

OPEN

Interrogating Mutant Allele Expression via Customized Reference Genomes to Define Influential Cancer Mutations

Adam D. Grant¹, Paris Vail¹, Megha Padi², Agnieszka K. Witkiewicz³ & Erik S. Knudsen⁴

Genetic alterations are essential for cancer initiation and progression. However, differentiating mutations that drive the tumor phenotype from mutations that do not affect tumor fitness remains a fundamental challenge in cancer biology. To better understand the impact of a given mutation within cancer, RNA-sequencing data was used to categorize mutations based on their allelic expression. For this purpose, we developed the MAXX (Mutation Allelic Expression Extractor) software, which is highly effective at delineating the allelic expression of both single nucleotide variants and small insertions and deletions. Results from MAXX demonstrated that mutations can be separated into three groups based on their expression of the mutant allele, lack of expression from both alleles, or expression of only the wild-type allele. By taking into consideration the allelic expression patterns of genes that are mutated in PDAC, it was possible to increase the sensitivity of widely used driver mutation detection methods, as well as identify subtypes that have prognostic significance and are associated with sensitivity to select classes of therapeutic agents in cell culture. Thus, differentiating mutations based on their mutant allele expression via MAXX represents a means to parse somatic variants in tumor genomes, helping to elucidate a gene's respective role in cancer.

Cancer is a complex disease, initiated by DNA alterations within genes that control multiple hallmarks of tumorigenesis, such as deregulated cell growth and genomic instability¹⁻³. Once the genomic architecture of a cancer cell is established, the cancer will continue to evolve to overcome additional regulatory mechanisms and eventually acquire the ability to progress to metastatic disease^{4,5}. A primary goal in cancer therapeutics is to target selective pathways that are critical to the tumor's growth and sustainability^{6,7}. However, during tumorigenesis, not only do mutations responsible for the cancer phenotype arise, but also random mutations that have no effect on the fitness of the tumor. The difficulty in distinguishing driver and passenger mutations represents a core challenge in distilling meaningful functional information from tumor sequencing data^{8,9}.

Many efforts are ongoing to determine which mutations contribute to the initiation and progression of cancer. To consolidate information regarding a mutation's impact on the progression of cancer, well curated databases such as COSMIC¹⁰ have been established. However, these databases are incomplete and depend heavily on a combination of DNA sequencing and computational software to identify a mutation's potential contribution towards the development of a tumor¹¹. Most driver mutation detection software is centered around one or a combination of three approaches: identifying genes that have an increased mutation rate among a subset of tumor samples¹², evaluating the functional impact of the mutation¹³⁻¹⁵, and using network analysis to identify gene interactions that have increased mutation rates^{16,17}. These computational approaches have contributed a great deal to our current understanding of how and which mutations are involved in the progression of cancer. However, there is a lack of concordance between positive results among the driver mutation identification approaches¹⁸. Also, a recent evaluation of commonly used driver mutation software demonstrated that in most cases, these approaches have high false positive rates¹⁹.

¹University of Arizona Cancer Center, Tucson, AZ, 85719, USA. ²Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ, 85719, USA. ³Center for Precision Medicine, Roswell Park Cancer Center, Buffalo, NY, 14263, USA. ⁴Department of Molecular and Cellular Biology, Roswell Park Cancer Center, Buffalo, NY, 14263, USA. Correspondence and requests for materials should be addressed to A.K.W. (email: agnieszka.witkiewicz@roswellpark.org) or E.S.K. (email: erik.knudsen@roswellpark.org)

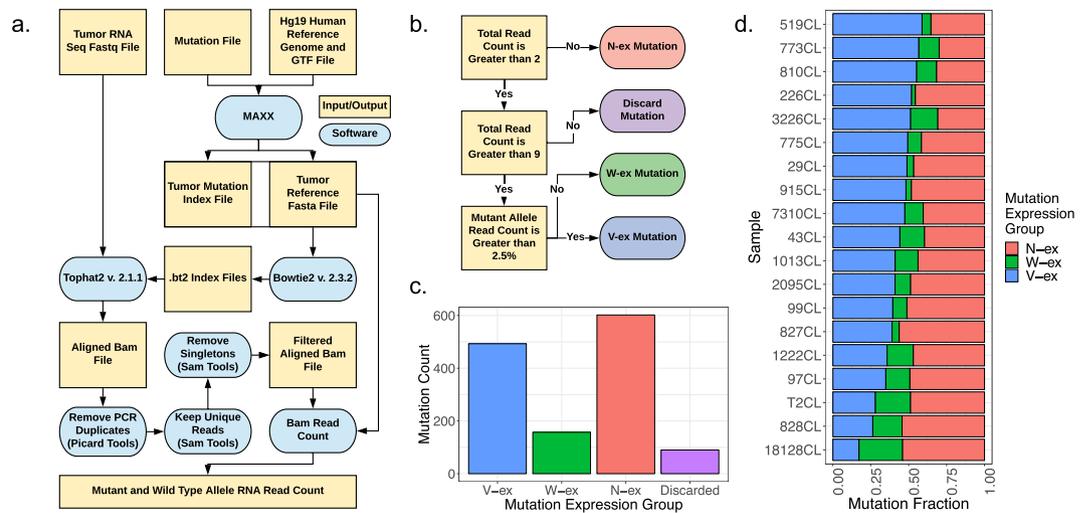


Figure 1. Developing an unbiased tool to assess the expression of somatic variants. **(a)** Flowchart of the MAXX pipeline, which ultimately identifies the RNA read count for the mutant allele and the wild type allele. **(b)** Methodology for mutation expression group placement, represented as V-ex (blue), W-ex (green), and N-ex (red); V-ex, mutations that express the mutant allele; W-ex, mutations that only express the wild type allele; N-ex, mutations that don't express the wild type allele or the mutant allele. **(c)** The number of mutations associated with each mutation expression group. These mutations were derived from 19 PDAC patient derived cell lines. **(d)** The fraction of each mutation expression group for all patient derived cell lines.

One major limitation of established driver identification software is their principle focus on data derived from DNA sequencing and inferred presence of the protein. These approaches provide insight on the distribution of the mutation in cancer and potential function but do not incorporate crucial information on the transcription of the mutated gene. Assumptions on the presence of mutated transcripts can cause inaccurate conclusions as to the mutation's impact within the tumor. Recent studies have demonstrated that integration of exome and transcriptome genomic profiles can reinforce the interpretation of mutation events^{20,21}. Thus, to increase the understanding of somatic mutations in cancer, we combined exome and RNA-sequencing data to segregate tumor mutations based on their allelic expression.

To enhance mutation allelic expression detection, the software MAXX (Mutation Allelic Expression Extractor) was developed. MAXX has the capability to identify the allelic expression for small insertion and deletion (indel) mutations and single nucleotide variants (SNV), while maintaining high alignment precision. Accurate alignment of RNA-seq reads sets up the groundwork to detect correct RNA allele frequencies for genetic variants^{22,23}. Evaluation of the allelic expression for thousands of mutations derived from pancreatic adenocarcinoma (PDAC) established that mutations can be separated into three expression groups based on the presence of RNA-sequencing reads from the mutant allele, wild type allele, or both alleles. These three mutation expression groups were established to have unique biological features that revealed expected characteristics of driver mutations, passenger mutations, and an understudied group of mutations that appeared to be selectively silenced within the tumor context. Mutation expression groups also assisted in the identification of PDAC subtypes that have prognostic relevance.

Results

Developing an unbiased tool to assess the allelic expression of somatic variants. While there are multiple approaches to delineating if a particular mutation is expressed, much of this work has involved gene-specific analysis instead of a holistic evaluation of the cancer genome. To attain a global assessment of mutation expression within tumors, we quantified the allelic expression of thousands of mutations (missense, nonsense, and small insertion/deletion) derived from PDAC patient-derived cell lines, patient-derived xenografts (PDX), and primary tumors obtained from The Cancer Genome Atlas (TCGA)^{24,25}. Because mutations can exist as either somatic single nucleotide variant (SNV) or small insertion/deletion (indel), a unique approach was developed to confidently identify the allelic expression of all mutations. Rather than depend on tumor RNA-sequencing reads to be correctly aligned onto a standardized reference genome (e.g. Hg19), the MAXX (MAXX: Mutation Allelic Expression Extractor) method creates a new reference genome that is tumor selective and specifically extracts mutant allele expression. The MAXX approach utilizes mutation calls derived from DNA sequencing to generate a tumor-specific reference genome (Fig. 1a). This precise reference genome was then used in conjunction with Tophat²⁶ to enhance the alignment of the tumor's RNA-sequencing reads to the mutant allele. The approach of MAXX enables the calculation of precise RNA mutant allele frequencies for essentially any variant type that is defined by the DNA sequencing approach, allowing a comprehensive and unbiased analysis of mutation allelic expression.

