TANDEM

A two-stage approach to maximize interpretability of drug response models based on multiple molecular data types

Aben, Nanne; Vis, Daniel J.; Michaut, Magali; Wessels, Lodewyk

Publication date
2016

Document Version
Final published version

Published in
Bioinformatics

Citation (APA)
Aben, N., Vis, D. J., Michaut, M., & Wessels, L. (2016). TANDEM: A two-stage approach to maximize interpretability of drug response models based on multiple molecular data types. *Bioinformatics*, *32*(17), i413-i420.

Bioinformatics, 32, 2016, i413–i420
doi: 10.1093/bioinformatics/btw449
ECCB 2016

OXFORD

TANDEM: a two-stage approach to maximize interpretability of drug response models based on multiple molecular data types

Nanne Aben[1,2], Daniel J. Vis[1], Magali Michaut[1,*] and Lodewyk F.A. Wessels[1,2,3,*]

[1]Division of Molecular Carcinogenesis, The Netherlands Cancer Institute, Amsterdam 1066CX, The Netherlands, [2]Faculty of EEMCS, Delft University of Technology, Delft 2628CD, The Netherlands and [3]Cancer Genomics Netherlands, Utrecht 3584CT, The Netherlands

*To whom correspondence should be addressed.

Abstract

Motivation: Clinical response to anti-cancer drugs varies between patients. A large portion of this variation can be explained by differences in molecular features, such as mutation status, copy number alterations, methylation and gene expression profiles. We show that the classic approach for combining these molecular features (Elastic Net regression on all molecular features simultaneously) results in models that are almost exclusively based on gene expression. The gene expression features selected by the classic approach are difficult to interpret as they often represent poorly studied combinations of genes, activated by aberrations in upstream signaling pathways.
Results: To utilize all data types in a more balanced way, we developed TANDEM, a two-stage approach in which the first stage explains response using upstream features (mutations, copy number, methylation and cancer type) and the second stage explains the remainder using downstream features (gene expression). Applying TANDEM to 934 cell lines profiled across 265 drugs (GDSC1000), we show that the resulting models are more interpretable, while retaining the same predictive performance as the classic approach. Using the more balanced contributions per data type as determined with TANDEM, we find that response to MAPK pathway inhibitors is largely predicted by mutation data, while predicting response to DNA damaging agents requires gene expression data, in particular *SLFN11* expression.
Availability and Implementation: TANDEM is available as an R package on CRAN (for more information, see http://ccb.nki.nl/software/tandem).
Contact: m.michaut@nki.nl or l.wessels@nki.nl
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Large-scale pharmacogenomics screens provide a wealth of information about potential mechanisms of drug response. In these screens, cell lines of different cancer types have been profiled molecularly (mutations, copy number alterations, DNA methylation and gene expression), as well pharmacologically (response to anti-cancer drugs) (Barretina *et al.*, 2012; Iorio *et al.*, 2016). Using drug response prediction models, statistical associations can be identified between the drug response and the molecular data. For example, the presence of a *BRAF* mutation predicts sensitivity to Vemurafenib in melanoma cell lines and a mutation in *TP53* predicts resistance to Nutlin-3a (Garnett *et al.*, 2012). By combining various data types in an integrative analysis, all molecular data can be employed to explain drug response. This is commonly achieved by performing Elastic Net regression (Zou and Hastie, 2005) on all molecular data types simultaneously (Barretina *et al.*, 2012; Costello *et al.*, 2014; Garnett *et al.*, 2012; Iorio *et al.*, 2016; Jang *et al.*, 2014). Throughout this work, we will refer to this approach as the 'classic approach' (Fig. 1A). While this approach could, in theory, use information from all molecular data types, we find that it typically leads to models that are mostly based on gene expression data. For instance, a *BRAF* mutation activates, via a cascade of signaling events, the transcription of many genes. As a result, the expression of these genes is tightly linked to the mutation status of the *BRAF* gene, and thus also predictive of response to Vemurafenib. When all molecular

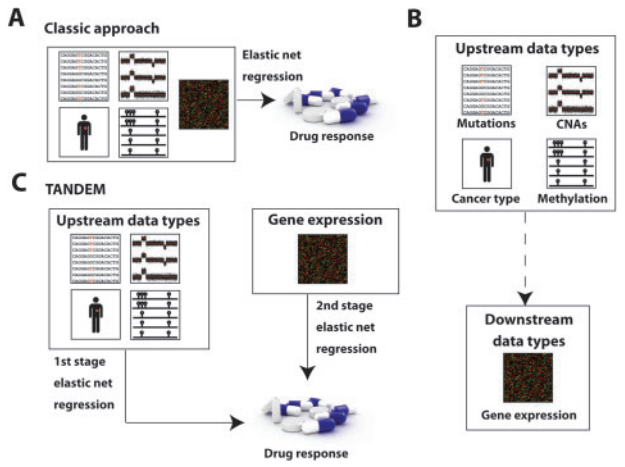

Fig. 1. Illustration of TANDEM and the classic approach. (A) The classic approach: an Elastic Net regression trained on all data types simultaneously. (B) The information predictive of drug response contained in the upstream data types is also present in the gene expression data. (C) TANDEM: our two-stage approach, which first uses the upstream data types to explain as much of the drug response as possible, and then uses the gene expression to explain the remainder

data are combined to build a predictive model for response to Vemurafenib, expression of these genes may be selected instead of the *BRAF* mutation, which would make the resulting model more difficult to interpret. Instead, selecting the *BRAF* mutation as a feature in the model would be more informative about the mechanism of the drug and thus lead to a more interpretable model.

