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

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

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

[§]In memory of Alexey D. Vyatkin, 13.07.1997-14.02.2021
*Corresponding author: belikov.research@gmail.com

Abstract

Background
There is a growing need to develop novel therapeutics for targeted treatment of cancer. The prerequisite to success is the knowledge about which types of molecular alterations are predominantly driving tumorigenesis – single nucleotide (SNA) or copy number (CNA), in oncogenes or in tumour suppressors, gains or losses of full chromosomes or chromosomal arms (aneuploidy). However, the number and proportion of various types of driver events per tumour is still not clear, neither for cancer in general, nor for individual cancer types, stages and patient demographics (age and gender).

Methods
To shed light onto this subject, we have utilized the largest database of human cancer mutations – TCGA PanCanAtlas, multiple established algorithms for cancer driver prediction (2020plus, CHASMplus, CompositeDriver, dNdScv, DriverNet, HotMAPS, IntOGen Plus, OncodriveCLUSTL, OncodriveFML) and developed four novel computational pipelines: SNADRIF (SNA DRIver Finder), GECNAV (Gene Expression-based CNA Validator), ANDRIF (ANeuploidy DRIver Finder) and PALDRIC (PAtient-Level DRIver Classifier). A unified workflow integrating all these pipelines, algorithms and datasets at cohort and patient levels was created.

Results
By integrating results of various driver prediction algorithms, we have found that there are on average 20 driver events per tumour, of which 1.5 are hyperactivating SNAs in oncogenes, 10.5 are amplifications of oncogenes, 2 are homozygous inactivating SNAs or deletions of tumour suppressors, 1.5 are driver chromosome losses, 2 are driver chromosome gains, 1 is a driver chromosome arm loss, and 1.5 are driver chromosome arm gains. The average number of driver events per tumour varies strongly between cancer types, from 1.7 in thyroid carcinoma to 42.4 in ovarian carcinoma. In females, the number of driver events increases most dramatically until the age of menopause (50 y.o.), whereas in males until 70 y.o. Moreover, in females, the number of driver events increases abruptly from Stage I to Stage II, after which stays more or less constant, and this increase is due to CNAs and aneuploidy but not due to SNAs. In tumours having only one driver event, this event is a SNA in an oncogene. However, with increasing number of driver events per tumour, the contribution of SNAs and tumour suppressor events decreases, whereas the contribution of oncogene amplifications and aneuploidy events increases. Patients with two driver events per tumour are the most frequent, and there are very few patients with more than 50 events.

Conclusions
As half of all driver events in a patient's tumour appear to be amplifications of oncogenes, we suggest that future therapeutics development efforts should be focused on targeting this alteration type. Therapies aimed at gains and losses of chromosomal arms and whole chromosomes also appear very promising. On the other hand, drugs aiming at point mutations and tumour suppressors are predicted to be less successful. Overall, our results provide valuable insights into the extent of driver alterations of different types in human tumours and suggest optimal targets for candidate therapeutics.

Introduction

Driver events are the molecular and cellular events that drive cancer progression, often called driver mutations (Pon & Marra, 2015). The knowledge about the spectrum and quantity of driver events in individual patients and groups of patients is crucial to inform design and selection of targeted therapeutics. Historically, most attention has been devoted to point mutations or single nucleotide alterations (SNAs), as most driver prediction algorithms work only with this class of driver events. As the average numbers of driver SNAs per tumour in various cancer types have been estimated before and shown to vary within 1-10 range, with the average across all cancer types of 4 (Martincorena et al., 2017), this defined current thinking that the number of driver events per tumour is relatively low.

However, SNA drivers represent only the tip of the iceberg, and are likely not the most crucial contribution to cancer progression. SNAs are very well tolerated by cancer cells (Martincorena et al., 2017). It is known that cancer cells contain large numbers of deletions and amplifications (often called copy number alterations, or CNAs), translocations, inversions, full chromosome and chromosomal arm gains and losses (aneuploidy), as well as epigenetic modifications (Ben-David & Amon, 2020; Dawson & Kouzarides, 2012; Zack et al., 2013), but the driver potential of these alterations has been left almost unexplored due to the scarcity of driver prediction algorithms for these classes of events. Here, we set the goal to predict, classify and quantify as many different classes of driver events as possible, using clear and straightforward principles, and developed custom computational pipelines for this purpose.

Another shortcoming of the majority of existing driver prediction algorithms is that they work at the cohort level, i.e. they predict driver mutations for large groups of patients, usually having a particular cancer type. This does not allow to look at the composition of driver events in individual patients. We wrote specific scripts to convert cohort-level predictions into patient-level events, which also allowed seamless integration of the results from various third-party algorithms. This is useful, as each individual driver prediction algorithm has its own strengths and shortcomings, and combining results from multiple algorithms allows to obtain more complete and balanced picture, ensuring that less driver mutations have been missed. Having patient-level data is also convenient for analysing driver event composition in various demographic and clinical groups (e.g. patients grouped by age, sex or cancer stage) without the need to rerun the driver prediction algorithms on subgroups of patients, which would otherwise not only consume additional time and resources but, most importantly, reduce the predictive power. Patient-level data also allow to study heterogeneity of cancer patient cohorts, e.g. by using distribution histograms.

Finally, to properly identify and count driver events at the individual patient level it is required to differentiate between oncogenes and tumour suppressors (Weinberg, 1994), as the former require only one molecular alteration (SNA or amplification) to become hyperactivated, whereas the latter require two. This is because both alleles of a tumour suppressor need to be inactivated for it to become a true functional driver. Thus, homozygous inactivation by two SNAs, two deletions or one SNA and one deletion is required. Tumour suppressor genes with only one allele affected should not be counted as drivers. As only a few existing driver prediction algorithms perform classification of genes into oncogenes and tumour suppressors, and even they leave a considerable percentage of genes unclassified, we devised and implemented a custom algorithm for this task, which allows to classify all genes in a dataset.

Overall, we created a universal workflow (Fig. 1) that ingests various TCGA PanCanAtlas datasets – SNA, CNA, mRNA, miRNA, aneuploidy, as well as sample quality and clinical annotations – and standardized results from third-party driver prediction algorithms, such as 2020plus, CHASMplus, CompositeDriver, dNdScv, DriverNet, HotMAPS, IntOGen Plus, OncodriveCLUSTL, and OncodriveFML, and then uniformly processes them through our four custom built pipelines - SNADRIF (SNA DRIver Finder), GECNAV (Gene Expression-based CNA Validator), ANDRIF (ANeuploidy DRIver Finder) and PALDRIC (PAtient-Level DRIver Classifier). These scripts are made publicly available, and allow anyone to conduct their own patient-level analyses of PanCanAtlas data, with the possibility of importing results from third-party driver prediction algorithms.

Methods

Source files and initial filtering

TCGA PanCanAtlas data were used. Files "Analyte level annotations - merged_sample_quality_annotations.tsv", "ABSOLUTE purity/ploidy file - TCGA_mastercalls.abs_tables_JSedit.fixed.txt", "Aneuploidy scores and arm calls file - PANCAN_ArmCallsAndAneuploidyScore_092817.txt", "Public mutation annotation file - mc3.v0.2.8.PUBLIC.maf.gz", "gzipped ISAR-corrected GISTIC2.0 all_thresholded.by_genes file - ISAR_GISTIC.all_thresholded.by_genes.txt", "RNA batch corrected matrix - EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2.geneExp.tsv", "miRNA batch corrected matrix - pancanMiRs_EBadjOnProtocolPlatformWithoutRepsWithUnCorrectMiRs_08_04_16.csv", were downloaded from https://gdc.cancer.gov/about-data/publications/PanCan-CellOfOrigin.

Using TCGA barcodes (see https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/ and https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/sample-type-codes), all samples except primary tumors (barcoded 01, 03, 09) were removed from all files. Based on the information in the column "Do_not_use" in the file "Analyte level annotations - merged_sample_quality_annotations.tsv", all samples with "True" value were removed from all files. All samples with "Cancer DNA fraction" <0.5 or unknown or with "Subclonal genome fraction" >0.5 or unknown in the file "TCGA_mastercalls.abs_tables_JSedit.fixed.txt" were removed from the file "PANCAN_ArmCallsAndAneuploidyScore_092817.txt". Moreover, all samples without "PASS" value in the column "FILTER" were removed from the file "mc3.v0.2.8.PUBLIC.maf.gz" and zeros in the column "Entrez_Gene_Id" were replaced with actual Entrez gene IDs, determined from the corresponding ENSEMBL gene IDs in the column

"Gene" and external database https://www.genenames.org/download/custom/. Filtered files were saved as "Primary_whitelisted_arms.tsv", "mc3.v0.2.8.PUBLIC_primary_whitelisted_Entrez.tsv", "ISAR_GISTIC.all_thresholded.by_genes_primary_whitelisted.tsv", "EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2-v2.geneExp_primary_whitelisted.tsv", "pancanMiRs_EBadjOnProtocolPlatformWithoutRepsWithUnCorrectMiRs_08_04_16_primary_whitelisted.tsv".

