HHS Public Access

Author manuscript

Cancer Res. Author manuscript; available in PMC 2016 June 06.

Published in final edited form as:

Cancer Res. 2013 July 15; 73(14): 4372–4382. doi:10.1158/0008-5472.CAN-12-3342.

The Exomes of the NCI-60 Panel: a Genomic Resource for Cancer Biology and Systems Pharmacology

Ogan D. Abaan[1,*], **Eric C. Polley**[3,*], **Sean R. Davis**[1,*], **Yuelin J. Zhu**[1], **Sven Bilke**[1], **Robert L. Walker**[1], **Marbin Pineda**[1], **Yevgeniy Gindin**[1], **Yuan Jiang**[1], **William C. Reinhold**[2], **Susan L. Holbeck**[3], **Richard M. Simon**[3], **James H. Doroshow**[2,3], **Yves Pommier**[2,**], and **Paul S. Meltzer**[1,**]

[1]Genetics Branch, Center for Cancer Research, NCI, NIH, Bethesda MD 20892 USA

[2]Laboratory of Molecular Pharmacology, Center for Cancer Research, NCI, NIH, Bethesda MD 20892 USA

[3]Division of Cancer Treatment and Diagnosis, NCI, NIH, Bethesda MD 20892 USA

Abstract

The NCI-60 cell lines are the most frequently studied human tumor cell lines in cancer research. This panel has generated the most extensive cancer pharmacology database worldwide. In addition, these cell lines have been intensely investigated, providing a unique platform for hypothesis driven research focused on enhancing our understanding of tumor biology. Here, we report a comprehensive analysis of coding variants in the NCI-60 panel of cell lines identified by whole exome sequencing (WES), providing a list of possible cancer specific variants for the community. Furthermore, we identify pharmacogenomic correlations between specific variants in genes like *TP53, BRAF, ERBBs* and *ATAD5* and anti-cancer agents such as nutlin, vemurafenib, erlotinib and bleomycin demonstrating one of many ways the data could be utilized to validate and generate novel hypotheses for further investigation. As new cancer genes are identified through large-scale sequencing studies, the data presented here for the NCI-60 will be an invaluable resource for identifying cell lines with mutations in such genes for hypothesis driven research. To enhance the utility of the data for the greater research community, the genomic variants are freely available in different formats and from multiple sources including the CellMiner and Ingenuity websites.

Introduction

The NCI-60 human tumor cell line panel ([1]) is used by a broad range of cancer investigators and by the NCI Developmental Therapeutics Program (DTP) to discover novel anticancer drugs ([2]). This panel represents an invaluable and publicly accessible platform of pharmacological, genomic, metabolomic, biochemical, and molecular datasets ([3–8]). The

**To whom correspondence should be addressed at: National Cancer Institute, 37 Convent Dr., Bethesda, MD 20982-4265, ; Email: pmeltzer@mail.nih.gov 301-496-5266 (Ph) 301-402-3241 (Fax), ; Email: pommier@nih.gov 301-496-5944 (Ph) 301-402-0752 (Fax)

*These authors contributed equally to this work

Conflict-of-interest: Authors declare no competing financial interests.

present study reports findings from whole exome sequencing (WES) of the NCI-60 panel of cell lines. In addition, pharmacogenomic analyses provide examples of a few of the many ways the variant data could be used to generate novel hypotheses. Our study complements two recently published large-scale cancer cell line sequencing studies, which utilized limited number of genes ([9], [10]), since our work provides the whole exome variants for the entire NCI-60 cell lines. The data is made available through the CellMiner, NCI-DTP and Ingenuity Systems' websites ([11]).

Materials and Methods

Cell lines

The list of cell lines in the NCI-60 panel and their tissue origins are given in Supplementary Figure 8. DNA was extracted from cells and fingerprinted as described before ([12]).

Exome capture and sequencing

Briefly, 37Mb of coding region for each cell line was captured using the Agilent SureSelect All Exon v1.0 kit (Agilent). Genomic DNA (3μg) was sheared using the Covaris S2 ultra-sonicator (Covaris) using the settings Duty Cycle 10%, Intensity 5%, Cycle/burst 200 and Time 60s, which yielded a fragment size distribution with a mean at 200bp. Libraries were generated using standard Illumina library protocol (Illumina) followed by size selection using ChromaSpin TE200 spin columns (Clonetech). Pre- and post-capture steps were performed following the manufacturers' protocol (Agilent). The samples were sequenced as paired-end 80-mer reads on an Illumina Genome Analyzer IIx instrument (Illumina) following the manufacturers' protocol.

Data processing and variant calls

Fastq files were aligned against the reference human genome build 19 (hg19) using BWA ([13]). Alignment files were base quality score recalibrated and locally realigned around indels with GATK ([14]) and marked for duplicates using PICARD tools (picard.sourceforge.net). Alignment files and variant calls can be accessed from the links provided ([11]). Consensus genotype calls were generated using samtools mpileup ([15]) and annotated using the Annovar package ([16]). Variants were further filtered for the SureSelect bait region, a minimum read depth of 6 and a minimum quality score of 30 for SNVs and 60 for indels, producing the final variant calls.

Drug activity determination

Drug activity was determined by the Developmental Therapeutics program (DTP) human cancer cell line screen ([11]). The concentration of agent required to cause 50% growth inhibition (GI50) as measured at 48 hours by the sulphorhodamine B assay ([17]) was determined.

Gene expression and other NCI-60 molecular characterization

mRNA expression, microRNA expression, copy number, and protein measurements are publicly available from DTP or from CellMiner (excluding the protein data) ([11]). The details pertaining to data acquisition and analysis are previously published ([18]).

Volcano plots

The x-axis of a volcano plot depicts the difference in mean log GI50 between the cell lines containing a mutation in the specified gene and the cell lines not containing such a mutation. The y-axis depicts the statistical significance level for the comparison of log GI50 for those two groups of cell lines with larger values indicating smaller p-values. On a volcano plot for a gene, the points represent the compounds. On a volcano plot for a compound, the points represent the genes. For a volcano plot representing a gene, the false discovery rate can be limited to 0.2 or less by restricting attention to the 310 clinical and investigational compounds with p values no greater than 0.0005. When examining all of the screening compounds, the false discovery rate will be greater unless attention is restricted by a more stringent significance cut-off (e.g 10^{-4}) and an imposed cut-off on difference in log GI50 between mutated and wild-type groups (e.g. ± 0.5). In general, however, the volcano plots are used to either confirm previously identify hypotheses or to generate hypotheses that require independent validation.

Super learner prediction models

Using GI50 data on the NCI-60 for 103 FDA approved and 207 investigational oncology drugs and the 711 genes with at least 5 cell lines containing a type 2 variant in the gene, we estimated a predictor for each drug using the Super Learner algorithm ([19]). The predictor is using the gene-level mutation profile to predict the log GI50 for each drug. The Super Learner is an ensemble based prediction methodology that combines different machine learning predictors into a single optimal predictor based on minimizing the cross-validated risk. The base algorithms for the Super Learner include elastic net regression, gradient boosting regression, bagging, CART, random forests, neural networks, and support vector machines. In total, 35 prediction algorithms were combined for the super learner ensemble. We do not expect a single prediction algorithm (e.g. elastic net regression) to be optimal across all 310 drugs and the Super Learner allows the final predictor to data-adaptively up-weight the best algorithms for the final predictor. Examining the weights for each algorithm across the 310 drugs (data not shown) shows great variability indicating we should see a benefit with the Super Learner ensemble approach. Within a drug, the Super Learner predicts the log GI50 based on the gene-level mutation profiles. To compare across the drugs with different potencies, the log GI50 values need to be normalized. We define the normalized log GI50 for a cell line as the log GI50 minus the mean log GI50 for that drug in all the other cell lines. For ROC analysis we classified a cell line as sensitive to a drug if its true normalized log GI50 was less than −0.5, and insensitive if the value was greater than 0.5.

Results

The variant calls were generated as described in Materials and Methods where we filtered variants with a minimum quality of 30 (60 for small insertions/deletions) and a minimum depth of 6 with at least 3 alternate alleles over the targeted 38Mb coding region. Since matched normals are not available for cell lines, we performed a more stringent filtering to identify potential cancer specific variants. Using this filtering, the variants were divided into two groups: Type 1 variants corresponding to common (and possibly germline) variants and Type 2 variants enriched for acquired cancer specific variants (Figures S1–S2). We obtained over 1.2 million Type 1 and 60,005 Type 2 variants in the NCI-60 cell lines.

