supplement 18. Comparative analysis of point mutation frequency among genes included in cancer gene sets, point mutations, GBM.

Figure supplement 19. Comparative analysis of point mutation frequency among genes included in cancer gene sets, point mutations, COAD.

Figure supplement 20. Comparative analysis of point mutation frequency among genes included in cancer gene sets, point mutations, BRCA.

Figure supplement 21. Boxplots comparing lengths of genes included in different sets versus rest of known genes.

evaluated via either q-value or PPV as described above. Only between 0.5% and 3% of the predictions were found in the artifact gene lists (upper right legends in new ***Supplementary file 5***). In the cohort-specific linear models (bottom left legends), the Bonferroni-adjusted p-values for the term 'known artifact' were lower than 0.05 only in five cases out of the 30. The artefactual genes are text labelled in the scatterplots when surpassed significance threshold of q=0.05 (1…13 genes per cohort) and marked in the last columns of the summary tables (***Supplementary file 8***). For comparison, these genes made up 1…9% of any other computational set in our analysis.

Novel findings

How many known drivers there are in individual cancer genomes and by how much the new method could expand this space? An earlier computational analysis estimated the number of point driver mutations as two to six per genome (***Kandoth et al., 2013***). In our study – by counting any genes included in the nine alternative sets (N=1434) – the modes (most frequent count values) ranged across the cohorts between *M*=1…3 in MB (known to have very low somatic mutation load) to M=55 in SKCM (having typically thousands mutated genes per sample). For NEAdriver [q(MutSet&PathReg)<0.05], respective values per genome were lower, ranging between *M*=0…1 (MB) to *M*=50 (SKCM) (***Figure 5A***). Overlaps between these two approaches were rather modest (*M*=0…8). In other words, the driver candidates identified by NEAdriver were mostly novel. The overlaps between 'alternative sets' and NEAdriver [q(MutSet&PathReg)<0.01] are also presented for individual cancer genomes (***Figure 5B***). The Jaccard coefficient values, with exceptions of MB and GBM, rarely exceeded 0.3, which confirmed that NEAdriver identified mostly novel genes.

Recalling that respective PPV estimates reached 50% and exceeded 75% when allowing for novel drivers, the predictions appeared fairly confident.

Clustering patient driver sets in pathway space revealed association with survival

One of the goals of tumour molecular profiling is to discover cancer subtypes which would be informative of disease outcome or clinically meaningful otherwise. The driver genes identified with our method were mostly rare and therefore not suitable as stand-alone subtype markers. However, using NEA we could generate 'DGS vs. FGS' scores which summarized signals from various disparate events and thus available for every patient. We explored if DGS profiles in the FGS space could partition the cohorts by differential survival.

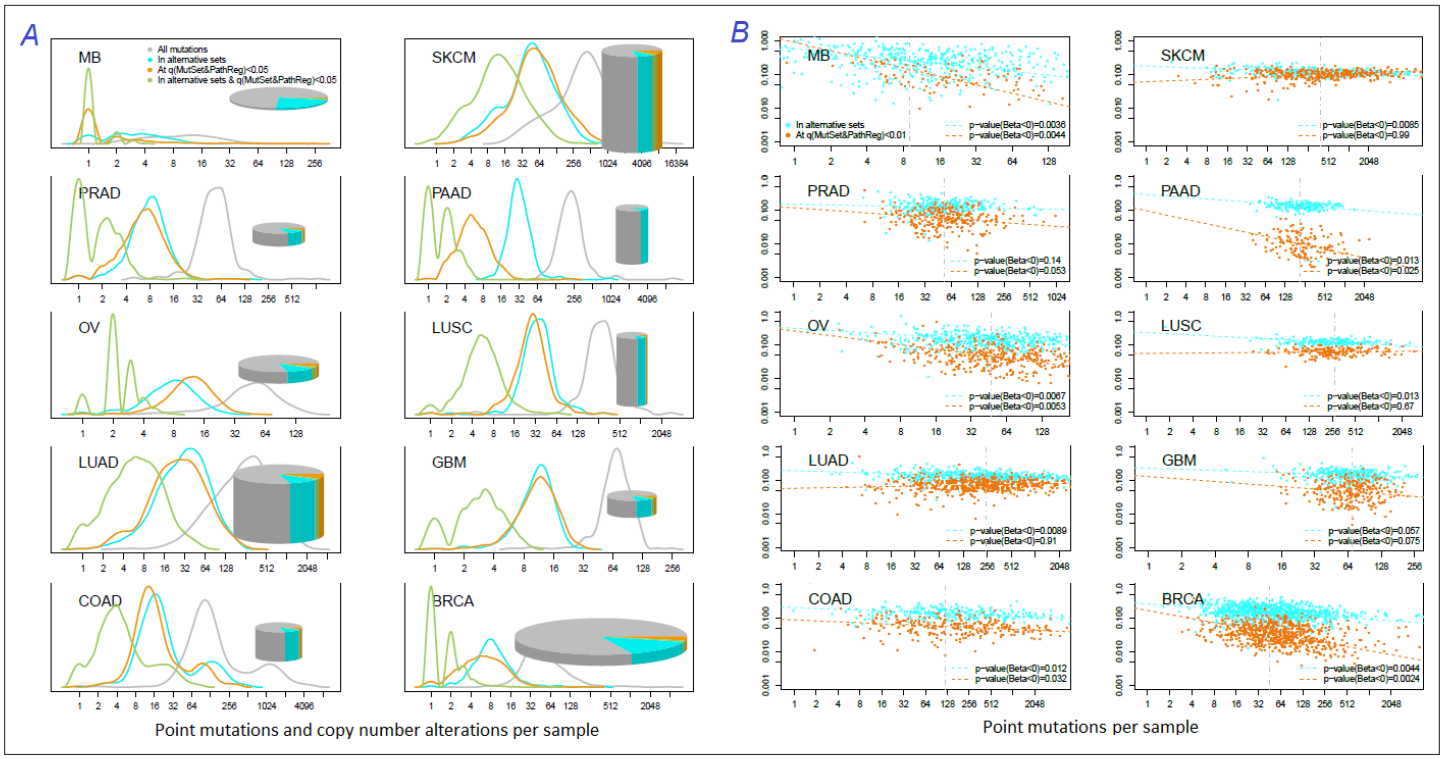

Figure 5. Distribution of somatic mutations versus drivers across genomic samples. (**A**) Relative density plots of mutations and declared drivers. Pie charts summarize counts per genomic sample in each of the ten cohorts (height: average number of reported mutations per sample; width: number of samples in the cohort). (**B**) Overlap between the predictions by MutSet&PathReg and the merge of alternative gene sets (1434 genes in total) color by Jaccard index (sets' intersection divided with sets' union). The MGS sizes (regardless of driver status) are expressed as marker size. Gaussian noise was added to marker coordinates for better readability.

Indeed, the DGSs [q(MutSet&PathReg)<0.05] were often informative on patient survival. We tested three different clustering techniques and found that in many cases DGS scores differentiated cohorts by survival: 7…14.8% of all tests yielded significant Cox proportional hazard models (Benjamini-Hochberg FDR <0.25). Furthermore, in up to 21.1% of all tested cases the significant partitions were recapitulated on test sets (while FDR estimates from Cox models were below 0.25) (see examples in ***Figure 6*** and full details in ***Supplementary file 6***). For comparison, splitting in the same framework by high vs. low tumor stage did not differentiate patients by survival (not shown).

NEA scores based on either drivers or gene expression point to same pathways associated with survival

Finally, we checked if association of specific FGS scores with survival could be traced at the level of mRNA transcription. To this end, we derived lists of 100 patient-specific genes with expression most deviating from the cohort mean (gene expression based AGS) and looked if their NEA scores for the same FGS would also be associated with survival. By testing the 10 cohorts, 2 survival types, 3 clustering methods, and the 1659 FGSs, we identified 31 cases where the association with survival was observed for both DGS-FGS and gene expression based AGS-FGS scores (***Figure 6—figure supplement 1*** to ***Figure 6***). The discovery of this many associations was significant in a random permutation test requiring Bonferroni-adjusted p-value <0.01 while permutation test-based p-value <0.0001 (***Figure 6—figure supplement 2*** to ***Figure 6***). Remarkably, in most of the cases opposite relations with survival between DGS and gene expression based AGSs were observed: better outcome was associated with high scores of the former while lower scores of the latter, or vice versa.

For example, MB and LUAD cohorts were differentiated by survival using NEA score profiles for two pathways (***Figure 7***). The cohort patients were represented first by DGSs (left) and then by gene expression based AGSs (right). The NEA scores reflected connectivity between pathway genes and patient genes (either DGS or gene expression based AGS). A higher NEA score would indicate that