RNA filtering of CNAs

Using the file "EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2-v2.geneExp_primary_whitelisted.tsv", the median expression level for each gene across patients was determined. If the expression for a given gene in a given patient was below 0.05x median value, it was encoded as "-2", if between 0.05x and 0.75x median value, it was encoded as "-1", if between 1.25x and 1.75x median value, it was encoded as "1", if above 1.75x median value, it was encoded as "2", otherwise it was encoded as "0". The file was saved as "EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2-v2.geneExp_primary_whitelisted_median.tsv." The same operations were performed with the file "pancanMiRs_EBadjOnProtocolPlatformWithoutRepsWithUnCorrectMiRs_08_04_16_primary_whitelisted.tsv", which was saved as "pancanMiRs_EBadjOnProtocolPlatformWithoutRepsWithUnCorrectMiRs_08_04_16_primary_whitelisted_median.tsv"
Next, the file "ISAR_GISTIC.all_thresholded.by_genes_primary_whitelisted.tsv" was processed according to the following rules: if the gene CNA status in a given patient was not zero and had the same sign as the gene expression status in the same patient (file "EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2-v2.geneExp_primary_whitelisted_median.tsv" or "pancanMiRs_EBadjOnProtocolPlatformWithoutRepsWithUnCorrectMiRs_08_04_16_primary_whitelisted_median.tsv" for miRNA genes), then the CNA status value was replaced with the gene expression status value, otherwise it was replaced by zero. If the corresponding expression status for a given gene was not found then its CNA status was not changed. The resulting file was saved as "ISAR_GISTIC.all_thresholded.by_genes_primary_whitelisted_RNAfiltered.tsv"

We named this algorithm GECNAV (Gene Expression-based CNA Validator) and created a Github repository: https://github.com/belikov-av/GECNAV

Aneuploidy driver prediction

Using the file "Primary_whitelisted_arms.tsv", the average alteration status of each chromosomal arm was calculated for each cancer type and saved as a matrix file "Arm_averages.tsv". By drawing statuses randomly with replacement (bootstrapping) from any cell of "Primary_whitelisted_arms.tsv", for each cancer type the number of statuses corresponding to the number of patients in that cancer type were generated and their average was calculated. The procedure was repeated 10000 times, the median for each cancer type was calculated and the results were saved as a matrix file "Bootstrapped_arm_averages.tsv".

P-value for each arm alteration status was calculated for each cancer type. To do this, first the alteration status for a given cancer type and a given arm in "Arm_averages.tsv" was compared to the median bootstrapped arm alteration status for this cancer type in "Bootstrapped_arm_averages.tsv". If the status in "Arm_averages.tsv" was higher than zero and the median in "Bootstrapped_arm_averages.tsv", the number of statuses for this cancer type in "Bootstrapped_arm_averages.tsv" that are higher than the status in "Arm_averages.tsv" was counted and divided by 5000. If the status in "Arm_averages.tsv" was lower than zero and the median in "Bootstrapped_arm_averages.tsv", the number of statuses for this cancer type in "Bootstrapped_arm_averages.tsv" that are lower than the status in "Arm_averages.tsv" was counted and divided by 5000, and marked with minus to indicate arm loss. Other values were ignored (cells left empty). The results were saved as a matrix file "Arm_Pvalues_cohorts.tsv".

For each cancer type, Benjamini–Hochberg procedure with FDR=5% was applied to P-values in "Arm_Pvalues_cohorts.tsv" and passing P-values were encoded as "DAG" (Driver arm gain) or "DAL" (Driver arm loss) if marked with minus. The other cells were made empty and the results were saved as a matrix file "Arm_drivers_FDR5_cohorts.tsv".

Alterations were classified according to the following rules: if the arm status in a given patient (file "Primary_whitelisted_arms.tsv") was "-1" and the average alteration status of a given arm in the same cancer type (file "Arm_drivers_FDR5_cohorts.tsv") was "DAL", then the alteration in the patient was classified as "DAL". If the arm status in a given patient was "1" and the average alteration status of a given arm in the same cancer type was "DAG", then the alteration in the patient was classified as "DAG". In all other cases an empty cell was written. The total number of DALs and DAGs was calculated, patients with zero drivers were removed, and the results were saved as a matrix file "Arm_drivers_FDR5.tsv".

Using the file "Primary_whitelisted_arms.tsv", the values for the whole chromosomes were calculated using the following rules: if both p- and q-arm statuses were "1" then the chromosome status was written as "1"; if both p- and q-arm statuses were "-1" then the chromosome status was written as "-1"; if at least one arm status was not known (empty cell) then the chromosome status was written as empty cell; in all other cases the chromosome status was written as "0". For one-arm chromosomes (13, 14, 15, 21, 22), their status equals the status of the arm. The resulting file was saved as "Primary_whitelisted_chromosomes.tsv".

The same procedures as described above for chromosomal arms were repeated for the whole chromosomes, with the resulting file "Chromosome_drivers_FDR5.tsv". Chromosome drivers were considered to override arm drivers, so if a chromosome had "DCL" (Driver chromosome loss) or "DCG" (Driver chromosome gain), no alterations were counted on the arm level, to prevent triple counting of the same event.

We named this algorithm ANDRIF (ANeuploidy DRIver Finder) and created a Github repository: https://github.com/belikov-av/ANDRIF

SNA driver prediction

Using the file "mc3.v0.2.8.PUBLIC_primary_whitelisted_Entrez.tsv" all SNAs were classified according to the column "Variant_Classification". "Frame_Shift_Del", "Frame_Shift_Ins", "Nonsense_Mutation", "Nonstop_Mutation" and "Translation_Start_Site" were considered potentially inactivating; "De_novo_Start_InFrame", "In_Frame_Del", "In_Frame_Ins" and "Missense_Mutation" were considered potentially hyperactivating; "De_novo_Start_OutOfFrame" and "Silent" were considered passengers; the rest were considered unclear. The classification results were saved as the file "SNA_classification_patients.tsv", with columns "Tumor_Sample_Barcode", "Hugo_Symbol", "Entrez_Gene_Id", "Gene", "Number of hyperactivating SNAs", "Number of inactivating SNAs", "Number of SNAs with unclear role", "Number of passenger SNAs".

Using this file, the sum of all alterations in all patients was calculated for each gene. Next, the Nonsynonymous SNA Enrichment Index (NSEI) was calculated for each gene as

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

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

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

saving the results as "SNA_classification_genes_NSEI_HISR.tsv".

Using the file "SNA_classification_patients.tsv", the gene-patient matrix "SNA_matrix.tsv" was constructed, encoding the "Number of hyperactivating SNAs", "Number of inactivating SNAs", "Number of SNAs with unclear role" and "Number of passenger SNAs" as one number separated by dots (e.g. "2.0.1.1"). If data for a given gene were absent in a given patient, it was encoded as "0.0.0.0". By drawing statuses randomly with replacement (bootstrapping) from any cell of "SNA_matrix.tsv" 10000 times for each patient, the matrix file "SNA_matrix_bootstrapped.tsv" was created. The sums of statuses in "SNA_matrix_bootstrapped.tsv" were calculated for each iteration separately, and then the corresponding NSEI and HISR indices were calculated and the results were saved as "SNA_bootstrapped_NSEI_HISR.tsv".

P-value for each gene was calculated as the number of NSEI values in "SNA_bootstrapped_NSEI_HISR.tsv" more extreme than its NSEI value in "SNA_classification_genes_NSEI_HISR.tsv" and divided by 10000. The results were saved as "SNA_classification_genes_NSEI_HISR_Pvalues.tsv". Benjamini–Hochberg procedure with FDR(Q)=5% was applied to P-values in "SNA_classification_genes_NSEI_HISR_Pvalues.tsv", and genes that pass were saved as "SNA_driver_gene_list_FDR5.tsv".

We named this algorithm SNADRIF (SNA DRIver Finder) and created a Github repository: https://github.com/belikov-av/SNADRIF

Driver prediction algorithms and conversion to the patient level

Lists of driver genes and mutations predicted by various algorithms (**Table 1**) applied to PanCanAtlas data were downloaded from https://gdc.cancer.gov/about-data/publications/pancan-driver (2020plus, CompositeDriver, DriverNet, HotMAPS, OncodriveFML), https://karchinlab.github.io/CHASMplus (CHASMplus), as well as received by personal communication from Francisco Martínez-Jiménez, Institute for Research in Biomedicine, Barcelona, francisco.martinez@irbbarcelona.org (dNdScv, IntOGen Plus, OncodriveCLUSTL, OncodriveFML). Additionally, a consensus driver gene list from 26 algorithms applied to PanCanAtlas data (Bailey et al., 2018) was downloaded from https://www.cell.com/cell/fulltext/S0092-8674(18)30237-X. All genes and mutations with q-value > 0.05 were removed. Entrez Gene IDs were identified for each gene using HUGO Symbol and external database ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz.

