SCIENTIFIC REPORTS
nature research

OPEN

Human transcription factor and protein kinase gene fusions in human cancer

Kari Salokas, Rigbe G. Weldatsadik & Markku Varjosalo✉

Oncogenic gene fusions are estimated to account for up-to 20% of cancer morbidity. Recently sequence-level studies have established oncofusions throughout all tissue types. However, the functional implications of the identified oncofusions have often not been investigated. In this study, identified oncofusions from a fusion detection approach (DEEPEST) were analyzed in detail. Of the 28,863 oncofusions, we found almost 30% are expected to produce functional proteins with features from both parent genes. Kinases and transcription factors were the main gene families of the protein producing fusions. Considering their role as initiators, actors, and termination points of cellular signaling pathways, we focused our in-depth analyses on them. Domain architecture of the fusions and their wild-type interactors suggests that abnormal molecular context of protein domains caused by fusion events may unlock the oncogenic potential of the wild type counterparts of the fusion proteins. To understand overall oncofusion effects, we performed differential expression analysis using TCGA cancer project samples. Results indicated oncofusion-specific alterations in gene expression levels, and lower expression levels of components of key cellular pathways, in particular signal transduction and transcription regulation. The sum of results suggests that kinase and transcription factor oncofusions deregulate cellular signaling, possibly via acquiring novel functions.

At any given moment, multitudes of molecular networks are activated in cells throughout the body. An important feature of these networks is highly concerted regulation of key signaling, and deviation from homeostasis can result in diseases, such as cancer. Cancer is a complex, progressive, multi-step disorder, which stems from mutations caused by genomic instability[1]. The accumulation of genetic and epigenetic abnormalities ultimately leads to the transformation of normal cells into malignant derivatives. Two highly enriched gene groups being mutated in the majority of cancer types are protein kinases (PKs) and transcription factors (TFs)[2,3]. PKs mediate most signal transduction events in cells by phosphorylation of specific substrates, thus modifying their activity, cellular localization, and/or association with other proteins. TFs are the "transistors" of the cellular signaling circuits, controlling the transcriptional outcome of activated signaling by binding to regulative elements of their corresponding target genes and driving or suppressing their expression. Therefore, it is easy to understand why mutational deregulation of these two gene groups can have such an impact on tumorigenesis.

In addition to harboring activating or inactivating somatic point mutations, PKs and TFs account for a large fraction of all human fusion genes involved in cancer (COSMIC, Catalogue of Somatic Mutations in Cancer, cancer.sanger.uk[4]; and dbCRID, Database of Chromosome Rearrangements in Disease[5]). Chromosomal translocations creating fusion genes are among the most common mutation class of known cancer genes, and they have long been identified as driver mutations in certain types of cancer[6]. Recently, oncogenic fusion genes (hereafter oncofusions, OFs) have been found in many hematological and solid tumors, demonstrating that translocations are a common cause of malignancy[7,8]. Fusion mutations occur when two different gene regions fuse together via translocation. Examples of consequences of chromosomal fusion to protein structure range from missense mutations to expression-change inducing promoter-gene –combinations to fully functional fusion proteins with neomorphic properties. A classic example of gained functions is the breakpoint cluster region-Abelson tyrosine-protein kinase 1 (BCR-ABL1) translocation in chronic myeloid leukemia[9]. Alternatively, a proto-oncogene is fused to a strong promoter, and thereby the oncogenic function is upregulated due to the strong promoter of the upstream fusion partner. This is common in lymphomas where oncogenes are juxtaposed to the promoters of the immunoglobulin genes[10], and also in prostate cancer where ETS TF (ERG) is fused with TMPRSS2 regulatory

Systems Pathology/Biology Research Group, Institute of Biotechnology, HiLIFE, University of Helsinki, Helsinki, Finland. ✉email: markku.varjosalo@helsinki.fi

sequences, thus obtaining androgen receptor (AR)-responsive expression[11]. The current understanding favors the aberrant gene function model rather than promoter-induced over-expression.

The frequency of recurrent OFs varies depending on the specific type of cancer[12–15], but identified translocations are estimated to account for up to 20% of cancer morbidity[8]. Recent fusion prioritization study found that in-frame transcripts were the most powerful predictor of driver fusions[16], confirming the intuition that in-frame transcripts are crucial to function. Notably, breakpoints were also observed to preferentially avoid splitting of domains. Together with frame-shift conservation, such trends could reflect a selection on fusion proteins to maintain protein stability and evade degradation pathways[17].

Next-generation sequencing (NGS) of genomes and transcriptomes from primary human cancer cells is constantly revealing new gene fusions that are involved in driving tumorigenesis; including examples found in colorectal carcinoma, bladder carcinoma, breast cancer and acute lymphoblastic leukemia (ALL)[15,18–20]. Furthermore, NGS has provided enough detailed sequence information of the fusion breakpoints allowing us to initiate systems-level research on human oncofusions. As a result, various algorithms have been developed to mine OFs from large cancer datasets such as TCGA. However, the concordance among the different algorithms is very low that metacaller approaches utilizing consensus calls have been employed[21], which limit novel OF discoveries. Recently a new statistical method, DEEPEST[22], was developed to overcome these limitations. In this study, oncofusions that involve PKs and TFs were selected from the data produced by DEEPEST applied to the TCGA dataset.

In most cases, it is not possible to draw definite conclusions about the mechanisms or extent by which individual translocations contribute to cancer. Predicting protein function from a sequence has proven an extremely difficult task. With gene fusions, the task is even more daunting. However, an unexpectedly large number of PKs and TFs have been found to be mutationally activated or have increased expression due to gene amplification or translocation in cancer[6]. The high number of PKs and TFs with relatively low individual mutational frequency suggests either that a large number of signaling pathways can contribute to cancer, or that many PKs and TFs can regulate the same pathways when activated unphysiologically. Some additional support for this hypothesis comes from the interconnectivity of the PK-/TF-oncofusions.

In this study, fusions predicted to produce in-frame proteins were analyzed to understand the protein-level implications of fusion events. The fusions were analyzed from the perspective of their domain architecture to understand likely modes of action of the novel proteins. Furthermore, known interactomes of the participating wild type proteins were used to determine possible mechanisms of action, pathways of interest, and possible treatment vectors for affecting as many different fusions as possible. As a result, multiple cellular signaling pathways were found to intersect with major subsets of these fusions, and multiple individual key interactors, such as NTRK1 with over 200 and EGFR with over 100 interacting fusions, were identified as potential targets of interest.

Materials and methods

Fusion selection and annotation. Fusions that involve protein kinase genes[23] and transcription factors[24] were selected from the 31,007 fusions that were identified by applying DEEPEST to the whole TCGA dataset[22]. Of these 28,862 were determined to be unique by considering Ensembl gene IDs, biotypes, chromosomal breakpoints, AGFusion assigned fusion effects, and resulting protein sequences. AGFusion was used to annotate these gene fusions to the human genome assembly GRCh38 v.89 from Ensembl. For analysis involving gene pairs, the pair entry was used in alphabetical order (e.g. ERG-TMPRSS2 instead of TMPRSS2-ERG) in all cases. Fusions were considered protein coding if both genes contributed over 30 amino acids to the product.

