Functional Synergies yet Distinct Modulators Affected by Genetic Alterations in Common Human Cancers.

Marina Bessarabova¹, Olga Pustovalova¹, Weiwei Shi¹, Tatiana Serebriyskaya¹, Alex Ishkin¹, Kornelia Polyak², Victor E. Velculescu³, Tatiana Nikolskaya¹, Yuri Nikolsky¹,⁴

¹Thomson Reuters, Healthcare & Life Science., 500 Renaissance Dr., St. Joseph, MI 49085
² Dana-Farber Cancer Institute, Harvard Medical School, Harvard University, 44 Binney St. D740C, Boston, MA 02115
³ Ludwig Center at Johns Hopkins, Johns Hopkins Medical School, Johns Hopkins University, 1650 Orleans St., Baltimore, MD 21231
⁴Correspondence author

Abstract
An important general concern in cancer research is how diverse genetic alterations and regulatory pathways can produce common signaling outcomes. In this study, we report the construction of cancer models that combine unique regulation and common signaling. We compared and functionally analyzed sets of genetic alterations, including somatic sequence mutations and copy number changes, in breast, colon, pancreatic cancer and glioblastoma that had been determined previously by global exon sequencing and SNP array analyses in multiple patients. The genes affected by the different types of alterations were mostly unique in each cancer type, affected different pathways, and were connected with different transcription factors, ligands and receptors. In our model, we show that distinct amplifications, deletions and sequence alterations in each cancer resulted in common signaling pathways and transcription regulation. In functional clustering, the impact of the type of alteration was more pronounced than the impact of the kind of cancer. Several pathways such as TGFbeta/SMAD signaling and PI3K-signaling were defined as synergistic (affected by different alterations in all four cancer types). Despite large differences at the genetic level, all datasets interacted with a common group of 65 "universal cancer genes" comprising a concise network focused on proliferation/apoptosis balance and angiogenesis. Using unique nodal regulators ("over-connected" genes), universal cancer genes, and synergistic pathways, the cancer models that we built could combine common signaling
with unique regulation. Our findings provide a novel integrated perspective on the complex signaling and regulatory networks which underlie common human cancers.

Précis.
This genetic bioinformatics study reports the development of a cancer model offering a unified perspective on the complex signaling and regulatory networks that comprise different human cancers.

Introduction
Cancer is a complex collection of diseases characterized by common phenotypic and molecular hallmarks, which result from a staged evolutionary process fueled by random genetic and epigenetic alterations and clonal selection (1). Advanced tumors feature a large number of DNA changes of different types, such as chromosomal rearrangements, deletions, gene amplifications, and point mutations. Recent large-scale re-sequencing projects of multiple cancers have shown that (a) cancers of different origins have different landscapes of somatic “driver” mutations, and (b) individual tumors of the same type are highly heterogeneous in the content of mutated, deleted and amplified genes. The sheer volume of sequencing data, the large number of genetic alterations in cancers and the multi-patient “genome-wide” scope of these studies requires novel interpretation methods, such as functional (systems) analysis, in which the sets of altered genes are interpreted in terms of protein networks, pathways and cellular processes (2-10). Arguably, the most important functional finding was that sets of “driver” somatic mutations vary widely between cancers of different origin but hit the same (surprisingly few) major pathways (11-17). In other words, clonally selected genetic perturbations are not random but follow the natural functional clustering of human proteins in cellular processes and pathways rather than individual genes (2, 11, 12, 14, 15, 18). The pathways are linked into a complex interconnected cellular network, which can be represented in different modules, depending on algorithms used and the source of protein interactions (19-26). Cancer subnetworks calculated by different means were tested as phenotype-predictive biomarkers (23-25) and for hypotheses generation on novel drug targets (27).