Different patterns of variant allelic expression in tumor models. To initially interrogate the efficacy of the MAXX pipeline, we generated mutation expression profiles, a file containing the DNA and RNA allele frequencies for all identified mutations, for 19 PDAC patient derived cell lines. These cell lines underwent both

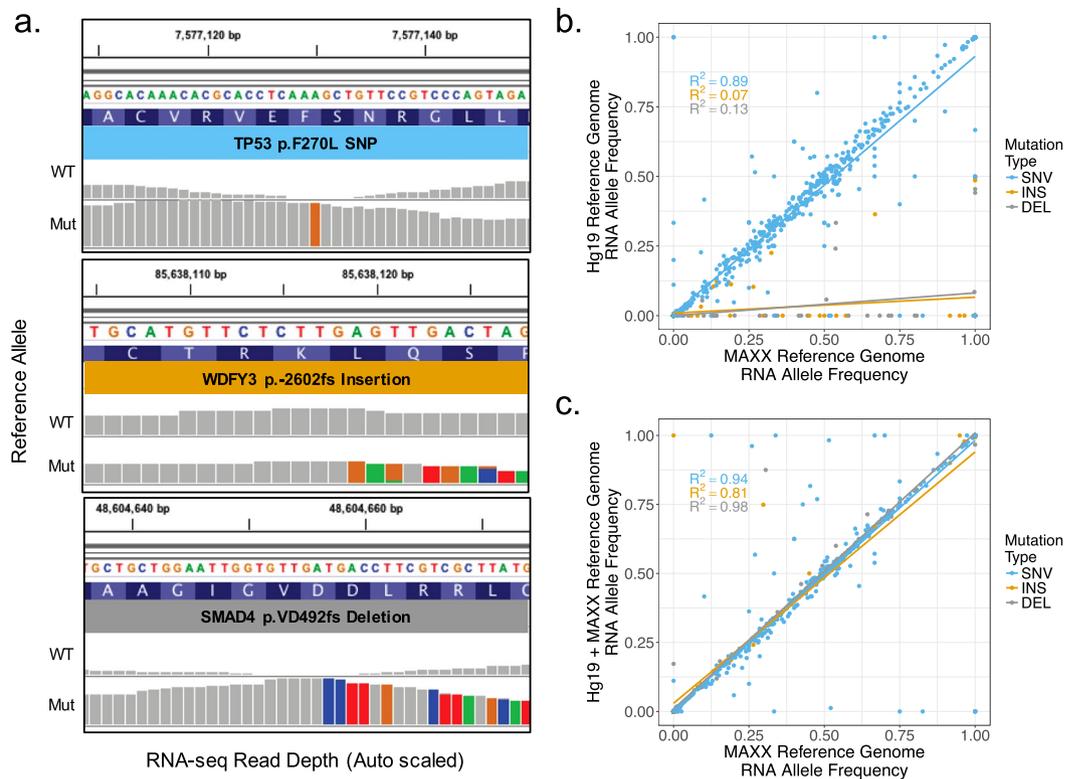


Figure 2. MAXX pipeline accurately maps indels and is computationally efficient. **(a)** Integrative genomic viewer visualization of raw RNA-seq reads that aligned to the wild type allele or the mutant allele for a SNV, insertion, and deletion mutation. Non-gray colors represent an alternative nucleotide compared to the Hg19 reference genome **(b)** Comparison of the RNA mutant allele frequencies calculated by using either the MAXX generated reference genome or the Hg19 reference genome. **(c)** The contrast between the RNA mutant allele frequencies identified by using either the MAXX generated reference genome or the Hg19 reference genome with an appended mutant genome created by MAXX.

RNA-sequencing and exome sequencing relative to the normal tissue. Cell line models were ideal for the primary evaluation of MAXX because of the absence of stromal contamination, substantially decreasing the confounding feature of tumor purity^{27,28}. Assessment of the nineteen mutation expression profiles identified that tumor mutations can be classified based on the number of reads that align to the wild type allele, mutant allele, or both alleles (Fig. 1b). In order to prevent misclassification between mutations that do and do not express the mutant allele, mutations that had a total RNA read count between three and nine were discarded from the study. All non-discarded mutations were placed into one of the following three mutation expression groups: 1. Mutations that express the variant allele, labeled as the variant expressed group (V-ex, Blue) 2. Mutations wherein only the wild type allele is expressed, labeled as the wild type expressed group (W-ex, Green). 3. Mutations that do not express the wild type or the mutant allele, labeled as the not expressed group (N-ex, Red). These three mutation expression groups have also been identified in patient derived cell lines from multiple myeloma patients²⁹. The distribution of the three mutation expression groups was summarized across all patient derived cell lines (Fig. 1c). The N-ex group made up half of all the variants, while only a small subset of mutations fell into the W-ex group. This overall distribution was observed in all individual models, suggesting that the expression distribution is not specific to a given sequencing run or tumor (Fig. 1d).

MAXX pipeline accurately maps indels and is computationally efficient. Previous studies have shown that RNA-sequencing aligners are capable of aligning SNV, but have poor alignment precision for reads containing an indel mutation^{30,31}. However, the unique workflow of MAXX allows the allelic expression to be confidently calculated for both SNV and indel mutations (Fig. 2a). To enhance indel alignment for RNA-sequencing data, which is a critical aspect in identifying accurate RNA allele frequencies²⁰, MAXX generates a tumor-specific reference genome based on mutation calls derived from DNA sequencing. When comparing RNA allele frequencies derived from bam files that were aligned with Tophat2 using either a MAXX or Hg19 reference genome and underwent important filtering steps for allelic expression analysis^{21,32}, it was identified that the MAXX reference genome significantly enhanced the allelic expression of indel mutations, compared to the Hg19 reference genome (Fig. 2b). There was relatively little variation between SNV RNA allele frequencies using the MAXX or Hg19 reference genomes.

Generating a reference genome for each tumor sample would require a significant amount of computational processing and storage for large studies. However, MAXX generated reference genomes are significantly smaller than the Hg19 reference genome. This decrease in size is due to the tumor reference genome containing only

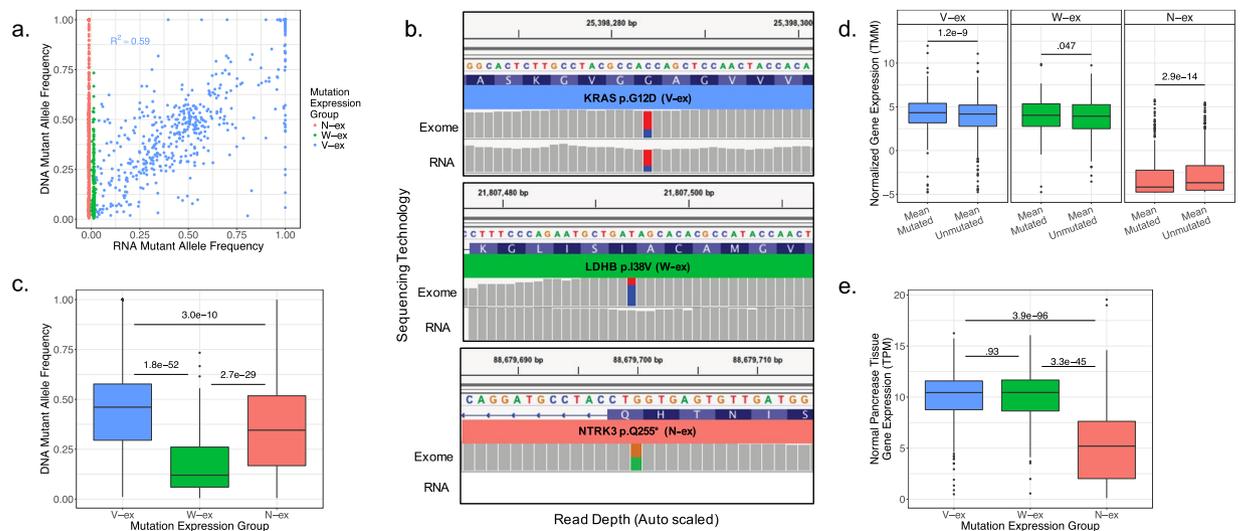


Figure 3. Mutant allele expression is associated with DNA mutant allele frequency, but not RNA expression level. **(a)** All mutations identified within patient derived cell lines DNA allele frequency plotted against their corresponding RNA allele frequency. **(b)** Integrative genomic viewer representation of the exome sequencing and RNA sequencing for a V-ex mutation, W-ex mutation, and N-ex mutation. Non-gray colors represent the presence of conflicting nucleotides aligned to the Hg19 reference genome. **(c)** The distribution of DNA mutant allele frequencies for the three mutation expression groups and statistical significance based on a two-sample t-test with a two-tail p-value. **(d)** Two sample paired t-test with a two-tail p-value between the average gene expression levels of samples with the mutated gene and samples without the mutated gene for each mutation expression group. **(e)** A box plot of the mean expression across 193 normal pancreas tissues of identified mutated genes within the patient derived cell lines. A two-sample t-test with a two-tailed p-value was used to compare mutation expression groups.

the wild type and mutant sequence for each gene harboring a mutation. Using a reference genome containing only the mutated genes significantly decreases the storage space and computational resources used to index the reference genome and run Tophat2 compared to the Hg19 reference genome (Supplementary Fig. 1). To determine if there is over selection of aligned reads due to the highly truncated genome, we compared RNA mutant allele frequencies established from an Hg19 reference genome with the appended MAXX mutant genome to a MAXX generated reference genome (Fig. 2c). These two reference genomes delivered virtually identical RNA mutant allele frequencies and placement of mutations into expression groups (Supplementary Fig. 2). Overall, this demonstrates that a simplified reference genome provides a robust substrate for differential alignment of transcript reads and can confidently be used to generate mutation expression profiles.

Mutant allele expression is associated with DNA mutant allele frequency, but not RNA expression level.

To resolve if the RNA mutant allele frequency is associated with the DNA mutant allele frequency in PDAC, we evaluated the dual relationship across the three mutation expression groups (Fig. 3a). In general, the variant allele expression of V-ex mutations correlated with their DNA allele frequency ($R^2 = 0.59$). This correlation was similarly identified in another study using CT26, B16F10, and 4T1 mouse cell lines³³. However, in our dataset it was observed that multiple genes with a V-ex mutation had a 100% RNA mutant allele frequency, despite having a DNA mutant allele frequency of ~50%. This discrepancy between DNA and RNA mutant allele frequencies suggests that the non-mutated allele of these genes is selectively silenced, which is a common phenomenon for tumor suppressors³⁴. N-ex and W-ex mutation expression groups have no expression of the mutant allele, and as expected they had an RNA mutant allele frequency of zero (Fig. 3b). To determine if a mutation expression group was biased towards a specific mutation type (missense, deletion, insertion, nonsense), as might be expected from processes such as nonsense mediated decay, the proportion of mutation types was measured between mutation expression groups (Supplementary Fig. 3). Statistical testing illustrated that there is no correlation between the mutation type and mutation expression groups. Interestingly, the variant allele frequency of mutation expression groups established that mutation expression provides insight on mutation selectivity (Fig. 3c). The V-ex group has a significantly higher mean variant allele frequency than the other two groups, suggesting that V-ex mutations have the greatest impact on tumor proliferation compared to W-ex and N-ex mutations. As for the N-ex group, most mutations had a variant allele frequency between 30% and 40%, suggesting that many of these mutations, whilst not expressed, are in fact largely clonal in the tumor. In contrast, the mean variant allele frequency of the W-ex group is considerably lower than the other two mutation expression groups. To interrogate if there is some selective feature of mutations that leads to transcript loss, we compared the mean gene expression of samples that did not have a mutated allele to the mean gene expression of samples that did contain a mutated allele (Fig. 3d). The overall expression of genes that contained a N-ex mutation is substantially less than genes with a V-ex or W-ex mutation, as expected. This confirms that genes with an N-ex mutation are rarely

expressed, and thus mutations within this class likely have little influence on PDAC proliferation. On the contrary, genes with a V-ex or W-ex mutation contained a relatively high gene expression within the tumor, suggesting that genes with a V-ex or W-ex mutation are important to the progression of the tumor. While significant differences were observed between mutated and unmutated genes among each expression group, the magnitude change was marginal. Interestingly, we identified a similar expression pattern of genes within each mutation expression group from 193 normal pancreas tissues obtained from the Genotype Tissue Expression (GTEx) database (Fig. 3e), further suggesting that genes with an N-ex mutation have little relevance to pancreas tissues compared to genes with a V-ex and W-ex mutation. However, in regard to tissue specificity, no mutation expression group was observed to contain a high fraction of pancreas tissue specific genes (Supplementary Fig. 4). To determine if an increased gene expression, yet an absence of mutant allele expression of W-ex mutations was due to RNA splicing events, the exon expression of mutated and non-mutated exons was quantified (Supplementary Fig. 5). Equivalent to the transcript analysis, there was little variance between the mutated and non-mutated exon expression. These findings imply that there is not a strong selection against transcript or exon expression among any of the mutation expression groups.

