

RESEARCH

# Causal modeling dissects tumour-microenvironment interactions in breast cancer

Leon Chlon\* and Florian Markowetz

\*Correspondence:

lc574@cam.ac.uk

Cancer Research UK Cambridge Institute, University of Cambridge, Robinson Way, CB2 0RE Cambridge, UK

Full list of author information is available at the end of the article

## Abstract

Elucidating interactions between cancer cells and their microenvironment is a key goal of cancer research with implications for understanding cancer evolution and improving immunotherapy. Previous studies used association-based approaches to infer relationships in transcriptomic data, but could not infer the direction of interaction. Here we present a causal modeling approach that infers directed interactions between signaling pathway activity and immune activity by anchoring the analysis on somatic genomic changes. Our approach integrates copy number profiles, transcriptomic data, image data and a protein-protein interaction network to infer directed relationships. As a result, we propose 11 novel genomic drivers of T cell phenotypes in the breast cancer tumour microenvironment and validate them in independent cohorts and orthogonal data types. Our framework is flexible and provides a generally applicable way to extend association-based analysis in other cancer types and to other data and clinical parameters.

**Keywords:** Breast Cancer; Cancer Immunology; Tumour Microenvironment; Causality; Integrative Analysis

## Background

Solid tumours like breast cancers are complex tissues consisting of a cell-autonomous compartment of cancer cells accumulating somatic changes and undergoing clonal evolution [1], and a non-cell-autonomous compartment containing lymphocytes, fibroblasts and other cell types forming the tumour microenvironment [2]. Both compartments are known to influence each other [3] but the details of how they communicate are still poorly understood. For example, it is unclear how breast cancer cells hijack signaling pathways and regulatory mechanisms to influence the recruitment and activity of tumour infiltrating lymphocytes (TILS) [4, 5].

Previous studies used association-based methods to characterise tumour-microenvironment interactions within the bulk [6, 7] and micro-dissected tumour transcriptomes [8]. For example, Ali *et al* [7] showed that patterns of immune infiltration varied between molecular subtypes of breast cancer in bulk tumour transcriptomes; and Oh *et al* [8] derived stromal-epithelial co-expression networks from micro-dissected tumour data to investigate crosstalk within the tumour microenvironment.

Data from transcriptomic studies are readily available, and association-based analyses have provided profound insights into the general structure of co-expression patterns within the data. However, association-based methods have several limitations. First of all, when analysing bulk data they are limited by the convolution of gene expression patterns arising from both the cell autonomous and non-autonomous

compartments [9]. Second, associations alone can not distinguish whether a change in gene expression is caused by a change in immune activity or is a response to it [10].

To overcome these limitations, we propose a statistical approach integrating data on somatic genomic changes, signaling pathway activity and immune activity in the tumour. Our approach assigns a direction to an association between signaling pathway activity and immune activity by anchoring the analysis on an underlying genomic event. We exemplify this general idea here by using copy number alterations (CNA) to measure genomic events, transcription factor (TF) activity inferred from transcriptional profiles to measure signaling activity, and expression of marker genes as well as imaging data to measure immune activity. Based on these data, our approach uses Bayesian networks and likelihood approaches to choose the best fitting model (similar to [10]), and reduces the search space by using protein-interaction networks to limit the number of potential models.

For discovery, we apply our approach to 1,980 breast tumour samples with paired genomic-transcriptomic data [11] to generate candidate models for transcriptomic phenotypes. For validation, we test these models in an independent genomic-transcriptomic dataset comprising 1,154 samples [12]. For further validation on an orthogonal measure of immune activity, we use data on lymphocyte infiltration derived from over 550 Haematoxylin & Eosin (H&E) stained whole tumour slides [13]. As a result, we find 12 models that robustly validate and propose 11 novel genomic drivers of T cell phenotypes in the breast cancer tumour microenvironment.

## Results

We developed a framework to formalise causal relationships between somatic genomic events, signaling pathway activity and immune activity in the tumour. Our approach is implemented in the statistical environment R [14] and all code to reproduce the results presented here is available as part of an annotated document in the supplementary information.

To measure genomic events we used the copy number profiles of 19,702 genes, as provided by the METABRIC [11] and TCGA projects [12]. To measure signaling pathway activity, we focused on 788 experimentally verified TFs [15] and used the paired transcriptional profiles from the same resources to apply VIPER [16], a method for network based prediction of transcription factor activity.