There are few studies on comparative analysis of genetic alterations in different cancers. In (28) the authors were trying to identify core cancer pathways on the basis of breast cancer, colon cancer, glioblastoma and pancreatic cancer mutomes. Although useful, these studies did not explicitly address a number of important issues in carcinogenesis, including the cancer specificity of mutations
and other alterations, the relative impact of sequence alterations, amplifications or deletions in cancer phenotype, and common vs. cancer-specific pathways, among others. Integrative analysis of multiple data types in cancer and cross-cancer comparison is a non-trivial task due to very poor gene content overlap between the datasets. Thus, the pooled sets of amplified and mutated genes in breast cancer have a lower gene content overlap than expected by chance (29). Only one gene, TP53, is commonly mutated in all four major cancer types (breast, colon, pancreatic, and glioblastoma) and not a single gene is commonly deleted or amplified. Standard cross-dataset comparative methods are not helpful, as they are based on correlation between gene content, while a functional analysis approach could be useful in these cases (9, 30).

Here, we attempted to answer some of the questions by analyzing 12 sets of genetically altered genes with amplifications, deletions and sequence alterations from four different cancer types – breast, colon, and pancreatic cancer, and glioblastoma, (11, 12, 14, 15) by applying several novel methods of functional “meta-analysis” of generally non-overlapping datasets, including synergy in ontology enrichment (29), common regulators and kappa similarity between ontology enrichment patterns (31).

**Materials and methods**

**Data preprocessing.**

Mutations, amplifications and deletions were identified previously (11, 12, 14, 15). For mutations, we converted HUGO Gene names into Entrez Gene IDs. Copy number alterations were identified in a similar manner in all of the corresponding papers (12, 14, 15). The Illumina SNP array data were analyzed as described in (15). Fluorescence intensity image files were processed using Illumina BeadStation software to provide normalized intensity values (R) for each SNP position. For each SNP, the normalized experimental intensity value (R) was compared to the intensity values for that SNP from a training set of normal samples and represented as a ratio (called the "Log R Ratio") of log2 (Rexperimental/Rtraining set).

Homozygous deletions (HDs) were defined as two or more consecutive SNPs with a Log R Ratio value of ≤ -2. The first and last SNPs of the HD region were considered to be the boundaries of the alteration for subsequent analyses. Adjacent homozygous deletions separated by two or fewer SNPs were considered to be part of the same deletion. To identify the target genes affected by HDs, we compared the location of coding exons in the RefSeq databases with the genomic coordinates of the observed HDs.
Amplifications were defined by regions with an average LogR ratio $\geq 0.9$, containing at least one SNP with a LogR ratio $\geq 1.4$ and at least one SNP with a LogR ratio $\geq 1$ every ten SNPs. Amplifications $>3$ Mb in size and groups of nearby amplifications (within 1 Mb) that were also $>3$ Mb in size were removed.

Evaluation of network topology
Degree was calculated as the average number of protein interactions per protein from a given set using MetaCore (Thomson Reuters) protein interaction database (32). Since the interactions are directed, the nodes were characterized by IN and OUT-degree, i.e. the average number of outgoing and incoming interactions (32).

The clustering coefficient captures the degree connectivity between the protein(node)’s neighbors (33). It is defined as: $C_i = \frac{2n_i}{k_i(k_i-1)}$, where $n_i$ is the number of interactions between the $k_i$ neighbors of node $i$.

Enrichment by protein function and ontology enrichment were performed using standard tools of MetaCore as described (6).

Upstream and downstream objects
For all gene lists, we identified upstream and downstream genes as their one-step interaction neighbors. As all interactions in MetaCore Global Network are directed, each network object has a set of upstream objects with outgoing links to the object, and downstream objects, with incoming links from the object. Therefore, the list of upstream (downstream) targets for each gene list of interest was defined as a complete set of upstream (downstream) objects for all genes in the list.

Relative connectivity of proteins inside the set (intra-connectivity) and between the set and the global interactome (inter-connectivity) were calculated using MetaCore protein interaction database as described (29).
Processes overview

To visually evaluate the degree of functional concordance between analyzed lists we ranked ontology entities (processes) according to the degree they discriminate one gene list of interest in comparison with other lists of interest. We took as input the results of enrichment analysis in functional ontologies for each gene list – lists of processes with p-value less than 0.05. Then we created a united process table, where processes were in rows and gene lists in columns. We set p-value\(i_j=1\) if definite process (i) was not in a list of processes for some gene list (j). For each cell of the table we calculated D-value: difference between \(-\log(p\text{Value})\) for a process in the current column and maximal \(-\log(p\text{Value})\) for the same process in other columns of table:

\[
D_{ij} = \left[-\log_{10}(p\text{Value}_{ij})\right] - \max_{x_j}(−\log(p\text{Value}_i)).
\]

The negative D values were set to zero. Therefore each process has a non-zero D value only for one gene list. Then, we ranked processes consequently for each list, so that D-value was sorted in decreasing order.