Table 1

Name	Ref.	Repository	Level	Principles
ANDRIF	This paper	https://github.com/belikov-av/ANDRIF	Chromosomal arm, chromosome	Recurrence
20/20plus	(Tokheim et al., 2016)	https://github.com/KarchinLab/2020plus	gene	Machine learning, trained on Cancer Genome Landscapes (20/20 rule); Nonsynonymous/Synonymous, clustering, conservation (uses UCSC's 46-way vertebrate alignment and SNVBox), impact (uses VEST), network (uses BioGrid), expression, chromatin, replication (uses MutSigCV)
CHASMplus	(Tokheim & Karchin, 2019)	https://github.com/KarchinLab/CHASMplus	mutation	Machine learning, trained on TCGA; clustering (uses HotMAPS 1D), conservation (uses UCSC Multiz-100-way and SNV box), network (uses Interactome Insider)
CompositeDriver	(Bailey et al., 2018)	https://github.com/mil2041/CompositeDriver	gene	Recurrence, impact (uses FunSeq2)
dNdScv	(Martincorena et al., 2017)	https://github.com/im3sanger/dndscv	gene	Nonsynonymous/Synonymous
DriverNet	(Bashashati et al., 2012)	https://github.com/shahcompbio/drivernet https://bioconductor.org/packages/release/bioc/html/DriverNet.html	gene	Network (uses MGSA and a human functional protein interaction network), impact (uses gene expression outliers)
HotMAPS	(Tokheim et al., 2016)	https://github.com/karchinlab/HotMAPS	mutation	3D clustering (uses Protein Data Bank and ModPipe)
IntOGen Plus	(Martínez-Jiménez et al., 2020)	https://www.intogen.org/search https://bitbucket.org/intogen/intogen-plus/src/master/	gene	Combination of dNdScv, CBaSE, OncodriveCLUSTL, HotMAPS, smRegions and OncodriveFML
OncodriveCLUSTL	(Arnedo-Pac et al., 2019)	http://bbglab.irbbarcelona.org/oncodriveclustl/analysis	gene	Clustering

		https://bitbucket.org/bbglab/oncodriveclustl/src/master/		
OncodriveFML	(Mularoni et al., 2016)	http://bbglab.irbbarcelona.org/oncodrivefml/analysis https://bitbucket.org/bbglab/oncodrivefml/src/master/	gene	Recurrence, Impact (uses CADD and RNAsnp)
SNADRIF	This paper	https://github.com/belikov-av/SNADRIF	gene	Nonsynonymous/Synonymous

To convert these population-level data to patient-level data, the following procedures were performed.

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

For lists of driver *mutations*, the procedures were the same, except that Ensembl Transcript ID and nucleotide/amino acid substitution were used for matching instead of Entrez Gene ID. These data (only columns "TCGA Barcode", "HUGO Symbol", "Entrez Gene ID") were saved as "AlgorithmName_output_SNA.tsv".

Additionally, all entries from the file "ISAR_GISTIC.all_thresholded.by_genes_primary_whitelisted.tsv" were removed except those that satisfied the following conditions simultaneously: "Locus ID" matches "Entrez Gene ID" in the driver list; cancer type (identified by matching Tumor Sample Barcode with "bcr_patient_barcode" and "acronym" in "clinical_PANCAN_patient_with_followup.tsv") matches "cohort" in the driver list or the driver list is for pancancer analysis; CNA values are "2", "1", "-1" or "-2". These data were converted from the matrix to a list format (with columns "TCGA Barcode", "HUGO Symbol", "Entrez Gene ID") and saved as "AlgorithmName_output_CNA.tsv".

Finally, the files "AlgorithmName_output_SNA.tsv" and "AlgorithmName_output_CNA.tsv" were combined, duplicate TCGA Barcode-Entrez Gene ID pairs were removed, and the results saved as "AlgorithmName_output.tsv".

Driver event classification and analysis

The file "Clinical with Follow-up - clinical_PANCAN_patient_with_followup.tsv" was downloaded from https://gdc.cancer.gov/node/905/. All patients with "icd_o_3_histology" different from XXXX/3 (primary malignant neoplasm) were removed, as well as all patients not simultaneously present in the following three files:
"mc3.v0.2.8.PUBLIC_primary_whitelisted_Entrez.tsv",

“ISAR_GISTIC.all_thresholded.by_genes_primary_whitelisted.tsv” and “Primary_whitelisted_arms.tsv”. The resulting file was saved as “clinical_PANCAN_patient_with_followup_primary_whitelisted.tsv”.

Several chosen “AlgorithmName_output.tsv” files were combined and duplicate TCGA Barcode-Entrez Gene ID pairs removed. Entries with TCGA Barcodes not present in “clinical_PANCAN_patient_with_followup_primary_whitelisted.tsv” were removed as well. Matching “Number of hyperactivating SNAs” and “Number of inactivating SNAs” for each TCGA Barcode-Entrez Gene ID pair were taken from the “SNA_classification_patients.tsv” file, in case of no match zeros were written. Matching HISR value was taken from “SNA_classification_genes_NSEI_HISR.tsv” for each Entrez Gene ID, in case of no match empty cell was left. Matching CNA status was taken from “ISAR_GISTIC.all_thresholded.by_genes_primary_whitelisted_RNAfiltered.tsv” for each TCGA Barcode-Entrez Gene ID pair, in case of no match zero was written.

Each TCGA Barcode-Entrez Gene ID pair was classified according to the **Table 2**:

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

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

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

Using the file "AnalysisName_patients.tsv", the number of patients with each integer total number of driver events from 0 to 50 was counted for each cancer type, also for males and females separately, and histograms with 5-bin moving average were plotted for each cancer type-gender combination. Using the same file "AnalysisName_patients.tsv", the average number of various types of driver events was calculated for each cancer type, tumor stage, age group, as well as for patients with each integer total number of driver events from 1 to 50. Analyses were performed for total population and for males and females separately, and cumulative histograms were plotted for each file.

We named this algorithm PALDRIC (PAtient-Level DRIver Classifier) and created a Github repository: https://github.com/belikov-av/PALDRIC

Results

Although there exist many driver prediction algorithms working with SNA data, we decided to create our own using clear and simple rules to have an internal reference standard. We called this algorithm SNADRIF - SNA DRIver Finder. It is a Python 3.7 software package that predicts cancer driver genes from the TCGA PanCanAtlas SNA data and classifies them into oncogenes and tumour suppressors. Driver prediction is based on calculating the ratio of nonsynonymous SNAs to silent SNAs (Martincorena et al., 2017), whereas driver classification is based on calculating the ratio of hyperactivating SNAs to inactivating SNAs (Vogelstein et al., 2013). Bootstrapping is used to calculate statistical significance and Benjamini–Hochberg procedure is used to keep false discovery rate under 5%.

As CNA data is obtained using quite crude methods that define large chunks of DNA as deleted or amplified, we wanted to clarify the actual copy number status of individual genes using mRNA and miRNA expression data available at TCGA PanCanAtlas. For this purpose, we created another Python 3.7 software package called GECNAV - Gene Expression-based CNA Validator. CNA validation is based on comparing the CNA status of a given gene in a given patient to expression of this gene in this patient relative to the median expression of this gene across all patients.

As we have not found any existing algorithms for predicting driver aneuploidies, we created our own called ANDRIF - ANeuploidy DRIver Finder. It is a Python 3.7 software package that predicts driver chromosomal arm or full chromosome gains or losses from the TCGA PanCanAtlas aneuploidy data. Driver prediction is based on calculating the average alteration status for each arm or chromosome in each cancer type. Bootstrapping is used to obtain the realistic distribution of the average alteration statuses under the null hypothesis. Benjamini–Hochberg procedure is performed to keep the false discovery rate under 5%.

Finally, we needed an algorithm to integrate all data on driver alterations from different algorithms - our own and third-party. We called this algorithm PALDRIC - PAtient-Level DRIver Classifier. It is a Python 3.7 software package that translates cohort-level lists of driver genes or mutations to the patient level, classifies driver events according to the molecular causes and functional consequences, and presents comprehensive statistics on various kinds of driver events in various demographic and clinical groups of patients. PALDRIC allows to natively combine outputs of various third-party algorithms to investigate optimal combinations.
Our overall workflow can be seen in **Fig. 1**.

To get three different perspectives on the number and composition of driver events, we performed three analyses. In the first one we used results of our own SNADRIF algorithm applied to the whole TCGA PanCanAtlas dataset. We will refer to this analysis as "Pancan SNADRIF". In the second analysis we used the combination of results from several third-party algorithms - 2020plus, CHASMplus, CompositeDriver, OncodriveFML and consensus results from (Bailey et al., 2018) – also applied to the whole TCGA PanCanAtlas dataset. We will refer to this analysis as "Pancan combined". In the third analysis we used the combination of results from 2020plus, CHASMplus, CompositeDriver, dNdScv, DriverNet, HotMAPS, IntOGen, OncodriveCLUSTL, OncodriveFML and consensus results from (Bailey et al., 2018), applied separately to each cancer cohort of TCGA PanCanAtlas. This approach is supposed to discover cancer type-specific drivers, but might have decreased power due to fewer patients analysed. However, combining the results of several algorithms with different underlying principles increases chances of discovery of all relevant drivers. We will refer to this analysis as "Cohorts combined".

Fig. 2 shows the average number of driver events of different classes in different cancer cohorts. It can be immediately seen that there is a dramatic difference in the number of driver events per patient between cancer types, regardless of the way we analysed data. However, the type of analysis affects the relative distributions as well as the absolute values. Indeed, "Pancan SNADRIF" and "Pancan combined" have very similar distributions but "Pancan combined" has ~30% higher absolute values. Interestingly, "Pancan SNADRIF" and "Cohorts combined" have similar absolute values despite having different distributions.