To measure immune activity, we used two orthogonal approaches: the first approach uses the mean expression of marker genes to define a cytolytic score (CS) [17] and a novel T-cell score (TCS). While the CS trait is a measure of lymphocyte activity, the TCS measures the degree to which they are represented in the tissue. We validated TCS on paired gene expression and flow cytometry blood sample data [9] and found that of the 9 leukocytes subsets tested, CD8+ T-cells demonstrated the only significant positive correlation with the TCS ( $\rho = 0.675$ ,  $P = 0.001$ ), while the remaining leukocytes show either negative or no significant correlation (see Additional file 1). The second approach uses paired H&E images from the METABRIC cohort to measure the absolute number of lymphocytes and their density per tumour [13]. The details of how we derived these measurements are given in the Materials and Methods section.

## A multi-step causal inference approach to assign directionality to signaling-immune associations

Our framework to assign directionality to signaling-immune associations is motivated by an established approach to order gene expression traits relative to one another and relative to other complex traits [10]. The key idea is to anchor the analysis on genomic variation and to systematically test whether DNA changes that lead to changes in signaling and immune activity support a causative, reactive or independent model of the interaction between signaling and immune activity.

We formalise different causal relationships in three different types of graphical models (Fig. 1A). Model 1 ( $M_1$ : the causative model) represents a case in which a genomic event changes immune activity by perturbing signaling activity. Model 2 ( $M_2$ : the reactive model) represents a case in which a genomic event leads to a change in immune activity, which then in turn changes signaling activity. Model 3 ( $M_3$ : the independence model) represents a case in which the genomic event influences immune activity and signaling activity independently of each other. We used standard assumptions of causal inference [18] to derive likelihood functions for each of the three models (see Materials and Methods). For each triplet consisting of one genomic locus, one transcription factor and one immune score, we maximise the likelihood function of the three models over their parameters, and finally choose the model with the smallest Akaike Information Criterion [19], a model selection criterion balancing goodness-of-fit with model complexity.

To limit the search space and reduce the number of triplets to test, we developed a multi-step causal model inference framework (CMIF) approach (Fig. 1). In a first step, CMIF filters only selects genes with experimentally verified protein-protein interaction (PPI) with a TF of interest (Fig. 1C). There are  $19,702 \times 788 = 15,525,176$  pairwise associations between copy number profiles and TF activities, and filtering them according to the StringDB database [20] results in just 2,333 potential models. This filtering substantially reduces the search space and enriches for biologically relevant drivers in groups of correlated genes that are jointly amplified or deleted.

In a second step, undirected skeletal association graphs are constructed for CNA events underlying both the TF activity and immune phenotype by computing pairwise correlation coefficients between the variables. Benjamini-Hochberg [21] p-value correction is applied and only skeletons where all pairwise associations are significant are passed to the final step (Fig. 1C). Finally, the likelihood function of each model consistent with these skeletons is maximised over its parameters and the model with the smallest Akaike Information Criterion is chosen. After filtering and model selection, CMIF provides as output the model of the causal relationship between the variables (Fig. 1D).

## Evaluating CMIF with CS/TCS immune metrics

In the first analysis we used the TCS and CS metrics derived from gene expression data as measurements of immune activity. Applying CMIF to the TCS/CS metrics, TF activities and copy number profiles resulted in 111 unique TFs that were significantly associated with either the TCS or CS measurements, and whose underlying CNA modulator correlates significantly with the immune phenotype. CNA events at loci corresponding to 176 genes significantly correlated with both the TF activity

and either immune phenotype, resulting in 475 undirected skeletal graphs. For each graph, we fit the likelihood models M1, M2 and M3 to the respective copy number profiles, TF activity and TCS/CS trait measurements (see Methods). 344 triplets (72%) were best represented by the causal model (M1) whereas the reactive model (M2) was the best in the remaining 131 (28%) cases. No M3 models were supported by the data, which can be explained by the degree to which the association and PPI filtering step of the method (Fig. 1C) identifies strictly causal or strictly reactive mechanisms.

*Validation in independent cohort* We validated these results in the independent TCGA cohort. Of the candidate triplets, 194 (54.6%) M1 and 24 (18.3%) M2 models validated in the TCGA cohort using CMIF (Fig. 2A). The higher validation percentage of M1 models over M2 models in TCGA data indicates that causal drivers of T cell infiltration are more robust and thus more frequently recapitulated in breast cancer populations. The higher prevalence of M1 over M2 models might be explained by cancer cells being immunoedited [22], a process in which somatic mutations break downstream pathways associated with a normal immune response. Over time, this would enable the tumour to exert more control over the immune system than vice-versa.