Although a limitation of cell line sequencing is the lack of available normal matched tissue for comparison, the NCI-60 panel does allow comparisons between cell lines from 9 distinct tissues of origin. NCI-60 cell lines with known microsatellite instability (MSI) (Figure S3) have very high Type 2 variant counts (Figure 1A). However, HCC2998, a colon cancer cell line not known to have MSI, has the highest number of Type 2 variants. In contrast to the known MSI cell lines, more than 98% of HCC2998 Type 2 variants are single nucleotide variants (Figure S4), suggesting that this hypermutator phenotype arises from a mechanism other than MSI. Of interest, HCC2998 carries a *POLE* exonuclease domain missence variant coding for a P286R mutation in POLε (Figure S5). Previous reports indicate that impaired POLε proofreading results in a high rate of single nucleotide substitutions and increased tumor formation ([20]) and *POLE* mutations in colorectal cancer has recently been reported ([21]). HCC2998 appears to exemplify this phenomenon, providing a reagent for further investigation and illustrating the utility of the NCI-60 WES data.

Given the diversity in the NCI-60 panel based on the tissue of origin, the WES data reveal important information regarding the etiology of each subgroup. As is evident from Figure 1B, there is a wide range of transition-to-transversion ratios (ti/tv) among the NCI-60 panel. Melanoma cell lines have the highest ti/tv (3.93) with higher C:G to T:A transitions, which is the major mode of change for UV-induced DNA damage ([22]). In contrast, lung cancer cell lines have a ti/tv (0.67) indicative of tobacco smoke-induced DNA damage ([23]). Thus, the WES data supports the prior notion that the NCI-60 panel retains disease etiology signatures ([7]).

Figure 2A shows a map of the 10 most frequently mutated genes in the Catalogue Of Somatic Mutations In Cancer (COSMIC) database ([24]). We annotated the WES variant calls as those present in the COSMIC database (v59) and those that are absent in COSMIC but predicted to be deleterious by the SIFT ([25]) or PolyPhen ([26]) algorithms. *TP53* is the most frequently mutated gene overall, while *BRAF* is the most frequently mutated gene among melanoma cell lines (Figure 2A). Although most of the variants identified in these 10 genes are already annotated in COSMIC, novel variants in these 10 genes were also observed. While, the lack of normal tissue makes it almost impossible to validate these as somatic changes, these variants were not observed in either the 1000 Genomes Project ([27]), or in the 5600 normal whole exomes available through the NHLBI Exome Sequencing Project ([28]). Besides the many well-defined cancer genes such as those in Figure 2A, large-scale tumor sequencing efforts by others continue leading to the discovery of novel cancer genes, such as

the 16 genes listed in Figure 2B. Because the NCI-60 cell lines are so well characterized and readily available, they are ideal tools for hypothesis driven research of these novel cancer genes/mutations identified by large-scale sequencing efforts. Details for these particular mutations or for any other gene mutation can be downloaded from the public domains, including CellMiner (Figure 3) or Ingenuity website (Figure S6).

To demonstrate the utility of this unique dataset and illustrate one of many ways to apply these data in hypothesis driven research, we carried out an integrated pharmacogenomic investigation. The fact that the NCI-60 panel has been used to screen thousands of compounds provides a rich resource for testing the relationship of variants in genes to drug response. Among 43,225 compounds screened for activity against the NCI-60 cell lines (as of September 2012), 15,898 showed high dynamic range in their GI50 estimates across all cell lines. For each gene with at least 5 cell lines containing a type 2 variant, we evaluated the association of log GI50 to variants in genes for all of the screened compounds. *TP53*, the most frequently mutated gene in the NCI-60 panel, demonstrates strong correlation with drug response. MDM-2 inhibitors are effective agents in cell lines with wild-type p53 (Figure 4A), where they can induce cell-death. Out of the 15,898 compounds and 310 FDA approved or investigational oncology drugs, the activities of two clinically relevant MDM-2 inhibitors show strong negative-correlation with mutant p53 (Figure 4B). Nutlin-3 gives the highest statistical significance score for its activity in p53 wild-type cell lines (Figure 4B and C). MI-219, a known MDM2 inhibitor exhibits a similar strong negative-correlation with mutant p53 (Figure 4B). In contrast, NSC-670177 (Table S1) shows significant selectivity for the p53 mutant cells. On the other hand, the proposed p53-specific compound RITA (NSC-652287) ([29]), initially identified as a DNA crosslinking agent ([30]) showed little evidence of selective activity for cell lines with p53 wild-type status and only limited correlation with Nutlin-3 (Figure S7A), questioning the claim that RITA acts specifically as a p53-reactivating compound. As for comparison, RITA displays far less selectivity for p53 wild-type cells than the classical DNA-targeted agent mithramycin. As expected, expression of the well-known components of the p53 pathway, MDM2 and miR-34a ([31]) correlate with p53 wild-type cell lines (Figures 4E and F). Additional pharmacogenomic correlations between *TP53* mutational status, microRNAs, mRNA transcripts or other agents are listed in Figure S7B. Integrating additional genomic datasets, such as gene and microRNA expression data ([18]) strengthens the value in all these comprehensive datasets for the NCI-60 panel.

We further supplemented this work with cross-validated multivariate analyses. For each of the 310 FDA approved or investigational oncology drugs, we developed a super learner ensemble machine learning model predicting log GI50 based on variants in genes. We included genes with type 2 variants in 5 or more cell lines across the NCI-60 panel. Leave-one-out cross-validation was used to evaluate the ability of such modeling to distinguish sensitive from insensitive cell lines for individual drugs and to select active drugs for individual cell lines. We developed these 310 models for each loop of a cross-validation in which one cell line was omitted and the remaining cell lines were used as a training set. Those models were then used to predict the log GI50 values for all drugs for the omitted cell line thereby predicting the most active drugs (smallest normalized log GI50; see Materials and Methods) against this cell line (Table S2). Using these models, we generated cross-

validated Receiver Operating Characteristic (ROC) curves for each cell line (Figure S8). The ROC curve plots sensitivity versus one minus specificity for identifying active drugs. The area under the curve (AUC) between the ROC curve and the diagonal line is a measure of the predictive accuracy of the WES based models. A large AUC value for a cell line indicates that the mutation spectrum of the cell line is informative for discriminating active from inactive drugs. The set of drugs analyzed, however, contains many cytotoxics, for which the predictive model based only on mutation spectrum was poorly informative. Our models included only mutation status and did not attempt to distinguish the confounding between mutation status and cell line lineage. Further studies with comprehensive models that include copy number, transcript abundance, and methylation status should yield more accurate predictions.

The ROC curves provide valuable insight into cancer biology. For instance, among the NCI-60 melanoma cell lines, SK-MEL-2 has the lowest AUC value (Figure 5A). This is particularly interesting since SK-MEL-2 is the only non-*BRAF-V600E* mutant melanoma cell line with an activating *NRAS-Q61R* mutation. As shown with the volcano plot in Figure 5B, the three *BRAF-V600E* specific inhibitors PLX-4720, vemurafenib (Figure 5C) and SB-590885 stand out with extremely high significance and differential mean GI50 in the *BRAF*-mutant cell lines. All the MEK inhibitors (blue font) including selumetinib (Figure 5D) and hypothemycin (Figure 5E) show highly significant selectivity and differential GI50, indicating their therapeutic value in cancer cells with activated MAP Kinase pathway. Notably, one compound, NSC-678518 showed extreme selectivity for the *BRA*F-mutated cells. NSC-678518, the anthrax lethal factor (LF), was identified in a screen for agents with similar inhibitory profiles to another MAP kinase kinase inhibitor, PD098059, and shown to proteolytically inactivate such kinases ([32]).

Parallel studies support the value of correlating genomics and targeted agents ([2], [9], [10]). Figure 5 (C through E) exemplifies that mutations in protein kinase target genes are strong indicators of response to clinically relevant targeted drugs. In addition, such observation could be generalized to key signaling pathways. Ten distinct kinase inhibitors from three major target classes cluster separately depending on the mutations in six genes: *BRAF, NRAS, PIK3CA, PIK3R1, PTEN* and *ERBB2* (Figure 5F). These effects can be viewed in the context of the MAP Kinase (MAPK) and PI3 Kinase (PI3K) pathways downstream of receptor tyrosine kinases (RTKs).

