Supplemental Material

ContrastRank: a new method for ranking putative cancer driver genes and classification of tumor samples

Rui Tian[1], Malay K Basu[1,2] and Emidio Capriotti[1,2,3].*

[1] Division of Informatics, Department of Pathology, University of Alabama at Birmingham, 619 19th St. South, 35249 Birmingham, AL (USA)
[2] Department of Clinical and Diagnostic Sciences, University of Alabama at Birmingham, 1705 University Boulevard, 35249 Birmingham, AL (USA)
[3] Department of Biomedical Engineering, University of Alabama at Birmingham, 1075 13th Street South, 35249 Birmingham, AL (USA)

*Corresponding author: Emidio Capriotti, emidio@uab.edu

Table of Content

Supplementary Methods

1. TCGA data collection and filtering

We collected TGGA whole exome sequencing data for three tumor types: colon, lung and prostate adenocarcinoma (COAD, LUAD and PRAD, respectively). The direct links to this restricted datasets available upon Data Use Certification Agreement are provided in the Supplementary Files. To avoid variability introduced by the use of multiple sequencing platforms, we used only whole-exome sequencing data obtained by Illumina platform, which represents the state of the art in the field and has lowest rate of false positives in the detection of SNVs (Quail, et al., 2012). For each tumor type, we analyzed the data extracted from the Variant Calling Format (VCF) files containing both whole-exome data from normal and tumor cells from the same patient.

Since the VCF files from Baylor College of Medicine (COAD) and those from Broad Institute (LUAD and PRAD) have been obtained with two different variant calling procedures, we adopted two alternative strategies to select germline and somatic variants, which are both important in our analysis. For the COAD samples, we used all the nsSNVs that pass Baylor College of Medicine filtering procedure and have the word "PASS" in the FILTER field of the VCF file. For the LUAD and PRAD samples, the germline variants included in the VCF files did not pass the MuTect (v.1) filter (Cibulskis, et al., 2013) used by Broad. To recover the germline variants in these samples, we filtered the VCF files selecting all the nsSNVs with average base quality (BQ) for reads supporting alternative alleles higher then 30 and fractions of reads (FA) higher than 0.05. Using this procedure, all the variants in tumor samples could be recovered including both the germline and somatic variants.

2. ContastRank benchmarking

2.1 Evaluation of the gene prioritization score

To test the performance of ContastRank in prioritizing cancer-causing genes we compared its performance against MutSigCV (Lawrence, et al., 2013). The lack of well-established "gold standard" set is one of the main issues in the evaluation of cancer-causing gene prioritization methods. To partially address this issue we assume as benchmark set three manually curated lists of cancer-associated genes. The largest list referred as "Bushman" is the union of 8 collections of cancer-related genes. This list, composed by 2,125 genes, is available online at http://www.bushmanlab.org/assets/doc/allonco_20130923.tsv (Bushman, 2013). The second is the cancer census gene list indicated as "COSMIC Census" downloaded from COSMIC database (Forbes, et al., 2011) website (http://cancer.sanger.ac.uk). This list contains 522 genes which mutations have been implicated in cancer. The latter list namely "Vogelstein" is smaller a more specific list of 125 driver genes affected by subtle mutations provided in the table S2A of the supplementary materials of a recent publication (Vogelstein, et al., 2013). This list is composed by 71 tumor suppressor genes and 54 oncogenes extracted from genome-wide sequencing studies of 3,284 tumors. Assuming these lists of genes as "true positive" and all the remaining genes as true negative we can calculate the true and false positive rates (defined in the following section) at different p-value cutoff. With this procedure we are able to estimate the ability of our method to correctly rank cancer-related genes among those with a p-value lower than a given cutoff. Accordingly, we were able to draw a receiver operating characteristic (ROC) curve and calculate the associate area under the curve (AUC) defined in the following section. Therefore we compared the performance of ContastRank and MutSigCV comparing the AUC of the obtained by the 2 methods using the Bushman, Census and Vogelstein as reference sets of cancer-related genes. The MutSigCV prioritization lists of cancer-causing genes for colon, lung and prostate adenocarcinomas have been calculated providing in input all types of somatic mutations found in the exonic regions in their relative cohorts. The list of somatic mutations has been obtained by removing from the genomic variants in tumor sample the variations detected in the associated normal sample. Using the output of MutSigCV we ranked the cancer-related genes according to their p-value. The MutSigCV cancer-related gene lists for the three tumor types are provided as supplementary files.

