Genomic characterization of additional cancer-driver genes using a weighted iterative regression accurately modelling background mutation rate

Lin Jiang[1,2,#], Jingjing Zheng[1,#], Johnny Sheung Him Kwan[9,10,11,#], Sheng Dai[1], Cong Li[1], Ka Fai TO[9,10,11], Pak Chung Sham[5,6,7,8,*], Yonghong Zhu[1,2,*] and Miaoxin Li[1,2,3,4,5,6,7*]

[1]Zhongshan School of Medicine,[2]First Affiliated Hospital, [3]Center for Genome Research, [4]Center for Precision Medicine, Sun Yat-sen University, Guangzhou 510080, China; [5]Key Laboratory of Tropical Disease Control (SYSU), Ministry of Education, Guangzhou 510080, China; [6]The Centre for Genomic Sciences, [7]Department of Psychiatry, [8]State Key Laboratory for Cognitive and Brain Sciences, the University of Hong Kong, Pokfulam, Hong Kong; [9]Department of Anatomical and Cellular Pathology, [10]State Key Laboratory in Oncology in South China, [11]Li Ka-Shing Institute of Health Sciences, The Chinese University of Hong Kong, New Territories, Hong Kong

Short title: a powerful approach to detect cancer-driver genes

* To whom correspondence should be addressed to

Miaoxin Li. Tel: +86-2087335080; Email: limiaoxin@mail.sysu.edu.cn Medical Science and Technology Building, Zhongshan School of Medicine, Sun Yat-sen University, China.

Yonghong Zhu. Tel: +86-20 87331451; Email: zhuyongh@mail.sysu.edu.cn Medical Science and Technology Building, Zhongshan School of Medicine, Sun Yat-sen University, China.

Pak-Chung Sham. Tel: +852-28315425; Fax: +852-28185653; Email: pcsham@hku.hk; The University of Hong Kong, 5 Sassoon Road, Pokfulam Hong Kong SAR".

[#] The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

Abstract:

Genomic aberrations in somatic cells are major drivers of cancers and cancers are of high genetic heterogeneity and most driver genes are only of moderate or small effect size. Existing bioinformatics methods poorly model background mutations and are underpowered to identify driver genes in typical-size samples. Here we propose a novel statistical approach, weighted iterative zero-truncated negative-binomial regression (WITER), to detect cancer-driver genes showing an excess of somatic mutations. This approach has a three-tier framework to improve power in small or moderate samples by accurately modelling background mutations. Compared to alternative methods, this approach detected more significant and cancer-consensus genes in all tested cancers. This technical advance enables the detection of driver genes in TCGA datasets as small as 30 subjects, rescuing genes missed by alternative tools. By introducing an advanced statistical model for accurately

estimating the background mutation rate even in small-to-moderate samples, the proposed method is more powerful approach for detecting cancer driver genes than current methods, helps provide a comprehensive landscape of driver genes in cancers.

Running title: powerful cancer-driver gene detection method

Keywords: iterative zero-truncated negative-binomial regression, cancer-driver genes, somatic mutations, passenger genes, small sample

Introduction

It is well known that genomic aberration in somatic cells makes an important contribution to the development of cancers(*1*). Mutations that confer selective growth advantage to cancer cells are known as cancer-drivers (*2*) (*3*); a gene harboring driver-mutations is defined as a cancer-driver gene. It has been established, for example, that non-synonymous mutations in the two famous driver genes TP53 and PIK3CA contribute to many types of cancer (*4*). However, cancers are known to be highly heterogeneous(*5*) and many driver genes for most cancers remain to be identified. A full landscape of driver-genes is critical for early diagnosis, identification of effective drug targets, and precise treatments of a cancer (*2*).

There are generally two existing strategies to detect cancer driver genes, background mutation rate (BMR) and ratiometric. The BMR-based methods evaluate whether a gene has more somatic mutations than expected; examples include MutSigCV (*6*) and MuSiC(*7*). The expected number of mutations is estimated from multiple predictors including base context, gene size and other variables of “background” passenger genes. In particular, MutSigCV proposed using three extra variables (DNA replication timing, transcriptional activity and chromatin state in cancer cells) to improve the prediction of the expected background mutation rate. The ratiometric-based methods detect cancer-driver genes according to the composition of mutation types normalized by the total number of mutations in a gene. For instance, the ratiometric 20/20 rule simply assesses the proportion of inactivating mutations (including synonymous mutations) and missense mutations(*3*). Oncodrive-fm(*8*) and OncodriveFML(*9*) integrate mutation functional impact into the evaluation. OncodriveCLUST considers the positional clustering of mutation patterns(*10*). Recently a method 20/20 plus (*11*) extended the ratiometric idea in the 20/20 rule and integrated 18 additional features of positive selection to predict cancer-driver genes by a machine learning approach. It also generated statistical p-values of the prediction scores by Monte Carlo simulations.

Although the general principles of both strategies are simple, technical issues remain especially when sample size is not sufficiently large. For example, a recent study (*11*) found that the statistical p-values produced by existing cancer-driver gene methods did not follow uniform distribution, implying the underfitting of background mutations. Although simulation or permutation can correct the distribution, adequate fitting of background genes is critical for accurate discrimination of true driver genes from noise background genes. This issue will become more severe when the sample is too small to generate a stable model. Moreover, existing statistical tests are generally underpowered to detect driver-genes with small or moderate effect size. This may be a reason why a supervised approach integrating common gene features beyond collected samples was also proposed. However, given the high heterogeneity in cancers (*6*), adding more common features may not work for unique driver genes; and the trained model for known driver genes may have limited power for detecting new driver genes. Lastly, the predicted cancer-driver genes by different tools do not generally agree with each other(*2*). It is often laborious and subjectively biased to combine their results. Therefore, more powerful methods are pressingly needed for unraveling a full spectrum of cancer-driver genes.

Here, we describe a new statistical method, weighted iterative zero-truncated negative-binomial regression (WITER), to detect cancer-driver genes by somatic mutations at non-synonymous variants. This approach belongs to the unsupervised category and therefore does not suffer from training bias. The method has a unique three-tier structure to accurately fit the number of somatic mutations in background genes. This structural advance enables it to detect more driver genes in both small and large samples regardless of cancer types. Although it is basically a type of BMR approach, it also adopts the ratiometric idea to use silent mutations as an explanatory variable in the regression model. We then investigated its performance in 34 cancers. A comprehensive landscape of driver-genes was constructed by WITER and analyzed to investigate the common and unique insights across cancers.

Results

Overview of the statistical framework

We propose a unified statistical framework, WITER, for detecting cancer-driver genes by somatic mutations in cancers. The main input is somatic mutations in samples from cancer patients. The output is a table of p-values for excess of somatic non-synonymous mutations at individual genes; a significant p-value would suggest a driver gene that has more somatic mutations because such mutations confer selective growth advantages to cancer cells. Compared to alternative approaches, it has a unique three-tier structure to accurately fit the number of somatic mutations in background genes (See its diagram in Figure 1). In the first tier, it has an advanced model, iterative zero-truncated negative-binomial distribution regression (ITER), to fit the zero-inflation and

overdispersion of mutation counts. As shown in the following section, the model can more accurately fit the background mutation counts compared to other widely-used models [Figure 3b and 3c]. In addition, p-values can be straightforwardly derived by deviance residuals of the regression [Figure 2 and 3a], so that conventional time-consuming simulations are not needed for significance evaluation. The iterative procedure reduces the distortion of the background mutation rates by the exclusion of the driver genes in the model, so that the relative excess of mutations can be measured more accurately. In the second tier, it can flexibly impose prior weights upon mutation counts to further boost statistical power. The prior weights are generated by a random forest model trained by a large dataset we curated from COSMIC(V83) [Figure S6]. The weighting scheme contributes to identification of extra significant genes which would be missed by same framework without weights [Figure 3b and 3c, Table S8]. In the third tier, it allows the integration of independent reference samples from either the same or different cancers to produce a stable background model. This solves the problem of model instability in small cancer samples. This feature enables WITER to produce statistically valid p-values (Figure S1) and to detect multiple significant driver-genes (Table 1) in datasets with around 30 subjects. The approach and auxiliary functions have been implemented into a user-friendly software tool which is publicly available at http://grass.cgs.hku.hk/limx/witer.

Distributions of *p*-values for background mutation genes

The p-values of the proposed approach approximately followed uniform distribution. When the overall divergence from uniform p-values was measured as the mean log fold change (MLFC) of Tokheim et al (2016), the MLFCs of MutSigCV and OncodriveFML deviate from zero in all the cancers substantially, suggesting a large deviation from uniform distribution (Figure 3a). Consistent with the QQ plots, ITER and WITER had very low absolute MLFC (<0.02) in all the 11 cancers (Figure 3a). Moreover, as shown in the QQ plots (Figure 2), the p-values produced by WITER, were close to the uniform distribution (corresponding to null hypotheses) in all the cancers. This was also true for the unweighted version, iterative zero-truncated negative-binomial regression (ITER). Invalid uniform distribution of *p*-values is a tricky problem in almost all existing approaches (*11*). We chose two alternative approaches which achieved the best performance among 7 widely-used unsupervised tools (*11*)for the comparison, MusigCV (*6*) and OncodriveFML (*9*). The MutSigCV produced deflated statistical p-values in 10 cancers except that it produced proximately uniform distribution p-values in melanoma (MEL) (Figure 2). The OncodriveFML also produced deflated statistical p-values in all the 11 cancers (*9*) (Figure 2).

Significant genes identified in the 11 cancers with relatively large mutation number

We compared the number of significant genes detected by the 4 un-supervised approaches (MutSigCV, OncodriveFML, ITER and WITER) in the 11 cancer datasets. Instead of following the conventional “pancancer” (all cancers) evaluation strategy (*11*), we made the comparison for

individual cancers, a more challenging scenario. The significant genes are determined according to a widely adopted cutoff in cancer-driver gene analysis, FDR<0.1(*11*). WITER always detected the largest number of significant genes in the 11 cancers among the 4 approaches (Figure 3b). ITER detected the second largest number of significant genes in 10 out of the 11 cancers. MutSigCV can be ranked at the third place according to the number of significant genes. The OncodriveFML detected the minimal number of significant genes in all the 11 cancers although it also integrated function prediction score, CADD (*12*). Compared to MutSigCV, WITER detected at least 8 extra significant genes in all cancers. The extra significant gene number increased to be at least 13 when comparing to OncodriveFML. WITER detected at least 12 more significant genes than ITER in 9 out the 11 cancers (bladder urothelial carcinoma, breast invasive carcinoma, colorectal adenocarcinoma, uterine corpus endometrial carcinoma, kidney renal clear cell carcinoma, lung adenocarcinoma, melanoma, ovarian carcinoma and stomach adenocarcinoma), suggesting the prior weights of frequent mutation potential at variants have great potential to improve the statistical power. Note that all the subjects in the testing cancer datasets were excluded from the COSMIC database to avoid circulating issues when building the prior weights for WITER.

Cancer consensus significant genes in the 11 cancers

We further checked the significant genes in the Cancer Gene Census (CGC) list (*13*) detected by the tools. Again, among the 4 methods, WITER always detected the largest number of genes in CGC list (Figure 3c). It detected at least 10 more CGC genes for 6 cancers than ITER. Anyhow, ITER was still the second-best method according to the number of CGC genes. It detected more CGC genes than MutSigCV in 10 cancers while it reported more CGC genes than OncodriveFML in all the 11 cancers. The OncodriveFML reported the smallest number of CGC genes in 10 cancers. Note we did not compare the percentage of the CGC genes in the total significant genes because the number of total significant genes by OncodriveFML were too few. Moreover, it should be noted that significant genes beyond CGC list are not necessarily spurious driver genes although a higher number of CGC genes is a strong sign of higher power. Take two non-CGC genes for examples. The AJUBA gene (p=8.1E-8 in head and neck cancer) is involved in the regulation of NOTCH/CTNNB1 signaling and is an important driver gene of head and neck cancer (*14*), (*15*). TLR4 (p=1.1E-4 in stomach adenocarcinoma) is an important member of Toll-like receptor (TLR) pathway and mutations in the gene may disrupt innate immune signaling and promote a microenvironment that favors tumorigenesis (*16*) and it was associated with gastric cancer in independent samples (*17*).

Unique significant genes by individual approaches

We also compared the number of unique significant genes by different tool. WITER detected the largest number of unique significant genes (FDR≤0.1) in the tested cancer types, which were insignificant and would be ignored by MutSigCV and OncodriveFML (Figure 3d). This was also true

for the unique significant CGC genes by WITER(Figure 3e). WITER detected in total 267 unique significant genes and 133 unique CGC genes for all 11 cancers. Each cancer had at least 8 unique significant genes (Figure 3d). The colorectal adenocarcinoma (COAD) had the largest number of unique significant genes by WITER, 44, among which 16 genes were CGC genes. For example, CTNNB1 is a well-known driver gene for colorectal adenocarcinoma (*18*). It had 11 non-synonymous somatic mutant alleles in the colorectal adenocarcinoma samples. WITER calculated a p-value 1.44E-10 at this gene. The p-values by MutSigCV and OncodriveFML were 0.001 and 0.51 respectively. In contrast, MutSigCV detected no unique significant genes (FDR≤0.1) in 6 cancers and ≤3 unique significant genes in 4 cancers. The only exception was the lung adenocarcinoma for which MutSigCV detected 8 unique significant genes out of the 16 significant genes (FDR≤0.1). Many of the 8 unique significant genes had either long coding region or multiple synonymous mutations or close chromatin states. After correcting for the explanatory variables, WITER produced an insignificant p-value. For example, FBN2 has 77 non-synonymous or splicing mutant alleles in the lung cancer patients and MutSigCV gave a p-value 4.28E-07. However, it had a 9.1 kb coding region, 10 synonymous mutant alleles and close chromatin state (scored 9), WITER gave an insignificant p-value 0.25 for the excess of corrected non-synonymous or splicing mutant alleles. Similarly, OncodriveFML also detected very few unique significant genes in each of the 11 cancers compared to WITER. In the comparison, we ignored ITER because all significant genes by ITER were also significant by WITER. These results suggest that the WITER has great power to detect many potential driver genes in many cancers, which might be ignored by widely-used alternative methods.

Moreover, we also investigated the enrichment significance of known cancer related genes in the unique significant genes by WITER. Using the 19198 protein coding gene as the population size and 699 CGC genes as the number of success states, we performed enrichment analysis by hypergeometric distribution test. As shown in Table S2, the unique driver genes by WITER in all the 11 cancers were significantly enriched with the CGC genes (p<5.56E-8). In addition, we also performed a rough *in-silico* validation for all genes by searching literature co-mentioning the gene symbols and the specific cancer names in titles and abstracts of papers from the NCBI PubMed database by July 10, 2018. As it is very time-consuming to check the hit papers for all coding genes, we drew a random gene set of the same size for each cancer and performed Fisher's exact test. Due to the small random sample size, it was much more conservative than the hypergeometric distribution test. The genes with three or more hit papers were counted. In the random gene set, most cancers had zero counts. The p-values were <0.01 in 7 cancers, showing a significant enrichment of cancer-related genes in the unique driver genes by WITER compared to the random gene sets. Noted that for less-studied cancers the significance tends to be less significant as well. This may be the reason why the significance varied from cancers to cancers. Nonetheless, the two analyses

convincingly suggested the unique significant genes by WITER were enriched with many functionally important genes for the corresponding cancers

Rescued significant genes in small samples by an alternative tool

We investigated the scenario in which the significant genes missed by a tool in a small sample can be rescued by another tool. We randomly drew 6 sub-samples of half size from the largest dataset, breast invasive carcinoma dataset, and detected cancer driver-genes by the three tools, MutSigCV, OncodriveFML and WITER. As shown in Table S8, MutSigCV detected 11 significant genes on average in sub-samples with half of the breast invasive carcinoma sample. In the same sub-samples, WITER rescued 4 genes on average, which were detected in the full sample by MutSigCV. Similarly, WITER rescued 3 genes on average, which were missed in half sample but were detected in the full sample by OncodriveFML. Using half of the breast invasive carcinoma sample, WITER detected similar number of significant genes (FDR<0.1) as it did in the full dataset. This was larger than that detected by MutSigCV and OncodriveFML in the full dataset. Moreover, MutSigCV and OncodriveFML rescued less than 1 gene on average which were missed in half sample and were detected in the full sample by WITER. This comparison shows WITER has enhanced power in small samples to detect driver genes that would be missed by alternative methods due to the small sample sizes.

Performance in 23 cancer datasets with relatively small samples

Another important advantage of WITER is its ability to detect cancer-driver genes in small samples with a usage of reference samples. We applied the approach to 23 cancers of small samples. We deliberately used two reference samples with very low and high background mutation rates to investigate how WITER is sensitive to the reference samples. The low background mutation rate cancer was the breast invasive carcinoma, and the high one was the melanoma. Four evaluations were carried out. First, the usage of the reference datasets substantially improved the distribution of p-values, compared to the analysis without reference samples. According to the QQ plots (Figure S1), the p-value distributions of the background genes (FDR>0.1) with reference samples were very close to the uniform distributions. In contrast, the p-values of the background genes without reference sample were weird and did not follow the uniform distribution. Second, WITER detected significant genes even for cancers with very small sample size (See the results in Table 1). Among the 21 cancers with one or more significant genes (FDR$\leqslant$0.1), 5 cancers had less than 50 subjects, e.g., B-cell lymphomas (n=26), small cell lung carcinoma (n=30), and cervical carcinoma (n=37). Third, it seemed the background mutation rate had a simple influence on the number of significant genes or statistical power. As expected, the low background mutation rate reference sample led to more significant genes than the high one. Moreover, we noted that almost all significant genes

according to the high background mutation rate reference sample were also significant according to the low background-mutation rate reference sample. Therefore, the false positive findings can be easily controlled by using a high background mutation rate reference sample in practice although this may increase the false negatives. Anyhow, the overlapping percentage of the significant genes were also generally high. For four cancers (acute myeloid leukemia, prostate adenocarcinoma, pancreatic adenocarcinoma and low-grade glioma) with at least 15 significant genes have 100%, 93%, 83% and 67% overlapped significant genes based on breast invasive carcinoma and melanoma reference samples respectively. Finally, WITER detected much more significant genes than ITER again (Table S3). WITER detected 5 to 23 more significant genes in 9 cancers than ITER regardless of different reference samples. These results suggest that the WITER is also powerful for datasets of small sample and the detected significant genes are not very sensitive to the reference datasets.

It should be also noted that extra significant genes according to the low background mutation rate reference sample are not necessarily false. For instance, MYCN was significant driver gene of neuroblastoma based on the breast invasive carcinoma reference (p=5.06E-8) but insignificant based on melanoma reference(p=0.0012). Actually, MYCN is a well-known driver gene of neuroblastoma (*19*). Anyhow, to reduce false positive results rigorously, we used the conservative results, i.e. significant genes according to the melanoma reference sample, for the subsequent analysis.

Analysis of explanatory variables for predicting background somatic mutations

We further investigated the contribution of the 6 explanatory variables to prediction of background mutations in the regression models (See coefficients and p-values in Table 2). The coding region length and number of mutant alleles at synonymous variants of a gene were the top two explanatory variables in terms of their statistical significance. Their p-values were extremely small in all the testing cancers. As expected, a gene having longer coding region and more synonymous variants(*20*) tended to have larger number of mutant alleles at non-synonymous variants and splicing variants in background genes, *n*. Interestingly, the significant p-values at both explanatory variables under the same model implied their independent contribution although they were also correlated (Spearman correlation≈0.4-0.5 in cancers). The replication time (measured in HeLa cells) was also positively related with *n* in most of cancers. This is consistent with the biological assumption that high replication leads to more somatic mutations (*21*) (*22*). The coefficients of constraint missense Z scores (*23*) were also positive in most of cancers, suggesting a gene with high *de novo* mutation potential in germline cells tends to have more somatic mutations as well. There were 2 explanatory variables, expression (averaged across 91 cell lines in the Cancer Cell Line Encylcopedia) and HiC (measured from HiC experiments in K562 cell), having negative coefficients. This is consistent with findings by Lawrence et al. (2013) in which genes with lower expression tended to have more somatic mutations (*22*). The negative coefficient of HiC implied that a gene with more densely packed DNA also tended to have less number of somatic mutations in cancer cells (*24*).

The zero-truncated negative binomial model outperforms other models

We also compared the performance of the zero-truncated negative binomial model with three alternative widely-used models for fitting the mutation counts. The three models are Poisson distribution model, negative binomial distribution model, zero-truncated Poisson distribution models respectively. It turned out the zero-truncated negative binomial model had the smallest Akaike information criterion (AIC) values in all the 11 cancers, suggesting it is the best fitting model for the counts of somatic mutations among the four models (Table 3). The zero-truncated Poisson distribution was the second best model although its averaged AIC values was still 3364 larger than that of the zero-truncated negative binomial distribution. For negative binomial distribution or Poisson distribution, the zero-truncated versions were much better than the original versions. For the negative binomial distribution, the averaged AIC value in 11 cancers of the zero-truncated ones was 3267 smaller than the un-truncated ones. The averaged AIC value of the zero-truncated Poisson distribution was 819 smaller than the un-truncated Poisson distribution. This implies that it is critical to exclude the influence of the zero-counts when constructing a regression model. A well-fitted model for mutation counts at background genes led to more accurate residues for evaluating the excess of mutations in a gene.

The numbers of significant genes are more related with the number of mutations than sample size

We also investigated factors influencing the number of significant genes among the 34 cancers by WITER, which implies factors affecting the power in real data. The number of significant genes was highly related with the number of somatic variants. In a linear prediction model, the number of somatic variants had a good prediction on the number of significant genes, with a coefficient of determination R^2, 0.36 (Figure S2). According to the prediction model, 57,000 somatic variants were needed to detect 30 significant genes. Because mutation rates are different in cancers, the corresponding sample sizes for such amount of mutations vary from cancers to cancers. Given the ratio of somatic variant number to sample size (Table S6), over 900 samples are needed to accumulate 57,000 variants in breast invasive carcinoma, kidney renal clear cell carcinoma, and ovarian serous cystadenocarcinoma. In contrast, for four cancers, less than 250 samples are sufficient, lung adenocarcinoma, melanoma, lung squamous cell carcinoma and bladder urothelial carcinoma. Compared to the number of somatic variants, sample size had less influence on the number of significant genes. In a linear regression model, coefficient of determination of sample size was only 0.17 (Figure S3). These results imply that the power of WITER may be determined by both sample size and somatic mutation rate.

The comprehensive landscape of driver-genes at 32 different cancers

WITER detected one or more significant genes in 32 cancers according to FDR<0.1. The total number of unique genes was 247 (See details in the Supplementary Excel File 1). Seventy-six genes occurred in two or more cancers. As expected, TP53 was the most common significant genes (in 27 cancer types), followed by PIK3CA, KRAS, FBXW7, NRAS, CTNNB1 and BRAF, each of which is

associated with 10 or more cancer types. Four cancers had over 40 significant genes, colon adenocarcinoma (COAD), uterine corpus endometrial carcinoma (UCEC), melanoma (MEL) and stomach adenocarcinoma (STAD). Most of the predicted driver genes are previously reported for the corresponding cancers. Interestingly, multiple PCDHA genes were significant in five cancers. Although the significance at multiple genes probably were probably caused by the highly overlapped coding regions, it at least suggested PCDHA gene family is associated with the cancers. PCDHA genes encode a family of cadherin-like cell surface proteins for cell-cell adhesion. There have been no studies showing its somatic mutations contribute to tumorigenesis. However, DNA hypermethylation on PCDHAs were detected in multiple cancers including prostate cancer(*25*) and small-cell lung cancer (*26*). In the *in-silico* validation in the NCBI PubMed, 21 cancers had 70% significant genes with hit papers (Summarized in Table S4).