We propose TANDEM, an approach that employs a two-stage analysis to improve the interpretability of prediction models by preferentially using the data types upstream of gene expression. To this end, we first split the molecular data types into 'upstream data' (somatic mutation, copy number alteration (CNA), methylation and cancer type) and 'downstream data' (gene expression) (Fig. 1B). This separation is based on the idea that mutation status, for example, affects the transcription of genes downstream of the pathway in which the mutation resides. TANDEM analyzes the upstream and downstream data 'in tandem': it first explains as much of the drug response as possible using the upstream (more interpretable) data and then explains the remainder using gene expression data (Fig. 1C). Applying TANDEM to a panel of 934 cell lines profiled across 265 drugs (Iorio *et al.*, 2016), we find that the upstream data types contribute more to the prediction than in the classic approach. At the same time, TANDEM retains the same predictive performance as the classic approach. The features selected by TANDEM result in twice as many significant pathway enrichments compared with the classic approach, implying that the selected features are more informative about the mechanisms of drug response. Additionally, using the more balanced contributions of the various data types, we find that response to MAPK targeting drugs is mostly explained by mutation data, while predicting response to DNA damaging agents requires gene expression data.

2 Methods

2.1 Data set

The Genomic Determinants of Sensitivity in Cancer 1000 (GDSC1000) data comprises a panel of cell lines screened for 265 anti-cancer drugs (Iorio *et al.*, 2016). This panel contains 926 cell lines that are fully characterized for point mutations, copy number alterations (CNAs), methylation status and gene expression profiles. Based on the human tumor data from The Cancer Gene Atlas (TCGA) (The Cancer Genome Atlas Research Network *et al.*, 2013), Iorio *et al.* (2016) have performed feature selection resulting in a set of 305 mutation, 409 CNA and 312 methylation features, all of which are binary. Additionally, we considered 29 binary features indicating the cancer type and 17 737 continuous gene expression features. The drug response was summarized by the IC50 (concentration that inhibits 50% of the target).

2.2 Drug response prediction using the classic approach

For drug response prediction models based on the classic approach, we used linear Elastic Net regression (Zou and Hastie, 2005) implemented in the R package glmnet (Friedman and Hastie, 2009). The hyper-parameter λ was optimized using 10-fold cross-validation and α was set to 0.5. Predictive performance estimates were made using double-loop cross-validation.

2.3 Predicting the binary value of upstream features from gene expression

We first identified upstream features that are associated with drug response using a Mann–Whitney U test, and only selecting features significantly associated with response to at least one drug (Benjamini–Hochberg corrected $p < 0.05$). For each of the identified upstream features, we then predicted its binary value using logistic regression of the gene expression data. Again, we used the implementation from the R package glmnet (Friedman and Hastie, 2009), optimized λ using 10-fold cross-validation and set α to 0.5. The classification performance (area-under-the-ROC, AUROC) was determined using double loop cross-validation. Because the classes are often highly unbalanced (i.e. a mutation typically only occurs in tens of samples out of 926), we used stratified cross-validation for the outer loop. This way, we ensured that each outer loop contains at least one sample per class. For the same reason, we omitted all upstream features that appear in fewer than ten samples in total.

2.4 Relative contribution of each data type to the prediction

In order to determine the relative contribution of each data source, we created a prediction per data source. We determined the relative contribution RC_i for each data source by dividing the sum-of-squares of a prediction from a certain data type by the sum-of-squares of the overall prediction (see Supplementary materials and Methods section). We only took into account drugs for which we achieved a predictive performance $r > 0.4$. This prevents models with poor predictive performance from confounding the analysis.

2.5 The TANDEM algorithm

We used a two-stage approach to predict drug response: (i) Fit an Elastic Net model to predict the drug response using the upstream data types; (ii) Fit an Elastic Net model to predict the residuals from the first stage using the gene expression data. Like in the classic approach, λ was optimized using cross-validation and α was set to 0.5. We used the same separation in cross-validation folds for both stages. Similar to the classic approach, we used a double-loop cross-validation to estimate performance.

2.6 Feature importance score

The feature importance FI for feature j was determined as follows:

$$\mathrm{FI} = \frac{\|X_j\beta_j\|_2^2}{\|\hat{y}\|_2^2} \quad (1)$$

where X_j is the column j of $\mathbf{X}$. Without loss of generality, we assume that all columns of $\mathbf{X}$ and the prediction $\hat{y}$ are mean-centered.

2.7 Pathway enrichment

We downloaded version 5 of the KEGG pathways from MSigDB (Subramanian *et al.*, 2005) and used a hypergeometric test to quantify the enrichment of selected features within a pathway. The *P*-values were controlled for FDR by applying Benjamini–Hochberg correction per drug. For more details, see Supplementary materials and Methods section.