2.2 Cross-validation procedures

We tested the performance of ContrastRank by implementing a simple binary classifier based on the score threshold to discriminate between normal and tumor samples. To avoid overfitting, we used a 2-fold cross-validation procedure randomly splitting each dataset (COAD, LUAD and PRAD) in two subsets, calculating the scores associated to the PIGs (s_g) in one subset and scoring the samples on the second subset. Reversing this process we obtain a score for all the samples in our dataset. Finally, we select the prediction threshold that maximizes the value of the Matthews correlation coefficient (Supplementary Methods 3). We repeated this procedure 10 times and calculated the average accuracy measures for each cancer type. This cross-validation procedure (CV Identifier) has been performed keeping the normal and tumor samples with the same identifier in the same subset. A second cross-validation procedure has been used to estimate the minimum level of accuracy in the case the putative defective rates (PDRs) of the putative impaired genes (PIGs) were extracted from an unrelated subset of normal samples. Accordingly, we implemented a 2-fold cross-validation test (CV Unseen) where the matching pairs of samples are divided into two groups and the normal samples are swapped between the groups. Thus, the normal samples in the first subset are used to calculate the PDRs to score the tumor samples in the second subset, and conversely, the normal samples in the second subset are used for scoring the tumor samples in the first subset. The CV Unseen test has been used to estimate the performance of ContrastRank when matching pairs of normal and tumor samples with same identifier are in disjoint sets.

2.3 Discriminating between normal and tumor samples

To evaluate the quality of our prioritization method (ContrastRank) we compared the performances using three alternative approaches:

1. ContrastRank, which uses the top ranking genes, sorted by the score described in section 2.3 of the main manuscript.
2. ContrastLow based on the lowest ranking genes in the previous list.
3. ContrastDiff, which prioritizes the genes using the difference between their PDRs in tumor and normal TCGA samples.

The standard ContrastRank approach relies on PIGs with highest score. Thus, given a list M putative impaired genes (PIGs) $G=\{g_1,g_2,\ldots\ldots,g_M\}$ with at least one putative deleterious variant (PDV) the total exome score for ContrastRank (S_{CR}) is calculated summing all the PIG score s_g as follows:

$$S_{CR} = \frac{1}{M}\sum_{i=1}^{M} s_{g_i} H(t) \qquad \text{where} \qquad H(t) = \begin{cases} 0, & s_{g_i} \le t \\ 1, & s_{g_i} > t \end{cases} \qquad [1]$$

where t is an arbitrary cutoff.

ContastLow score (S_{CL}) is the sum of all the PIG scores s_g below an arbitrary cutoff *t*

$$S_{CL} = \frac{1}{M}\sum_{i=1}^{M} s_{g_i}\overline{H}(t) \quad \text{where} \quad \overline{H}(t) = \begin{cases} 0, & s_{g_i} > t \\ 1, & s_{g_i} \le t \end{cases} \qquad [2]$$

ContrastLow provides an estimation of the lower bound performance obtained removing the highest ranking putative impaired genes (PIGs).

In ContrastDiff method, the total score S_{CD} is given by the Eq.1 where the s_g is calculated as follows:

$$s_g = \tau_g - \pi_g \qquad [4]$$

where τ_g *and* π_g are the putative defective rates (see section 2.1 in main manuscript) calculated over the subsets of tumor and normal samples respectively. Comparing the performance of ContrastDiff with ContrastRank we evaluate the improvement resulting from the use of our prioritization score based on the binomial distribution. In the table 1 of the main manuscript the s_g cutoff has been arbitrarily set to 3 that corresponds a probability of 0.001. The performances of ContrastRank and ContastLow reported in Supplementary Tables S5, S6 and S7 have been selected in using a decreasing cutoff *t* that allows us to consider an increasing number of highest scored PIGs in the calculation of ContrastRank score (Eq. 2) and decreasing number of highest scored PIGs in the calculation of ContrastLow score (Eq. 3).

2.4 Discriminating different tumor samples

In this work we also evaluated the ability of our scoring system to discriminate between one type of adenocarcinoma from the others. For this purpose, we created randomly sampled sets composed by 50% tumor samples under study and the remaining 50% equally divided between the other two tumor types. For example, to estimate the ability of ContrastRank to discriminate colon adenocarcinoma from lung and prostate adenocarcinoma samples we build a dataset composed by 50% COAD, 25% LUAD and 25% PRAD samples.

To discriminate between different tumor types, we calculated the ContastRank score for the cancer type under study and for the mixture of the remaining tumor types. Thus, the final score for each PIG is the difference between its score for the tumor under study and the mixture of remaining tumors. Both scores are obtained comparing the gene putative defective rates (PDRs) in subset of tumors samples and the background PDR. With this procedure, important genes in the cancer type under study will have high positive value and important genes in other tumor types will have negative scores. Therefore, the total cancer exome score (S_T) for this test is calculated as follows

$$S_T = \frac{1}{M}\sum_{i=1}^{M} \Delta s_{g_i} H^*(t) \quad \text{where} \quad H^*(t) = \begin{cases} 0, & |\Delta s_{g_i}| \le t \\ 1, & |\Delta s_{g_i}| > t \end{cases} \qquad [5]$$

and Δs_g is the difference between the PIG scores calculated on the subset of tumor samples under study and the mixture of samples from other tumors.

This test is again performed using a 2-fold cross-validation procedure. Since both high positive and negative score genes are import to discriminate between tumor types, for this task we calculated the performance of ContrastRank using a list *2j* which is composed by the first *j* genes with high positive scores and the last *j* genes with high negative scores.

3. Measures of Performance

In all measures of performance (assuming that positives indicate tumor and negatives indicate either normal or an alternative tumor type), TP (true positives) are correctly predicted tumor genotypes, TN (true negatives) are correctly normal genotypes, FP (false positives) normal genotypes predicted as tumor and FN (false negatives) are tumor genotypes predicted to be normal.