Cancer clusters according to overlapped significant genes

According to multiple overlapped significant genes(Table S7), cancers were clustered into groups (Figure 4). Consistent with a recent study(*2*), some cancers in a group had either similar tissue or similar cell of origins. A group contained 4 blood cell related cancers, multiple myeloma(MM), diffuse large B-cell lymphoma(DLBCL), chronic lymphocytic leukemia (CLL), acute myeloid leukemia (LAML). DLBCL and LAML had a uniquely overlapped gene, EZH2, which had been widely studied for both diseases (*27*), (*28*). In another group, two nervous system related cancers, low grade glioma (LGG) and glioblastoma multiforme (GBM), had 7 overlapped significant genes and formed a sub-group. The two female cancers, uterine corpus endometrial carcinoma (UCEC) and breast invasive carcinoma (BRCA) had 14 overlapped significant genes and formed a sub-group. Moreover, there were also multiple sub-groups which did not look so related biologically. For example, in a group, lung squamous cell carcinoma(LUSC) and head and neck squamous cell carcinoma(HNSC) had 9 overlapped genes and formed a sub-group. There have been multiple studies suggesting that the two types of tumors had similar pathological features (*29*), (*30*). In another group, ovarian serous cystadenocarcinoma (OV) and bladder urothelial carcinoma (BLCA) had 8 overlapped genes. The prostate adenocarcinoma (PRAD) and pancreatic adenocarcinoma (PAAD) had even 15 overlapped genes and formed a sub-group. These high overlapping patterns imply pathogenic connection of different cancers although larger samples and more experiments are needed to investigate the possible mechanistic link.

Genes significant only in an individual cancer

Besides the overlapped genomic signatures for clustering cancers, it is also interesting to find out the unique significant genes of a cancer for the characterization. Among the 34 cancers, we found 23 cancers having one or more unique significant genes (See details in Table 4 and Table S5). Cancers with more significant genes tended to have more unique significant genes, implying their high

heterogeneity. For instance, for the 7 cancers with over 30 significant genes, each had over 10 unique significant genes (See the cancer names in Table 4). The numbers of hit papers in the NCBI PubMed database are summarized in Table 4 and the detailed PubMed IDs of hit papers are listed in Table S5.

Take colon adenocarcinoma for an example, it had 20 unique significant genes. Three genes (CXCR4, TCF7L2 and GNAS) had over 10 hit papers, suggesting that they are well-studied genes for colon cancer. For instance, there are at least 100 papers mentioning the relation of CXCR4 with colon adenocarcinoma. CXCR4 encodes a CXC chemokine receptor specific for stromal cell-derived factor-1. A very recent study suggested that the level of CXCR4 can determine the effects of ALDH1A3 on *in vitro* proliferation and invasion in colon cancer (*31*). Zheng et al suggested CXCR4 may play a key role in colorectal adenocarcinoma progression via the mediation of tumor cell adhesion (*32*). However, in literature, CXCR4 was also associated with lymphoplasmacytic lymphoma (*33*). However, in the 34 collected cancers, 31 cancers had totally insignificant p-values ($p>0.18$) expect for multiple myeloma ($p=0.0025$) and lung squamous cell carcinoma ($p=0.046$). These results suggest mutant CXCR4 may at least have relatively larger susceptibility to colon cancer than to most of other cancers. The gene TCF7L2 encodes a transcription factor 7-like2/transcription factor 4 that plays a key role in the Wnt/β-catenin signaling pathway (*34*) and was reported to be associated with colon adenocarcinoma (*35*). Similarly, except for a suggestively significant p-value in stomach cancer ($p=8.47E-4$), it had totally insignificant p-values in the other 32 cancers ($p>0.42$) although it was also reported to be associated with other cancers, such as breast cancer (*36*). The gene, GNAS, encodes guanine nucleotide binding protein (G Protein) and alpha stimulating activity polypeptide complex. In human protein atlas (HPA, http://www.proteinatlas.org/ENSG00000087460-GNAS/tissue) database, this gene has been categorized as a cancer-related gene. (See the PubMed IDs of related papers in Supplementary Table 4). In addition, 7 genes had one or several hit papers related to colon adenocarcinoma. For example, the unique significant gene PCBP1 ($p=6.17e-07$) had two hit papers. One paper suggested that PCBP1 was a molecular marker of Oxaliplatin (a standard treatment for colorectal adenocarcinoma) resistance in colorectal adenocarcinoma and a promising target for colorectal adenocarcinoma therapy (*37*). The other paper suggested that PCBP1 was responsible for stabilizing gastrin mRNA which was highly expressed in colorectal adenocarcinoma (*38*). PCBP1 represses autophagy-mediated cell survival and inhibition of tumor cell autophagy and the PCBP1 upregulation may be an effective therapeutic strategy to colon tumor with low PCBP1 expression (*39*). LIFR($p=4.90e-04$) had 4 hit papers and encodes protein that belongs to the type I cytokine receptor family. One of the studies used the meta-analysis with public cancer methylome data verified the colon cancer specificity of LIFR promoter methylation (*40*). Kim et al suggested that a missense mutation of LIFR rs3729740 may be useful as a biomarker for predicting whether metastatic colorectal adenocarcinoma patients were sensitive to relevant target regimens(*41*).

Five cancers had only one unique significant gene, chronic lymphocytic leukemia (CLL), cervical carcinoma (CESC), multiple myeloma (MM), rhabdoid tumor (RHAB) and thyroid carcinoma (THCA). The genes of two cancers (RHAB and CLL) had multiple hit papers. The unique significant gene of RHAB, SMARCB1, had even 100 hit papers. SMARCB1 encodes part of a complex that relieves repressive chromatin structures to allow the transcriptional machinery to access its targets effectively. It is a known tumor suppressor gene, and its mutations have been associated with malignant RHAB (*42*). After first discovered in RHAB, mutant SMARCB1was subsequently found in multiple cancers (e.g., renal medullary carcinoma) (*43*). Almost all the cancers with mutant SMARCB1 were characterized by the presence of 'rhabdoid cells' featuring large vesicular nuclei and large paranuclear filamentous cytoplasmic inclusion (*44*). The gene FGFR1 for Astrocytoma had 16 hit papers. FGFR1 encodes a fibroblast growth factor receptor. Studies suggested genomic alterations in FGFR1 can account for most pathogenic alterations in low-grade neuroepithelial tumors, including pilocytic astrocytomas (*45*). The unique significant gene of CLL, MYD88 (p=1.34E-09), had 40 hit papers. MYD88 encodes cytosolic adapter protein, an essential signal transducer in the interleukin-1 and Toll-like receptor signaling pathways (*46*). Except for a suggestively significant p-value in diffuse large B-cell lymphoma (DLBCL) (p=8.47E-4), it had totally insignificant p-values in the other 32 cancers (p>0.42). In fact, a lot of studies have suggested MYD88 as a driver gene for the two cancers (*47*), (*48*). The single unique genes of three other cancers had no hit papers by far in PubMed and are subject to validation in the future.

Pathway analysis of driver genes among multiple cancers

We performed pathway enrichment analysis by DAVID 6.7 (https://david.ncifcrf.gov/, Figure S7) among 8 cancers which had more than 10 significant driver genes. Two pathways, ErbB signaling pathway and Neurotrophin/Trk signaling, were enriched by the predicted driver-genes in most cancers. The ErbB signaling pathway was significant in all the 8 cancers. ErbB family of receptor tyrosine kinases (RTKs) are involved in intracellular signaling pathways to regulate diverse biologic responses, including proliferation, differentiation, cell motility and survival (*49*). Several well-known cancerous pathways, such as MAPK pathway and PI-3K pathway, are the downstream of the ErbB receptors (*50*). This result suggests that ErbB signaling pathway may have a common driver role in genesis of many tumors. The neurotrophin signaling pathway was significant in 6 cancers. The Neurotrophin/Trk signaling is regulated by connecting a variety of intracellar signaling cascades, which include MAPK pathway, PI-3K pathway, and PLC pathway, transmitting positive signals like enhanced survival and growth (*51*). Therefore, Neurotrophin/Trk signaling may be commonly involved in the development of multiple tumors. Another interesting pattern was that MEL (melanoma) and STAD (stomach adenocarcinoma) had many shared pathways although they only had 10 shared predicted driver genes. Quite a few of the shared pathways are related to immune

response, such as Chemokine signaling pathway, Fc epsilon RI signaling pathway, and Natural killer cell mediated cytotoxicity (*52*). Besides, another shared pathway, focal adhesive, plays essential roles in important biological processed including cell motility, proliferation, differentiation (*53*). The shared pathways provide interesting clues to common pathogenesis of cancers, which are subject to be investigated by more experiments.

Discussion

Accurately modeling counts of somatic mutations at background genes in small samples has long been a fundamental technical challenge in genomic characterization of cancer-driver genes (*2*, *11*). The proposed approach, WITER, has four unique advantages to address this issue. First, it has an advanced model, zero-truncated negative binomial regression, to fit the number of somatic mutations at background genes. In small samples, one often sees an inflation of zero mutation genes and overdispersion of mutation counts. Particularly, the inflated zero values make it difficult to fit the distribution of genomic counts by conventional distributions. The zero-truncated negative binomial distribution subtly circumvents both the zero inflation and the overdispersion issues. This is also the reason why zero-truncated negative binomial model always achieved the minimal AIC among four alternative models. Moreover, the deviance residuals in the regression model lead to statistically valid *p* values for rapid analysis. This solves the common problem of alternative methods that time-consuming simulation or permutation is needed to obtain valid p-values (Figure 3b and c) for hypothesis tests. Secondly, the iteration of the regression diminishes the influence of driver genes on the background mutation models. The progressive exclusion of likely driver genes results in a "purer" background mutation model, in the contribution of somatic mutations from driver genes will become less prominent. Third, the method also has an advantage of using an independent sample as reference to boost statistical power. When the sample size is small, there will be limited number of mutations and the resulting model for background genes will be unstable. This may be a common problem of existing cancer-driver gene tests. The usage of reference sample solves the problem of small samples. More importantly, we found the number of significant genes detected by WITER was generally not sensitive to the reference samples in most cancers (Table 1). Finally, it can impose prior weights to treat potential driver mutations and passenger mutations differently. Due to the iterative design, the resulting model will be fitted mainly by the passenger mutations.

The weighting scheme contributed much to the finding of extra significant genes. In the real data analysis of 34 cancers, the weighted version (WITER) always detected more significant genes and cancer-consensus genes than the unweighted version (ITER) and two other widely-used methods (MutSigCV and OncodriveFML). Note OncodriveFML also integrated functional impact scores (e.g., CADD). We also have demonstrated the WITER and ITER had similar and valid p-value distributions (Figure 3a), implying the imposed weights do not statistically invalidate the p-values. In

the present study, we simply used the predicted highly frequent (n>15) mutation potential in COSMIC database as prior weights with the assumption that highly frequent somatic mutations in cancer cells are more likely to be driver-mutations. Although it is hard to say the assumption works for every somatic mutation, the prior weights substantially enhanced the power in all cancers (Figure 3b and c). Theoretically, this property should be applicable to other types of prior weights. The more accurate weights in terms of the probability of being a cancer driver-mutation, the more improved power WITER will have.

We compared the proposed method with two widely-used and well-performed approaches(*11*), both of which belong to the unsupervised category. Another category of methods is the supervised approaches for detecting cancer driver genes. According to Tokheim et al (2016)(*11*), the supervised method 20/20plus outperformed the unsupervised methods (including MutSigCV and OncodriveFML) in terms of p-value distributions and the number of significant genes. However, a supervised strategy has learning bias toward the training samples in nature(*54*). If the training sample is not representative of all sample, the trained model will have low power for new samples. This would be particularly true for cancers because of their high genetic heterogeneity(*5*). Second, the 20/20plus also used many common genomic features of a gene (e.g., evolutionary conservation, predicted functional impact of variants, and gene interaction network connectivity) in the prediction (*11*). Although the usage of common genomic features will add information to prioritize common cancer-driver genes, it also runs the risk of diluting the information in local sample for identifying unique cancer driver genes, which would be important for a precision diagnosis and treatment of the tested cancers. Finally, the 20/20plus resorted time-consuming permutation procedure to generate p-values for statistical test. In contrast, the WITER and ITER are much faster than 20/20plus because it calculates p-values directly. Nevertheless, we also made additional comparisons between WITER and 20/20plus approach in the 11 cancers. In 4 cancer datasets, the p-value distribution of background ground genes produced by WITER were a little bit closer to uniform distribution than that by 20/20plus. (See QQ plots in Figure S4). WITER also detected more significant and cancer-consensus genes in 6 out of the 11 cancers (See details in Figure S5) and rescued more missed genes by other tools(See details in Table S8). These results suggest WITER may have slightly better performance than 20/20plus generally.

Applying the powerful approach, WITER, we generated a landscape of driver genes in 32 cancers. Although it would be more informative if samples were larger, the landscape has already showed some common and unique patterns of cancers. According to the overlapped significant genes, we saw many cancer subgroups, say UCEC and BRCA. Although the underlying mechanism of common driver genes between the different cancers remains elusive, highly overlapped genes in these subgroups unlikely occur by chance. Identifying the common causes of a subgroup cancers

may help find the pathogenic and metastatic relationship of the cancers and facilitate development of common treatments. On the other hand, the unique significant genes in the landscape have potential to characterize individual cancers. There are 24 cancers with one or more unique significant genes. Although some significant genes of a cancer may become no longer unique after sample size get increased, it may at least imply a relatively high susceptibility of the gene in a reported cancer, say SMARCB1 for rhabdoid tumor. Clearly, some of these unique significant genes will be very helpful for characterizing the tumor types, say MYD88 for lymphoma (*55*), which is important for precision diagnosis and treatment of the tumors.

Methods and Materials

The unified statistical framework

The unified statistical framework has a three-tier structure to examine driver genes by using somatic mutations in cancer cells (See the diagram in Figure 1). The first tier is an iterative zero-truncated negative-binomial regression which estimates expected non-synonymous and splicing mutation counts of a gene under background mutation model. The second tier is a weighting scheme to generate and integrate prior weights for prioritizing variants of high somatic mutation potential in cancer samples. The third tier is a schedule of adopting independent reference samples to stabilize the regression model in small samples. These methods work from different angles to improve the model of background mutations in passenger genes for a more powerful evaluation of driver genes.

Tier I: The iterative zero-truncated negative-binomial regression

We proposed an approach, ITER, to estimate somatic mutation counts of each gene on the genome. The difference between the observed mutation counts and the estimated counts of a gene measures the excess of somatic mutations at a gene in a cancer. The mutation types of interest are non-synonymous mutations and splicing mutations, which assumes a gene with significant excess of these types of mutations may confer selective growth advantage in cancer as a driver gene(*6*). Denote the mutant allele counts at a non-synonymous or a splicing variant *j* in a background gene *i* as $c_{i,j}$ and the total alleles of $m_i$variants in this gene is, y_i. We assume y_i follows a negative binomial (NB) distribution(*56*):

$$y_i = \sum_{j=1}^{m_i} c_{i,j} \sim NB(\mu_i, \theta),$$

where μ_i is the expected number of mutations and θ is a dispersion parameter. The probability mass function (PMF) is $f(x|\mu_i, \theta) = \frac{\Gamma(x+\theta)}{\Gamma(\theta)\cdot x!} \cdot \frac{\mu_i^x \theta^\theta}{(\mu_i^x+\theta)^{x+\theta}}$, where $\Gamma(\quad)$ is the gamma function and x=0,1,2, ….

As somatic mutation is a rare event, many genes have no somatic mutations in a sample of typical size. While the negative binomial model includes a probability mass at *x*=0, this is often much less than the number of genes with no somatic mutations in real data. This inflation of zeros makes it very difficult to fit the negative binomial distribution to the counts of somatic mutations. Therefore, we proposed to use a zero-truncated negative binomial (TNB) distribution to model the mutant allele counts of background gene *i*. The PMF of TNB is:

$$g(x|\mu_i,\theta) = \frac{f(x|\mu_i,\theta)}{1-f(0|\mu_i,\theta)}, \quad x = 1,2,\cdots .$$

Based on the TNB, we constructed a generalized linear regression model to estimate mutant allele of non-synonymous or splicing variants in a gene *i* by 6 covariables:

$\eta = \log(\mu_i) = \beta_0 + \beta_1 \times [x_1, \text{number of mutant alleles at synonymous variants}]$
$+\beta_2 \times [x_2, \text{length of unique coding region}]$
$+\beta_3 \times [x_3, \text{constraint score for de novo mutation potential}]$
$+\beta_4 \times [x_4, \text{expression in cell lines in the Cancer Cell Line Encylcopedia}]$
$+\beta_5 \times [x_5, \text{DNA replication timing in HeLa cells}]$
$+\beta_6 \times [x_6, \text{long} - \text{range chromatin interactions by HiC in K562 cell}],$
where $\log(\mu_i)$ is the link function and the $\beta_0, \cdots, \beta_6$ are the coefficients.

The number of mutant alleles at synonymous variants was counted in the local samples. The length of unique coding region was calculated from gene model defined by a reference gene model database, RefGene. The gene's constraint scores were from Samocha et al (2014) (*23*). The last three covariates were adopted from MutSigCV (*6*). The expression values were averaged expression across 91 cell lines in the Cancer Cell Line Encylcopedia (CCLE). The replication time of a gene was measured in HeLa cells, ranging from 100 (very early) to 1000 (very late). The chromatin state of a gene was measured from HiC experiments in K562 cells, ranging approximately from -50 (very closed) to +50 (very open). Because some covariables had missing values, a widely-used nonparametric missing value imputation method based on Random Forest, missForest, in a R package was used to impute missing values. This model is also open for other covariables as long as they can improve the prediction accuracy.

The parameters can be estimated by maximum likelihood with a quasi-Newton method. In our study, we called the maximum likelihood method in a R package countreg (https://r-forge.r-project.org/R/?group_id=522) to estimate the coefficients. The dispersion parameter θ is jointly estimated with the regression coefficients, $\beta_0, \cdots, \beta_6$. The model is fitted only for genes with non-zero counts.

With the established model, the logarithm of the expected mutation counts, $log(\hat{\mu}_i)$, at non-synonymous or splicing variants in a gene *i* can be calculated by:

$$log(\hat{\mu}_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \hat{\beta}_3 x_{i,3} + \hat{\beta}_4 x_{i,4} + \hat{\beta}_5 x_{i,5} + \hat{\beta}_6 x_{i,6}$$

where $\hat{\beta}_0, \cdots, \hat{\beta}_6$ are the fitted coefficients.

Given the fitted parameters, the probability of zero mutation gene *i* is: $p_{i,0} = \left(\frac{\hat{\theta}}{\hat{\theta}+\hat{\mu}_i}\right)^{\hat{\theta}}$.

Under zero-truncated model, the raw residual at gene *i* is:

$$r_i = y_i - \frac{\hat{\mu}_i}{1-p_{i,0}}.$$

The deviance residual of the model at gene *i* is:

$$e_i = sign(r_i) * \sqrt{2 * |ll(y_i|\mu_i^*, \hat{\theta}) - ll(y_i|\hat{\mu}_i, \hat{\theta})|}\ ,$$

where $sign(x)$ is the standard sign function, $ll(\mu, \theta)$ is the natural logarithm of the likelihood function of the zero-truncated negative binomial distribution,

$$ll(y_i|\mu, \theta) = \ln[g(y_i|\mu, \theta)].$$

and μ_i^* is the estimated mean given the observed count y_i and estimated $\hat{\theta}$ of a saturated model, obtained by solving the following equation:

$$y_i = \frac{\mu_i^*}{1 - (\frac{\hat{\theta}}{\hat{\theta}+\mu_i^*})^{\hat{\theta}}}$$

The deviance residuals are further standardized by the estimated mean $\hat{\mu}_e$ and standard deviation $\hat{\sigma}_e$ of the deviance residuals,

$$\acute{e}_i = \frac{e_i - \hat{\mu}_e}{\hat{\sigma}_e}.$$

In real data analysis [Figure 2, 3a and S1], we demonstrated the standard normal distribution can be used to approximate the corresponding p-values of the standardized deviance residual $\acute{e}_i$:

$$p_i = 1 - \Phi(\acute{e}_i),$$

where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution.

The assumption is that most majority of genes are background passenger genes. So, the ITER models the expected mutant alleles at somatic non-synonymous or splicing variants under null hypothesis. A large $\acute{e}_i$ means the observed number of somatic mutations is much larger than the expected number of mutations from the null hypothesis.

In order to reduce distortion of driver genes in the null-hypothesis regression model, we proposed to perform the regression under an iterative procedure:

Step 1: perform ITER to calculate p-vales for all genes.
Step 2: exclude significant genes by a cutoff corresponding to false discovery rate (FDR)$\leqslant$0.1.
Step 3: perform ITER to calculate p-vales for the retained genes.
Step 4: repeat Step 2 and 3 until there is no extra significant genes according to FDR$\leqslant$0.1.
The fitted ITER model in the last iteration is closest to the null hypothesis model and is then used to re-calculate deviance residuals and p-values of all genes (including the ones excluded during iteration).

Tier II: The weighted iterative zero-truncated negative-binomial regression

We further extend ITER to a WITER, which integrates prior weights at variants to boost power. Assume a variant *j* of gene *i* has a score, $s_{i,j} \in [0,1]$, implying its cancer driver potential. We bin $s_{i,j}$ as an integer score, $w_{i,i}$, by the ceiling function of $s_{i,j}/0.1$, i.e., $w_{i,j} = \lceil s_{i,j}/0.1 \rceil$. The integer score is then used as prior weights for the variant. The ITER is a special case of WITER when $w_{i,j} = 1$ for all variants. The weighted mutation allele count is:

$$\grave{y}_i = \sum_{j=1}^{m_i} c_{i,j} * w_{i,j}.$$

We now assume the weighted counts $\grave{y}_i$ follow a negative binomial (NB) distribution:

$$\grave{y}_i \sim NB(\grave{\mu}_i, \grave{\theta}),$$

where $\grave{\mu}_i$ is the expected weighted counts of mutations and $\grave{\theta}$ is a dispersion parameter of the NB distribution. After replacement of original counts (y_i) with weighted counts ($\grave{y}_i$), the same iterative zero-truncated negative-binomial regression procedure is carried out to test whether a gene has excess of weighted mutant alleles at non-synonymous or splicing variants.

In the present study, we built a model to predict high-frequency cancer driver potential to use as prior weights, in the form of a random forest (ensemble of 500 decision trees) trained by a large cancer somatic mutation database, COSMIC (V83). To avoid circular bias, all subjects (n=7,916) in our collected testing samples of the 34 cancers were excluded from COSMIC database. We collected 4,320 somatic mutation variants occurring over 15 times in primary cancer tissues to constitute a positive variant set in COSMIC(V83). A negative control variant set containing 258,846 somatic mutation variants was randomly sampled from the COSMIC as well. Each of the control variant occurred only once in primary cancer tissues. The predictors at each variant include 19 deleterious or conservation scores from the database dbNSFP v3.5 (*57*), (e.g., MutationTaster2 (*58*) and FATHMM (*59*), see the names of all tools in Supplementary Figure S6). The area under the receiver operating characteristic curve of the random forest model was 79%, which was much better than a multivariate logistic regression model and individual predictors (Figure S6). The random forest prediction scores,

s, ranged from 0 to 1. For variants without prediction scores due to missing values, the average score in the gene was used.

Tier III: ITER or WITER with reference samples in analysis for small cancer samples

When the number of somatic variants is small (say <28,000), it is difficult to build a stable regression model. However, note that the key idea of ITER and WITER is to build a prediction model for background passenger genes. When the mutation rates of passenger genes of two cancers are similar, it may be workable to integrate background genes of one cancer for the other cancer. We proposed a reference sample strategy for building a stable ITER or WITER model in small sample dataset. This is carried out into two stages.