3 Results

3.1 The information in all data types is captured in the gene expression data

For each of the 265 drugs of the GDSC1000 pharmacogenomics panel, we first built drug response models for each drug and each data type separately using Elastic Net regression. We assessed the predictive performance of these models using the Pearson correlation coefficient between the observed and the predicted IC50s (Supplementary Table S1). The most predictive data type was found to be gene expression data: the median predictive performance of these models is higher compared to models based on other data types (Fig. 2A). This finding is consistent with previous work by Costello *et al.* (2014) and Jang *et al.* (2014). Subsequently, we built drug response models using all data types simultaneously, referred to as the 'classic approach'. We found that the predictive performance of models based on only gene expression and models based on the classic approach was nearly identical (median difference across drugs: 0.001, Fig. 2A). The predictive performance of these two methods is not only comparable at the median, but it is also highly correlated across all drugs in the panel (Pearson correlation coefficient across drugs: 0.99, Supplementary Fig. S1A), indicating that both methods achieve similar performance for the same drugs. Altogether, we found that adding upstream data does not improve a model based on gene expression only, implying that the information from the upstream data types is already contained in the gene expression data.

To investigate the possible redundancy between the upstream and the downstream data, we attempted to predict the upstream features (e.g. aberration status or cancer type) from downstream data (gene expression). For the 503 upstream features associated with drug response, predicting the aberration status or cancer type from gene expression resulted in a median AUROC of 0.88 (Supplementary Fig. S1B). Hence, we found that it is indeed possible to predict the upstream features with high accuracy from downstream data, which further corroborates that the information in the upstream features is also present in the gene expression data.

Finally, we investigated the relative contribution of each data type to models based on the classic approach. To assess the relative contribution of a given data type, we determined what fraction of the prediction using all data types is explained by that particular data type (Methods section and Supplementary Table S1). Despite the redundancy between the upstream and the downstream data, the models preferentially select gene expression features (Fig. 2B). For

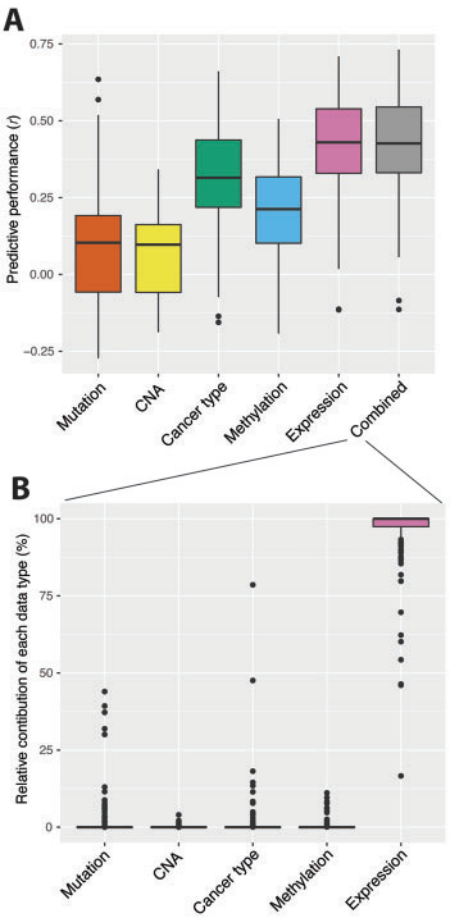

Fig. 2. Predictive performance of individual molecular data types. (A) Predictive performance (Pearson correlation between measured IC50s and predictions from the classic approach) across 265 drugs using individual data types (mutation, CNA, tissue of origin, methylation, gene expression) or a combination of all data types (combined) with the classic approach. (B) Relative contribution of each data type in the combined models, across all drugs for which we achieved a predictive performance $r > 0.4$

89% of the drugs, more than 90% of the variation in the prediction was attributed to gene expression. To investigate whether the high dimensionality and the continuous nature of the gene expression data had an effect on this result, we reduced the number of features and discretized the gene expression (Supplementary methods). In both cases, we still observed the domination of the gene expression in the models (Supplementary Fig. S1C). We concluded that neither the dimensionality nor the continuous nature of the data explain the high relative contribution of gene expression in the models based on the classic approach.

Altogether, we have shown that, in the context of drug response prediction, gene expression recapitulates the information contained in upstream data. Thus, we set out to exploit the redundancy between the upstream and downstream data to create more interpretable models.

3.2 TANDEM produces a more balanced contribution of different data types while maintaining the same performance

To utilize the information from gene expression data, without allowing it to completely dominate the models, we propose a two-stage approach to predict drug sensitivity. In the first stage,

TANDEM constructs a model to predict as much of the variation in the drug response as possible using the—more interpretable—upstream data types only. In the second stage, TANDEM explains the remainder of the variation in the drug response using gene expression data.

We illustrate the results of our method and its differences with the classic approach using three well-characterized drugs: Trametinib (a MEK inhibitor), Nutlin-3a (an MDM2 inhibitor) and Nilotinib (a BCR-ABL inhibitor). Using the classic approach, gene expression accounts for most of the prediction (Fig. 3A). For Trametinib, 94% of the prediction is attributed to gene expression data and only 6% is attributed to the upstream data types. In contrast, using TANDEM, we obtain a model where 32% of the prediction is attributed to gene expression and 68% to the upstream data types (Fig. 3B). The same holds for Nutlin-3a and Nilotinib: when employing TANDEM, the contribution of upstream data types increases dramatically, albeit in different proportions, while maintaining the same level of predictive performance (Fig. 3A and B).