Kappa statistics for inter-rater agreement was applied as published recently to compare the poorly overlapping gene lists or their congruency (31).

Results

We analyzed three types of DNA alterations (amplifications, homozygous deletions, and “driver” somatic point mutations) described in four large-scale studies on exon re-sequencing of four common cancers (breast cancer (B), colon cancer (C), glioblastoma (G), pancreatic cancer (P)) (11, 12, 14, 15) (Supplementary Table S1; Supplementary Excel File S1). The 12 gene lists (three alteration types in four cancers) represented pooled data on all genomic alterations detected in primary tumors of 11 patients each in both breast and colon cancers, 22 patients in GBM and 24 in pancreatic cancer.

We performed a multi-dimensional functional analysis at two levels: 1) DNA alteration type level, aiming to understand commonalities and differences between sets of amplifications (A), point mutations (M), and deletions (D) between cancers (cross-cancer analysis); 2) cancer type level – to describe the differences between the three types of DNA alterations for each cancer (cross-data type analysis). We also analyzed the non-redundant A+M+D unions for each cancer and B+C+G+P unions for each alteration type in order to describe the common functional traits for each level.
Genetically modified gene content is generally unique across cancers and data types

First, we compared all 12 gene lists, based on the assumption that common functionality is defined by similar sets of altered genes (Fig. 1). However, the datasets had very limited overlap both cross-cancer and cross-data type. Not a single common gene was found among all three data types (A, M, D) in any of the four cancers. This observation may suggest different functional roles for amplifications, deletions and mutations in carcinogenesis. Only one gene was altered at any level in all four cancers: the mutated TP53 tumor suppressor gene. The cross-cancer intersections between A, M, and D sets were small, with the majority of amplified, mutated, and deleted genes being unique for each cancer.

This observation is illustrated by the similarity tree between four cancers based on gene content using Cohen kappas as the similarity metric (Fig. 2, A; Supplementary Excel File S2). Although the gene intersections are small, the tree suggests that the gene content similarity is rather defined by the DNA alteration type rather than by the type of cancer. The impact of the DNA alteration type over the cancer type was also higher in the comparison of ontology enrichment patterns between the sets (31; Fig. 2, B; Supplementary Excel File S2). Thus, sets of amplifications for P, B, C and G clustered together on multiple pathway and cell process ontologies level. Mutation and deletion sets also mostly grouped together between different cancers.

Altered genes are highly connected and alteration type – biased in encoded functions.

We evaluated the degree of connectivity of proteins encoded by the A, M, and D datasets and compared it to the general connectivity of the human “global network” (6; Supplementary Excel File S3). In general, genetically modified genes averaged more interactions than expected (10 out of 12 datasets, except for amplifications and deletions in breast cancer) (Fig. 3). In all cancers, the sets of genes with “driver” mutations were more enriched in “hubs” than the sets of amplified and deleted genes. On average, mutated genes had three to seven fold more interactions than an average human protein, even after removal the major hub TP53 from the lists (Supplementary Excel File S3).

Protein function profiles were significantly different between the sets (Supplementary Excel File S4). For instance, in colon cancer, there were twice as many mutated receptors and amplified transcription factors and kinases as expected. Unsupervised hierarchical clustering of functional enrichment profiles shows that the sets clustered according to alteration level rather than by cancer type (Fig. 4). Thus, deletion sets in all cancers except for colorectal formed a tight cluster with over-abundant ligands and under-represented enzymes and proteases, which further confirmed our
hypothesis about differences in the functional landscape of colon cancer alterations. Mutation sets for the same three cancers (B, P, G) formed another cluster.