One of the most clinically relevant RTK is the Epidermal Growth Factor Receptor (EGFR). However, as demonstrated by Garnett *et al.* ([10]), it is critical to integrate genomic mutation data with transcript levels to correlate and possibly predict drug responses. The NCI-60 provides a solid background for studying gene expression ([18]) (see MDM2 example in Figure 4E), and its large drug database offers unique opportunities to query drug response parameters. To test this possibility, we examined the EGFR inhibitor, erlotinib, whose activity is highly correlated with gefitinib and lapatinib in the NCI-60 [see Figure 6 in ([18])]. Overall, high expression of EGFR (ERBB1) and ERBB2 is a determinant of cellular response to erlotinib (Figure 6B). However, the colon and CNS cell lines are generally insensitive to erlotinib in spite of high EGFR and ERBB2 expression. This can be rationalized by taking into account mutations in the MAPK or PI3K pathways, a common

mechanism of resistance ([33]), which are present in all 7 colon and 4 out of 6 CNS cell lines (Figure 6B).

Additional examples of correlations between type 2 variants and the 16,208 compounds, including the 310 FDA approved or investigational oncology drugs are included in Supplemental figures 9 and 10. Supplemental figure 9 contains volcano plots for type 2 variants in 44 other genes of interest with the corresponding list of significant NSC numbers in supplemental table 3. Supplemental figure 10 shows volcano plots for 28 selected drugs that are in clinical use or clinical trials. Together, these data again demonstrate the potential value of the NCI-60 drug and genomic databases for systems pharmacology.

The power of whole exome sequencing, instead of focused sequencing of preselected genes as published ([9, 10]), was revealed when we coincidentally found a significant correlation between a germline in-frame deletion (delCAATGT) in *ATAD5* (rs72427574) in certain cell lines and their increased sensitivity to DNA damaging agent Bleomycin. In addition, Zorbamycin (NSC-146208), and Peplomycin (NSC-276382), which are both Bleomycin analogues, show strong activity towards these cell lines. *ATAD5*, the human homolog of yeast *ELG1* is essential for maintaining genome stability through its functions in deubiqitinating PCNA and is known to be mutated in endometrial cancer ([34–36]). Genotype calls revealed ten cell lines where 5 are heterozygous and 5 are homozygous for delCAATGT. Out of the ten cell lines 3 are renal (ACHN, CAKI-1, RXF–393) where earlier work suggests dimethane sulfonate analogues, such as DMS612, as effective agents against renal cancer ([37]) and are being investigated phase I trials in renal cancer patients (#09-C-0111). Interestingly, there are additional germline variants in *ATAD5*, that are also present exclusively in the same set of 10 cell lines. When we looked for possible haplotypes in the Hapmap database, we discovered a region of linkage-disequilibrium spanning over 300 kb (Figure 7B). Therefore, this particular haplotype could be a response modifier during chemotherapy with DNA damaging agents. These results illustrate the discovery potential of exonic variant data when integrated with previously available NCI-60 databases.

Discussion

In this study, we provide WES analysis of the widely used NCI-60 cell line panel. We show that the overall pattern of mutation is strikingly divergent between cell lines, ranging from 172 to 9205 type 2 variants. As expected, higher variant rates are observed in MSI cell lines, but remarkably the highest number of SNV's was observed in HCC2998, a colon cancer cell line in which we discovered a defect in the proofreading domain of POLε. The signature of specific carcinogens is readily discernible in lung cancer and melanoma, which show very low (0.67) and high (3.93) transition-to-transversion ratio, respectively. Variants in established cancer genes are abundantly represented in the NCI-60, and numerous examples of variants in recently implied cancer genes are also present.

In addition to the mutational data provided in this manuscript, substantial drug sensitivity data for tens of thousands of compounds and multiple other types of biological data are available for the NCI-60. Using straightforward approaches (see Fig. 3) together with more sophisticated analyses, we were able to demonstrate the influence of specific variants for

TP53, BRAF, KRAS, NRAS, PIK3CA, PTEN and *ERBBs* on the response to clinically-relevant targeted agents (nutlin, vemurafenib, selumetinib, hypothemycin, rapamycin, wortmannin, perifosine, erlotinib, afatinib, lapatinib and neratinib) and to identify aspects of those results that may merit further study. For example, even though targeted inhibitors of activated *BRAF-V600E* have been widely studied, the comprehensive NCI-60 datasets offers a unique opportunity to identify additional mechanisms of resistance and possibly offer novel means to overcome acquired resistance. The power of the NCI-60 WES variants is apparent from the observation that common variants in the human population may have a profound effect on drug response. Of course, our observation with *ATAD5* gene locus requires further studies, however, it opens up a completely new perspective on common variants and their phenotypes in the context of DNA damaging agents and the ongoing clinical trials with DMS612 ([37]).

In comparison to the two recent studies conducted with more cell lines [947 in ([9]) and 639 in ([10])], our study integrate far more drugs [approximately 20,000 vs. 24 in ([9]) and 130 in ([10])] (see Volcano plots in Figures 4, 5, [7] and Supplemental Figures) and provides a comprehensive dataset of all exonic variants for the NCI-60 cell lines, whereas 1600 genes were sequenced in ([9]) and 64 cancer-related genes in ([10]). Given the availability of extensive biological and pharmacological data and the vast number of NCI-60 variants identified in this study, such comprehensive analyses as performed by these two studies offer enormous opportunities. The WES data that we are providing for the NCI-60 also enables the vast compound activity database to be used as a resource for drug development to complement genomic studies conducted using larger cell line panels. That is, when one discovers a genomic variant as a molecular target using other cell line resources, using the WES data for the NCI-60 one can potentially identify screened compounds with selective activity for that target. We have limited our work to the exploration of certain aspects of this invaluable data, and made this dataset public for the greater community to utilize and analyze. This is critical for expanding our knowledge in understanding tumorigenesis and the genomic bases of drug sensitivity in years to come as many more cancer related gene aberrations are discovered.

Importantly, the availability of this sequencing data will allow increased precision in the use of these common cell lines as experimental models and, as indicated above, to expand the utility of other cell line panels for drug development. In order to enable this important step forward the complete dataset is readily accessible in two forms, the easily searchable CellMiner database and a pre-filtered, annotated Ingenuity Systems database. Through these portals, cancer investigators will be able to select precisely the cell line models most genetically suited to their research. The availability of the variant information allows the formulation and testing of hypotheses arising from the entire range of projects using the NCI-60 or its components. In conclusion, our datasets add substantial depth to the already extensive characterization of the NCI-60 tumor cell panel and provide an invaluable resource for ongoing investigations in cancer cell biology and pharmacology.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Bill Kopp, NCI-Frederick, for DNA purification and validation. This study was supported by the Division of Cancer Treatment and Diagnosis (DCTD), and the Center for Cancer Research (CCR) of the National Cancer Institute, NIH. The authors would like to thank the NHLBI GO Exome Sequencing Project and its ongoing studies which produced and provided exome variant calls for comparison: the Lung GO Sequencing Project (HL-102923), the WHI Sequencing Project (HL-102924), the Broad GO Sequencing Project (HL-102925), the Seattle GO Sequencing Project (HL-102926) and the Heart GO Sequencing Project (HL-103010).

References

1. Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer. 2006; 6:813–823. [PubMed: 16990858]
2. Weinstein JN. Drug discovery: Cell lines battle cancer. Nature. 2012; 483:544–545. [PubMed: 22460893]
3. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, et al. A gene expression database for the molecular pharmacology of cancer. Nat Genet. 2000; 24:236–244. [PubMed: 10700175]
4. Staunton JE, Slonim DK, Coller HA, Tamayo P, Angelo MJ, Park J, et al. Chemosensitivity prediction by transcriptional profiling. Proc Natl Acad Sci U S A. 2001; 98:10787–10792. [PubMed: 11553813]
5. Szakacs G, Annereau JP, Lababidi S, Shankavaram U, Arciello A, Bussey KJ, et al. Predicting drug sensitivity and resistance: profiling ABC transporter genes in cancer cells. Cancer Cell. 2004; 6:129–137. [PubMed: 15324696]
6. Zoppoli G, Solier S, Reinhold WC, Liu H, Connelly JW Jr, Monks A, et al. CHEK2 genomic and proteomic analyses reveal genetic inactivation or endogenous activation across the 60 cell lines of the US National Cancer Institute. Oncogene. 2011
7. Liu H, D'Andrade P, Fulmer-Smentek S, Lorenzi P, Kohn KW, Weinstein JN, et al. mRNA and microRNA expression profiles of the NCI-60 integrated with drug activities. Mol Cancer Ther. 2010; 9:1080–1091. [PubMed: 20442302]
8. Weinstein JN, Pommier Y. Connecting genes, drugs and diseases. Nat Biotechnol. 2006; 24:1365–1366. [PubMed: 17093484]
9. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012; 483:603–607. [PubMed: 22460905]
10. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature. 2012; 483:570–575. [PubMed: 22460902]
11. NCI60_WES_data_links. [cited; Available from: Cellminer: http://discover.nci.nih.gov/cellminerexome Ingenuity: http://www.ingenuity.com/NCI60_WES BAM_files: http://watson.nci.nih.gov/projects/nci60/wes/BAMS/ DTP_drug_screen: http://dtp.nci.nih.gov/branches/btb/ivclsp.html DTP_molecular_targets_screen: http://dtp.nci.nih.gov/mtargets/mt_index.html
12. Lorenzi PL, Reinhold WC, Varma S, Hutchinson AA, Pommier Y, Chanock SJ, et al. DNA fingerprinting of the NCI-60 cell line panel. Mol Cancer Ther. 2009; 8:713–724. [PubMed: 19372543]
13. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]
14. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011; 43:491–498. [PubMed: 21478889]
15. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–2079. [PubMed: 19505943]
16. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38:e164. [PubMed: 20601685]