It can also be seen that the relative contribution of the different types of driver events is generally similar between cancer types and analysis methods. Clearly, the majority of driver events can be attributed to amplifications of oncogenes. Hyperactivating point mutations comprise a much smaller percentage of total events, especially in "Pancan SNADRIF" analysis. Inactivation of tumour suppressors is similarly rare. However, a quite large contribution comes from aneuploidy events (these data did not differ between analysis types as they come directly from ANDRIF pipeline). "Pancan SNADRIF" tends to reveal fewer point mutations in oncogenes and tumour suppressors but a higher number of homozygous deletions in tumour suppressors, compared to other methods. It is also interesting to note that while the SNA event distribution of "Pancan combined" is similar to "Cohorts combined", the distribution of the total number of events is nevertheless most similar to "Pancan SNADRIF".

The distributions for males and females separately are available as **Supplementary Figures.**

Fig. 3, 4 and **5** show the distribution of driver events in various demographic and clinical groups of patients. **Fig. 3** demonstrates that there is no clear difference in driver event distribution between males and females, besides females having a higher number of SNAs participating in homozygous tumour suppressor inactivation. This is most likely due to their overabundance in UCEC (see **Fig. 2**).

Fig. 4 demonstrates an increase in the number of driver events with age. It can be seen that in female patients the number of driver events increases dramatically from the earliest age until approximately 50 y.o., i.e. the age of menopause. After this the growth slows down. In males the picture is different. There is a high number of driver events in the earliest age group, due to the frequent occurrence of testicular cancer (TGCT) in this age group and a high number of

driver events in this cancer type (see **Fig. 2**). The next age group has much lower number of drivers, and like in females, their number increases with age, but until an older age, 70 y.o., where it reaches a plateau.

Fig. 5 demonstrates that in females, the number of driver events increases dramatically from Stage I to Stage II, after which stays more or less constant or even decreases. It is interesting that this increase is due to CNAs and aneuploidy but not due to point mutations. The number of SNAs actually decreases. In males, the increase in driver events with progression from Stage I to Stage IV is more gradual and chaotic. For unclear reason there appears to be less driver events in Stage III than in Stage II tumours. However, Stage IV tumours have more events than Stage II. As in females, the increase is due to CNAs and aneuploidy but not due to point mutations.

In **Fig. 6** we aimed to show how different classes of driver events increase proportionally to the total number of driver events in a patient. It is interesting to note that contribution from aneuploidy and amplifications of oncogenes increases whereas contribution from tumour suppressor events and hyperactivating point mutations in oncogenes actually decreases. This is even more striking, considering that if a tumour has only one driver event it is almost certain to be a hyperactivating point mutation in an oncogene.

The distributions for males and females separately are available as **Supplementary Figures.**

Fig. 7 shows the actual distributions of driver event quantities amongst patients of all cancer types combined. It can be seen that these distributions have a sharp peak at 2 driver events per tumour. Females have an additional, smoother peak at 8-12 events. Males seem to have no additional clearly defined peaks. Further behaviour of the distributions depends on the analysis method, but generally there are very few patients with more than 50 driver events per tumour.

The distributions for individual cancer types are available as **Supplementary Figures.**

Discussion

Our results pose several interesting questions. First, why there is such a great difference in the number of driver events between cancer types? Why only few driver events are sufficient to initiate thyroid cancer but several dozen are observed in lung carcinomas? Are these differences explained by different tissue microenvironments to which these tumours have to adapt? Could it be that we are missing a lot of epigenetic drivers?

We also observe that, using our methodology to estimate the number of driver events, the set of driver prediction algorithms is not as important as whether they are applied to each cancer type (cohort) individually or to all cohorts combined (Pancan analysis). Applying algorithms to individual cohorts allows to discover cancer type-specific drivers and avoid “contamination” by false positives, i.e. driver genes discovered during Pancan analysis that do not in reality play any role in a given cancer type. On the other hand, much fewer patients are available for cohort-specific analysis, and this decreases statistical power to discover new driver genes. Of note, our SNADRIF algorithm works best for Pancan analysis and struggles with small cohorts, due to scarcity of point mutations. However, when a combination of driver prediction algorithms is used, there are less chances of missing an important driver gene even in a cohort-specific analysis, as algorithms based on differing principles complement each other.

It is a striking discovery that half of all driver events in a patient are amplifications of oncogenes. On the other hand, there is logic to it. CNAs are very common in cancer and each of CNA regions encompasses dozens of genes (Zack et al., 2013). Thus, if some gene plays a driver role in a given cancer type there is a high probability that it is within one of the CNAs. If the gene is an oncogene and the CNA is an amplification then this is a functional driver event. The situation with tumour suppressors is more complex. As both alleles of a tumour suppressor have to be inactivated to procure a functional driver, either two deletions, two inactivating SNAs or one deletion and one inactivating SNA are required. However, homozygous deletions of multiple genes are very likely to completely remove some housekeeping genes or have other negative consequences for cell survival and thus are rare. Combination of an inactivating point mutation and a deletion, as well as two inactivating SNAs in the same gene is also a statistically rare event due to relative scarcity of inactivating SNAs compared to deletions. Hyperactivating SNAs are also rare compared to amplifications. Thus, oncogene amplifications are naturally predisposed to be the most common type of functional driver alterations, and should be chosen as the main target of anticancer therapies.

It is also quite notable that more than a quarter of all driver events are due to aneuploidy, which makes them the second best target for anticancer therapeutics. Full chromosome events are more common that arm events, and gains are more common than losses. This is in accord with our finding on oncogene amplifications, as a chromosome (arm) duplication is equivalent to amplifications of all oncogenes present on this chromosome (arm), with the exception of an unintended amplification of all tumour suppressors on this chromosome (arm). Likewise, a chromosome (arm) loss is equivalent to deletions of all tumour suppressors present on this chromosome (arm), with the exception of unintended deletion of all oncogenes on this chromosome (arm). Because such aneuploidy events are combined with CNA events (Pihan & Doxsey, 2003), e.g. loss of one allele of a tumour suppressor due to the chromosome loss and loss of the other allele due to the CNA, we decided to estimate the actual copy number status of each gene using RNA expression data in addition to CNA data (see our GECNAV pipeline). In the end, changes in the actual expression of the driver gene(s) are what makes a CNA or aneuploidy event functional.

It is interesting to speculate why the number of driver events steadily increases in women until 50 y.o. and in men until 70 y.o. As 50 y.o. is the average age of menopause whereas males are fertile until much older ages, it is logical to suggest that driver accumulation is somehow promoted by high levels of sex hormones. If we concur with the theory that most driver events accumulate in stem cells due to errors in DNA replication and chromosome segregation (Tomasetti et al., 2017), and combine it with the knowledge that sex hormones increase the rate of stem cell division (Asselin-Labat et al., 2010; Joshi et al., 2010; Nakada et al., 2014), the link becomes clear. It is known that individuals with higher levels of sex hormones have higher rates of cancer (Folkerd & Dowsett, 2010).

The analysis of the distribution of various classes of driver events across cancer stages leads to an interesting observation of the sequential order in which these alterations appear. All SNA drivers appear already at the first stage and no new SNA drivers appear at the later stages. All other classes of driver events are also present already at the first stage cancer, but their quantity increases at further stages. It is especially clear in females, where CNA and aneuploidy events increase prominently at the second stage, but do not increase further at later stages. In males, they increase also at the fourth stage. Moreover, the analysis of the contribution of

different classes of driver events with increasing number of total driver events shows that the first driver event is always a hyperactivating SNA in an oncogene, and then the number of CNA and aneuploidy events increases dramatically, whereas the number of SNAs only slightly (with the possible exception of hypermutators). Thus, it appears that a hyperactivating SNA in an oncogene is the seed which triggers the appearance of CNAs and aneuploidy events, the latter two driving the progression of cancer. It is therefore even more important to develop therapies aimed at CNAs and aneuploidy, because at the time of diagnosis SNAs are unlikely to play any role in disease progression.

The analysis of the distributions of driver event quantities amongst patients of all cancer types combined shows that most often patients have just two events. On the other hand, there are still plenty of patients with more than 30 driver events. This begs the question: why for some patients two driver events are sufficient to develop a detectable tumour whereas the others are not diagnosed until dozens of events are accumulated? This question resonates with the question asked in the beginning of this Discussion, but cancer types cannot fully explain the observed peaks, as most cancer types have much more than two driver events on average. One possible explanation is that in patients with low number of predicted driver events some actually important drivers have not been discovered due to technical reasons, from sequencing to bioinformatic analysis. Some of them might even be epigenetic (Lyu et al., 2020), a class of drivers we have not analysed in this study. Another possible explanation is that in patients with high numbers of predicted drivers some of those drivers are actually false positives due to imperfections of driver prediction algorithms. Although both of these explanations are likely true to some degree, their contribution does not seem to be so large as to dramatically affect the results. The most probable explanation is the inherent stochasticity of carcinogenesis processes (Belikov, 2017) combined with unequal strength of different driver events (Zhang et al., 2021). For example, patients having tumours with only two driver events were unlucky that these events happened (by chance) in the most crucial genes such as *TP53* or *KRAS*. On the other hand, other patients did not develop a detectable tumour until they have accumulated 30 events because those events involved only weak drivers. It would be interesting to develop a bioinformatic approach to determine the strength of driver genes, e.g. based on how often a given gene is found amongst a few sufficient ones.

Overall, our study brought some clarity to the distributions of driver events of various classes in various demographic and clinical groups of cancer patients and highlighted the importance of oncogene amplifications and aneuploidy events. Many crucial questions, such as the reason behind the dramatic variability in the number of driver events between cancer types, remain to be answered.