PDAC associated genes mainly fall into the V-ex and W-ex groups. To more fully understand the significance of mutation expression profiles, we evaluated the mutated genes within the three expression groups. In the context of PDAC tumors, it is well known that there are four genes that are frequently mutated and considered disease drivers: KRAS, TP53, SMAD4, and CDKN2A³⁵. These four genes were repeatedly observed to contain V-ex mutations, skewing the mutation frequency distribution of the V-ex group as shown in the violin plots (Fig. 4a). In comparison, the frequency distribution of mutated genes in W-ex and N-ex groups is heavily centered around one, implying little recurrence of such genes within PDAC. Oncogenes such as KRAS generally have a DNA mutant allele frequency of ~50%, because only one allele is required to be mutated for the gene to act as an oncogene³⁶. However, in select cases there is specific selection for the expression of only the KRAS mutant allele. Interestingly, the ~100% expression of the KRAS mutant allele within these samples is largely due to an increased DNA mutant allele frequency, as opposed to allele specific gene expression (Fig. 4b). In contrast, tumor suppressors TP53, SMAD4, and CDKN2A have mutation allele frequencies that approach 100% and mainly express only the mutated allele. In the case of SMAD4 and CDKN2A, this would be expected. However, as TP53 mutations can have gain of function mutations, this data suggests there is exceedingly strong pressure to lose the wild-type allele during tumor development within the pancreas.

In order to delineate the putative significance to cancer, we concentrated on the fraction of mutations in each mutation expression group that were present in either the COSMIC.v80¹⁰ or Tamborero *et al.*³⁷ cancer associated gene datasets (Fig. 4c). As expected, due to the frequency of KRAS, TP53, SMAD4, and CDKN2A, the V-ex group contained the most statistically significant proportion of mutated genes which are associated with cancer. Comparing the W-ex group to the N-ex group demonstrated that the W-ex group had a significantly higher proportion of mutated genes that are associated with cancer. To determine if there is a relationship between mutated cancer associated genes and mutation expression groups, network analysis was performed (Fig. 4d). Evaluation of the generated network indicates that cancer associated genes with a W-ex mutation are well integrated with cancer associated genes containing a V-ex mutation. As for genes containing an N-ex mutation, they were commonly found on the outsides of the network, having little contribution to the structure of the network. These results support the hypothesis that W-ex mutations occur in genes that are important to the tumor, despite their lack of clonal selection and mutant allele expression. In contrast, N-ex mutations appear to be in genes that have little to no impact on tumor progression.

Most, if not all, current driver mutation detection methods do not integrate allelic expression information when predicting influential cancer mutations. Thus, to determine if incorporation of mutation allelic expression can increase cancer gene specificity, we started by employing three commonly used driver mutation detection methods. The three methods selected were 2020+¹⁹, Muffin¹⁷, and OncodriveFM¹⁵. These methods were chosen because of their capability to handle indel mutations and relatively small sample sizes. When comparing the mutation expression group distribution of the top 50 ranked mutated genes identified by 2020+, Muffin and OncodriveFM, there was a consistent tradeoff between V-ex and N-ex groups (Fig. 4e). The driver mutation detection method 2020+ had the highest proportion of V-ex mutated genes, while OncodriveFM had the lowest proportion of V-ex mutated genes. To test the ability of each method to identify cancer associated genes from the Cosmic and Tamborero datasets, a Fisher's Exact test was performed with the assumption that only the top 50 mutated genes predicted by each method were cancer associated (Fig. 4f). Overall, Muffin outperformed 2020+ and OncodriveFM. When we removed the N-ex mutations from the top 50 putative driver mutations, all three methods had an increase of statistical significance in predicting cancer associated mutated genes. This suggests that incorporation of mutations' allelic expression can assist in reducing false positive discovery rates of driver mutation detection methods and improve downstream analyses.

Conservation of mutational expression features from cell lines to PDX tumors. To determine the efficacy of the MAXX pipeline in producing mutation expression profiles for tumors, the MAXX pipeline was evaluated on PDX models. Eight of the nineteen patient derived cell lines have matched PDX models²⁴; therefore, these eight PDX models shared the same tumor specific reference genome as their corresponding cell line. However, these models had their own transcriptome sequenced. Similar to the cell line data, the three mutation expression groups were determined based on wild type and mutant allele read counts, and the RNA mutant allele frequency was plotted against the DNA mutant allele frequency (Fig. 5a). The PDX models demonstrated that mutation expression groups are present in both *in-vitro* and *in-vivo*. Equivalent to the cell lines, RNA mutant allele frequency is strongly correlated with DNA mutant allele frequency. Amongst the V-ex mutations, there was exceedingly high commonality in expression features between PDX and cell lines (Fig. 5b). This conservation

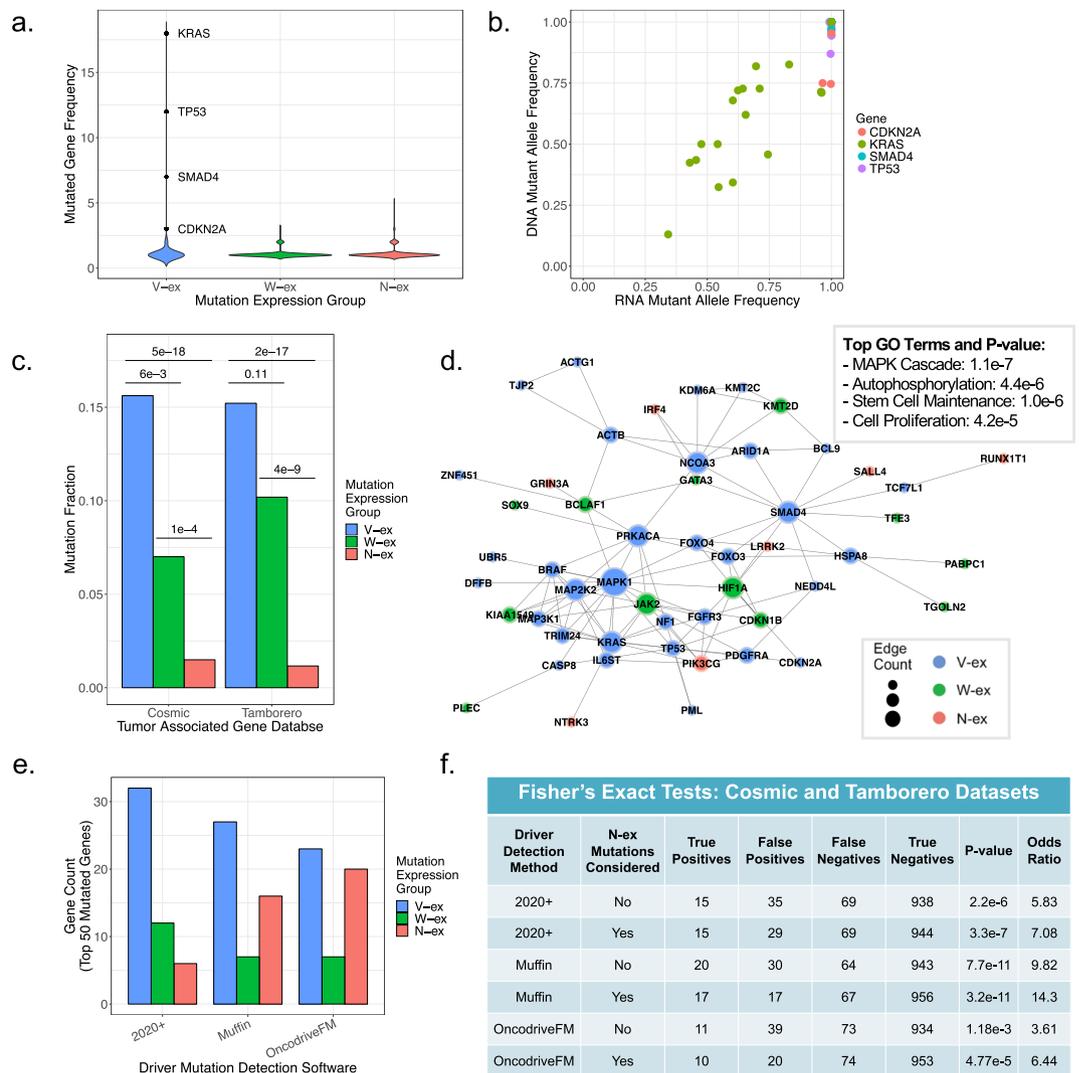


Figure 4. PDAC associated genes mainly fall into the V-ex and W-ex groups. **(a)** The gene mutation frequency of all V-ex, W-ex and N-ex mutations. **(b)** Comparison of the DNA allele frequency and RNA allele frequency for the most well-known PDAC mutations (KRAS, CDKN2A, SMAD4, TP53). **(c)** Fraction of mutations that are associated with the Cosmic and Tamborero cancer gene datasets for each mutation expression group. Statistical significance was performed using a two-proportion z-test between each of the mutation expression groups. **(d)** A network generated from the V-ex, W-ex, and N-ex mutated genes that were present in either the Cosmic or Tamborero datasets. **(e)** The distribution of mutation expression groups from the top 50 ranked mutated genes outputted by the driver mutation detection methods 2020+, Muffin, and OncodriveFM. **(f)** A Fisher's exact test was performed on three driver mutation detection methods to quantify their ability to identify cancer associated genes from the Cosmic and Tamborero, based on their 50 top-ranked mutated genes. Cancer gene statistical significance was calculated for either all top 50 mutated genes or the top 50 mutated genes that did not contain an N-ex mutation.

was even more extreme for N-ex mutations (Fig. 5c). However, W-ex mutations demonstrated more variability between PDXs and their corresponding cell lines, where most of the unique W-ex mutations fell within the PDX samples (Fig. 5d). While evaluating how mutation allele frequency associates with the common vs unique between PDX and cell lines, V-ex variants demonstrated that there is clearly a relationship (Fig. 5e). Although not statistically significant, W-ex mutations demonstrated an opposite relationship. This result suggests that in part, the lack of conservation of mutation expression is a reflection of the sub-clonal feature of the specific variant.