Clinical data. Clinical data for TCGA samples was obtained from the GDC data repository. The data was matched to AGFusion output data based on TCGA barcode (e.g. TCGA-WB-A80K) using a custom in-house python script. Stage information from the clinical data was simplified where possible (e.g., Stage IIA was changed to Stage II). Entries such as Stage 0, Stage X and I/II NOS were ignored. Tissue entries were simplified from detailed ICD-O 3 topographical codes to more general, e.g. C56.9 → C56, and mapped to names accordingly. Chromosomal sequence information from GRCh38 v.89 was used to categorize breakpoints into 5% chromosomal interval groups.

Interactor analysis. Interactors for wild type proteins of all fusion partners were obtained from IMEx consortium[25] and any interactions that were not confirmed to be physical by experimental methods were discarded. Interactors were added to the interactor set from each fusion, while leaving out the fusion pair genes themselves. Annotations for interactors were obtained using Uniprot and Reactome. From Reactome, mappings to all levels of pathway hierarchy were used. Dijkstra's algorithm[26] implemented with a custom python script was then used to establish shortest paths to Reactome root nodes for each network node. A weight of 1 was used for all network edges.

Domain annotation. For the protein producing fusions, sequences of the protein products were produced using the AGFusion tools. Duplicate fusions based on fusion genes and protein sequence were discarded. Domains were taken from AGFusion output, and mapped to protein sequence in the wild type protein. The intactness of domains was then determined by matching the WT domain sequence to the predicted fusion protein sequence, and only full length, intact domains were picked for further study. A domain was classified as PK- or TF-specific if ≥ 95% of all its occurrences were in PK or TF proteins, respectively.

Data visualization. Data illustrations were made with CorelDRAW, Excel, and in-house python scripts using Matplotlib and Seaborn. Cytoscape[27] was used for creating network figures.

Differential expression analysis. Gene expression quantification HTSeq-counts –files were downloaded from GDC data portal. Samples where OFs with intact, full-length PK, or TF domains were detected were grouped together based on fusion gene pairs. The groups were then analyzed with DESeq2[28] using other samples with protein producing non-PK/-TF fusions as controls. For each pair group, differential expression analysis against an equal number of control samples picked from samples in which other protein-producing fusions were found. Analysis was repeated 1,000 times for each fusion pair. For the resulting significantly differentially expressed genes (q-value < 0.05), basemean and expected values were averaged across all runs, and a fold change value calculated based on these. GO annotations were then added from ensembl annotations, and Reactome pathways from first mapping ensembl gene IDs to Uniprot via Ensembl BioMart, and then to Reactome lowest level pathway terms via Reactome. Z-score value for pathway level over-/underexpression was calculated by a method used in GOplot[29] i.e.by deducting the number of underexpressed genes from the number of overexpressed genes and dividing the result by the square root of the number of significantly changed genes (FDR corrected $p \leq 0.05$).

Results

Detection of oncofusions from TCGA dataset reveals enrichment of PK and TF fusions. In this study, we focused on protein producing OF genes. Translocation of chromosomal regions can result in either in-frame or out-of-frame OFs (Fig. 1A). To characterize the proteins produced by known OFs in the TCGA dataset, which currently contains data from 33 different cancer projects, we launched an analysis to understand the potential functional space of the protein producing fusions (Fig. 1B), and especially those that involve either a PK or a TF, or both (PK-TF fusions).

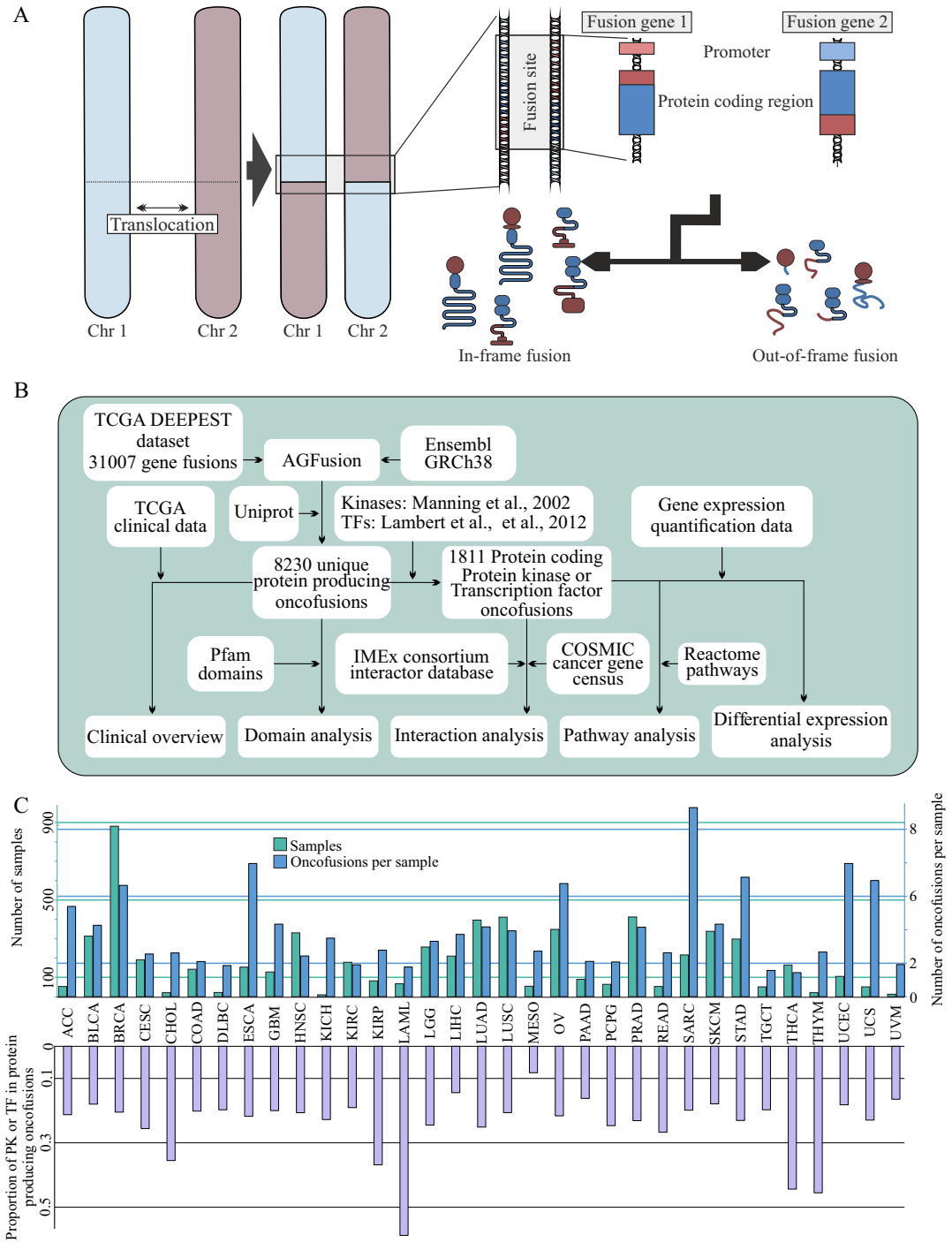
The DEEPEST dataset included 31,007 fusions detected from 6,123 cancer samples. Of these, 28,862 were unique fusions (Fig. 1C, upper panel). Among the unique OFs, 29% (8,230) were predicted to retain frame, and also produce potentially functional proteins, where both genes contributed over 30 in-frame amino acids (Fig. 1B, Supplementary table S1). The limit of 30 amino acids was the length of the shortest non-repeat domain present in the fused proteins. Examining the resulting protein producing OF set, we noticed an abundance of those involving PK or TF. Indeed, these fusions constituted 1,811 protein producing OFs (Fig. 1C, lower panel). Generally the proportion of PK/TF fusions was under 0.3, except in the PK/TF—fusion prone cancers acute myeloid leukemia, cholangiocarcinoma, thyroid carcinoma, and thymoma. The number of OFs per sample varied across cancer types. The types most prone to protein producing fusions were sarcoma (SARC) with an average of 3.7 protein producing fusions per sample, esophageal carcinoma (ESCA: 3.5 fusions), uterine corpus endometrial carcinoma (UCEC: 2.9), stomach adenocarcinoma (STAD: 2.8), breast invasive carcinoma (BRCA: 2.7), uterine carcinosarcoma (UCS: 2.6), and ovarian serous cystadenocarcinoma (OV: 2.5).