- At the first stage, the above ITER or WITER is used to produce p-values for excess of somatic mutations at genes in a reference sample which have sufficient number of variants. Genes with p-values less than a very loose cutoff, say FDR 0.8, are excluded.
- At the second stage, the somatic mutations of retained genes are integrated with the local small sample and input into ITER or WITER to build a new regression model. The excess of somatic mutations and corresponding p-values at genes are calculated based on the new model.

Performance comparison with alternative tools

There have been multiple tools for detecting cancer-driver genes(*2*). According to an evaluation study(*11*), 2 tools (MutSigCV(*6*) and OncodriveFML(*9*)) and 1 tool (20/20plus(*11*)) had relatively better performance were chosen for comparisons in the present study. We compared their p-value distributions and number of significant genes with ITER and WITER. The MutSigCV was developed based on the background mutation rate while the OncodriveFML and 20/20+ were developed based on the ratio-metric. According to another classification, MutSigCV and OncodriveFML used an unsupervised strategy to predict cancer driver genes while 20/20plus used a supervised strategy. So, the unsupervised methods were chosen as the main targets for the performance comparison. MutSigCV is a powerful method for detecting genes mutated more often than expected by chance. It used a local regression model to estimate the expected mutant alleles by multiple genomic features of a gene in cancer cells including its expression level, replication time and 3D chromatin interaction capture (HiC). The online MutsigCV version (1.2) was used through the Broad website (http://genepattern.broadinstitute.org/gp/pages/index.jsf?lsid=MutSigCV). The recommended exome coverage file (https://genepattern.broadinstitute.org/gp/data/xchip/gpprod/shared_data/example_files/MutSigCV_1.3/exome_full192.coverage.txt) and gene covariates file (https://genepattern.broadinstitute.org/gp/data//xchip/gpprod/shared_data/example_files/MutSigCV_1.3/gene.covariates.txt) were used. OncodriveFML is a method designed to estimate the accumulated functional impact bias of tumor somatic mutations in both coding and non-coding genomic regions,

based on a simulation process. It used CADD scores to predict mutational impacts. The results were produced according to coding DNA sequence (CDS) regions. The genome reference and CDS files were downloaded from the website (https://bitbucket.org/bbglab/oncodrivefml) as the authors recommended. The default parameters of OncodriveFML were used to produce the results. The 20/20 plus is a machine-learning-based method integrating multiple features to predict driver genes, including sample mutational clustering, evolutionary conservation, predicted functional impact of variants, mutation consequence types, gene interaction network connectivity, etc. It used computer simulation to generate p-values for statistical significance. The 20/20plus v1.1.3 was downloaded and installed according to the website tutorial (http://2020plus.readthedocs.io/en/latest/index.html). The necessary files were also collected as the authors suggested (http://probabilistic2020.readthedocs.io/en/latest/tutorial.html#gene-bed-file and http://probabilistic2020.readthedocs.io/en/latest/tutorial.html#pre-computed-scores-optional). The data were analyzed by a pipeline to predict the cancer drivers under the default parameters. The 20/20 plus took 1.5 hours on average to analyze a dataset on a computer with 12 CPU (1.70GHz) cores and 64G RAM. The number of simulations was 10000.

Evaluation metrics in the performance comparison

We adopted four evaluation metrics for performance comparison, number of significant genes predicted, overlap with Cancer Gene Census (CGC) (*13*), observed vs. theoretical *p* values, and unique significant genes by a tool. The former 3 were also major metrics in an evaluation framework of cancer driver gene prediction method(*11*). The CGC dataset contained 699 manually curated cancer genes by Dec. 16, 2017. The departure of p-values from uniform distribution was measured by the mean absolute log2 fold change (MLFC) (*11*). The widely-used cutoff, Benjamini and Hochberg FDR 0.1, was used to report significant genes. A valid statistical test should lead to a MLFC close to zero in background (or passage) gene. We also used the distribution of Quantile-Quantile (QQ) plot to examine the distribution of p-values at the tail of small p-values.

Dataset of somatic mutations

We partitioned a curated full somatic mutation dataset by Tokheima and colleagues (*11*) into 34 sub-datasets according to the cancer types (See Table S6). Eleven sub-datasets contain 2,800 or more variants and were called relatively larger cancer dataset throughout the paper. Their sample sizes ranged from 142 to 1093. The ratios of variant number to sample size in the 11 cancers ranged from 50 to 327. The 23 other cancers with less number of variants are called relatively smaller cancer sets. The names, variant number and sample sizes of all cancers can be seen in Table S6.

In silico validation by PubMed search

We used PubMed search function to coarsely validate the relation between significant genes and a specific cancer. The underlying assumption is that the papers co-mentioning the gene and the cancer name in the title or abstract are likely to implicate the relatedness between the gene and the cancer. The more hit papers, the more likely the gene is related to the cancer. This is a quick *in-silico* validation although it may be rough. We employed the web application programming interfaces (APIs) of PubMed to execute the search. The search link was, http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term="DiseaseNames(inlcuding homonymies)"[tiab]%29+AND+"GeneSymbol (including RefSeq mRNA IDs)" [tiab]. The search responsed PubMed ID and relevant data of the papers, if available, in extensible markup language (XML).

Tool availability

The statistical framework has been implemented into a Java standalone application and is available at http://grass.cgs.hku.hk/limx/witer/.

Acknowledgements

This work was funded by National Natural Science Foundation of China (31771401), Science and Technology Program of Guangzhou (201803010116), Hong Kong Health and Medical Research Fund (02132236). Hong Kong General Research Fund 17124017, 17121414 and TRS T12C-714/14-R. We thank Tokheima and colleagues for sharing the high-quality curated somatic mutations in 32 cancers from multiple resources.

Contributions

M.L., J.K., L.J., Y.Z. and P.S. conceived the study. M.L. oversaw all aspects of the study. L.J. M.L. and J.K. developed the models. J.Z., S.D. and C.L. performed extensive computational analyses for performance comparison. Y.Z. and K.T. analyzed landscape of cancer-driver genes. M.L. and L.J. wrote the manuscript with input from J.Z. and S.D. All authors edited and approved of the final manuscript.

Competing interests

The authors declare no competing interests.

Figure legends

Figure 1: The diagram of the statistical framework for detecting cancer-driver genes

This framework includes three tiers denoted by the dashed rectangles. The first tier is an iterative zero-truncated negative-binomial regression (ITER). The second tier is a weighted iterative zero-truncated negative-binomial regression (WITER). The third tier is the ITER or WITER integrating reference samples. Tier 1 is part of tier 2. Tiers 1 and 2 are also part of tier 3. The unique components of each tier are marked by different colors. The major inputs are somatic mutations at non-synonymous, splicing and synonymous variants of different cancer patients. The outputs are *p*-values for excess of somatic mutations of individual gene in the cancer samples.

Figure 2: QQ plot of background gene p-values produced by 4 methods in 11 cancers

The p-values less than a cutoff according to FDR 0.1 were excluded. Among the 34 collected cancers, 11 cancers have 25,000 variants with somatic mutations in the data sets and were used for the comparison.

Figure 3: Performance comparison of different methods for detecting cancer driver mutation in 11 cancers

a: The MLFC of 4 methods; b: the number of significant genes; c: cancer consensus significant genes; d: the number of unique significant genes; e: the number of unique significant genes overlapped with the cancer consensus gene set.

The p-values less than a cutoff according to FDR 0.1 were excluded. Cancer name labels: BLCA: Bladder Urothelial Carcinoma; BRCA: Breast invasive carcinoma; COAD: Colon adenocarcinoma; UCEC: Uterine corpus endometrial carcinoma; HNSC: Head and Neck Squamous Cell Carcinoma; KIRC: Kidney renal clear cell carcinoma; LUAD: Lung Adenocarcinoma; LUSC: Lung Squamous Cell Carcinoma; MEL: Melanoma; OV: Ovarian serous cystadenocarcinoma; STAD: Stomach Adenocarcinoma.

Figure 4. Circos plot displays 247 significant genes in 32 cancers

Notes: The innermost ring denotes dendrogram of genes. The next ring contains significant genes (marked in red) and cancer clusters. It is followed by a ring of counts cancers in which the genes are significant. The outmost ring contains gene symbols. The bashed rectangles denote the main clusters of the cancers according to the overlapped genes. ALL: Acute lymphoblastic leukemia, AT: Astrocytoma, BLCA: Bladder Urothelial Carcinoma, BRCA: Breast invasive carcinoma, CARC: Carcinoid Cancer, CESC: Cervical Carcinoma, CLL: Chronic lymphocytic leukemia, COAD: Colon adenocarcinoma, DLBCL: Diffuse large B-cell lymphoma, ESCA: Esophageal carcinoma, GBM: Glioblastoma Multiforme, HNSC: Head and Neck Squamous Cell Carcinoma, KICH: Kidney chromophobe carcinoma, KIRC: Kidney renal clear cell carcinoma, KIRP: Kidney Papillary Cell Carcinoma, LAML: Acute myeloid leukemia, LB: B-cell lymphomas, LGG: Low Grade Glioma, LIHC: Liver Hepatocellular carcinoma, LUAD: Lung Adenocarcinoma, LUSC: Lung Squamous Cell Carcinoma, LUSE: Small cell lung carcinoma, MED: Medulloblastoma, MEL: Melanoma, MM: Multiple myeloma, NB: Neuroblastoma, OV: Ovarian serous cystadenocarcinoma, PAAD:

Pancreatic Adenocarcinoma, PRAD: Prostate Adenocarcinoma, RHAB: Rhabdoid tumor, STAD: Stomach Adenocarcinoma, STS: Soft Tissue Sarcoma, THCA: Thyroid Carcinoma, UCEC: Uterine corpus endometrial carcinoma

Figure 1

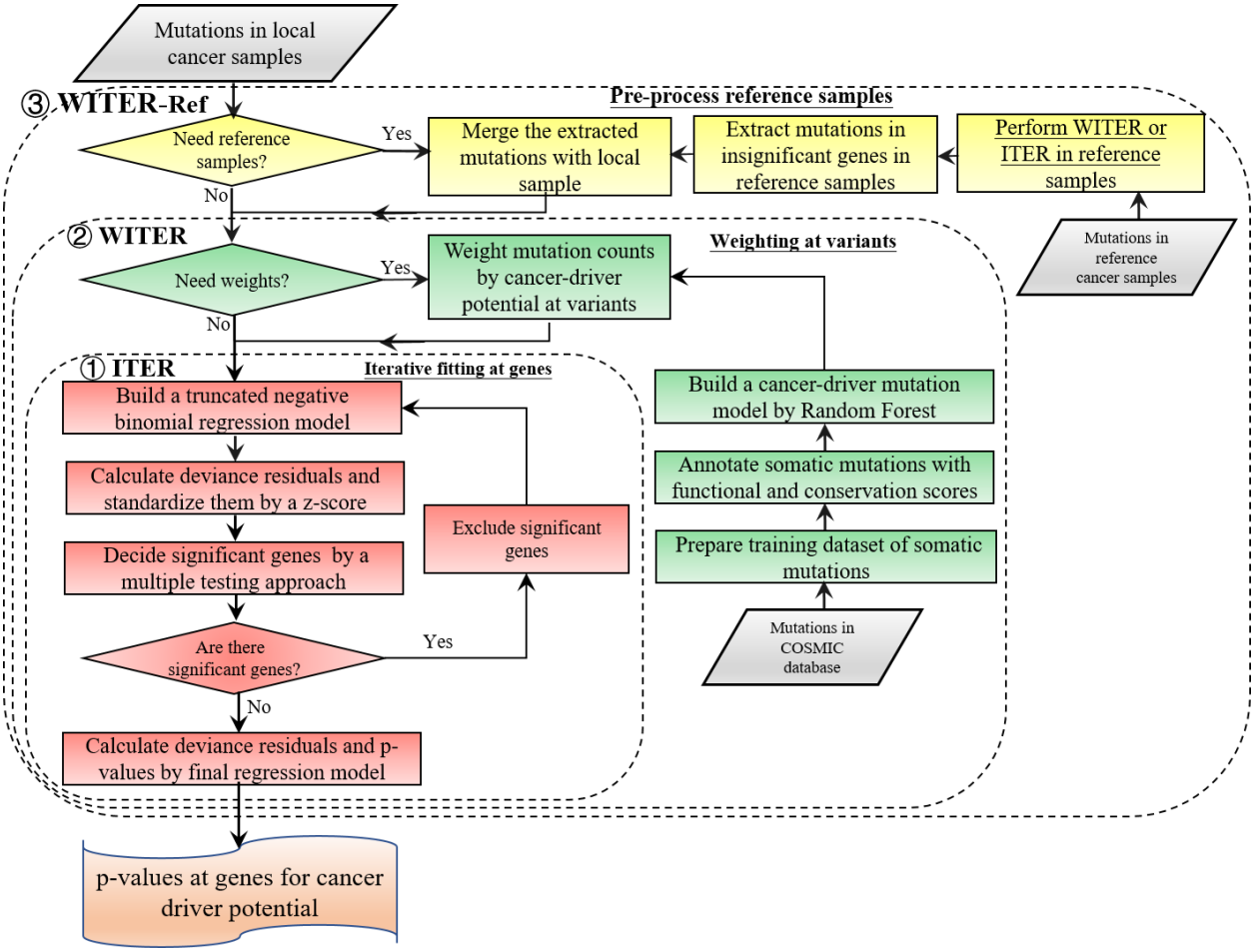


Figure 2

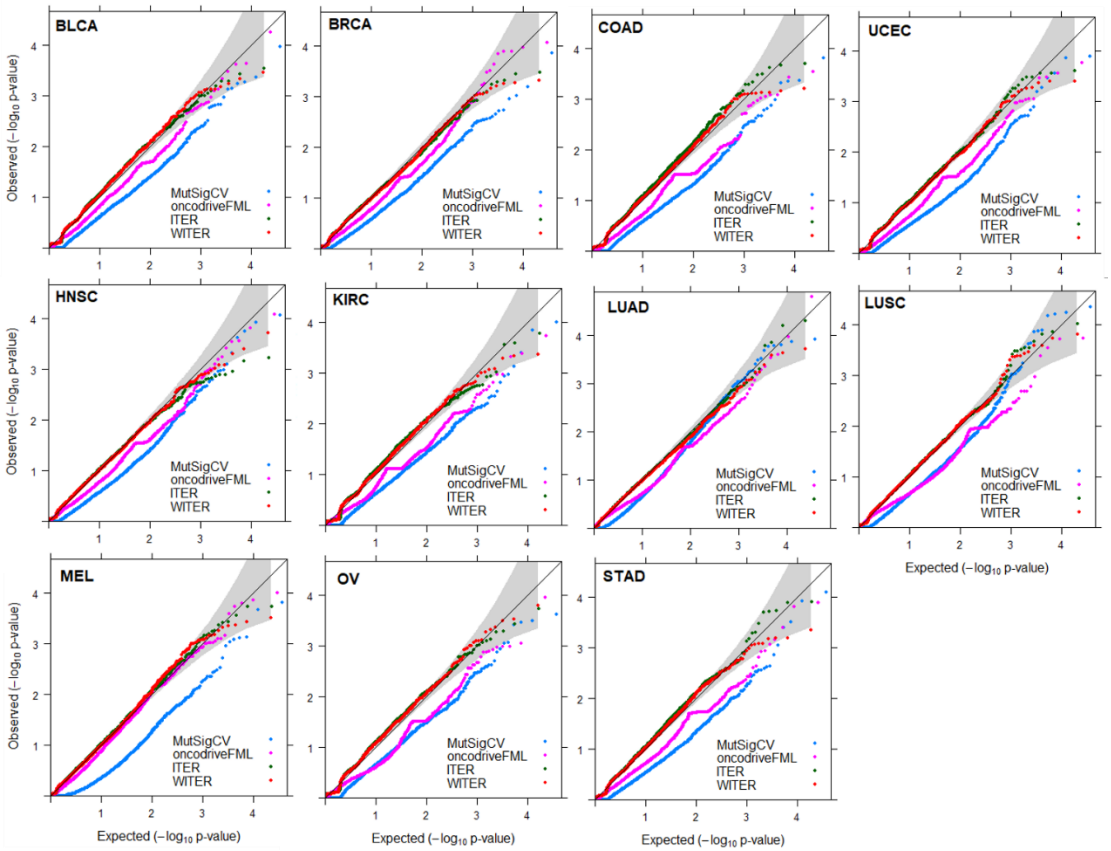
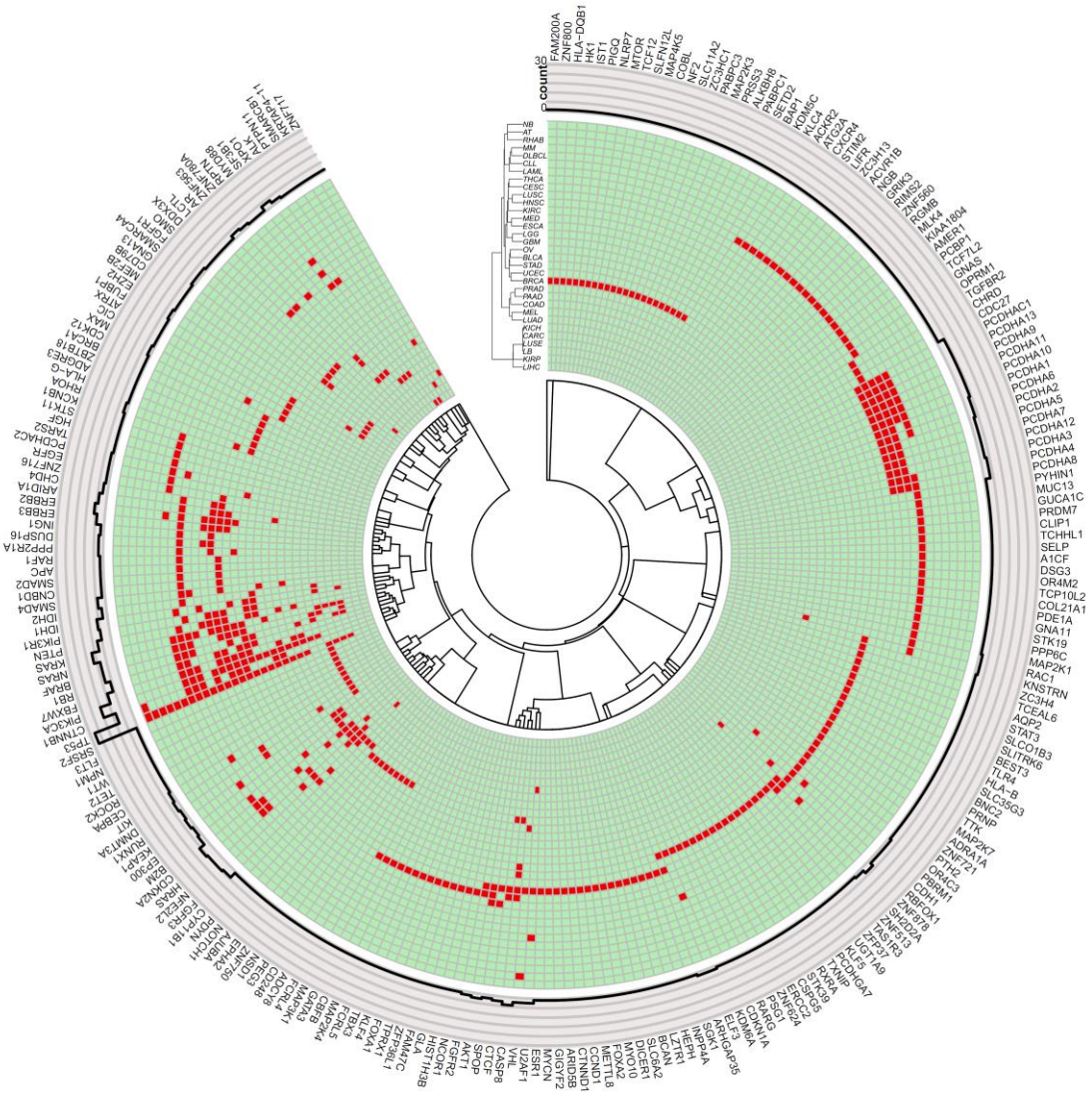


Figure 3

Figure 4



Tables:

Table 1 Significant genes by WITER in 23 cancers with small samples

Cancer	Sample Size	Sig. Genes with Melanoma Ref.	Sig. Genes with Breast invasive carcinoma Ref	Overlapped Genes
Prostate Adenocarcinoma	420	25	27	25
Glioblastoma Multiforme	365	10	13	9
Neuroblastoma	351	2	3	2
Medulloblastoma	331	8	8	7
Thyroid Carcinoma	325	8	10	8
Pancreatic Adenocarcinoma	233	21	23	19
Low Grade Glioma	227	10	15	10
Chronic lymphocytic leukemia	223	5	8	5
Multiple myeloma	205	7	11	7
Acute myeloid leukemia	196	17	17	17
Esophageal carcinoma	160	6	7	6

Liver Hepatocellular Carcinoma	150	2	2	2
Kidney Papillary Cell Carcinoma	111	2	4	2
Kidney chromophobe carcinoma	65	1	1	1
Diffuse large B-cell lymphoma	56	7	14	7
Acute lymphoblastic leukemia	55	0	0	0
Carcinoid Cancer	54	1	2	1
Astrocytoma	42	1	1	1
Cervical Carcinoma	37	3	5	2
Rhabdoid tumor	32	1	1	1
Small cell lung carcinoma	30	1	1	1
B-cell lymphomas	26	1	3	1
Soft Tissue Sarcoma	15	0	0	0

Note: Sig. Genes: Significant genes according to their p-values (FDR<0.1). Ref.: reference sample.

Table 2. The significance level of covariates in 11 cancer datasets

	#Minor alleles at synonymous variants	Coding Length	Expression	Replication Time	HiC	Constraint Score
Bladder Urothelial Carcinoma	0.072(1.40E-09)	0.245(2.41E-298)	-0.05(1.56E-03)	1.61E-07(9.98E-01)	-2.46E-03(5.74E-05)	0.026(1.39E-05)
Breast invasive carcinoma	0.071(2.67E-23)	0.244(0)	-0.078(3.27E-12)	2.05E-04(2.71E-04)	-1.56E-03(4.02E-04)	0.025(1.25E-08)
Colon adenocarcinoma	0.176(6.46E-56)	0.185(5.91E-180)	-0.12(5.36E-14)	4.67E-04(1.05E-08)	-2.62E-03(4.26E-05)	0.052(1.50E-16)
Uterine corpus endometrial carcinoma	0.106(5.02E-30)	0.211(0)	-0.11(3.48E-15)	8.59E-05(2.22E-01)	6.63E-07(9.99E-01)	0.054(2.17E-24)
Head and Neck Squamous Cell Carcinoma	0.104(2.87E-69)	0.23(0)	-0.109(6.16E-25)	4.84E-04(3.50E-19)	-2.83E-03(2.74E-11)	0.029(4.95E-11)
Kidney renal clear cell carcinoma	0.088(2.13E-09)	0.214(1.53E-252)	-0.09(8.64E-08)	2.76E-05(7.41E-01)	6.60E-05(9.21E-01)	0.02(1.27E-03)
Lung Adenocarcinoma	0.135(0)	0.205(0)	-0.092(1.38E-25)	6.26E-04(5.72E-48)	-3.90E-03(6.46E-29)	-0.003(4.43E-01)
Lung Squamous Cell Carcinoma	0.149(5.55E-116)	0.209(0)	-0.106(4.10E-21)	6.12E-04(4.72E-27)	-4.15E-03(4.25E-20)	-0.007(1.57E-01)
Melanoma	0.139(0)	0.179(1.99E-254)	-0.067(1.12E-10)	4.76E-04(2.83E-21)	-2.91E-03(1.51E-12)	-0.028(3.76E-10)
Ovarian serous cystadenocarcinoma	0.121(2.16E-16)	0.213(1.84E-221)	-0.089(6.49E-08)	3.13E-04(1.96E-04)	-7.26E-04(2.68E-01)	0.026(3.76E-05)
Stomach Adenocarcinoma	0.142(4.97E-68)	0.215(1.25E-294)	-0.128(3.02E-23)	4.97E-04(1.05E-13)	-3.29E-03(2.71E-10)	0.036(6.19E-12)

Note: The coefficients are produced by WITER using R package “countreg”.The values in the brackets are p-values for the significance of the coefficients calculated by Wald test.