Unique and common processes and pathways affected in different datasets.
We conducted an ontology enrichment analysis (EA) of all 12 datasets in order to determine the commonality and differences between the individual sets of alterations (data type x cancer), between cancers (unions of A+M+D) and data types (unions of B+C+P+G) (Supplementary Excel File S5-S9). Ontology entities were ranked according to the degree they discriminate one set of alterations from all others by D-score (Supplementary Excel File S10). As expected, all 12 sets were heavily enriched in cellular processes related to carcinogenesis. However, the vast majority of top ranked pathways and processes were unique for each set (Supplementary Excel File S11). Thus, deletions in glioblastoma were highly enriched in immune response pathways, such as the Classical complement pathway. This is probably due to co-deletion with the CDKN2A tumor suppressor, which was the case in 9 out of 14 glioblastoma samples. Deleted genes in pancreatic cancer mostly took part in apoptosis-related pathways, while mutations in pancreatic cancer hit mostly development pathways.

Core synergistic pathways are shared between cancers despite different gene content
In order to determine whether different sets of altered genes alter biological functions cooperatively, we evaluated synergy in enrichment distributions between data types and cancers (Supplementary Excel File S12). A pathway or a process was considered as synergistically affected if its enrichment p-value for the non-redundant union of examined sets was lower than each p-value for individual sets (29). Synergy ranking, therefore, reflected the degree of functional “co-operation” between the altered genes of different type, regardless of their gene overlap. The top scored synergistic pathway maps for all 12 sets included EGFR signaling via PIP3, G1/S transition in cell cycle, PKA signaling, and multiple development pathways (Leptin, ESR1, ESR2, HGF, and WNT signaling) (Supplementary Table S2).

Ontology synergy was pronounced both at the cancer level and at the DNA alteration level. Each cancer featured a particular distribution of synergistic pathways and processes. The top synergistic pathways for breast cancer were G1/S transition in cell cycle, transport of CDKs/cyclins, and Brca1 regulation of DNA damage control. In colon cancer, these were EGFR signaling via both PIP3 and small GTPases, gastrin signaling, and ERBB family signaling; in pancreatic cancer –
apoptosis, PKA signaling, and cytoskeleton remodeling; and in glioblastoma – apoptotic pathways, G1/S transition, and p53-dependent transcription (Supplementary Table S2).

Although differentially affected in different cancers, the synergistic pathways are interconnected and form remarkably similar core network structures in all four cancers (Supplementary Fig. S1-S4). All “core pathway reconstructions” include regulation of cell proliferation and cell cycle via TGF-beta receptors, SMAD2,3,4, PIP3K signaling, and CDK2,4,6. Different versions of PIP3K (which is frequently mutated) signaling are also common in all four cancers. The core pathways are then linked by protein interactions with the pathways affected by alterations in a cancer-specific manner, including Brca1 signaling in breast cancer, PKA and ESR-Erk signaling in pancreatic cancer; plasmin-fibronectin-laminin pathways in colon cancer, and NGF-PIP3K signaling in glioblastoma.

Key upstream and downstream interactions of genetically altered genes and “universal cancer genes”

Datasets that are functionally similar but different in gene content often have common regulator genes (31). We identified the sets of one-step upstream and downstream objects (proteins and complexes) for all 12 alteration datasets and 7 unions using the MetaCore collection of 300,000 protein interactions as the source of connectors (Supplementary Excel File S13). We expected that the sets of upstream objects would provide clues about common and unique regulation of amplified, mutated, and deleted genes, and that the set of downstream objects would provide information about the common and unique processes and pathways affected by DNA alterations in different cancers (Supplementary Excel File S14).

First, we compared similarity between 12 upstream and downstream datasets using the kappa-based similarity approach (Fig. 2, C; Supplementary Excel File S2). The downstream sets clustered in a DNA-level favored grouping pattern, similar to the patterns described above for gene content and functional enrichment categories. However, the upstream datasets tended to group based on cancer rather than DNA alteration type. Such patterns may suggest common regulation of DNA level changes in a cancer-type dependent manner.