Due to the prevalence of PK and TF genes in the fusions, we next investigated if they are enriched in particular cancers. While in most cancers PK/TF fusions made up around 20–25% of all protein producing OFs, the percentage reached 60% in acute myeloid leukemia (LAML) samples , 46% in thymoma (THYM), 45% in thyroid carcinoma (THCA), and 37% and 36% in kidney renal papillary cell carcinoma (KIRP) and cholangiocarcinoma (CHOL) respectively (Fig. 1C). Acute myeloid leukemia is well known as an OF-prone cancer[30] However, aside from the four fusions detected between ABL1 and BCR, the high percentage was mostly TF-driven, with KMT2A, RUNX1, and RARA being found in 9, 6, and 4 fusions respectively. This is in contrast to the peak in THCA, which is driven by 12 BRAF fusions, 11 fusions of RET, 6 of NTRK1, and 5 of NTRK3, among 8 other protein kinases.

Reading frame retention is common in PK and TF oncofusions. The 31,007 fusions consisted of 23,354 unique gene pairs and 14,632 individual genes; 14,338 of the pairs did not have any protein producing fusions. The top protein producing fusion was RPS6KB1-VMP1 with 13 unique protein producing fusions in the dataset, all the others having less than 10 (Fig. 2A). There were 47 fusion gene pairs that were predicted to produce protein in at least 4 fusions, 159 in 3 fusions, 835 in 2, and 7,975 in 1 fusion. Out of the 32 fusion gene pairs that produced 4 or more unique proteins, 15 were PK/TF fusions.

To better understand the behavior of prolific gene pairs, we next mapped tissue annotations from TCGA to fusions of each gene pair based on barcodes from samples, where a fusion of the gene pair was present. In contrast to RPS6KB1-VMP1 and ITGB6-RBMS1, which were seen in samples from 6 different cancers, 7,055 pairs were seen in samples of only one cancer type. Out of these cancer-specific fusions, 38 were predicted to produce 2 or more unique proteins (with ERG-TMPRSS2 predicted to produce 4 different unique proteins, supplementary table S2). PK/TF fusions featured 1,449 different PK or TF genes, ERG being the most common TF, and ERRB2 the most common PK (supplementary table S3). Between 84 and 97 percent of oncofusions in each cancer project were unique, highest being sarcoma with 97% unique gene pair combinations, and thyroid carcinoma the lowest with 84% (supplementary table S4). Protein producing fusions followed a similar theme, unique protein producing fusions making up between 23 and 52% of all oncofusions in each given cancer project (supplementary table S4).

We next looked in more detail what cancer stages PK and TF fusions were detected in. The most prominent group was stage II breast invasive carcinoma, which also had the most samples in the data set (Fig. 2B). In total, of the 1,811 PK and TF fusions, 271 were found in stage I samples, 444 in stage II, 303 in stage III, and 130 in stage IV. On average, samples had 0.30 PK/TF fusions per sample. However, in some cancers, PK or TF fusions are enriched towards the more severe stages. Discounting stage groups with less than 10 samples, 4 groups had more than 0.6 fusions per sample. ESCA stage III samples in particular had 0.76 PK/TF fusions per sample, while STAD and BRCA stage IV samples had 0.69 and 0.65 respectively, and STAD stage I had 0.61 (Supplementary table S5). The distribution of protein producing OFs mirrored that of PK/TF fusions quite closely (supplementary figure S1A). In terms of chromosomal breakpoint locations, those in the PK/TF fusions varied compared



ACC	-	Adrenocortical carcinoma	KICH	-	Kidney chromophobe	PRAD	-	Prostate adenocarcinoma
BLCA	-	Bladder urothelial carcinoma	KIRC	-	Kidney renal clear cell carcinoma	READ	-	Rectum adenocarcinoma
BRCA	-	Breast invasive carcinoma	KIRP	-	Kidney renal papillary cell carcinoma	SARC	-	Sarcoma
CESC	-	Cervical squamous cell carcinoma and endocervical adenocarcinoma	LAML	-	Acute myeloid leukemia	SKCM	-	Skin cutaneous melanoma
			LGG	-	Brain lower grade glioma	STAD	-	Stomach adenocarcinoma
CHOL	-	Cholangiocarcinoma	LIHC	-	Liver hepatocellular carcinoma	TGCT	-	Testicular germ cell tumors
COAD	-	Colon adenocarcinoma	LUAD	-	Lung adenocarcinoma	THCA	-	Thyroid carcinoma
DLBC	-	Lymphoid neoplasm diffuse large B-cell lymphoma	LUSC	-	Lung squamous cellcarcinoma	THYM	-	Thymoma
			MESO	-	Mesothelioma	UCEC	-	Uterine corpus endometrial carcinoma
ESCA	-	Esophageal carcinoma	OV	-	Ovarian serous cystadenocarcinoma			
GBM	-	Glioblastoma multiforme	PAAD	-	Pancreatic adenocarcinoma	UCS	-	Uterine carcinosarcoma
HNSC	-	Head and neck squamous cell carcinoma	PCPG	-	Pheochromocytoma and paraganglioma	UVM	-	Uveal melanoma

◂**Figure 1.** Schematic illustration of the gene fusions, workflow, and the number of gene fusions in human cancer. (**A**) Schematic description of gene fusions formation. Fusions are formed mainly via balanced and unbalanced chromosomal rearrangements, such as translocations, deletions, inversions and insertions. This usually leads to formation of a fusion gene with the 5′ end of Gene 1 and 3′ end of Gene 2. If the fusion occurs between two protein coding genes, depending on whether the reading frame is violated, and where exactly the fusion occurs, a fusion protein may be transcribed with features and domains from both partners. Other possible outcomes include full or truncated 3′ gene under the control of the promoter of the 5′ gene. (**B**) Workflow used in this study. Analysis progressed from the total set of fusions discovered by the DEEPEST method[22] and moved towards more specific kinase / TF containing, protein producing oncofusions. We started with TCGA data-based fusion set from Dehghannasiri et al. (2019), for which we generated protein sequences with AGFusion. Domains were added by matching sequence to Uniprot proteins annotated with Pfam domains, after which non-unique entries were dropped. Fusions were classified as protein producing, if both gene fragments were predicted to produce > 30 AA of protein sequence. From this set, the two most prominent protein groups were protein kinases and transcription factors, and thus we focused further analysis on the 1,811 unique protein kinase or transcription factor containing fusions, using the full protein producing fusion set for comparison. Known interactions for wild type fusion proteins were obtained from IMEx consortium, and used for estimating maximal foreseeable effect on signaling pathways from Reactome. Finally, TCGA gene expression quantification data was used to probe observable effects of kinase/TF fusions, using other protein producing fusions as background. (**C**) Top: Breakdown of samples and fusion mutations by TCGA project. Largest single contributor of samples with fusions was TCGA breast invasive carcinoma project (BRCA), which had the highest number of samples and identified fusion mutations. Bottom: Proportion of protein producing fusions that include PK or TF genes.

to all protein producing fusion mutations, but the prominent role of PK/TF fusions is illustrated by overlapping hotspots (supplementary figure S2).