Mutation expression profiles from primary tumors. Primary tumors frequently contain stromal contamination, which decreases the sensitivity of mutation calls and diminishes the RNA-sequencing reads associated with the tumor³⁸. Consequently, stromal contamination has been reported to cause inaccurate results when performing DNA methylation and subtyping analyses^{28,39}. Thus, to limit the issue of tumor purity, we focused on the 75 TCGA PDAC cases that had the highest tumor purity score determined by the software ESTIMATE⁴⁰ and had a V-ex mutation percentage of at least 10%. Similar to the cell line and PDX models, the three mutation

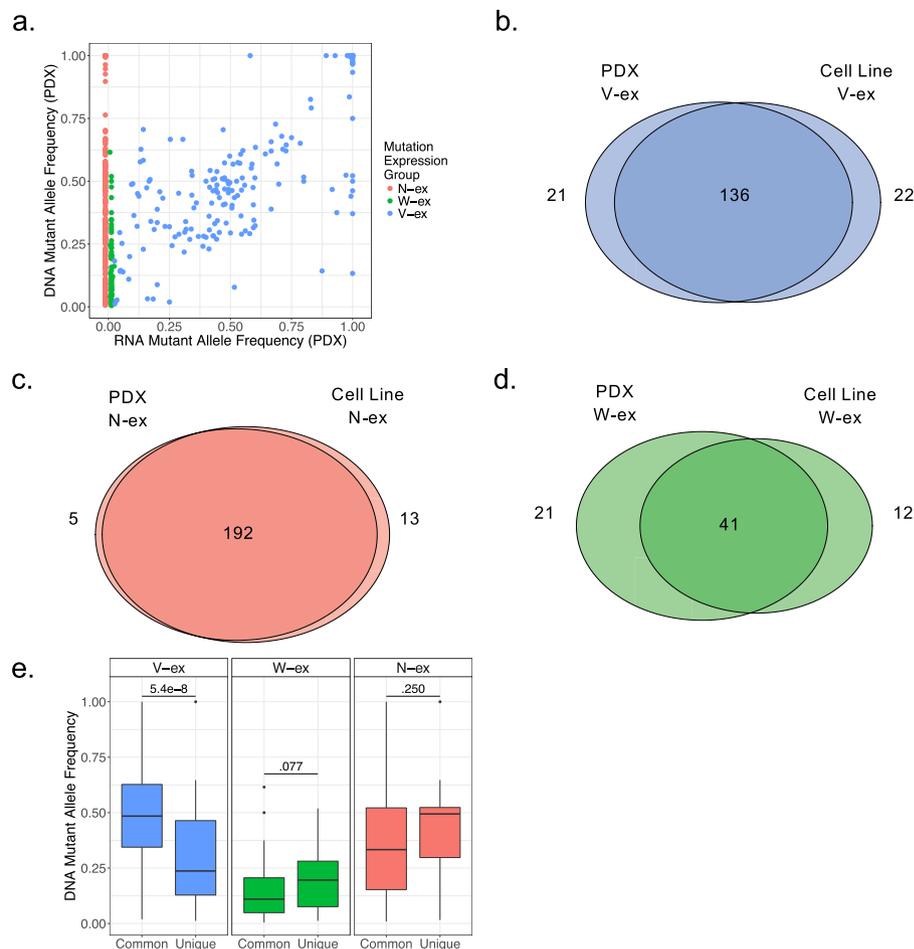


Figure 5. Conservation of mutational expression features from cell lines to PDX tumors. **(a)** Comparison of the DNA allele frequency and RNA allele frequency of each mutation derived from the PDX samples. **(b–d)** The overlap of V-ex, W-ex, and N-ex mutations between the corresponding cell line and PDX samples. Mutations classified as “discarded” in either the cell line or PDX data were excluded. **(e)** A two-tailed Wilcoxon Mann Whitney test between the DNA allele frequency of mutations that are present in both the cell line and PDX samples and the DNA allele frequency of mutations that are unique to either the cell line or PDX samples was performed for each mutation expression group.

expression groups were present in the TCGA samples (Fig. 6a). However, the distribution of the tumor DNA mutant allele frequency is mostly between 0–50%, rather than 0–100% as it was for the cell lines (Figs 6a vs. 3a). To determine if stromal contamination also altered the RNA allele frequency, we focused on KRAS, CDKN2A, TP53, and SMAD4 mutations (Fig. 6b). While the range of the RNA allele frequency for KRAS and CDKN2A is comparable to that observed in cell line data, the range significantly changed for TP53 and SMAD4. This suggests that SMAD4 and TP53 expression is observed in both the tumor specimen and the stromal compartment. In spite of this complexity, most mutations in these canonical PDAC mutations were identified to be expressed within the tumor, as seen in the cell line and PDX data.

Using mutation expression and methylation profiles from TCGA samples, we were able to interrogate the levels of DNA methylation across mutation expression groups. To measure the methylation of genes containing a mutation, the mean methylation percentage of each recorded CpG within the first exon was calculated⁴¹. Similar to the mutant gene expression analysis, the mean first exon methylation of samples that did not have a mutation in the gene was compared to the mean first exon methylation of samples that did contain the mutated gene (Fig. 6c). Evaluation of the coefficient of determination (R^2) for each expression group demonstrated that there is little deviation between wild type and mutated gene first exon methylation. However, when comparing the average mutated first exon methylation between mutation expression groups, the mean first exon methylation of N-ex mutations is significantly higher than the V-ex and W-ex mutations (Fig. 6d). This finding suggests that hypermethylation of the gene is a probable explanation for lack of expression among genes with an N-ex mutation. When comparing the first exon methylation between the V-ex and the W-ex groups, genes containing a W-ex mutation were identified to have a statistically higher first exon methylation. One possible explanation of why the W-ex group had a statistically lower first exon methylation than the N-ex group, yet a statistically higher first exon methylation than the V-ex group, is that these genes are susceptible to allele specific methylation⁴². Allele specific methylation of genes with a W-ex mutation would also explain why the mutant allele transcript from these genes is undetectable.

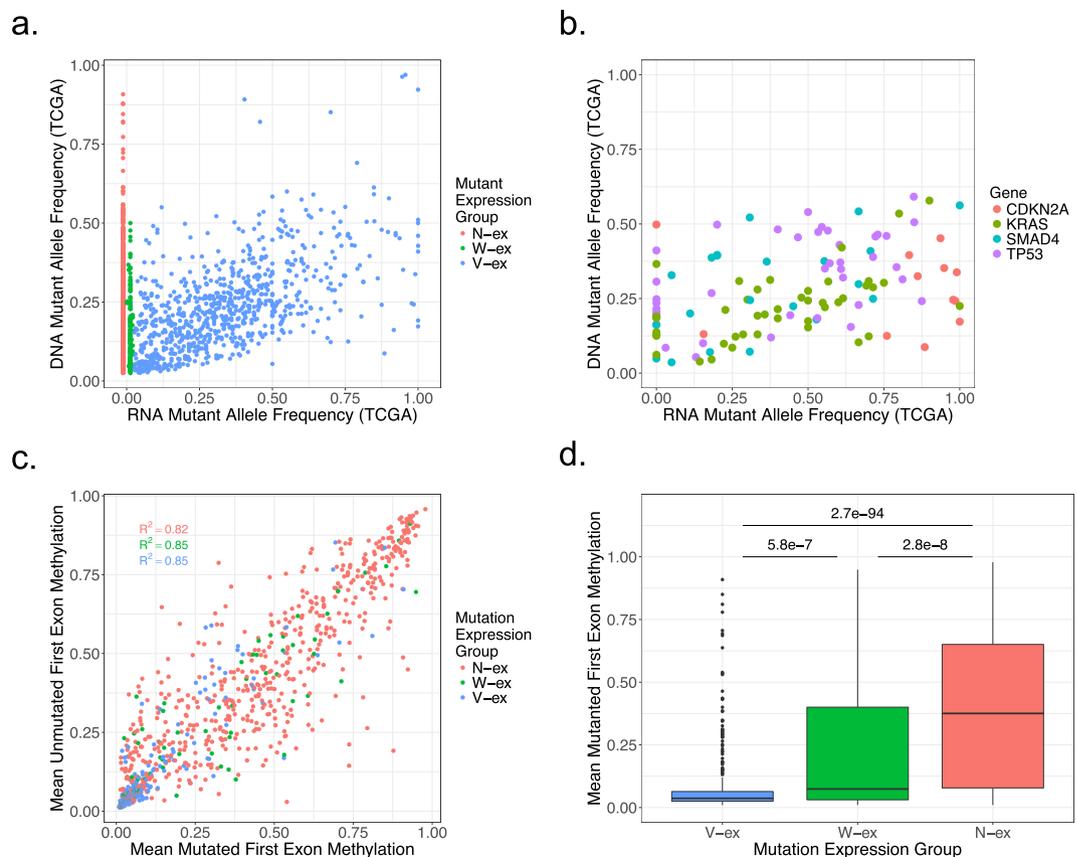


Figure 6. Mutation expression profiles from primary tumors. **(a)** Comparison of the DNA allele frequency and RNA allele frequency of each mutation derived from the TCGA samples. **(b)** Correlation of the DNA allele frequency and the RNA allele frequency of CDKN2A, KRAS, SMAD4, and TP53 mutations identified within the TCGA samples. **(c)** The average first exon methylation of samples that didn't contain the mutated gene was plotted against the average first exon methylation of samples that did contain the mutated gene. This was performed for non-discarded mutated genes that had available first exon methylation data. **(d)** A two-sample t-test with a two-tail p-value was performed on the mutant first exon methylation between each mutation expression group.

