Bioinformatics, 31(16), 2015, 2607–2613
doi: 10.1093/bioinformatics/btv164
Advance Access Publication Date: 24 March 2015
Original Paper

OXFORD

Sequence analysis

Integration of somatic mutation, expression and functional data reveals potential driver genes predictive of breast cancer survival

Chen Suo[1,2], Olga Hrydziuszko[2], Donghwan Lee[3], Setia Pramana[2,4], Dhany Saputra[2], Himanshu Joshi[2], Stefano Calza[2,5] and Yudi Pawitan[2,*]

[1]School of Life Sciences, Peking University, Beijing, China, [2]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden, [3]Department of Statistics, Ewha Womans University, Seoul, South Korea, [4]Department of Computational Statistics, Institute of Statistics, Jakarta, Indonesia and [5]Department of Molecular and Translational Medicine, University of Brescia, Brescia, Italy

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on November 18, 2014; revised on February 25, 2015; accepted on March 16, 2015

Abstract

Motivation: Genome and transcriptome analyses can be used to explore cancers comprehensively, and it is increasingly common to have multiple omics data measured from each individual. Furthermore, there are rich functional data such as predicted impact of mutations on protein coding and gene/protein networks. However, integration of the complex information across the different omics and functional data is still challenging. Clinical validation, particularly based on patient outcomes such as survival, is important for assessing the relevance of the integrated information and for comparing different procedures.

Results: An analysis pipeline is built for integrating genomic and transcriptomic alterations from whole-exome and RNA sequence data and functional data from protein function prediction and gene interaction networks. The method accumulates evidence for the functional implications of mutated potential driver genes found within and across patients. A driver-gene score (DGscore) is developed to capture the cumulative effect of such genes. To contribute to the score, a gene has to be frequently mutated, with high or moderate mutational impact at protein level, exhibiting an extreme expression and functionally linked to many differentially expressed neighbors in the functional gene network. The pipeline is applied to 60 matched tumor and normal samples of the same patient from The Cancer Genome Atlas breast-cancer project. In clinical validation, patients with high DGscores have worse survival than those with low scores ($P = 0.001$). Furthermore, the DGscore outperforms the established expression-based signatures MammaPrint and PAM50 in predicting patient survival. In conclusion, integration of mutation, expression and functional data allows identification of clinically relevant potential driver genes in cancer.

Availability and implementation: The documented pipeline including annotated sample scripts can be found in http://fafner.meb.ki.se/biostatwiki/driver-genes/.

Contact: yudi.pawitan@ki.se

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Analysis of genome and transcriptome sequencing experiments provides a comprehensive tool for molecular studies of cancers. The Cancer Genome Atlas (TCGA) breast cancer project (Cancer Genome Atlas Network, 2012) presents a rich dataset of whole-exome and RNA sequence data of matched tumor and normal samples of the same patient; these allow us to accurately infer tumor-specific alterations including somatic mutations and isoform-level differential expression. However, it is still a challenge how to subsequently integrate the complex information across the different omics data, while also exploiting the rich functional data such as protein prediction and gene/protein interaction networks.

Most carcinomas are driven to develop by a few genetic alterations, whereas the majority of the remaining genetic changes have neutral or less deleterious effect in cancer development (Futreal *et al*., 2004). To date, potential driver genes are identified largely based on finding recurrent copy-number alteration or mutations in a specific region across multiple samples (Akavia *et al*., 2010; Ciriello *et al*., 2012; Lazar *et al*., 2013) but not on patient-specific genomic alterations. In this article, we develop a full analysis pipeline to combine DNA and RNA sequencing data together with functional data, starting from the preprocessing of raw aligned sequencing reads, to search for potential driver genes and eventually to summarize the effects of these genes into a single driver-gene score (DGscore). In addition to using existing bioinformatics tools, including GATK (McKenna *et al*., 2010), SnpEff (Cingolani *et al*., 2012) and Sequgio (Suo *et al*., 2014), we suggest a novel method based on the network enrichment analysis (NEA; Alexeyenko *et al*., 2012) to integrate the genomic and transcriptomic profiles.

There is a growing literature on searching for driver genes based on data integration. We can state at least two novel aspects in this study: (i) isoform-level analysis: because exon lengths are not multiples of three, the potential protein-coding impact of one mutation is different for different isoforms. Hence isoform-level assessment is necessary. (ii) Clinical validation: we assess the clinical relevance of the potential drivers in terms of correlation with patient outcomes such as survival. Most current integrative methods in identifying cancer driver genes are based on whole genes and are primarily validated based on previously reported drivers and pathway analysis, but not clinically. Akavia *et al.* (2010) identified some known drivers of melanoma, and two of the predicted drivers were demonstrated to be critical for tumor growth with knockdown experiments. Youn and Simon (2011) proposed a method that accounts for the functional impact of mutations on protein coding, a sample variation in background mutation rate and the redundancy of the genetic code. They analyzed a dataset of non-small cell lung tumors to show that the identified driver genes were also deemed important previously.