Across all drugs for which we obtained a predictive performance $r > 0.4$, the median percentage of variation attributed to gene expression was 100% when using the classic approach, while it dropped to 52% when using TANDEM (Fig. 3C). In the latter case, the median percentage of variation explained by mutations, CNAs, methylation status and cancer type was 3%, 2%, 20% and 11%, respectively (Fig. 3C). In addition, TANDEM obtains virtually the same predictive performance as the classic approach (Fig. 3D) (Pearson correlation: 0.99, median difference: 0.002). In summary, TANDEM results in models that use all data types in a more balanced fashion, while retaining the same predictive performance as the classic approach.

3.3 TANDEM produces more interpretable models

TANDEM produces models that are mostly based on upstream data features. As these upstream features are more likely causally related to drug response, the resulting models are easier to interpret. To demonstrate the improved interpretability, we performed a pathway enrichment analysis of the genes identified by TANDEM as being associated with drug response. Using the KEGG pathways (Kanehisa and Goto, 2000; Kanehisa *et al.*, 2014), we tested all drug-pathway pairs for enrichment of predictive genes (i.e. genes associated with response to the drug in our model) among the genes annotated to this pathway. Since TANDEM preferentially uses the upstream data, which is enriched for well-studied genes, we were concerned with selection bias when testing for pathway enrichment against a genome-wide background distribution. To account for this bias, we instead defined the background distribution using only genes present in at least one KEGG pathway (Methods section). After correcting for multiple testing, TANDEM yielded more than twice (164 versus 64) the number of significant enrichments as compared to the classic approach (Supplementary Fig. S2A and B). The features selected by TANDEM can thus be related to existing knowledge (pathways) more easily than those selected by the classic approach, implying that the resulting models are more easily interpreted.

We illustrate these results using two significant enrichments from TANDEM: the features in the MAPK pathway associated with response to the MEK inhibitor Trametinib (Benjamini–Hochberg FDR corrected *P*: 1.0e−3, Fig. 4A) and the features in the B cell receptor signaling pathway associated with the HDAC6 inhibitor Tubastatin (Benjamini–Hochberg FDR corrected *P*: 5.3e−5, Fig. 5B). In both examples, the features selected by TANDEM resulted in a significant enrichment, whereas the features selected by the classic approach did not.

For Trametinib (a MEK inhibitor), both methods identified *KRAS*, *NRAS* and *BRAF* mutations to be associated with sensitivity (Supplementary Fig. S3A–C). This is expected as these mutations all activate MAPK signaling through MEK, and inhibition of MEK shuts down the pathway, thereby mitigating their effect and rendering mutated cell lines sensitive to Trametinib. TANDEM selected two additional mutations in the pathway: *HRAS* and *MYC* (Supplementary Fig. S3D and E). Like the aforementioned mutations, HRAS signals through MEK and hence *HRAS* mutations are associated with sensitivity. Myc proteins can harbor a mutation in their regulatory phosphorylation site, which allows them to escape ubiquitin/proteasome-mediated turnover and leads to accumulation of Myc protein (Bahram *et al.*, 2000). Because the mutated Myc proteins activate the downstream targets of the pathway independently of MEK, mutated cell lines are insensitive to the MEK inhibitor. Thus, this mutation is associated with resistance to MEK inhibition. In addition, both methods identified *DUSP6* as a predictive feature (Supplementary Fig. S3F). *DUSP6* transcription is induced by ERK activation (Furukawa *et al.*, 2008). Hence, by proxy, high *DUSP6* expression is an indication of high phospho-ERK levels. Since phospho-ERK can be attenuated by MEK inhibition, high *DUSP6* expression is associated with sensitivity to MEK inhibition (Jing *et al.*, 2012). *DUSP6* is an example of a gene expression feature

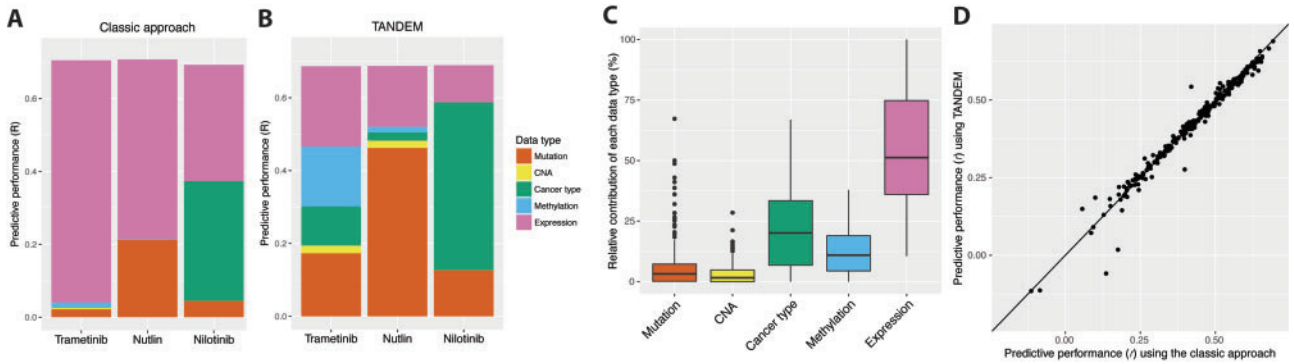

Fig. 3. Data type contribution and predictive performance. Relative contribution of each data type (indicated by the colors) and predictive performance (*r*, the Pearson correlation between observed and predicted IC50s) for three example drugs, using (A) the classic approach for data integration and (B) TANDEM. (C) Relative contribution of each data type in TANDEM, across 265 drugs, across all drugs for which we achieve a predictive performance $r > 0.4$. (D) Predictive performance of the classical approach versus TANDEM