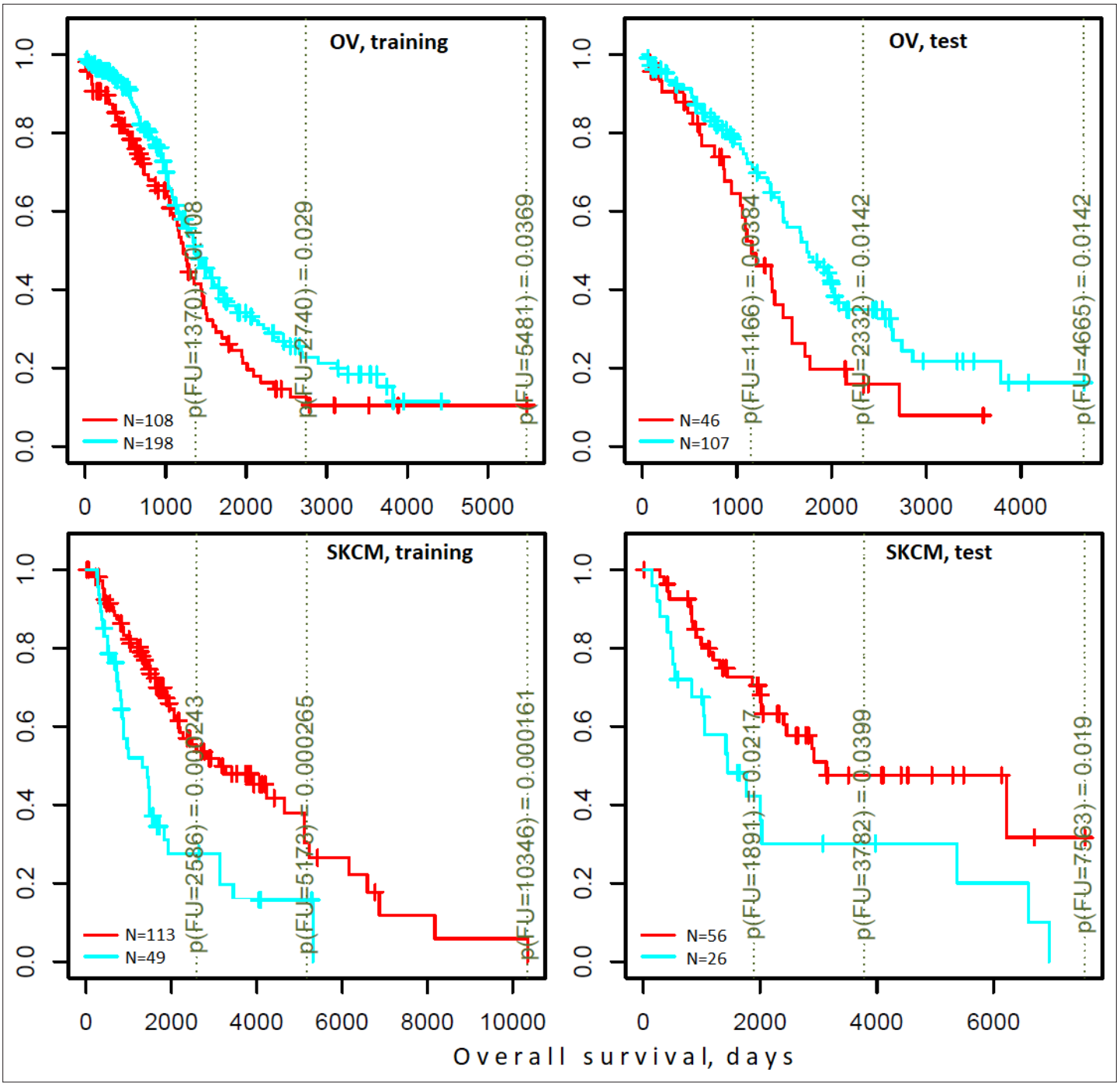

Figure 6. Differential survival of patients stratified in pathway space created by network enrichment analysis of driver gene sets. Vertical captions (brown) convey Cox proportional hazard p-values for three follow-up intervals.

The online version of this article includes the following figure supplement(s) for figure 6:

Figure supplement 1. Analysis of significance across survival curves.

Figure supplement 2. Distribution of p-values for survival correlations from different clustering methods and agreement of P-values on train versus test data sets.

relatively many patient-specific genes were linked to the given pathway. MB cells are known to sometimes produce granulocyte colony-stimulating factor (***Pietsch et al., 2008***), which can affect influx of granulocytes (***Vermeulen et al., 2017***) and disease prognosis (***Paul et al., 2020***). With regard to 'Biocarta granulocytes pathway', the MB patients where stratified so that higher DGS scores indicated

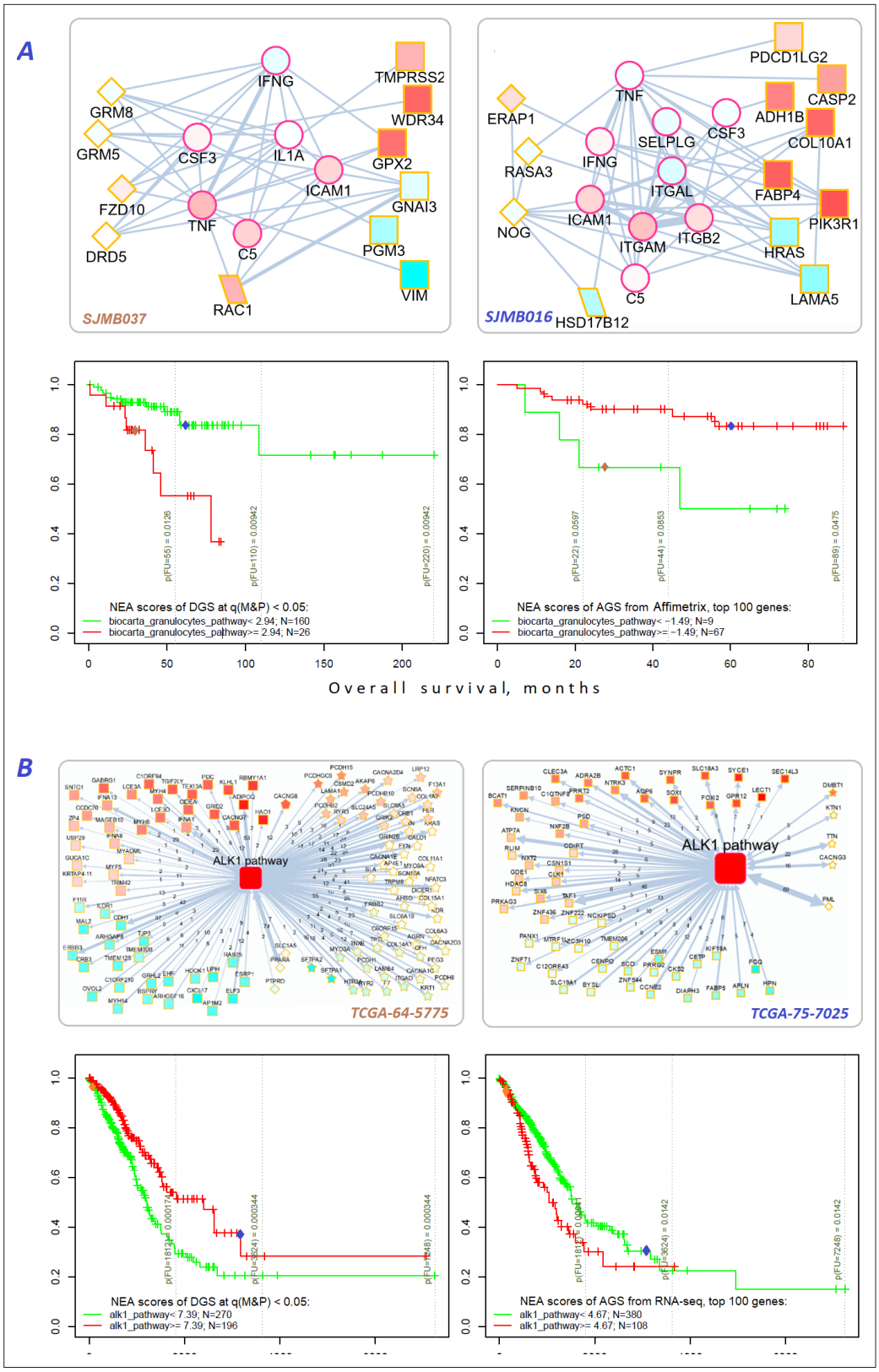

Figure 7. Network enrichment and survival analyses of patient specific lists of drivers and differentially expressed genes. (**A**) Example from MB cohort. (**B**) Example from LUAD cohort. Yellow borders: patient-specific gene sets including (***Torkamani et al., 2009***) driver alterations (*q(MutSet&PathReg)*<0.05): either point mutations (stars) or copy number changes (diamonds) (***Hanahan and Weinberg, 2011***) genes with mRNA expression most deviating

Figure 7 continued on next page

Figure 7 continued

compared to the rest of the cohort (rectangles) (***Lawrence et al., 2014***) both categories 1 and 2 (rhomboids). Magenta borders: pathway genes (circles). Each gene is colored by expression in the given patient sample compared to the cohort mean. Note that pathway genes usually did not manifest genomic or strong expression changes. **In figure** (**B**) the edges combine individual network links between genes. Links within pathway not shown. Clinical and NEA data for the patients.

The online version of this article includes the following source data for figure 7:

Source data 1. Clinical survival data and NEA scores for the MB and LUAD patients.

poorer survival, whereas higher gene expression based AGS scores were associated with better survival. Subnetwork patterns for two patients exemplify this analysis (***Figure 7A***). ALK fusion events are a well-established target for non-small cell lung cancer therapy (***Ross et al., 2017***). While none of the patients were treated with an ALK inhibitor in LUAD cohort, 'Biocarta ALK1 pathway' scores for both DGS and gene-expression-based AGS were informative on overall survival within 6-year follow-up interval. Again, relations to survival were opposite for DGS versus gene-expression-based AGS scores.

Novel categories of cancer driver genes

We noticed that a large part of the connectivity with regard to functional hallmarks (***Figure 2B***), which distinguished the NEAdriver predictions from other computational gene sets, was due to multiple collagens, laminins, and integrins predicted in most of the cohorts. These genes are typically rather long, within the 2nd quartile of protein coding gene list ranked by CDS length. Their median mutation frequencies per base pair of CDS length were just 1.5…2.5 times higher than that of all protein coding genes, which likely explains their escape from computational analyses so far. Nonetheless, across the ten studied cohorts genomic alterations occurred on average in 2–7 genes of these families per sample (***Figure 8A***). Using logistic regression with total mutation burden per sample as a covariate, we found that point mutations patterns of these genes also significantly (at FDR <0.05 after adjustment for multiple testing) co-occurred pairwise: there were e.g. 84, 10, 23, and 308 such pairs in BRCA, COAD, LUAD, and PAAD cohorts, respectively.