References

Arnedo-Pac, C., Mularoni, L., Muiños, F., Gonzalez-Perez, A., & Lopez-Bigas, N. (2019). OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. *Bioinformatics*, *35*(22), 4788–4790. https://doi.org/10.1093/bioinformatics/btz501

Asselin-Labat, M.-L., Vaillant, F., Sheridan, J. M., Pal, B., Wu, D., Simpson, E. R., Yasuda, H., Smyth, G. K., Martin, T. J., Lindeman, G. J., & Visvader, J. E. (2010). Control of mammary stem cell function by steroid hormone signalling. *Nature*, *465*(7299), 798–802. https://doi.org/10.1038/nature09027

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A.,

Colaprico, A., Wendl, M. C., Kim, J., Reardon, B., Ng, P. K.-S., Jeong, K. J., Cao, S., Wang, Z., Gao, J., Gao, Q., Wang, F., Liu, E. M., Mularoni, L., … Ding, L. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*, *173*(2), 371-385.e18. https://doi.org/10.1016/j.cell.2018.02.060

Bashashati, A., Haffari, G., Ding, J., Ha, G., Lui, K., Rosner, J., Huntsman, D. G., Caldas, C., Aparicio, S. A., & Shah, S. P. (2012). DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biology*, *13*(12), R124. https://doi.org/10.1186/gb-2012-13-12-r124

Belikov, A. V. (2017). The number of key carcinogenic events can be predicted from cancer incidence. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-12448-7

Ben-David, U., & Amon, A. (2020). Context is everything: aneuploidy in cancer. *Nature Reviews Genetics*, *21*(1), 44–62. https://doi.org/10.1038/s41576-019-0171-x

Dawson, M. A., & Kouzarides, T. (2012). Cancer Epigenetics: From Mechanism to Therapy. *Cell*, *150*(1), 12–27. https://doi.org/10.1016/j.cell.2012.06.013

Folkerd, E. J., & Dowsett, M. (2010). Influence of Sex Hormones on Cancer Progression. *Journal of Clinical Oncology*, *28*(26), 4038–4044. https://doi.org/10.1200/JCO.2009.27.4290

Joshi, P. A., Jackson, H. W., Beristain, A. G., Di Grappa, M. A., Mote, P. A., Clarke, C. L., Stingl, J., Waterhouse, P. D., & Khokha, R. (2010). Progesterone induces adult mammary stem cell expansion. *Nature*, *465*(7299), 803–807. https://doi.org/10.1038/nature09091

Lyu, J., Li, J. J., Su, J., Peng, F., Chen, Y. E., Ge, X., & Li, W. (2020). DORGE: Discovery of Oncogenes and tumoR suppressor genes using Genetic and Epigenetic features. *Science Advances*, *6*(46), eaba6784. https://doi.org/10.1126/sciadv.aba6784

Martincorena, I., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Van Loo, P., Davies, H., Stratton, M. R., & Campbell, P. J. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*, *171*(5), 1029-1041.e21. https://doi.org/10.1016/j.cell.2017.09.042

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*(10), 555–572. https://doi.org/10.1038/s41568-020-0290-x

Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A., & López-Bigas, N. (2016). OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biology*, *17*(1), 128. https://doi.org/10.1186/s13059-016-0994-0

Nakada, D., Oguro, H., Levi, B. P., Ryan, N., Kitano, A., Saitoh, Y., Takeichi, M., Wendt, G. R., & Morrison, S. J. (2014). Oestrogen increases haematopoietic stem-cell self-renewal in females and during pregnancy. *Nature*, *505*(7484), 555–558. https://doi.org/10.1038/nature12932

Pihan, G., & Doxsey, S. J. (2003). Mutations and aneuploidy: Co-conspirators in cancer? *Cancer Cell*, *4*(2), 89–94. https://doi.org/10.1016/S1535-6108(03)00195-8

Pon, J. R., & Marra, M. A. (2015). Driver and Passenger Mutations in Cancer. *Annual Review of Pathology: Mechanisms of Disease*, *10*(1), 25–50. https://doi.org/10.1146/annurev-pathol-012414-040312

Tokheim, C., Bhattacharya, R., Niknafs, N., Gygax, D. M., Kim, R., Ryan, M., Masica, D. L., & Karchin, R. (2016). Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. *Cancer Research*, *76*(13), 3719 LP – 3731. https://doi.org/10.1158/0008-5472.CAN-15-3190

Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B., & Karchin, R. (2016). Evaluating the evaluation of cancer driver genes. *Proceedings of the National Academy of Sciences*, *113*(50), 14330 LP – 14335. https://doi.org/10.1073/pnas.1616440113

Tokheim, C., & Karchin, R. (2019). CHASMplus Reveals the Scope of Somatic Missense

Mutations Driving Human Cancers. *Cell Systems*, *9*(1), 9-23.e8. https://doi.org/10.1016/j.cels.2019.05.005

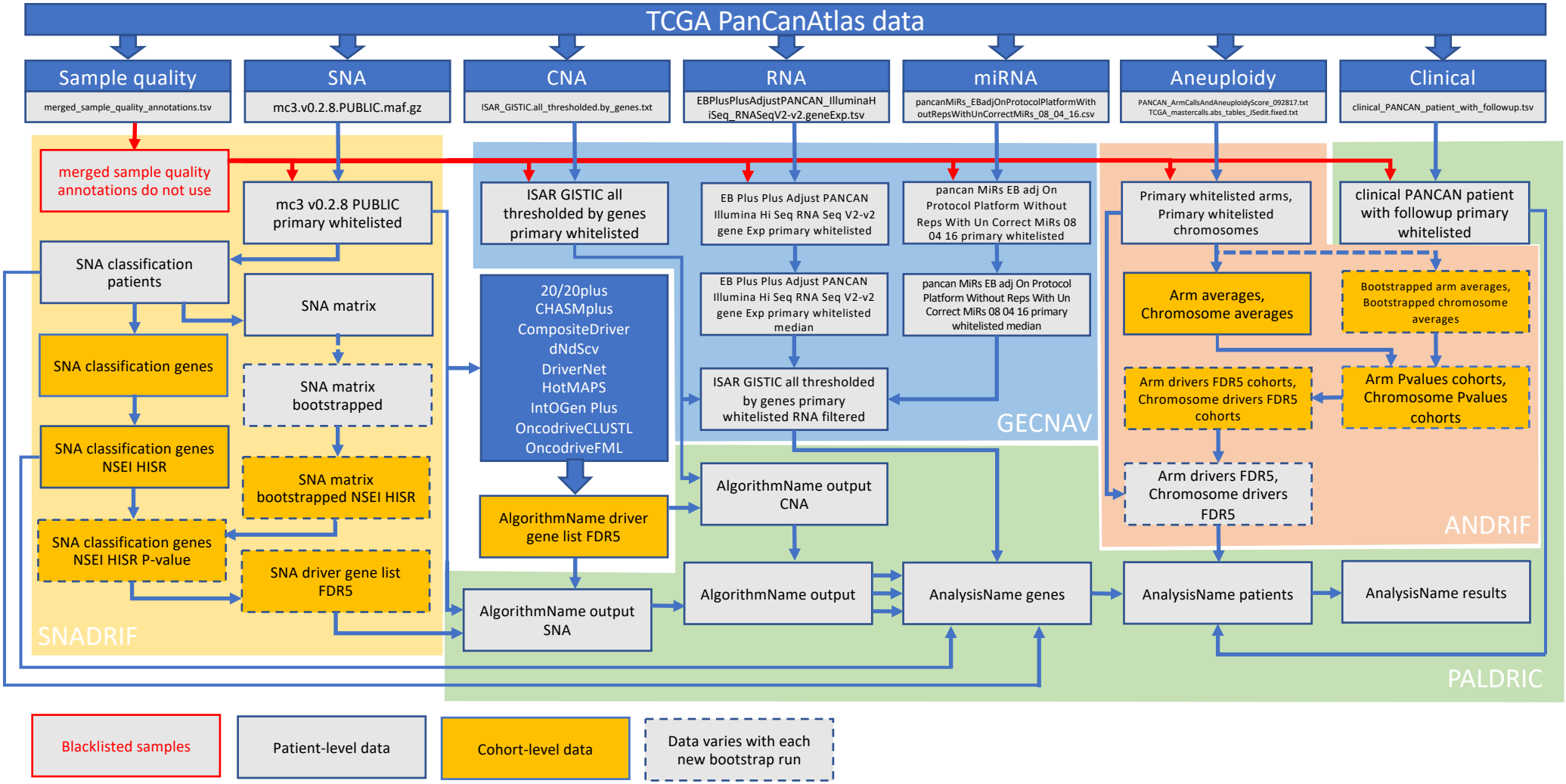
Tomasetti, C., Li, L., & Vogelstein, B. (2017). Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science*, *355*(6331), 1330 LP – 1334. https://doi.org/10.1126/science.aaf9011