The correlation between TF activity and the individual immune traits agreed well between METABRIC and TCGA (TCS:  $\rho = 0.98$ ,  $P < 2.2 \times 10^{-16}$ ; CS:  $\rho = 0.984$ ,  $P < 2.2 \times 10^{-16}$ ) (Fig. 1B), highlighting robust co-dependence relationships between lymphocyte infiltration/activation and TF activity. Of the validated models, 118 were shared between the TCS and CS traits, with 47 unique to the TCS (165 total) and 53 unique to the CS (171 total). (Fig. 2C). This high degree of concordance is reassuring considering that many molecular pathways facilitating lymphocyte aggregation will also directly or indirectly influence the cytolytic activity and vice-versa.

*Validation by literature* Many of the top predictions generated by CMIF are well supported by the literature. For example, when ranked by the strength of their correlation with the immune phenotype, CMIF analysis highlighted *IRF1* as the strongest causal mediator of the TCS phenotype across both METABRIC and TCGA, whose TF activity is down-regulated by the amplification of *PIAS3* (Fig. 2D). This is consistent with reports that *PIAS3* induces transcriptional repression of *IRF1* through binding to it as a SUMO-1 ligase [23]. Furthermore, *IRF1* has been shown to play a crucial role in driving anti-tumour immune response [24] and thus this model's categorisation as causal for TCS is well supported by the strong body of literature surrounding the relationship between the variables.

*RUNX3*, a well known tumour suppressor gene [25], was identified as the second strongest causal modulator of the TCS. *RUNX3* activity has been found experimentally to mediate lymphocyte chemotaxis through the TGF-*B* pathway [26, 27]. The positive association found between *TAL1* and *RUNX3* has been also been confirmed in studies demonstrating that the *RUNX* genes are direct targets of *TAL1* [28]. Additionally, CMIF identified *ETS1* as a causal mediator for the TCS, which is unsurprising given that its activity has been shown to regulate the transcription of chemokines and cytokines directly involved in lymphocyte migration [29].

*Validation with image-derived features* Another way of validating the robustness of our model is testing how well it predicts lymphocyte infiltration in an orthogonal dataset. To facilitate this, we used the paired tumour whole tissue section slides stained with Haematoxylin & Eosin (H&E) provided by the METABRIC study [11], which provide an estimate of lymphocytic infiltration independent of the gene expression based estimates used above.

We evaluated the predictive utility of 165 validated TCS M1 models on an image cohort consisting of 534 samples. We used the results of previous image analyses to segment and quantify the absolute number of lymphocytes and the lymphocyte density [13], with further normalisation techniques applied in our study to generate traits from these features (See Methods).

We combined these image features, TF activities and copy number profiles in our CMIF approach and computed the overlap between the image-based causal models and those from the transcriptomic phenotype set. 18 (10.9%) of the initial predictors of the TCS were also predictive of the image features, with the majority (15/18) belonging to the lymphocyte density trait. The extent to which TF activity significantly correlated with both the TCS and the image lymphocyte density feature simultaneously was weaker ( $\rho = 0.45$ ,  $P < 2.2 \times 10^{-16}$ ) than that of the TCS between METABRIC and TCGA. This might be due to the fact that the transcriptomic T cell features are not necessarily perfect proxies to lymphocyte features extracted from images. For example, while the TCS makes measurements about T cells exclusively, the feature information in H&E images is not sufficiently descriptive to differentiate NK cells from T cells leading to a weaker correlation. Furthermore, there are confounding systematic errors that may have arisen during the segmentation and classification process used to generate the image features that render the correlation with the transcriptomic phenotypes weaker than expected.

For consistency, we only considered causal models that demonstrated the same directionality of association between the TF activity and both the TCS feature and image feature. The image validated model list was comprised of 12 triplets, revealing 11 unique DNA loci exerting influence over lymphocyte infiltration through activity perturbation to 8 TFs (Table 1). Notably, 8 out of the top 10 strongest causal models for the TCS phenotype (as ranked by association with TF activity) validated for the lymphocyte density trait, highlighting the excellent predictive potential of our models in the image cohort. The remaining 2 triplets negatively associated with the trait and demonstrated a stronger predictive power in the images.