Intact, in-frame domains are commonly retained in OFs. To understand the contribution of each OF to the overall development or survival of cancerous cells, the functional consequences of any given mutation and its impact on the pathways the proteins are involved in must be understood. To this end, we analyzed all identified unique protein producing fusions, and the full-length, in-frame domains of the fusion proteins.

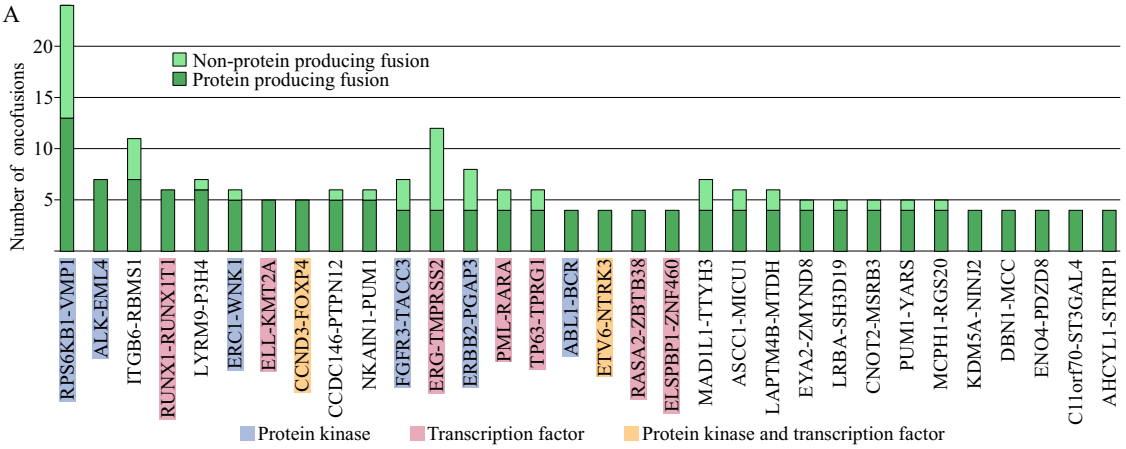
While AGFusion does predict protein sequence for each fusion partner, and corresponding conserved or lost domains, a domain is counted as conserved already if only 5 amino acids are included in the sequence. To adapt this to the study of full-length domains, we first mapped the Pfam identifiers of the domains to sequences in the wild type proteins from Uniprot. The domains were then defined as conserved only if the full sequence was present in the fusion protein. This resulted in 10,100 conserved domains in all protein producing fusions. Over 50% (5,373) of these domains are in PK/TF fusions, which account for 22% of all protein producing fusions (supplementary tables S1, S6), suggesting overall domain count strongly favors PK and TF genes, perhaps indicating that these fusions produce more functional proteins in comparison to all protein producing fusions.

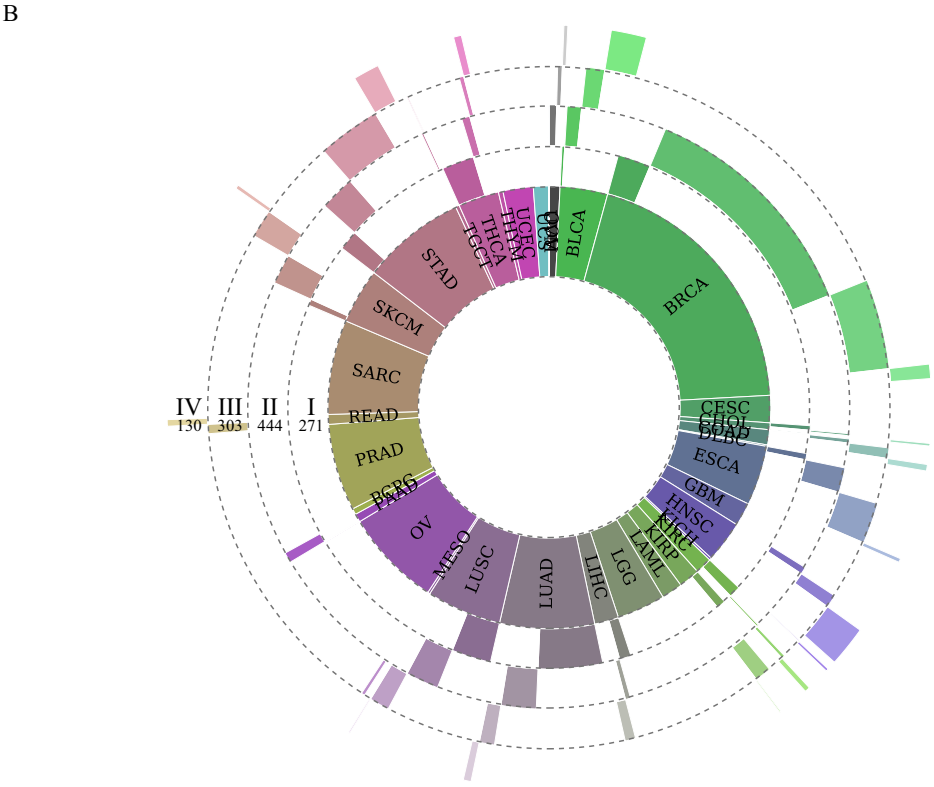
The most conserved domain was the protein tyrosine kinase domain (Fig. 3A, supplementary table S6), which was conserved in 159 fusions. This was followed by the PH domain, a common domain in intracellular signaling proteins and proteins of the cytoskeleton, and the protein kinase domain. To assess retention of non-obvious PK or TF domains, we classified domains as PK or TF specific if over 95% of the copies were found in PK or TF halves of the fusion proteins. This resulted in 622 copies of 131 different TF-specific domains predicted to exist in the fusions, compared to 455 copies of 44 PK-specific domains. Most common TF domains were zinc finger C2H2 type, KRAB, and HLH DNA binding domains, present in 59, 45, and 43 copies respectively. Many TF domains, such as KRAB, are involved in both transcriptional activation and repression, depending on the molecular context.

On average, protein producing fusions in samples of most cancer projects tended to have close to 1 intact, full length domain per protein producing OF. PK/TF fusions on average had more intact domains in all except for 5 projects (Fig. 3B). On average, fusions in all projects tended to have between 1 and 2 intact domains, while PK/TF fusions featured a slightly higher average. Although some cancers do appear to have particularly many domains, this is mostly due to low count of fusions detected in the project. Exception seems to be acute myeloid leukemia, with 47 detected protein producing fusions, 28 of which contain either a PK or a TF. Most striking differences being seen in mesothelioma, thyroid carcinoma, rectum adenocarcinoma, and uveal melanoma with 1.17, 1.14, 1.0, and 1.0 more retained domains on average in PK/TF fusions compared to protein producing fusions, respectively.