Predictor performance was evaluated using the following metrics: positive and negative predicted values respectively PPV, NPV), true positive and negative rates (respectively TPR, TNR), and overall accuracy (Q_2; Eq. 5)

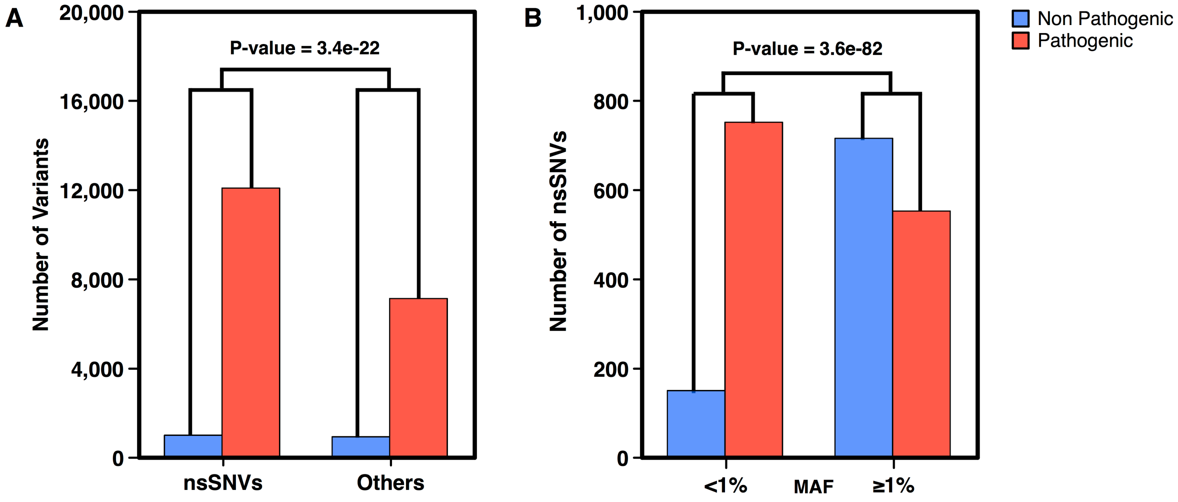
$$\begin{gathered} PPV = \frac{TP}{TP+FP} \quad TPR = \frac{TP}{TP+FN} \\ NPV = \frac{TN}{TN+FN} \quad TNR = \frac{TN}{TN+FP} \\ Q_2 = \frac{TP+TN}{TP+FP+TN+FN} \end{gathered} \qquad [6]$$

We also computed the Matthew's correlation coefficient *C* (Eq. 6) as:

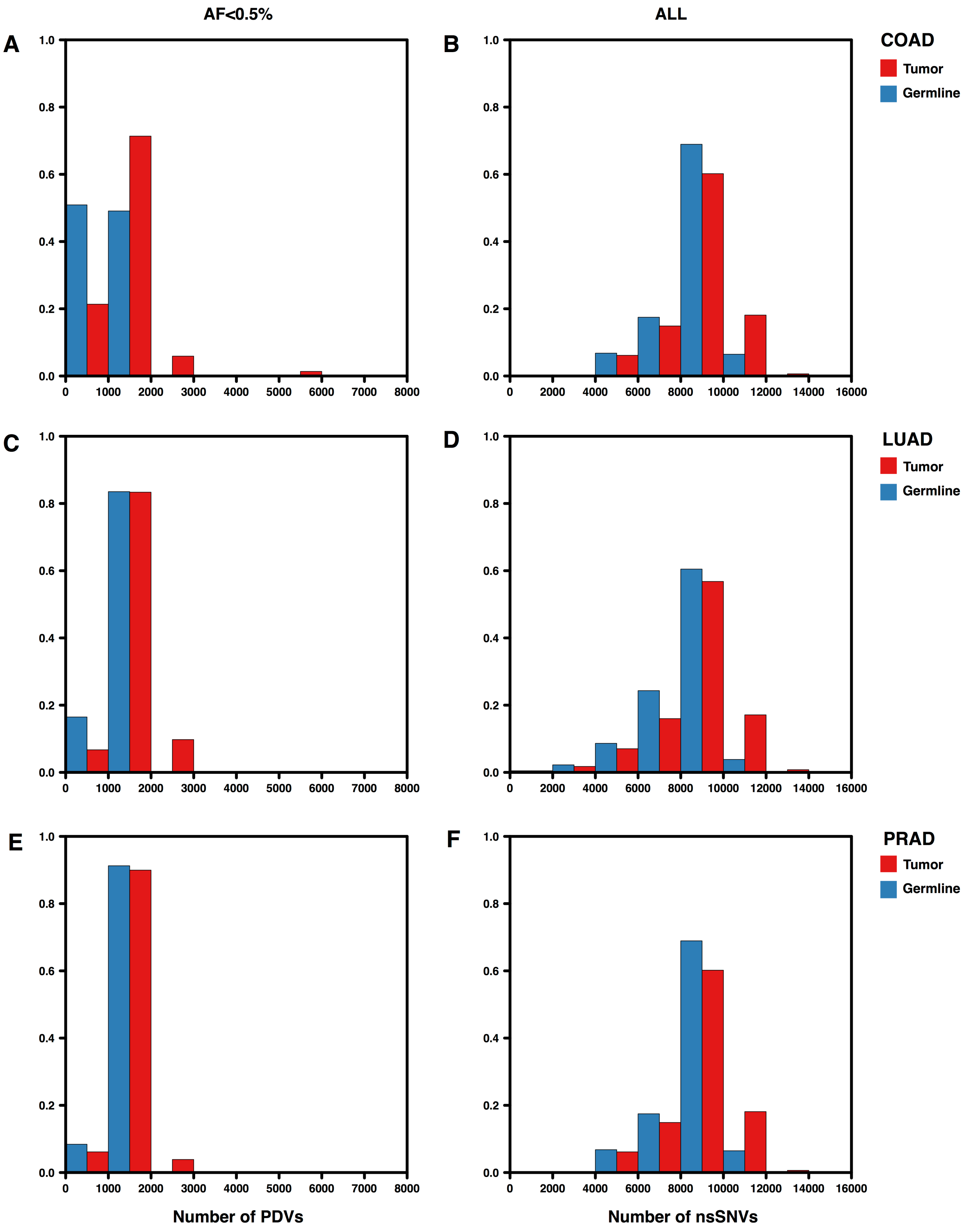
$$C = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \qquad [7]$$