17. Rubinstein LV, Shoemaker RH, Paull KD, Simon RM, Tosini S, Skehan P, et al. Comparison of in vitro anticancer-drug-screening data generated with a tetrazolium assay versus a protein assay against a diverse panel of human tumor cell lines. J Natl Cancer Inst. 1990; 82:1113–1118. [PubMed: 2359137]
18. Reinhold WC, Sunshine M, Liu H, Varma S, Kohn KW, Morris J, et al. CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. Cancer Res. 2012; 72:3499–3511. [PubMed: 22802077]
19. van der Laan MJ, Polley EC, Hubbard AE. Super learner. Stat Appl Genet Mol Biol. 2007; 6 Article 25.
20. Albertson TM, Ogawa M, Bugni JM, Hays LE, Chen Y, Wang Y, et al. DNA polymerase epsilon and delta proofreading suppress discrete mutator and cancer phenotypes in mice. Proc Natl Acad Sci U S A. 2009; 106:17101–17104. [PubMed: 19805137]
21. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487:330–337. [PubMed: 22810696]
22. Ikehata H, Ono T. The Mechanisms of UV Mutagenesis. J Radiat Res (Tokyo). 2011; 52:115–125. [PubMed: 21436607]
23. DeMarini DM. Genotoxicity of tobacco smoke and tobacco smoke condensate: a review. Mutat Res. 2004; 567:447–474. [PubMed: 15572290]
24. Forbes S, Clements J, Dawson E, Bamford S, Webb T, Dogan A, et al. Cosmic 2005. Br J Cancer. 2006; 94:318–322. [PubMed: 16421597]
25. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001; 11:863–874. [PubMed: 11337480]
26. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7:248–249. [PubMed: 20354512]
27. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]
28. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013; 493:216–220. [PubMed: 23201682]
29. Issaeva N, Bozko P, Enge M, Protopopova M, Verhoef LG, Masucci M, et al. Small molecule RITA binds to p53, blocks p53-HDM-2 interaction and activates p53 function in tumors. Nat Med. 2004; 10:1321–1328. [PubMed: 15558054]
30. Nieves-Neira W, Rivera MI, Kohlhagen G, Hursey ML, Pourquier P, Sausville EA, et al. DNA protein cross-links produced by NSC 652287, a novel thiophene derivative active against human renal cancer cells. Mol Pharmacol. 1999; 56:478–484. [PubMed: 10462535]
31. He L, He X, Lim LP, de Stanchina E, Xuan Z, Liang Y, et al. A microRNA component of the p53 tumour suppressor network. Nature. 2007; 447:1130–1134. [PubMed: 17554337]
32. Duesbery NS, Vande Woude GF. Anthrax lethal factor causes proteolytic inactivation of mitogen-activated protein kinase kinase. J Appl Microbiol. 1999; 87:289–293. [PubMed: 10475971]
33. Wheeler DL, Dunn EF, Harari PM. Understanding resistance to EGFR inhibitors-impact on future treatment strategies. Nat Rev Clin Oncol. 2010; 7:493–507. [PubMed: 20551942]
34. Bell DW, Sikdar N, Lee KY, Price JC, Chatterjee R, Park HD, et al. Predisposition to cancer caused by genetic and functional defects of mammalian Atad5. PLoS Genet. 2011; 7:e1002245. [PubMed: 21901109]
35. Davidson MB, Katou Y, Keszthelyi A, Sing TL, Xia T, Ou J, et al. Endogenous DNA replication stress results in expansion of dNTP pools and a mutator phenotype. EMBO J. 2012; 31:895–907. [PubMed: 22234187]
36. Fox JT, Lee KY, Myung K. Dynamic regulation of PCNA ubiquitylation/deubiquitylation. FEBS Lett. 2011; 585:2780–2785. [PubMed: 21640107]
37. Mertins SD, Myers TG, Holbeck SL, Medina-Perez W, Wang E, Kohlhagen G, et al. In vitro evaluation of dimethane sulfonate analogues with potential alkylating activity and selective renal cell carcinoma cytotoxicity. Mol Cancer Ther. 2004; 3:849–860. [PubMed: 15252146]

38. Dalgliesh GL, Furge K, Greenman C, Chen L, Bignell G, Butler A, et al. Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. Nature. 2010; 463:360–363. [PubMed: 20054297]

39. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008; 455:1069–1075. [PubMed: 18948947]

40. Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, et al. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature. 2010; 465:473–477. [PubMed: 20505728]

41. Varela I, Tarpey P, Raine K, Huang D, Ong CK, Stephens P, et al. Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. Nature. 2011; 469:539–542. [PubMed: 21248752]

42. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, et al. The genomic complexity of primary human prostate cancer. Nature. 2011; 470:214–220. [PubMed: 21307934]

43. Ley TJ, Ding L, Walter MJ, McLellan MD, Lamprecht T, Larson DE, et al. DNMT3A mutations in acute myeloid leukemia. N Engl J Med. 2010; 363:2424–2433. [PubMed: 21067377]

44. Jones S, Wang TL, Shih Ie M, Mao TL, Nakayama K, Roden R, et al. Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. Science. 2010; 330:228–231. [PubMed: 20826764]

45. Wei X, Walia V, Lin JC, Teer JK, Prickett TD, Gartner J, et al. Exome sequencing identifies GRIN2A as frequently mutated in melanoma. Nat Genet. 2011; 43:442–446. [PubMed: 21499247]

46. Solomon DA, Kim T, Diaz-Martinez LA, Fair J, Elkahloun AG, Harris BT, et al. Mutational inactivation of STAG2 causes aneuploidy in human cancer. Science. 2011; 333:1039–1043. [PubMed: 21852505]

47. Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, et al. The consensus coding sequences of human breast and colorectal cancers. Science. 2006; 314:268–274. [PubMed: 16959974]

48. Vassilev LT. MDM2 inhibitors for cancer therapy. Trends Mol Med. 2007; 13:23–31. [PubMed: 17126603]

49. Ikediobi ON, Davies H, Bignell G, Edkins S, Stevens C, O'Meara S, et al. Mutation analysis of 24 known cancer genes in the NCI-60 cell line set. Mol Cancer Ther. 2006; 5:2606–2612. [PubMed: 17088437]

50. Kohn KW, Aladjem MI. Circuit diagrams for biological networks. Mol Syst Biol. 2006; 2 2006 0002.

51. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005; 21:263–265. [PubMed: 15297300]

