Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Network and Pathway Analysis of Cancer Susceptibility (A)

Predicting Cancer Prognosis Using Functional Genomics Data Sets

Jishnu Das[1,2], Kaitlyn M. Gayvert[3] and Haiyuan Yu[1,2]

[1]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY, USA. [2]Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY, USA. [3]Tri-Institutional Training Program in Computational Biology and Medicine, New York, NY, USA.

ABSTRACT: Elucidating the molecular basis of human cancers is an extremely complex and challenging task. A wide variety of computational tools and experimental techniques have been used to address different aspects of this characterization. One major hurdle faced by both clinicians and researchers has been to pinpoint the mechanistic basis underlying a wide range of prognostic outcomes for the same type of cancer. Here, we provide an overview of various computational methods that have leveraged different functional genomics data sets to identify molecular signatures that can be used to predict prognostic outcome for various human cancers. Furthermore, we outline challenges that remain and future directions that may be explored to address them.

KEYWORDS: cancer prognosis prediction, functional genomics, gene expression, cellular networks, somatic mutations

SUPPLEMENT: Network and Pathway Analysis of Cancer Susceptibility (A)

CITATION: Das et al. Predicting Cancer Prognosis Using Functional Genomics Data Sets. *Cancer Informatics* 2014:13(S5) 85–88 doi: 10.4137/CIN.S14064.

RECEIVED: June 29, 2014. **RESUBMITTED:** September 17, 2014. **ACCEPTED FOR PUBLICATION:** September 19, 2014.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Review

FUNDING: This work was funded by National Institute of General Medical Sciences grant GM104424 and National Cancer Institute grant CA167824. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: haiyuan.yu@cornell.edu; jd327@cornell.edu

Introduction

The genetic complexity across the landscape of human cancers makes it extremely difficult to find concordant molecular signatures – there is a tremendous amount of genetic heterogeneity across various cancer types and subtypes.[1] It is well accepted that the development of cancer entails six biological capabilities acquired during tumorigenesis – "sustaining proliferative signaling," "evading growth suppressors," "activating invasion and metastasis," "enabling replicative immortality," "inducing angiogenesis," and "resisting cell death."[2,3] A complex combination of these events causes major changes in phenotype – tumor cells have drastically rewired cellular machinery such as altered messenger RNA (mRNA) and microRNA expression, copy number variations, and epigenetic modifications. The molecular elucidation of all these changes is a highly challenging task that has necessitated global international collaborations combining different areas of expertise.[4,5] It is widely believed that comprehensive molecular characterization of key causal driver events that are concordant across cancer types holds the key in developing successful therapeutic regimens for cancer.[3]

Thus, one primary focus of many efforts has been to develop bioinformatic tools to analyze the different kinds of data that are being made available by these projects and identify global signatures. A key challenge addressed by many of these tools is to utilize different data sets and predict prognostic outcome for one or more types of cancer. This is an extremely difficult problem because it is not clear as to which measurements are most informative of disease outcome. This had led researchers to use several functional genomics data sets to predict prognosis – expression, protein networks, somatic mutation profiles, and epigenetic modifications. Conceptually, these data sets can be grouped into three broad categories – changes in expression, nucleotide modifications, and nucleotide alterations. These three sets can be used individually or in the context of the underlying cellular networks to predict prognosis. In the course of this review, we outline different studies that have adopted these approaches to identify a molecular signature that correlates with disease outcome and use it to predict prognosis.

Prognosis Prediction Using Expression Profiles

Since cancer mutations significantly rewire transcriptional machinery, it is well established that gene expression profiles of tumors undergo major changes.[6] This enables the classification of tumors into subtypes determined by their expression signatures.[6] In two seminal studies, van 't Veer et al.[7] and van de Vijver et al.[8] used microarrays to obtain the transcriptomic profile of 295 breast cancer patients and clustered the expression matrix to identify a 70-gene signature that could be used to predict disease outcome. They concluded that gene expression profile was a better predictor of outcome than standard clinical and histological criteria. These studies were a major breakthrough at different levels – they showed the limitations of existing clinical parameters in predicting disease outcome, the utility of an expression data set in overcoming these limitations, and established the concept of a gene signature for predicting prognosis.

Wang et al.[9] measured the expression profile of a different cohort of 286 breast cancer patients. Using hierarchical clustering, they identified a 76-gene signature comprising 60 genes for estrogen receptors (ER)–positive and 16 genes for ER-negative patients. Although both groups comprised breast cancer patients, surprisingly there was very little overlap between the two gene signatures. This led Wang et al.[9] to hypothesize that "because of differences in patients, techniques, and materials used" the signatures were vastly divergent. This also suggested that even within the same cancer type, different subtypes could have completely different gene signatures that correlate with disease outcome. This has been valuable in helping clinicians and researchers appreciate that there may be major molecular differences across cancer subtypes.