Ciriello *et al.* (2012) developed the so-called MEMo algorithm and showed that it was able to recapitulate previously identified pathways in glioblastoma and ovarian cancer. Potential driver genes were used for histological subtyping in lung cancer (Lazar *et al*., 2013) or to reveal intrinsic subtype-specific mutations in breast cancer (Cancer Genome Atlas Network, 2012). CAERUS focuses on investigation of biological network disruptions linked to cancer outcomes at the protein domain level, but it does not take into account the impact of mutations on RNA expression and protein coding (Zhang and Ouellette, 2011). DawnRank, a recently published algorithm, directly prioritizes altered genes on a single-patient scope regardless of mutation frequency (Hou and Ma, 2014). DawnRank requires information of a gene interaction network, somatic genomic alterations and the differential gene expression profile, but it does not consider the predicted biological impact of a mutation on protein. Helios incorporates somatic copy number alterations, point mutations, gene expression and RNAi screens to pinpoint driver genes with large recurrently amplified regions of DNA (Sanchez-Garcia *et al*., 2014). In summary, these methods do not utilize isoform-level information and the potential drivers are generally not validated in terms of patients' clinical outcomes such as survival.

As genetic instability caused by mutations in multiple critical genes may be central to tumor progression (Loeb and Loeb, 2000; Salomon *et al*., 2013), we summarize the effects of potential driver genes into a single value DGscore and assess its clinical value as prognostic biomarker. In our analysis of the TCGA breast cancer data, there is a strong evidence that patients carrying more mutated genes with functional implications and extreme expression pattern have worse survival rates than those with less mutated potential driver genes. Over the last decade, a number of studies have developed gene signatures for prognostic classification of breast cancer. Among the most well known, Mammaprint (van 't Veer *et al*., 2002) captures the expression profile of 70-gene associated with prognosis and has been shown to be a better predictor for distant metastasis than the clinical parameters. The so-called PAM50 signature (Parker *et al*., 2009) classifies breast cancer into intrinsic molecular subtypes and has been shown to add significant prognostic value. For the TCGA breast cancer data, as predictor of patient survival, the DGscore based on the potential driver genes outperforms the 70-gene MammaPrint and the 50-gene PAM50 signatures. In summary, integration of omics and functional data have the potential to produce clinically valuable information.

2 Patients and methods

2.1 TCGA60 paired samples

From the TCGA breast cancer project, we identified a total of 60 female patients diagnosed with invasive breast carcinoma, for which both DNA and RNA sequencing data were available, and both adjacent normal breast tissue and blood samples were also collected. The age of patients at initial pathological diagnosis ranged from 31 to 90 years, with a median of 58. A majority of the patients (77%) had infiltrating ductal carcinoma. Seven patients were stage I, 36 stage II, 14 stage III, 1 stage IV and 2 patients' stage information was not available. Thirty-seven patients had estrogen-receptor positive tumors. After a median follow-up of 25 months, with a range from 0 to 147 months, 18 patients had died. Other clinical variables and information on generating the sequencing data are summarized in the Supplementary Material (section Patients and tissue samples, Sequencing data processing and Supplementary Table S1). The TCGA samples IDs that used in the analysis are provided in Supplementary Table S2 in the Supplementary Material.

2.2 Integrative statistical analysis

An overview of the scheme to identify putative cancer drivers, including procedures on data processing, analysis and integration, is shown in Figure 1.

First, we use the Genome Analysis Toolkit (McKenna *et al*., 2010) and PICARD (http://picard.sourceforge.net/) for refining initial reads, followed by a χ^2 test to assess the statistical significance of association between allelic counts and tumor/normal (T/N) status for somatic variant calling. To increase statistical confidence, a variant must have P-value < 0.01 to be considered further, and up to 1000 variants are allowed for each patient. This is likely to be more

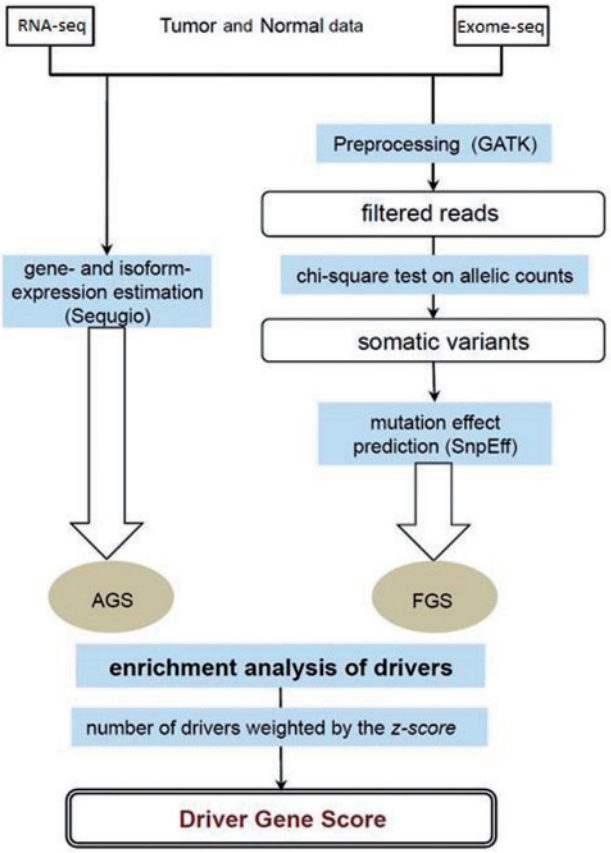

Fig. 1. Flowchart of the identification of potential driver genes and derivation of the DGscore; see text for detailed description