We also calculated the area under the receiver operating characteristic (ROC) curve (AUC). This is calculated by plotting the True Positive Rate as a function of the False Positive Rate (1-TNR) at different score thresholds of predicting a genotypes as tumor. All the measures of performance have been calculated using ROCR package (Sing, et al., 2005).

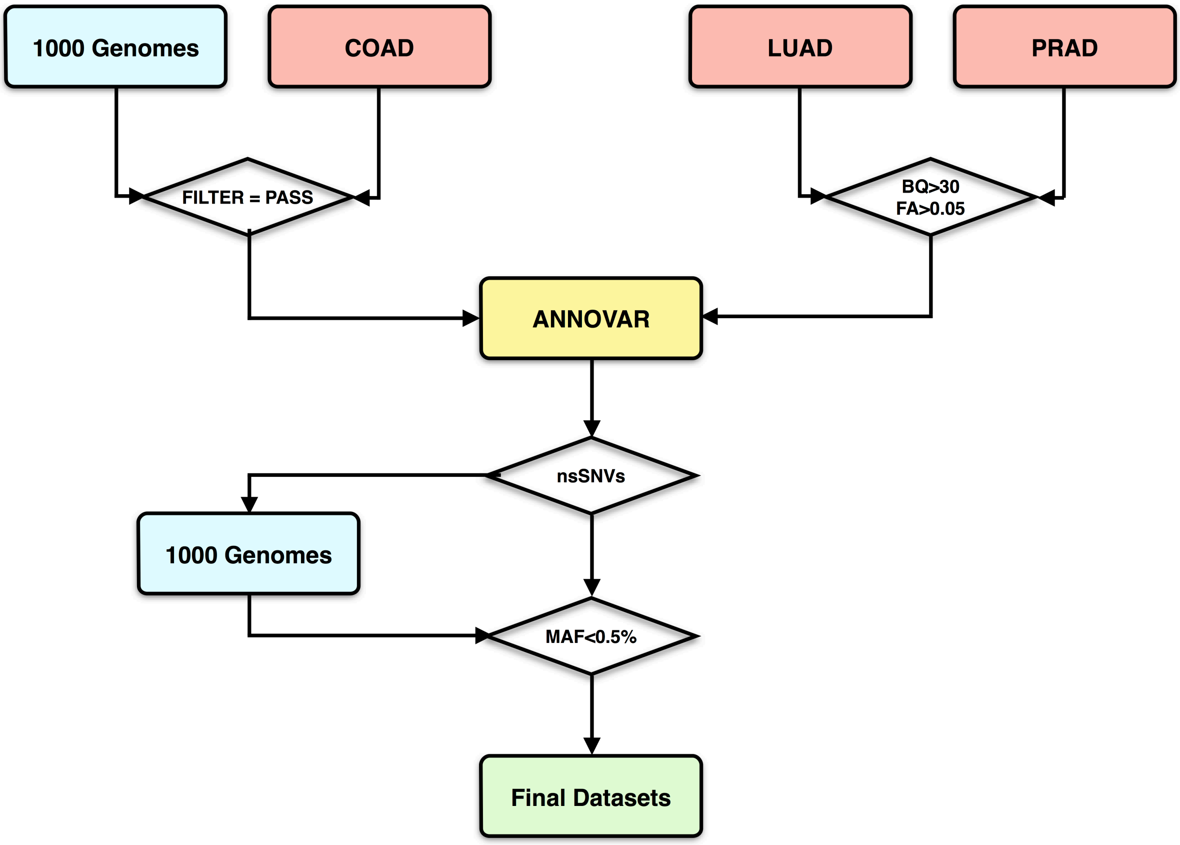
Supplementary Figures



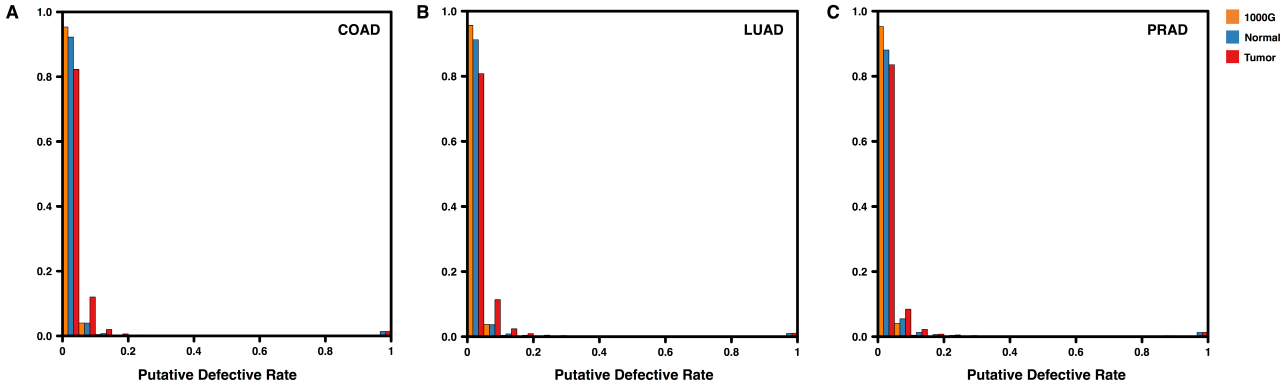
Supplementary Fig. S1. Distributions of pathogenic (red) and non-pathogenic (red) variants in dbSNP (Sherry, et al., 2001), for the subsets of nsSNVs and other SNP types (panel A). Distributions of pathogenic (red) and non-pathogenic (blue) nsSNVs