IMPORTANCE OF RARE COPY NUMBER ALTERATIONS FOR PERSONALIZED TUMOR CHARACTERIZATION

Michael Seifert¹, Betty Friedrich² & Andreas Beyer¹*  
Dresden University of Technology, Germany¹, ETH Zurich, Switzerland², University of Cologne, Germany³, *andreas.beyer@uni-koeln.de

Copy number alterations (CNAs) of large genomic regions are frequent in many tumor types, but only few of them are assumed to be relevant for the cancerous phenotype. It has proven exceedingly difficult to ascertain rare mutations that might have strong effects in individual patients. Here, we show that a genome-wide transcriptional regulatory network inferred from gene expression and gene copy number data of 708 human cancer cell lines can be used to quantify the impact of individual patient-specific gene CNAs on cancer-specific survival signatures. The model was highly predictive for gene expression in 4,548 clinical samples originating from 13 different tissues. Focused analysis of tumors from six tissues revealed that in an individual patient a combination of up to 100 gene CNAs directly or indirectly affected the expression of clinically relevant survival signature genes. Importantly, rare patient-specific mutations (< 1% in a given cohort) often had stronger effects on signature genes than frequent mutations. Subsequent integration with genomic data suggests that frequency variation among high-impact genes is mainly driven by gene location rather than gene function. Our framework contributes to the individualized quantification of cancer risk, along with determining individual key risk factors and their downstream targets.

INTRODUCTION

Although only a relatively small fraction of all mutations in any given cancer cell contributes to tumorigenesis, it is emerging that many more genes than previously thought determine clinically relevant endpoints such as proliferation rates, metastatic potential, or drug resistance¹,². Clearly, hundreds of genes have the potential to contribute to tumor phenotypes¹, but we are still far from being able to quantify individual cancer risks. The frequency at which genes are mutated in a certain cancer cohort is an indicator of clinical importance. Even though frequent mutations (i.e. mutations that are more frequent than expected by chance in a specific cohort) are more likely to have tumor-related effects, individual cancer risks are not fully explained by frequent mutations alone. Rare mutations could act in combination with frequent mutations or they may, entirely independent from frequent mutations, establish a significant risk for the patient on their own. We do not know how important rare mutations are in comparison to frequently observed mutations, simply because we are lacking the means to quantify their effects. The specific pattern of small mutations (SNPs, small indels) in candidate genes can be used to prioritize putative driver genes without using epidemiological information²,⁴. Here, we present an approach exploiting the additional information contained in gene expression data in order to quantify potential effects of rare copy number alterations (CNAs). Our framework rests on the notion that regulatory relationships between genes are fairly robust across tumors, whereas the specific functional pattern of a given tumor is virtually private⁵,⁶. Put differently: most CNAs increase or decrease the activity of genes, while potentially only a small fraction of them alter the regulatory relationships between genes. Hence, using large compendia of expression- and mutation datasets we can establish regulatory relationships between genes in cancer cells and quantify the effects of CNAs on gene expression. Such a model can subsequently be used to analyze individual tumors with known mutational patterns to quantify the impact of specific CNAs on global expression. Further, by relating those expression changes to clinical endpoints we are able to quantify the effects of single CNAs on the survival of an individual patient.

Using this framework we can quantify direct (cis-) effects and indirect (trans-) effects of CNAs, we can identify key regulators in CNA regions ('driver genes') with particularly strong impact on the expression of clinically relevant genes, we can compare the importance of rarely mutated genes with frequently mutated genes, and we can quantify the combined effects of all CNAs on survival risk for an individual patient. Our analysis shows that usually many mutations together influence individual patient survival by further impacting on common molecular pathways. At the individual level, it turns out that rare copy number mutations (< 1% frequency in a given cancer cohort) can be as important as frequent mutations and we are able to specifically pinpoint the most risky rare and frequent CNAs in individual patients.