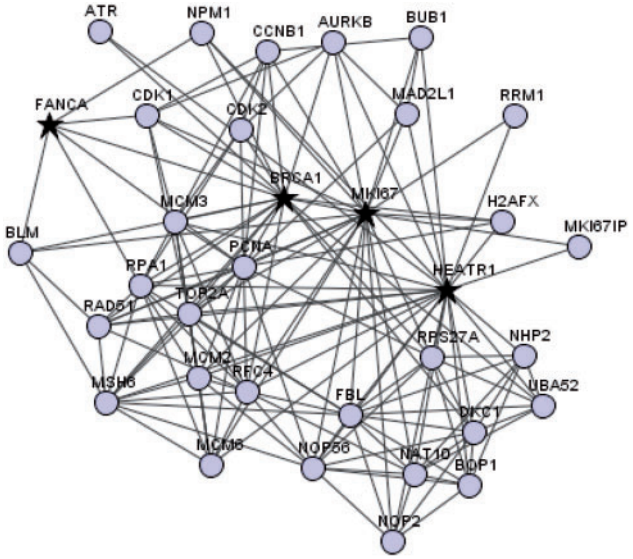

Fig. 2. Gene network around four potential driver genes *FANCA, BRCA1, MKI67* and *HEATR1* (black stars). This network contains 30 other genes (circles) connected to the potential driver genes. There is a total of 159 links, including protein–protein interactions, metabolic and signaling links from the functional coupling network (Alexeyenko and Sonnhammer, 2009)

mutations than expected in a tumor, but at this point, we need to preserve sensitivity, so specificity is deliberately compromised, since it will be improved in the subsequent filtering steps.

Non-synonymous mutation is considered more likely to have a direct role in phenotypic change, so the SnpEff is then used to predict the impact of these mutations on protein coding as 'high', 'moderate' or 'low', based on the functional consequences of the mutation on known genes (Cingolani *et al.*, 2012). A high impact mutation, for instance, leads to exon deletion, frame shift and stop lost. Moderate impact includes non-synonymous coding, codon insertion or deletion, etc. Synonymous start/stop, synonymous coding, etc, are categorized as low impact. Isoforms mutated in at least 10 patients and predicted to have high or moderate impact in more than two patients are kept on the list of potential drivers genes. (See the Supplementary Material, section Selection of frequently mutated isoforms for details.) Although extremely rare mutations may be neglected by considering recurrently mutated genes, setting a cutoff to filter out low frequency mutations may reduce false positive mutation calls and highlight primary genes in tumor development.

Gene- and isoform-level expression is estimated using Sequgio (Suo *et al.*, 2014). A median and variance scaling is then performed to normalize the expression on the scale of $\log_2$ T/N ratio.

While most of our analysis is at isoform level, we often still have to refer to genes. In general, if at least one isoform of a potential driver gene meets the necessary requirement, the gene is included in the next step of the pipeline.

The integration of mutation and expression data is done using a network enrichment analysis (NEA) as follows. The key idea is that the functional impact of each mutation is assessed in terms of the number of differentially expressed (DE) neighbors in a gene network. We use a comprehensive network (Alexeyenko *et al.*, 2012) containing approximately 1.4 millions functional interactions between 16 288 HUPO genes/proteins. To be consistent with the terminology used by Alexeyenko *et al.* (2012), each mutated gene is referred to as a functional gene set (FGS, Fig. 1). While such a set usually contains multiple genes, in our application each mutated gene is assessed independently for its central functional role in modulating the expression of its interacting neighbors in the network. Also following the NEA terminology, a collection of DE genes based on the expression data is called the altered gene set (AGS). We compute a quantitative enrichment score (z-score) as

$$z = \frac{d_{\mathrm{AF}} - \mu_{\mathrm{AF}}}{\sigma_{\mathrm{AF}}},$$

where d_{AF} is the number of network links between genes in the AGS and the FGS, and μ_{AF} and σ_{AF} are the expected mean and standard deviation of d_{AF}. These parameters are computed using a randomized network under the null hypothesis of no enrichment (Alexeyenko *et al.*, 2012). See also Figure 2 for an illustration of AGS links to some mutated genes. The z-score measures the over-representations of direct links between the AGS and the FGS, hence the transcriptomic impact of the mutation. This in turn allows us to prioritize the mutated genes. We only consider genes with positive z-value >2 as potential driver genes.

Because the underlying gene/protein interaction networks are available only at gene level, the NEA must be done at gene level. In setting up the AGS and FGS, we simply use the gene associated with the isoform. If there are multiple isoforms involved, the corresponding gene is set only once in the AGS or FGS.

Let us take *BRCA1* as an illustration of how the integrative algorithm works up to this point. *BRCA1* has six isoforms, all of which contain at least one point mutation with a mutation frequency of 30% across the patients. Three of the isoforms, NM_007294, NM_007297 and NM_007300, are predicted to have a moderate functional impact on the protein coding, thus kept in the subsequent analysis. Hence, *BRCA1* gene is considered as an FGS in the NEA.