for the subsets of rare (MAF<1%) and common nsSNVs (panel B). The Minor Allele Frequency (MAF) refers to the frequency at which the least common allele occurs in the population.



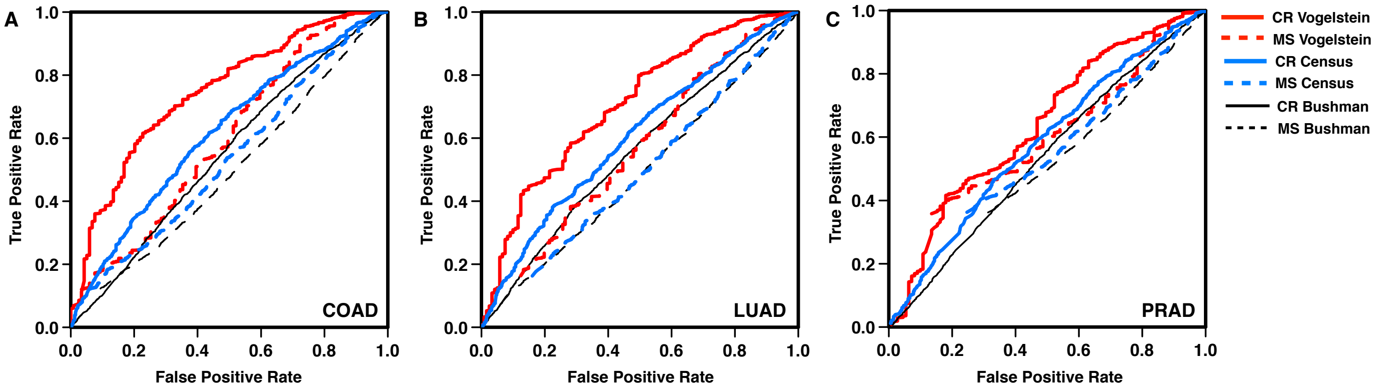
Supplementary Fig. 2. Distribution of nsSNVs in normal (blue) and tumor (red) samples for COAD LUAD and PRAD (Panels B, D and F). Same distributions for putative deleterious variants (PDV= nsSNV with MAF<0.5%) for COAD LUAD and PRAD (Panels A, C and E).



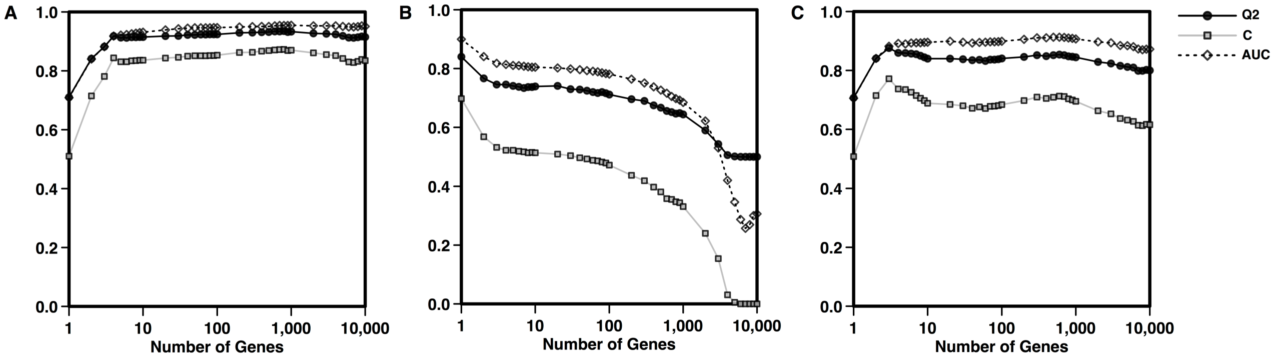
Supplementary Fig. S3. Flow chart of the procedure for the selection of putative deleterious variants (PDVs) in COAD LUAD and PRAD normal and tumor samples. MAF=Minor Allele Frequency