Song et al.[10] used microRNA and mRNA expression profiles for prognosis prediction in gastric cancer. With 90 cancer tissue samples and 10 normal samples, the authors first used consensus clustering to identify candidate microRNAs and targets that are potential biomarkers – ie, correlate significantly with disease outcome. These were then further validated and evaluated using 385 samples. Specifically, the authors found that miR-200c, miR-200b, and miR-125b and the corresponding target expression correlated most with survival outcome and are thus potential prognostic biomarkers for gastric cancer.

Combining Expression Data with Interactome Networks for Prognosis Prediction

One key limitation to expression-based approaches is the treatment of genes as independent variables. The previously described methods used genes that are differentially expressed between patients with good and bad prognoses. However, in the cellular environment, these genes do not act in isolation. They are part of different complex cellular networks, one of which is the protein interactome network. In this network, nodes represent proteins and edges represent physical interactions between these proteins.[11] Complex phenotypes are best explained by dysregulation of gene sets[12] rather than isolated genes since alteration of expression levels of genes affects not only the protein products encoded by those genes but also those that are in the network neighborhood of those proteins. This has led to the idea of "guilt-by-association"[13] – a concept that has been often used to identify and prioritize disease genes.

In the context of cancer prognosis prediction, Chuang et al.[14] were the first to attempt network-based classification of breast cancer prognosis at a genomic scale. Compared to previous studies that had identified gene-based markers, the authors elucidated subnetwork markers. For each subnetwork, they calculated a patient-specific activity score by averaging the normalized expression values for genes in that subnetwork. The "discriminative potential" of a subnetwork was defined as the mutual information between its activity score and disease status (metastatic or not) across all patients. They found that significant subnetworks thus identified were more reproducible than individual gene markers and were enriched for biological processes known to be important in cancer progression. These could also be used to predict prognostic outcome.

Another key property of these protein interactome networks is their modularity.[15] Specifically, there are two kinds of hubs in these networks – intramodular and intermodular hubs. It has been shown that the biological properties of these hubs are very different.[15–17] Taylor et al.[18] showed that differential expression of these hub groups in the protein network can be used to predict breast cancer prognosis. Specifically, the co-expression of intermodular hubs with their interactors is tissue specific but the co-expression of intramodular hubs with their interactors is mostly generic and tissue independent. The authors used this dynamic modularity to identify hubs whose relative expression with their interactors was significantly correlated with prognosis. These hubs were then subjected to affinity propagation clustering[19] to identify top exemplars that could be used to predict prognosis. The results of this study suggest that molecular changes in a tumor are reflected by alterations in disease network modularity. Since these changes are significantly correlated with prognosis, measuring them would improve the predictive power of previously used clinical prognostic variables.

Edwin Wang and colleagues developed an algorithm – Multiple Survival Screening (MSS) – to identify breast cancer metastasis drivers using a combination of expression and interaction data.[20] Their key hypothesis was that the variability of gene expression in tumor cells is much higher than that in normal cells. Thus, only a small fraction of the altered expression profiles can be explained by driver metastatic events. Thus, a "one-step clustering" of expression profiles is likely to identify mostly passenger events. They used this to explain the low overlap between the gene signatures identified for the two cohorts of breast cancer patients described earlier.[8,9] To circumvent this, MSS used three discrete gene signatures to identify different risk categories – low, intermediate, and high. Combining these signatures with the protein interactome also yielded driver modules recurrent in breast cancer tumors.

Wu et al.[21] used a previously constructed functional interaction network to identify prognostic biomarkers for breast and ovarian cancer.[22] The functional interaction network incorporates gene ontology annotations and domain–domain interactions in addition to protein–protein interactions and gene expression.[21] To predict prognosis, the authors used Markov clustering to identify gene expression modules from this network. They then identified linear combinations of modules that are maximally correlated with patient survival.

Chowdhury et al.[23] and Patel et al.[24] developed Combinatorially dysRegulAted subNEtworks – a neural network–based approach for prognosis prediction. Their approach identified subnetworks that are dysregulated in tumor and metastatic samples. They found that these subnetworks can be used as accurated predictors of prognostic outcome for colorectal cancer[23] and glioblastoma multiformae.[24]

Other Functional Genomics Data Sets for Prognosis Prediction

With the rapid growth of sequencing technology, whole-genome and whole-exome sequencing studies have become an indispensable tool to identify drivers of cancer.[1,25] Typically, these studies identify thousands of mutations that are enriched in the individuals afflicted with disease when compared to normal controls.[26] However, most of these are passenger mutations and not causative.[26] One commonly accepted way to search for cancer drivers is to look for hypermutated genes and combine it with orthogonal lines of functional evidence wherever available.[27] While a large amount of somatic mutation data have recently become available,[28] it is still unclear in many cases as to how or why these driver events lead to vastly different outcomes. Although the issue of understanding molecular outcomes is extremely complex, the large number of recently sequenced cancer samples has helped build comprehensive catalogs of cancer genes.[25] It has also been estimated that 600–5,000 samples per tumor type may lead to "saturation" in terms of analyzing gene-specific mutational spectrums.[25]