Depending on the way we define the AGS, NEA can be used to identify mutated potential driver genes that are either common driver genes or patient-specific driver genes. For the common drivers, the AGS is derived from the top 100 genes for which at least one of their isoforms exhibit the greatest DE among all isoforms in terms of isoform-level expression between all tumors versus the paired normal. As stated earlier, each gene appears only once in the AGS; for example, two isoforms of the *NUF2* gene, NM_145697 and NM_031423, are ranked as the 34th and 36th top DE isoforms, respectively, so *NUF2* is set once in the AGS. For the patient-specific drivers, the AGS contains the top 100 genes for which at least one of their isoforms exhibit the largest fold change, computed as tumor/normal expression ratio in a specific patient.

While NEA measures how important a potential driver is in the gene network, extreme expression may be another indicator for the mutated potential driver to have a dramatic functional impact in activating or silencing expression of the neighbor genes. So to serve successfully as a putative driver, over- or under-expression, ranked over the top 90th percentile of the absolute $\log_2$ T/N expression distribution for which at least one of their isoforms is also required.

Although a single mutation may not be associated with poor cancer survival (Goodwin *et al.*, 2012), genomic instability accumulated from a number of cancer-related mutations may result in a worse prognosis (Salomon *et al.*, 2013). So we compute the total number of drivers weighted by their corresponding NEA *z*-scores as a measure of mutation load, so a driver with a higher *z*-score would be assigned a larger weight. We term the weighted sum a 'driver-gene score' (DGscore).

A more detailed description of the integrative algorithm can be found in the Supplementary Material (section Integrative algorithm). We provide a website http://fafner.meb.ki.se/biostatwiki/driver-genes/ with examples of scripts to use the various tools in the pipeline.

2.3 Clinical validation and comparisons

We use the DGscore derived from the integrative analysis to assess the prognosis of the 60 TCGA patients, where we shall compare patients with DGscore larger than the median versus those lower than the median. We note that the DGscore has been computed independently from the survival information, so this survival analysis is unbiased. Two well-known prognostic signatures are compared: (i) the 70-gene MammaPrint (van 't Veer *et al.*, 2002) and (ii) the 50-gene PAM50 classifier (Parker *et al.*, 2009). These are described later. Survival times are visualized with Kaplan-Meier plots and the *P*-value for the difference between the survival curves is calculated by the log-rank test.

The expression profiles and clinical datasets for 117 breast cancer patients in the study conducted by van 't Veer *et al.* (2002) can be obtained from the R package mammaPrintData at http://astor.som.jhmi.edu/~marchion/breastTSP.html. Forty-four patients remained disease-free of at least 5 years after their initial diagnosis. The intensity values on *log* scale are standardized such that every patient has a mean intensity value of zero and a standard deviation of one. Out of the 70 prognostic markers in MammaPrint, 47 are present in the TCGA expression data; the rest appear to be expressed sequence tags (ESTs) that are not mapped in standard databases. The risk score is computed based on the correlation of the 47 genes from each TCGA patient to the average expression profile of the good-prognosis patients in the original van 't Veer *et al.* study (2002). The 60 patients in TCGA study are then split into two prognostic groups using median correlation coefficient as a cutoff.

The 50 genes in PAM50 are used for classification of breast cancer into five intrinsic molecular subtypes: basal-like, HER2-enriched, luminal A (LumA), luminal B (LumB) and normal-like. Two versions of risk of relapse (ROR) scores are proposed by Parker *et al.* (2009): (i) ROR-S uses the intrinsic subtypes predicted by the 50-gene PAM50 classifier and (ii) ROR-C combines the predicted subtypes and tumor size. These are

$$\texttt{ROR-S} = 0.05\texttt{basal} + 0.12\texttt{HER2} - 0.34\texttt{LumA} + 0.23\texttt{LumB}$$
$$\texttt{ROR-C} = 0.05\texttt{basal} + 0.11\texttt{HER2} - 0.23\texttt{LumA} + 0.09\texttt{LumB} + 0.17\,\texttt{Size}.$$

The centroids predicted by the 50-gene classifier are obtained from R package genefu (Haibe-Kains *et al.*, 2012). Correlation between the expression profile of the 50 genes and centroids is calculated for each TCGA60 paired samples. Using the correlation for the subtypes, ROR score is computed for each patient according to the coefficients listed earlier. Threshold is determined from the median of the ROR score to split the patients into two groups, for which survival time is compared.

3 Results

3.1 TCGA60 paired samples

Among the TCGA60 paired samples, we identify 16 recurrent mutated isoforms with high impact and 245 with moderate impact. Among these frequently mutated variants, the NEA analysis finds 17 common potential driver genes, namely, *CHD1L*, *ADCY10*, *HEATR1*, *MUC4*, *DSPP*, *PKHD1*, *LPA*, *COL14A1*, *MKI67*, *OVCH1*, *RNF17*, *DNAH3*, *CDH11*, *CRISPLD2*, *FANCA*, *BRCA1* and *LAMA1*. Figure 2 illustrates the gene network involving *FANCA*, *BRCA1*, *MKI67* and *HEATR1* genes. By construction in the algorithm, these genes are enriched in their links to genes that are differently expressed between tumors and normals.