Of our validated models, notable examples include a process by which *PIAS3* copy number aberration was found to attenuate the TCS/lymphocyte density through *TFEB* and *IRF1* activity repression, both of which positively associate with the mentioned traits (Table 1). *CREBBP* and *EP300* were found to exert similar causal pressure on the traits through their action on the TF *ETS1*. This follows from experimental evidence demonstrating that *CREBBP* and *EP300* form a protein complex *CBP/p300* that is recruited by *ETS1* to facilitate its transcription factor functionality [30].

Among the top 8 candidate TFs (Table 1) we identified *NR3C1*: this gene encodes for the glucocorticoid receptor, and influences immune activity through inflammation [31]. This TF was ranked only 95th of 510 in the list of image associations and

27th of 510 in the TCS associations. Its function as a driver of immune infiltration was only elucidated once the causal relationships between genome, signaling and immune phenotypes were modeled together, highlighting the advantage of CMIF over standard pairwise-association approaches.

#### Causal model case studies and mechanisms

Our results provide several specific biological examples of causal models of the interaction between cancer signaling and immune activity.

*EP300 and NCOR1 modulate cytolytic activity through ETS1/SPI1/TP53 network perturbation* Copy number amplification of *EP300* and *NCOR1* were found to underlie the cytolytic activity trait in both the METABRIC and TCGA cohorts. Interestingly, the original study by Rooney and colleagues [17] found that single nucleotide variants in these genes correlated positively with cytolytic activity in cancer types other than breast. The CMI's ability to elucidate these mechanisms in breast cancer may be due the higher prevalence of CNA mutations over SNPs in the disease [32]. Furthermore, our method extends the understanding of the association between these DNA-level drivers and the cytolytic score by suggesting they act through perturbations to the activity of *ETS1*, *SPI1* and *TP53*.

Our discovery of a positive association between *SPI1* activity and the CS ( $\rho = 0.7$ ) is consistent with studies demonstrating that *SPI1* transcribes *CCL5*, a key player in cytolytic activity [33]. Similarly, *ETS1* deletion in mice has been linked to decreased cytolytic activity in NK cells [34], consistent with our observed positive correlation ( $\rho = 0.58$ ). The association between TP53 activity and cytolysis is not well characterised, although it has been shown that mutant TP53 attenuates cytolytic activity in ovarian and other cancers [17].

The direct correlations between cytolytic activity and *EP300* ( $\rho = 0.141$ ) and *NCOR1* ( $\rho = 0.08$ ) are weak, and the additional causal context provided by CMIF was needed to highlight *EP300* and *NCOR1* amplification as drivers of cytolytic activity in breast cancer through TF network perturbation.

*TF drivers of immune localisation regulate adaptive immune pathways* Functional annotation of TF transcriptional targets can elucidate which molecular pathways are over- or underrepresented in the presence of an immune phenotype. We investigated this by partitioning our model set into those where the activity of a TF positively associated with lymphocyte infiltration and those that were negatively associated. We then aggregated the transcription targets of each TF and applied GO term enrichment to functionally annotate the gene sets (Fig. 3A).

For TFs positively associated with lymphocyte recruitment, the term "T cell activation" and "adaptive immune response" (full term list in Additional file 1) were among the top associated pathways (adjusted  $P = 1.7 \times 10^{-33}$  and  $2.4 \times 10^{-28}$  respectfully). Interestingly, "antigen processing and presentation" was also ranked highly on the list (adjusted  $P = 1.3 \times 10^{-14}$ ) highlighting the importance of comprehensive antigen recognition in facilitating lymphocyte recruitment. All positively associative TFs in our model space demonstrated target overlap with one another (Fig. 3B(i)). TFs negatively associated with lymphocyte recruitment displayed disjoint sets of target genes (Fig. 3B(ii)) whose aggregate functional annotation was predominantly associated with pathways involved in innate immune cell regulation.

*Systems driving lymphocyte recruitment stratify by ER status* Differences in magnitude and prognostic relevance of lymphocytic infiltration between ER stratified breast cancer patients have been widely observed [35, 36, 7], but little is known regarding the causal chain of events that gives rise to this discrepancy. We hypothesised that suppressed TF activity in ER+ samples could suggest a possible mechanism by which ER+ tumours evade immune destruction. To investigate this, we used the clinical data for the METABRIC cohort samples to investigate whether genomic drivers of lymphocytic recruitment stratify by ER status.