Figure 8B displays a typical subnetwork of genes that encode collagens, laminins, and integrins interconnected with heparan sulphate, fibronectin as well as a few signaling proteins – all affected with point mutations or copy number changes in the same cancer genome. This pattern explains high enrichment in network links to epithelial mesenchymal transition, apical junction, and angiogenesis hallmarks presented in ***Figure 2B***. Previously, roles of these families in for example cell migration, epithelial mesenchymal transition, or angiogenesis were rather well characterized at the structural (***Ahmed et al., 2005***; ***Rousselle and Scoazec, 2020***; ***Moilanen et al., 2017***) and tissue-specific transcriptional (***Bretaud et al., 2020***; ***Mammoto et al., 2013***) levels, and even suggested as markers for cancer diagnostics (***Risteli et al., 2014***) and targets for treatment (***Tsuruta et al., 2008***). However, they were not recognized by computational analyses, with a few exceptions: COL1A1 (34), COL5A1, COL5A3, ITGB7 (13), COL18A1 and ITGA6 (42), and none were so far included in the Cancer Gene Census or FoundationOne targeted sequencing panel.

Discussion

So far, most of projects presenting novel cancer drivers s generalized the discovery: either globally, within the pan-cancer paradigm (***Campbell et al., 2020***; ***The Cancer Genome Atlas Research Network et al., 2013***) or within site/organ specific tumour types (***Berger et al., 2018***), sometimes delineating subtype-specific drivers (***Pugh et al., 2012***; ***Sweet-Cordero and Biegel, 2019***). Such approaches possess lower statistical power regarding short genes. We found that shorter genes were underrepresented in all alternative sets considered in this study, except the curated cancer pathways (Figure Supplement to ***Figure 4***). Meanwhile, even rarely mutated genes can be drivers for example in absence of alterations in a 'major' gene, such as TP53 – but identifying such associations would require genome-wide studies at an unaffordable scale (***Stracquadanio et al., 2016***). This situation apparently contradicts the individualized approach to cancer treatment, which suggests molecular pathological analyses for disease prognostication, administration of targeted drugs (***Remke et al.,***

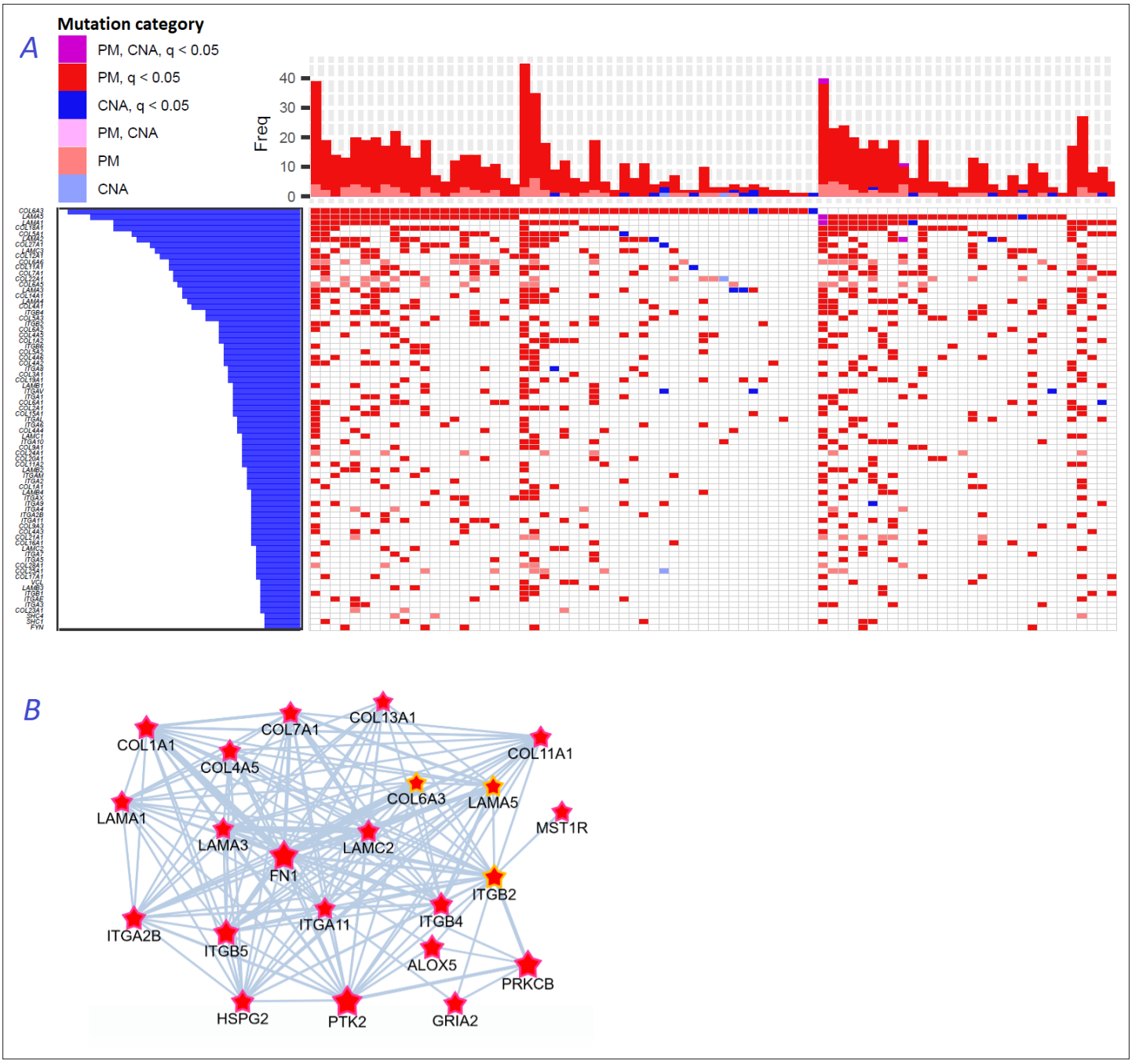

Figure 8. Novel gene families in cancer driver context. (**A**) Upper left fragment of a waterfall plot containing top 75 genes for collagens, laminins, integrins, and a few signaling proteins most frequently mutated in COAD cohort (269 samples in total). Genes with point mutations (PM) and copy number alterations (CNA) are colored according to gaining significance as q(MutSet&PathReg)<0.05 or not. (**B**) Point mutations connected with each other in the global network and identified as potential drivers in one genomic sample TCGA-CM-6171–01 (COAD) at q(MutSet&PathReg)<0.05. In particular, point mutations in COL6A3, ITGB2, and LAMA5 were also detected as significantly co-occurring across COAD cohort in a general linear model accounting for total mutation burden as covariate (FDR <0.01).

2013), and discovery of novel drug targets. Currently, the majority of patients are not amenable to any approved targeted treatment since respective matching mutations occur with low prevalence. Further development of precision cancer medicine requires considering functional context of cancer genome in each patient (***Wheeler and Wang, 2013***).

Novel approaches to driver identification is therefore urgent: while every cancer genome is expected to possess driver mutations, many cases lack any alterations in known cancer genes – which

is counterintuitive and undermines the ground for targeted therapy. Due to the high mutational heterogeneity of cancer samples, frequency-based methods have reached their limits of statistical power to detect novel cancer drivers.

One should distinguish between mutation frequency as a tool to discover driver genes and the biological mechanism via which driverness is implemented. A mechanism for a given gene would be implemented at the level of individual tumors and therefore does not have to be directly associated with cohort-level statistics. Unlike of most of the methods (including many network-based ones), NEAdriver itself did not lean on the mutation frequency features. Instead, individual mutation events became input to NEAdriver and were evaluated independently of each other.

We pursued this approach via network analysis of mutated genes, by which patient-specific driver constellations should be discerned from the background of passenger burden. The network analysis – an already popular method to identify cancer genes using functional context – in our implementation was relatively less biased toward network hubs and thus more sensitive to novel driver genes. The previous guilt-by-association analyses (***Köhler et al., 2008***; ***Cava et al., 2018***; ***Cho et al., 2016***; ***Reyna et al., 2020***) predicted gene function based on functional connections to known functional categories. Confidence of such predictions alone, for example in absence of experimental data was very low due to rare occurrence of actual mutations and thus lower true discovery rate, as shown by John Ioannidis (***Ioannidis, 2005***). NEAdriver was more focused due to considering concrete molecular phenotypes (*de facto* alterations in individual genomes) and combining relevant evidence from two network analysis channels. MutSet channel possesses an extra advantage: it can be directly, without pre-training on whole-cohort data, applied to a sequencing dataset of a single novel patient.

Since capabilities of both MutSet and PathReg could only be implemented on driver constellations of sufficient size, it was important to test NEAdriver on cancer genomes with different mutation loads. This feature varied from a few affected genes in MB to thousands in lung and skin cancers. Despite the variability of this and other biological parameters, both statistical performance and candidate drivers' functional profiles proved to be rather close across the 10 cohorts. As an example, three pathways were systematically included in the models: hsa04020:Calcium_signaling_pathway (CS), hsa05412:Arrhythmogenic_right_ventricular_cardiomyopathy_(ARVC), and hsa05414:Dilated_cardiomyopathy (DCM). Although the role of CS in cancer was rarely considered central, cell migration and adhesion do involve modulation of cell motility and shape where ion channels and pumps play major roles, so that CS genes are known for both downregulation and functional implication in cancers (***Tajada and Villalobos, 2020***; ***Phan et al., 2017***; ***Litan and Langhans, 2015***). ARVC and DCM are functionally close to CS, although have little overlap in member genes. At the first glance, the major factor behind MGS-CS interrelations might be the frequently mutating titin TTN. Its involvement in cardiopathies and cancer has been long argued because of its extremely long coding sequence, thus likely prone to spurious alterations (~50% of LUAD samples). However, there were many individually rare mutations, which together revealed emergent network patterns between MGSs and either CS, or ARVC, or DCM: cadherins, laminins, integrins, metalloproteases, nitric oxide synthases, ryanodine receptors, adenylate cyclases, subunits of protein kinase A etc. They contributed to the NEA scores with multiple network links so that for example the median edge counts between genes of MGSs and of the pathways were 22…45 in MB and 497…890 in LUAD. We therefore did not exclude genes commonly supposed to be 'artefactual' but instead labelled them in the final tables.