To identify the patient-specific drivers that may be neglected in the common-driver genes, we perform the NEA within each patient, where the altered gene-set is defined as the top 100 isoforms having the largest fold-change in expression between tumor and normal. In total, we found 27 patient-specific drivers, but 15 of them are already among the common drivers; the other 12 are *TP53*, *NOMO1*, *AIFM1*, *BCLAF1*, *HMCN1*, *NBPF9*, *ABCA1*, *CAPN9*, *GPR98*, *LMO7*, *TTN* and *KRT14*. Illustrating the strictness of the algorithm, the *PIK3CA* gene, which is mutated in 14 patients, is not identified either as a common driver or as a patient-specific driver in any of the patients.

The computed DGscore ranges between 0 and 11.8, and a high DGscore is defined as larger than the median value of 0.54. Figure 3a shows that breast cancer patients with a high DGscore have relatively poor survival ($P = 0.001$). As shown in Figure 3b, had we simply reported the crude (uncharacterized) number of mutations, with 394 exhibiting high impact and 6738 with moderate impact, we would not be able to predict the patients' survival ($P = 0.25$), demonstrating the importance of filtering by frequency of mutation pattern, expression level and functional characterization by NEA.

Isoforms of a gene are templates for producing distinct but related proteins. Many isoforms have been found to be implicated in a wide range of human diseases (Nagao *et al.*, 2005). Thus, it is crucial to separate the isoforms of the same gene. Summarized gene expression may cancel out the different expression pattern between isoforms and subsequently affect the counting of drivers. In fact, DGscore calculated at gene-level is not a significant prognostic

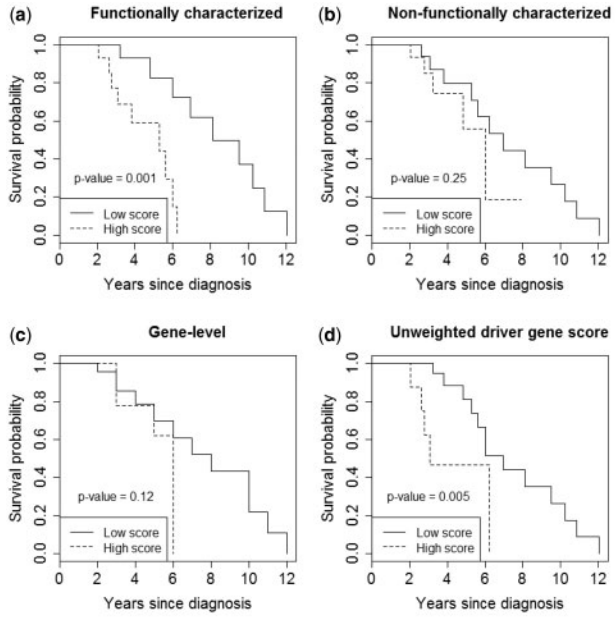

Fig. 3. Comparison of patient survival for DGscore derived by the full pipeline versus DGscore calculated when some steps in the pipeline are compromised. A high score (dashed black line) is defined as DGscore larger than the median value. (**a**) Functionally characterized drivers refer to the 17 common-driver genes and all patient-specific drivers identified by fully complying with the proposed algorithm. (**b**) Non-functionally characterized mutations refer to all the 394 mutated isoforms with high mutational impact and 6738 isoforms with moderate impact. (**c**) The DGscore is calculated based on gene expression, rather than isoform expression. (**d**) The DGscore is calculated following the proposed algorithm closely, except that the number of drivers is counted without any weighting scheme

factor (Fig. 3c, $P = 0.12$), suggesting the importance of isoform-level expression. A higher z-score of a candidate driver from NEA implies a more strongly connected network between a driver and its differentially expressed neighbors, so it seems sensible to assign a large weight to a high z-score driver. In Figure 3d, we observe that an unweighted DGscore could predict patient survival, although yielding a slightly less significant P-value of 0.005, when compared with the P-value of 0.001 derived from the weighted score (Fig. 3a).

As tumor stage, hormone-receptor status and age are known prognostic factors in breast cancer, next we investigate whether the DGscore has an independent prognostic effect. Table 1 summarizes the results of the Cox regression analysis. The first four rows of the table show the individual predictors in univariate regression, showing DGscore to be the only significant predictor. The last three lines indicate that DGscore remains highly significant after adjusting for estrogen-receptor status, tumor stage or age.

An important feature of the DGscore is that it integrates information on mutation and expression, each of which provides a different view of potential molecular defects in cancer. To include a gene as a potential driver, over- or under-expression is required for it to be counted in the DGscore. If we ignore this criterion, all common drivers and patient-specific driver genes characterized by NEA would be counted in the DGscore even though not expressed, potentially resulting in false positive cancer drivers. Figure 4a shows that the incomplete DGscore is not associated with patient survival ($P = 0.72$).