METHODS

In order to predict potential effects of copy number variations in the specific environment of tumor cells, we computationally inferred a genome-wide transcriptional regulatory network from human cancer cell lines of 24 different tumor sites⁶. We termed this model the Cancer Cell Transcriptional Network (CCTN). To identify putative regulator genes, we modeled the expression level of each gene (target gene) as a linear combination of the gene-specific copy number and the expression levels of all other potential regulator genes. Sparse regression based on lasso (least absolute shrinkage and selection operator) was used to prioritize the inclusion of direct effects into the model. CCTN is characterized by few central hub genes that have a large number of incoming and outgoing edges. Well known cancer genes (e.g. TNFRSF17, FUS, IKZF1, GATA1, PAX8, SF3Q, IRF4, KLF2, COL1A1, MSL2, HSP90AB1, PHOX2B, CD79B, LYL1) were significantly overrepresented among the 219 hub genes with more than 20 trans-acting regulatory edges to or from other genes (Fisher's exact test: p-value < 0.006). Regulator
genes with a large number of outgoing edges (i.e. major regulators) were enriched for known transcription factors and signaling pathway genes suggests that lasso successfully enriched for direct effects. We further validated CCTN on tumor data from 13 different cancer cohorts from the TCGA consortium (4,548 tumor patients) and by using in vitro single-gene perturbation data (50,306 knockdown or over expression experiments). Next, we devised a method to compute the impact of specific gene perturbations on the expression levels of all other genes in the network. We validated the impact prediction using independent patient cohorts not used for the training and based in in vitro experiments. Finally, we used a Random Forest-based approach to identify signature genes indicative of patient survival and predicted the impact of each gene’s mutation in individual patients on the survival of that patient. Again, these predictions were validated using data from independent patient cohorts.

RESULTS & DISCUSSION

We applied this framework to six TCGA cohorts of sufficient size (AML: acute myeloid leukemia, GBM: glioblastoma multiforme, HNSC: head and neck squamous cell carcinoma, LUAD: lung adenocarcinoma, OV: ovarian serous cystadenocarcinoma, SKCM: skin cutaneous melanoma) and quantified the number of gene CNAs contributing to survival in individual patients. Up to 100 gene CNAs were individually contributing to patient survival. Thus, although less than 10 genes might be required for the initial neoplastic transformation, many more genes seem to contribute to patient survival. Next, we analyzed the relationship between the frequency of gene CNAs in a cancer cohort and their impact on survival. As expected, more frequent mutations had on average higher impacts than low-frequency mutations and high-frequency mutations were more enriched for known cancer genes. However, although low-frequency mutations (< 1%) had on average lower impacts, occasionally their impacts were as strong or even stronger than those of frequent mutations (see Figure 1 for GBM as an example; similar observations were made for the other 5 tumor types). In order to understand why genes with similar impacts are mutated at largely different frequencies we investigated their functions and genomic positions. Instead of function, genomic positioning of genes better explains the frequencies of CNAs. For example, we observed that frequently mutated CNA genes tend to be closer to fragile sites and closer to frequently germ-line mutated regions than low-frequency genes. Likewise, tumor suppressor genes were less likely to be deleted if they were close to proto-oncogenes or essential genes. In addition, we noticed striking differences between tissues or tumor types. For example, the correlation between CNA frequencies and genomic features was highly dependent on the tumor type. Further, we found many genes that are well established cancer genes in one tissue to be also mutated (with large predicted impact) in other tumors. However, the CNA frequency in those 'new' tissues was mostly low, explaining why many of these genes have not been detected as being relevant in those tumors before. These observations imply that tissue-specific factors such as chromatin state, cell-cycle rates, exposure to DNA damaging agents, number of stem cell divisions, or even the expression of specific genes could considerably impact on mutational mechanisms.

CONCLUSIONS

Although expression variation of individual regulators changes the activity of molecular sub-networks, the topology of regulatory relationships as such turns out to be remarkably robust across cell types. Because of that we were able to quantify the importance of gene CNAs for individual tumor risks leading to the observation that rare variants can be as important as frequent variants. Importantly, the frequency at which a high-impact gene gets mutated seems to be determined by factors that are independent of its function or impact. Thus, the fact that some high-impact genes have higher CNA frequencies may simply be due to their placement in genomic regions that are more amenable for CNAs than others.

Our work contributes to the quantification of individual cancer risks established by patterns of frequent and infrequent copy number alterations. The availability of a regulatory model facilitates the detection of genes that are commonly affected by different rare CNAs, which opens a window of opportunity for developing therapeutic strategies against such rare mutations.

FIGURE 1. Impact of copy number altered (CNA) genes in glioblastoma multiform (GBM) tumors on patient survival.

Impact of copy number alterations on the expression of survival signature genes versus their frequency in the TCGA GBM population. Mutations with frequencies at 0.2% (leftmost) occur in only one patient. Some of these mutations have impacts comparable to high-impact frequent mutations. Vertical dashed line: 1% threshold used to separate low and high frequency gene CNAs.

REFERENCES