Second, we identified transcription factors (TFs), ligands, receptors, kinases, proteases and proteins of other functions that were over-connected with the proteins encoded by altered genes. P-value ranking of over-connected proteins provides an intuitive hypothesis-generation overview of the most relevant upstream regulators and downstream effectors for a given dataset (29). In general, intra-
connectivity was lower than expected within all sets, which suggests a lack of functional dependency between mutated or amplified or deleted proteins in any cancer (Supplementary Excel File S15). We observed several trends. First, there were disproportionally more over-connected objects (OCOs, i.e. proteins and protein complexes) for the subsets of unique genes than subsets of common genes in all 12 cases. Each set of altered genes had a unique set of interacting transcription factors, kinases, receptors and ligands. For instance, in breast cancer, amplifications were over-connected with LXR-beta and TR-alpha, whereas the mutated genes were over-connected with TFIIE (Supplementary Excel File S16). Second, every type of cancer (defined as a union of A, M, D for each cancer) and every type of alteration (union of B, C, G, P for each type) were associated with a unique set of TFs, kinases, receptors and ligands (Supplementary Table S3). At the DNA level, the unique TFs for amplifications were LXE-beta and TR-beta; for mutations – STAT1, ERF7, p73 (Supplementary Excel File S17).

In order to reveal common genes regulating or regulated by genetically altered genes in each cancer, we intersected sets of upstream and downstream objects correspondingly for all 12 sets. Despite low overlap between altered genes in the sets, we found 34 common downstream and 47 common upstream genes which we called “universal cancer genes” (Supplementary Excel File S18). 16 genes were common for two sets, which is expected, given the highly-connected hub proteins participating in the overlapping signaling pathways. The “universal” genes had several distinct features. First, they were disproportionally enriched with oncogenes and tumor suppressor genes, many of which were altered in at least one cancer. In total, both lists contained 20 tumor suppressor genes and 35 known oncogenes. Interestingly, all altered tumor suppressors were either mutated or deleted, which is expected in invasive cancers. However, only 33% of oncogenes were amplified (an alteration assumed to “match” with the phenotype; the remaining 67% were either deleted or mutated). Second, both sets featured a disproportionally high fraction of transcription factors: 62% for upstream genes and 36% for downstream genes, a 10- and 5-fold enrichment compared to the 5.3% TF content of the human proteome (MetaCore database). Third, essentially all universal genes formed a tight network module, inter-connected by one-step physical protein interactions. The network mostly connected two series of pathways: regulation of proliferation and apoptosis and regulation of angiogenesis (Fig. 5). Finally, ontology enrichment distributions for upstream and downstream universal genes had exceptionally low p-values, which indicate high degree of functional connectivity between the encoded proteins, and many common pathways and processes and were highly synergic (Supplementary Excel File S19, S20). Not surprisingly, “universal” cancer genes most densely...
populated key signaling pathways involved in carcinogenesis: Brca1-regulated transcription in DNA damage, PTEN pathway, TGF and WNT pathway in cytoskeleton remodeling, G1/S cell cycle transition, p53 signaling, AKT signaling and many others.

**Integrative network models for each cancer type**

As described above, we produced three “derivative” gene sets linked with the altered genes in each cancer type: 1. Genes from “synergistic” pathways where mutated, amplified and deleted genes co-operate; 2. A set of “universal” cancer genes defined as genes which commonly interact with all 12 datasets; 3. A set of “over-connected” genes defined as the genes with a disproportionately high number of protein interactions with the genetically modified genes in each cancer. We combined all three sets to come up with pathway models showing common and unique features for each cancer (Fig. 6, 7). The most common feature for all cancers comprises two overlapping sub-networks involved in regulation of the balance between cell proliferation and apoptosis (with the largest hub being p53) and regulation of angiogenesis (with the main pathway VEGFR – PKC – MEK1/2 – IRK ½ - NF-kB – c-Jun). A third, smaller sub-network includes a transcription modulation complex of Histone H3, GCNS, p300, SAP30, and histone deacetylase class I, which is coupled with NOTCH signaling and NF-kB signaling. These common signaling and effector pathways of carcinogenesis are downstream of cancer-specific signaling defined by unique “over-connected” genes for each cancer. In the case of breast cancer, the most prominent unique signaling is through RNF20 kinase; for colon cancer it is endothelin-EGF-VEGFA signaling; for pancreatic cancer thrombospondin 1 – CD36 - p38MAPK signaling for apoptosis, DLL10 NOTCH for transcription, netrin 1 – DSCAM – UNC5B signaling leading to nerve ingrowth process accompanying vascular ingrowth. In the case of glioblastoma, the central unique element is nucleostemin, which is shown to participate in carcinogenesis. This gene is connected with MDM2, TRF1 and p14 ARF, which are linked to p53 to carry out regulation of the balance between apoptosis and proliferation (Fig. 5,B).