Similarly, as the 17 common-driver genes are not mutated in every patient, we should be careful in accounting for the contribution of mutated common-driver genes to the DGscore. We should

Table 1. Cox proportional-hazard regression models of survival

Variable	Hazard ratio	P*
DGscore	7.26	0.004
Tumor stage[a]	1.67, 2.22	0.44, 0.24
ER status	0.90	0.86
Age	0.98	0.30
DGscore + ER status[b]	20.23; 0.82	0.005; 0.75
DGscore + age	8.04; 1.01	0.007; 0.75
DGscore + tumor stage[a]	6.12; 1.39, 1.77	0.009; 0.62, 0.42

ER, estrogen receptor.
[a]Stage II and stage III+ are compared against stage I.
[b]ER negative is used as reference group.
*P-values from the Wald test.

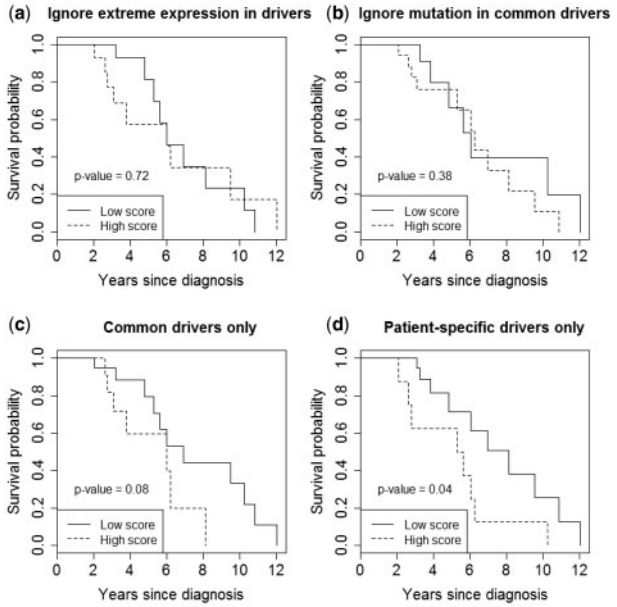

Fig. 4. Comparison of patient survival for DGscore derived based on partial molecular data. A high score (dashed black line) is defined as DGscore larger than the median value. (**a**) Extreme expression is not an inclusion criterion; the DGscore takes more driver genes into account than in the proposed algorithm. (**b**) Always include all 17 common drivers in the DGscore regardless of mutation status. (**c**) The DGscore takes only common drivers into account. (**d**) The DGscore takes only patient-specific drivers into account

re-evaluate the survival characteristics associated with the DGscore in the absence of mutation status for the common-driver genes. To do so, every common-driver gene is optimistically assumed to contain a somatic mutation, while patient-specific driver genes are defined and accounted for in the usual way. The survival of patients with the resulting high DGscore shows no difference to those with low DGscore (Fig. 4b, $P = 0.38$). This confirms our hypothesis that both types of molecular data, the somatic mutation and outlying expression of the potential driver genes, carry prognostic information.

Another important feature of the DGscore is that it summarizes the recurrent and individualized signatures in tumors, by incorporating information contained in common-driver genes and patient-specific driver genes. When we neglect patient-specific genes and focus only on common-driver genes, we find that the partial list of potential drivers cannot predict patient survival well (Fig. 4c, $P = 0.08$). We would also miss out on well-established cancer genes like *TP53*, which is observed as a patient-specific driver gene in three patients. Alternatively, when we use NEA to select potential

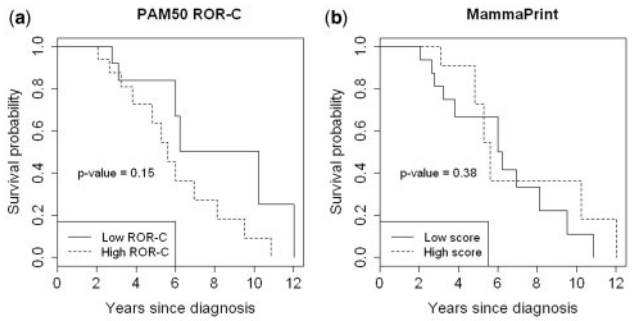

Fig. 5. (**a**) Performance of PAM50 based on the ROR-C score. (**b**) Performance of MammaPrint

driver genes based on T/N expression profiles within individual patient only, the survival curves of the high and low DGscore groups exhibit only a borderline significant difference ($P = 0.04$, Fig. 4d).

3.2 Prediction comparison

We now compare with the performance of PAM50 ROR-C and MammaPrint in predicting survival using the TCGA60 samples: see Figure 5a and b. For the PAM50 signatures, there is little difference in the prediction based on combined tumor size and intrinsic subtypes (ROR-C, $P = 0.15$) compared that based on subtypes only (ROR-S, $P = 0.19$), so we only show the former. MammaPrint also does not produce statistical significant result ($P = 0.38$). Thus DGscore is the most significant predictor (Fig. 3a, $P = 0.001$).

We further investigate whether ROR-C and DGscore predict prognosis independently. Given barely any correlation between these two signatures, we fit a Cox proportional hazards regression model with both as predictors. The *P*-values for DGscore and ROR-C are 0.001 and 0.04, respectively, indicating DGscore is an independent prognostic factor of survival.