The cohort- and sample-specific NEAdriver predictions agreed well with nearly all tested alternative sets (***Figure 2A***). The best agreement was found for cancer-site-specific gold standard sets while less so for pan-cancer sets. The computational, mostly frequency-based sets performed worse than curated sets (***Figure 3A and B***). In the functional space, NEAdriver predictions were positioned differently compared to computational and database sets, but close to curated cancer pathway sets (***Figure 2B***), which confirmed both novelty and relevance of NEAdriver findings. The latter also differed from most of the alternative sets in having much less bias in regard of network node degrees and gene length.

A realistic estimate of the true discovery rate for NEAdriver predictions was obtained by accounting for putative drivers not included of the gold standard sets, so that PPV could be as high as 78%...88% at the q(mutSet&PathReg)=0.05, which was verified by the two alternative ways of PPV calculation using four different gold standards (***Figure 3E and F***). The efficiency of NEAdriver was confirmed by

the ability of DGS to stratify patients by survival (***Figure 6***) and the striking tendency of same pathways being associated with survival via both mutation- and expression-based patient scores (***Figure 7***).

Obviously, NEAdriver alone would miss events detectable by other methods, for example when certain 'stand-alone drivers' impose strong effects on their own (TP53, APC etc.), without apparent interaction with other genes. This creates an incentive for creating a combined methodology and a toolbox in the future. But already now our analysis identified hundreds of somatic gene alterations that had not been deemed functional in previous research. After evaluation performed in multiple ways, the body of predictions appears confident, providing a set of provable research hypotheses and suggesting new strategies for cancer prognosis and individualized treatment.

Materials and methods

Medulloblastoma meta-cohort

We collected data from publications presenting large-scale datasets (***Jones et al., 2012***; ***Northcott et al., 2017***; ***Pugh et al., 2012***; ***Robinson et al., 2012***) and two public datasets available online (PBCA-DE and PEME-CA). We retrieved available exome sequencing profiles as well as copy number alterations, gene expression, and clinical data. We translated gene identifiers into gene symbols according to ENSEMBL annotations v.93 and then made sure all the gene symbols are found in the network and are up to date according to GeneCards (***Stelzer et al., 2011***) annotations.

For consistency with the publication datasets, we excluded the following types of mutations from PBCA-DE and PEME-CA sets: intron variant, upstream gene variant, 3_prime_UTR_variant, 5_prime_UTR_variant, intergenic region, downstream gene variant, synonymous variant, and splice region variant. For a few patient IDs that were found in more than one dataset, their mutation profiles were merged (if different).

Overall survival data was collected from the published datasets. A few patients with discrepant data (for instance, ICGC_MB193 was 2.3 years old according to Northcott dataset, but 70 years old according to PBCA-DE dataset) were excluded. For 18 samples with different follow-up, we accepted the newest survival time values from Northcott dataset.

Data from all the datasets were combined into one cohort dubbed MB(union), so that 541 patients were covered with both clinical and exome sequencing data.

TCGA cohorts

The TCGA data were obtained via https://portal.gdc.cancer.gov/. Clinical profiles were used according to the most recent update (***Liu et al., 2019***).

Network enrichment analysis

Network enrichment between two gene sets of interest S_a and S_b is estimated by comparing the actual number of network edges $\hat{\varepsilon}_{S_a \leftrightarrow S_b}$ that connect nodes of S_a with nodes of S_b in the real, biological network G_B=(*E*,*V*) with a number expected by chance $\hat{\varepsilon}_{S_a \leftrightarrow S_b}$ in a random network G_R=(*E*,*V*) where particular node degrees *k* of genes $\forall g_i \in \boldsymbol{S_a}$; $\forall g_j \in \boldsymbol{S_b}$; $g_i \neq g_j$ equal to those of the actual network (which implicitly assumes that the whole degree sequences of G_B and G_R are identical, too). In an earlier work (***Alexeyenko et al., 2012***), series of randomized instances of G_R were created using an algorithm of explicit edge permutation (***Maslov and Sneppen, 2002***) and used for estimating expected variance of ε. Later, it was demonstrated (***Jeggari and Alexeyenko, 2017***) that $\epsilon_{i \hat{\leftrightarrow} GS}$ can be calculated analytically in a fast and unbiased manner:

$$\hat{\varepsilon}_{S_a \leftrightarrow S_b} = \left(\sum_{i=1}^{|S_a|} \mathrm{k}_i * \sum_{j=1}^{|S_b|} \mathrm{k}_j \right) / 2|\mathrm{E}|;$$

Then the difference between the actual and expected edge counts

$$\Delta\varepsilon = \varepsilon_{S_a \leftrightarrow S_b} - \hat{\varepsilon}_{S_a \leftrightarrow S_b};$$

is used to estimate significance of the relation $S_a \leftrightarrow S_b$ with a $\chi 2$ statistic:

$$\chi^2 = \frac{\Delta\varepsilon_i^2}{\hat{\varepsilon}_{S_a \leftrightarrow S_b}} + \frac{\Delta\varepsilon^2}{|E| - \hat{\varepsilon}_{S_a \leftrightarrow S_b}},$$

The $\chi 2$ does not follow Gaussian distribution, but it can be conveniently converted to Z-scores and then used safely for downstream calculations in for example linear models.

In the simplest NEA case one of the sets is a single gene *i*:

$$\hat{\varepsilon}_{i \leftrightarrow \mathbf{S}} = (K_i * \sum_g k_g)/2|E|; \;\; \Delta\varepsilon_i = \varepsilon_{i \leftrightarrow \mathbf{S}} - \hat{\varepsilon}_{i \leftrightarrow \mathbf{S}};$$

$$\chi^2 = \frac{\Delta\varepsilon_i^2}{\hat{\varepsilon}_{i \leftrightarrow \mathbf{S}}} + \frac{\Delta\varepsilon_i^2}{|\mathrm{E}| - \hat{\varepsilon}_{i \leftrightarrow \mathbf{S}}},$$

which simplified calculation and – within this work – enabled estimation of network enrichment for a mutated gene against the (rest of) mutations in the same cancer genome, called mutated gene set (MGS) in MutSet method or functional gene set (FGS, or simply pathways) in case of PathReg.

Network

For NEA we merged network of top 1 million edges, ranked by confidence (i.e. Final Bayesian Score) from FunCoup 3 (***Schmitt et al., 2014***) and all edges of Pathway Commons 9 (***Cerami et al., 2010***; ***Rodchenkov et al., 2020***). We made sure that all genes reported as altered in at least one of the ten cancer cohorts had up-to-date gene symbols in network. That resulted in a network of 19,035 nodes (unique gene symbols) connected with 1,731,648 unique edges.

Mutation gene sets and driver gene sets

We defined mutated gene sets (MGS) as lists of all genes of a given tumour sample reported with somatic mutations (SM) in the MAF files. MGSs were used as whole sets in driver evaluation of MutSet channel. MGSs ***did not include*** copy number altered (CNA) genes.

The analysis included all mutations reported in the TCGA MAF files, regardless of predicted functional impact. Indeed, although synonymous and intronic mutations were often disregarded in cancer research, their involvement in carcinogenesis seems likely and has been recently demonstrated (***Sharma et al., 2019***). In our datasets, frequency of silent mutations was somewhat lower than of non-synonymous ones, but still significant so that many frequently mutated genes showed elevated rates in both categories (***Supplementary file 7***). Therefore, each altered gene *i* from a given sample, either SM or CNA, was evaluated against the MGS and received a MutSet q-value. Significantly altered genes with were included in final driver gene sets $DGS_{0.05}$ and $DGS_{0.01}$ under conditions $q(MutSet\&PathReg)<0.05$ and $q(MutSet\&PathReg)<0.01$, respectively.

Functional gene sets

For the PathReg predictor, we used 318 KEGG pathways from version as of 16 August 2018. Considering the importance of SHH and WNT pathways in e.g. medulloblastoma, alongside with respective KEGG pathways we included these two also in Biocarta (***Nishimura, 2001***) versions (the versions were very different in size and length). We updated gene symbols in the sets in the same way as described for the mutations above.

For the survival analysis, the FGS collection consisted of 1,659 entries from BioCarta (***Nishimura, 2001***), KEGG (***Kanehisa et al., 2002***), Reactome (***Croft et al., 2014***), WikiPathways (***Pico et al., 2008***), MetaCyc (***Caspi et al., 2014***), and MSigDB hallmarks (***Liberzon et al., 2015***).

Altered gene sets (transcriptomics)

The gene expression data was used from the available cohort data sets:

- Affymetrix for MB (***Robinson et al., 2012***);
- Agilent for OV and GBM;
- IlluminaHiSeq_RNASeqV2 for the rest of TCGA cohorts.

The AGS were compiled as sample-specific lists of top *N* genes (*N*=[50,100,200]) with normalized mRNA expression most different from the respective cohort mean using function samples2ags(…, method = "topnorm") from R package NEArender (***Jeggari and Alexeyenko, 2017***).

NEAdriver: algorithm

The driver discovery algorithm combined results from two NEA-based channels, MutSet and PathReg.

MutSet channel

The MutSet values quantified network enrichment between each gene *m* having a somatic point mutation in genome *j* and the set MGS of all other point mutations in the same genome ($m \in MGS_j$). They were calculated as NEA scores $Z_{m\leftrightarrow MGS_j}$ so that within a cohort the same gene might receive multiple, sample-specific MutSet values. Respective p-values were obtained from the normally distributed $Z_{m\leftrightarrow MGS_j}$ values with a trivial R function pnorm (***Jeggari and Alexeyenko, 2017***).