Separating mutations by expression significantly enhances PDAC subtyping. In order to determine if mutation expression groups provide insight into prognostic relevance, the Network Based Stratification (NBS)⁴³ pipeline was utilized on both the 19 patient-derived cell lines and 75 TCGA samples. NBS uses a protein-protein interaction network to separate samples into a predefined number of subtypes based on a mutation sample matrix. To quantify the prognostic capabilities of mutation expression groups, a survival log rank statistic $-\text{Log}_{10}(\text{p-value})$ was calculated for only the TCGA samples. The log rank statistic p-value was calculated for all possible combinations of mutation expression groups, using 3–8 predefined number of subtypes (Fig. 7a). The combination of mutation expression groups that had the strongest statistical prediction of prognosis was the V-ex and W-ex mutations. This held true regardless of the number of predefined subtypes. These findings further support that both the V-ex and W-ex mutations are relevant to tumor biology and occur within similar neighborhoods of the protein-protein interaction network. Comparatively, N-ex mutations have little influence on the progression of the tumor and appear to be randomly distributed within a protein-protein interaction network. To compare the prognostic prediction of the V-ex and W-ex mutations to gene rankings of popular driver mutation detection methods, we performed a similar NBS analysis on results produced by 2020+, Muffin and OncodriverFM (Fig. 7b). To perform an unbiased comparison between all methods, the same input data was used for each method to produce a ranked list of mutated genes. Then for the NBS analysis, the same number of ranked mutated genes was used to generate a mutation sample matrix ($n = 1,349$). Compared to each driver mutation detection method, the combination of the V-ex and W-ex mutations provided the best prognostic prediction for each number of predefined subtypes. Thus, these results imply that similar to Fig. 4f, current driver mutation detection methods report many false positives which contaminate results and affect downstream analyses. However, by separating mutations based on their allelic expression, it was possible to remove a large subset of mutations that appear to have little to no influence on the progression of the tumor and significantly enhance prognosis prediction with an emphasis on mutation data.

The most statistically significant number of stratified subtypes based on the V-ex and W-ex groups was three. The Kaplan-Meier plot generated from this result is shown in Fig. 7c. To determine which genes were unique to the

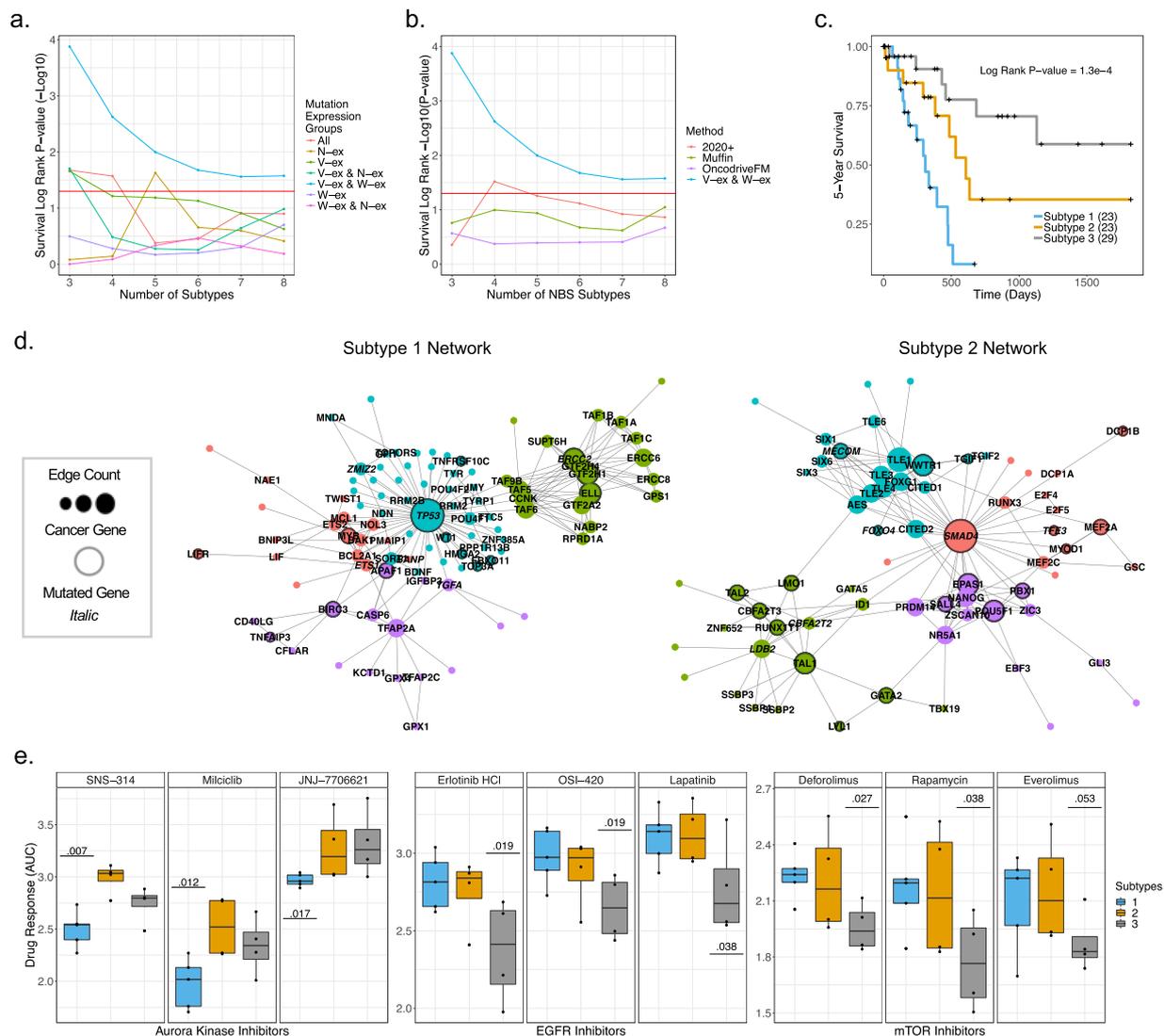


Figure 7. Separating mutations by expression enhances PDAC subtyping. **(a)** The log rank statistic $-\log_{10}(p\text{-value})$ based on the NBS results for 75 TCGA samples. The log rank statistic $-\log_{10}(p\text{-value})$ was calculated for the NBS results of all pairwise combinations of mutation expression groups from 3–8 number of predefined subtypes. **(b)** The log rank statistic $-\log_{10}(p\text{-value})$ based on the NBS results for 75 TCGA samples using the stratified samples based on results from 2020+, Muffin, OncodriveFM or V-ex and W-ex mutations. **(c)** Kaplan Meier plot for TCGA samples based on the NBS results using V-ex and W-ex mutations with number of subtypes equal to three. **(d)** Networks signifying the mutated pathways that are unique to subtype 1 and subtype 2. **(e)** Cell line drug response data, grouped by identified subtypes and statistical significance via a Wilcoxon Mann Whitney less than test.

newly established subtypes, a Wilcoxon Mann Whitney greater than test was performed on the NBS gene smooth scores between each two-way comparison of the three subtypes (e.g. subtype 1 vs subtype 2 and subtype 3). After adjusting the p-values using the Benjamini/Yekutieli method⁴⁴, networks were generated for each subtype using genes that contained an adjusted p-value less than 0.01 (Fig. 7d). Subtype 1 and subtype 2 had uniquely defined networks centered around TP53 and SMAD4, respectively. GO term analysis identified that subtype 1 mutations influence apoptosis pathways, while subtype 2 mutations are involved in cell differentiation pathways. In contrast, samples classified as subtype 3 had only a few statistically significant mutated interactions, and thus a well-structured network could not be generated. Interestingly, the sole mutation of either TP53 or SMAD4 was not sufficient to subtype the samples used in this study (Supplementary Fig. 6). This data supports the use of pathway and MAXX analysis to decipher new prognostic subtypes based on mutation data.

In order to evaluate if such subtyping could have relevance to treatment of PDAC, statistical analysis was performed using drug response data from 13 of our 19 patient derived cell lines (Fig. 7e). Among the cell lines with drug sensitivity data, five were categorized as subtype 1, four were categorized as subtype 2, and four were categorized as subtype 3. To determine the significance of drug treatments between subtypes, a Wilcoxon Mann Whitney less than test was performed on the drug response area under the curve (AUC) values between each

two-way comparison of the three subtypes. The drugs that were most consistent between subtype 1 and subtype 3 were, respectively, aurora kinase and EGFR inhibitors. Subtype 3 was also identified to respond better to mTOR inhibitors relative to the other subtypes. Unexpectedly, subtype 2 did not display selective sensitivity to any drug treatments. These results suggest that MAXX-derived data can be applied to predict drug sensitivity, in addition to prognosis.

Discussion

This study shows that tumor specific reference genomes generated via MAXX provide a substrate for a comprehensive analysis of mutation allelic expression. In comparison to standardized reference genomes, tumor specific reference genomes significantly enhance RNA-sequencing alignment of reads containing indel mutations. This feature allows accurate allelic expression to be detected for all mutation types. In the context of our analysis using nineteen patient derived cell lines, the capability to confidently detect allelic expression of indel mutations increased the mutation sample size by ~20%. MAXX also provided an unbiased allelic expression analysis among genes that commonly exhibit both SNV and indel variants (e.g. TP53 and SMAD4). The robustness of MAXX was measured on patient derived cell lines, PDX models, and primary tumors. Performing these analyses among different tumor models demonstrated that the MAXX pipeline is a comprehensive, consistent and computationally efficient method to identify mutation allelic expression. MAXX is also capable of generating customizable gene specific reference genomes that can be used as input for an alignment software to effectively query DNA or RNA allele frequencies for specific gene mutations from any organism with a reference genome and GTF file.