On the cancer project level, thyroid carcinoma had the highest percentage of PK domains (19% of all domains identified in the 97 samples of the project, supplementary table S7, supplementary Figure S3), which totaled to 34, only exceeded by breast invasive carcinoma with 92 PK specific domains (4% of all BRCA domains), and lung adenocarcinoma (LUAD) with 35 (6%). Proportion of TF domains varied less. Kidney renal papillary cell carcinoma had 15% of its intact domains in the TF-specific set, followed by acute myeloid leukemia with 12%, and rectum adenocarcinoma and prostate adenocarcinoma, both at 11%. Aside from prostate adenocarcinoma, these projects had <50 samples in the TCGA dataset.

Interactors of fusion partners can point to impact of OFs. To understand what kind of implications the functional changes of lost / conserved PK or TF specific domains in new combinations could have for the





ACC - Adrenocortical carcinoma
BLCA - Bladder urothelial carcinoma
BRCA - Breast invasive carcinoma
CESC - Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHOL - Cholangiocarcinoma
COAD - Colon adenocarcinoma
DLBC - Lymphoid neoplasm diffuse large B-cell lymphoma
ESCA - Esophageal carcinoma
GBM - Glioblastoma multiforme
HNSC - Head and neck squamous cell carcinoma
KICH - Kidney chromophobe
KIRC - Kidney renal clear cell carcinoma
KIRP - Kidney renal papillary cell carcinoma
LAML - Acute myeloid leukemia
LGG - Brain lower grade glioma
LIHC - Liver hepatocellular carcinoma
LUAD - Lung adenocarcinoma
LUSC - Lung squamous cellcarcinoma
MESO - Mesothelioma
OV - Ovarian serous cystadenocarcinoma
PAAD - Pancreatic adenocarcinoma
PCPG - Pheochromocytoma and paraganglioma
PRAD - Prostate adenocarcinoma
READ - Rectum adenocarcinoma
SARC - Sarcoma
SKCM - Skin cutaneous melanoma
STAD - Stomach adenocarcinoma
TGCT - Testicular germ cell tumors
THCA - Thyroid carcinoma
THYM - Thymoma
UCEC - Uterine corpus endometrial carcinoma
UCS - Uterine carcinosarcoma
UVM - Uveal melanoma

Figure 2. Clinical characterization of protein producing oncofusions by cancer stage. (**A**) The most common protein producing gene pairs in oncofusions. In total, protein producing fusions were comprised of 23,354 unique gene pairs predicted to produce one or more unique protein products. The most common pair was RPS6KB1-VMP1, with over 10 unique proteins, followed by ITGB6-RBMS2 and ALK-EML4 with 7 each, and LYRM9-P3H4 and RUNX1-RUNX1T1 at 5. Kinase and TF fusions were common in top protein producing gene pairs, illustrated by blue shading for the presence of a protein kinase in gene pair, red for TF, and orange for both. (**B**) Sunburst diagram of project and stage distribution of PK/TF oncofusions. The innermost layer represents the number of fusions in each project. The layers radiating out are the proportion of fusions detected in Stage I, II, III, and IV samples, in order from in to out. Total numbers of fusions from each stage is marked under the stage indicators.

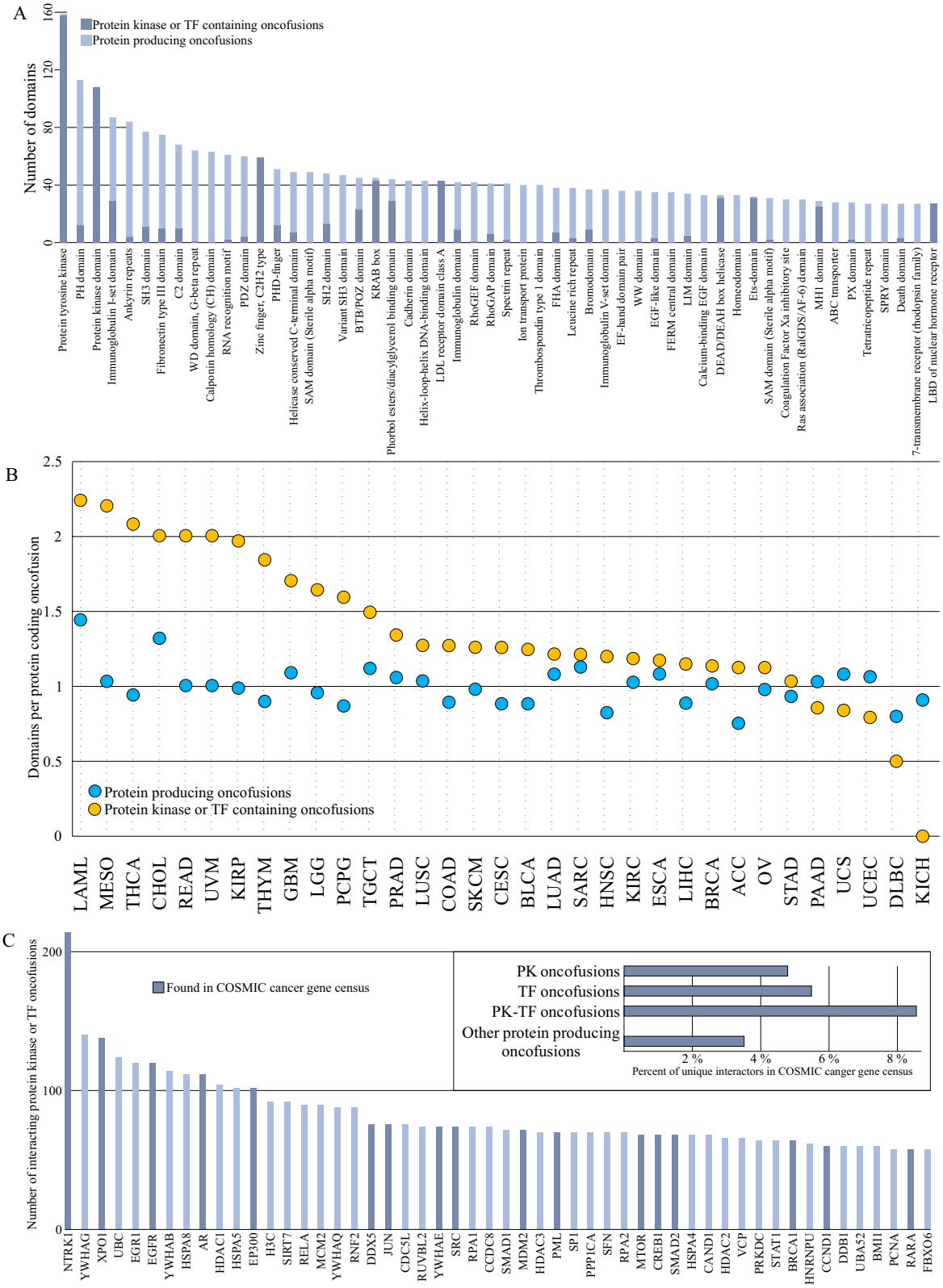


Figure 3. Domain analysis of protein producing fusions. (**A**) Intact, full-length domains identified in unique protein producing fusions. In total, 10,100 intact domains were detected. The protein tyrosine kinase domain was the most prevalent with 159 identifications. In addition, protein kinase domain was detected with 108 copies each. Kinase or TF specific domains included 44 and 131 unique domains, respectively. 455 copies of kinase-specific domains were seen, and 622 of TF specific. Kinase domains focused more on the two top kinase domains, whereas TF domains were a much more evenly distributed group, the top TF-specific domain, C2H2 type zinc finger, having 59 copies. (**B**) The number of domains per protein coding oncofusion in the TCGA projects. (**C**) Most common interactors of protein kinase/TF fusions. Y-axis describes numbers of unique protein producing fusions, where one or both of the fusion partner WT genes interact with the protein. Top right inset: Proportion of interactors found in COSMIC cancer gene census is higher in both protein kinase and transcription factor fusions, and most common in fusions between protein kinase and transcription factor genes.