Table 3: Akaike information criterion (AIC) of the various regression models

	Poisson	Zt- Poisson	Negative Binomial	Zt-Negative Binomial

Bladder Urothelial Carcinoma	32574.08	30840.35	33492.05	29827.37
Breast invasive carcinoma	48124.84	47794.76	47552.87	44544.85
Colon adenocarcinoma	29239.13	27224.3	29842.38	26295.96
Uterine corpus endometrial carcinoma	35991.2	34559.58	36589.3	32903.45
Head and Neck Squamous Cell Carcinoma	50038.23	49927.61	49332.66	46188.44
Kidney renal clear cell carcinoma	27669.52	25839.32	29178.44	25416.79
Lung Adenocarcinoma	76294.46	77151.63	67903.28	65669.84
Lung Squamous Cell Carcinoma	44860.29	44692.98	45806.77	42342.03
Melanoma	67107.65	68103.84	60022.15	58037.14
Ovarian serous cystadenocarcinoma	28326.67	26517.45	29772.79	25979.07
Stomach Adenocarcinoma	37883.63	36445.78	38534.62	34881.70

Note: The glm() function in R was used to fit the generalized linear model (GLM) of Poisson distribution. The glm.nb() function in the R package of MASS was used to fit the GLM of Negative Binomial distribution. The other two models were fitted by the R package of countreg.

Table 4, **Genes significant only in an individual cancer**

Cancer	Total	[a]Uni.	Genes
KIRC	34	23	BAP1[9.49e-28(56)], KDM5C[3.86e-15(13)], SETD2[6.89e-14(43)], PABPC1[6.48e-13(0)], ALKBH8[1.18e-08(0)], PRSS3[1.47e-08(0)], MAP2K3[1.26e-07(0)], PABPC3[3.79e-07(0)], ZC3HC1[6.16e-07(0)], SLC11A2[1.36e-06(0)], NF2[4.35e-06(3)], COBL[5.70e-06(0)], MAP4K5[1.48e-05(0)], SLFN12L[2.13e-05(0)], TCF12[3.88e-05(0)], MTOR[4.32e-05(65)], NLRP7[4.52e-05(0)], PIGQ[7.30e-05(0)], IST1[1.14e-04(0)], HK1[2.66e-04(0)], HLA-DQB1[3.20e-04(0)], ZNF800[3.47e-04(0)], FAM200A[3.57e-04(0)]
COAD	57	20	GNAS[1.23e-07(11)], OPRM1[1.95e-07(0)], TCF7L2[5.64e-07(65)], PCBP1[7.12e-07(2)], AMER1[1.52e-06(5)], KIAA1804[6.64e-06(2)], MLK4[9.89e-06(1)], RGMB[2.30e-05(2)], ZNF560[8.45e-05(1)], RIMS2[1.36e-04(0)], GRIK3[1.77e-04(0)], NGB[2.30e-04(1)], ACVR1B[3.01e-04(1)], ZC3H13[3.12e-04(1)], LIFR[4.14e-04(4)], STIM2[4.77e-04(3)], CXCR4[5.21e-04(100)], ATG2A[5.78e-04(0)], ACKR2[5.86e-04(0)], KLC4[5.96e-04(0)]
STAD	45	18	PTH2[8.78e-15(0)], OR4C3[3.63e-08(0)], ZNF721[7.39e-08(0)], ADRA1A[3.69e-07(0)], MAP2K7[4.33e-07(0)], TTK[4.89e-07(0)], PRNP[2.85e-06(0)], BNC2[5.56e-06(0)], SLC35G3[7.86e-05(0)], HLA-B[8.88e-05(0)], TLR4[1.15e-04(4)], BEST3[1.36e-04(0)], SLITRK6[1.62e-04(0)], SLCO1B3[1.63e-04(0)], STAT3[1.75e-04(6)], AQP2[2.61e-04(0)], TCEAL6[2.80e-04(0)], ZC3H4[3.03e-04(0)]
MEL	41	17	PPP6C[2.10e-21(10)], MAP2K1[3.01e-16(14)], STK19[1.56e-15(5)], GNA11[1.52e-10(100)], PDE1A[8.29e-07(1)], COL21A1[2.35e-06(1)], TCP10L2[8.27e-06(0)], OR4M2[1.51e-05(0)], DSG3[2.32e-05(1)], A1CF[3.67e-05(0)], SELP[3.72e-05(1)], TCHHL1[4.08e-05(0)], CLIP1[9.97e-05(0)], PRDM7[1.54e-04(0)], GUCA1C[1.71e-04(0)], MUC13[1.81e-04(0)], PYHIN1[2.87e-04(1)]
BLCA	35	16	CDKN1A[1.11e-12(2)], KDM6A[3.91e-08(3)], RARG[8.89e-08(0)], PSG1[1.63e-07(0)], ZNF624[1.65e-07(0)], ERCC2[9.69e-07(3)], CSPG5[3.91e-06(0)], STK39[1.91e-05(0)], RXRA[2.17e-05(1)], TXNIP[5.80e-05(0)], PCDHGA7[9.39e-05(0)], KLF5[1.00e-04(1)], UGT1A9[1.08e-04(0)], ZFP37[1.42e-04(0)], TAS1R3[2.09e-04(0)], ZNF513[2.80e-04(0)]
UCEC	42	16	MYCN[6.07e-11(3)], ESR1[5.10e-09(35)], GIGYF2[8.62e-09(0)], ARID5B[2.15e-08(2)], CTNND1[1.86e-06(0)], CCND1[7.15e-06(20)], METTL8[1.05e-05(0)], FOXA2[3.27e-05(6)], MYO10[4.00e-05(1)], DICER1[7.58e-05(7)], SLC6A2[8.32e-05(1)], BCAN[1.16e-04(0)], LZTR1[1.24e-04(0)], HEPH[1.35e-04(0)], INPP4A[1.79e-04(0)], SGK1[3.74e-04(2)]
BRCA	33	14	GATA3[5.08e-41(100)], MAP3K1[7.40e-23(78)], CBFB[7.99e-13(7)], MAP2K4[2.53e-10(16)], FCRL5[1.95e-06(0)], TBX3[3.17e-06(32)], KLF4[1.07e-05(87)], FOXA1[1.09e-05(100)], TPRX1[2.41e-05(0)], ZFP36L1[5.67e-05(2)], FAM47C[1.11e-04(0)], GLA[1.40e-04(35)], HIST1H3B[1.41e-04(0)], NCOR1[2.17e-04(22)]
HNSC	24	9	AJUBA[8.10e-08(4)], NOTCH1[2.95e-07(41)], EPHA2[6.94e-07(4)], ZNF750[4.76e-06(1)], NSD1[3.03e-05(4)], PEG3[5.65e-05(0)], CD248[5.87e-05(0)], ADCY8[1.52e-04(0)], FCRL4[1.68e-04(0)]
LAML	17	7	FLT3[5.79e-61(100)], SRSF2[1.30e-12(45)], NPM1[1.77e-12(100)], WT1[1.14e-09(100)], TET2[1.32e-06(100)], ROCK2[1.39e-05(0)], CEBPA[1.60e-05(100)]

OV	17	5	BRCA1[3.35e-07(100)], CDK12[7.56e-06(10)], ZBTB18[2.41e-05(0)], ADGRE3[5.11e-05(0)], HLA-G[1.39e-04(14)]
LUAD	29	4	STK11[1.35e-14(88)], KCNB1[7.22e-05(0)], HGF[8.67e-05(100)], TARS2[1.46e-04(0)]
DLBCL	7	3	CD79B[5.33e-10(42)], GNA13[2.74e-05(13)], MEF2B[4.51e-05(10)]
LGG	10	3	ATRX[1.51e-10(11)], FUBP1[3.39e-06(2)], CIC[4.10e-05(1)]
PAAD	21	3	CHRD[1.20e-06(0)], CDC27[4.28e-05(1)], TGFBR2[1.44e-04(15)]
LUSC	13	2	PDYN[1.78e-05(0)], CYP11B1[3.67e-05(1)]
MED	8	2	SMO[2.38e-15(77)], DDX3X[4.64e-14(10)]
NB	2	2	ALK[4.07e-18(100)], PTPN11[7.12e-08(14)]
PRAD	25	2	LCTL[2.85e-06(0)], AR[5.33e-05(100)]
CLL	5	1	MYD88[1.34e-09(40)]
CESC	3	1	KRTAP4-11[1.45e-05(0)]
MM	7	1	ZNF717[9.14e-07(0)]
RHAB	1	1	SMARCB1[2.52e-07(100)]
THCA	8	1	RPTN[2.17e-05(0)]
ALL	0	0	-
AT	1	0	-
CARC	1	0	-
ESCA	6	0	-
GBM	10	0	-
KICH	1	0	-
KIRP	2	0	-
LIHC	2	0	-
LUSE	1	0	-
LB	1	0	-
STS	0	0	-

Note: The values in square brackets are p-values by WITER. The significant genes are determined according to the p-values (FDR<0.1). The number in the brackets are the number of papers co-mentioning the disease name and gene symbol according to search API in PubMed database, http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term="DiseaseNames(inlcuding homonymies)"[tiab]%29+AND+"GeneSymbol (including RefSeq mRNA IDs)" [tiab]. For genes with over 100 papers, only the most recent 100 papers are shown. The PubMed ID of the papers are in supplementary Table 3. "-" denotes no unique significant genes. "a": number of total significant genes. "b": number of unique significant genes. ALL: Acute lymphoblastic leukemia, AT: Astrocytoma, BLCA: Bladder Urothelial Carcinoma, BRCA: Breast invasive carcinoma, CARC: Carcinoid Cancer, CESC: Cervical Carcinoma, CLL: Chronic lymphocytic leukemia, COAD: Colon adenocarcinoma, DLBCL: Diffuse large B-cell lymphoma, ESCA: Esophageal carcinoma, GBM: Glioblastoma Multiforme, HNSC: Head and Neck Squamous Cell Carcinoma, KICH: Kidney chromophobe carcinoma, KIRC: Kidney renal clear cell carcinoma, KIRP: Kidney Papillary Cell Carcinoma, LAML: Acute myeloid leukemia, LB: B-cell lymphomas, LGG: Low Grade Glioma, LIHC: Liver Hepatocellular carcinoma, LUAD: Lung Adenocarcinoma, LUSC: Lung Squamous Cell Carcinoma, LUSE: Small cell lung carcinoma, MED: Medulloblastoma, MEL: Melanoma, MM: Multiple myeloma , NB: Neuroblastoma, OV: Ovarian serous cystadenocarcinoma, PAAD: Pancreatic Adenocarcinoma, PRAD: Prostate Adenocarcinoma, RHAB: Rhabdoid tumor, STAD: Stomach Adenocarcinoma, STS: Soft Tissue Sarcoma, THCA: Thyroid Carcinoma, UCEC: Uterine corpus endometrial carcinoma

References

1. S. F. Bunting, A. Nussenzweig, End-joining, translocations and cancer. *Nat Rev Cancer* **13**, 443-454 (2013).
2. M. H. Bailey, C. Tokheim, E. Porta-Pardo, S. Sengupta, D. Bertrand, A. Weerasinghe, A. Colaprico, M. C. Wendl, J. Kim, B. Reardon, P. K. Ng, K. J. Jeong, S. Cao, Z. Wang, J. Gao, Q. Gao, F. Wang, E. M. Liu, L. Mularoni, C. Rubio-Perez, N. Nagarajan, I. Cortes-Ciriano, D. C. Zhou, W. W. Liang, J. M. Hess, V. D. Yellapantula, D. Tamborero, A. Gonzalez-Perez, C. Suphavilai, J. Y. Ko, E. Khurana, P. J. Park, E. M. Van Allen, H. Liang, M. C. W. Group, N. Cancer Genome Atlas Research, M. S. Lawrence, A. Godzik, N. Lopez-Bigas, J. Stuart, D. Wheeler, G. Getz, K. Chen, A. J. Lazar, G. B. Mills, R. Karchin, L. Ding, Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385 e318 (2018).
3. B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, Jr., K. W. Kinzler, Cancer genome landscapes. *Science* **339**, 1546-1558 (2013).
4. S. Kim, New and emerging factors in tumorigenesis: an overview. *Cancer Manag Res* **7**, 225-239 (2015).
5. K. Cyll, E. Ersvaer, L. Vlatkovic, M. Pradhan, W. Kildal, M. Avranden Kjaer, A. Kleppe, T. S. Hveem, B. Carlsen, S. Gill, S. Loffeler, E. S. Haug, H. Waehre, P. Sooriakumaran, H. E.

Danielsen, Tumour heterogeneity poses a significant challenge to cancer biomarker research. *Br J Cancer* **117**, 367-375 (2017).

6. M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, A. Kiezun, P. S. Hammerman, A. McKenna, Y. Drier, L. Zou, A. H. Ramos, T. J. Pugh, N. Stransky, E. Helman, J. Kim, C. Sougnez, L. Ambrogio, E. Nickerson, E. Shefler, M. L. Cortes, D. Auclair, G. Saksena, D. Voet, M. Noble, D. DiCara, P. Lin, L. Lichtenstein, D. I. Heiman, T. Fennell, M. Imielinski, B. Hernandez, E. Hodis, S. Baca, A. M. Dulak, J. Lohr, D. A. Landau, C. J. Wu, J. Melendez-Zajgla, A. Hidalgo-Miranda, A. Koren, S. A. McCarroll, J. Mora, B. Crompton, R. Onofrio, M. Parkin, W. Winckler, K. Ardlie, S. B. Gabriel, C. W. M. Roberts, J. A. Biegel, K. Stegmaier, A. J. Bass, L. A. Garraway, M. Meyerson, T. R. Golub, D. A. Gordenin, S. Sunyaev, E. S. Lander, G. Getz, Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218 (2013).
7. N. D. Dees, Q. Zhang, C. Kandoth, M. C. Wendl, W. Schierding, D. C. Koboldt, T. B. Mooney, M. B. Callaway, D. Dooling, E. R. Mardis, R. K. Wilson, L. Ding, MuSiC: identifying mutational significance in cancer genomes. *Genome Res* **22**, 1589-1598 (2012).
8. A. Gonzalez-Perez, N. Lopez-Bigas, Functional impact bias reveals cancer drivers. *Nucleic Acids Res* **40**, e169 (2012).
9. L. Mularoni, R. Sabarinathan, J. Deu-Pons, A. Gonzalez-Perez, N. Lopez-Bigas, OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol* **17**, 128 (2016).
10. D. Tamborero, A. Gonzalez-Perez, N. Lopez-Bigas, OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238-2244 (2013).
11. C. J. Tokheim, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, R. Karchin, Evaluating the evaluation of cancer driver genes. *Proc Natl Acad Sci U S A* **113**, 14330-14335 (2016).
12. M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, J. Shendure, A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-315 (2014).
13. P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, M. R. Stratton, A census of human cancer genes. *Nat Rev Cancer* **4**, 177-183 (2004).
14. T. N. Beck, E. A. Golemis, Genomic insights into head and neck cancer. *Cancers Head Neck* **1**, (2016).
15. C. R. Pickering, J. H. Zhou, J. J. Lee, J. A. Drummond, S. A. Peng, R. E. Saade, K. Y. Tsai, J. L. Curry, M. T. Tetzlaff, S. Y. Lai, J. Yu, D. M. Muzny, H. Doddapaneni, E. Shinbrot, K. R. Covington, J. Zhang, S. Seth, C. Caulin, G. L. Clayman, A. K. El-Naggar, R. A. Gibbs, R. S. Weber, J. N. Myers, D. A. Wheeler, M. J. Frederick, Mutational landscape of aggressive cutaneous squamous cell carcinoma. *Clin Cancer Res* **20**, 6582-6592 (2014).
16. D. R. Fels Elliott, J. Perner, X. Li, M. F. Symmons, B. Verstak, M. Eldridge, L. Bower, M. O'Donovan, N. J. Gay, O. Consortium, R. C. Fitzgerald, Impact of mutations in Toll-like receptor pathway genes on esophageal carcinogenesis. *PLoS Genet* **13**, e1006808 (2017).

17. E. Garza-Gonzalez, F. J. Bosques-Padilla, S. I. Mendoza-Ibarra, J. P. Flores-Gutierrez, H. J. Maldonado-Garza, G. I. Perez-Perez, Assessment of the toll-like receptor 4 Asp299Gly, Thr399Ile and interleukin-8 -251 polymorphisms in the risk for the development of distal gastric cancer. *BMC Cancer* **7**, 70 (2007).
18. W. M. Grady, S. D. Markowitz, Genetic and epigenetic alterations in colon cancer. *Annu Rev Genomics Hum Genet* **3**, 101-128 (2002).
19. J. I. Fletcher, D. S. Ziegler, T. N. Trahair, G. M. Marshall, M. Haber, M. D. Norris, Too many targets, not enough patients: rethinking neuroblastoma clinical trials. *Nat Rev Cancer* **18**, 389-400 (2018).
20. P. Evans, S. Avey, Y. Kong, M. Krauthammer, Adjusting for Background Mutation Frequency Biases Improves the Identification of Cancer Driver Genes. *IEEE Transactions on NanoBioscience* **12**, 5 (2013).
21. N. Donley, M. J. Thayer, DNA replication timing, genome stability and cancer: late and/or delayed DNA replication timing is associated with increased genomic instability. *Semin Cancer Biol* **23**, 80-89 (2013).
22. J. A. Stamatoyannopoulos, I. Adzhubei, R. E. Thurman, G. V. Kryukov, S. M. Mirkin, S. R. Sunyaev, Human mutation rate associated with DNA replication timing. *Nat Genet* **41**, 393-395 (2009).
23. K. E. Samocha, E. B. Robinson, S. J. Sanders, C. Stevens, A. Sabo, L. M. McGrath, J. A. Kosmicki, K. Rehnstrom, S. Mallick, A. Kirby, D. P. Wall, D. G. MacArthur, S. B. Gabriel, M. DePristo, S. M. Purcell, A. Palotie, E. Boerwinkle, J. D. Buxbaum, E. H. Cook, Jr., R. A. Gibbs, G. D. Schellenberg, J. S. Sutcliffe, B. Devlin, K. Roeder, B. M. Neale, M. J. Daly, A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-950 (2014).
24. K. D. Makova, R. C. Hardison, The effects of chromatin organization on variation in mutation rates in the genome. *Nat Rev Genet* **16**, 213-223 (2015).
25. M. B. Cook, Z. Wang, E. D. Yeboah, Y. Tettey, R. B. Biritwum, A. A. Adjei, E. Tay, A. Truelove, S. Niwa, C. C. Chung, A. P. Chokkalingam, L. W. Chu, M. Yeager, A. Hutchinson, K. Yu, K. A. Rand, C. A. Haiman, G. C. African Ancestry Prostate Cancer, R. N. Hoover, A. W. Hsing, S. J. Chanock, A genome-wide association study of prostate cancer in West African men. *Hum Genet* **133**, 509-521 (2014).
26. H. Ono, N. Motoi, H. Nagano, E. Miyauchi, M. Ushijima, M. Matsuura, S. Okumura, M. Nishio, T. Hirose, N. Inase, Y. Ishikawa, Long noncoding RNA HOTAIR is relevant to cellular proliferation, invasiveness, and clinical relapse in small-cell lung cancer. *Cancer Med* **3**, 632-642 (2014).
27. R. C. Lindsley, B. G. Mar, E. Mazzola, P. V. Grauman, S. Shareef, S. L. Allen, A. Pigneux, M. Wetzler, R. K. Stuart, H. P. Erba, L. E. Damon, B. L. Powell, N. Lindeman, D. P. Steensma, M. Wadleigh, D. J. DeAngelo, D. Neuberg, R. M. Stone, B. L. Ebert, Acute myeloid leukemia ontogeny is defined by distinct somatic mutations. *Blood* **125**, 1367-1376 (2015).
28. Y. Liu, K. Yu, M. Li, K. Zeng, J. Wei, X. Li, Y. Liu, D. Zhao, L. Fan, Z. Yu, Y. Wang, Z. Li, W. Zhang, Q. Bai, Q. Yan, Y. Guo, Z. Wang, S. Guo, EZH2 overexpression in primary

gastrointestinal diffuse large B-cell lymphoma and its association with the clinicopathological features. *Hum Pathol* **64**, 213-221 (2017).

29. T. W. Geurts, P. M. Nederlof, M. W. van den Brekel, L. J. van't Veer, D. de Jong, A. A. Hart, N. van Zandwijk, H. Klomp, A. J. Balm, M. L. van Velthuysen, Pulmonary squamous cell carcinoma following head and neck squamous cell carcinoma: metastasis or second primary? *Clin Cancer Res* **11**, 6608-6614 (2005).
30. A. Lal, R. Panos, M. Marjanovic, M. Walker, E. Fuentes, G. J. Kubicek, W. D. Henner, L. J. Buturovic, M. Halks-Miller, A gene expression profile test to resolve head & neck squamous versus lung squamous cancers. *Diagn Pathol* **8**, 44 (2013).
31. H. Feng, Y. Liu, X. Bian, F. Zhou, Y. Liu, ALDH1A3 affects colon cancer in vitro proliferation and invasion depending on CXCR4 status. *Br J Cancer* **118**, 224-232 (2018).
32. F. Zheng, Z. Zhang, V. Flamini, W. G. Jiang, Y. Cui, The Axis of CXCR4/SDF-1 Plays a Role in Colon Cancer Cell Adhesion Through Regulation of the AKT and IGF1R Signalling Pathways. *Anticancer Res* **37**, 4361-4369 (2017).
33. S. P. Treon, Y. Cao, L. Xu, G. Yang, X. Liu, Z. R. Hunter, Somatic mutations in MYD88 and CXCR4 are determinants of clinical presentation and overall survival in Waldenstrom macroglobulinemia. *Blood* **123**, 2791-2796 (2014).
34. C. Zhu, K. Yamaguchi, T. Ohsugi, Y. Terakado, R. Noguchi, T. Ikenoue, Y. Furukawa, Identification of FERM domain-containing protein 5 as a novel target of beta-catenin/TCF7L2 complex. *Cancer Sci* **108**, 612-619 (2017).
35. B. Zhang, W. H. Jia, K. Matsuda, S. S. Kweon, K. Matsuo, Y. B. Xiang, A. Shin, S. H. Jee, D. H. Kim, Q. Cai, J. Long, J. Shi, W. Wen, G. Yang, Y. Zhang, C. Li, B. Li, Y. Guo, Z. Ren, B. T. Ji, Z. Z. Pan, A. Takahashi, M. H. Shin, F. Matsuda, Y. T. Gao, J. H. Oh, S. Kim, Y. O. Ahn, Genetics, C. Epidemiology of Colorectal Cancer, A. T. Chan, J. Chang-Claude, M. L. Slattery, S. Colorectal Transdisciplinary, S. B. Gruber, F. R. Schumacher, S. L. Stenzel, R. Colon Cancer Family, G. Casey, H. R. Kim, J. Y. Jeong, J. W. Park, H. L. Li, S. Hosono, S. H. Cho, M. Kubo, X. O. Shu, Y. X. Zeng, W. Zheng, Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. *Nat Genet* **46**, 533-542 (2014).
36. X. P. Lu, G. N. Hu, J. Q. Du, H. Q. Li, TCF7L2 gene polymorphisms and susceptibility to breast cancer: a meta-analysis. *Genet Mol Res* **14**, 2860-2867 (2015).
37. J. Guo, C. Zhu, K. Yang, J. Li, N. Du, M. Zong, J. Zhou, J. He, Poly(C)-binding protein 1 mediates drug resistance in colorectal cancer. *Oncotarget* **8**, 13312-13319 (2017).
38. P. T. Lee, P. C. Liao, W. C. Chang, J. T. Tseng, Epidermal growth factor increases the interaction between nucleolin and heterogeneous nuclear ribonucleoprotein K/poly(C) binding protein 1 complex to regulate the gastrin mRNA turnover. *Mol Biol Cell* **18**, 5004-5013 (2007).
39. W. Zhang, H. Shi, M. Zhang, B. Liu, S. Mao, L. Li, F. Tong, G. Liu, S. Yang, H. Wang, Poly C binding protein 1 represses autophagy through downregulation of LC3B to promote tumor cell apoptosis in starvation. *Int J Biochem Cell Biol* **73**, 127-136 (2016).