**Discussion.**

We compared the sets of three genetic alterations: sequence alterations, deletions and amplifications in breast, colon, and pancreatic cancer and glioblastoma using multi-patient pooled re-sequencing data (11, 12, 14, 15). We aimed to understand the relative impact of two factors: the cancer
type (tissue and microenvironment) and the alteration type in each dataset and to identify the commonalities and uniqueness in the functional landscape evolved in cancer alterations.

Most of the genetically modified genes were unique in each dataset with very limited overlap. Vast majority of altered genes were altered uniquely in just one cancer type but not the others. Enrichment analysis also showed that the narrowly defined pathways and processes are uniquely affected in each set. This observation supports the general knowledge of high heterogeneity among human cancers.

By definition, each set of altered genes was impacted by both the type of genetic alteration and the type of cancer. Using kappa statistics (31), we calculated the relative impact of both “cancer” and “alteration” trends in the datasets and found that the type of alteration is relatively more important than the type of cancer at both the gene content and functional distribution levels. However, we found that common upstream regulation of the altered genes is cancer-specific rather than alteration type-specific.

None of the 12 datasets was functionally self–sufficient, i.e. the altered genes of any type would not form concise networks and did not belong to a particular process or pathway. This assumes functional cooperation between different types of altered proteins in carcinogenesis.

As most altered genes were unique for each set, the sets were essentially incomparable at gene level. However, we applied several methods of functional analysis to establish connectivity and functional co-operation between the sets. Thus, we calculated the degree of co-operation using the enrichment synergy approach (29) and found that TGF-beta signaling, EGFR signaling via PIP3, G1/S transition in cell cycle and several development pathways are the most synergistic among all cancers. These are among the very few signaling pathways shown to be commonly affected by driver somatic mutations in different cancers (11-13, 15). The synergy findings further confirm the importance of these key pathways in carcinogenesis.

Although the primary alteration sets did not statistically significantly overlap, their interaction partners formed a cohesive set of 65 “universal cancer genes” (37 up-stream and 47 down-stream genes, with 16 overlapping genes). Unlike the primary alteration sets, the “universal” genes formed a highly significant direct interaction network and, as a set, were highly enriched in cancerogenesis pathways and processes. Many “universal” genes are known as cancer suppressor genes and oncogenes. Interestingly, all “universal” tumor suppressors happen to be either mutated or deleted (but never amplified) in at least one cancer, which indirectly confirms the high quality of the original data. In contrast, oncogenes were equally likely to be amplified, deleted or mutated.
We have also identified the sets of transcription factors, ligands, receptors, kinases, and proteases over-connected with the genes from the set via protein interactions. The degree of over-connectivity reflects the relative functional importance of a gene for the dataset (29). In general, intra-connectivity within the sets was lower than expected, which suggests a low functional correlation between mutated, amplified, or deleted proteins in any cancer and the necessity of co-operation between differently modified proteins and normal proteins in pathways and processes. This is in line with our earlier observation of a lack of functional self-sufficiency within amplicons in breast cancer (29). Interestingly, the 12 alteration sets were clustered mostly in accordance to cancer type, not alteration type, when compared by common upstream regulators.