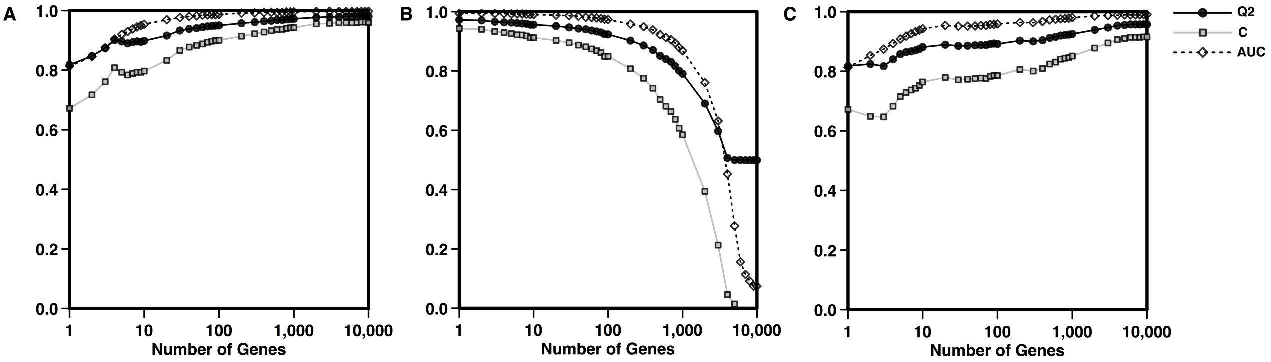
Supplementary Fig. S4. Distribution of putative defective rates (PDRs) in 1000 Genomes (yellow), TCGA normal (blue) and tumor (red) for colon, lung and prostate adenocarcinomas (COAD, LUAD and PRAD respectively).



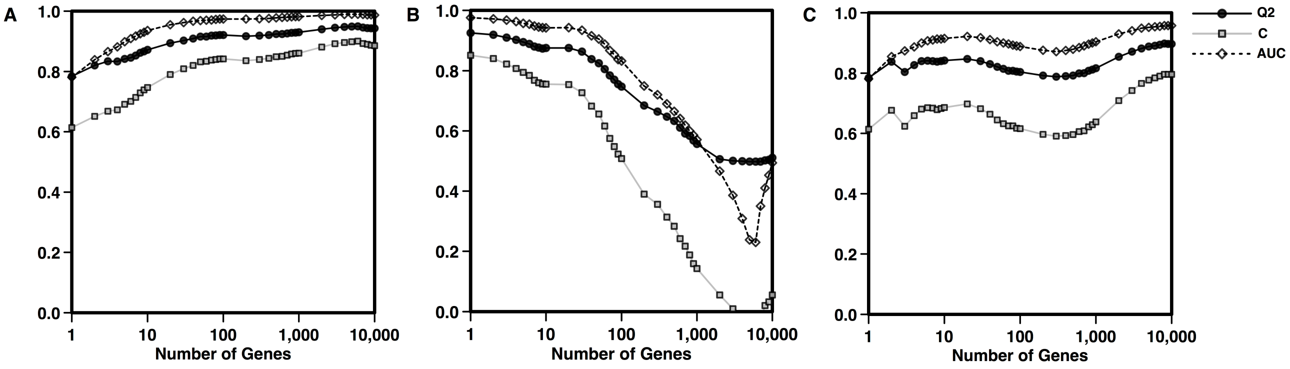
Supplementary Fig. S5. Performances of ContastRank (CR solid lines) and MutSigCV (MS dotted lines) in the prioritization of cancer-related genes for colon lung and prostate adenocarcinomas (respectively COAD, LUAD and PRAD). The ROC curves have been calculated using three lists of cancer-related genes from the BushmanLab website (Bushman, in black), COSMIC cancer census (Census in blue) and a recent publication from Vogelstein et al. (Vogelstein in red). More information about the 3 lists is provided Supplementary Methods 2.1. The values of the AUCs are reported in Supplementary Table S2.