3.3 Non-paired TCGA data

The availability of both DNA and RNA-seq data of paired tumor-normal tissue in the TCGA60 samples is unique. We cannot find other public breast cancer datasets that contain as rich information as in the TCGA60 paired samples. Another set of 671 TCGA breast-cancer samples, which do not overlap with the TCGA60 samples, is available from International Cancer Genome Consortium (Zhang *et al.*, 2011).The dataset is described in detail in the Supplementary Material (section Non-paired TCGA data). But unfortunately, only expression data are available for tumor tissues and estimated only at gene-level. We use the 17 common-driver genes as discovered in the TCGA60 samples, while patient-specific driver genes are assessed for the 671 samples. This is because the 17 common-driver genes are identified based on a list of AGS derived from the top 100 genes whose isoforms exhibit the greatest DE between all tumor versus all normal tissues. Thus, we wanted to check if these 17 common-driver genes are generalizable to other breast cancer patients.

In the absence of matched normal tissue, expression fold-change of T/N cannot be obtained. So the expression-altered gene set in NEA is defined as those genes whose absolute expression has the top 100 ranks among patients, to access the patient-specific driver genes. Vital status data are available, but not survival time, so only the Wilcoxon rank-sum test is carried out. The test shows no significant difference in DGscores between patients who have died and those still alive ($P = 0.78$). This result is consistent with what we observe in Figure 3c, demonstrating the necessity of isoform-level quantification in order for the algorithm to be able to identify potential cancer drivers. Furthermore using microarray-derived gene expression, we investigate the properties of the proposed algorithm to identify putative driver genes, under the circumstances of no mutation data nor isoform-level expression. We obtain a negative result of *P*-value 0.75 which is also in line with our expectation, since mutation status and isoform-level information is not available. The dataset and result are described in detail in the Supplementary Material (section Swedish microarray data).

4 Discussion

We have developed and illustrated an analytical framework to exploit the various molecular data that will likely be commonly gathered from cancer patients. This framework allows integration of mutation, expression and functional data, a weighting scheme for counting putative driver genes and a method for identifying drivers in a global and patient-specific manner based on the network enrichment analysis. The resulting score is shown to have clinical relevance in terms of significant association with patient survival that is stronger compared with other existing expression-based signatures.

Some of the identified potential common-driver genes, such as *FANCA*, *CHD1L*, *ADCY10*, *MUC4*, *PKHD1*, *MKI67*, *OVCH1* and *BRCA1*, are implicated in many important functional roles in tumorigenesis, such as DNA repair, cell proliferation and differentiation (Castella *et al.*, 2011; Cooke and Brenton, 2011; Flacke *et al.*, 2013; Mukhopadhyay *et al.*, 2013; Nishimiya *et al.*, 2014; Pines *et al.*, 2012; Tang *et al.*, 2014; Zhang *et al.*, 2012). The algorithm also allows the identification of patient-specific driver genes, which are candidates for contributing to personalized treatment. The identified patient-specific genes include the extensively reported tumor suppressor gene *TP53*. In a recent study, sustained expression of mutant *TP53* has been shown to be required to drive pancreatic cancer metastasis (Weissmueller *et al.*, 2014), implying that for a tumor suppressor to be deleterious, its expression does not have to be downregulated. This study also supports our criteria in selecting potential cancer drivers, for which both mutation and abnormal expression are required, regardless of the direction of regulation.

Unlike some existing expression-based signatures, our method does not simply produce prognostic biomarkers. What we gain from the integration of molecular and functional data is a list of patient-specific potential cancer drivers, so it has specific implications for personalized cancer therapy. A well-characterized true driver will also provide biologists with insights into cancer etiology.

Our study has its limitations. First, it is based on a small sample. Presently, we are not aware of large-scale study using the paired tumor-normal samples with both DNA and RNA-sequence data. But even in this small study, the DGscore is highly correlated with patient survival, more so than some existing prognostic factors, such as tumor stage and estrogen-receptor status, and some existing expression-based signatures. Applying MammaPrint and PAM50 gene signatures to the TCGA expression data, we observe some differences in the survival curves, although none of the *P*-values are significant. The failure of these established prognostic markers is most likely due to the small sample size but perhaps also because they were trained using distant recurrence/metastasis rather than survival outcome. Only 67% of the 70 gene identifiers in MammaPrint is mapped to the RefSeq gene names in the TCGA data, whereas the unmapped ones are those not annotated to standard databases. Hence, it is possible that the fewer number of mapped genes relative to the original study results in the suboptimal performance of MammaPrint.

Second, since only exome sequencing, but not whole–genome sequencing, is available, the data cannot capture driver mutations residing in non-coding regions. Furthermore, the annotation and functional characterization of genes rely on (i) known databases, (ii) non-directional biological network and (iii) predicted impact on protein coding. These are most likely incomplete and not necessarily validated. With more progress in completing the functional annotation, our approach would make an even more accurate identification of the potential drivers. Experimentally validated directional networks would be useful for assessing the causal relationship between a putative driver gene and its neighbors. Future works will include applying the scheme to identify putative cancer drivers in large-scale datasets in different cancers.