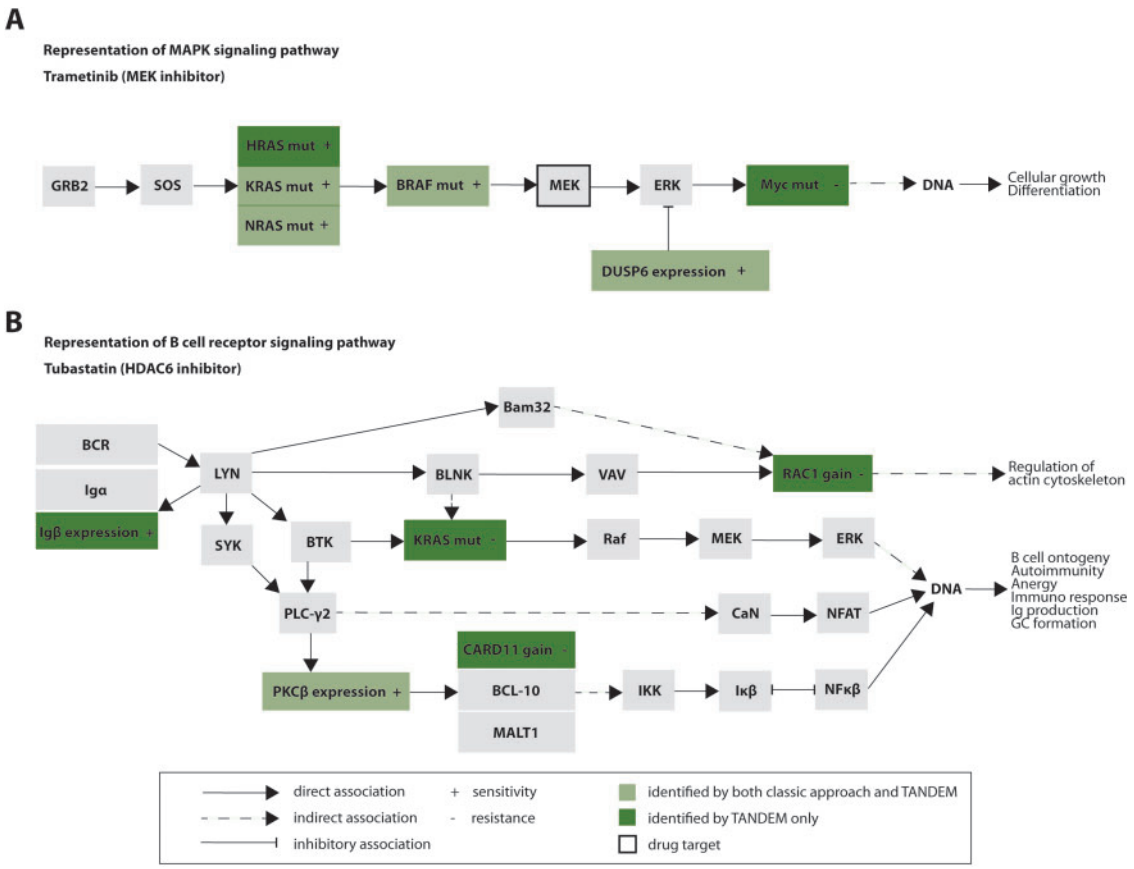

Fig. 4. Features selected by TANDEM in the context of two pathways. Representation of (A) the MAPK signaling pathway and (B) the B cell receptor signaling pathway from KEGG. Indicated in color are the genes associated with response to (A) Trametinib or (B) Tubastatin by TANDEM (dark green) or by both approaches (light green)

whose selection not only increases the predictive performance but also benefits the interpretability.

Our second example models the response to the HDAC6 inhibitor Tubastatin (Fig. 4B), an anti-inflammatory drug (Butler *et al.*, 2010; Vishwakarma *et al.*, 2013) that has shown anti-cancer potential (Hideshima *et al.*, 2005; Minucci and Pelicci, 2006). Unlike other members of the HDAC family, HDAC6 is exclusively localized in the cytoplasm and hence does not have a histone deacetylase