PathReg channel

As the independent variable for training the PathReg predictor, we employed vectors anchor.summary. Specific NEA scores were calculated for every gene *i* present in the network (N=19035) versus every MGS in the given cancer cohort *c*. The anchor.summary values μ_{ic} were then obtained by summing up over all N_c available samples, regardless of mutation status of *i* in genome *j*:

$$\mu_{ic} = \sqrt{\log \frac{\sum_{j=1}^{j\leq N_c} Z_{i\leftrightarrow MGS_j}}{N_c}};$$

Since the score $Z_{i\leftrightarrow MGS_j}$ is derived from the network patterns of mutated genes across the cohort and does not depend on the mutation profile of *i* itself, the μ_{ic} value would reflect a general propensity of *i* to interact with constellations of putative cancer genes. We note that the algorithm is not given any information on previously identified cancer driver genes and works in the assumption that passenger mutations would not produce relevant signal. The transformations via $\chi 2 \rightarrow \mathrm{Z}$, log, and square root were imposed in order to render distributions closer to Gaussian.

The μ_{ic} profiles were rather scarce due to rare occurrence in MGS of true drivers that would interact with a given gene *i*. We thus further improved the gene specific values via modeling μ_{ic} with pathway NEA scores $Z_{i\leftrightarrow FGS}$. These were calculated for 320 FGS versus each of the *N* network genes and then used as a matrix of dependent variables $\mathbf{\Phi}$ in PathReg training.

Then sparse regression models were created using function cv.glmnet from R package glmnet (***Friedman et al., 2010***).The chosen package implements elastic net models for solving the problem:

$$\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} w_i l\left(y_i, \beta_0 + \beta^T x_i\right) + \lambda \left[\frac{(1-\alpha)\|\beta\|_{l2}^2}{2} + \alpha \|\beta\|_{l1}\right],$$

where α is a mixing parameter for balance between lasso and ridge regression (whereby α=0 and α=1 would lead to plain ridge and lasso regressions, respectively). In our case (α=1), glmnet solved just the lasso problem:

$$\min_{\beta_0,\beta} R_\lambda\left(\beta_0, \beta\right) = \min_{\beta_0,\beta}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\mu_{ic} - \beta_0 - \beta^T\Phi\right) + \lambda\|\beta\|_{l1}\right)$$

Parameter λ determines complexity of the multivariate regression model, i.e. what subset of initially submitted variables of $\mathbf{\Phi}$ should receive non-zero coefficients. Under 3-fold cross-validation, function cv.glmnet tested a series of λ values while controlling the cross-validation mean squared error (CVM). The cohort-specific optimum λ_c was found as a trade-off between model precision and complexity using Bayesian information criterion (BIC), which was deemed preferable (***Giraud, 2015***) over Akaike information criterion in the context of favourable dimensionality $\left(n_m \gg p; p = 320\right)$. The optimal λ_c was set at the number of FGS variables with non-zero coefficients *k* as the smallest possible within 2 standard errors of BIC from the lowest BIC value:

$$k:\ \lambda = \arg\min_{\lambda}\left[BIC < \left(\inf\left(BIC\right) + \frac{2\sigma_{BIC}}{\sqrt{n}}\right)\right];$$

$$m_{ic} = \beta_0 + \sum_{j=1}^{k}\beta_j f_{ij};\ \forall f \in \Phi; |\beta \ni \beta \neq 0| = k;$$

After this training and model selection step, the retained test subsets were used to check how the original values μ_{ic} correlate with the predicted values m_{ic} (***Supplementary file 1***).

The distribution of m_{ic} values was non-parametric but close to Gaussian. Therefore, respective p-values were modelled via a normal distribution where mean and standard deviation were estimated as median and 84.2th percentile of the empirical distribution, respectively:

$$\bar{m} = \tilde{m};\ \sigma = P_{84.2}\left(m\right)$$

(in the Gaussian distribution 84.2% of values are within $\bar{m} \pm \sigma$).

Integration of channels

The p- values from both MutSet and PathReg were adjusted to with Benjamini-Hochberg method (***Benjamini and Hochberg, 1995***). These *q*-values were equivalent to false discovery rate which conveys the probability of a given driver prediction to be false. These values were integrated into the final value as a product $q(MutSet\&PathReg)=q_{MutSet}*q_{PathReg}$, which presented the probability that neither channel have produced true predictions. Therefore, 1 – *q(MutSet&PathReg)* was the probability of either channel to be true and we convened to trust a driver prediction if *q(MutSet&PathReg)*<*c* (*c*=[0.01, 0.05]).

Gold standard and alternative driver sets

Literature-based sets

As gold standard for **MB**, we compiled a list of unique 516 gene symbols (Hg19 human genome version), of which 12 were found in OMIM database, 140 in Disease Ontology, and 399 in MB-related publications found in PubMed:

- Cavalli, F. M. G. et al. Intertumoral Heterogeneity within Medulloblastoma Subgroups. *Cancer Cell* **31**, 737–754.e6 (2017).
- Ellison, D. W. et al. Medulloblastoma: clinicopathological correlates of SHH, WNT, and non-SHH/WNT molecular subgroups. *Acta Neuropathol. (Berl.)* **121**, 381–396 (2011).
- Gajjar, A. et al. Pediatric Brain Tumors: Innovative Genomic Information Is Transforming the Diagnostic and Clinical Landscape. *J. Clin. Oncol.* **33**, 2986–2998 (2015).
- Hovestadt, V. et al. Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. *Nature* **510**, 537–541 (2014).
- Jones, D. T. W. et al. Dissecting the genomic complexity underlying medulloblastoma. *Nature* **488**, 100–105 (2012).
- Northcott, P. A. et al. Medulloblastomics: the end of the beginning. *Nat. Rev. Cancer* **12**, 818–834 (2012).
- Northcott, P. A. et al. Subgroup-specific structural variation across 1,000 medulloblastoma genomes. *Nature* **488**, 49–56 (2012).
- Northcott, P. A., Dubuc, A. M., Pfister, S. & Taylor, M. D. Molecular subgroups of medulloblastoma. *Expert Rev. Neurother.* **12**, 871–884 (2012).
- Parsons, D. W. et al. The genetic landscape of the childhood cancer medulloblastoma. *Science* **331**, 435–439 (2011).
- Pugh, T. J. et al. Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. *Nature* **488**, 106–110 (2012).
- Robinson, G. et al. Novel mutations target distinct subgroups of medulloblastoma. *Nature* **488**, 43–48 (2012).
- Taylor, M. D. et al. Molecular subgroups of medulloblastoma: the current consensus. *Acta Neuropathol. (Berl.)* **123**, 465–472 (2012).

We included all altered genes mentioned in these publications. Mechanisms of alteration included changed methylation, gene copy number, and point mutations.

For the TCGA cohorts, we employed the dedicated KEGG pathways:

- BRCA <- hsa05224:Breast_cancer;
- GBM <- hsa05214:Glioma;
- LUAD <- hsa05223:Non-small_cell_lung_cancer;
- LUSC <- hsa05223:Non-small_cell_lung_cancer;
- SKCM <- hsa05218:Melanoma;
- PRAD <- hsa05215:Prostate_cancer;

- PAAD <- hsa05212:Pancreatic_cancer;
- COAD <- hsa05210:Colorectal_cancer;
- BLCA <- hsa05219:Bladder_cancer;
- OV <- hsa05213:Endometrial_cancer, since origins of these two are intertwined (***Merritt and Cramer, 2010***).

Gene sets from computational analyses

Bailey2018.PANCAN, N=200 (***Bailey et al., 2018***) PanCancer and PanSoftware analysis over 9,423 tumor exomes from 33 of The Cancer Genome Atlas projects using 26 computational tools.

HCD, N=291 (***Tamborero et al., 2013***) discovered drivers in 12 TCGA cohorts (of which six overlapped with our analysis) with a combination of four algorithms that prioritized mutated genes based on mutation rate, functional impact, positional 'hotspots', and specific enrichment in phosphorylation sites.

IntOGen, N=37…73 (***Martínez-Jiménez et al., 2020***) cancer-type-specific lists of cancer genes (per type) predicted by Integrative OncoGenomics (IntOGen) pipeline.

Martincorena-2017, N=369 (***Martincorena et al., 2017***) a list of known cancer genes used in the molecular evolution (positive selection) analysis and applied 7664 tumors across 29 cancer types.

MuSiC, N=127 (***Kandoth et al., 2013***) identified drivers based on relative point mutation frequency, assisted with expression analysis and database annotations in 12 TCGA cancers (of which six overlapped with our analysis). We used the 'PanCancer' list.

MutSig, N=260 (***Lawrence et al., 2015***) performed a comprehensive point mutation frequency analysis using exome sequencing data from 21 cancer cohorts (of which nine overlapped with our study), while accounting for mutation burden, clustering, and functional impact.

NetSig5000, N=62 (***Horn et al., 2018***) gathered evidence for potential driverness of each gene via functional coupling to frequently mutated genes in the global network. The genes' own mutation frequencies were incorporated into NetSig scores at a separate step.

Database gene sets

Cancer Gene Census N=12…64 (***Futreal et al., 2004***) cancer type specific lists were downloaded on 7th of February, 2019.

KEGG#05200:Pathways_in_cancer, N=395 (***Kanehisa et al., 2002***) was a 'pan-cancer' version including a curated selection of organ-specific cancer pathway gene lists.

Five 'general' cancer pathways, N=457 (***Kanehisa et al., 2002***) was created as a union of the following cancer related KEGG pathways:

- hsa05202:Transcriptional_misregulation_in_cancer;
- hsa05203:Viral_carcinogenesis;
- hsa05204:Chemical_carcinogenesis;
- hsa05205:Proteoglycans_in_cancer;
- hsa05206:MicroRNAs_in_cancer.

FoundationOne, N=330 (https://www.foundationmedicine.com/resources) is the targeted sequencing panel used for cancer diagnostics, created as a merge of 'general' and 'rearrangements' sections.

Gene sets from individualized analyses

OncoIMPACT (N=162…695 per cohort) (***Bertrand et al., 2015***) used an expression-driven approach to verify driver roles of point mutations in five TCGA cohorts (of which four overlapped with our analysis). The paper reported cohort-specific driver ranks rather than individual, sample-level estimates. The genes that received a rank were used in comparisons with NEAdriver results.