5 Conclusion

We develop a practical analysis pipeline to perform somatic variant calling, gene and isoform expression quantification and the integration of genomic and transcriptomic profiles based on known biological network and the functional impact on protein coding. Using this methodology, we show that breast cancer patients who carry more mutated genes with functional implications and extreme expression pattern have worse survival than those with less potential driver genes.

Funding

This project was supported by grants from the Swedish Science Council and the Swedish Cancer Foundation.

Conflict of Interest: none declared.

References

Akavia,U.D. *et al.* (2010) An integrated approach to uncover drivers of cancer. *Cell*, **143**, 1005–1017.

Alexeyenko,A. and Sonnhammer,E. (2009) Global networks of functional coupling in eukaryotes from comprehensive data integration. *Genomes Res.*, **19**, 1109–1116.

Alexeyenko,A. *et al.* (2012) Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. *BMC Bioinformatics*, **13**, 226.

Cancer Genome Atlas Network. (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

Castella,M. *et al.* (2011) Origin, functional role, and clinical impact of Fanconi anemia FANCA mutations. *Blood*, **117**, 3759–3769.

Cingolani,P. *et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly(Austin)*, **6**, 80–92.

Ciriello,G. *et al.* (2012) Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.*, **22**, 398–406.

Cooke,S.L. and Brenton,J.D. (2011) Evolution of platinum resistance in high-grade serous ovarian cancer. *Lancet Oncol.*, **12**, 1169–1174.

Flacke,J.P. *et al.* (2013) Type 10 soluble adenylyl cyclase is overexpressed in prostate carcinoma and controls proliferation of prostate cancer cells. *J. Biol. Chem.*, **288**, 3126–3135.

Futreal,P.A. *et al.* (2004) A cencus of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.

Goodwin,P.J. *et al.* (2012) Breast cancer prognosis in BRCA1 and BRCA2 mutation carriers: an international prospective breast cancer family registry population based cohort study. *J. Clin. Oncol.*, **30**, 19–26.

Haibe-Kains,B. *et al.* (2012) genefu: Relevant Functions for Gene Expression Analysis, Especially in Breast Cancer. R package version 1.6.1. http://compbio.dfci.harvard.edu.

Hou,J.P. and Ma,J. (2014) DawnRank: discovering personalized driver genes in caner. *Genome Med.*, **6**, 56.

Lazar,V. *et al.* (2013) Integrated molecular portrait of non-small cell lung cancers. *BMC Med. Genomics*, **6**, 53.

Loeb,K.R. and Loeb,K.A. (2000) Significance of multiple mutations in cancer. *Carcinogensis*, **21**, 379–385.

McKenna,A. *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.

Mukhopadhyay,P. *et al.* (2013) MUC4 overexpression augments cell migration and metastasis through EGFR family proteins in triple negativebreast cancer cells. *PLoS One*, **8**, e54455.

Nagao,K. *et al.* (2005) Detecting tissue-specific alternative splicing and disease-associated aberrant splicing of the PTCH gene with exon junction microarrays. *Hum. Mol. Genet.*, **14**, 3379–3388.

Nishimiya,H. *et al.* (2014) Prognostic significance of Ki-67 in chemotherapy-naive breast cancer patients with 10-year follow-up. *Anticancer Res.*, **34**, 259–268.

Parker,J.S. *et al.* (2009) Supervised predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.*, **27**, 1160–1167.

Pines,A. *et al.* (2012) PARP1 promotes nucleotide excision repair through DDB2 stabilization and recruitment of ALC1. *J. Cell. Biol.*, **199**, 235–249.

Salomon,A.V. *et al.* (2013) Genome instability: a stronger prognostic marker than proliferation for early stage luminal breast carcinomas. *PLoS One*, **8**, e76496.

Sanchez-Garcia,F. *et al.* (2014) Integration of genomic data enables selective discovery of breast cancer drivers. *Cell*, **159**, 1461–1475.

Suo,C. *et al.* (2014) Joint estimation of isoform expression and isoform-specific read distribution using multi-sample RNA-seq data. *Bioinformatics*, **30**, 506–513.

Tang,M.K. *et al.* (2014) BRCA1 deficiency induces protective autophagy to mitigate stress and provides a mechanism for BRCA1 haploinsufficiency in tumorigenesis. *Cancer Lett.*, **346**, 139–147.

van 't Veer,L.J. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Weissmueller,S. *et al.* (2014) Mutant p53 drives pancreatic cancer metastasis through cell-autonomous PDGF receptor β signalling. *Cell*, **157**, 382–394.

Youn,A. and Simon,R. (2011) Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics*, **27**, 175–181.

Zhang,K.X. and Ouellette,F. (2011) CAERUS: Predicting CAncER oUtcomeS using relationship between protein structural information, protein networks, gene expression data, and mutation data. *PLoS Comput. Biol.*, **7**, e1001114.

Zhang,J. *et al.* (2011) International Cancer Genome Consortium Data Portal - a one-stop shop for cancer genomics data. *Database*, doi:10.1093/database/bar026.

Zhang,D. *et al.* (2012) Exome sequencing identifies compound heterozygous PKHD1 mutations as a cause of autosomal recessive polycystic kidney disease. *Chin. Med. J.*, **125**, 2482–2486.