40. K. Jeon, B. Min, J. S. Park, Y. K. Kang, Simultaneous Methylation-Level Assessment of Hundreds of CpG Sites by Targeted Bisulfite PCR Sequencing (TBPseq). *Front Genet* **8**, 97 (2017).
41. J. C. Kim, Y. J. Ha, S. A. Roh, E. Y. Choi, Y. S. Yoon, K. P. Kim, Y. S. Hong, T. W. Kim, D. H. Cho, S. Y. Kim, Y. S. Kim, Feasibility of proposed single-nucleotide polymorphisms as predictive markers for targeted regimens in metastatic colorectal cancer. *Br J Cancer* **108**, 1862-1869 (2013).
42. C. E. Fuller, All things rhabdoid and SMARC: An enigmatic exploration with Dr. Louis P. Dehner. *Semin Diagn Pathol* **33**, 427-440 (2016).
43. S. N. Kalimuthu, R. Chetty, Gene of the month: SMARCB1. *J Clin Pathol* **69**, 484-489 (2016).
44. J. Masliah-Planchon, I. Bieche, J. M. Guinebretiere, F. Bourdeaut, O. Delattre, SWI/SNF chromatin remodeling and human malignancies. *Annu Rev Pathol* **10**, 145-171 (2015).
45. I. Qaddoumi, W. Orisme, J. Wen, T. Santiago, K. Gupta, J. D. Dalton, B. Tang, K. Haupfear, C. Punchihewa, J. Easton, H. Mulder, K. Boggs, Y. Shao, M. Rusch, J. Becksfort, P. Gupta, S. Wang, R. P. Lee, D. Brat, V. Peter Collins, S. Dahiya, D. George, W. Konomos, K. M. Kurian, K. McFadden, L. N. Serafini, H. Nickols, A. Perry, S. Shurtleff, A. Gajjar, F. A. Boop, P. D. Klimo, Jr., E. R. Mardis, R. K. Wilson, S. J. Baker, J. Zhang, G. Wu, J. R. Downing, R. G. Tatevossian, D. W. Ellison, Genetic alterations in uncommon low-grade neuroepithelial tumors: BRAF, FGFR1, and MYB mutations occur at high frequency and align with morphology. *Acta Neuropathol* **131**, 833-845 (2016).
46. J. J. Castillo, Z. R. Hunter, G. Yang, S. P. Treon, Novel approaches to targeting MYD88 in Waldenstrom macroglobulinemia. *Expert Rev Hematol* **10**, 739-744 (2017).
47. D. D. W. Twa, A. Mottok, K. J. Savage, C. Steidl, The pathobiology of primary testicular diffuse large B-cell lymphoma: Implications for novel therapies. *Blood Rev* **32**, 249-255 (2018).
48. L. Mansouri, N. Papakonstantinou, S. Ntoufa, K. Stamatopoulos, R. Rosenquist, NF-kappaB activation in chronic lymphocytic leukemia: A point of convergence of external triggers and intrinsic lesions. *Semin Cancer Biol* **39**, 40-48 (2016).
49. I. Alroy, Y. Yarden, The ErbB signaling network in embryogenesis and oncogenesis: signal diversification through combinatorial ligand-receptor interactions. *FEBS Lett* **410**, 83-86 (1997).
50. S. Grant, L. Qiao, P. Dent, Roles of ERBB family receptor tyrosine kinases, and downstream signaling pathways, in the control of cell growth and survival. *Front Biosci* **7**, d376-389 (2002).
51. A. Patapoutian, L. F. Reichardt, Trk receptors: mediators of neurotrophin action. *Curr Opin Neurobiol* **11**, 272-280 (2001).
52. I. Tassi, J. Klesney-Tait, M. Colonna, Dissecting natural killer cell activation pathways through analysis of genetic mutations in human and mouse. *Immunol Rev* **214**, 92-105 (2006).
53. A. P. Gilmore, L. H. Romer, Inhibition of focal adhesion kinase (FAK) signaling in focal adhesions decreases cell motility and proliferation. *Mol Biol Cell* **7**, 1209-1224 (1996).

54. M. DesJardins, D. F. Gordon, Evaluation and selection of biases in machine learning. *Machine Learning Journal* **5**, 17 (1995).
55. V. N. Ngo, R. M. Young, R. Schmitz, S. Jhavar, W. Xiao, K. H. Lim, H. Kohlhammer, W. Xu, Y. Yang, H. Zhao, A. L. Shaffer, P. Romesser, G. Wright, J. Powell, A. Rosenwald, H. K. Muller-Hermelink, G. Ott, R. D. Gascoyne, J. M. Connors, L. M. Rimsza, E. Campo, E. S. Jaffe, J. Delabie, E. B. Smeland, R. I. Fisher, R. M. Braziel, R. R. Tubbs, J. R. Cook, D. D. Weisenburger, W. C. Chan, L. M. Staudt, Oncogenically active MYD88 mutations in human lymphoma. *Nature* **470**, 115-119 (2011).
56. H. W. Deng, Y. X. Fu, Counting mutations by parsimony and estimation of mutation rate variation across nucleotide sites - A simulation study. *Math Comput Model* **32**, 83-95 (2000).
57. X. Liu, X. Jian, E. Boerwinkle, dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat* **32**, 894-899 (2011).
58. J. M. Schwarz, D. N. Cooper, M. Schuelke, D. Seelow, MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods* **11**, 361-362 (2014).
59. H. A. Shihab, J. Gough, D. N. Cooper, I. N. Day, T. R. Gaunt, Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics* **29**, 1504-1510 (2013).

Supplementary Figures and Tables

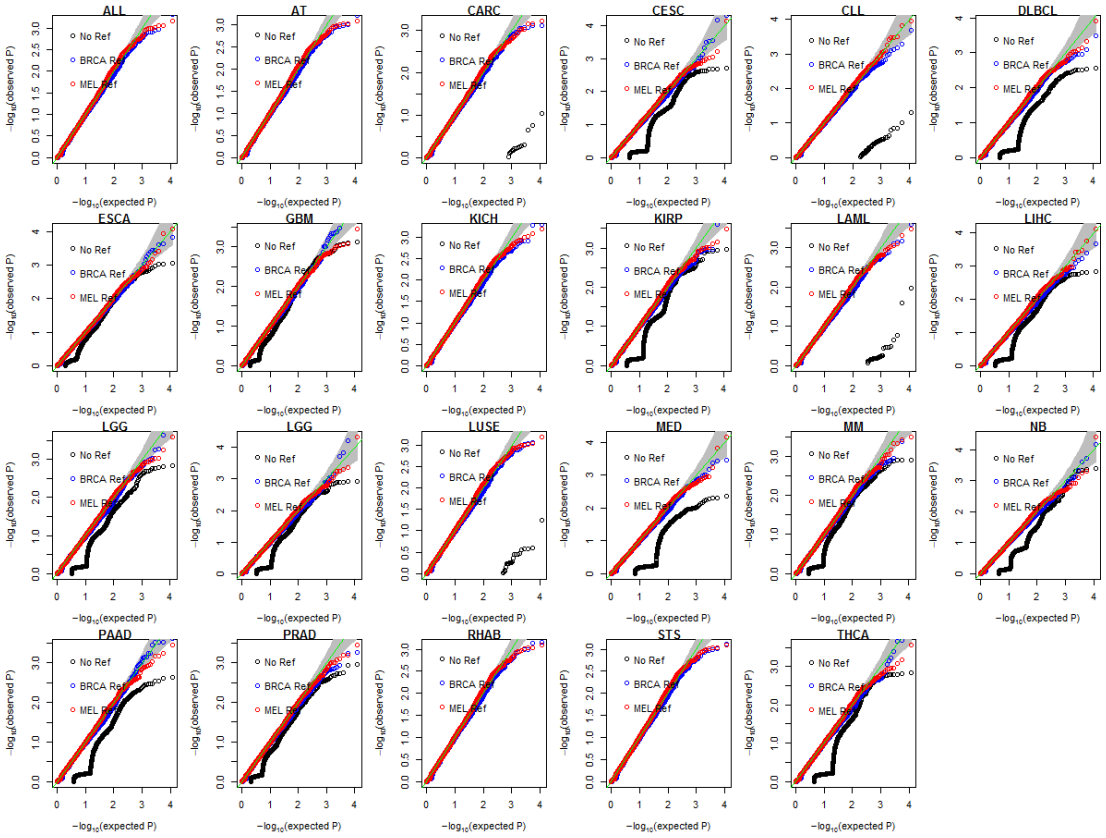


Figure S1: the QQ plots of p-values in 23 sample by WITER with and without reference samples
Note: The p-values less than a cutoff according to FDR 0.1 were excluded.

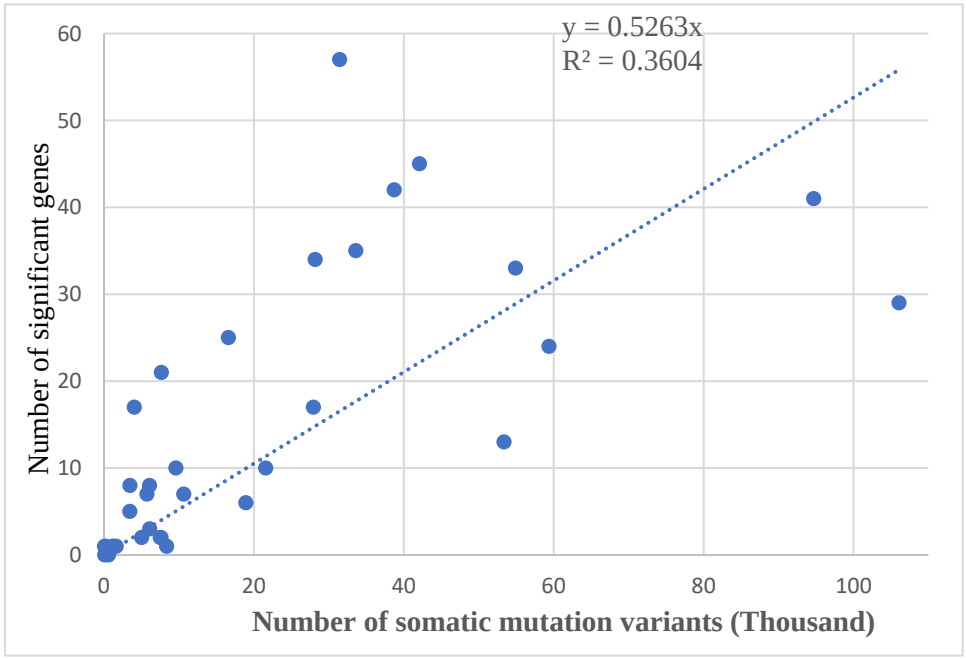


Figure S2: The linear regression between number of significant genes and number of somatic mutation variants
The dashed line is the fitted line. R2 is the coefficient of determination.

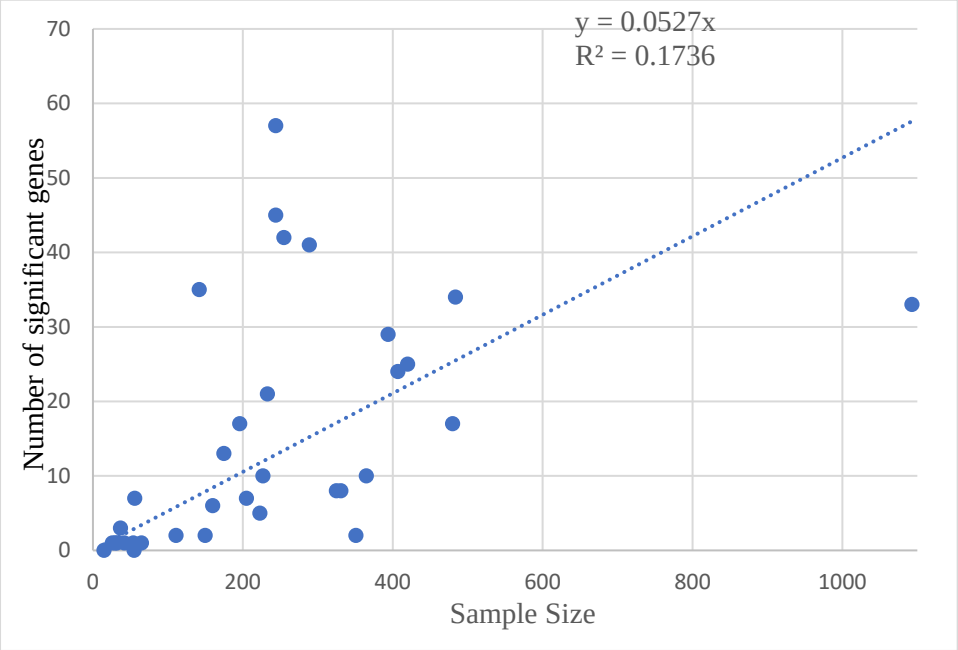


Figure S3: The linear regression between number of significant genes and sample size
The dashed line is the fitted line. R2 is the coefficient of determination.

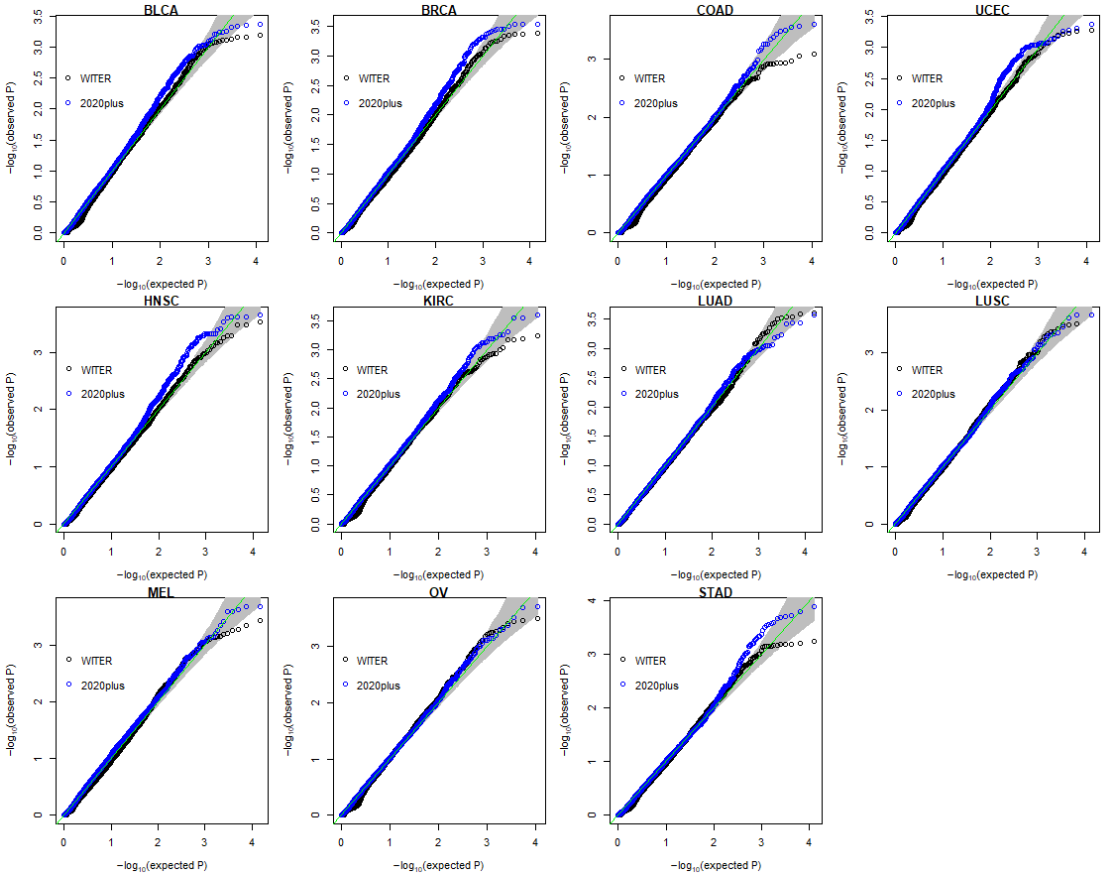


Figure S4. QQ plot of background gene p-values produced by WITER and 2020 plus methods in 11 cancers. The p-values less than a cutoff according to FDR 0.1 were excluded. Cancer name labels: BLCA: Bladder Urothelial Carcinoma; BRCA: Breast invasive carcinoma; COAD: Colon adenocarcinoma; UCEC: Uterine corpus endometrial carcinoma; HNSC: Head and Neck Squamous Cell Carcinoma; KIRC: Kidney Clear Cell Carcinoma; LUAD: Lung Adenocarcinoma; LUSC: Lung Squamous Cell Carcinoma; MEL: Melanoma; OV: Ovarian serous cystadenocarcinoma; STAD: Stomach Adenocarcinoma.

Figure S5: Performance comparison of WITER and 20/20+ for detecting cancer driver mutation in 11 cancers

a: the number of significant genes; b: cancer consensus significant genes. The p-values less than a cutoff according to FDR 0.1 were excluded. Cancer name labels: BLCA: Bladder Urothelial Carcinoma; BRCA: Breast invasive carcinoma; COAD: Colon adenocarcinoma; UCEC: Uterine corpus endometrial carcinoma; HNSC: Head and Neck Squamous Cell Carcinoma; KIRC: Kidney Clear Cell Carcinoma; LUAD: Lung Adenocarcinoma; LUSC: Lung Squamous Cell Carcinoma; MEL: Melanoma; OV: Ovarian serous cystadenocarcinoma; STAD: Stomach Adenocarcinoma.

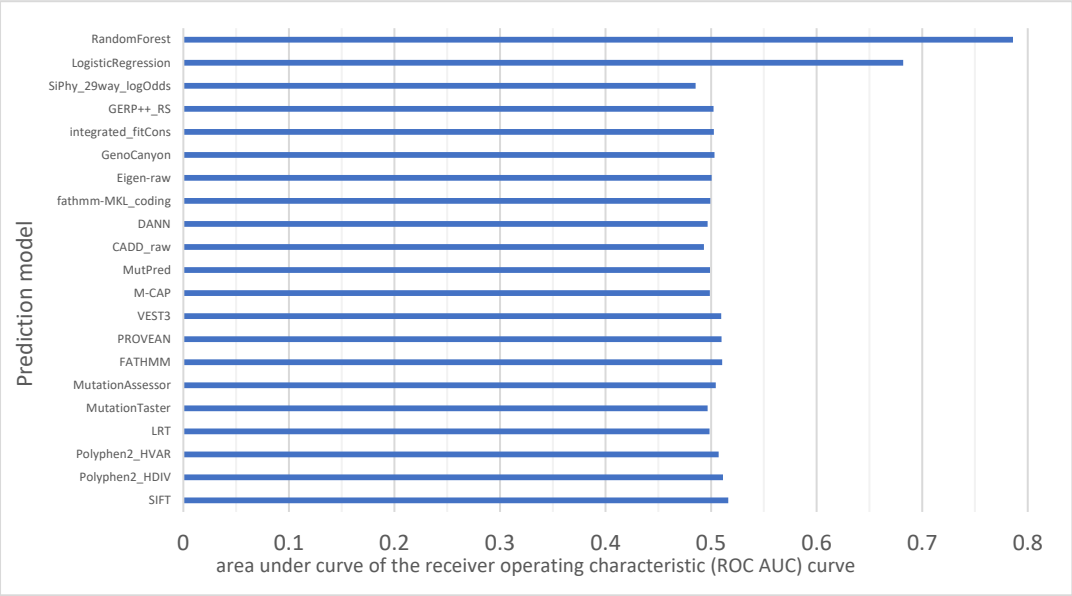


Figure S6: Performance of predicting frequent mutant variants by individual and combined methods. The 5 folder-cross validation was used to generate the AUC. The evaluation was carried out by a Java package WEKA(V3, https://www.cs.waikato.ac.nz/ml/weka/)

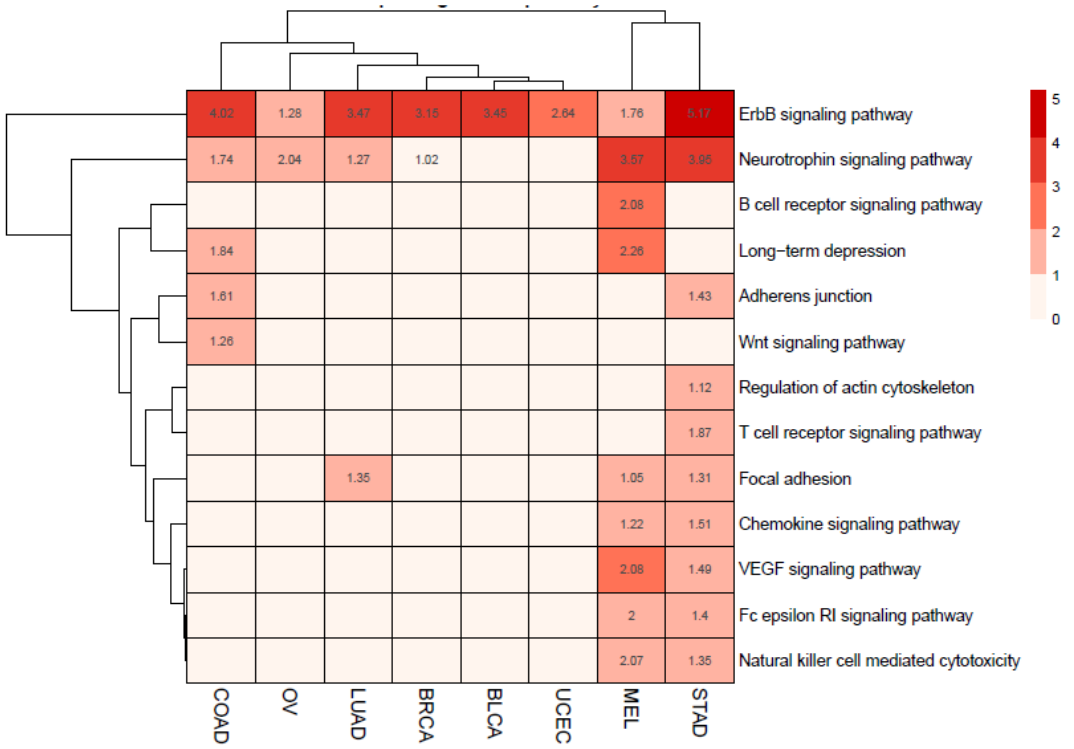


Figure S7: Heatmap of biological pathways enriched by driver-genes The values of the cells are -$\log_{10}$(FDR q-values). The colors indicate the magnitude of the values. The cancers and pathways are grouped by an heterarchical bi-cluster analysis in R