A
Transcriptional regulation pathways
RNA Polymerase II transcription
Signaling by Rho GTPases
MAPK family signaling cascade
GPCR signaling
Signaling by Notch
Generic transcription pathway
TGF-β signaling
Signal transduction
Hedgehog pathway
Signaling by Receptor tyrosine kinases
Transcriptional regulation by TP53
Integrin signaling
Intracellular signaling by second messengers
Death receptor signaling
Wnt signaling
Node proteins interact with
TF fusions
Kinase fusions
B
Cytokine Signaling in Immune system
Developmental Biology
Axon guidance
Metabolism
PIP3 activates AKT signaling
Intracellular signaling by second messengers
PTEN Regulation
Regulation of PTEN gene transcription
Oxidative Stress Induced Senescence
HDMs demethylate histones
SUMOylation of transcription cofactors
Transcriptional Regulation by E2F6
SUMOylation of chromatin organization proteins
Transcriptional regulation by RUNX1
Chromatin organization
Chromatin modifying enzymes
Cellular Senescence
SUMO E3 ligases SUMOylate target proteins
SUMOylation
Interferon alpha/beta signaling
EPH-Ephrin signaling
Phospholipid metabolism
EPH-ephrin mediated repulsion of cells
EPHA-mediated growth cone collapse
PI Metabolism
Synthesis of PIPs at the plasma membrane
Interferon Signaling
Metabolism of lipids
Oncogenic MAPK signaling
Signaling by moderate kinase activity BRAF mutants
Paradoxical activation of RAF signaling by kinase inactive BRAF
Signaling downstream of RAS mutants
rRNA modification in the nucleus and cytosol
RHO GTPases Activate Formins
Major pathway of rRNA processing in the nucleolus and cytosol
rRNA processing in the nucleus and cytosol
rRNA processing
Processing of Capped Intron-Containing Pre-mRNA
mRNA Splicing
mRNA Splicing - Major Pathway
Macroautophagy
Autophagy
Toll Like Receptor 3 (TLR3) Cascade
MyD88-independent TLR4 cascade
TRIF(TICAM1)-mediated TLR4 signaling
Toll-like Receptor Cascades
Toll Like Receptor 4 (TLR4) Cascade
Signaling by TGF-beta family members
Signaling by Nuclear Receptors
TCF dependent signaling in response to WNT
Signaling by WNT
Signaling by Receptor Tyrosine Kinases
Diseases of signal transduction
MAPK family signaling cascades
FLT3 Signaling
MAPK1/MAPK3 signaling
RAF/MAP kinase cascade
Signaling by Rho GTPases
RHO GTPase Effectors
M Phase
Cell Cycle Checkpoints
Hemostasis
Vesicle-mediated transport
Membrane Trafficking
Metabolism of RNA
Signaling by Interleukins
Transcriptional Regulation by Tp53
Innate Immune System
Adaptive Immune System
Infectious disease
Cell Cycle, Mitotic
Cell Cycle
Signal Transduction
Immune System
Metabolism of proteins
Generic Transcription Pathway
RNA Polymerase II Transcription
Gene expression (Transcription)
Disease
Post-translational protein modification
Cellular responses to external stimuli
Cellular responses to stress
0
1
Proportion of pathway in all pathways per disease
ACC
TGCT
DLBC
PAAD
UCS
LAML
UCEC
LUAD
SKCM
LIHC
ESCA
STAD
PRAD
OV
LUSC
BRCA
SARC
GBM
BLCA
KIRC
LGG
COAD
HNSC
CESC
CHOL
THCA
KIRP
READ
PCPG

Figure 4. Functional potential of Kinase/TF fusion interactors. (**A**) Interactors mapped to Reactome pathways. The interactors produced hits in almost 2000 pathways. Most prominent hits were centered around signal transduction pathways, which links to transcription events via TGF-β signaling pathway. The size of the node is directly proportional to number of fusions with interactors identified with the annotation from Reactome database. The used annotation file contained annotations for all levels of Reactome hierarchy. Included in the figure are pathways up to 7 steps away from the signal transduction root node. The node size is directly proportional to the sum of oncofusion interactors, and the count of fusions that interact with them. (**B**) Relative frequency of each pathway per TCGA project on a scale from 0 to 1 (1 being the pathway with most interactors). While pathways with the most potential interactors of fusions identified are the same in majority of the projects, different subpathways are seen in different projects, such as oncogenic MAPK signaling in DLBC and KIRP, or PI3-Akt signaling in UCS and CHOL.

cell and the organism as a whole, we next analysed the interaction networks of the wild type proteins in PK/TF fusion set.

Although PK or TF fusion proteins are likely to lose domains necessary for these interactions to form, they are also likely to instead gain domains facilitating new interactions. We took the known, experimentally validated interactomes of the wild type proteins from the IMEx consortium[25]. We treated the resulting interactor set as the hypothetical maximal foreseeable effect set, which consisted of interaction partners that may have an effect on the fusion protein, or that the fusion protein may have an effect on.

From this set, we found interactors that were particularly prominent. PKs, such as NTRK1, EGFR, and SRC, as well as various TFs, like NFKB3 and SP1 were among the top results (Fig. 3C, supplementary table S8). The list is mostly made up of other kinases or transcription factors, with NTRK1 potentially interacting with over 200 individual, unique fusions. Genes found in the COSMIC cancer gene census were more common towards high numbers of potentially interacting fusions. Through these interactions it is possible to identify significant central nodes through which multiple different fusions in different cancers may affect the growth of the tumor. For example, the second most common possible interactor, YWHAG, is a common regulator of signaling pathways. Approximately 7.7% of all interactors of protein producing fusions were found in the COSMIC cancer gene census, whereas the percentage rises to 29%, and to 40% if we consider only the 100 and 10 most common protein producing fusions respectively. Interactors of PK or TF fusions were more often seen in the cancer gene census, than those of other protein producing fusions (Fig. 3C upper right inset).

Pathway analysis of OF interactors highlights signal transduction and regulatory functions. Next, we combined Reactome pathway data to the interactor set, and built hierarchic networks of the found pathways (Fig. 4A). Considering the dataset, we focused on one network root node: signal transduction, and its descendants up to 7 links away. Another root node, gene transcription, can already be seen on this scale, which is unsurprising considering the inclusion of many TF fusions, and the interplay of signal transduction and gene transcription. For each Reactome pathway, we calculated an interactor count by adding together the number of potentially interacting fusions for each protein in the pathway.