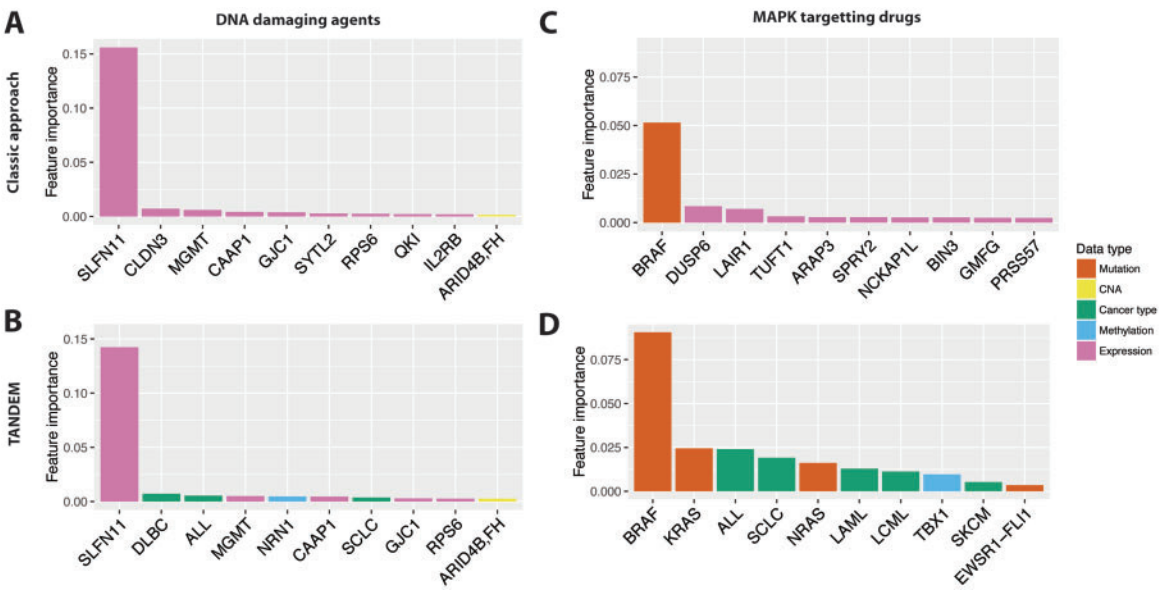

Fig. 5. Most important features for predicting response to DNA damaging agents and MAPK pathway inhibitors. Top 10 most important features (based on their average feature importance score) for predicting response to MAPK-targeting drugs (A and B) or DNA damaging agents (C and D) using the classic approach (A and C) or TANDEM (B and D)

function (Gao *et al.*, 2007; Hubbert *et al.*, 2002). Instead, Gao *et al.* (2007) have proposed that HDAC6 is required for efficient Rac1 activation. Interestingly, TANDEM identifies *RAC1* amplifications to be associated with resistance to Tubastatin (Supplementary Fig. S3G), whereas the classic approach does not. This could mean that when Rac1 is available in abundant levels, efficient activation of Rac1 by HDAC6 is not required anymore and hence HDAC6 inhibition has little effect, causing resistance. Both methods associated *PKCβ* expression with sensitivity to Tubastatin (Supplementary Fig. S3H). One additional gene expression feature was uniquely identified using TANDEM: the expression of *Igβ* (Supplementary Fig. S3I). As *PKCβ* and *Igβ* both reside in the B cell receptor signaling pathway, their selection could mean that Tubastatin is especially potent in B cell-derived lymphoid cancers with active B cell receptor signaling. This is further supported by a negative correlation between the expression of *Igβ* (a component of the B cell receptor) and response to Tubastatin within the 68 B cell derived lymphoid cell lines in the GDSC1000 data set (Pearson correlation coefficient: −0.49, Supplementary Fig. S3J).

Altogether, we found that the features identified by TANDEM can be interpreted in the context of pathways. Due to the more balanced contributions of upstream and downstream data types, we show that our method leads to improved interpretability of the drug response models, while achieving the same predictive performance.

3.4 Different data types predict response to different drug classes

To test if certain data types better predicted response to certain classes of drugs, we used the drug classification provided with the GSDC1000 data (Iorio *et al.*, 2016), where all 265 drugs are categorized into 21 classes, based on either the mechanism of action (e.g. DNA damaging agents) or the pathway in which the drug target resides (e.g. MAPK pathway). For a given drug class, we considered the relative contribution each data type makes to the prediction using TANDEM, using only the drugs for which a model could be built with predictive performance $r > 0.4$. Using these relative contributions, we tested each drug class for association with each data type (Supplementary Fig. S4). We further investigated two associations: the most significant association using upstream data (MAPK pathway inhibitors and mutation data) and the most significant association using downstream data (DNA damaging agents (DDAs) and gene expression). For these drug classes, we determined the top 10 most important features using both the classic approach and the TANDEM. The feature importance was assessed based on the size of the regression coefficient, corrected for the variance of the corresponding feature (Supplementary Table S1).

3.5 Gene expression data is the best predictor of response to DNA-damaging agents

For the 10 drugs from the DDA drug class, the response models produced by TANDEM had a higher contribution of gene expression compared to other drug classes (Benjamini–Hochberg corrected *P*: 0.046, one-tailed Mann–Whitney test, Supplementary Fig. S5A). Given that our method preferentially uses upstream features, we found it intriguing that gene expression still accounts for a median 76% of the explained variation. In fact, the contribution of gene expression is mostly due to the expression of *SLFN11*, which is the most important predictor of response to DDAs in both the classic approach and TANDEM (Fig. 5A and B). Part of the information contained in the expression of *SLFN11* is also present in some upstream features, which results in a lower feature importance for *SLFN11* when using TANDEM. For example, *SLFN11* expression is significantly higher in the ALL (*P*-value: 5.2e−9, Supplementary Fig. S5B). However, as TANDEM selects *SLFN11* expression after the acute lymphoid leukemia (ALL) cancer type has been selected, we can rule out that *SLFN11* is merely selected as a proxy for ALL. Altogether, this points to an important role for *SLFN11* in DDA response. Indeed, Zoppoli *et al.* (2012) have found that knockdown of *SLFN11* leads to increased resistance to many DDAs, indicating a causative role for *SLFN11* expression.