The copy number profiles of all genomic drivers in our list of models (Table 1) significantly stratified by ER status with the exception of HAX1 (Fig. 4A). For example, genes such as *PIAS3*, *POU2F1* and *CREBBP* were significantly more amplified in ER+ patients over ER-. These genes downregulate TFs positively associated with lymphocytic activity such as *TFEB*, *IRF1*, *NR3C1* and *ETS1* leading to significantly less activity relative to ER- samples (Fig. 4B). Further to this, *TAL1* and *CBFB* were both significantly more amplified in ER- samples over ER+, and subsequently, significantly higher activity observed in the activity of the respective TFs they modulate in ER- samples over ER+. Significantly lower cytolytic activity was also observed in ER+ samples relative to ER- (see Additional file 1), which is unsurprising given that transcriptional targets of TFs regulating the TCS were shown to modulate T cell activation (Fig. 3B).

The stratification of these causal events provides compelling evidence of a genomic basis for the ER stratification of lymphocyte infiltration and activity. These results are difficult to infer from association studies alone, highlighting a chief advantage of a deriving causal frameworks from large datasets.

## Discussion & Conclusions

The aim of our study was to dissect interactions between cancer cells and their microenvironment. To achieve this aim, we developed a multistep methodology to inferring directed relationships between signaling activity and immune infiltration in the tumour microenvironment that overcomes limitations of conventional association studies by anchoring the analysis on somatic genomic events.

Our approach uses established methods to estimate TF activity and lymphocyte infiltration (CS) from gene expression data, and proposed a novel score for lymphocytic activity (TCS). Since genes tend to be amplified/deleted together, we use a PPI network as a biological prior to isolate drivers from a list of genes correlated with an immune trait. Causal inference is achieved using a likelihood test, which returns the most likely relationship between the genotype, TF activity and phenotype given the data. Association based methods are widely used in cancer research and our methodology is a step towards a causal and mechanistic understanding of these relationships.

Our analysis consisted of identifying models for the TCS/CS traits in a discovery cohort (METABRIC [11]), validating them in a large independent cohort (TCGA [12]) and further evaluating their predictive utility using orthogonal measures of lymphocyte infiltration from H&E images. The final model list revealed 11 driver genes modulating lymphocyte recruitment into the microenvironment through perturbation to the activity of 8 TFs. Whilst most TFs in our models

have been experimentally linked to lymphocyte infiltration, the majority of driver genes we found are novel, highlighting a principal advantage of causal driver discovery over standard association studies. This was further realized with the discovery of *EP300/NCOR1* copy number alterations as drivers of cytolytic activity, whereas SNV mutations in these genes were previously found not to correlate with the trait in breast cancer. Drivers of lymphocytic infiltration were found to stratify by ER status, leading to significant stratification of activity profiles of TFs found to be causal for lymphocyte infiltration. This observation provides evidence supporting a genomic basis for the observed stratification of lymphocyte infiltration and prognostic utility by ER status.

Our approach has several limitations, many of which are technical and relate to the assumptions we had to make for statistical modelling. One such limitation involves measurement errors within the individual data inputs to our integrative analysis. If the margin of error for one variable is wider than that of another, it could potentially lead to the misclassification of the causal-reactive relationship between the two variables. Another limitation arises from the simplicity of the DAG models we design for the interaction between two variables. TFs causal for a trait will regulate genes that are interacting within the context of a much larger network and with feedback controls that need to be accounted for. Although our models successfully predict lymphocyte infiltration in image cohorts, a stronger validation would involve knockdown experiments in mice to directly observe changes in TF profile and our trait of interest. While our methodology is generally applicable, the details of the statistical model will have to be tailored to the specific type of data used in the study, which might not all be normally distributed.

A conceptual limitation is that the whole study is based on one major assumption: genomic events in the cell-autonomous compartment drive the development of the cancer and can thus be used as anchors for causal analysis. This assumption is shared by almost all cancer genomics studies, in particular those that aim to identify genomic drivers of the disease [11, 32]. At the same time, we acknowledge that there could be situations in which microenvironmental changes like inflammation cause genomic events, rather than being caused by them.

Despite these limitations, our method has demonstrated power to recapitulate known mechanisms and has shown results that stay robust when using independent data sets and orthogonal data types. Given our method's high validation rate between two large and independent datasets and its capacity to predict results supported by the literature, we believe it to be a robust predictor of cancer-immune communication mechanisms in transcriptomic data.

Our analysis was focused on breast cancer, but large efforts like The Cancer Genome Atlas (TCGA) or the International Cancer Genome Consortium (ICGC) provide the same types of data for many other kinds of cancer and thus our methodological framework can be easily applied to many other forms of the disease. Additionally, the framework is not confined to the CS/TCS metrics as measures of immune activity and can be applied to any other available feature of the microenvironment. Thus, in summary, we have presented an integrative analysis of genomic events, signaling activity and immune markers, which is flexible and can form the foundation for a more mechanistic understanding of tumour-microenvironment interactions across cancer types.