Hofree et al.[29] devised a Network-Based Stratification (NBS) approach on somatic mutations from sequencing data to predict subtypes of different cancers. NBS uses somatic mutation profiles in conjunction with an interactome network to divide patients into subtypes where patients with mutations in similar network neighborhoods are assigned to the same group. Conceptually, there are three steps in NBS. First, somatic mutations for each patient are projected onto a network and their influence is propagated using an algorithm previously established by Vanunu et al.[30] The resulting profiles are then subsampled and clustered into a fixed number of subtypes using nonnegative matrix factorization.[31] The subsamples are then combined into a single aggregate matrix using consensus clustering.[32] The authors showed that subtypes identified by NBS are clinically relevant as they correlate well with prognostic outcomes for ovarian, uterine, and lung adenocarcinoma.

Wen et al. found that aberrant DNA methylation of transcription factors can be used to identify genes causing colorectal cancer.[33] They combined methylation data with known cancer genes to identify candidate causal genes. This information was used in conjunction with a protein network weighted by corresponding mRNA co-expression values to define an activity matrix for network modules. Using this matrix, the authors identified seven important modules that could serve as biomarkers for colorectal cancer. Although the authors did not use their method to predict survival, since the identified modules try to infer underlying causality, the expression profiles of these key network modules may provide insights into severity of prognostic outcome. Another key difference of this study from previous studies was the incorporation of epigenomic information in addition to expression and network data.

Kai Tan and colleagues[34] also used DNA methylation data in conjunction with gene expression and the protein network as inputs to a support vector machine classifier to predict prognosis for glioblastoma multiformae patients with reasonably high accuracy. They found 10 expression-based and 3 methylation-based network markers that have a significant effect on outcome. Based on these network markers, they hypothesized that two key pathways have a significant effect on prognosis – GTPase-mediated trafficking and ubiquitination-dependent degradation.

Zhang and Ouellette developed CAERUS - a modified Naive Bayes classifier that combines domain architecture with gene expression, protein networks and somatic mutations to predict breast and ovarian cancer prognosis[35]. They were the first to incorporate protein structural information in prognosis prediction.

Discussion

The methods discussed above represent excellent efforts to predict cancer prognosis across cancer types and subtypes. They have been validated on large sample sizes and in general are quite successful both in terms of prediction and identification of certain genes or groups of genes that help determine prognostic outcome. However, there are several issues that limit the use of these computational methods in conjunction with existing clinical practices. Typically, these have been validated on single-patient cohorts. Thus, it is unclear how successful the same method would be for a different cohort. For example, as discussed earlier, the gene signature identified by the van de Vijver et al. study has minimal overlap with the Wang et al. study. Ideally, it is important to validate a method across multiple sets of patients. Moreover, studies that combine multiple functional genomics data sets sometimes do not focus on how well each individual data set performs. This information can often be critical as all data sets may not be available for different cohorts.

In general, it seems clear that integration of multiple layers of information helps prediction accuracy. Efforts such as the TCGA Pan-Cancer Analysis Project are already trying to combine mutation, copy number, gene expression, methylation, microRNA, proteomic arrays, and clinical data to distinguish driver from passenger mutations.[4] The complementarity of these different data sets has also been recognized – eg, direct study of protein levels often provides insights that genomic or transcriptomic data sets cannot.[36]

However, several questions remain unanswered. First, our understanding of which data sets are most informative for a particular cancer type/subtype is still highly limited. Being able to identify certain data sets a priori certain is likely to significantly boost prognosis prediction performance of existing algorithms. Second, there has been a significant amount of effort in the community to translate genetic insights into druggable targets – numerous repositories of drug-gene targets such as DrugBank,[37] PharmGKB,[38] and Therapeutic Target DB[39] catalog these efforts. However, still only a small fraction of all the US Food and Drug Administration–approved drugs have gene targets based on well-elucidated mechanisms (ie, are etiology specific) and the majority of drugs today are still palliative.[40] Thus, it remains to be seen whether the mechanistic insights generated by these prognosis prediction methods can be used to guide rational drug design. Finally, it is still unclear whether these prognosis prediction schemes can be directly applied in a clinical setting to create personalized treatment regimens. While there have been a few cases where insights have directly guided clinical treatment schemes,[41] these are still the exception rather than the norm. In general, there is still a wide chasm between researchers who develop these in silico approaches and clinicians who actually decide how to treat cancer patients. It is important for the in silico approaches to focus on generating biological insights that are translatable to concrete mechanistic hypotheses. Finally, crosstalk between these two communities and development of customized computational pipelines that clinicians are willing to trust and adopt in specific settings can go a long way in bridging this gap.