SCS (***Guo et al., 2018***) reported driver role evaluation in individual samples. They reported only top 50 genes after global, cohort-specific ranking. In parallel, the tables contained top 50 genes from OncoIMPACT (***Bertrand et al., 2015***), DriverNet (***Bashashati et al., 2012***), DawnRank (***Hou and Ma, 2014***), and HotNet2 (***Reyna et al., 2018***; ***Leiserson et al., 2015***) methods, which we also imported and used in our comparison.

Sets of artefactual driver genes

Lawrence2013.FPs, N=19 (***Lawrence et al., 2013***) list of genes frequently presented in literature as false positive cancer drivers.

Martincorena2017.artifacts, N=49 (***Martincorena et al., 2017***) list of genes which are usually heavily-affected by sequencing artifacts.

IntOGen.KnownArtifacts, N=19 (***Martínez-Jiménez et al., 2020***) list of genes labelled as "Known artifact" in their resulting table.

Normalization of mutation frequencies

The number of samples where each given gene was mutated was normalized by dividing it with its CDS length. When necessary, the sample-specific total mutation burden values were accounted for as total number of point mutations reported in all genes per sample.

Code availability

Documented code is available under the modified BSD license on GitHub: https://github.com/avev-iort/NEArender-2.x, (copy archived at swh:1:rev:5829beb819c689790359f199547362a31d1a1d54; ***Petrov, 2022***).

Acknowledgements

The authors are grateful to Vetenskapsrådet for provided funding. The analysis used data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Additional information

Funding

Funder	Grant reference number	Author
Vetenskapsrådet	2016-04940	Iurii Petrov Andrey Alexeyenko

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Iurii Petrov, Data curation, Formal analysis, Investigation, Resources, Software, Validation, Writing - review and editing; Andrey Alexeyenko, Conceptualization, Funding acquisition, Investigation, Methodology, Software, Supervision, Visualization, Writing – original draft

Author ORCIDs

Andrey Alexeyenko http://orcid.org/0000-0001-8812-6481

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.74010.sa1
Author response https://doi.org/10.7554/eLife.74010.sa2

Additional files

Supplementary files

- Supplementary file 1. Performance of PathReg multiple regression models. Observed vs. predicted values of anchor.summary values represents performance of created models on continuous scale. In absence of a strict cut-off, performance was measured as a correlation between anchor.summary observed for each gene in the given cohort versus the a value predicted by the multiple regression model. In heatmaps, values next to gene names indicate number of samples with mutations in the given gene.
- Supplementary file 2. Size vs node degree of cancer genes included in NEAdriver driver

predictions, compared against the union of all alternative sets. X coordinates for the alternative sets represent their sizes. For NEAdriver (which in total, across all cohort samples typically predicted hundreds genes, most being very rare) the sets represent samples n=[50, 100, 200, 400] genes most frequently predicted in each cohort.

- Supplementary file 3. Comparison of NEAdriver results with OncoIMPACT (4 cohorts).
- Supplementary file 4. Comparison of NEAdriver results with other network-based methods (4 cohorts, 50 top genes from each method).
- Supplementary file 5. Dependence of NEAdriver q-values from covariates. The five pages present relations between MutSet&PathReg q (Y axis) and no. of mutations per cohort, gene length, and normalized mutation frequency, replication rate, and gene expression rates, respectively (X axis). Top left legend: Spearman rank R and Kendall tau represent overall correlations between X and Y coordinates regardless of other factors. Bottom left legend: terms' significance in 4-way linear models. Colored points: genes suggested as potential artifacts in literature; those receiving q<0.05 are text-labeled.
- Supplementary file 6. Kaplan-Meier plots for 10 cohorts: different survival metrics, dichotomized by NEA scores for either DGS or GE AGS.
- Supplementary file 7. No of mutations versus normalized mutation frequency: relative frequencies of silent and non-silent mutations per gene.
- Supplementary file 8. Summary tables over each of the ten cohorts. PathReg, MutSet, and combined values per sample, per gene, in each cohort. MutSet q-values are accompanied with no. of network links observed between the given gene and point mutations in the sample, as well as respective NEA Z and NEA p-value. All the MutSet values are sample-specific. PathReg q-values are accompanied with respective anchor.summary, PathReg score, and PathReg p-values. All the PathReg values are cohort-specific and not sample-specific. NEAdriver q-value is product of MutSet q-value and PathReg q-value. The last 3 columns indicate genes listed as possible artifacts in literature.
- Transparent reporting form

Data availability

All data generated or analysed during this study are included in the manuscript and supporting files.

The following previously published dataset was used:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
TCGA consortium	2008	The Cancer Genome Atlas	https://www.cancer.gov/tcga	TCGA, Genome-Atlas

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nature Methods* **7**:248–249. DOI: https://doi.org/10.1038/nmeth0410-248, PMID: 20354512

Ahmed N, Riley C, Rice G, Quinn M. 2005. Role of integrin receptors for fibronectin, collagen and laminin in the regulation of ovarian carcinoma functions in response to a matrix microenvironment. *Clinical & Experimental Metastasis* **22**:391–402. DOI: https://doi.org/10.1007/s10585-005-1262-y, PMID: 16283482

Alexeyenko A, Lee W, Pernemalm M, Guegan J, Dessen P, Lazar V, Lehtiö J, Pawitan Y. 2012. Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. *BMC Bioinformatics* **13**:226. DOI: https://doi.org/10.1186/1471-2105-13-226, PMID: 22966941

Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, Kwok-Shing Ng P, Jeong KJ, Cao S, Wang Z, Gao J, Gao Q, Wang F, Liu EM, Mularoni L, Rubio-Perez C, et al. 2018. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **174**:1034–1035. DOI: https://doi.org/10.1016/j.cell.2018.07.034, PMID: 30096302

Barabási AL, Gulbahce N, Loscalzo J. 2011. Network medicine: a network-based approach to human disease. *Nature Reviews. Genetics* **12**:56–68. DOI: https://doi.org/10.1038/nrg2918, PMID: 21164525

Bashashati A, Haffari G, Ding J, Ha G, Lui K, Rosner J, Huntsman DG, Caldas C, Aparicio SA, Shah SP. 2012. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biology* **13**:12. DOI: https://doi.org/10.1186/gb-2012-13-12-r124, PMID: 23383675

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society* **57**:289–300. DOI: https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berger AC, Korkut A, Kanchi RS, Hegde AM, Lenoir W, Liu W, Liu Y, Fan H, Shen H, Ravikumar V, Rao A, Schultz A, Li X, Sumazin P, Williams C, Mestdagh P, Gunaratne PH, Yau C, Bowlby R, Robertson AG, et al. 2018. A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell* **33**:690-705.. DOI: https://doi.org/10.1016/j.ccell.2018.03.014, PMID: 29622464

Bertrand D, Chng KR, Sherbaf FG, Kiesel A, Chia BKH, Sia YY, Huang SK, Hoon DSB, Liu ET, Hillmer A, Nagarajan N. 2015. Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. *Nucleic Acids Research* **43**:e44. DOI: https://doi.org/10.1093/nar/gku1393, PMID: 25572314

Bretaud S, Guillon E, Karppinen SM, Pihlajaniemi T, Ruggiero F. 2020. Collagen XV, a multifaceted multiplexin present across tissues and species. *Matrix Biology Plus* **6–7**:100023. DOI: https://doi.org/10.1016/j.mbplus.2020.100023, PMID: 33543021

Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD. 2020. Pan-cancer analysis of whole genomes. *Nature* **578**:82–93. DOI: https://doi.org/10.1038/s41586-020-1969-6

Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, Holland TA, Keseler IM, Kothari A, Kubo A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Subhraveti P, Weaver DS, Weerasinghe D, Zhang P, Karp PD. 2014. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research* **42**:D459–D471. DOI: https://doi.org/10.1093/nar/gkt1103, PMID: 24225315

Cava C, Bertoli G, Colaprico A, Olsen C, Bontempi G, Castiglioni I. 2018. Integration of multiple networks and pathways identifies cancer driver genes in pan-cancer analysis. *BMC Genomics* **19**:223. DOI: https://doi.org/10.1186/s12864-017-4423-x, PMID: 5756345

Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, Schultz N, Bader GD, Sander C. 2010. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research* **39**:D685–D690. DOI: https://doi.org/10.1093/nar/gkq1039, PMID: 21071392

Cho H, Berger B, Peng J. 2016. Compact Integration of Multi-Network Topology for Functional Analysis of Genes. *Cell Systems* **3**:540–548. DOI: https://doi.org/10.1016/j.cels.2016.10.017, PMID: 27889536

Ciriello G, Cerami E, Sander C, Schultz N. 2012. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Research* **22**:398–406. DOI: https://doi.org/10.1101/gr.125567.111, PMID: 21908773

Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR, Jassal B, Jupe S, Matthews L, May B, Palatnik S, Rothfels K, Shamovsky V, Song H, Williams M, Birney E, et al. 2014. The Reactome pathway knowledgebase. *Nucleic Acids Research* **42**:D472–D477. DOI: https://doi.org/10.1093/nar/gkt1102, PMID: 24243840

Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, Wilson RK, Ding L. 2012. MuSiC: identifying mutational significance in cancer genomes. *Genome Research* **22**:1589–1598. DOI: https://doi.org/10.1101/gr.134635.111, PMID: 22759861

Doncheva NT, Kacprowski T, Albrecht M. 2012. Recent approaches to the prioritization of candidate disease genes. *Wiley Interdisciplinary Reviews. Systems Biology and Medicine* **4**:429–442. DOI: https://doi.org/10.1002/wsbm.1177, PMID: 22689539

Erten S, Bebek G, Ewing RM, Koyutürk M. 2011. DADA: Degree-Aware Algorithms for Network-Based Disease Gene Prioritization. *BioData Mining* **4**:19. DOI: https://doi.org/10.1186/1756-0381-4-19, PMID: 21699738

Fang H, Gough J. 2014. The "dnet" approach promotes emerging research on cancer patient survival. *Genome Medicine* **6**:64. DOI: https://doi.org/10.1186/s13073-014-0064-8, PMID: 25246945

Franco M, Jeggari A, Peuget S, Böttger F, Selivanova G, Alexeyenko A. 2019. Prediction of response to anti-cancer drugs becomes robust via network integration of molecular data. *Scientific Reports* **9**:2379. DOI: https://doi.org/10.1038/s41598-019-39019-2, PMID: 30787419

Freudenberg J, Propping P. 2002. A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics (Oxford, England)* **18 Suppl 2**:S110–S115. DOI: https://doi.org/10.1093/bioinformatics/18.suppl_2.s110, PMID: 12385992

Friedman JH, Hastie T, Tibshirani R. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33**:1–22. DOI: https://doi.org/10.18637/jss.v033.i01, PMID: 20808728

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. 2004. A census of human cancer genes. *Nature Reviews Cancer* **4**:177–183. DOI: https://doi.org/10.1038/nrc1299, PMID: 14993899

Giraud C. 2015. Introduction to High-Dimensional Statistics. Florida, United States: CRC Press.