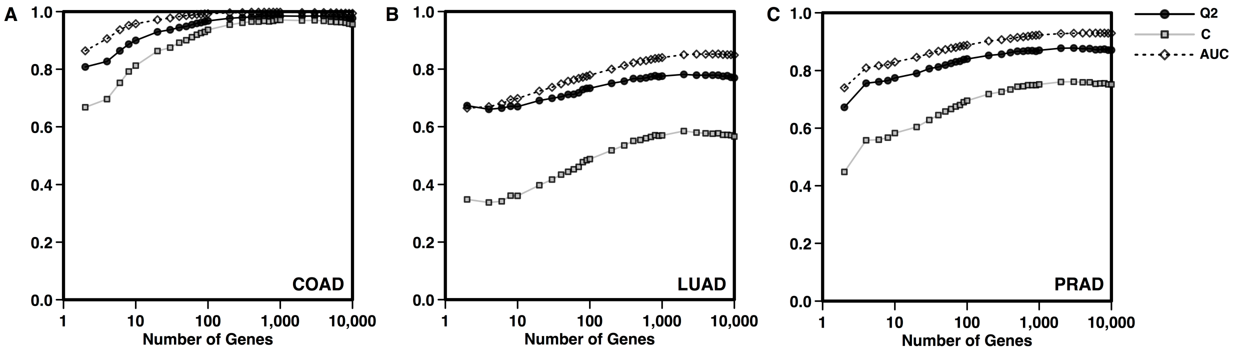
Supplementary Fig. S6. Average accuracy measures for ContrastRank (A), ContrastLow (B) and ContrastDiff (C) methods in the classification of COAD tumor and normal samples. In panel A and C the each point represent the performance of ContrastRank and ContrastDiff as a function of the number of genes use in the to calculate the global score assigned to each genome. In panel B accuracy of ContrastRank as a function of the number of top scored genes removed from the whole set of putative impaired genes. Q2=overall accuracy, C= Matthews correlation and AUC= area under the (ROC) curve.



Supplementary Fig. S7. Average accuracy measures for ContrastRank (A), ContrastLow (B) and ContrastDiff (C) methods in the classification of LUAD tumor and normal samples. In panel A and C the each point represent the performance of ContrastRank and ContrastDiff as a function of the number of genes used to calculate the global score assigned to each genome. In panel B accuracy of ContrastRank as a function of the number of top scored genes removed from the whole set of putative impaired genes. Q2=overall accuracy, C= Matthews correlation and AUC= area under the (ROC) curve.



Supplementary Fig. S8. Average accuracy measures for ContrastRank (A), ContrastLow (B) and ContrastDiff (C) methods in the classification of PRAD tumor and normal samples. In panel A and C the each point represent the performance of ContrastRank and ContrastDiff as a function of the number of genes used to calculate the global score assigned to each genome. In panel B accuracy of ContrastRank as a function of the number of top scored genes removed from the whole set of putative impaired genes. Q2=overall accuracy, C= Matthews correlation and AUC= area under the (ROC) curve.



Supplementary Fig. S9. Average accuracy measures for discriminating each type of tumor samples (COAD, LUAD and PRAD). In all the panels each point represent the performance of our method based on the difference of ContrastRank scores as a function of the number of genes used to calculate the global score assigned to each genome. In this case we consider and even number of highly positive and low negative scored genes. Q2=overall accuracy, C= Matthews correlation and AUC= area under the (ROC) curve.

Supplementary Tables

Supplementary Table S1

Tumor	Samples	N_{VAR}	N_{PDV}	N_{GEN}	N_{PIG}
COAD	220	9067/9362	996/1276	5057/5198	643/880
LUAD	625	8225/8656	1202/1599	4733/4930	751/1041
PRAD	309	8782/9019	1321/1540	4967/5063	819/953
1000G	1092	10559	318	5829	305

Samples = total pairs of normal/tumor samples. N_{VAR}= average number of nsSNVs in normal/tumor samples, N_{PDV}= number of putative deleterious variants (PDVs) with allele frequency <0.5%. N_{GEN} = number of genes with nsSNVs in normal/tumor samples. N_{PIG} = number of putative impaired genes (PIGs) with at least one PDV. COAD = Colon Adenocarcinoma, LUAD = Lung Adenocarcinoma, PRAD = Prostate Adenocarcinoma and 1000G= genotypes from 1000 Genomes Consortium.

Supplementary Table S2

Method	Bushman			COSMIC Census			Vogelstein		
	COAD	LUAD	PRAD	COAD	LUAD	PRAD	COAD	LUAD	PRAD
MutSigCV	0.49	0.60	0.56	0.53	0.49	0.53	0.60	0.56	0.60
ContastRank	0.55	0.75	0.71	0.62	0.60	0.58	0.75	0.71	0.64

Performances of ContastRank and MutSigCV in the prioritization of cancer-related genes for colon lung and prostate adenocarcinomas (respectively COAD, LUAD and PRAD). The values represents the AUCs of the ROCs in Supplementary Figure S5 that have been calculated using Bushman, COSMIC Census and Vogelstein lists of cancer-related genes respectively from the BushmanLab website (Bushman, 2013), COSMIC database (Forbes, et al., 2011) and a recent publication from Vogelstein and colleagues (Vogelstein, et al., 2013). More information about the lists and the procedure used for this test are available in Supplementary Methods 2.1.