Author Contributions

Contributed to the writing of the manuscript: JD. Jointly developed the structure and arguments for the paper: JD, KMG, HY. All authors reviewed and approved of the final manuscript.

REFERENCES

1. Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*. 2013;499:214–8.
2. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell*. 2000;100:57–70.
3. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011;144:646–74.
4. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*. 2013;45:1113–20.
5. International Cancer Genome Consortium, Hudson TJ, Anderson W, et al. International network of cancer genome projects. *Nature*. 2010;464:993–8.
6. Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature*. 2000;406:747–52.
7. van 't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002;415:530–6.
8. van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*. 2002;347:1999–2009.
9. Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*. 2005;365:671–9.
10. Song F, Yang D, Liu B, et al. Integrated microRNA network analyses identify a poor-prognosis subtype of gastric cancer characterized by the miR-200 family. *Clin Cancer Res*. 2014;20:878–89.
11. Vidal M. Interactome modeling. *FEBS Lett*. 2005;579:1834–8.
12. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102:15545–50.
13. Oliver S. Guilt-by-association goes global. *Nature*. 2000;403:601–3.
14. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. *Mol Syst Biol*. 2007;3:140.
15. Han JD, Bertin N, Hao T, et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*. 2004;430:88–93.
16. Das J, Mohammed J, Yu H. Genome-scale analysis of interaction dynamics reveals organization of biological networks. *Bioinformatics*. 2012;28:1873–8.
17. Das J, Vo TV, Wei X, et al. Cross-species protein interactome mapping reveals species-specific wiring of stress response pathways. *Sci Signal*. 2013;6:ra38.
18. Taylor IW, Linding R, Warde-Farley D, et al. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotechnol*. 2009;27:199–204.
19. Frey BJ, Dueck D. Clustering by passing messages between data points. *Science*. 2007;315:972–6.
20. Li J, Lenferink AE, Deng Y, et al. Identification of high-quality cancer prognostic markers and metastasis network modules. *Nat Commun*. 2010;1:34.
21. Wu G, Feng X, Stein L. A human functional protein interaction network and its application to cancer data analysis. *Genome Biol*. 2010;11:R53.
22. Wu G, Stein L. A network module-based method for identifying cancer prognostic signatures. *Genome Biol*. 2012;13:R112.
23. Chowdhury SA, Nibbe RK, Chance MR, Koyuturk M. Subnetwork state functions define dysregulated subnetworks in cancer. *J Comput Biol*. 2011;18:263–81.
24. Patel VN, Gokulrangan G, Chowdhury SA, et al. Network signatures of survival in glioblastoma multiforme. *PLoS Comput Biol*. 2013;9:e1003237.
25. Lawrence MS, Stojanov P, Mermel CH, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*. 2014;505:495–501.
26. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature*. 2013;500:415–21.
27. Futreal PA, Coin L, Marshall M, et al. A census of human cancer genes. *Nat Rev Cancer*. 2004;4:177–83.
28. Forbes SA, Bindal N, Bamford S, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res*. 2011;39:D945–50.
29. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. *Nat Methods*. 2013;10:1108–15.
30. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol*. 2010;6:e1000641.
31. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature*. 1999;401:788–91.
32. Monti S. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach Learn*. 2003;52:91–118.
33. Wen Z, Liu ZP, Liu Z, Zhang Y, Chen L. An integrated approach to identify causal network modules of complex diseases with application to colorectal cancer. *J Am Med Inform Assoc*. 2013;20:659–67.
34. Kim J, Gao L, Tan K. Multi-analyte network markers for tumor prognosis. *PLoS One*. 2012;7:e52973.
35. Zhang KX, Ouellette BFF. CAERUS: Predicting cancer outcomes using relationship between protein structural information, protein networks, gene expression data, and mutation data. *PLoS Comput Biol*. 2011;7:1001114.
36. Akbani R, Ng PK, Werner HM, et al. A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nat Commun*. 2014;5:3887.
37. Law V, Knox C, Djoumbou Y, et al. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res*. 2014;42:D1091–7.
38. Whirl-Carrillo M, McDonagh EM, Hebert JM, et al. Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther*. 2012;92:414–7.
39. Qin C, Zhang C, Zhu F, et al. Therapeutic target database update 2014: a resource for targeted therapeutics. *Nucleic Acids Res*. 2014;42:D1118–23.
40. Yildirim MA, Goh KI, Cusick ME, Barabasi AL, Vidal M. Drug-target network. *Nat Biotechnol*. 2007;25:1119–26.
41. Worthey EA, Mayer AN, Syverson GD, et al. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet Med*. 2011;13:255–62.