Gnad F, Baucom A, Mukhyala K, Manning G, Zhang Z. 2013. Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics* **14 Suppl 3**:S7. DOI: https://doi.org/10.1186/1471-2164-14-S3-S7, PMID: 23819521

Guo W-F, Zhang S-W, Liu L-L, Liu F, Shi Q-Q, Zhang L, Tang Y, Zeng T, Chen L. 2018. Discovering personalized driver mutation profiles of single samples in cancer by network control strategy. *Bioinformatics (Oxford, England)* **34**:1893–1903. DOI: https://doi.org/10.1093/bioinformatics/bty006, PMID: 29329368

Hanahan D, Weinberg RA. 2011. Hallmarks of cancer: the next generation. *Cell* **144**:646–674. DOI: https://doi.org/10.1016/j.cell.2011.02.013, PMID: 21376230

Horn H, Lawrence MS, Chouinard CR, Shrestha Y, Hu JX, Worstell E, Shea E, Ilic N, Kim E, Kamburov A, Kashani A, Hahn WC, Campbell JD, Boehm JS, Getz G, Lage K. 2018. NetSig: network-based discovery from cancer genomes. *Nature Methods* **15**:61–66. DOI: https://doi.org/10.1038/nmeth.4514, PMID: 29200198

Hou JP, Ma J. 2014. DawnRank: discovering personalized driver genes in cancer. *Genome Medicine* **6**:56. DOI: https://doi.org/10.1186/s13073-014-0056-8, PMID: 25177370

Ioannidis JPA. 2005. Why most published research findings are false. *PLOS Medicine* **2**:e124. DOI: https://doi.org/10.1371/journal.pmed.0020124, PMID: 16060722

Jeggari A, Alexeyenko A. 2017. NEArender: an R package for functional interpretation of "omics" data via network enrichment analysis. *BMC Bioinformatics* **18**:118. DOI: https://doi.org/10.1186/s12859-017-1534-y, PMID: 28361684

Jimenez-Sanchez G, Childs B, Valle D. 2001. Human disease genes. *Nature* **409**:853–855. DOI: https://doi.org/10.1038/35057050, PMID: 11237009

Jones DTW, Jäger N, Kool M, Zichner T, Hutter B, Sultan M, Cho Y-J, Pugh TJ, Hovestadt V, Stütz AM, Rausch T, Warnatz H-J, Ryzhova M, Bender S, Sturm D, Pleier S, Cin H, Pfaff E, Sieber L, Wittmann A, et al. 2012. Dissecting the genomic complexity underlying medulloblastoma. *Nature* **488**:100–105. DOI: https://doi.org/10.1038/nature11284, PMID: 22832583

Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson MDM, Miller CA, Welch JS, Walter MJ, Wendl MC, Ley TJ, Wilson RK, Raphael BJ, Ding L. 2013. Mutational landscape and significance across 12 major cancer types. *Nature* **502**:333–339. DOI: https://doi.org/10.1038/nature12634, PMID: 24132290

Kanehisa M, Goto S, Kawashima S, Nakaya A. 2002. The KEGG databases at GenomeNet. *Nucleic Acids Research* **30**:42–46. DOI: https://doi.org/10.1093/nar/30.1.42, PMID: 11752249

Köhler S, Bauer S, Horn D, Robinson PN. 2008. Walking the interactome for prioritization of candidate disease genes. *American Journal of Human Genetics* **82**:949–958. DOI: https://doi.org/10.1016/j.ajhg.2008.02.013, PMID: 18371930

Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, McKenna A, Drier Y, Zou L, Ramos AH, Pugh TJ, Stransky N, Helman E, Kim J, et al. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**:214–218. DOI: https://doi.org/10.1038/nature12213, PMID: 23770567

Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES, Getz G. 2014. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**:495–501. DOI: https://doi.org/10.1038/nature12912, PMID: 24390350

Lawrence MS, Sougnez C, Lichtenstein L, Cibulskis K, Lander E, Gabriel SB. 2015. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**:576–582. DOI: https://doi.org/10.1038/nature14129, PMID: 25631445

Leiserson MDM, Vandin F, Wu H-T, Dobson JR, Eldridge JV, Thomas JL, Papoutsaki A, Kim Y, Niu B, McLellan M, Lawrence MS, Gonzalez-Perez A, Tamborero D, Cheng Y, Ryslik GA, Lopez-Bigas N, Getz G, Ding L, Raphael BJ. 2015. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics* **47**:106–114. DOI: https://doi.org/10.1038/ng.3168, PMID: 25501392

Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. 2015. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Systems* **1**:417–425. DOI: https://doi.org/10.1016/j.cels.2015.12.004, PMID: 26771021

Litan A, Langhans SA. 2015. Cancer as a channelopathy: ion channels and pumps in tumor development and progression. *Frontiers in Cellular Neuroscience* **9**:86. DOI: https://doi.org/10.3389/fncel.2015.00086, PMID: 4362317

Liu D, Schilling B, Liu D, Sucker A, Livingstone E, Jerby-Arnon L, Zimmer L, Gutzmer R, Satzger I, Loquai C, Grabbe S, Vokes N, Margolis CA, Conway J, He MX, Elmarakeby H, Dietlein F, Miao D, Tracy A, Gogas H, et al. 2019. Integrative molecular and clinical modeling of clinical outcomes to PD1 blockade in patients with metastatic melanoma. *Nature Medicine* **25**:1916–1927. DOI: https://doi.org/10.1038/s41591-019-0654-5, PMID: 31792460

Mammoto T, Jiang A, Jiang E, Panigrahy D, Kieran MW, Mammoto A. 2013. Role of collagen matrix in tumor angiogenesis and glioblastoma multiforme progression. *The American Journal of Pathology* **183**:1293–1305. DOI: https://doi.org/10.1016/j.ajpath.2013.06.026, PMID: 23928381

Martelotto LG, Ng CK, De Filippo MR, Zhang Y, Piscuoglio S, Lim RS, Shen R, Norton L, Reis-Filho JS, Weigelt B. 2014. Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. *Genome Biology* **15**:84. DOI: https://doi.org/10.1186/s13059-014-0484-1, PMID: 4232638

Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, Davies H, Stratton MR, Campbell PJ. 2017. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**:1029-1041.. DOI: https://doi.org/10.1016/j.cell.2017.09.042, PMID: 29056346

Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, Mularoni L, Pich O, Bonet J, Kranas H, Gonzalez-Perez A, Lopez-Bigas N. 2020. A compendium of mutational cancer driver genes. *Nature Reviews Cancer* **20**:555–572. DOI: https://doi.org/10.1038/s41568-020-0290-x, PMID: 32778778

Maslov S, Sneppen K. 2002. Specificity and Stability in Topology of Protein Networks. *Science* **296**:910–913. DOI: https://doi.org/10.1126/science.1065103, PMID: 11988575

Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. 2011. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology* **12**:R41. DOI: https://doi.org/10.1186/gb-2011-12-4-r41, PMID: 21527027

Merritt MA, Cramer DW. 2010. Molecular pathogenesis of endometrial and ovarian cancer. *Cancer Biomarkers* **9**:287–305. DOI: https://doi.org/10.3233/CBM-2011-0167, PMID: 22112481

Moilanen JM, Löffek S, Kokkonen N, Salo S, Väyrynen JP, Hurskainen T, Manninen A, Riihilä P, Heljasvaara R, Franzke C-W, Kähäri V-M, Salo T, Mäkinen MJ, Tasanen K. 2017. Significant Role of Collagen XVII And Integrin

β4 in Migration and Invasion of The Less Aggressive Squamous Cell Carcinoma Cells. *Scientific Reports* **7**:45057. DOI: https://doi.org/10.1038/srep45057, PMID: 28327550

Nishimura D. 2001. BioCarta. *Biotech Software & Internet Report* **2**:117–120. DOI: https://doi.org/10.1089/152791601750294344

Northcott PA, Buchhalter I, Morrissy AS, Hovestadt V, Weischenfeldt J, Ehrenberger T, Gröbner S, Segura-Wang M, Zichner T, Rudneva VA, Warnatz H-J, Sidiropoulos N, Phillips AH, Schumacher S, Kleinheinz K, Waszak SM, Erkek S, Jones DTW, Worst BC, Kool M, et al. 2017. The whole-genome landscape of medulloblastoma subtypes. *Nature* **547**:311–317. DOI: https://doi.org/10.1038/nature22973, PMID: 28726821

Oliver S. 2000. Guilt-by-association goes global. *Nature* **403**:601–603. DOI: https://doi.org/10.1038/35001165, PMID: 10688178

Ozturk K, Dow M, Carlin DE, Bejar R, Carter H. 2018. The Emerging Potential for Network Analysis to Inform Precision Cancer Medicine. *Journal of Molecular Biology* **430**:2875–2899. DOI: https://doi.org/10.1016/j.jmb.2018.06.016, PMID: 29908887

Page L, Brin S, Motwani R, Winograd T. 1999. The PageRank Citation Ranking: Bringing Order to the Web. California, United States: Stanford InfoLab. http://ilpubs.stanford.edu:8090/422/