Table S1: Unique significant genes by WITER in the 11 cancers

Cancer	Gene	MutSigCV	oncodriveFML	ITER	WITER	CancerConsensus
BLCA	ELF3	0.010103	2.30E-04	1.61E-09	5.76E-09	1
BLCA	NFE2L2	5.59E-04	0.9962	6.68E-07	1.36E-16	1
BLCA	ERBB3	0.001646	0.24978	8.98E-06	1.64E-15	1
BLCA	ERBB2	0.08132	0.96319	0.002115	3.50E-07	1
BLCA	ERCC2	0.001459	0.92244	3.60E-07	9.69E-07	1
BLCA	NRAS	0.133981	0.92125	0.244118	3.86E-08	1
BLCA	IDH1	0.078409	0.53155	0.094995	9.54E-05	1
BLCA	FBXW7	4.28E-04	0.22664	3.09E-05	3.59E-17	1
BLCA	FGFR3	0.027388	0.18021	1.32E-05	8.23E-23	1
BLCA	PIK3CA	0.001452	0.65818	1.57E-06	2.15E-33	1
BLCA	KRAS	0.398044	0.99787	0.283061	1.41E-07	1
BLCA	HRAS	0.02805	0.89445	0.003506	1.62E-25	1
BLCA	CASP8	0.017869	0.39864	0.033465	1.29E-04	1

BLCA	CTNNB1	0.684378	0.57518	0.416055	1.86E-07	1
BLCA	SH2D2A	0.20842	0.79357	0.068549	2.88E-05	0
BLCA	CSPG5	0.182609	0.00472	0.107078	3.91E-06	0
BLCA	RXRA	0.00108	0.86402	1.24E-05	2.17E-05	0
BLCA	TAS1R3	0.045749	0.22505	1.32E-04	2.09E-04	0
BLCA	PSG1	0.05198	0.91418	0.012434	1.63E-07	0
BLCA	KLF5	0.181665	0.3705	6.55E-05	1.00E-04	0
BLCA	PCDHGA7	0.028043	0.28091	0.009189	9.39E-05	0
BLCA	RARG	0.010501	0.29293	4.14E-07	8.89E-08	0
BLCA	ZFP37	0.004457	0.49592	7.15E-05	1.42E-04	0
BLCA	TXNIP	0.604767	2.40E-04	3.49E-05	5.80E-05	0
BLCA	ZNF878	0.96682	0.95306	0.169874	1.65E-10	0
BLCA	STK39	0.771555	1	0.359025	1.91E-05	0
BLCA	ZNF624	0.512218	0.26753	3.57E-08	1.65E-07	0
BLCA	ZNF513	0.039256	0.02061	1.65E-04	2.80E-04	0
BLCA	UGT1A9	0.10839	0.59112	0.100226	1.08E-04	0
BRCA	FAM47C	0.087477	0.36442	0.009041	1.11E-04	1
BRCA	DNMT3A	0.843839	0.50455	0.064191	2.64E-05	1
BRCA	HIST1H3B	0.002763	0.22537	8.65E-05	1.41E-04	1
BRCA	ERBB3	6.32E-04	0.14789	9.74E-05	2.02E-08	1
BRCA	ERBB2	0.00302	0.45049	2.27E-05	2.42E-26	1
BRCA	FOXA1	0.024566	0.00308	5.67E-06	1.09E-05	1
BRCA	XPO1	0.798226	0.16802	0.068682	2.85E-05	1
BRCA	KLF4	0.962375	0.72487	0.256697	1.07E-05	1
BRCA	FBXW7	0.691241	0.49823	0.004454	3.04E-08	1
BRCA	FGFR2	0.12402	0.41424	0.012374	1.61E-10	1
BRCA	PIK3R1	0.007757	0.0056	1.39E-07	2.37E-07	1
BRCA	KRAS	0.001827	0.33858	0.00379	1.08E-26	1
BRCA	GLA	0.329366	0.80435	0.089082	1.40E-04	0
BRCA	FCRL5	0.277234	0.50848	0.008545	1.95E-06	0
BRCA	DUSP16	0.051057	0.01668	0.007051	7.55E-05	0
BRCA	TPRX1	0.253056	0.37048	0.071328	2.41E-05	0
BRCA	ZFP36L1	0.002974	1.30E-04	3.73E-05	5.67E-05	0
COAD	CHD4	0.078877	0.20769	0.002289	2.45E-06	1
COAD	PPP2R1A	0.141973	0.72775	0.141875	1.06E-07	1
COAD	CNBD1	0.110659	0.86018	0.010546	2.40E-05	1
COAD	ERBB3	0.1158	0.64679	0.00142	1.41E-13	1
COAD	ERBB2	0.065614	0.9027	0.005465	9.95E-14	1
COAD	LIFR	0.006573	0.29661	1.96E-04	4.14E-04	1
COAD	CXCR4	0.011262	0.86731	0.031564	5.21E-04	1
COAD	IDH2	0.045499	0.22688	0.014912	7.01E-06	1

COAD	BRAF	4.82E-04	0.32704	8.91E-08	1.82E-50	1
COAD	RAF1	0.244126	0.72739	0.009653	5.35E-15	1
COAD	PIK3CA	0.001561	0.98003	6.31E-16	9.59E-90	1
COAD	PIK3R1	0.360783	0.034	0.044889	1.61E-05	1
COAD	PCBP1	0.02945	0.22172	0.004449	7.12E-07	1
COAD	ARID1A	0.003605	0.28983	0.001281	2.92E-04	1
COAD	GNAS	0.242151	0.82802	0.009841	1.23E-07	1
COAD	CTNNB1	0.001028	0.50651	1.03E-04	1.44E-10	1
COAD	ACVR1B	0.005239	0.3425	1.48E-04	3.01E-04	0
COAD	ZC3H13	0.002089	0.09886	3.47E-04	3.12E-04	0
COAD	NGB	0.002664	0.80948	0.01041	2.30E-04	0
COAD	ATG2A	0.802205	0.61151	0.232782	5.78E-04	0
COAD	KLC4	0.039784	0.7423	2.99E-04	5.96E-04	0
COAD	DUSP16	0.087652	0.52282	0.018478	1.65E-04	0
COAD	PCDHA1	0.372299	0.69821	1.38E-04	7.00E-20	0
COAD	PCDHA5	0.643394	0.52408	0.001556	2.91E-18	0
COAD	PCDHA4	0.555239	0.26458	0.002888	1.02E-16	0
COAD	PCDHA3	0.106343	0.79539	4.69E-05	2.27E-20	0
COAD	PCDHA2	0.48885	0.68016	0.003121	1.54E-16	0
COAD	PCDHA9	0.571704	0.61216	7.01E-04	2.74E-18	0
COAD	PCDHA8	1	0.97541	0.049066	7.49E-15	0
COAD	PCDHA7	0.641067	0.61617	6.92E-04	3.53E-19	0
COAD	PCDHA6	0.645205	0.67044	0.001918	8.60E-18	0
COAD	PCDHAC2	0.822949	0.63764	0.005493	3.00E-15	0
COAD	PCDHAC1	0.356993	0.11488	0.001222	1.35E-15	0
COAD	KIAA1804	0.145079	0.60659	3.25E-04	6.64E-06	0
COAD	PCDHA13	0.064687	0.05377	4.32E-05	5.81E-18	0
COAD	PCDHA12	0.3295	0.00718	6.95E-05	3.80E-21	0
COAD	PCDHA11	0.481675	0.35176	0.001229	1.18E-18	0
COAD	PCDHA10	0.216798	0.20336	5.62E-05	1.51E-21	0
COAD	RGMB	0.564593	0.03731	0.348865	2.30E-05	0
COAD	ING1	0.003298	0.26098	0.004505	2.00E-07	0
COAD	OPRM1	0.174087	0.75567	0.073868	1.95E-07	0
COAD	RIMS2	0.669523	0.87839	0.187879	1.36E-04	0
COAD	ZNF560	0.091638	0.38193	0.002806	8.45E-05	0
COAD	STIM2	0.459306	0.3714	0.100589	4.77E-04	0
HNSC	FCRL4	0.007489	0.24308	1.14E-04	1.68E-04	1
HNSC	SMAD4	0.002188	0.36564	0.006628	1.82E-07	1
HNSC	CTCF	0.001286	0.03975	2.24E-05	3.07E-05	1
HNSC	KEAP1	0.001599	0.57845	2.75E-05	3.94E-05	1
HNSC	RAC1	0.014924	0.63907	0.017661	9.06E-05	1

HNSC	FGFR3	0.370749	0.58282	0.009941	1.49E-04	1
HNSC	EP300	2.42E-04	0.74769	0.001763	2.17E-06	1
HNSC	CD248	0.681864	0.64177	4.14E-05	5.87E-05	0
HNSC	ADCY8	0.010859	0.73408	1.24E-04	1.52E-04	0
HNSC	ZNF750	0.004355	0.00107	2.96E-06	4.76E-06	0
HNSC	PEG3	0.048453	0.3124	4.29E-05	5.65E-05	0
HNSC	ZNF563	0.257063	0.21164	0.007018	9.62E-05	0
KIRC	DNMT3A	0.005229	0.14019	1.00E-04	1.88E-04	1
KIRC	NFE2L2	0.002882	0.44296	0.001841	1.28E-06	1
KIRC	NF2	0.085351	0.14288	0.024529	4.35E-06	1
KIRC	SMARCA4	0.051244	0.82687	0.002104	1.97E-05	1
KIRC	TCF12	0.115914	0.70984	0.016125	3.88E-05	1
KIRC	CDKN2A	0.016215	0.20555	0.071108	1.36E-04	1
KIRC	MTOR	1.40E-04	0.70182	3.42E-05	4.32E-05	1
KIRC	PIK3CA	0.038472	0.72737	6.76E-06	8.48E-32	1
KIRC	ARID1A	0.177427	0.00908	8.99E-05	1.14E-04	1
KIRC	PABPC1	0.999712	0.75376	2.10E-04	6.48E-13	1
KIRC	ZC3HC1	0.200287	0.05711	0.019925	6.16E-07	0
KIRC	SLFN12L	0.128629	0.67547	0.025289	2.13E-05	0
KIRC	HK1	0.903897	0.43214	0.069491	2.66E-04	0
KIRC	SLC11A2	0.392912	0.02879	0.279041	1.36E-06	0
KIRC	HLA-DQB1	0.35833	0.88843	0.283265	3.20E-04	0
KIRC	COBL	0.231746	0.34013	0.023333	5.70E-06	0
KIRC	MAP2K3	0.127763	0.70998	0.069393	1.26E-07	0
KIRC	MAP4K5	0.204151	0.01573	0.03684	1.48E-05	0
KIRC	FAM200A	0.023997	0.54103	1.67E-04	3.57E-04	0
KIRC	PABPC3	0.240863	0.66656	0.014623	3.79E-07	0
KIRC	ALKBH8	0.609278	0.15275	0.027668	1.18E-08	0
KIRC	PRSS3	0.043846	0.38443	0.227373	1.47E-08	0
KIRC	ZNF800	0.036792	0.15307	1.55E-04	3.47E-04	0
KIRC	ZNF563	0.145925	0.96191	0.2047	5.47E-05	0
KIRC	NLRP7	0.86958	0.56637	0.179508	4.52E-05	0
KIRC	PIGQ	0.067671	0.75672	0.007288	7.30E-05	0
LUAD	ERBB2	0.209199	0.31662	0.006193	4.22E-06	1
LUAD	NRAS	0.122752	0.59442	0.192335	5.67E-09	1
LUAD	BRAF	0.001509	0.9526	0.001325	1.73E-13	1
LUAD	FBXW7	0.075704	0.49917	0.067721	1.17E-04	1
LUAD	PIK3CA	0.557858	0.62573	0.005158	2.60E-17	1
LUAD	EGFR	0.187768	0.07497	3.20E-05	1.46E-25	1
LUAD	HRAS	0.30991	0.4113	0.440817	5.10E-06	1

LUAD	B2M	0.009884	0.12132	0.276422	1.46E-04	1
LUAD	CTNNB1	0.059615	0.77653	0.009174	5.73E-14	1
LUAD	KCNB1	0.002193	0.16821	6.26E-05	7.22E-05	0
LUAD	TARS2	0.001074	0.69406	1.21E-04	1.46E-04	0
LUAD	HGF	5.61E-04	0.73882	8.01E-05	8.67E-05	0
LUAD	PCDHA1	0.209752	0.27592	0.001493	3.17E-06	0
LUAD	PCDHA2	0.827439	0.73373	0.006186	1.30E-06	0
LUAD	PCDHA9	0.987765	0.28907	0.04019	1.46E-05	0
LUAD	PCDHA6	0.947894	0.54997	0.035802	1.24E-04	0
LUAD	PCDHAC2	0.96184	0.27743	0.061438	1.56E-04	0
LUAD	PCDHAC1	0.981846	0.00265	0.053546	1.12E-04	0
LUAD	PCDHA13	0.644272	0.12318	0.009487	1.77E-06	0
LUAD	PCDHA11	0.661556	0.88795	0.011187	1.56E-05	0
LUAD	PCDHA10	0.466271	0.01717	0.012936	1.04E-05	0
LUAD	ZNF716	9.10E-04	0.47729	0.002249	4.26E-05	0
LUSC	BRAF	0.111669	0.92827	0.02411	1.11E-06	1
LUSC	FBXW7	0.71995	0.32652	0.040881	3.21E-10	1
LUSC	FGFR3	0.711763	0.60777	0.071165	4.27E-09	1
LUSC	PIK3CA	6.58E-05	0.70456	2.01E-07	2.56E-41	1
LUSC	HRAS	0.004331	0.60182	0.001246	3.00E-26	1
LUSC	EP300	0.566185	0.44866	0.023434	4.04E-06	1
LUSC	CYP11B1	0.018104	0.3285	7.75E-05	3.67E-05	0
LUSC	PDYN	1.38E-04	0.3027	9.77E-05	1.78E-05	0
MEL	KIT	0.998462	0.03646	0.366527	7.44E-05	1
MEL	IDH1	0.002983	0.66861	0.001578	1.05E-20	1
MEL	RB1	0.003552	0.62742	0.026078	1.72E-04	1
MEL	SF3B1	0.867726	0.0909	0.113301	2.09E-04	1
MEL	FBXW7	0.227594	0.37101	0.041834	3.17E-06	1
MEL	MAP2K1	0.01392	0.65035	1.26E-04	3.01E-16	1
MEL	PIK3CA	0.997668	0.99968	0.145884	1.08E-07	1
MEL	A1CF	0.150594	0.56116	0.042353	3.67E-05	1
MEL	CLIP1	0.93025	0.893	0.099724	9.97E-05	1
MEL	KRAS	0.253845	0.97933	0.157189	2.37E-05	1
MEL	TP53	1.50E-04	0.10517	1.23E-12	1.49E-99	1
MEL	HRAS	0.787203	0.33192	0.386905	9.68E-06	1
MEL	GNA11	0.305116	0.00179	0.038123	1.52E-10	1
MEL	CTNNB1	0.229997	0.08914	0.006305	2.75E-11	1
MEL	TCHHL1	7.39E-04	0.24143	7.57E-05	4.08E-05	0
MEL	PRDM7	0.473125	0.02803	0.150664	1.54E-04	0
MEL	DSG3	0.007989	0.54472	1.05E-05	2.32E-05	0
MEL	COL21A1	0.005546	0.17557	1.50E-05	2.35E-06	0

MEL	GUCA1C	0.999985	0.94228	0.354276	1.71E-04	0
MEL	TCP10L2	0.771295	0.98607	0.108327	8.27E-06	0
MEL	MUC13	0.13666	0.63281	0.002076	1.81E-04	0
MEL	PYHIN1	0.089041	0.85172	0.041391	2.87E-04	0
MEL	SELP	0.991729	0.66347	0.022625	3.72E-05	0
MEL	PDE1A	0.020903	0.24727	0.003439	8.29E-07	0
MEL	PCDHA1	0.736402	0.49221	6.69E-04	6.41E-07	0
MEL	PCDHA4	0.664652	0.69132	0.014475	9.36E-05	0
MEL	PCDHA9	0.985803	0.89353	0.045793	2.11E-05	0
MEL	PCDHA8	0.943164	0.45249	0.004418	4.75E-06	0
MEL	PCDHAC1	0.975452	0.90427	0.022801	1.60E-04	0
MEL	PCDHA13	0.974912	0.53566	0.038056	1.01E-04	0
MEL	PCDHA11	0.832659	0.79157	0.015775	2.61E-05	0
MEL	PCDHA10	0.619007	0.6205	0.013421	8.04E-06	0
MEL	OR4M2	0.248195	0.70563	0.040744	1.51E-05	0
MEL	STK19	0.015007	0.40515	0.001674	1.56E-15	0
OV	PPP2R1A	0.035963	0.70978	0.02113	8.49E-08	1
OV	ERBB2	0.09813	0.60963	0.013541	1.52E-07	1
OV	NRAS	0.002679	0.19963	0.01214	2.27E-10	1
OV	BRAF	0.027588	0.88872	0.147142	1.53E-06	1
OV	FBXW7	0.054434	0.55663	0.026804	1.70E-08	1
OV	PIK3CA	0.764045	0.62605	0.34417	2.83E-06	1
OV	KRAS	0.031988	0.78275	0.064287	5.91E-15	1
OV	BRCA1	0.046156	0.00213	3.99E-07	3.35E-07	1
OV	RHOA	0.017567	0.49302	0.041488	4.64E-06	1
OV	CASP8	0.098064	0.16397	0.020871	3.70E-05	1
OV	B2M	0.00113	1	0.184225	4.06E-08	1
OV	HLA-G	0.009173	0.98089	0.035958	1.39E-04	0
STAD	CHD4	0.245523	0.93725	0.001411	4.80E-05	1
STAD	PBRM1	0.035327	0.37478	0.007447	3.63E-04	1
STAD	CNBD1	0.22133	0.33913	0.004851	1.17E-05	1
STAD	SMAD2	1.17E-04	0.183	0.007402	2.24E-08	1
STAD	ERBB3	0.054637	0.43843	1.52E-05	2.56E-14	1
STAD	ERBB2	0.378157	0.51653	1.25E-04	7.70E-22	1
STAD	PTEN	0.002746	0.76119	2.43E-04	5.34E-10	1
STAD	NRAS	0.232912	0.55772	0.122884	6.70E-08	1
STAD	BRAF	0.723083	0.2195	0.11387	6.04E-06	1
STAD	FBXW7	0.30357	0.07034	0.057405	5.96E-07	1
STAD	RAF1	0.556977	0.65231	0.135408	3.37E-07	1
STAD	CDKN2A	0.002944	0.01257	0.006938	4.80E-05	1
STAD	STAT3	0.999879	0.85839	0.033901	1.75E-04	1

STAD	PIK3R1	0.087327	0.00732	0.038981	2.61E-05	1
STAD	APC	3.95E-04	0.00153	0.007943	2.59E-05	1
STAD	CTNNB1	0.080782	0.36951	0.003409	1.85E-15	1
STAD	ZC3H4	0.554437	0.43913	0.042694	3.03E-04	0
STAD	SH2D2A	0.210084	0.68834	0.053048	1.14E-05	0
STAD	TCEAL6	0.289818	0.66131	0.132711	2.80E-04	0
STAD	SLITRK6	0.156137	0.40283	8.37E-05	1.62E-04	0
STAD	ADRA1A	0.061514	0.29788	0.015719	3.69E-07	0
STAD	PRNP	0.24016	0.93164	0.211218	2.85E-06	0
STAD	PTH2	0.016192	0.08521	0.038948	8.78E-15	0
STAD	HLA-B	0.034437	0.02137	0.008293	8.88E-05	0
STAD	BEST3	0.034339	0.35638	7.28E-05	1.36E-04	0
STAD	BNC2	0.33532	0.54984	0.029017	5.56E-06	0
STAD	MAP2K7	0.016319	0.44553	3.26E-04	4.33E-07	0
STAD	TTK	0.884283	0.03814	0.063077	4.89E-07	0
STAD	AQP2	0.016088	0.97596	0.215925	2.61E-04	0
STAD	TLR4	0.00703	0.18558	4.89E-05	1.15E-04	0
STAD	OR4C3	0.141493	0.81999	0.034835	3.63E-08	0
STAD	ZNF878	0.789422	1	0.383649	4.69E-07	0
STAD	ZNF721	0.534072	0.97744	0.012705	7.39E-08	0
STAD	ZNF716	0.089027	0.49855	0.014667	2.19E-04	0
STAD	SLCO1B3	0.028308	0.88891	5.61E-04	1.63E-04	0
UCEC	U2AF1	0.209206	0.18068	0.334132	3.73E-07	1
UCEC	NFE2L2	0.048451	0.94445	3.41E-05	1.16E-08	1
UCEC	ESR1	0.344057	0.14186	0.00915	5.10E-09	1
UCEC	ERBB3	0.213392	0.47654	7.62E-05	4.72E-10	1
UCEC	ERBB2	0.942756	0.88756	0.063658	4.10E-05	1
UCEC	MAX	0.029877	0.19557	0.010747	5.53E-16	1
UCEC	LZTR1	0.035479	0.84019	7.56E-05	1.24E-04	1
UCEC	AKT1	1.33E-04	0.83264	0.003355	1.36E-14	1
UCEC	NRAS	0.003672	0.9416	0.001754	8.38E-23	1
UCEC	DICER1	0.824804	0.28748	0.070705	7.58E-05	1
UCEC	RB1	0.054063	0.88633	0.017768	2.41E-04	1
UCEC	CCND1	3.45E-04	0.1409	3.04E-06	7.15E-06	1
UCEC	VHL	0.145151	0.72076	0.246007	1.48E-06	1
UCEC	SGK1	0.411208	0.22921	2.61E-04	3.74E-04	1
UCEC	KRAS	0.159112	0.72248	9.04E-25	6.68E-190	1
UCEC	MYCN	0.123471	0.556	0.013539	6.07E-11	1
UCEC	CASP8	0.006749	0.58603	0.002024	3.77E-06	1
UCEC	CTNND1	0.520007	0.13848	5.16E-04	1.86E-06	1
UCEC	BCAN	0.033629	0.49353	6.53E-05	1.16E-04	0

UCEC	GIGYF2	0.751871	0.07728	7.68E-04	8.62E-09	0
UCEC	FOXA2	6.29E-04	1.70E-04	1.54E-05	3.27E-05	0
UCEC	MYO10	0.712137	0.10595	0.058648	4.00E-05	0
UCEC	DUSP16	0.013243	0.43285	0.01384	1.86E-04	0
UCEC	HEPH	0.187478	0.62103	0.020346	1.35E-04	0
UCEC	ING1	2.74E-04	0.17255	0.018412	3.31E-06	0
UCEC	SLC6A2	0.525061	0.2611	0.047213	8.32E-05	0
UCEC	METTL8	0.019233	0.87109	0.109738	1.05E-05	0
UCEC	INPP4A	0.556185	0.01036	1.08E-04	1.79E-04	0

The p-values less than a cutoff according to FDR 0.1 were excluded. Cancer name labels: BLCA: Bladder Urothelial Carcinoma; BRCA: Breast invasive carcinoma; COAD: Colon adenocarcinoma; UCEC: Uterine corpus endometrial carcinoma; HNSC: Head and Neck Squamous Cell Carcinoma; KIRC: Kidney Clear Cell Carcinoma; LUAD: Lung Adenocarcinoma; LUSC: Lung Squamous Cell Carcinoma; MEL: Melanoma; OV: Ovarian serous cystadenocarcinoma; STAD: Stomach Adenocarcinoma.