Particularly enriched were proteins related to signal transduction, where interactors were detected in 15 branches from the root. Especially prominent pathways are those relating to receptor tyrosine kinase signaling with potential interactors from 1,230 unique PK/TF oncofusions), PI3K-Akt (1,220), Rho GTPases (1,095), Integrin (1,156) and GPCR (905) signaling, as well as MAPK family signaling cascades (1,053). Multiple smaller, but significant pathways such as Hedgehog, Notch and Wnt pathways are also seen. Generic transcription pathways and their related pathways, such as transcriptional regulation and RNA polymerase II transcription, are very prominent as well, with 1,503 PK/TF oncofusions.

The proportion of interactors from each pathway varied slightly between different cancers (Fig. 4B). While signal transduction was the most common pathway in most cancers, different signaling cascades, such as MAPK cascades or TLR cascades featured much more variation, pointing to relative enrichment of different pathways in different fusions, and perhaps to cancer-specific effects of unique gene-pair mutations in said cancers.

Oncofusions lead to distinct changes in gene expression. To understand if intact PK or TF domains had a recognizable and distinct downstream effect on gene expression, differential expression analysis was performed. Gene expression quantification result files were downloaded from GDC data portal, and divided into groups based on PK/TF fusion gene pairs.

Only gene pairs with conserved kinase or TF domains and at least two expression level quantification result files available were used (517 pairs). The analysis was repeated 1,000 times for each gene pair with random non-PK/TF protein producing control set, and results were filtered based on q value under 0.05. Results and expected values were then averaged across the replicates, and a fold change calculated.

In the results, 48,657 differentially expressed genes were thus identified (Fig. 5A, supplementary table S9). Overall results indicate mostly downregulation of the expression of majority of genes. The most common overexpressed genes were MTRNR2L1, SCGBID2, and CTAG2, seen in 346, 323, and 321 pair groups, respectively. Other common overexpressed genes were detected in under 300 pair groups. The most common underexpressed were PSPHP1, GSTM1, and PPP2R2C, detected in 245, 234, and 227 pair groups, respectively (Fig. 5B). Of all the differentially expressed genes, 713 were in the COSMIC cancer gene census. Most common overexpressed census genes were WIF1, SSX1, Pax7, PTPRT, and S100A7. These were identified as overexpressed in 285, 272, 262, 258, and 257 gene pairs, respectively. The most often underexpressed likewise were CDKN2A, SIX2, CCNE1, CNTNAP2, and MNX1, identified as underexpressed in 187, 174, 169, 162, and 161 fusion pair groups, respectively (supplementary table S10).

To gain a more complete image of what pathways each fusion mutation specifically affects, we performed analysis with Reactome annotations, and calculated a z-score for each annotation term. Only lowest level Reactome terms were used for mapping. This resulted in a total of 1887 pathways identified with a non-zero z-score (supplementary table S11). Of these, 53 displayed z-scores above 7.5 or below −7.5 (Fig. 5C), based on which gene pair groups fall into roughly four groups (Fig. 5 C, supplementary table S12), based on what kind of pathways the proteins the expression of which fusions up- or downregulate are in. The four groups are centered around complement and immunity regulation, ribosomal and exon junction complex functions, mitochondrial functions, and cell cycle related pathways.

Finally, we identified a total of 25 tumor suppressor genes (according to COSMIC cancer gene census), 30 common fusion genes, and 29 other oncogenes in these four pathway groups. Majority of these (23 fusion genes, 13 tumor suppressors, and 26 other oncogenes) were in the first group of immune system related pathways.

Discussion

We examined the gene fusion landscape in human cancer (from TCGA datasets). Gene fusions are among the most common mutation classes of known cancer genes[6], found both in hematological and solid tumors. Although the fusions can drive cancer via expression level changes when an oncogene is fused with a strong promoter such as TMPRSS2-ERG fusions in prostate cancer[31], we find that the majority 19,911 of the 28,863 oncofusions are in-frame mutations between exonic regions of two protein coding genes. In total, over 9,000 gene pairs were seen participating in fusions that were predicted to produce intact, potentially functional proteins. TCGA solid tumor samples tended to have fusions producing in-frame proteins of adequate length approximately 20% of the time in all diseases (Fig. 1C). Equal distribution across stages may hint at protein producing fusions being early events in the development of the tumors from which they were identified. We identified several particularly prolific fusion gene pairs, among them capturing also several that have been featured prominently in literature. Most prolific protein producing OF gene pairs were found to feature either a protein kinase or a transcription factor (Fig. 2C), further validating the previously suggested idea that protein kinase and transcription factor fusions constitute to a major fraction of the oncofusions.

We next moved on to characterize the structure of produced fusion proteins, to understand the protein-level consequences of the mutations, and thus the possible impact on protein activity in a cellular context. Particularly abundant protein groups among all protein producing fusions were PKs and TFs, which has been noted in previous studies as well[30]. We therefore decided to focus on their fusions in particular. Intact, full-length domains were abundant in the predicted fusion proteins. Especially protein kinase domain was very prominent, featured in 159 unique protein producing fusions (Fig. 3, supplementary table S6). As the kinase domain is usually in the C terminal of the protein, fusion mutations can easily cause kinase domains to be mislocalized due to localization signals from the fusion partner protein, or deleted membrane-spanning regions of the original kinase, for example. In addition to the protein tyrosine kinase domain, 43 other PK-specific domains were identified, bringing the total number of PK-specific domains to 455. In comparison, TF-specific domains consisted of a wider variety of individual domains, with 622 copies of 131 different domains.

Although fusion proteins are likely to lose domains that facilitate the validated interactions of the wild-type proteins, this is not necessarily the case. Receptor tyrosine kinases for example are commonly at the 3′ end of the new fusion gene, and the breakpoint often occurs just on the 3′ side of the region coding for the transmembrane part of the receptor. This could cause the kinase domain and intracellular protein–protein interaction domains to end up in a localization dictated by the 5′ gene. This, in turn, would lead to activation in an inappropriate place and/or at an inappropriate time. Similarly affected may be proteins shuttling between nucleus and cytoplasm as a response to an outside signal: they may end up perpetually trapped in the cytoplasm or the nucleus, or shuttled between the two in atypical conditions. To understand what kind of impact protein domains in novel environments might have, we next looked at already known interactors of all fusions (Figs. 3C,4). By grouping together interactors of both wild type proteins of each fusion, we were able to estimate the maximal set of currently foreseeable interactors of the fusion protein. I.e. phosphorylation targets, complex components etc. We found genes mentioned in the COSMIC cancer gene census to be enriched in both PK and TF fusions when comparing to other protein producing fusions, and occurring at a much higher rate in the interactor set of fusions between PK and TF genes. Same trend is reflected in CGC genes being more common the higher the number of potentially interacting fusions is. Possible roles of the interactors were then investigated via the Reactome pathway database. Interactors of PK and TF fusions were heavily concentrated around signal transduction pathways (Fig. 4).