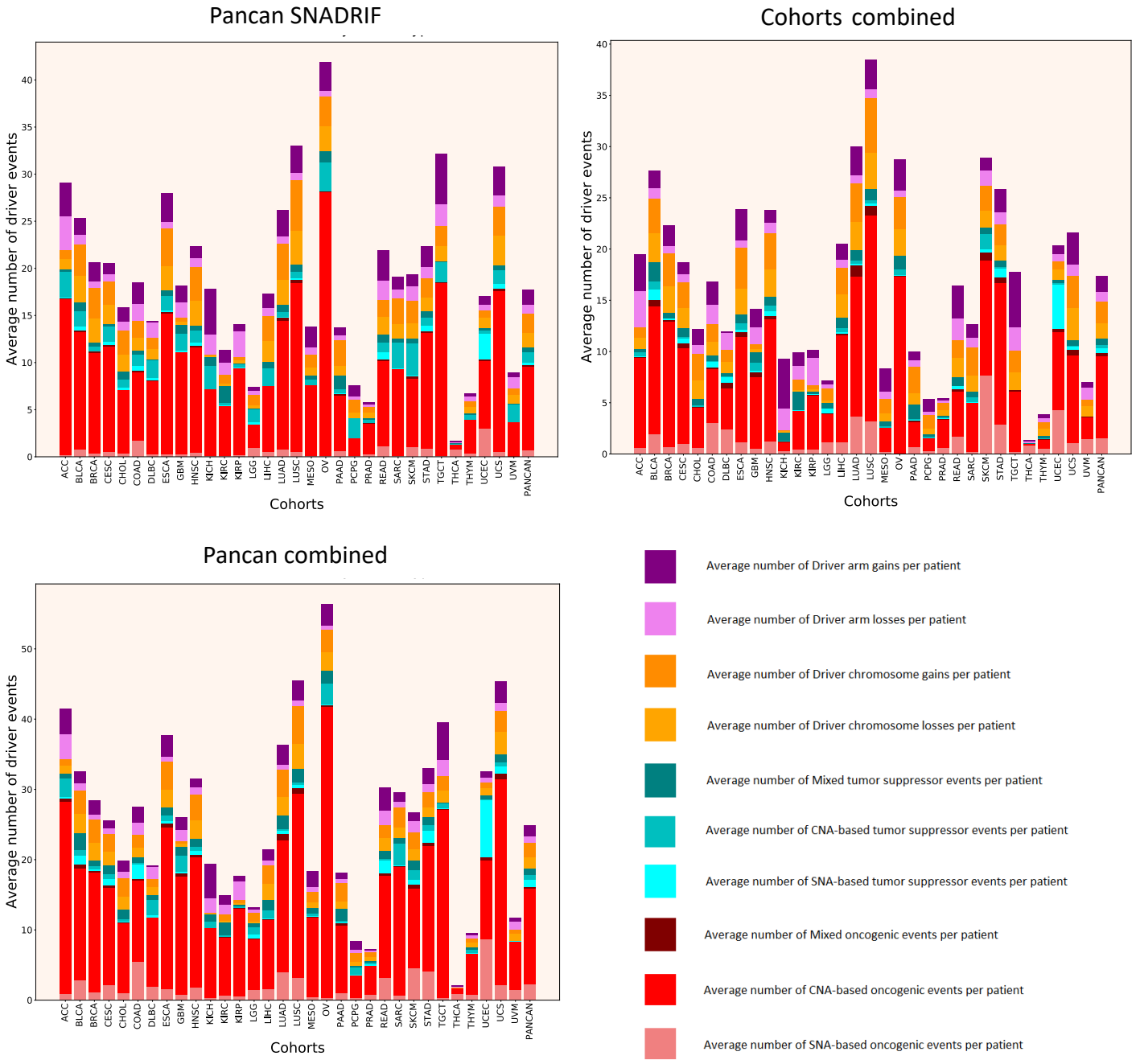
Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., & Kinzler, K. W. (2013). Cancer Genome Landscapes. *Science*, *339*(6127), 1546 LP – 1558. https://doi.org/10.1126/science.1235122

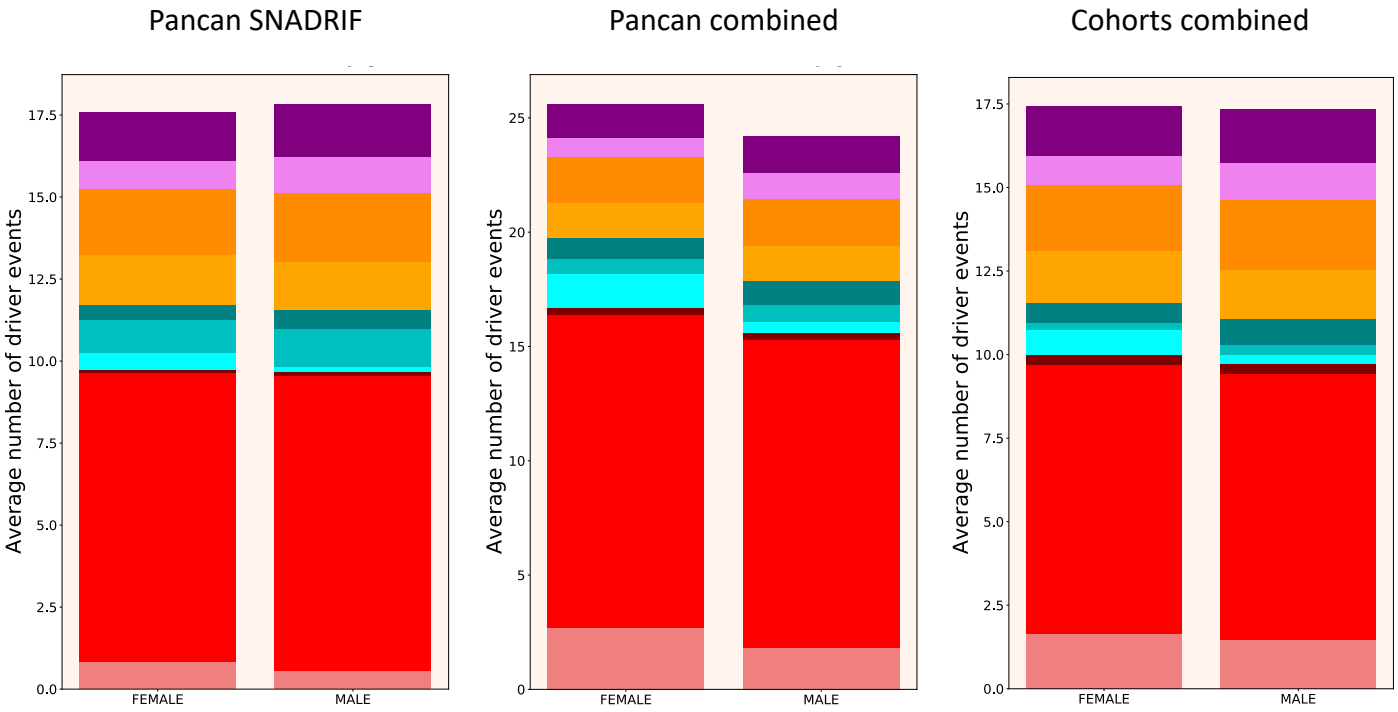
Weinberg, R. A. (1994). Oncogenes and tumor suppressor genes. *CA: A Cancer Journal for Clinicians*, *44*(3), 160–170. https://doi.org/https://doi.org/10.3322/canjclin.44.3.160

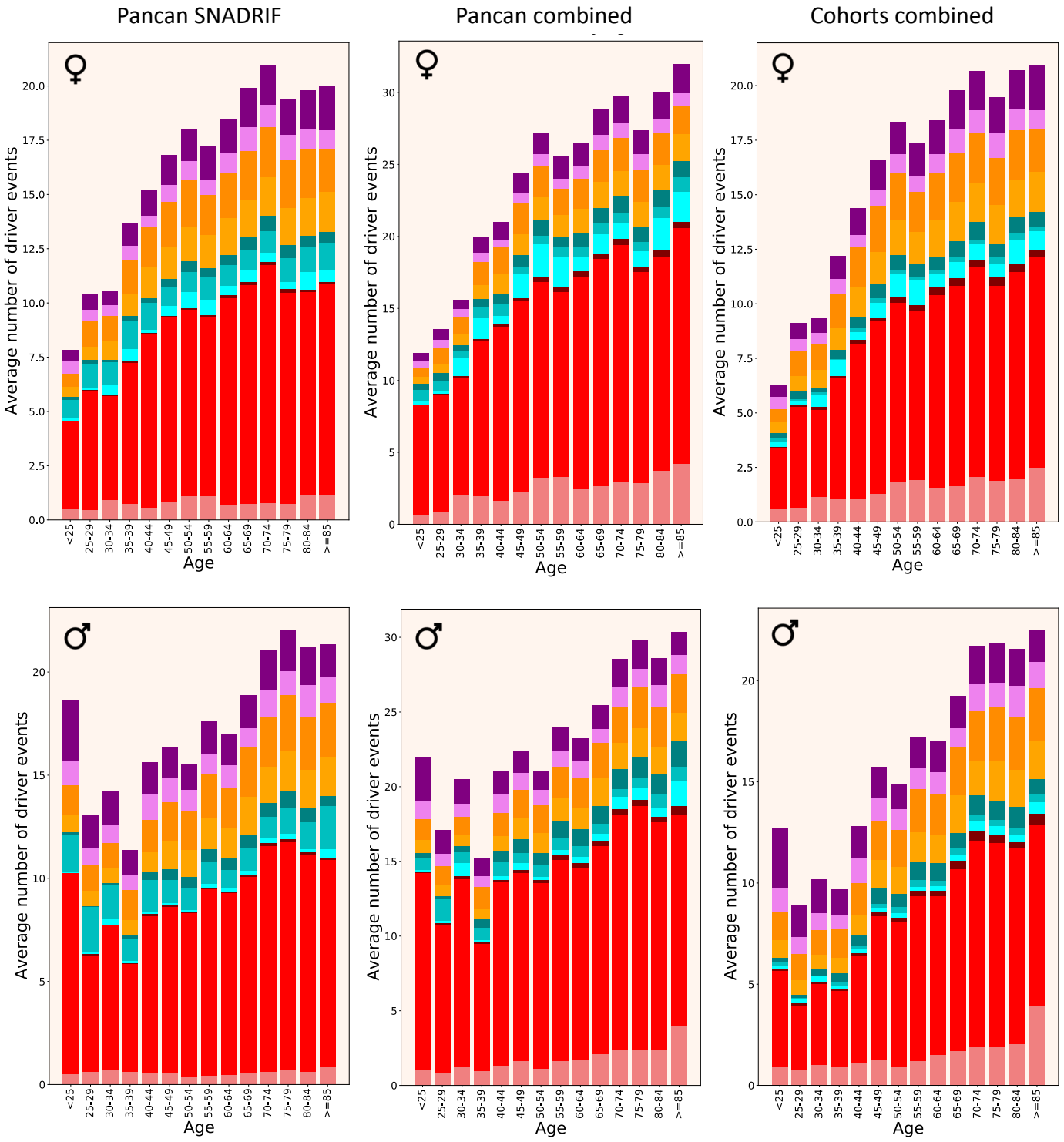
Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, A. D., Saksena, G., Tabak, B., Lawrence, M. S., Zhang, C.-Z., Wala, J., Mermel, C. H., Sougnez, C., Gabriel, S. B., Hernandez, B., Shen, H., Laird, P. W., Getz, G., Meyerson, M., & Beroukhim, R. (2013). Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*, *45*(10), 1134–1140. https://doi.org/10.1038/ng.2760