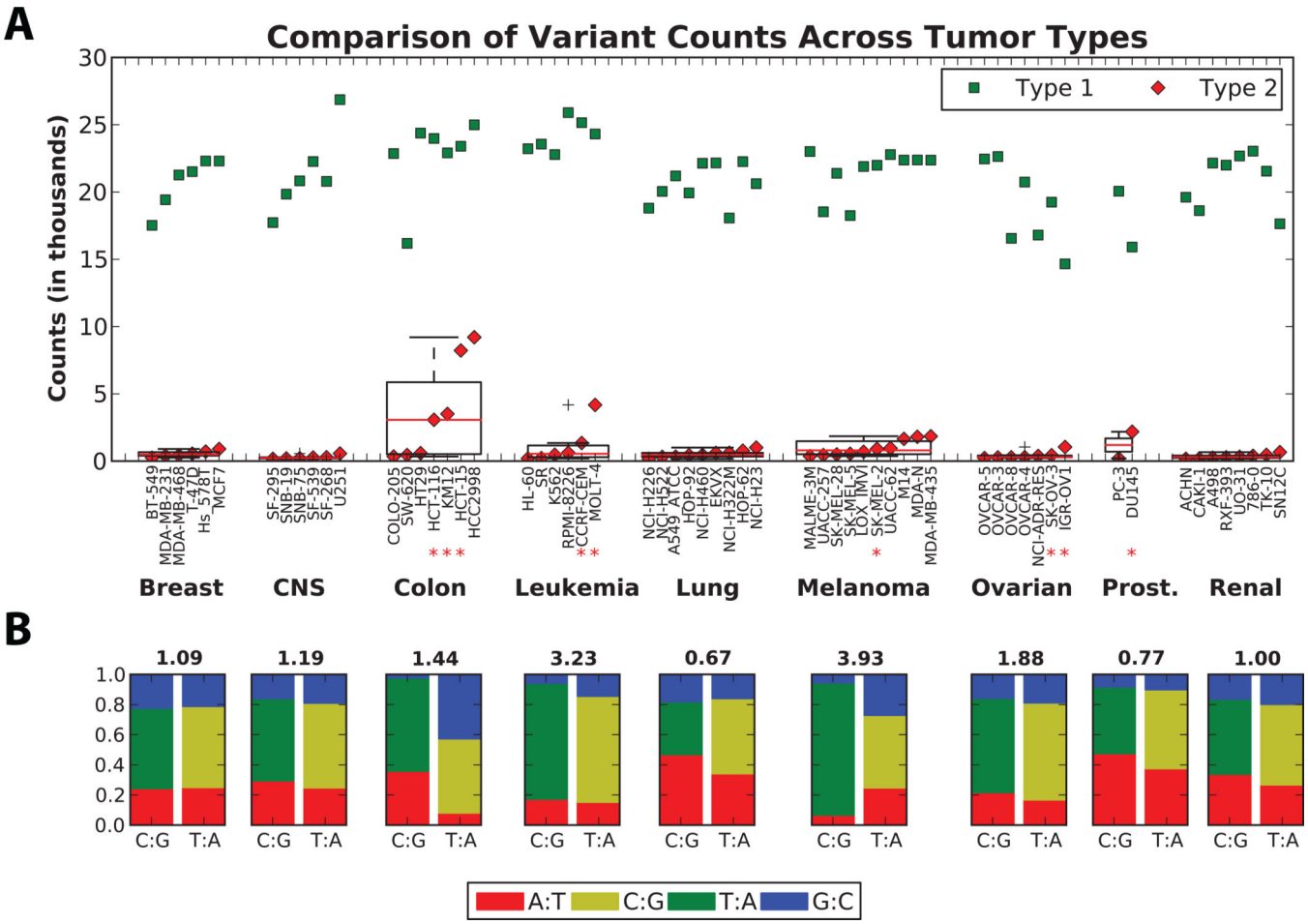

Figure 1. Results of WES variant calling

(A) Variant counts for each cell line from each tumor type are plotted for Type 1 and Type 2 fraction as green squares and red diamonds, respectively. Within each tumor type, the variant counts are sorted from lowest to highest, and a box blot is superimposed to demonstrate subgroup mean and spread. Microsatellite unstable cell lines are marked with a red asterisk. (B) Base transition-to-transversion (ti/tv) ratio is plotted for each tumor type in the NCI-60 panel for Type 2 variants that may likely be tumor specific. The y-axis represents the fraction of base conversions from a C:G or a T:A base pair to any other possible base pair change, which cumulatively equals 1. See also Figure S1 for additional details.

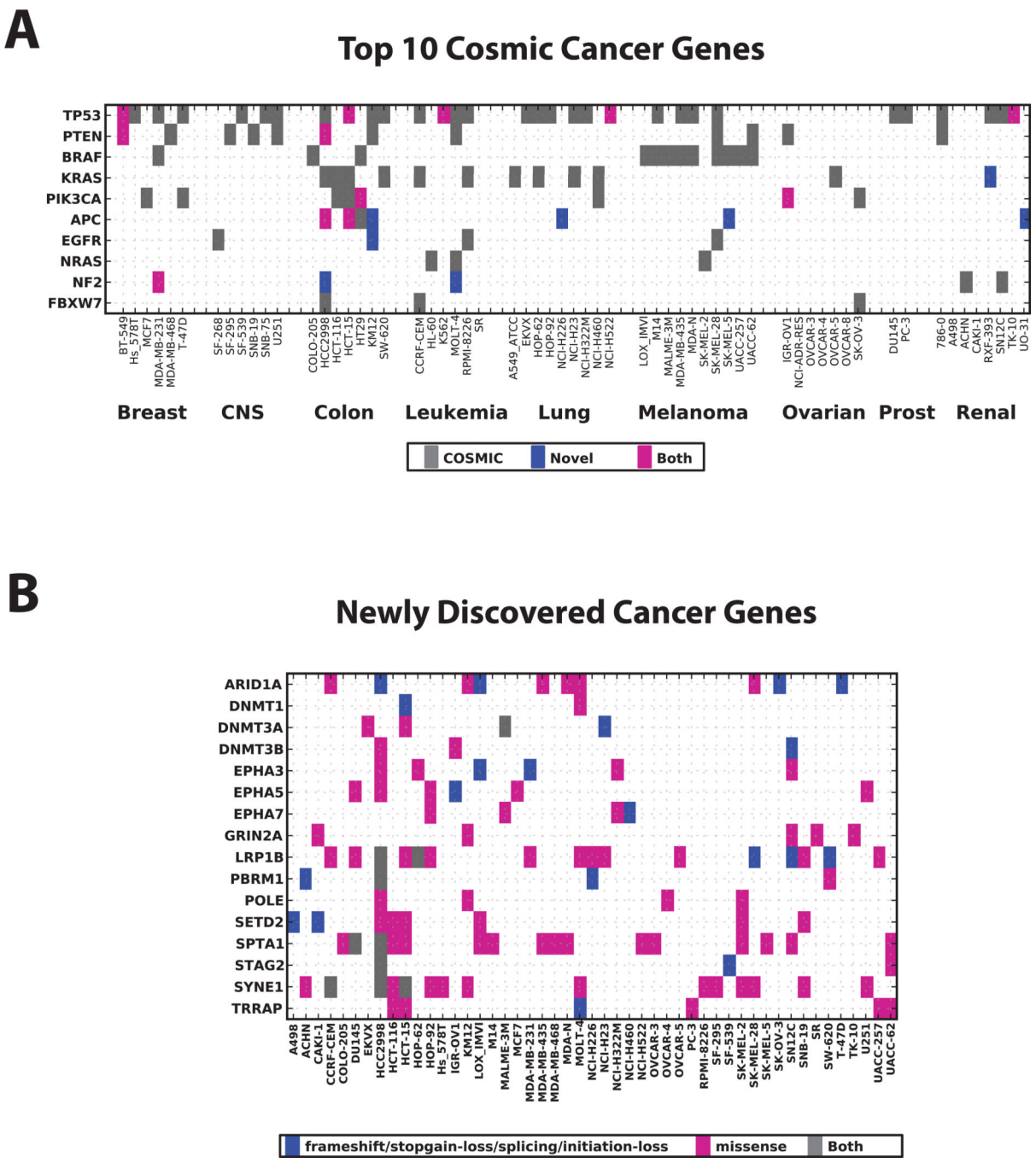

Figure 2. Mutation spectrum for the Top 10 most frequently mutated genes and novel cancer related genes in the NCI-60

(A) The Top 10 Cosmic Census Cancer genes (sorted by the number of occurrences in the NCI-60 panel) were scored for the presence of mutations in each cell line. Gray marks variants annotated in the COSMIC v59 database. Blue marks variants that are not in the COSMIC database but identified in the current study and predicted to be of deleterious in nature (either sift score < 0.05 or polyphen2 score > 0.85). Magenta marks cases where a cell lines harbors at least one COSMIC annotated and at least one novel variant in a particular gene (a gray and a blue mark). (B) New cancer genes identified in recent large-

scale sequencing studies such as: *SETD2* ([38]), *LRP1B* ([39, 40]), *PBRM1* ([41]), *SPTA1* ([42]), *DNMT3A* ([43]), *ARID1A* ([44]), *GRIN2A* ([45]), *TRRAP* ([45]), *STAG2* ([46]), *EPHA3/5/7* ([39]), POLE ([21]), and *SYNE1* ([47]). Blue boxes represent likely loss-of-function mutations (e.g. nonsense, splice site, initiation loss, and frame shift insertions or deletions) while magenta indicates missense mutations. Cases with co-occurrence of both types are labeled in gray.