Table S2 Enrichment significance of related genes in the unique significant genes by WITER

Cancer	Sample	With 3 or more hit papers	Without or less 3 hit papers	[a]p-value	In consensus list	Out consensus list	[b]p-value
BLCA	Real	10	19	0.00077	14	15	0
	Random	0	29		-	-	
BRCA	Real	12	5	0.0013	12	5	2.59E-14
	Random	2	15		-	-	
COAD	Real	15	29	0.0028	16	28	1.30E-12
	Random	3	41		-	-	
UCEC	Real	15	13	4.6E-6	18	10	0
	Random	0	28		-	-	
HNSC	Real	5	7	0.037	7	5	5.56E-08
	Random	0	12		-	-	
KIRC	Real	6	20	0.022	10	16	1.20E-08
	Random	0	26		-	-	
LUAD	Real	10	12	0.00052	9	13	3.47E-08
	Random	0	22		-	-	
LUSC	Real	3	5	0.2	6	2	6.00E-08
	Random	0	8		-	-	
MEL	Real	13	21	0.0087	14	20	4.49E-12
	Random	3	31		-	-	
OV	Real	12	0	9.6E-6	11	1	1.55E-15
	Random	1	11		-	-	
STAD	Real	6	29	0.025	16	19	1.72E-14
	Random	0	35		-	-	

a: The p-values was calculated by the Fisher exact test. b: The p-values was calculated by the hypergeometric distribution test. Cancer name labels: BLCA: Bladder Urothelial Carcinoma; BRCA: Breast invasive carcinoma; COAD: Colon adenocarcinoma; UCEC: Uterine corpus endometrial carcinoma; HNSC: Head and Neck Squamous Cell Carcinoma; KIRC: Kidney Clear Cell Carcinoma; LUAD: Lung Adenocarcinoma; LUSC: Lung Squamous Cell Carcinoma; MEL: Melanoma; OV: Ovarian serous cystadenocarcinoma; STAD: Stomach Adenocarcinoma.

Table S3: The number of significant genes in 23 cancers with two different extreme samples by ITER

Gene	Melanoma Ref	Breast invasive carcinoma Ref	Overlapped Gene Number	Overlapped Percentage
Acute lymphoblastic leukemia	0	0	0	-
Astrocytoma	0	0	0	-
Carcinoid Cancer	0	0	0	-
Chronic lymphocytic leukemia	0	0	0	-
Cervical Carcinoma	2	1	1	0.5
Diffuse large B-cell lymphoma	0	2	0	0
Esophageal carcinoma	1	1	1	1
Glioblastoma Multiforme	5	6	5	0.83
Kidney chromophobe carcinoma	0	1	0	0
Kidney Papillary Cell Carcinoma	0	0	0	-
Acute myeloid leukemia	10	13	10	0.77
Liver Hepatocellular Carcinoma	1	1	1	1
Low Grade Glioma	4	4	4	1
Small cell lung carcinoma	1	1	1	1
B-cell lymphomas	0	1	0	0
Multiple myeloma	2	2	2	1
Medulloblastoma	3	3	3	1
Neuroblastoma	0	0	0	-
Pancreatic Adenocarcinoma	3	3	3	1
Prostate Adenocarcinoma	2	2	2	1
Rhabdoid tumor	0	0	0	-
Soft Tissue Sarcoma	0	0	0	-
Thyroid Carcinoma	2	2	2	1

Note: The percentage of overlapped genes is calculated by #overlapped genes/(#sig. genes in set1 +#sig. genes in set2-#overlapped genes).

Table S4: The number of significant genes in 34 cancers with hit papers in PubMed database

Cancer	AllSig.Gene	HitSig.Gene	Proportion
AT	1	1	1.000
CLL	5	5	1.000
DLBCL	7	7	1.000
ESCA	6	6	1.000
GBM	10	10	1.000
KICH	1	1	1.000
KIRP	2	2	1.000
LIHC	2	2	1.000
LUSE	1	1	1.000
LB	1	1	1.000
NB	2	2	1.000
RHAB	1	1	1.000
LAML	17	16	0.941

OV	17	15	0.882
BRCA	33	29	0.879
MED	8	7	0.875
LUSC	13	11	0.846
LGG	10	8	0.800
HNSC	24	19	0.792
UCEC	42	31	0.738
MM	7	5	0.714
CESC	3	2	0.667
LUAD	29	18	0.621
MEL	41	25	0.610
COAD	57	34	0.596
BLCA	35	18	0.514
THCA	8	4	0.500
KIRC	34	14	0.412
PRAD	25	9	0.360
STAD	45	15	0.333
PAAD	21	6	0.286
CARC	1	0	0.000
ALL	0	0	-
STS	0	0	-

AllSig.Gene: the number of all significant genes. HitSig.Gene: the number of significant genes with hit papers. ALL: Acute lymphoblastic leukemia, AT: Astrocytoma, BLCA: Bladder Urothelial Carcinoma, BRCA: Breast invasive carcinoma, CARC: Carcinoid Cancer, CESC: Cervical Carcinoma, CLL: Chronic lymphocytic leukemia, COAD: Colon adenocarcinoma, DLBCL: Diffuse large B-cell lymphoma, ESCA: Esophageal carcinoma, GBM: Glioblastoma Multiforme, HNSC: Head and Neck Squamous Cell Carcinoma, KICH: Kidney chromophobe carcinoma, KIRC: Kidney Clear Cell Carcinoma, KIRP: Kidney Papillary Cell Carcinoma, LAML: Acute myeloid leukemia, LB: B-cell lymphomas, LGG: Low Grade Glioma, LIHC: Liver Hepatocellular carcinoma, LUAD: Lung Adenocarcinoma, LUSC: Lung Squamous Cell Carcinoma, LUSE: Small cell lung carcinoma, MED: Medulloblastoma, MEL: Melanoma, MM: Multiple myeloma , NB: Neuroblastoma, OV: Ovarian serous cystadenocarcinoma, PAAD: Pancreatic Adenocarcinoma, PRAD: Prostate Adenocarcinoma, RHAB: Rhabdoid tumor, STAD: Stomach Adenocarcinoma, STS: Soft Tissue Sarcoma, THCA: Thyroid Carcinoma, UCEC: Uterine corpus endometrial carcinoma

Table S5. Uniquely significant genes of different cancers

Cancer	Total	[a]Uni.	Genes
KIRC	34	23	BAP1[9.49e-28(29617669;29558292;29426696;29266978;29158875;29118224;28900502;28812986;28779136;28765116;28753773;28731045;28723536;28618948;28488170;28473526;28459210;28408295;28327121;28284891;28212566;27751729;27556922;27085487;26891804;26864202;26854086;26839909;26484545;26452128;26300492;26300218;26166446;26111976;25972334;25873528;25826081;28326264;25479927;25465300;25126716;25124064;24821879;24382589;24166983;24158655;24128712;24076305;24029645;23709298;23620406;23277170;23036577;22949125;22805307;22461374)], KDM5C[3.86e-15(28779136;28723536;28408295;28212566;27751729;27556922;26484545;25124064;24029645;23036577;22949125;21725364;20054297)], SETD2[6.89e-14(29674707;29558292;28812986;28779136;28754676;28753773;28731045;28723536;28445125;28408295;28260718;27764136;27751729;27288695;27556922;27292023;26891804;26864202;26575290;26559293;26537074;26452128;26166446;26111976;26073078;25873528;25853938;25714014;28326264;25124064;24821879;24186201;24166983;24158655;24029645;23792563;23620406;23036577;22949125;22805307;22461374;20501857;20054297)], PABPC1[6.48e-13(?)], ALKBH8[1.18e-08(?)], PRSS3[1.47e-08(?)], MAP2K3[1.26e-07(?)], PABPC3[3.79e-07(?)], ZC3HC1[6.16e-07(?)], SLC11A2[1.36e-06(?)], NF2[4.35e-06(25893302;20528227;20054297)], COBL[5.70e-06(?)], MAP4K5[1.48e-05(?)], SLFN12L[2.13e-05(?)], TCF12[3.88e-05(?)], MTOR[4.32e-05(29610387;29536301;29508215;29479523;29202733;29158991;29144820;29118224;29079709;28927098;28779136;28765116;28723536;28680592;28545465;28396841;28329682;28257806;28247252;28197812;27751729;27738339;27615548;27574806;27453294;27405474;26974204;26787754;26609489;26408740;26255626;26224474;25997916;25948777;25783986;25625928;25520878;25426415;25351205;25293974;25186283;25152703;24929890;24821879;24575852;24565854;24504440;24496460;24495452;24136229;23797736;23633458;23290145;22322364;21798997;21644050;20709527;20437403;20022054;19956876;19843858;19663736;19657325;19265534;17987219)], NLRP7[4.52e-05(?)], PIGQ[7.30e-05(?)], IST1[1.14e-04(?)], HK1[2.66e-04(?)], HLA-DQB1[3.20e-04(?)], ZNF800[3.47e-04(?)], FAM200A[3.57e-04(?)]
COAD	57	20	GNAS[1.23e-07(29069792;28749961;27938333;27756406;27568332;26074686;24498230;24475022;23857251;23403822;20126641)], OPRM1[1.95e-07(?)], TCF7L2[5.64e-07(29511559;29389519;29301589;29245969;29135090;29131639;29118424;29050326;28949031;28811361;28472810;28450117;28362475;28343235;28143522;28117551;28060743;28002797;27792933;27761963;27504909;27755946;27709738;27527215;27398792;26474385;26243311;26191083;26165840;26060019;25913757;25277775;25205133;25131200;25050608;24943349;24913975;24836286;24828199;24670930;24608966;24398765;24338422;24317174;23951231;23817222;23796952;23319804;22895193;22419714;22

			108803;22005519;21983179;21956205;21892161;20056134;19924301;19760027;19561607;18992263;18621708;18478343;18398040;18268068;18268006)], PCBP1[7.12e-07(28076324;17928403)], AMER1[1.52e-06(28551381;28002797;26527806;26071483;24251807)], KIAA1804[6.64e-06(29453410;24811787)], MLK4[9.89e-06(17016444)], RGMB[2.30e-05(27384995;26029998)], ZNF560[8.45e-05(24137329)], RIMS2[1.36e-04(?)], GRIK3[1.77e-04(?)], NGB[2.30e-04(25683270)], ACVR1B[3.01e-04(26497569)], ZC3H13[3.12e-04(21388066)], LIFR[4.14e-04(28751909;23579219;21617854;21239504)], STIM2[4.77e-04(28087343;25143380;22125164)], CXCR4[5.21e-04(29739049;29719205;29641987;29481800;29408658;29257331;29235568;29117108;29063978;28936810;28900008;28739729;28629469;28621237;28549801;28515923;28176874;28123896;28061765;27835898;27835894;27798120;27668882;27657827;27575851;27517626;27489286;27481098;27409174;27330310;27314328;27212031;27176720;27160279;27085904;27059706;26998092;26985708;26959057;26745593;26744523;26727921;26678887;26552750;26406413;26318034;26078947;26020117;25987090;25960214;25899003;25884903;25866254;25846512;25735334;25669980;25580640;26978017;25413075;25396735;25359494;25301403;25280997;25232565;25150783;24804700;24647809;24629239;24532280;24395653;24375277;24342323;24330809;24255072;24238971;24189753;24085800;23885267;23766363;23758411;23744532;23730510;23708747;23544165;23385555;23313233;23245395;23192271;23188729;23098564;23071744;22970002;22959209;22923991;22888250;22871210;22689289;22639890;22433494;22409183)], ATG2A[5.78e-04(?)], ACKR2[5.86e-04(?)], KLC4[5.96e-04(?)]
STAD	45	18	PTH2[8.78e-15(?)], OR4C3[3.63e-08(?)], ZNF721[7.39e-08(?)], ADRA1A[3.69e-07(?)], MAP2K7[4.33e-07(?)], TTK[4.89e-07(?)], PRNP[2.85e-06(?)], BNC2[5.56e-06(?)], SLC35G3[7.86e-05(?)], HLA-B[8.88e-05(?)], TLR4[1.15e-04(28531216;28093329;25371568;17645528)], BEST3[1.36e-04(?)], SLITRK6[1.62e-04(?)], SLCO1B3[1.63e-04(?)], STAT3[1.75e-04(27930339;22834702;22581828;15730617;15682485;14996748)], AQP2[2.61e-04(?)], TCEAL6[2.80e-04(?)], ZC3H4[3.03e-04(?)]
MEL	41	17	PPP6C[2.10e-21(26868000;25857817;25486434;24755198;26263704;24341237;24336958;23729733;22842228;22817889)], MAP2K1[3.01e-16(29461977;28881731;26913480;26684394;26673799;26343386;26018731;23639941;23444215;23174022;22197931;22105811;21726664;20526349)], STK19[1.56e-15(27184836;25857817;26263704;24341237;22817889)], GNA11[1.52e-10(29738114;29726589;29689622;29570931;29490280;29371009;29206651;29059311;28982892;28881731;28809862;28700778;28594900;28486107;28444874;28409567;28399339;28229253;28228113;28223438;28203054;28074614;28018010;27934878;27914687;27745836;27660484;27507190;27499153;27486988;27354579;27348266;27273450;27239460;27218826;27148356;27123562;27117140;27116551;27089179;27058448;27044592;26994139;26991400;26825879;26791842;26769193;26744134;26743513;26743478;26683228;26601868;26403583;26397223;26275246;26217306;26113083;26086698;26084293;26076063;25976133;25769001;25764247;25695059;25653058;27308390;27188223;25526026;25413220;25399693;25361747;25315378;25304237;25113308;25030020;24994677;24970262;24938562;24899684;24882516;24842760;24755198;24713608;24697775;24563540;24423917;24345920;24274719;24141786;24077403;23981010;23975010;23887304;23877823;23825798;23778528;23752084;23714557;23685997;23478236)], PDE1A[8.29e-07(22045655)], COL21A1[2.35e-06(20667089)], TCP10L2[8.27e-06(?)], OR4M2[1.51e-05(?)], DSG3[2.32e-05(11422052)], A1CF[3.67e-05(?)], SELP[3.72e-05(23648484)], TCHHL1[4.08e-05(?)], CLIP1[9.97e-05(?)], PRDM7[1.54e-04(?)], GUCA1C[1.71e-04(?)], MUC13[1.81e-04(?)], PYHIN1[2.87e-04(25199457)]
BLCA	35	16	CDKN1A[1.11e-12(28802642;23571005)], KDM6A[3.91e-08(29573965;28339163;23887298)], RARG[8.89e-08(?)], PSG1[1.63e-07(?)], ZNF624[1.65e-07(?)], ERCC2[9.69e-07(28802642;27479538;25096233)], CSPG5[3.91e-06(?)], STK39[1.91e-05(?)], RXRA[2.17e-05(26008846)], TXNIP[5.80e-05(?)], PCDHGA7[9.39e-05(?)], KLF5[1.00e-04(28915599)], UGT1A9[1.08e-04(?)], ZFP37[1.42e-04(?)], TAS1R3[2.09e-04(?)], ZNF513[2.80e-04(?)]
UCEC	42	16	MYCN[6.07e-11(23167388;14654551;11433525)], ESR1[5.10e-09(29546395;28578502;27527851;27160768;27018308;26957478;26594762;26431491;26330482;25884434;25546926;25437045;25048628;24337234;24023309;23843231;23624782;23593326;23319822;23019147;22633539;22404101;21543766;21472251;21272446;20381444;20018910;19438492;19319135;18990228;18923163;18788074;18720455;18403104;16707768)], GIGYF2[8.62e-09(?)], ARID5B[2.15e-08(27346418;23636398)], CTNND1[1.86e-06(?)], CCND1[7.15e-06(29232554;28408839;27831653;27648123;27349856;27105504;26366417;26353976;25546926;25214561;24779718;24337234;24126431;23733133;23731275;21454826;16569247;15069681;12955092;10473073)], METTL8[1.05e-05(?)], FOXA2[3.27e-05(29546371;29442045;28940304;27538367;25994056;22945641)], MYO10[4.00e-05(18783612)], DICER1[7.58e-05(28529604;28459098;28381177;23680357;23392577;22252463;21425145)], SLC6A2[8.32e-05(29693365)], BCAN[1.16e-04(?)], LZTR1[1.24e-04(?)], HEPH[1.35e-04(?)], INPP4A[1.79e-04(?)], SGK1[3.74e-04(28177128;22911820)]
BRCA	33	14	GATA3[5.08e-41(29662164;29609951;29593425;29546532;29535312;29510139;29462945;29435983;29431200;29416660;29408697;29358704;29351903;29262572;29207126;29202657;29123100;29053396;28966727;28965624;28945747;28884749;28810293;28805661;28789340;28752189;28722108;28703335;28693516;28690657;28611201;28581515;28580595;28574279;28514748;28428285;28423734;28394898;28351929;28288473;28273452;28258171;28211079;28078827;28077797;28066512;28038704;28027327;27997592;27917009;27904775;27900363;27867016;27829216;27809618;27666519;27654269;27588951;27556500;27556158;27514395;27473079;27356755;27354564;27338760;27283966;27184484;27154416;27093921;27041579;27018307;26998104;26960396;26922637;26907767;26852374;26825466;26772397;26768031;26730200;26719157;26682631;26657142;26648682;26637396;26603012;26510790;26486740;26467651;26465236;26451490;26428280;26313026;26249178;26160249;26028330;26008846;25994056;25906123;25850943)], MAP3K1[7.40e-23(29559730;29372690;29371908;29339359;29296238;29139094;28985766;28757652;28672935;28608266;28580595;28491135;28408616;28344865;28178648;28029147;28027327;27572905;26920143;26803517;26770289;26759750;26695891;26458823;26094658;25798844;25529635;24993294;24759887;24743323;24595411;24386504;24340245;24253898;24218030;24177593;23634849;23577780;23544012;23225170;23000897;22993404;22965832;22910930;22722202;22722201;22722193;22532573;22452962;21996731;21

			791674;21748294;21475998;21445572;21415360;21197568;21118973;20809358;20690207;20605201;20554749;19887619;19843670;19656774;19617217;19607694;19232126;19094228;19092773;19088016;19028704;18973230;18785201;18612136;18437204;18355772;17997823;17529967)], CBFB[7.99e-13(28077088;28027327;26870154;26643573;22722202;22722193;9586906)], MAP2K4[2.53e-10(28491135;28446401;28344865;28027327;27792260;26907767;26249178;25086928;24194916;22722193;22522925;19593635;19404734;15578079;12097290;11754110)], FCRL5[1.95e-06(?)], TBX3[3.17e-06(29344954;28238063;28215225;27632063;27553211;27100732;26920143;26579496;26451490;26249178;26215676;25552398;25343378;23733266;23624936;22722201;22535523;22532574;22039763;21098263;20942798;21779450;19858224;19828084;19403417;19218121;18245468;18025091;16049973;15781639;15289316;11255752)], KLF4[1.07e-05(29614984;29552326;29322784;29200954;29185119;29133945;28988130;28656310;28565864;28423718;28422735;28289232;28167342;28068319;27721402;27609189;27590511;27502039;27323810;27300169;27109463;27082853;26998096;26840086;26825466;26729194;26657485;26459242;26420673;26356142;26191205;26110566;26053033;25879941;25834779;25819032;25789974;25652398;25616642;25481840;25428807;25417726;25368523;25220908;25202123;25127259;25122612;24824039;24675390;24532790;24531713;24386275;24088818;24037901;23974095;23770845;23737434;23451207;23384942;23376074;23374354;23019226;22908280;22751119;25436680;22528804;22489015;22389506;22037779;21750654;21674249;21586797;21518959;21263130;21261996;21242971;20937839;20356845;19503094;19276356;18376139;17908689;17472751;17308127;16244670;15102675;11551969)], FOXA1[1.09e-05(29416660;29396764;29358704;29180470;29123100;28943920;28884749;28867731;28865492;28816236;28789340;28756535;28658208;28534958;28514748;28455227;28361702;28350011;28336670;28273452;28270510;28215225;27997592;27959926;27926873;27835577;27185372;27791031;27672107;27524420;27514395;27499099;27496708;27473079;27390128;27378691;27284343;27233940;27212698;27197147;27103403;27062924;27045898;27034986;27005559;26926684;26919034;26708273;26541755;26537518;26527523;26510790;26476779;26451490;26431101;26404658;26363213;26298189;26260807;26160249;26008846;25995231;25994056;25762479;25755696;25752574;25716347;25707489;25652398;25531315;25435372;25422910;25415051;25264199;25248036;25234841;25223786;25175082;25155268;25145671;25122612;25100862;25071007;25016694;24962896;24891455;24887547;24830797;24802759;24758297;24639548;24596378;24596370;24564526;24549642;24528009;24484401;24434785;24415069;24392136)], TPRX1[2.41e-05(?)], ZFP36L1[5.67e-05(19146866;17855657)], FAM47C[1.11e-04(?)], GLA[1.40e-04(29689288;28464803;28212442;27009385;25980823;25335329;24754877;24289581;22516725;21475864;16264182;15763439;15607568;15208499;15138577;15138562;14521914;12810158;12684675;12538085;11291069;10780877;10699943;9218004;8669881;7491296;8375111;8210965;8319825;8435199;1399132;2216462;3288258;3702424;6601977)], HIST1H3B[1.41e-04(?)], NCOR1[2.17e-04(27699256;27499907;26920143;26219265;25670202;24563328;22722201;21731475;20003447;19781322;19183483;19122196;18768663;17130524;16886664;16609009;16529049;16019133;15225781;15225779;12684393;12124798)]
HNSC	24	9	AJUBA[8.10e-08(29053175;28126323;29034103;25303977)], NOTCH1[2.95e-07(29489439;29340043;29331751;29232766;29146722;29068587;29053175;28195818;27965308;27595504;27380877;27117272;27035284;27028310;26927514;29034103;25836654;25633867;25588898;25580884;28324520;25440877;25303977;25275298;25234595;24787294;24670651;24667986;24292195;24277457;24001612;23750501;23714515;23645351;23607916;22773520;21798897;21798893;20175927;20127005;19550121)], EPHA2[6.94e-07(24864260;22455776;21955398;18425361)], ZNF750[4.76e-06(26949921)], NSD1[3.03e-05(29636367;29340043;29213088;29053175)], PEG3[5.65e-05(?)], CD248[5.87e-05(?)], ADCY8[1.52e-04(?)], FCRL4[1.68e-04(?)]
LAML	17	7	FLT3[5.79e-61(29721667;29716633;29696374;29692343;29688850;29682194;29665898;29664232;29663558;29654398;29654265;29643943;29625580;29624746;29573577;29563537;29556023;29551027;29541391;29534404;29530994;29507660;29505696;29491461;29487059;29472722;29472720;29472718;29463564;29463558;29437468;29431743;29416774;29408852;29384595;29372308;29343975;29339551;29336115;29330746;29310020;29309772;29306105;29304116;29296935;29286103;29286055;29274134;29262547;29257272;29254227;29249819;29231051;29227476;29212189;29209600;29206680;29193057;29188605;29187377;29172276;29166740;29166738;29142066;29100302;29090521;29080039;29079128;29074603;29069784;29059168;28989589;28980058;28978861;28978821;28967922;28940816;28933735;28923882;28914261;28895560;28893624;28884855;28883285;28883284;28882949;28881711;28858244;28851457;28841206;28836868;28835438;28830460;28823257;28810324;28799432;28793301;28767575;28753595;28748750)], SRSF2[1.30e-12(29721207;29549983;29549529;29472724;29309772;29249818;29181548;29148089;28953917;28751771;28555081;28255022;28152414;28054536;27137476;27486981;27256388;27135740;27023522;26849014;26848861;26848006;26820131;26812887;26799334;26542416;26514544;26115659;25553291;25533824;25445211;25412851;25220401;24989313;24970933;24923295;23996481;23645565;23558522;23349007;22823977;22773603;22722453;22431577;22389253)], NPM1[1.77e-12(29721667;29696374;29665898;29661468;29625580;29624746;29622865;29573577;29563537;29556023;29541391;29534404;29530994;29519869;29491461;29472722;29441887;29435155;29423110;29408852;29402726;29343273;29330746;29310020;29286103;29283500;29274134;29254789;29254227;29249819;29238371;29224316;29221119;29219176;29193057;29188605;29172276;29166740;29166738;29157973;29111347;29090521;29079128;29069784;28978861;28971903;28923882;28920929;28882949;28841206;28836868;28835438;28830460;28823257;28753595;28740552;28710806;28698788;28679652;28618016;28574487;28569789;28475434;28473620;28471807;28456748;28452374;28411256;28407515;28384310;28380436;28368672;28362701;28341738;28318150;28315400;28297624;28294102;28245376;28219218;28210583;28167452;28163010;28152414;28111462;28106537;28090023;28070990;28055106;28017614;27995876;27994664;27983727;27906185;27899775;27865970;27864740;27841873;27595757;27581357)], WT1[1.14e-09(29739109;29573577;29563537;29551027;29452230;29434724;29408852;29407184;29388165;29386195;29306105;29296935;29286103;29240258;29227476;29166742;29166740;29152069;29096332;29070097;29041012;28994041;28980766;28954349;28949050;28923882;28846953;28830889;28830460;28810324;28567073;28521413;28477011;28475434;28454430;28400619;28395566;28321480;28211167;28163010;28159598;28139337;28125133;28114959;28114350;28074068;28024475;27974109;27941286;27893200;27889611;27866185;27821287;27801325;27636548;27694926;27659531;27612989;27575502;27544285;27512765;27499136;27478011;27359055;27342485;27285584;27252512;27225156;27197573;27149388;2711