Supplementary Table S3

Tumor	CV	Q_2	PPV	TPR	NPV	TNR	C	AUC
COAD	Identifier	0.92	0.97	0.86	0.87	0.97	0.84	0.94
	Unseen	0.77	0.92	0.60	0.72	0.95	0.58	0.78
LUAD	Identifier	0.97	0.97	0.96	0.96	0.97	0.93	0.99
	Unseen	0.95	0.96	0.94	0.94	0.96	0.90	0.98
PRAD	Identifier	0.91	0.92	0.91	0.91	0.92	0.83	0.97
	Unseen	0.82	0.85	0.77	0.80	0.86	0.64	0.88

Performance of ContrastRank in discriminating normal from tumor samples of colon, lung and prostate adenocarcinomas (respectively COAD, LUAD and PRAD) using putative impaired genes (PIGs) with score higher than 3. The results have been obtained using two different 2-fold cross-validation procedures (CV Identifier and CV Unseen) described in Supplementary Methods 2.2. The following accuracy measures are defied in Supplementary Methods 3. Q_2=Overall accuracy, PPV and NPV=Positive and Negative Predicted Values, TPR and TNR=True Positive and Negative Rates. MCC=Matthew's correlation, AUC=area under the (ROC) curve.

Supplementary Table S4

	Bushman			COSMIC Census			Vogelstein		
Tumor	**High**	**Low**	***p***	**High**	**Low**	***p***	**High**	**Low**	***p***
COAD	37/102	1,748/15,120	$7*10^{-8}$	20/119	397/16,471	$2*10^{-10}$	15/124	107/16,761	$7*10^{-14}$
LUAD	47/272	1,868/16,024	0.01	20/299	422/17,470	$1*10^{-4}$	16/303	109/17,783	$6*10^{-10}$
PRAD	18/79	1,754/14,915	0.01	8/89	405/16,264	$3*10^{-3}$	5/92	109/16,560	$5*10^{-4}$

High and Low represent the number of high (>3) and low (≤3) scored putative impaired genes. For each High and Low columns we reported the number of gene found in a manually annotated cancer-related gene list versus the remaining number of genes included in ContastRank list. The curated gene lists included in our test are: i) the Bushman list (Bushman, 2013); ii) the COSMIC Census list of cancer genes in COSMIC database (Forbes, et al., 2011) ; iii) the Vogelstein's list of driver genes affected by subtle mutations provided in Table S2A of a recently published paper (Vogelstein, et al., 2013). The p-value (*p*) is calculated comparing the distribution of high and low scored genes in manually curated and ContrastRank lists using the Fisher's exact test.

Supplementary Files

ContrastRank scores for COAD LUAD and PRAD are listed respectively in coad_normal_scores.txt, luad_normal_scores.txt and prad_normal_scores.txt files.

MutSigCV scores for COAD, LUAD and PRAD are listed respectively in coad_mutsigcv_scores.txt, luad_mutsigcv_scores.txt and prad_mutsigcv_scores.txt files.

ContrastRank scores used to discriminate between one tumor type and the remaining two are listed respectively in coad_mix_scores.txt, luad_mix_scores.txt and prad_mix_scores.txt files

These files are available online at http://snps.biofold.org/data/supfiles.tar.gz

External Link of data used in this work

Direct link to restricted TCGA datasets used in this work.

COAD:https://tcga-data-secure.nci.nih.gov/tcgafiles/tcga4yeo/tumor/coad/gsc/hgsc.bcm.edu/illuminaga_dnaseq_cont/mutations_protected/hgsc.bcm.edu_COAD.IlluminaGA_DNASeq_Cont.Level_2.1.5.0.tar.gz

LUAD:https://tcga-data-secure.nci.nih.gov/tcgafiles/tcga4yeo/tumor/luad/gsc/broad.mit.edu/illuminaga_dnaseq_cont/mutations_protected/broad.mit.edu_LUAD.IlluminaGA_DNASeq_Cont.Level_2.0.4.0.tar.gz

PRAD:https://tcga-data-secure.nci.nih.gov/tcgafiles/tcga4yeo/tumor/prad/gsc/broad.mit.edu/illuminaga_dnaseq_cont_curated/mutations_protected/broad.mit.edu_PRAD.IlluminaGA_DNASeq_Cont_curated.Level_2.1.4.0.tar.gz.md5

References

Bushman, F. (2013) Cancer Gene List. http://www.bushmanlab.org/links/genelists.

Cibulskis, K.*, et al.* (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples, *Nature biotechnology*, **31**, 213-219.

Forbes, S.A.*, et al.* (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer, *Nucleic acids research*, **39**, D945-950.

Lawrence, M.S.*, et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes, *Nature*, **499**, 214-218.

Quail, M.A.*, et al.* (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers, *BMC genomics*, **13**, 341.

Sherry, S.T.*, et al.* (2001) dbSNP: the NCBI database of genetic variation, *Nucleic Acids Res*, **29**, 308-311.

Sing, T.*, et al.* (2005) ROCR: visualizing classifier performance in R, *Bioinformatics*, **21**, 3940-3941.

Vogelstein, B.*, et al.* (2013) Cancer genome landscapes, *Science*, **339**, 1546-1558.