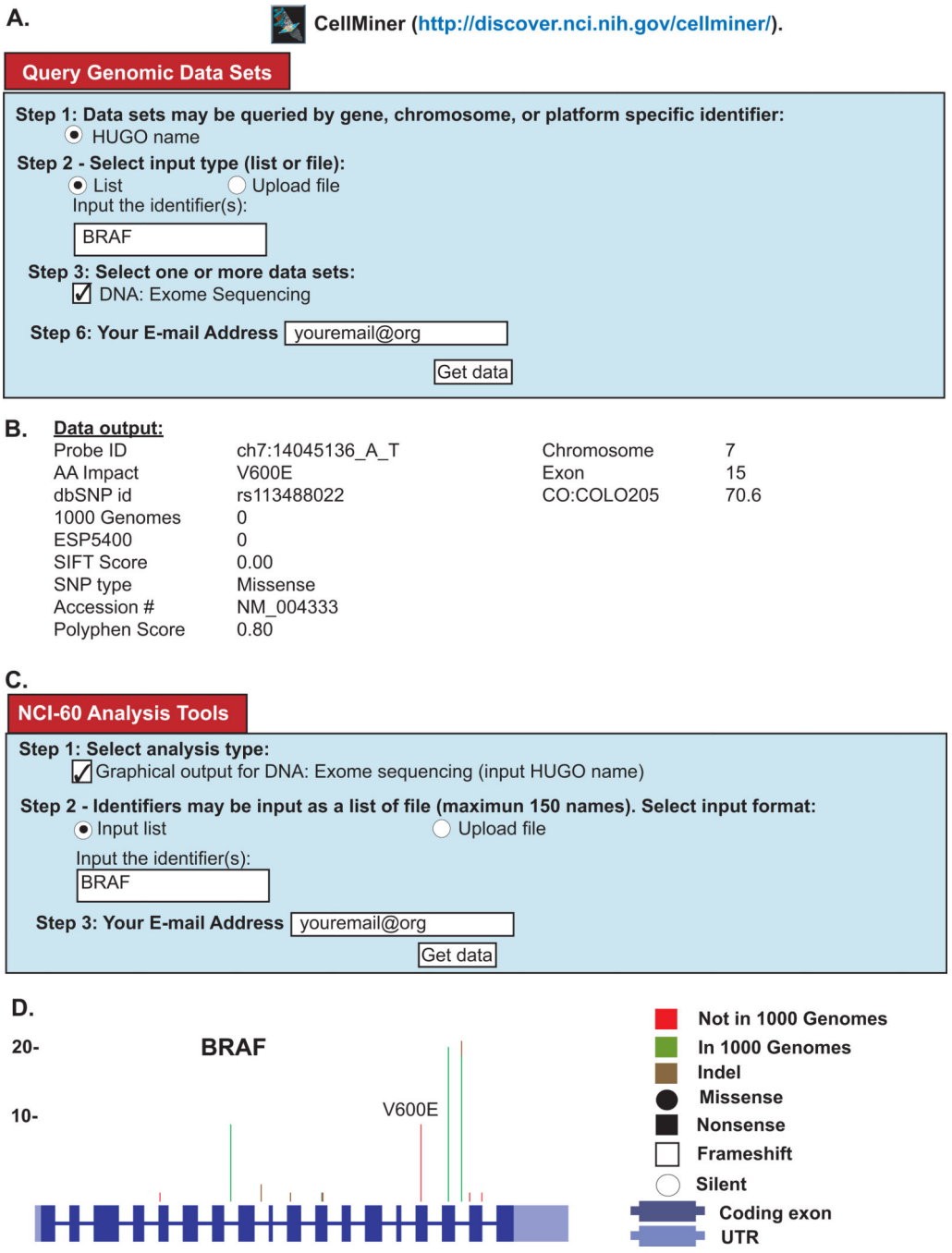

Figure 3. Snapshot from the Cell-miner Website

Figure 3: Snapshot from the Cell-miner Website. (A) To access tabular data, first click on the "Query Genomic Data Sets" tab. Specify data you want by: i) identifying the query type in Step 1 (HUGO name is required); ii) choosing whether you wish to type in your identifier, or upload your identifier(s) as a file in Step 2; iii) identifying the data set being queried in Step 3 (in this case exome sequencing); iv) entering your E-mail address in Step 6; and clicking "Get data". (B) The tabular data sent to you will include a full set of the data for all 60 cell lines (only one cell line is included for reasons of space). Within the output, i) The probe ID

denotes the chromosome number, start location, and the nucleotide change, ii) AA is amino acid, iii) dbSNP id iv) allele frequency in 1000-genomes, v) allele frequency in ESP5400, vi) SIFT score, vii) NCBI accession number, viii) Polyphen2 score. (C) To access graphical data, first click on the “NCI-60 Analysis Tools” tab. Choose the graphical output tool by, i) clicking “Graphical output for DNA:Exome sequencing” in Step 1; ii) choosing whether you wish to type in a your identifier, or upload your identifier(s) as a file in Step 2; iii) identifying the gene being queried, also in Step 2; iv) entering your E-mail address in Step 3; and clicking “Get data”. (D) The graphical data will be sent as an html, with accompanying pngs. The summary of all variants in BRAF is shown (individual cell lines are also included). The number of variants at each location are depicted by the vertical green, red, or brown lines.

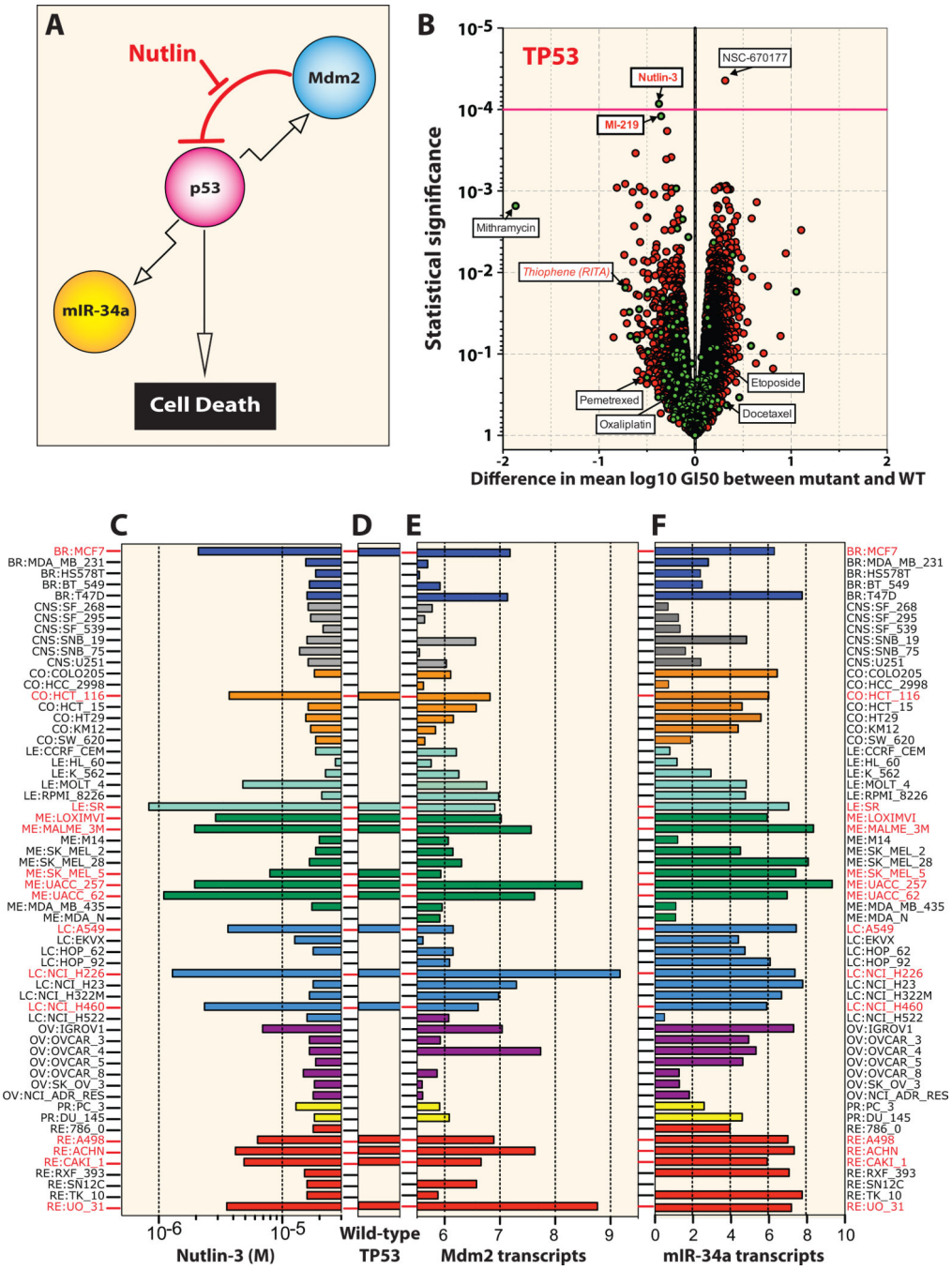

Figure 4. Correlation of *TP53* wild-type cells with nutlin-3 and other p53 pathway modulators