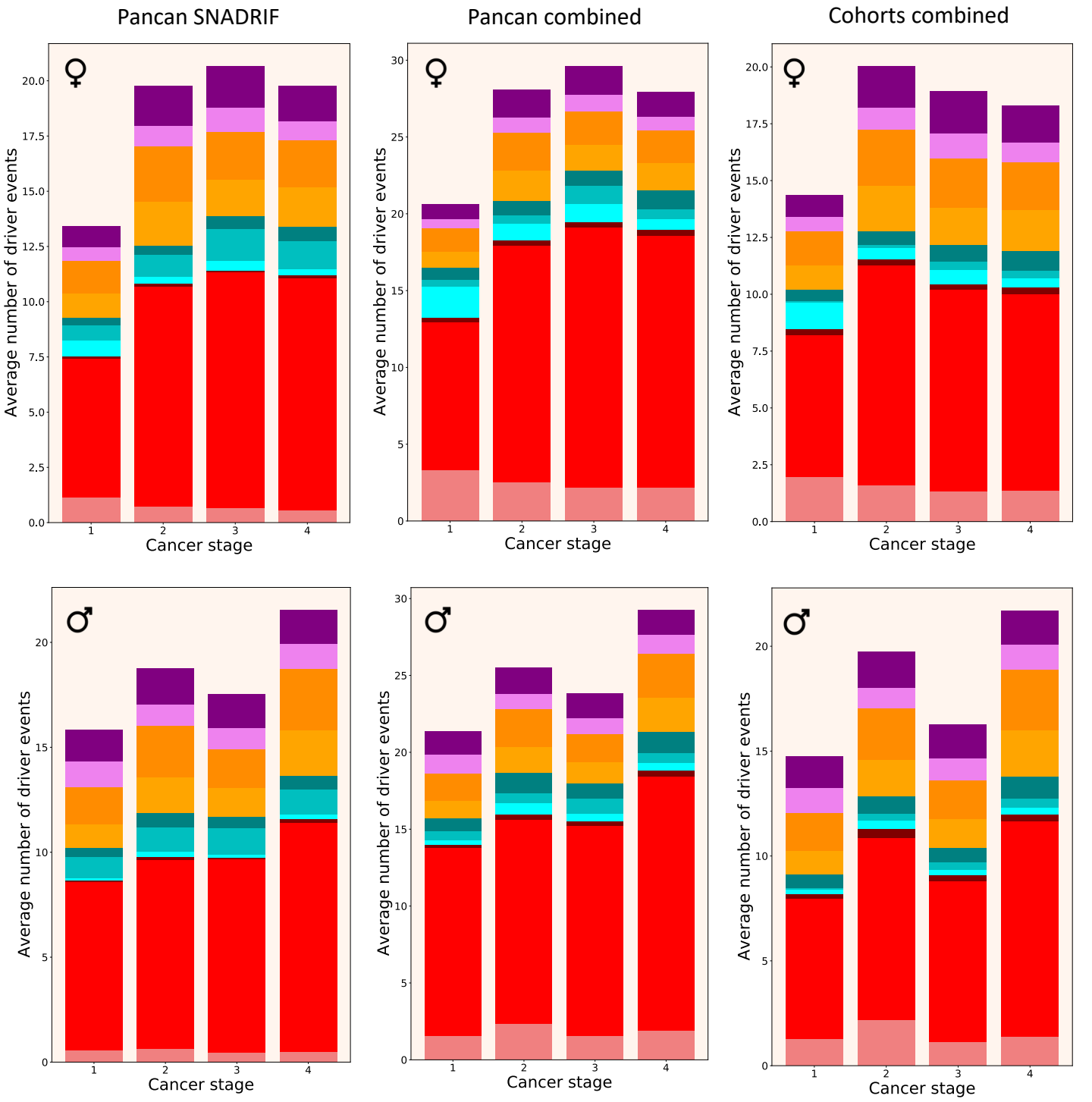
Zhang, M., Jang, H., & Nussinov, R. (2021). PI3K Driver Mutations: A Biophysical Membrane-Centric Perspective. *Cancer Research*, *81*(2), 237 LP – 247. https://doi.org/10.1158/0008-5472.CAN-20-0911

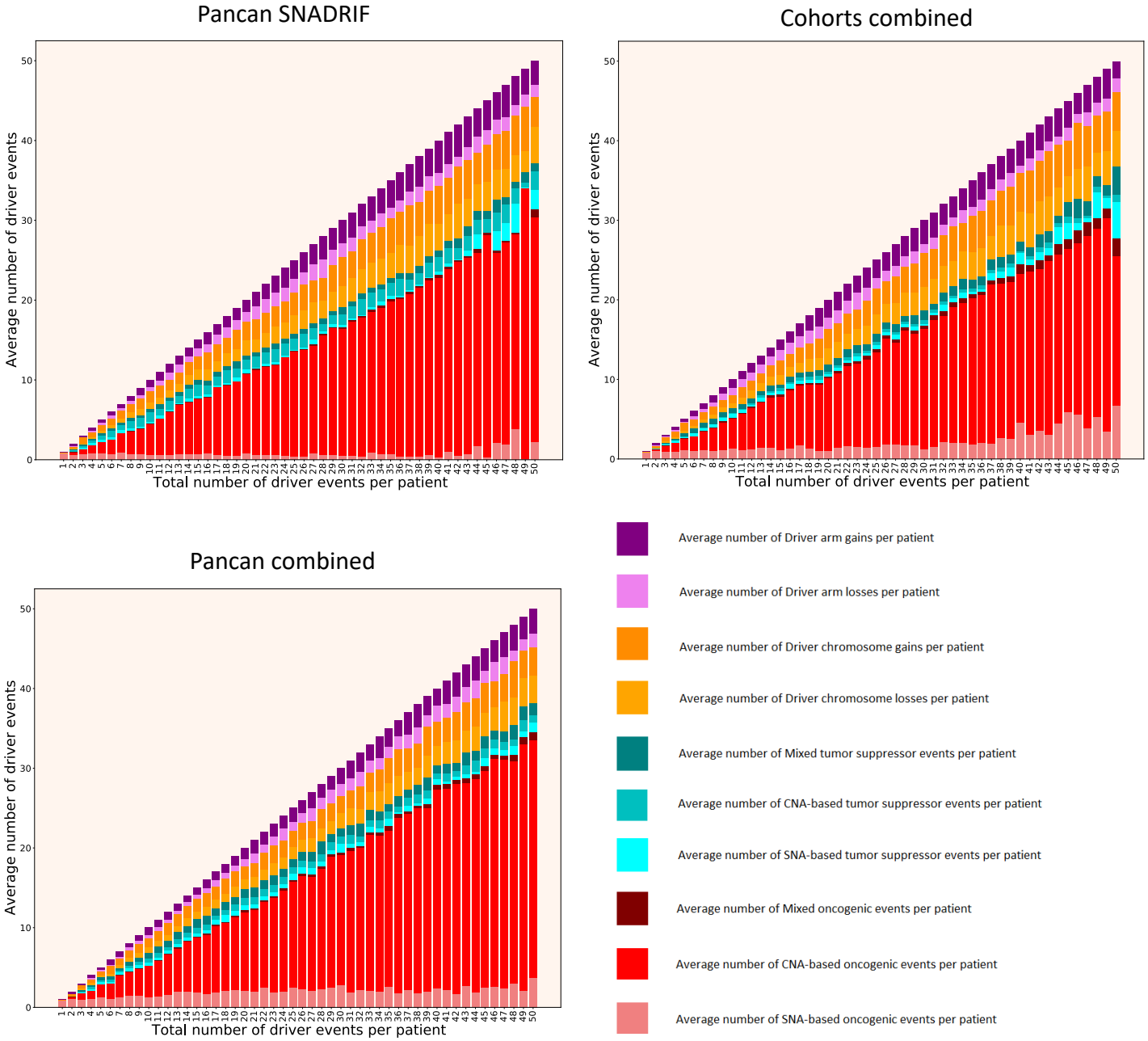

TCGA PanCanAtlas data
Sample quality
merged_sample_quality_annotations.tsv
SNA
mc3.v0.2.8.PUBLIC.maf.gz
CNA
ISAR_GISTIC.all_thresholded.by_genes.txt
RNA
EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2-v2.geneExp.tsv
miRNA
pancanMiRs_EBadjOnProtocolPlatformWithoutRepsWithUnCorrectMiRs_08_04_16.csv
Aneuploidy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