Centered on variants derived from PDAC, we report the discovery that mutation allelic expression can be used to separate mutations based on their respective impact on the tumor phenotype. The mutant expression group that was mainly responsible for the progression and clonality of the tumor was the V-ex group. This expression group consisted of not only the major PDAC oncogene (KRAS), but also all of the well-known PDAC tumor suppressor genes (TP53, SMAD4, and CDKN2A). Regardless of their assumed loss of function, our data identified that it is common for tumor suppressors to be transcriptionally present. This finding supports that RNA allele frequencies derived from allelic expression analysis can effectively be used to identify bona fide tumor suppressors. To illustrate, the well-known tumor suppressor gene SMAD4 commonly loses its expression of the wild type allele due to a homozygous deletion. Consequently, SMAD4's role as a tumor suppressor can be identified using both DNA mutant allele frequencies and RNA mutant allele frequencies. However, within our cell and PDX data, it was observed that multiple mutated genes had an RNA mutant allele frequency of 100%, despite having a DNA mutant allele frequency of ~50%. Thus, MAXX derived RNA mutant allele frequencies are more suited to identify genes that lose their expression of the wild type allele, which is an expected feature of tumor suppressor genes, compared to DNA mutant allele frequencies. Two genes from our data set that had a ~50% variant DNA allele frequency, a 100% variant RNA allele frequency, and have been demonstrated to act as tumor suppressors within the literature were DUSP5⁴⁵ and EI24⁴⁶. Additional genes that had similar DNA and RNA frequency as DUSP5 and EI24 but no previous research on their potential role as a tumor suppressor were UBXN11, CAPN15, C6orf62, GPRIN1, KIAA2018, PCDHGA10, and PCGF1.

The mutation expression group that appeared to have little effect on tumor progression was the N-ex group. Despite their relatively high mutant allele frequency, N-ex mutations were identified to be within genes that had a significantly lower transcript level than genes containing a V-ex or W-ex mutation. Using TCGA methylation data, we identified that a large proportion of genes containing an N-ex mutation had a hypermethylated first exon, regardless of the presence or absence of a mutation. This finding suggests that genes harboring an N-ex mutation are genes that are typically the target of epigenetic silencing across pancreatic cancer. Thus, the minimal impact of N-ex mutations on tumor progression is partially explained by their occurrence within genes that are developmentally silenced or are within genes that are more susceptible to transcription silencing among tumor samples. We observed that some N-ex classified genes have been previously identified to be associated with cancer. However, in comparison to cancer associated genes with a V-ex or W-ex mutation, cancer associated genes with an N-ex mutation had little impact on well-characterized mutated PDAC pathways such as MAPK cascade, autophosphorylation and stem cell maintenance. It was also observed that when taking into consideration whether a mutation was classified as an N-ex mutation, the mutation detection methods 2020+, Muffin, and OncodriveFM had an increased specificity for pre-classified cancer associated genes. Thus, we suggest that unexpressed genes that are the target of a mutation provide little insight into the current progression and therapeutic response of the tumor, relative to expressed genes that contain a mutation.

In addition to identifying groups of mutations that resemble driver mutations (V-ex) or passenger mutations (N-ex), we found a third mutation group, the W-ex group, which appeared to be an understudied subset of mutations. Despite the lack of the mutant allele transcripts, W-ex mutations appear to be biologically relevant to the progression of the tumor, similar to V-ex mutations. In comparison to the N-ex group, the W-ex group had a significant proportion of cancer associated genes and high gene expression levels. However, W-ex mutations were more sub clonal and infrequent within tumor samples, relative to the V-ex and N-ex groups. These unique features of W-ex mutations suggest that this mutation expression group is unlike traditional passenger or driver mutations. Based on our results, one possible explanation of why only the wild type allele is expressed in genes containing a W-ex mutation is that this particular mutation impairs the functionality or translation of a gene that is important to the tumor's progression. Thus, the tumor selectively silences the mutant allele and only expresses the wild type allele to maintain its phenotype. A potential example of this phenomena is the mutated gene LDHB in sample 810CL. A recent study that performed an RNA interference screen in KRAS dependent lung adenocarcinomas identified that LDHB was a strong regulator of cell proliferation in these tumors⁴⁷. LDHB has also been shown to be responsible for altering the metabolic addiction in PDAC⁴⁸. Interestingly, the mutated LDHB gene in sample 810CL had a DNA allele frequency of 62% and 1,587 reads that aligned to the wild type allele, but zero reads that aligned to the mutant allele. This supports the evidence that the non-mutated LDHB gene is potentially

a critical aspect within tumor sample 810CL. Additional confidently classified W-ex mutations from the cell line data that were noted to act in a similar manner as LDHB and have been shown to positively influence tumor progression are BCLAF1, JAK2, and KMT2D^{49–51}. However, to identify exactly what impact W-ex mutations have on tumor progression, additional research is necessary. In regards to how the mutant allele of W-ex mutations is repressed, we attempted to identify if this event was due to exon skipping or nonsense mediated decay due to having a higher proportion of nonsense and indel mutations⁵². Interestingly, neither of these mechanisms seemed to be involved in the silencing of the mutant allele. Alternatively, the mechanism of RNA-editing may be responsible for the inability to identify the mutant allele of W-ex mutations⁵³. However, based on our statistical testing between the first exon methylation of V-ex mutations and W-ex mutations, we would suggest that the most probable mechanism of W-ex mutant allele repression is allele specific methylation⁴².

By classifying PDAC mutations based on their allelic expression, we were able to identify ~50% of all mutations within this study as likely passenger mutations, increase the sensitivity of mutation detection methods, significantly enhance the prediction of PDAC prognosis using mutation data, and identify subtypes that are statistically sensitive to specific drug therapeutics in cell culture models. These results support the concept that both V-ex and W-ex mutations contribute to the biological features of the tumor, while N-ex mutations have little influence on the progression of the tumor. Thus, we conclude that the underutilized technique of allelic expression analysis of tumor mutations provides an effective top-level method to separate mutations based on their respective tumor impact. Based on these results within PDAC, we expect that allele expression analysis will be able to assist in other aspects of cancer biology. For example, this approach can be used to explore cancer epitope detection⁵⁴ and interrogation of tumor clonality^{54,55}. Because the MAXX pipeline enables accurate identification of allelic expression for all mutation types by increasing RNA-sequencing alignment, it represents the most inclusive method to perform mutation allelic expression analyses. We anticipate that custom gene specific reference genomes generated via MAXX will become a prevalent aspect in performing mutation allelic expression analyses and will be employed to classify mutations for advancement of precision medicine.

Methods

Expanded overview of MAXX pipeline. MAXX is a freely available tool (<https://github.com/Adglink/MAXX>) that generates gene specific reference genomes to identify allelic expression. As input, MAXX requires a mutation file, a reference genome, and the reference genome's corresponding GTF file. The reference files used in this study were GENCODE v19 for the patient derived models and GENCODE GRCh v21 for the TCGA data⁵⁶. In brief, MAXX uses information from the GTF file to identify the sequence of the input genes within the fasta file. Once the sequence of the genes within the mutation list is identified, MAXX first uses the gene name to create a header tag (e.g. >KRAS). Then the header tag and wild type sequence are written to the new fasta file. Next, the mutated sequence and its header are generated. The mutation sequence is created based on the information from the mutation file, while the header is the gene's name with a _Mut tag (e.g. >KRAS_Mut). The _Mut tag allows reads to be separated based on their alignment to either the wild type or mutated sequence. In cases where a gene contains multiple mutations, all mutations are inserted into the gene via a shifting algorithm. Because this study emphasized the allelic expression of mutations, MAXX is designed to only output the wild type and mutant sequence of genes presented in the mutation file. This approach dramatically decreases the size of the fasta file, which allows storage space and alignment run time to be more efficient compared to the Hg19 reference genome (Supplementary Fig. 1), while maintaining accurate RNA-sequencing alignment (Fig. 2c). In addition to the custom-made reference genome, MAXX also outputs an index file that identifies the new positions of the mutations in both the wild type and mutant sequence.

Because all of the PDAC samples within this study had undergone exome sequencing, we were able to use MAXX to generate a tumor specific reference genome for each sample. However, in situations where exome sequencing is not available, a reference genome could be generated using predefined mutations derived from other sequencing projects. Our tumor specific reference genomes were indexed using Bowtie2 v. 2.3.2⁵⁷. Then Tophat2 v. 2.1.1²⁶ was used to align the sample's corresponding RNA-sequencing reads. In addition to the default parameters for Tophat2, the commands “-p 10,-no-coverage-search,-b2-very-sensitive” were used. To ensure the accuracy of calculated RNA-allele frequencies, the BAM files generated by Tophat2 underwent multiple filtering steps^{23,32}. First, the PCR duplicates were removed, using the Picard Tools (<http://broadinstitute.github.io/picard/>) MarkDuplicates command where the input parameter REMOVE_DUPLICATES was set to “true”. Second, SAMtools v. 1.4⁵⁸ was used to keep the uniquely aligned reads via the command “samtools view -q 50 -b input.bam >output.bam”. Finally, because our data was paired-end, the additional step of singleton removal using the SAMtools v. 1.4 command “samtools view -F 8 -b input.bam >output.bam” was performed. The filtered bam file, tumor specific reference genome, and tumor mutation index file were then used as input for Bam-Readcount (<https://github.com/genome/bam-readcount>), which identified the number of reads that aligned to the mutant and wild type sequences.