Paul MR, Huo Y, Liu A, Lesperance J, Garancher A, Wechsler-Reya RJ, Zage PE. 2020. Characterization of G-CSF receptor expression in medulloblastoma. *Neuro-Oncology Advances* **2**:62. DOI: https://doi.org/10.1093/noajnl/vdaa062, PMID: 32642714

Petrov I. 2022. NEArender-2.x. swh:1:rev:5829beb819c689790359f199547362a31d1a1d54. GitHub. https://github.com/AveViort/NEArender-2.x

Phan NN, Wang CY, Chen CF, Sun Z, Lai MD, Lin YC. 2017. Voltage-gated calcium channels: Novel targets for cancer therapy. *Oncology Letters* **14**:2059–2074. DOI: https://doi.org/10.3892/ol.2017.6457, PMID: 28781648

Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C. 2008. WikiPathways: pathway editing for the people. *PLOS Biology* **6**:e184. DOI: https://doi.org/10.1371/journal.pbio.0060184, PMID: 18651794

Pietsch T, Mempel K, Menzel Th, Öckler R, Welte K. 2008. Medulloblastoma Cells Constitutively Produce Granulocyte Colony-Stimulating Factor*. *Klinische Pädiatrie* **202**:235–239. DOI: https://doi.org/10.1055/s-2007-1025526, PMID: 1697636

Pugh TJ, Weeraratne SD, Archer TC, Pomeranz Krummel DA, Auclair D, Bochicchio J, Carneiro MO, Carter SL, Cibulskis K, Erlich RL, Greulich H, Lawrence MS, Lennon NJ, McKenna A, Meldrim J, Ramos AH, Ross MG, Russ C, Shefler E, Sivachenko A, et al. 2012. Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. *Nature* **488**:106–110. DOI: https://doi.org/10.1038/nature11329, PMID: 22820256

Remke M, Ramaswamy V, Taylor MD. 2013. Medulloblastoma molecular dissection: the way toward targeted therapy. *Current Opinion in Oncology* **25**:674–681. DOI: https://doi.org/10.1097/CCO.0000000000000008, PMID: 24076581

Reva B, Antipin Y, Sander C. 2011. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Research* **39**:e118. DOI: https://doi.org/10.1093/nar/gkr407, PMID: 21727090

Reyna MA, Leiserson MDM, Raphael BJ. 2018. Hierarchical HotNet: identifying hierarchies of altered subnetworks. *Bioinformatics* **34**:i972–i980. DOI: https://doi.org/10.1093/bioinformatics/bty613, PMID: 30423088

Reyna MA, Haan D, Paczkowska M, Verbeke LPC, Vazquez M, Kahraman A, Pulido-Tamayo S, Barenboim J, Wadi L, Dhingra P, Shrestha R, Getz G, Lawrence MS, Pedersen JS, Rubin MA, Wheeler DA, Brunak S, Izarzugaza JMG, Khurana E, Marchal K, et al. 2020. Pathway and network analysis of more than 2500 whole cancer genomes. *Nature Communications* **11**:729. DOI: https://doi.org/10.1038/s41467-020-14367-0, PMID: 32024854

Risteli L, Koivula MK, Risteli J. 2014. Procollagen assays in cancer. *Advances in Clinical Chemistry* **66**:79–100. DOI: https://doi.org/10.1016/b978-0-12-801401-1.00003-7, PMID: 25344986

Robinson G, Parker M, Kranenburg TA, Lu C, Chen X, Ding L, Phoenix TN, Hedlund E, Wei L, Zhu X, Chalhoub N, Baker SJ, Huether R, Kriwacki R, Curley N, Thiruvenkatam R, Wang J, Wu G, Rusch M, Hong X, et al. 2012. Novel mutations target distinct subgroups of medulloblastoma. *Nature* **488**:43–48. DOI: https://doi.org/10.1038/nature11213, PMID: 22722829

Rodchenkov I, Babur O, Luna A, Aksoy BA, Wong JV, Fong D, Franz M, Siper MC, Cheung M, Wrana M, Mistry H, Mosier L, Dlin J, Wen Q, O'Callaghan C, Li W, Elder G, Smith PT, Dallago C, Cerami E, et al. 2020. Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. *Nucleic Acids Research* **48**:D489–D497. DOI: https://doi.org/10.1093/nar/gkz946, PMID: 31647099

Ross JS, Ali SM, Fasan O, Block J, Pal S, Elvin JA, Schrock AB, Suh J, Nozad S, Kim S, Jeong Lee H, Sheehan CE, Jones DM, Vergilio J-A, Ramkissoon S, Severson E, Daniel S, Fabrizio D, Frampton G, Miller VA, et al. 2017. *ALK* Fusions in a Wide Variety of Tumor Types Respond to Anti-ALK Targeted Therapy. *The Oncologist* **22**:1444–1450. DOI: https://doi.org/10.1634/theoncologist.2016-0488, PMID: 29079636

Rousselle P, Scoazec JY. 2020. Laminin 332 in cancer: When the extracellular matrix turns signals from cell anchorage to cell movement. *Seminars in Cancer Biology* **62**:149–165. DOI: https://doi.org/10.1016/j.semcancer.2019.09.026, PMID: 31639412

Schmitt T, Ogris C, Sonnhammer ELL. 2014. FunCoup 3.0: database of genome-wide functional coupling networks. *Nucleic Acids Research* **42**:D380–D388. DOI: https://doi.org/10.1093/nar/gkt984, PMID: 24185702

Sharma Y, Miladi M, Dukare S, Boulay K, Caudron-Herger M, Groß M, Backofen R, Diederichs S. 2019. A pan-cancer analysis of synonymous mutations. *Nature Communications* **10**:1–14. DOI: https://doi.org/10.1038/s41467-019-10489-2

Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. 2012. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research* **40**:W452–W457. DOI: https://doi.org/10.1093/nar/gks539, PMID: 22689647

Stelzer G, Dalah I, Stein T, Satanower Y, Rosen N, Nativ N, Oz-Levi D, Olender T, Belinky F, Bahir I, Krug H, Perco P, Mayer B, Kolker E, Safran M, Lancet D. 2011. In-silico human genomics with GeneCards. *Human Genomics* **5**:709. DOI: https://doi.org/10.1186/1479-7364-5-6-709, PMID: 22155609

Stracquadanio G, Wang X, Wallace MD, Grawenda AM, Zhang P, Hewitt J, Zeron-Medina J, Castro-Giner F, Tomlinson IP, Goding CR, Cygan KJ, Fairbrother WG, Thomas LF, Sætrom P, Gemignani F, Landi S, Schuster-Böckler B, Bell DA, Bond GL. 2016. The importance of p53 pathway genetics in inherited and somatic cancer genomes. *Nature Reviews. Cancer* **16**:251–265. DOI: https://doi.org/10.1038/nrc.2016.15, PMID: 27009395

Sweet-Cordero EA, Biegel JA. 2019. The genomic landscape of pediatric cancers: Implications for diagnosis and treatment. *Science (New York, N.Y.)* **363**:1170–1175. DOI: https://doi.org/10.1126/science.aaw3535, PMID: 30872516

Tajada S, Villalobos C. 2020. Calcium Permeable Channels in Cancer Hallmarks. *Frontiers in Pharmacology* **11**:968. DOI: https://doi.org/10.3389/fphar.2020.00968, PMID: 32733237

Tamborero D, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandoth C, Reimand J, Lawrence MS, Getz G, Bader GD, Ding L, Lopez-Bigas N. 2013. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Scientific Reports* **3**:2650. DOI: https://doi.org/10.1038/srep02650, PMID: 24084849

The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* **45**:1113–1120. DOI: https://doi.org/10.1038/ng.2764

Torkamani A, Schork NJ. 2009. Identification of rare cancer driver mutations by network reconstruction. *Genome Research* **19**:1570–1578. DOI: https://doi.org/10.1101/gr.092833.109, PMID: 19574499

Torkamani A, Verkhivker G, Schork NJ. 2009. Cancer driver mutations in protein kinase genes. *Cancer Letters* **281**:117–127. DOI: https://doi.org/10.1016/j.canlet.2008.11.008, PMID: 19081671

Tranchevent LC, Capdevila FB, Nitsch D, De Moor B, De Causmaecker P, Moreau Y. 2011. A guide to web tools to prioritize candidate genes. *Briefings in Bioinformatics* **12**:22–32. DOI: https://doi.org/10.1093/bib/bbq007, PMID: 21278374

Tsuruta D, Kobayashi H, Imanishi H, Sugawara K, Ishii M, Jones JCR. 2008. Laminin-332-integrin interaction: A target for cancer therapy? *Current Medicinal Chemistry* **15**:1968–1975. DOI: https://doi.org/10.2174/092986708785132834, PMID: 18691052

Vermeulen JF, Van Hecke W, Adriaansen EJM, Jansen MK, Bouma RG, Villacorta Hidalgo J, Fisch P, Broekhuizen R, Spliet WGM, Kool M, Bovenschen N. 2017. Prognostic relevance of tumor-infiltrating lymphocytes and immune checkpoints in pediatric medulloblastoma. *OncoImmunology* **7**:e1398877. DOI: https://doi.org/10.1080/2162402X.2017.1398877, PMID: 29399402

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. 2013. Cancer genome landscapes. *Science (New York, N.Y.)* **339**:1546–1558. DOI: https://doi.org/10.1126/science.1235122, PMID: 23539594

Wheeler DA, Wang L. 2013. From human genome to cancer genome: the first decade. *Genome Research* **23**:1054–1062. DOI: https://doi.org/10.1101/gr.157602.113, PMID: 23817046

Winter C, Kristiansen G, Kersting S, Roy J, Aust D, Knösel T, Rümmele P, Jahnke B, Hentrich V, Rückert F, Niedergethmann M, Weichert W, Bahra M, Schlitt HJ, Settmacher U, Friess H, Büchler M, Saeger H-D, Schroeder M, Pilarsky C, et al. 2012. Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLOS Computational Biology* **8**:e1002511. DOI: https://doi.org/10.1371/journal.pcbi.1002511, PMID: 22615549