Pancan SNADRIF
Cohorts combined
Pancan combined
Average number of driver events
Cohorts
ACC
BLCA
BRCA
CESC
CHOL
COAD
DLBC
ESCA
GBM
HNSC
KICH
KIRC
KIRP
LGG
LIHC
LUAD
LUSC
MESO
OV
PAAD
PCPG
PRAD
READ
SARC
SKCM
STAD
TGCT
THCA
THYM
UCEC
UCS
UVM
PANCAN
Average number of Driver arm gains per patient
Average number of Driver arm losses per patient
Average number of Driver chromosome gains per patient
Average number of Driver chromosome losses per patient
Average number of Mixed tumor suppressor events per patient
Average number of CNA-based tumor suppressor events per patient
Average number of SNA-based tumor suppressor events per patient
Average number of Mixed oncogenic events per patient
Average number of CNA-based oncogenic events per patient
Average number of SNA-based oncogenic events per patient

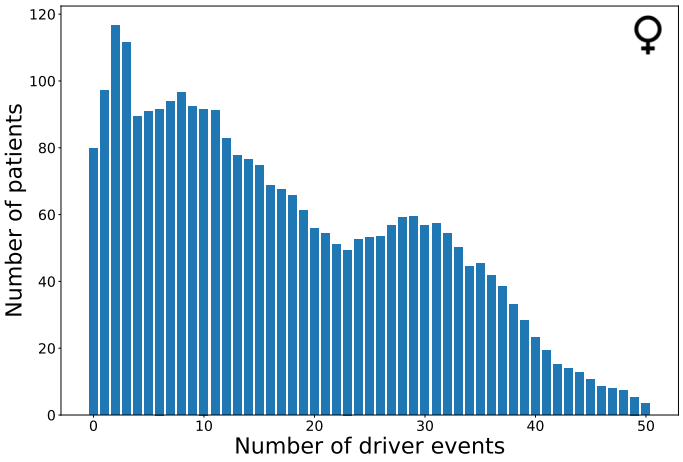

Pancan SNADRIF
Pancan combined
Cohorts combined
Average number of driver events
FEMALE
MALE

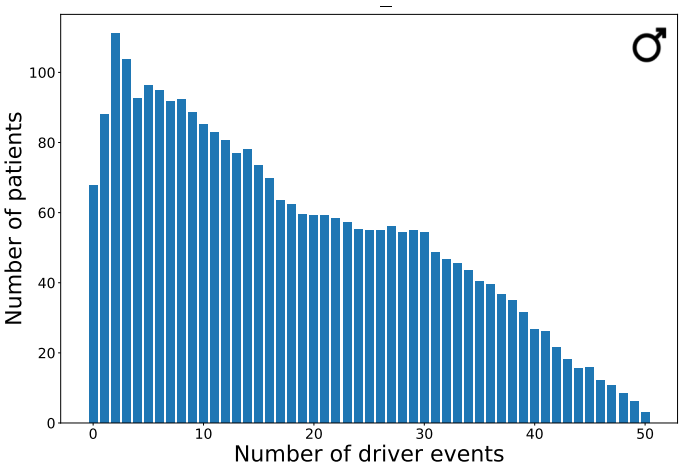

Pancan SNADRIF
Pancan combined
Cohorts combined
♀
♂
Average number of driver events
Age
<25
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80-84
>=85


Pancan SNADRIF
Pancan combined
Cohorts combined
♀
♂
Average number of driver events
Cancer stage
1
2
3
4

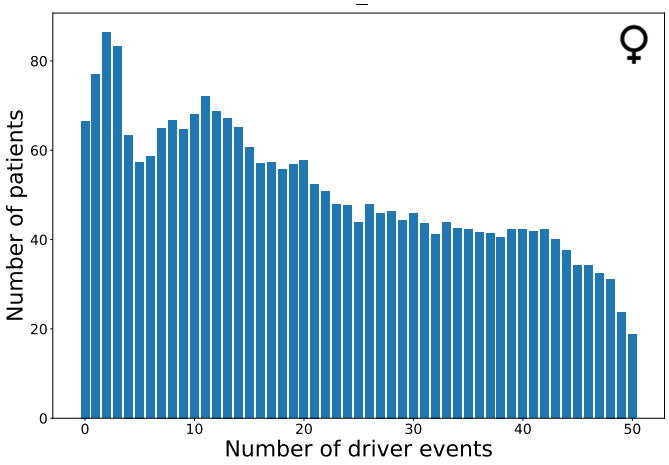

Pancan SNADRIF
Cohorts combined
Pancan combined
Average number of driver events
Total number of driver events per patient
Average number of Driver arm gains per patient
Average number of Driver arm losses per patient
Average number of Driver chromosome gains per patient
Average number of Driver chromosome losses per patient
Average number of Mixed tumor suppressor events per patient
Average number of CNA-based tumor suppressor events per patient
Average number of SNA-based tumor suppressor events per patient
Average number of Mixed oncogenic events per patient
Average number of CNA-based oncogenic events per patient
Average number of SNA-based oncogenic events per patient

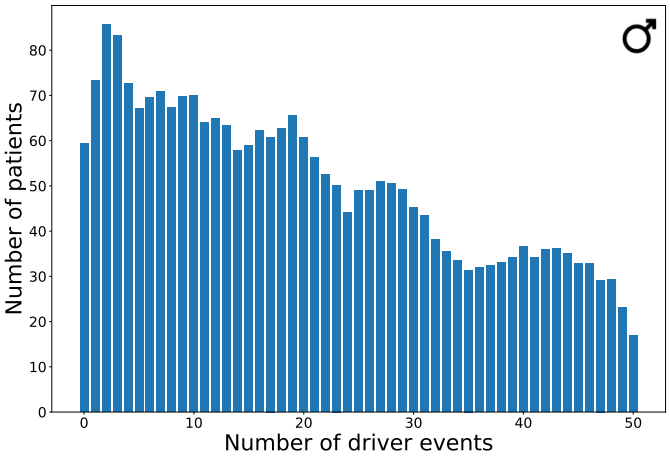
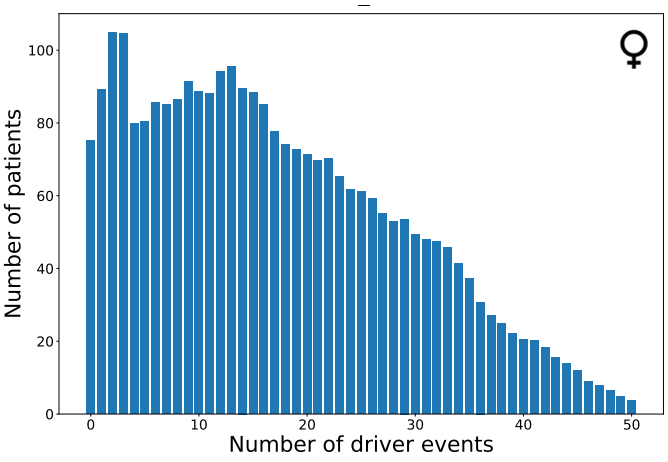
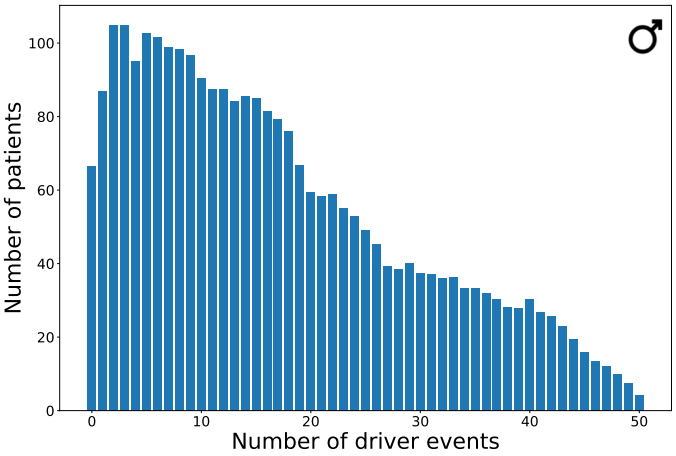
Pancan SNADRIF





Pancan combined





Cohorts combined