To gain insight into whether the deductions so far were valid, we next used TCGA transcription quantification data to dig into the effects of specific gene-pair fusions with intact kinase or transcription factor specific domains. We discovered observable changes in gene expression, when comparing fusion groups against other protein producing fusions (Fig. 5), and effects that were seen to produce a noticeable impact on a pathway level as well, pointing to lower activity of various regulatory functions with many of the domain-containing gene pair oncofusions. Despite very heterogeneous gene-level differential expression patterns (Fig. 5 A), many genes are either over- or underexpressed in hundreds of gene pairs, with the most common ones being seen in the list of up- or downregulated genes in over half of all the gene pair groups studied. Considering distinct, pathway-level implications of these expression level changes, the fusions fall into two distinct groups based on which Reactome pathways the proteins they over-/underexpress function in (Fig. 5C). What is clear is that even though fusions may produce results that look alike, they each bring their own variation, and perhaps the specific cellular pathway-level effects of the expression changes are as distinct as fusion pairs themselves. Indeed, although the top pathways identified in the fusions formed roughly four groups, with some individual pathways outside of them, the full pathway annotation (supplementary table S11) includes a very individualistic figure of each fusion pair. The four pathway groups, however, may point to interesting findings about the fusion pairs themselves: Most of

Figure 5. Results of differential gene expression analysis. (**A**) Overall differential gene expression analysis results of the 517 fusion gene pair groups. Over 48,000 differentially expressed genes were found with q-value filter of 0.05. (**B**) Most common overexpressed (left) and underexpressed (right) genes in the differential expression data. (**C**) Reactome pathways detected and enriched in the differential expression data. Only pathways with z-score above 7.5 or below -7.5 are shown. Four groups with similar fusion pair patterns are highlighted in different colors.

them overexpress proteins of either the immune system related pathway group, or the other three groups, while underexpressing those of the other.

The conservation of domains may suggest conserved active functions, such as those of the kinase domain, potentially linked to inappropriate dimerization domains, target recognition domains, or domains that alter the entire molecular context of the novel fusion protein by targeting it to the wrong cellular compartment, membrane, or membrane raft. Through analyzing interactors of wild type proteins, we have identified multiple common interactors. If these interactions rely on the intact domains of the fusion protein, we can assume they represent possible pan-fusion drug targets, with which a multi-cancer effect may be achieved.

Taken together, we have now created and characterized the largest dataset of kinase and transcription factor oncofusions. This database will work as the foundation for molecular cloning and characterization of the PK- and TF-oncofusions using biochemistry, proteomics and cell biology—and a baseline hypothesis for the expected results.

Data availability

The datasets generated during and analyzed in this study are available either in the supplementary information or from the corresponding author on reasonable request.

Received: 21 April 2020; Accepted: 30 July 2020
Published online: 25 August 2020

References

1. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674. https://doi.org/10.1016/j.cell.2011.02.013 (2011).
2. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158. https://doi.org/10.1038/nature05610 (2007).
3. Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucl. Acids Res.* **45**, D777–D783. https://doi.org/10.1093/nar/gkw1121 (2017).
4. Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucl. Acids Res.* **45**, D777–D783. https://doi.org/10.1093/nar/gkw1121%JNucleicAcidsResearch (2016).
5. Kong, F. *et al.* dbCRID: a database of chromosomal rearrangements in human diseases. *Nucl. Acids Res.* **39**, D895–D900. https://doi.org/10.1093/nar/gkq1038 (2011).
6. Futreal, P. A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183. https://doi.org/10.1038/nrc1299 (2004).
7. Mitelman, F., Johansson, B. & Mertens, F. Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. *Nat. Genet.* **36**, 331–334. https://doi.org/10.1038/ng1335 (2004).
8. Mitelman, F., Johansson, B. & Mertens, F. The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer* **7**, 233–245. https://doi.org/10.1038/nrc2091 (2007).
9. Nowell, P. C. & Hungerford, D. A. Chromosome studies on normal and leukemic human leukocytes. *J. Natl. Cancer Inst.* **25**, 85–109 (1960).
10. Vega, F. & Medeiros, L. J. Chromosomal translocations involved in non-Hodgkin lymphomas. *Arch. Pathol. Lab. Med.* **127**, 1148–1160. https://doi.org/10.1043/1543-2165(2003)127%3c1148:CTIINL%3e2.0.CO;2 (2003).
11. Tomlins, S. A. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644–648. https://doi.org/10.1126/science.1117679 (2005).
12. Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214–220. https://doi.org/10.1038/nature09744 (2011).
13. Cancer Genome Atlas Research, N. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525. https://doi.org/10.1038/nature11404 (2012).
14. Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107–1120. https://doi.org/10.1016/j.cell.2012.08.029 (2012).
15. Stephens, P. J. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005–1010. https://doi.org/10.1038/nature08645 (2009).
16. Abate, F. *et al.* Pegasus: a comprehensive annotation and prediction tool for detection of driver gene fusions in cancer. *BMC Syst. Biol.* **8**, 97. https://doi.org/10.1186/s12918-014-0097-z (2014).
17. Suzuki, S. *et al.* The role of the amino-terminal domain in the interaction of unliganded peroxisome proliferator-activated receptor gamma-2 with nuclear receptor co-repressor. *J. Mol. Endocrinol.* **45**, 133–145. https://doi.org/10.1677/JME-10-0007 (2010).
18. Cancer Genome Atlas, N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337. https://doi.org/10.1038/nature11252 (2012).
19. Cancer Genome Atlas Research, N. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315–322. https://doi.org/10.1038/nature12965 (2014).
20. Zhang, J. *et al.* The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature* **481**, 157–163. https://doi.org/10.1038/nature10725 (2012).
21. Liu, S. *et al.* Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data. *Nucl. Acids Res.* **44**, e47. https://doi.org/10.1093/nar/gkv1234 (2016).
22. Dehghannasiri, R. *et al.* Improved detection of gene fusions by applying statistical methods reveals oncogenic RNA cancer drivers. *J. Proc. Nat. Acad. Sci.* **116**, 15524–15533. https://doi.org/10.1073/pnas.1900391116% (2019).
23. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298**, 1912–1934. https://doi.org/10.1126/science.1075762 (2002).
24. Lambert, S. A. *et al.* The human transcription factors. *Cell* **175**, 598–599. https://doi.org/10.1016/j.cell.2018.09.045 (2018).
25. Orchard, S. *et al.* Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods* **9**, 345. https://doi.org/10.1038/nmeth.1931 (2012).
26. Dijkstra, E. W. A note on two problems in connexion with graphs. *Numer. Math.* **1**, 269–271. https://doi.org/10.1007/BF01386390 (1959).
27. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504. https://doi.org/10.1101/gr.1239303 (2003).
28. 28Love, M. I., Huber, W. & Anders, S. J. G. B. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. **15**, 550. https://doi.org/10.1186/s13059-014-055008 (2014).
29. Walter, W., Sanchez-Cabo, F. & Ricote, M. GOplot: an R package for visually combining expression data with functional analysis. *Bioinformatics (Oxford, England)* **31**, 2912–2914. https://doi.org/10.1093/bioinformatics/btv300 (2015).

30 Gao, Q. *et al.* Driver fusions and their implications in the development and treatment of human cancers. *Cell Rep.* **23**, 227–238. https://doi.org/10.1016/j.celrep.2018.03.050 (2018).
31. Khemlina, G., Ikeda, S. & Kurzrock, R. Molecular landscape of prostate cancer: implications for current clinical trials. *Cancer Treat. Rev.* **41**, 761–766. https://doi.org/10.1016/j.ctrv.2015.07.001 (2015).

Author contributions

K.S., R.W. and M.V. conceived the study. R.W. did fusion set preparation and ran AGFusion. K.S. did domain, interactor, pathway & differential expression analyses. K.S., R.W., and M.V. wrote and prepared the manuscript. K.S. prepared the figures with input from R.W. and M.V.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-71040-8.

Correspondence and requests for materials should be addressed to M.V.

Reprints and permissions information is available at www.nature.com/reprints.

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