(A) Schematic representation of the p53-MDM2 feedback loop with p53 acting as a positive transcription factor for MDM2 and microRNA-34a while nutlin-3 acts as an MDM2 antagonist ([48]), blocking MDM2-mediated p53 degradation and killing of wild-type p53 cell lines. (B) The volcano plots show the difference in mean log GI50 between the cell lines containing a type 2 variant in *TP53* versus those cell lines not containing a variant along the x-axis and the −log10 p-value on the y-axis. Each red point represents one of the 15,989 compounds tested from the NCI screening data plus 310 approved and investigational drugs

(green points). A magenta guideline is given at significant p-value 10^{-4}. The NSC numbers or names for the statistically significant and for comparison some non-significant compounds are annotated on the plot. *TP53*-reactivating compounds from literature and in red. (C) Antiproliferative activity of nutlin-3 across the NCI-60 cell lines, where the bar graph is color coded by tissues of origins. (D) The *TP53* wild-type cells are marked with horizontal bars and red thick-marks. (E) MDM2 expression is highest in the *TP53* wild-type cells and those targeted by nutlin-3 (note mirror image profiles). (F) The expression profile of microRNA 34a, an established p53 target. See also Figure S6 for additional correlations.

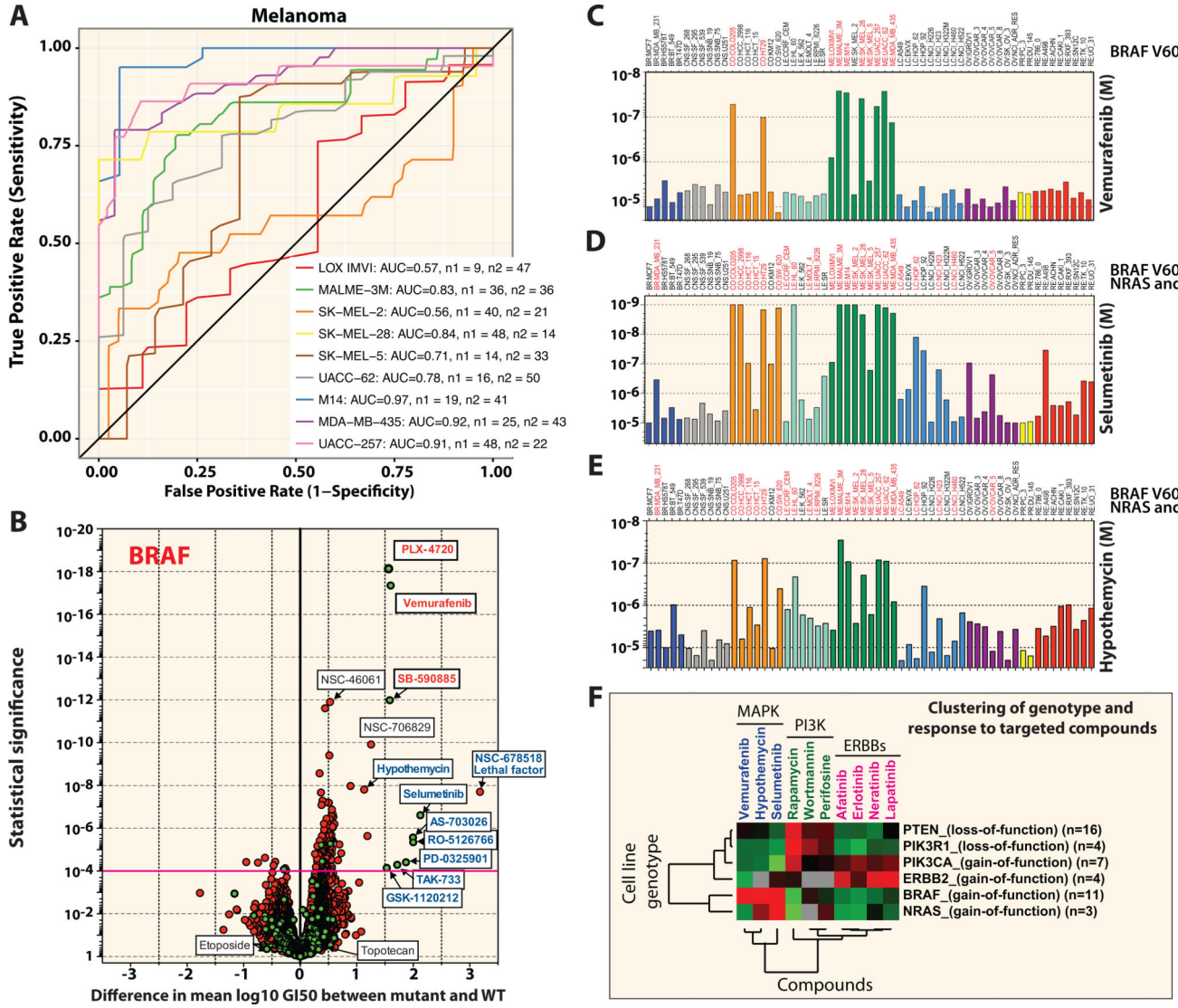

Figure 5. Correlation between MAP kinase pathway mutations and drug response to compounds that target this pathway in the NCI-60 panel

(A) Receiver operating curves (ROC) for cross-validated drug predictors for melanoma cells. Cross-validated ROC curves are shown for each cell line. The inset reports the area under the ROC curve (AUC) for each cell line and the number of inactive drugs (n1) and active drugs (n2). (B) Same volcano plot as in Figure 4B, for BRAF variants. A magenta guideline is given at significant p-value 10^{-4}. The NSC numbers or names for the statistically significant and for comparison some nonsignificant compounds are annotated on the plot. Drug response for (C) the *BRAF V600E* inhibitor vemurafenib, (D) the MEK inhibitor selumetinib and (E) the MEK/ERK inhibitor hypothemycin. Cell lines with mutations are labeled in red for the gene(s) indicated to the right. (F) Heat map showing correlations between mutations in key signaling intermediates (*PTEN, PIK3R1, PIK3CA, ERBB2, BRAF* and *NRAS*) versus drugs that target these pathways; MAPK pathway inhibitors (blue), PI3K pathway inhibitors (green), EGFR/ERBB inhibitors (magenta). Values for each drug represent the mean GI50 for each cell line with the particular gene mutations, including

previously published deletions and small mutations ([49]). The number of cell lines with the particular mutation is given in parentheses.