RNA mutant allele frequency calculation. To determine RNA mutant allele frequencies from the Bam-Readcount output, the number of reads that aligned to the mutant position of the mutated sequence was first calculated, then divided by the sum of the reads that aligned to the mutant position of both the wild type and mutant sequences. Due to the shifting of nucleotides from indel mutations, there were slight variations in the way the reads were calculated for the wild type allele and mutant allele between the different mutation types. For SNV mutations, the same position for the mutant allele and wild type allele was used to calculate the number of reads for each allele. As for deletion mutations, the wild type allele reads were calculated based on the average number of reads mapped at each deleted nucleotide, while the first position of the non-deleted nucleotide was used to calculate the number of mutant allele reads. To calculate the wild type and mutant allele read count for insertions, the average read count of reads aligned to each inserted nucleotide was used for the mutant allele, while the first position of the non-inserted nucleotide was used to calculate the number of reads for the wild type allele.

Mutant expression classification expounded. Based on the variant's total read count and RNA mutant allele frequency, the mutation was placed into a mutation expression group (Fig. 1b). To prevent misalignments and poor sequencing depth at specific regions from biasing mutation expression group placement, mutations that had a total number of reads between three and nine were discarded from the study. It was also noted that a cut off of the number of reads that aligned to the mutant allele was required to differentiate W-ex mutations from V-ex mutations. Rather than have a distinct read count cutoff, the RNA mutant allele frequency of 2.5% was used to distinguish V-ex mutations from W-ex mutations.

Patient-derived cell lines gene expression. Gene read counts were obtained using HTSeq⁵⁹ on BAM files aligned to the Hg19 reference. The R package edgeR⁶⁰ was then used to normalize the gene read counts.

Patient-derived cell lines exon expression. Exon read counts were obtained using DEXseq⁶¹ on BAM files aligned to the Hg19 reference. The R package edgeR was then used to normalize the exon read counts.

Tissue specificity of patient-derived cell line variants. The 193 GTEx pancreas tissue gene expression profiles were obtained from the website https://zenodo.org/record/838734#.XSe_GS3Myw4⁶² under the R object "exp". Information classifying whether the expression of a gene is specific to pancreas tissue was also obtained from the previously mentioned website, under the R object expTS.

Patient-derived cell lines and PDX models. The establishment and variant calls of patient-derived cell lines and PDX models employed have been previously described^{24,63}. The samples were exome sequenced and variant calls were identified relative to the normal tissue from the patient from which the models were derived. There is a high level of conservation between the mutations present in the models and that shown in the primary tumor.

Sequencing of samples. The patient-derived cell lines and PDX models were subjected to exome sequencing and RNA sequencing as previously published^{24,63}.

Visualization of exome and RNA-sequencing reads. Bam files derived from exome or RNA sequencing alignment were uploaded into the Integrative Genomic Viewer software v. 2.3.93⁶⁴. The parameters filter duplicate reads and secondary alignments were set to true.

Generation and clustering of networks. The Cytoscape⁶⁵ environment, in conjunction with the Reactome database, was used to generate the networks presented in this study. After downloading the ReactomeFI⁶⁶ plugin for Cytoscape, our gene lists were uploaded using the "Gene Set/Mutation Analysis" option. The 2016 Reactome FI network version was used to visualize the results. The clustering analysis performed on the subtype networks was accomplished using the ReactomeFI plugin "Cluster FI Network" option. Only the clusters that contained at least 10% of the genes within the network were analyzed. GO Terms and associated p-values for all networks were calculated via the ReactomeFI plugin "Analyze Module Functions" option. Network visualization was enhanced using the R package ggnet2.

Driver mutation detection methods. As stated previously, the driver mutation detection methods 2020+¹⁹, Muffin¹⁷, and OncodriveFM¹⁵ were selectively chosen because of their capability to handle indel mutations and relatively small sample sizes. Each method was run two times using either the 19 patient derived cell line data or a combination of the 19 patient derived cell line data and 75 samples from TCGA. In addition to missense, frame shift deletion, frame shift insertion, and nonsense mutations, silence mutations were included in the input data for 2020+ and OncodriveFM, as recommended. 2020+ was run using the cancer type specific analysis procedure found at <http://2020+.readthedocs.io>. Muffin was run using the default parameters and the results produced by NDMAX with the Humannet network were used. OncodriveFM was run with the parameter "-gt 1", and as input the functional impact scores produced by SIFT¹⁴, PolyPhen-2¹³, Vest⁶⁷ and Chasni⁶⁸ were used.

MAXX mutation expression profiles for TCGA data. To generate a tumor specific reference genome for each sample within the TCGA-PAAD project, the somatic mutation data was obtained by downloading the PDAC Mutect2 annotation file from the website <https://portal.gdc.cancer.gov/repository>. This file, containing a list of all mutations for each TCGA PDAC sample, was used to generate MAXX appropriate mutation files for all available samples. The mutation files, GRCh38 v21 reference genome and GTF file were then used to generate tumor specific genomes via MAXX. Because the TCGA RNA-sequencing data needed to be re-aligned to the tumor specific genome, the RNA-seq fastq files were downloaded from <https://portal.gdc.cancer.gov/legacy-archive/> and aligned using Tophat2. The rest of the analysis follows the same protocol as the cell line and PDX data to calculate mutation expression groups. Due to the confounding feature of tumor purity, we focused on the 75 TCGA PDAC cases that had the highest tumor purity score and at least 10% of the mutations were classified as V-ex mutations. To assess the tumor purity of each TCGA PDAC sample, purity scores established by Yoshihara *et al.*⁴⁰ were downloaded from <http://bioinformatics.mdanderson.org/main/ESTIMATE>.

Methylation analysis expounded. The Illumina Infinium HumanMethylation450 platform level 3 generated DNA methylation files for the TCGA samples used in this study were downloaded from <https://portal.gdc.cancer.gov/repository>. Calculation of the first exon methylation was performed using a custom python script. The Hg19 reference genome and corresponding GTF file was first used to identify the first exon position of all genes containing a mutation; then the first exon methylation percentage was determined by taking the average beta values for all CpG nucleotides within the first exon. Genes with mutations that were identified to be within multiple mutation expression groups between TCGA samples were excluded from the methylation analysis, as well as

mutations that were classified as “discarded” (Fig. 1b). Also, it was observed that approximately half of the genes within each mutation expression group did not contain first exon methylation data. This may occur because the CpG methylation was not available or analyzed in the first exon of these genes within the downloaded data.

Network based stratification analysis. The NBS software was used to stratify the cell line and TCGA samples based on mutations associated with each combination of mutation expression groups. The NBS algorithm was developed by Hofree *et al.*⁴³, but the implementation of NBS by Morvan *et al.*⁶⁹ was used to stratify our samples. The source code for this NBS implementation can be found at the github repository <https://github.com/marineLM/NetNorM>. Multiple methods of stratification are provided, but to replicate the NBS method, the *stratification_NMF.py* script was used. This script requires a sample mutation matrix and a node edge matrix. The Pathway Commons v6 network⁷⁰ (Commons.6.All.EXTENDED_BINARY_SIF.tsv) was used to generate the node edge matrix, while different combinations of mutation expression groups were used to generate the sample mutation matrix. To replicate our results, the following commands should be used as input for the *stratification_NMF.py* function: `method_rep = 'smoothing'`, `method_norm = 'qn'`, `k = 'NA'`, `alpha = '0.5'`, `randomized = 'False'`, `rs_rand = 'NA'`, and `N = '3'`.

Survival curves. Survival information for the TCGA PDAC samples was obtained from the nationwideshildrens.org_clinical_patient_paad.txt file, downloaded from <https://portal.gdc.cancer.gov/legacy-archive/> on September 6, 2017. The R package survival was then used to calculate the log rank statistic, log rank p-value, and generate the Kaplan Meier plot.

Data Availability

The MAXX software, which generates tumor specific references genomes based on a mutation list is available for download at <https://github.com/Adglink/MAXX>.

References

- Loeb, K. R. & Loeb, L. A. Significance of multiple mutations in cancer. *Carcinogenesis* **21**, 379–385 (2000).
- Loeb, L. A., Loeb, K. R. & Anderson, J. P. Multiple mutations and cancer. *Proc Natl Acad Sci USA* **100**, 776–781, <https://doi.org/10.1073/pnas.0334858100> (2003).
- Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421, <https://doi.org/10.1038/nature12477> (2013).
- Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724, <https://doi.org/10.1038/nature07943> (2009).
- Tomasetti, C., Marchionni, L., Nowak, M. A., Parmigiani, G. & Vogelstein, B. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proc Natl Acad Sci USA* **112**, 118–123, <https://doi.org/10.1073/pnas.1421839112> (2015).
- Santarpia, L. *et al.* Deciphering and Targeting Oncogenic Mutations and Pathways in Breast Cancer. *Oncologist* **21**, 1063–1078, <https://doi.org/10.1634/theoncologist.2015-0369> (2016).
- Yap, T. A., Omlin, A. & de Bono, J. S. Development of therapeutic combinations targeting major cancer signaling pathways. *J Clin Oncol* **31**, 1592–1605, <https://doi.org/10.1200/JCO.2011.37.6418> (2013).
- Pon, J. R. & Marra, M. A. Driver and passenger mutations in cancer. *Annu Rev Pathol* **10**, 25–50, <https://doi.org/10.1146/annurev-pathol-012414-040312> (2015).
- Sharma, S. V. & Settleman, J. Oncogene addiction: setting the stage for molecularly targeted cancer therapy. *Genes Dev* **21**, 3214–3231, <https://doi.org/10.1101/gad.1609907> (2007).
- Bamford, S. *et al.* The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br J Cancer* **91**, 355–358, <https://doi.org/10.1038/sj.bjc.6601894> (2004).
- Raphael, B. J., Dobson, J. R., Oesper, L. & Vandin, F. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med* **6**, 5, <https://doi.org/10.1186/gm524> (2014).
- Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501, <https://doi.org/10.1038/nature12912> (2014).
- Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248–249, <https://doi.org/10.1038/nmeth0410-248> (2010).
- Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812–3814 (2003).
- Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res* **40**, e169, <https://doi.org/10.1093/nar/gks743> (2012).
- Leiserson, M. D. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet* **47**, 106–114, <https://doi.org/10.1038/ng.3168> (2015).
- Cho, A. *et al.* MUFFINN: cancer gene discovery via network analysis of somatic mutation data. *Genome Biol* **17**, 129, <https://doi.org/10.1186/s13059-016-0989-x> (2016).
- Zhang, J. *et al.* Identifying driver mutations from sequencing data of heterogeneous tumors in the era of personalized genome sequencing. *Brief Bioinform* **15**, 244–255, <https://doi.org/10.1093/bib/bbt042> (2014).
- Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B. & Karchin, R. Evaluating the evaluation of cancer driver genes. *Proc Natl Acad Sci USA* **113**, 14330–14335, <https://doi.org/10.1073/pnas.1616440113> (2016).
- Fleck, J. L., Pavel, A. B. & Cassandras, C. G. Integrating mutation and gene expression cross-sectional data to infer cancer progression. *BMC Syst Biol* **10**, 12, <https://doi.org/10.1186/s12918-016-0255-6> (2016).
- Gerstung, M. *et al.* Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nat Commun* **6**, 5901, <https://doi.org/10.1038/ncomms6901> (2015).
- Stevenson, K. R., Coolon, J. D. & Wittkopp, P. J. Sources of bias in measures of allele-specific expression derived from RNA-sequence data aligned to a single reference genome. *BMC Genomics* **14**, 536, <https://doi.org/10.1186/1471-2164-14-536> (2013).
- Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biol* **16**, 195, <https://doi.org/10.1186/s13059-015-0762-6> (2015).
- Witkiewicz, A. K. *et al.* Integrated Patient-Derived Models Delineate Individualized Therapeutic Vulnerabilities of Pancreatic Cancer. *Cell Rep* **16**, 2017–2031, <https://doi.org/10.1016/j.celrep.2016.07.023> (2016).
- Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat Genet* **45**, 1113–1120 (2013).
- Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013).