Analysis of the “derivative” gene lists allowed us to hypothesize that carcinogenesis may be carried out as a combination of common signaling and “effector” pathways (cell cycle regulation, cytoskeleton remodeling, development signaling etc.) populated by highly connected hub proteins (encoded by “universal” cancer genes such as p53, Brca1, beta-catenin, SMADs etc.) and unique “boutique” pathways defined by cancer-specific proteins with very few interactions. This model is similar to a road map which combines highly travelled (but few) freeways and hierarchically organized networks of small country roads, unique for the geography. It has been shown that, indeed, a combination of common hubs and proteins with few interactions defines the uniqueness of biological networks and their so-called “small world and scale-free” behavior (34). This model also helps to explain poor overlap among the genetically altered gene sets and very high heterogeneity of genetic modifications in individual tumors in general. We have built cancerogenesis models for all four cancers using the universal cancer genes, unique over-connected genes and common “synergistic” pathways as building blocks. The models, although different in details, are all centered on the two subnetworks consisting of “universal genes” which control regulation of key decision points in carcinogenesis, including apoptosis, cell proliferation, and angiogenesis.

Sampling error may play a role in data interpretation. Very likely, there would be more unique cancer- and alteration-level genes identified with the larger number of primary tumor samples re-sequenced, due to high genetic heterogeneity of human cancers. However, we believe that the main findings of our study would not be challenged as we already deal with to a large extend unique datasets (cancer type x alteration type) as gene content does not substantially overlap. It is an interesting statistical problem to calculate the number of samples needed to exhaust the pool of genetic alterations for each cancer type but this is beyond the scope of the current study.
Acknowledgements

This study was supported in part by National Institutes of Health grants CA134175 and CA121113.
References:


10. Vogelstein B. The sequence of all 185,000 coding exons in each of 100 human tumors: What has it taught us? AACR 101st Annual Meeting 2010


32. Diestel R. Graph Theory. 2005: Springer.
**Figure legends**

**Figure 1.** Gene content overlap between the sets of genetically modified genes. A. Overlap at the level of cancer type; B. Overlap at the DNA alteration level; C. Overlap between non-redundant unions.

**Figure 2.** Kappa-based similarity trees between sets of mutations, deletions and amplifications for colon, breast, pancreatic cancers and glioblastoma. A. similarity of gene content; B. similarity of enrichment distribution in the GO processes; C. Similarity of the sets of upstream regulators for each alteration set. Upstream regulators were identified as the one-step “neighbors” of the genes in the alteration sets defined by directed binary protein-protein interactions. Red boxes - similarity is predominately defined by the DNA alteration level; blue boxes – similarity is mostly defined by the type of cancer.

**Figure 3.** Network topology analysis of genetically modified genes in four cancer types.

**Figure 4.** Hierarchical clustering of protein function enrichment. Red color on the heat map denotes a high positive z-score (over-abundance of a particular protein class); green color denotes a negative z-score (protein class is under-represented). Deletion sets in all cancers except for colorectal formed a tight cluster (blue box). Mutation sets for the same cancers formed another cluster (red box).

**Figure 5.** Universal cancer genes (UCGs). A. A direct interaction network for UCGs. The network is functionally divided into regulation of proliferation and apoptosis and regulation of angiogenesis. B. Connectivity between UCGs and unique over-connected genes for each cancer. The network objects represent proteins and complexes; the links between objects – biological interactions (green links for activation, red links for inhibition). Please see Supplementary Fig. S5 for a complete legend on objects and links.

**Figure 6.** Examples of cancer-specific connectivity between universal genes, unique overconnected genes and pathways co-operatively affected by genetic alterations for breast cancer and colon cancer. A. Breast cancer; B. Colon cancer. The network objects represent proteins and complexes; the links
between objects – biological interactions (green links for activation, red links for inhibition). Please see Supplementary Fig. S5 for a complete legend on objects and links.

**Figure 7.** Examples of cancer-specific connectivity between universal genes, unique over-connected genes and pathways co-operatively affected by genetic alterations for pancreatic cancer and glioblastoma. A. Pancreatic cancer; B. Glioblastoma. The network objects represent proteins and complexes; the links between objects – biological interactions (green links for activation, red links for inhibition). Please see Supplementary Fig. S5 for a complete legend on objects and links.
Functional synergies yet distinct modulators affected by genetic alterations in common human cancers


Cancer Res  Published OnlineFirst March 11, 2011.