3.6 Mutations are the best predictors of response to MAPK pathway inhibitors

For the 16 drugs from the MAPK pathway inhibition class, the response models produced by TANDEM had a significantly higher contribution of mutation data compared to other drug classes (Benjamini–Hochberg corrected *P*: 1.1e−5, one-tailed Mann–Whitney test, Supplementary Fig. S5C). Investigating the most important features obtained using both methods (Fig. 5C and D), we found that they both identified the *BRAF* mutation as the strongest predictor of response, as expected (Downward, 2003). The remaining part of the top 10 features is completely different between the two methods: for the classic approach, it solely consists of gene expression features, whereas for TANDEM it consists of upstream features. TANDEM identifies *KRAS* and *NRAS*, two canonical mutations known to modulate response to MAPK pathway inhibitors (Downward, 2003), while the gene expression features identified by the classic approach do not give clear insight into the mechanisms of drug response. Consistent with the literature, TANDEM also associates a number of cancer types with response to MAPK inhibition: melanoma (SKCM), acute myeloid leukemia (LAML) and chronic myeloid leukemia (LCML) are associated with sensitivity (Geest and Coffer, 2009; Inamdar *et al.*, 2010), whereas small cell lung cancer (SCLC) is associated with resistance (DeGregori, 2006; Ravi *et al.*, 1998).

3.7 TANDEM prevents cancer type specific expression from confounding the results

Using cancer type as an upstream feature, TANDEM avoids the selection of genes whose expression is specific to one cancer type. In the MAPK inhibitors example above, the classic approach selects *LAIR1* and *PRSS57* as important features (positions 3 and 10 in the top 15 classic approach features). However, these genes are preferentially expressed in LAML and LCML ($P < 2.2e{-}16$, Mann–Whitney *U* test, Supplementary Fig. S6A and B). Thus, the selection of LAML and LCML cancer types as important features by TANDEM is much more informative. Similarly, the classic approach selects *BIN3* expression, but *BIN3* is preferentially expressed in SKCM ($P < 2.2e{-}16$, Mann–Whitney *U* test, Supplementary Fig. S6C). The selection of SKCM by TANDEM is, therefore, more informative.

To further look for a possible link between expression of these genes and drug response as identified by the classic approach, we investigated whether these three genes are involved in the resistance mechanism in the cell lines of the corresponding cancer type. To do this, we tested the correlation between these genes and response to MAPK pathway inhibitors within the respective cancer type. None of these genes showed a significant correlation with the drug response (Supplementary Fig. S6D–F) (Benjamini–Hochberg corrected $P > 0.05$, Pearson correlation). Unless this is due to small sample size and multiple testing correction, this supports the conclusion that these gene expression features are selected as a proxy for cancer

type and are not directly associated with drug response. Hence, TANDEM more accurately indicates the cancer type as a predictive feature.

Altogether, we have shown that by using the different data types in a more balanced fashion, TANDEM replaces part of the gene expression signatures by various upstream features, such as mutations and cancer type features (MAPK pathway inhibitors). At the same time, for the gene expression features that are selected by TANDEM, such as *DUSP6* (Trametinib) and *SLFN11* (DDAs), we can rule out that they are merely selected as a proxy for a specific cancer type.

4 Discussion

Large-scale pharmacogenomics screens can offer insights into relations between molecular data and drug response. By integrating the various data types, the molecular data can be comprehensively associated to drug response. However, we have shown that the classic approach for data integration (Elastic Net regression on all molecular data types simultaneously) results in models that are largely based on gene expression. This can be attributed to the redundancy in information between the upstream and downstream data. Here, we introduced TANDEM, an approach that preferentially uses the upstream data types, and only adds gene expression when necessary. The resulting models have a much larger contribution of upstream data types, while retaining the same predictive performance as the classic approach.

The main advantage of TANDEM is that the resulting models are more interpretable. By focusing on the upstream data types first, the analysis is prevented from being confounded by the expression of genes that are either specific to the cancer type or serve as 'signatures' of the aberration status of upstream genes. Yet, because the model uses gene expression in the second stage, our method also identifies relevant genes, such as *SLFN11* (DNA damaging agents) or *DUSP6* (Trametinib), based on their gene expression patterns.

De Bin *et al.* (2014) have investigated additional strategies to combine redundant data, in particular clinical and molecular data. In their 'favoring' strategy, they remove the regularization penalty from the clinical data to 'favor' clinical data over the rest. This approach was not feasible in our setting, as the upstream data is high-dimensional and removing the regularization would result in the inversion of a singular matrix. Similar to their 'dimension reduction' strategy, we reduced the dimensionality of the gene expression data, but we found that this still leads to models that are dominated by gene expression data (Supplementary Fig. S1C). For the combination of multiple molecular data types, we found that a two-stage approach (in their terminology: a 'residuals strategy') works well to combine upstream and downstream data types.

Redundancy between molecular data types has been explored before. Wang *et al.* (2013) have shown that the information from methylation status is captured in gene expression profiles. Although they did not study drug response prediction in cell lines, but rather investigated clinical outcome in patients, their results support our idea of redundancy captured by upstream and downstream data types. In the model by Wang *et al.*, the gene expression is decomposed in two parts, based on whether it can be modulated by methylation. This can provide insight in relations between methylation and gene expression features. Explicitly modeling the relations between gene expression and upstream data could be an interesting extension for TANDEM.