27. Mitra, A., Mishra, L. & Li, S. Technologies for deriving primary tumor cells for use in personalized cancer therapy. *Trends Biotechnol* **31**, 347–354, <https://doi.org/10.1016/j.tibtech.2013.03.006> (2013).
28. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat Commun* **6**, 8971, <https://doi.org/10.1038/ncomms9971> (2015).
29. Rashid, N. U. *et al.* Differential and limited expression of mutant alleles in multiple myeloma. *Blood* **124**, 3110–3117 (2014).
30. Engström, P. G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature methods* **10**, 1185–1191 (2013).
31. Sun, Z., Bhagwate, A., Prodduturi, N., Yang, P. & Kocher, J. A. Indel detection from RNA-seq data: tool evaluation and strategies for accurate detection of actionable mutations. *Brief Bioinform* **18**, 973–983, <https://doi.org/10.1093/bib/bbw069> (2017).
32. Quinn, E. M. *et al.* Development of strategies for SNP detection in RNA-seq data: application to lymphoblastoid cell lines and evaluation using 1000 Genomes data. *PLoS One* **8**, e58815, <https://doi.org/10.1371/journal.pone.0058815> (2013).
33. Castle, J. C. *et al.* Mutated tumor alleles are expressed according to their DNA frequency. *Sci Rep* **4** (2014).
34. Kazanets, A., Shorstova, T., Hilmi, K., Marques, M. & Witcher, M. Epigenetic silencing of tumor suppressor genes: Paradigms, puzzles, and potential. *Biochim Biophys Acta* **1865**, 275–288, <https://doi.org/10.1016/j.bbcan.2016.04.001> (2016).
35. Maitra, A. & Hruban, R. H. Pancreatic cancer. *Annu Rev Pathol* **3**, 157–188, <https://doi.org/10.1146/annurev.pathmechdis.3.121806.154305> (2008).
36. Krasinskas, A. M., Moser, A. J., Saka, B., Adsay, N. V. & Chiosea, S. I. KRAS mutant allele-specific imbalance is associated with worse prognosis in pancreatic cancer and progression to undifferentiated carcinoma of the pancreas. *Mod Pathol* **26**, 1346–1354, <https://doi.org/10.1038/modpathol.2013.71> (2013).
37. Tamborero, D. *et al.* Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci Rep* **3**, 2650, <https://doi.org/10.1038/srep02650> (2013).
38. Griffith, M. *et al.* Optimizing cancer genome sequencing and analysis. *Cell Syst* **1**, 210–223, <https://doi.org/10.1016/j.cels.2015.08.015> (2015).
39. Zhang, W., Feng, H., Wu, H. & Zheng, X. Accounting for tumor purity improves cancer subtype classification from DNA methylation data. *Bioinformatics* **33**, 2651–2657, <https://doi.org/10.1093/bioinformatics/btx303> (2017).
40. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* **4**, 2612, <https://doi.org/10.1038/ncomms3612> (2013).
41. Brenet, F. *et al.* DNA methylation of the first exon is tightly linked to transcriptional silencing. *PLoS One* **6**, e14524, <https://doi.org/10.1371/journal.pone.0014524> (2011).
42. Meaburn, E. L., Schalkwyk, L. C. & Mill, J. Allele-specific methylation in the human genome: implications for genetic studies of complex disease. *Epigenetics* **5**, 578–582, <https://doi.org/10.4161/epi.5.7.12960> (2010).
43. Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nature methods* **10**, 1108–1115 (2013).
44. Benjamini, Y. & Yekutieli, D. The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics* **29**, 1165–1188 (2001).
45. Rushworth, L. K. *et al.* Dual-specificity phosphatase 5 regulates nuclear ERK activity and suppresses skin cancer by inhibiting mutant Harvey-Ras (HRasQ61L)-driven SerpinB2 expression. *Proc Natl Acad Sci USA* **111**, 18267–18272, <https://doi.org/10.1073/pnas.1420159112> (2014).
46. Mork, C. N., Faller, D. V. & Spanjaard, R. A. Loss of putative tumor suppressor EI24/PIG8 confers resistance to etoposide. *FEBS Lett* **581**, 5440–5444, <https://doi.org/10.1016/j.febslet.2007.10.046> (2007).
47. McClelland, M. L. *et al.* Lactate dehydrogenase B is required for the growth of KRAS-dependent lung adenocarcinomas. *Clin Cancer Res* **19**, 773–784, <https://doi.org/10.1158/1078-0432.CCR-12-2638> (2013).
48. Blum, R. & Kloog, Y. Metabolism addiction in pancreatic cancer. *Cell Death Dis* **5**, e1065, <https://doi.org/10.1038/cddis.2014.38> (2014).
49. Zhou, X. *et al.* BCLAF1 and its splicing regulator SRSF10 regulate the tumorigenic potential of colon cancer cells. *Nat Commun* **5**, 4581, <https://doi.org/10.1038/ncomms5581> (2014).
50. Thomas, S. J., Snowden, J. A., Zeidler, M. P. & Danson, S. J. The role of JAK/STAT signalling in the pathogenesis, prognosis and treatment of solid tumours. *Br J Cancer* **113**, 365–371, <https://doi.org/10.1038/bjc.2015.233> (2015).
51. Guo, C. *et al.* KMT2D maintains neoplastic cell proliferation and global histone H3 lysine 4 monomethylation. *Oncotarget* **4**, 2144–2153, <https://doi.org/10.18632/oncotarget.1555> (2013).
52. Frischmeyer, P. A. & Dietz, H. C. Nonsense-mediated mRNA decay in health and disease. *Hum Mol Genet* **8**, 1893–1900 (1999).
53. Gagnidze, K., Rayon-Estrada, V., Harroch, S., Bulloch, K. & Papavasiliou, F. N. A New Chapter in Genetic Medicine: RNA Editing and its Role in Disease Pathogenesis. *Trends Mol Med* **24**, 294–303, <https://doi.org/10.1016/j.molmed.2018.01.002> (2018).
54. Karasaki, T. *et al.* Prediction and prioritization of neoantigens: integration of RNA sequencing data with whole-exome sequencing. *Cancer Sci* **108**, 170–177, <https://doi.org/10.1111/cas.13131> (2017).
55. Peterson, E. A. *et al.* Enhancing cancer clonality analysis with integrative genomics. *BMC Bioinformatics* **16**(Suppl 13), S7, <https://doi.org/10.1186/1471-2105-16-S13-S7> (2015).
56. Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760–1774 (2012).
57. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357–359, <https://doi.org/10.1038/nmeth.1923> (2012).
58. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
59. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
60. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140, <https://doi.org/10.1093/bioinformatics/btp616> (2010).
61. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res* **22**, 2008–2017, <https://doi.org/10.1101/gr.133744.111> (2012).
62. Sonawane, A. R. *et al.* Understanding Tissue-Specific Gene Regulation. *Cell Rep* **21**, 1077–1088, <https://doi.org/10.1016/j.celrep.2017.10.001> (2017).
63. Knudsen, E. S. *et al.* Pancreatic cancer cell lines as patient-derived avatars: genetic characterisation and functional utility. *Gut* **67**, 508–520, <https://doi.org/10.1136/gutjnl-2016-313133> (2018).
64. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24–26, <https://doi.org/10.1038/nbt.1754> (2011).
65. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res* **13**, 2498–2504 (2003).
66. Wu, G., Dawson, E., Duong, A., Haw, R. & Stein, L. ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis. *F1000Res* **3**, 146, <https://doi.org/10.12688/f1000research.4431.2> (2014).
67. Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14**(Suppl 3), S3, <https://doi.org/10.1186/1471-2164-14-S3-S3> (2013).
68. Carter, H. *et al.* Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res* **69**, 6660–6667, <https://doi.org/10.1158/0008-5472.CAN-09-1133> (2009).
69. Le Morvan, M., Zinovyev, A. & Vert, J. P. NetNorM: Capturing cancer-relevant information in somatic exome mutation data with gene networks for cancer stratification and prognosis. *PLoS Comput Biol* **13** (2017).
70. Cerami, E. G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res* **39**, D685–690 (2011).

Acknowledgements

The authors thank all members of the Knudsen and Witkiewicz laboratory for thought-provoking discussion, technical assistance and help with manuscript preparation. We also thank the reviewers for their insightful comments and suggestions, which have improved the quality of this study. This project was supported by the National Institute of Health grant R01CA211878.

Author Contributions

A.D.G. is responsible for creating the MAXX pipeline, data generation and writing most of the manuscript under the mentorship of E.S.K., A.K.W. and M.P. P.V. provided fundamental scripts to perform gene expression and survival analyses. P.V. also contributed to the development and edits of the study.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-019-48967-8>.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019