			1858;27062340;27055875;26992216;26970379;26941285;26893773;26725349;26644203;26531831;26520650;26519872;26499507;26451309;26284582;26234722;26224397;26221900;26138637;26137066;26054017;26046002;26012361;25956466;25932436;25890432;25841655;25835542;25807502;25805812)], TET2[1.32e-06(29721207;29702001;29664232;29661468;29624746;29573577;29491461;29472724;29343972;29331774;29309772;29306105;29285580;29274134;29249818;29219176;29150453;29148089;29029424;28992762;28978861;28978821;28923882;28823558;28823257;28642303;28555081;28407691;28400619;28315400;28297624;28255022;28242787;28167452;28152414;28074068;28070990;28053194;27881874;27821287;27658049;27486981;27477909;27449473;27424808;27391574;27389053;27359055;27352183;27285584;27215596;27055875;27050425;27023522;26984174;26941285;26876596;26849014;26848006;26828965;26812887;26789100;26725349;26703470;26666714;26666262;26586702;26568194;26524018;26414667;26375248;26277372;26234722;26118500;25956466;25886910;25873173;25700647;25699704;25601757;25482556;25473623;25426838;25412851;25381129;25311741;25276435;25246247;25200248;25022553;24989313;24986689;24970933;24898826;24859829;24816242;24778653;24726781;24659740;24609756)], ROCK2[1.39e-05(?)], CEBPA[1.60e-05(29624746;29622865;29573577;29541391;29534404;29515250;29483711;29435155;29431622;29402726;29343483;29310020;29306105;29286103;29238371;29193057;29188605;29032147;29025912;28978861;28923882;28900037;28895127;28882949;28830460;28753595;28745571;28663557;28504718;28473620;28452374;28380436;28357685;28341738;28299657;28250006;28249600;28210006;28203345;28186500;28179278;28144729;28090023;28074068;28070990;27899775;27812248;27694926;27626217;27512765;27367478;27359055;27350755;27288520;27285584;27129260;27062340;27040395;27034432;27023522;27012040;26992835;26940274;26876264;26802049;26725349;26721895;26708912;26693794;26676635;26586702;26537612;26496024;26488113;26466372;26460249;26450903;26419342;26408402;26386075;26377688;26376842;26375248;26374622;26239249;26234722;26174629;26167872;26071459;26053097;26031527;26025484;25987038;25976969;25938608;25932436;25794001;25787321;25732229;25659730)]
OV	17	5	BRCA1[3.35e-07(29731958;29712865;29707124;29707112;29673794;29671401;29665859;29661778;29660759;29659587;29618939;29617664;29617652;29615458;29610032;29606854;29602379;29580810;29567272;29566657;29558274;29550970;29550896;29545475;29534594;29522266;29511213;29493783;29487695;29483665;29479477;29470806;29464354;29464067;29460478;29453736;29447163;29445031;29429842;29428045;29427345;29409816;29409476;29405995;29404838;29383094;29371908;29368626;29367421;29361001;29356917;29348823;29344385;29340030;29335712;29319983;29310832;29307397;29302806;29298688;29297111;29286205;29282716;29278246;29275357;29273311;29271107;29270046;29262038;29259228;29254167;29252925;29236593;29236234;29215753;29203787;29189915;29170526;29168504;29164969;29153097;29146938;29143969;29138572;29137324;29133618;29132681;29121898;29109859;29096890;29094253;29084914;29082457;29081841;29063517;29061375;29058922;29054568;29054544;29053726)], CDK12[7.56e-06(28950147;27905519;27662623;27241520;26247403;25429106;24554720;24240700;22012619;21720365)], ZBTB18[2.41e-05(?)], ADGRE3[5.11e-05(?)], HLA-G[1.39e-04(26942060;26687271;24987709;23228395;21858813;20001801;19910891;19692629;19660509;18498645;17846022;17681474;15589578;8612864)]
LUAD	29	4	STK11[1.35e-14(29575851;29540834;29535211;29337640;29307989;29279706;29219616;29198084;29191602;29168346;29066508;28914263;28911955;28884744;28754670;28652249;28619094;28538732;28435024;28413430;28387316;28336552;28205554;28145643;27923066;27687306;27565922;27467949;27299180;27218826;27151654;27121209;26960398;26917230;26833127;26829311;26625312;26599269;26477306;26463840;26420428;26350096;26124082;26119936;26087898;26069186;26066407;25982285;25969368;25964588;25695224;25634010;25477232;25444907;25278450;25122068;25036236;25031567;24828662;24482041;24468202;24448687;24297535;24236184;24086281;24077454;24054548;23276293;23047306;22768234;22590557;21532627;20057966;19661141;19483050;19353596;19176640;19165201;17711506;17676035;17216011;16912160;16580634;15639728;15021901;12097271;11212897;10508479)], KCNB1[7.22e-05(?)], HGF[8.67e-05(29717265;29558956;29371783;29253515;29187584;29168346;29125233;29063069;29058790;29050231;28944826;28940757;28938541;28903317;28843992;28559461;28554854;28485480;28469968;28416482;28404966;28373408;28332364;28260071;28192876;28164089;28164087;28121629;28096505;28064454;28061464;28038979;27873490;27863726;27803065;27716616;27566197;27525306;27422710;27374174;27133742;27071409;27015549;26983447;26923077;26919104;26919096;26870265;26811313;26719536;26701889;26695082;26639195;26579470;26542886;26463323;26416301;26153496;26138771;26115510;26063323;26045672;26038598;26011628;25992382;25992367;25936889;25925948;25919140;25889721;25877780;25806289;25798262;25757678;25640943;25575814;25543140;25522765;25504327;25502629;25449774;25444907;28548075;25314153;25266653;25249428;25130970;25057941;24983493;24959087;24952482;24867356;24828661;24722154;24710956;24687921;24628993;24378092;24355409;24327519)], TARS2[1.46e-04(?)]
DLBCL	7	3	CD79B[5.33e-10(29734251;29641966;29289361;29262531;29245897;28993276;28884033;28864640;28841204;28664939;28619981;28479318;28319526;28153771;28073005;28011673;27923841;27915469;27048211;26773040;26647218;26639163;26515759;25991819;25925619;25708834;25623213;25612555;25529125;25391967;25275047;25030036;24327543;24240734;23372794;23285191;22961721;21173233;20054396;16531332;12651942;11396639)], GNA13[2.74e-05(28302137;27980305;26989201;26819451;26773040;26616858;26608593;25991819;25274307;23699601;23292937;23143597;22343534)], MEF2B[4.51e-05(29309299;28851661;27166360;26245647;26089142;25769544;23974956;23292937;22343534;21796119)]
LGG	10	3	ATRX[1.51e-10(29111096;29091765;29027701;28980701;28419269;28392842;27758882;26508407;25664944;24710217;23373454)], FUBP1[3.39e-06(29606613;23373454)], CIC[4.10e-05(23373454)]
PAAD	21	3	CHRD[1.20e-06(?)], CDC27[4.28e-05(25912578)], TGFBR2[1.44e-04(29393426;28809762;28373289;26279302;26255562;25791160;23690952;23378339;23237571;23103869;22523087;17031113;12615714;9850059;9598801)]

LUSC	13	2	PDYN[1.78e-05(?)], CYP11B1[3.67e-05(27347096)]
MED	8	2	SMO[2.38e-15(29531057;29378965;29348431;29274272;29208776;29055107;28923910;28873303;28833911;28716052;28618224;28605510;28487292;27785591;27495899;27236920;27069629;26960983;26891329;26691947;26633513;26450969;26371509;26323341;26286140;26169613;26113054;26080084;25859932;25636740;25505589;25485584;25484239;25376612;25355313;25306392;25131638;24994715;24973920;24951114;24871706;24276242;24068730;23872071;23671675;23662017;22966790;22923130;22869526;22851551;22452947;24451804;22084163;21618411;21501498;21325292;21143927;21123452;20881279;20524040;20493695;20386868;20024066;19726788;19701203;18826648;18502968;18288402;17413002;17017853;16707575;16618744;15806168;12192414;11965540;10984056;10564585)], DDX3X[4.64e-14(29582169;29222110;27180681;27058758;26290144;25724843;24608801;22832583;22820256;22722829)]
NB	2	2	ALK[4.07e-18(29660984;29642598;29638111;29600072;29559559;29556564;29555900;29535836;29515255;29505958;29492199;29466695;29455642;29441070;29380702;29378002;29374774;29371588;29357780;29321660;29317532;29296183;29290991;29203817;29184034;29084134;29081033;29069774;29027209;29018329;28915622;28915608;28871274;28800395;28756644;28676342;28674118;28666189;28665006;28662353;28604107;28602975;28546523;28521285;28458126;28425916;28423360;28350380;28338501;28326957;28178969;28163672;28139105;28069802;28030793;27997549;27879258;27830764;27707976;27684973;27655666;27604320;27573755;27483357;27471553;27285993;27179218;27165366;27076624;27013922;27009859;27009842;26986945;26925973;26893860;26835380;26829053;26826611;26794043;26750252;26735175;26687816;26633716;26630010;26616860;26539795;26517508;26503946;26468446;26388126;26309160;26299615;26206265;26122839;26067621;26059187;26005112;25979929;25950466;25925003)], PTPN11[7.12e-08(29189514;28947394;28329685;27655895;27362227;24628801;23813970;23334666;21548061;20461756;18328949;16631468;16518851;15604238)]
PRAD	25	2	LCTL[2.85e-06(?)], AR[5.33e-05(29734647;29733466;29725990;29721186;29716963;29712835;29707651;29707137;29699261;29700003;29695920;29693622;29693262;29691406;29686105;29684818;29682197;29682196;29670000;29668110;29666833;29666783;29666302;29665325;29662238;29658587;29641940;29633296;29632047;29618577;29594945;29588330;29581250;29579692;29574703;29572225;29571584;29568400;29566488;29562689;29562494;29555975;29555663;29552052;29542849;29541371;29540675;29535823;29530947;29527701;29523594;29508425;29490263;29488772;29477539;29477409;29474983;29473182;29464071;29463549;29462692;29456113;29453313;29449534;29448139;29444261;29441606;29438990;29438723;29436611;29431615;29429990;29427323;29425687;29423094;29421751;29417861;29402932;29398263;29395951;29388326;29386530;29383186;29383141;29383125;29381490;29379164;29378906;29372107;29371946;29367197;29366632;29360794;29359890;29358171;29353883;29346776;29339080;29334357;29332354)]
CLL	5	1	MYD88[1.34e-09(29286214;29242635;28994094;28892161;28664939;28424451;28399885;28255015;28241765;27959900;27742074;27633522;27491692;27198719;27060156;26630574;26482097;26454445;26316624;26230596;26181643;26136429;26053404;25696845;25605254;25480502;24943833;24943832;24782504;24767771;24103588;23935380;23684423;23665546;23477936;23419703;23246696;23178471;22150006;19050243)]
CESC	3	1	KRTAP4-11[1.45e-05(?)]
MM	7	1	ZNF717[9.14e-07(?)]
RHAB	1	1	SMARCB1[2.52e-07(29696793;29670784;29602769;29528755;29512865;29428974;29397238;29339179;29324471;29316066;29280680;29271065;29258531;29228610;29110337;28966010;28945250;28824165;28812319;28789476;28777153;28714904;28521298;28434767;28382842;28338502;28109176;28108836;28084340;27966820;27783942;27734605;27695363;27639430;27467095;27380723;27356182;27338635;27267444;27218413;27095948;27092963;27013922;26920892;26755072;26646792;26578851;26567940;26557502;26407663;26370283;26363008;26342593;26234633;26073604;25751458;25638158;25494491;25479928;25312828;25307865;25274825;25268025;25262118;25246033;25200863;25169151;25114695;25053104;25018128;25016934;24972932;24853101;24585572;24555876;24503755;24423609;24418192;24327545;24308011;24287458;24141276;24075062;23880166;23364536;23190500;23154773;23084579;23074045;23060122;22997201;22814326;22201954;22180295;21934399;21775180;21724432;21566516;21417895;21412926)]
THCA	8	1	RPTN[2.17e-05(?)]
ALL	0	0	-
AT	1	0	-
CARC	1	0	-
ESCA	6	0	-
GBM	10	0	-
KICH	1	0	-
KIRP	2	0	-
LIHC	2	0	-
LUSE	1	0	-
LB	1	0	-
STS	0	0	-

Note: The values in square brackets are p-values by WITER. The significant genes are determined according to the p-values (FDR<0.1). The number in the brackets are the PubMed ID of papers co-mentioning the disease name and gene symbol according to search API in PubMed database, http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term="DiseaseNames(inlcuding homonymies)"[tiab]%29+AND+"GeneSymbol

(including RefSeq mRNA IDs)" [tiab]. For genes with over 100 papers, only the most recent 100 papers are shown. "-" denotes no unique significant genes. "a": number of total significant genes. "b": number of unique significant genes. ALL: Acute lymphoblastic leukemia, AT: Astrocytoma, BLCA: Bladder Urothelial Carcinoma, BRCA: Breast invasive carcinoma, CARC: Carcinoid Cancer, CESC: Cervical Carcinoma, CLL: Chronic lymphocytic leukemia, COAD: Colon adenocarcinoma, DLBCL: Diffuse large B-cell lymphoma, ESCA: Esophageal carcinoma, GBM: Glioblastoma Multiforme, HNSC: Head and Neck Squamous Cell Carcinoma, KICH: Kidney chromophobe carcinoma, KIRC: Kidney Clear Cell Carcinoma, KIRP: Kidney Papillary Cell Carcinoma, LAML: Acute myeloid leukemia, LB: B-cell lymphomas, LGG: Low Grade Glioma, LIHC: Liver Hepatocellular carcinoma, LUAD: Lung Adenocarcinoma, LUSC: Lung Squamous Cell Carcinoma, LUSE: Small cell lung carcinoma, MED: Medulloblastoma, MEL: Melanoma, MM: Multiple myeloma , NB: Neuroblastoma, OV: Ovarian serous cystadenocarcinoma, PAAD: Pancreatic Adenocarcinoma, PRAD: Prostate Adenocarcinoma, RHAB: Rhabdoid tumor, STAD: Stomach Adenocarcinoma, STS: Soft Tissue Sarcoma, THCA: Thyroid Carcinoma, UCEC: Uterine corpus endometrial carcinoma

Table S6: Sample size and variant number of 34 cancer datasets

Full Name	Abbreviation	Sample Size	Variant Number	Ratio	[a]Estimated Sample Size
Lung Adenocarcinoma	LUAD	394	106088	269.26	212
Melanoma	MEL	289	94715	327.73	174
Head and Neck Squamous Cell Carcinoma	HNSC	407	59389	145.92	391
Breast invasive carcinoma	BRCA	1093	54932	50.26	1134
Lung Squamous Cell Carcinoma	LUSC	175	53384	305.05	187
Stomach Adenocarcinoma	STAD	244	42113	172.59	330
Uterine corpus endometrial carcinoma	UCEC	255	38733	151.89	375
Bladder Urothelial Carcinoma	BLCA	142	33623	236.78	241
Colon adenocarcinoma	COAD	244	31461	128.94	442
Kidney Clear Cell Carcinoma	KIRC	484	28199	58.26	978
Ovarian serous cystadenocarcinoma	OV	480	27946	58.22	979
Glioblastoma Multiforme	GBM	365	21601	59.18	-
Esophageal carcinoma	ESCA	160	18935	118.34	-
Prostate Adenocarcinoma	PRAD	420	16618	39.57	-
Multiple myeloma	MM	205	10663	52.01	-
Low Grade Glioma	LGG	227	9620	42.38	-
Small cell lung carcinoma	LUSE	30	8377	279.23	-
Pancreatic Adenocarcinoma	PAAD	233	7674	32.94	-
Liver Hepatocellular carcinoma	LIHC	150	7614	50.76	-
Kidney Papillary Cell Carcinoma	KIRP	111	7513	67.68	-
Cervical Carcinoma	CESC	37	6104	164.97	-
Thyroid Carcinoma	THCA	325	6095	18.75	-
Diffuse large B-cell lymphoma	DLBCL	56	5742	102.54	-
Neuroblastoma	NB	351	5042	14.36	-
Acute myeloid leukemia	LAML	196	4052	20.67	-
Medulloblastoma	MED	331	3483	10.52	-
Chronic lymphocytic leukemia	CLL	223	3455	15.49	-
Carcinoid Cancer	CARC	54	1650	30.56	-
Kidney chromophobe carcinoma	KICH	65	1263	19.43	-
B-cell lymphomas	LB	26	1168	44.92	-
Acute lymphoblastic leukemia	ALL	55	613	11.15	-
Rhabdoid tumor	RHAB	32	240	7.5	-
Soft Tissue Sarcoma	STS	15	117	7.8	-
Astrocytoma	AT	42	95	2.26	-

Note: a: the estimated sample size is for detecting 30 significant genes by WITER. –: the sample sizes are not estimated because of the unreliable ratio derived in small samples.

Table S7: The overlapped significant genes among multiple cancers detected by WITER

Cancer	ALL	AT	BLCA	BRCA	CARC	CLL	CESC	COAD	DLBCL	UCEC	ESCA	GBM	HNSC	KICH	KIRC	KIRP	LAML	LIHC	LGG	LUAD	LUSE	LUSC	LB	MM	MED	MEL	NB	OV	PAAD	PRAD	RHAB	STS	STAD	THCA
ALL	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AT	0	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BLCA	0	0	35	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BRCA	0	0	8	33	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CARC	0	0	1	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CLL	0	0	2	4	1	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CESC	0	0	2	1	0	0	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
COAD	0	0	10	8	1	2	1	57	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DLBCL	0	0	2	2	1	2	0	3	7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
UCEC	0	0	13	14	1	2	2	14	2	42	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ESCA	0	0	4	3	1	2	1	5	2	4	6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GBM	0	0	5	5	1	1	1	4	2	5	2	10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
HNSC	0	0	7	5	1	1	2	4	1	6	4	3	24	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
KICH	0	0	1	1	1	1	0	1	1	1	1	1	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
KIRC	0	0	4	4	1	1	2	3	1	6	3	3	5	1	34	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
KIRP	0	0	1	1	1	1	0	1	1	2	1	1	1	1	2	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
LAML	0	0	4	4	1	2	0	4	3	4	2	2	1	1	2	1	17	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
LIHC	0	0	2	1	1	1	0	2	1	2	2	1	1	1	1	1	1	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
LGG	0	0	3	4	1	1	1	4	1	5	2	5	2	1	3	1	3	1	10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
LUAD	0	0	9	6	1	2	1	17	3	9	5	5	7	1	3	1	4	2	2	29	-	-	-	-	-	-	-	-	-	-	-	-	-	-
LUSE	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-
LUSC	0	0	6	4	1	1	2	4	2	5	3	5	9	1	5	1	1	1	3	7	1	13	-	-	-	-	-	-	-	-	-	-	-	-
LB	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-	-	-	-	-	-	-	-	-	-	-
MM	0	0	5	2	1	2	0	4	3	3	2	4	2	1	1	1	4	1	2	4	1	3	1	7	-	-	-	-	-	-	-	-	-	-
MED	0	1	4	3	1	1	1	4	1	4	3	2	3	1	3	1	1	2	2	4	1	3	1	1	8	-	-	-	-	-	-	-	-	-
MEL	0	0	9	7	1	3	1	15	3	8	5	6	6	1	4	1	5	2	4	16	1	7	1	5	4	41	-	-	-	-	-	-	-	-
NB	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	-	-	-	-	-	-	-
OV	0	0	8	7	1	2	1	8	3	9	3	4	5	1	2	1	3	1	2	9	1	4	1	4	3	7	0	17	-	-	-	-	-	-
PAAD	0	0	2	3	1	3	0	17	2	2	3	1	2	1	1	1	2	1	1	10	1	1	1	2	1	11	0	2	21	-	-	-	-	-
PRAD	0	0	5	4	1	1	1	17	1	6	3	4	2	1	3	1	2	2	4	11	1	3	1	2	3	13	0	2	15	25	-	-	-	-
RHAB	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-	-	-
STS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	-
STAD	0	0	11	9	1	2	1	17	3	12	6	6	5	1	6	1	3	2	4	10	1	6	1	4	4	10	0	8	3	5	0	0	45	-
THCA	0	0	4	1	0	0	1	3	1	2	1	2	3	0	2	0	1	0	1	4	0	3	0	2	1	4	0	3	0	3	0	0	4	8

Notes: ALL: Acute lymphoblastic leukemia, AT: Astrocytoma, BLCA: Bladder Urothelial Carcinoma, BRCA: Breast invasive carcinoma, CARC: Carcinoid Cancer, CESC: Cervical Carcinoma, CLL: Chronic lymphocytic leukemia, COAD: Colon adenocarcinoma, DLBCL: Diffuse large B-cell lymphoma, ESCA: Esophageal carcinoma, GBM: Glioblastoma Multiforme, HNSC: Head and Neck Squamous Cell Carcinoma, KICH: Kidney chromophobe carcinoma, KIRC: Kidney Clear Cell Carcinoma, KIRP: Kidney Papillary Cell Carcinoma, LAML: Acute myeloid leukemia, LB: B-cell lymphomas, LGG: Low Grade Glioma, LIHC: Liver Hepatocellular carcinoma, LUAD: Lung Adenocarcinoma, LUSC: Lung Squamous Cell Carcinoma, LUSE: Small cell lung carcinoma, MED: Medulloblastoma, MEL: Melanoma, MM: Multiple myeloma , NB: Neuroblastoma, OV: Ovarian serous cystadenocarcinoma, PAAD: Pancreatic Adenocarcinoma, PRAD: Prostate Adenocarcinoma, RHAB: Rhabdoid tumor, STAD: Stomach Adenocarcinoma, STS: Soft Tissue Sarcoma, THCA: Thyroid Carcinoma, UCEC: Uterine corpus endometrial carcinoma

Table S8: Rescued genes in the Breast invasive carcinoma dataset by different tools

Major tool[a]	RandomSampleID	Sig.Genes[b]	MutSigCV	oncodriveFML	WITER	20/20plus
MutSigCV (17)	1	11	-c	0	3	2
	2	11	-	0	3	3
	3	13	-	0	4	3
	4	10	-	0	5	3
	5	10	-	1	4	3
	6	10	-	0	4	4
oncodriveFML (11)	1	11	0	-	0	0
	2	6	1	-	4	2
	3	9	1	-	3	1
	4	6	1	-	3	3
	5	6	0	-	2	2
	6	8	2	-	4	2
WITER (33)	1	35	0	2	-	0
	2	36	0	0	-	1
	3	34	1	1	-	1
	4	40	0	0	-	0
	5	34	0	0	-	0
	6	35	1	0	-	1
20/20plus (38)	1	23	1	0	4	-
	2	17	2	0	6	-
	3	26	1	0	3	-
	4	23	0	0	2	-
	5	32	1	0	2	-
	6	31	0	0	1	-

a: The significant genes detected by a tool in the full dataset. b:The significant genes detected by a tool in the extracted random samples. c: This item is inapplicable.