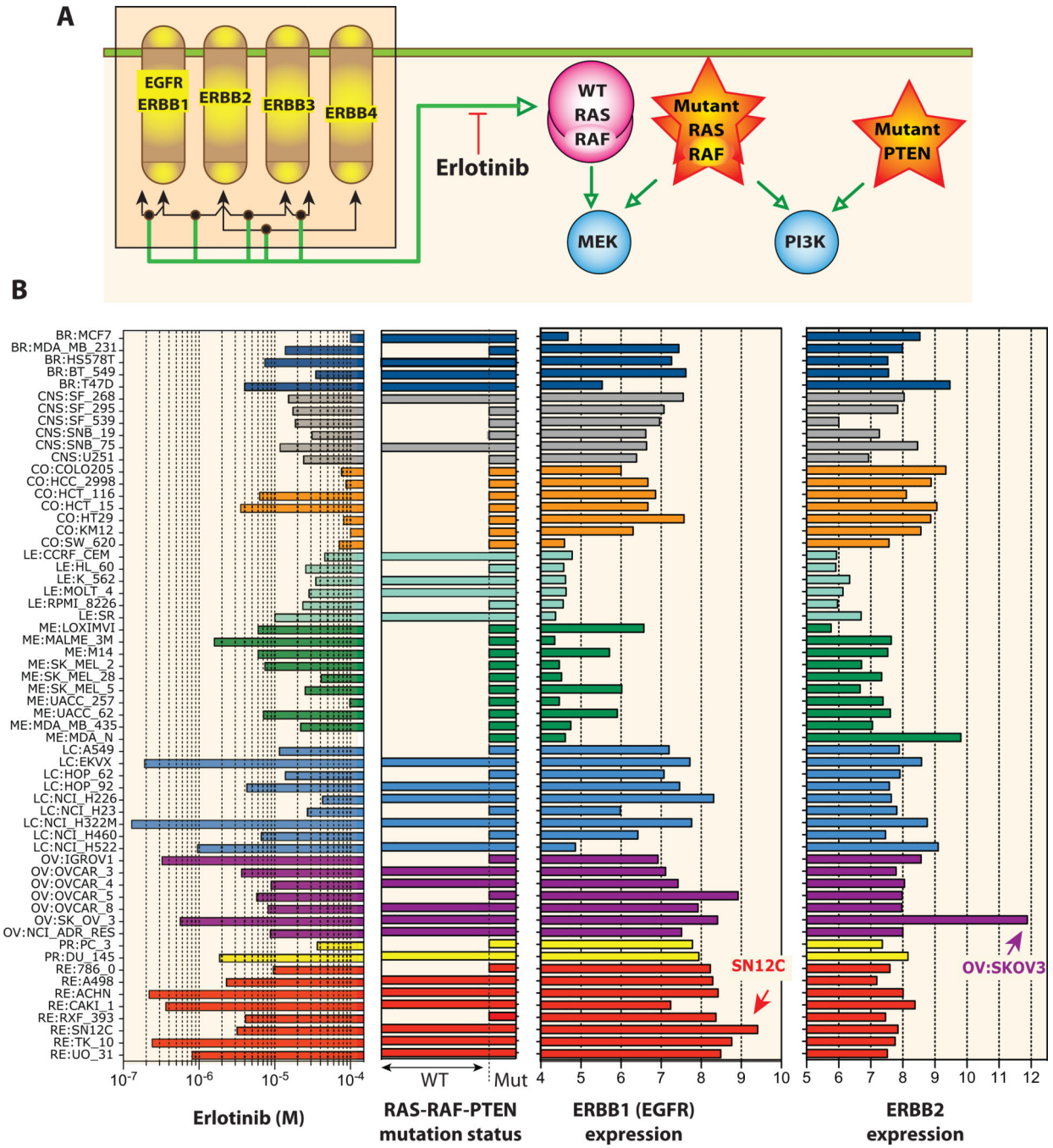

Figure 6. Correlation between erlotinib response and EGFR pathway gene expression and RAS-RAF-PTEN mutations in the NCI-60 panel

(A) Schematic representation of the EGFR pathway with its four components: ERBB1 (EGFR), ERBB2, ERBB3 and ERBB4. Dimerization complexes are indicated as nodes on the double-ended arrows according the Kohn's MIM nomenclature convention ([50]). Activations are shown as green arrows. Activating mutations of RAS or RAF directly activate MEK and render cells resistant to erlotinib ([33]). Similarly, inactivation of PTEN confers resistance by direct activation of PI3 kinase. (B) Left: antiproliferative activity of erlotinib across the NCI-60. The cell lines are color-coded by tissues of origins. Center left:

The *RAS-RAF-PTEN* wild-type (WT) cells are marked as full horizontal bars. Mutant cells (Mut) are shown as short bars. Center right: ERBB1 expression is highest in many of the cells targeted by erlotinib (note mirror image profiles). Far right: ERBB2 expression profile. The cell lines identified by arrows have focal amplification for *ERBB1* (RE:SN12C) and *ERBB2* (OV:SKOV3) (unpublished data).

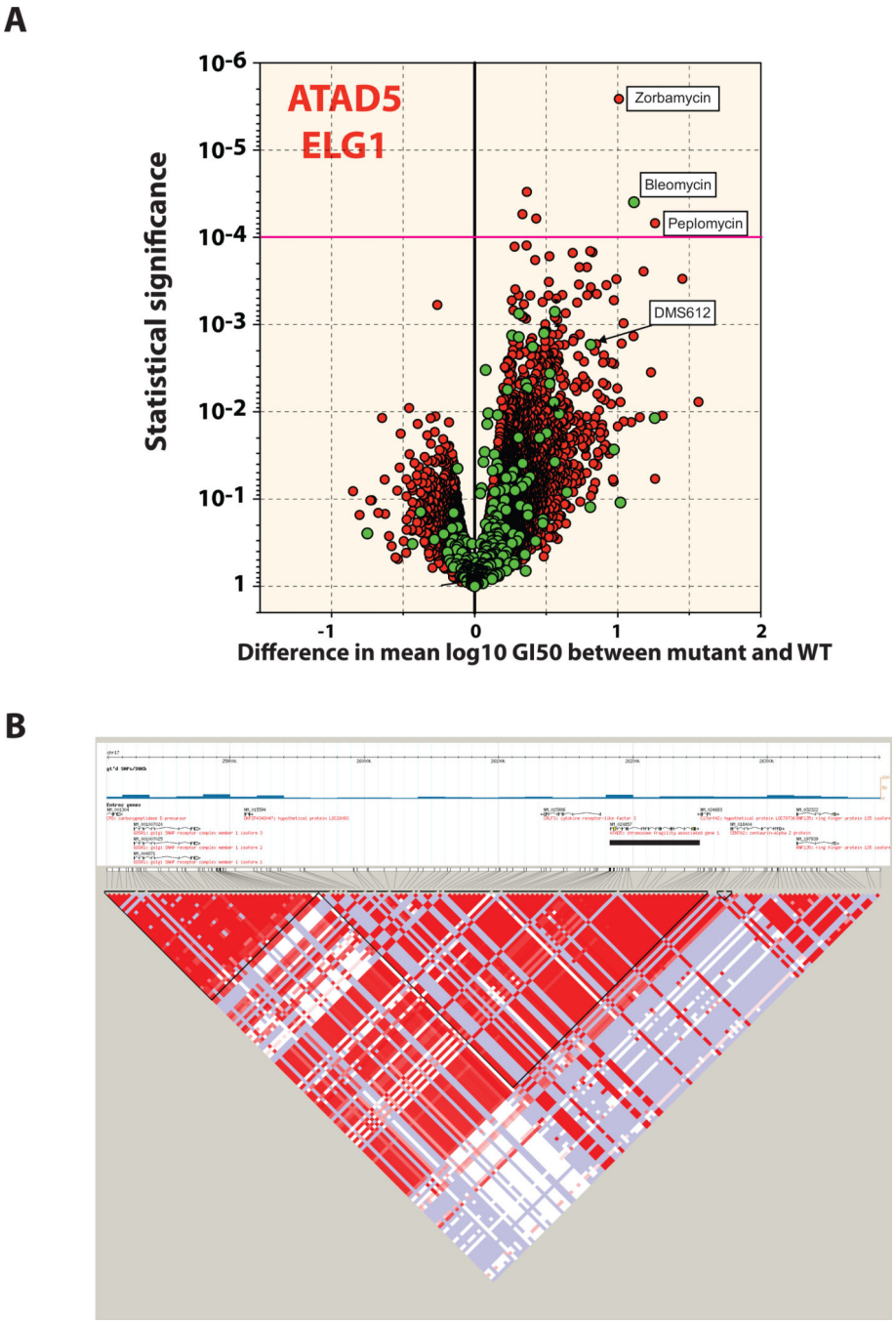

Figure 7. ATAD5 locus as a response-modifier for DNA-damaging agents

(A) Same volcano plot as in Figure 4B, for ATAD5 delCAATGG (rs72427574). A magenta guideline is given at significant p-value 10^{-4}. The names for the statistically significant compounds are annotated on the plot. (B) Linkage disequilibrium (LD) plot characterizing haplotype blocks in the *ATAD5* locus. The black bar marks the *ATAD5* gene location. The haplotype blocks were created using HaploView program ([51]), version 4.2