Similar to the redundancy between methylation and gene expression, Iorio *et al.* (2016) observed that, in GDSC1000, the gene expression data captures a large fraction of the information regarding the cancer type. In agreement with the observations made by Iorio *et al.* (2016), we found that the cancer type features show the strongest redundancy with gene expression. We extended these ideas by considering not only the redundancy between gene expression and either methylation or cancer type, but by jointly considering all other data types. In the future, it would be interesting to assess whether gene expression also captures information from other molecular effects, such as miRNAs.

In this work, we have introduced TANDEM, a two-stage approach that improves the interpretability of the resulting drug response models by focusing on upstream features, while retaining good predictive performance. We believe that advances in the integrated analysis of multiple molecular data types will lead to a better understanding of the mechanisms of drug response and ultimately to improved treatments in the clinic.

Acknowledgements

We thank Ultan McDermott and Mathew Garnett for early pre-publication access to the GDSC1000 pharmacogenomics data set. We also thank Gergana Bounova and Remco Nagel for critically reading the manuscript and providing feedback.

Funding

This work was funded by the ERC Synergy Project CombatCancer.

Conflict of Interest: none declared.

References

Bahram,F. *et al.* (2000) c-Myc hot spot mutations in lymphomas result in inefficient ubiquitination and decreased proteasome-mediated turnover. *Blood*, **95**, 2104?2110.

Barretina,J. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603?607.

Butler,K.V. *et al.* (2010) Rational design and simple chemistry yield a superior, neuroprotective HDAC6 inhibitor, tubastatin A. *J. Am. Chem. Soc.*, **132**, 10842?10846.

Costello,J.C. *et al.* (2014) A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.*, **32**, 1202?1212.

De Bin,R. *et al.* (2014) Investigating the prediction ability of survival models based on both clinical and omics data: two case studies. *Stat. Med.*, **33**, 5310?5329.

DeGregori,J. (2006) Surprising dependency for retinoblastoma protein in ras-mediated tumorigenesis. *Mol. Cell. Biol.*, **26**, 1165?1169.

Downward,J. (2003) Targeting RAS signalling pathways in cancer therapy. *Nat. Rev. Cancer*, **3**, 11?22.

Friedman,J. and Hastie,T. (2009) glmnet: Lasso and elastic-net regularized generalized linear models.

Furukawa,T. *et al.* (2008) Feedback regulation of DUSP6 transcription responding to MAPK1 via ETS2 in human cells. *Biochem. Biophys. Res. Commun.*, **377**, 317?320.

Gao,Y.S. *et al.* (2007) Histone deacetylase 6 regulates growth factor-induced actin remodeling and endocytosis. *Mol. Cell. Biol.*, **27**, 8637?8647.

Garnett,M.J. *et al.* (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, **483**, 570?575.

Geest,C.R. and Coffer,P.J. (2009) MAPK signaling pathways in the regulation of hematopoiesis. *J. Leukoc. Biol.*, **86**, 237?250.

Hideshima,T. *et al.* (2005) Small-molecule inhibition of proteasome and aggresome function induces synergistic antitumor activity in multiple myeloma. *Proc. Natl. Acad. Sci. U S A*, **102**, 8567?8572.

Hubbert,C. *et al*. (2002) HDAC6 is a microtubule-associated deacetylase. *Nature*, **417**, 455?458.

Inamdar,G.S. *et al*. (2010) Targeting the MAPK pathway in melanoma: why some approaches succeed and other fail. *Biochem. Pharmacol*., **80**, 624?637.

Iorio,F. *et al*. (2016) A landscape of pharmacogenomic interactions in cancer. *Cell*, DOI: 10.1016/j.cell.2016.06.017.

Jang,I.S., *et al*. (2014) Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. *Pac. Symp. Biocomput*., 63?74.

Jing,J. *et al*. (2012) Comprehensive predictive biomarker analysis for MEK inhibitor GSK1120212. *Mol. Cancer Ther*., **11**, 720?729.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*., **28**, 27?30.

Kanehisa,M. *et al*. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res*., **42(Database issue)**, D199?D205.

Minucci,S. and Pelicci,P.G. (2006) Histone deacetylase inhibitors and the promise of epigenetic (and more) treatments for cancer. *Nat. Rev. Cancer*, **6**, 38?51.

Ravi,R.K. *et al*. (1998) Activated Raf-1 causes growth arrest in human small cell lung cancer cells. *J. Clin. Invest*., **101**, 153?159.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U S A*, **102**, 15545?15550.

The Cancer Genome Atlas Research Network *et al*. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet*., **45**, 1113?1120.

Vishwakarma,S. *et al*. (2013) Tubastatin, a selective histone deacetylase 6 inhibitor shows anti-inflammatory and anti-rheumatic effects. *Int. Immunopharmacol*., **16**, 72?78.

Wang,W. *et al*. (2013) iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, **29**, 149?159.

Zoppoli,G. *et al*. (2012) Putative DNA/RNA helicase Schlafen-11 (SLFN11) sensitizes cancer cells to DNA-damaging agents. *Proc. Natl. Acad. Sci. U S A*, **109**, 15030?15035.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. Roy. Stat. Soc. B*, **67**, 301?320.