## Materials and Methods

### CMIF Step 1: Data preparation

*Gene Expression Data* Microarray transcriptomic profiles corresponding to 1980 patients from the METABRIC cohort were downloaded from the European Genome-Phenome archive under the accession id: EGAD00010000268 (<https://ega-archive.org/datasets/EGAD00010000268>). The issue of multiple probes mapping to the same gene was addressed by selecting the probe with the highest variance. RNA-seq count data comprising 1154 BRCA samples was downloaded from the TCGA archive (<https://tcga-data.nci.nih.gov/tcga>) and processed using a two-step process: applying the variance stabilising transform and quantile normalising the matrix with respect to the METABRIC gene expression distribution. This was done to correct for the large heteroscedasticity between genes and make the expression distributions more comparable.

*DNA copy number aberrations* METABRIC Affymetrix SNP 6.0 data were downloaded from the same resource as the transcriptomic data. SNP array genomic positions were mapped to gene symbols using the hg18 build. TCGA GISTIC2 gene-level, zero-centered, focal copy number calls for each patient were accessed from GDAC Firehose (<http://gdac.broadinstitute.org/>).

*Transcription Factor Network Enrichment* To infer TF activity, we used a co-expression network [37] for 788 experimentally verified TFs derived using mutual information and used this to calculate the activity of each TF for each sample using the R package viper (virtual inference of protein activity by enriched regulon analysis) [16] using default parameters. VIPER tests the activity of each TF by examining the relative transcript abundance of known targets genes, collectively referred to as a "regulon". An activity score for each TF is computed using an enrichment analysis method that includes a probabilistic weighting using TF-gene association likelihoods.

*Immune trait inference from transcriptomic data* To gauge the degree of T cell infiltration in transcriptomic data, we introduce the T cell score (TCS) as the geometric mean of *CD3D*, *CD4*, *CD8A* and *CD8B* expression. In addition to being well established T cell markers, they are expressed primarily in cells of a haematopoietic lineage with little noise contamination from the rest of the tumour. To measure cytolytic activity from transcriptomic data, we define the cytolytic score (CS) as the geometric mean of *GZMA* and *PRF1* as per work done by Rooney and colleagues [17]. In both cases, the geometric mean is chosen over the arithmetic mean given its reduced sensitivity to outliers.

*Protein-protein interactions* A network detailing protein-protein interactions was downloaded from the STRINGdb resource (<http://string-db.org/>) and ENSEMBL identifiers were mapped to HUGO gene symbols using the R biomaRt package.

### CMIF Step 2: Triplet Initialisation

Here we describe our approach for deriving pairwise associations between CNA profiles, TF activity and immune phenotypes. These triplet skeletal graphs are the input to the model scoring step of the CMIF.

*Significance of associations* All associations in this manuscript are computed using the Pearson correlation coefficient and  $p$ -values are calculated using Student's test unless explicitly stated otherwise.

*Computing undirected triplet graphs* Undirected skeletal graphs were constructed by computing pairwise associations between the CNA profiles, TF activity and the phenotype of interest. To reduce the number of hypotheses to test, we only considered pairs of TFs and CNAs, if the corresponding proteins showed protein-protein interaction.

METABRIC was taken as our discovery cohort and thus association  $P$ -values were adjusted using the Benjamini–Hochberg (BH) procedure [21] and a conservative significance threshold defined at  $P \leq 10^{-3}$ . Findings were said to be validated if reproduced at unadjusted  $P \leq 0.05$  in the TCGA cohort.

Skeletal graphs between triplets of variables were passed to the model scoring step if all three pairwise associations in the graph demonstrated significance.

### CMIF Step 3: Model scoring

*Likelihood Function Definitions* Our approach uses Bayesian networks and likelihood models to determine which relationship between the variables is best supported by the data. Assuming the conditional probability distribution of the future state of the system depends only on the present state, we can write our joint probability distribution of models  $M_1$ ,  $M_2$  and  $M_3$  as

$$M_1. \quad P(X, Y, Z) = P(X)P(Y|X)P(Z|Y), \quad (1)$$

$$M_2. \quad P(X, Y, Z) = P(X)P(Z|X)P(Y|Z), \quad (2)$$

$$M_3. \quad P(X, Y, Z) = P(X)P(Z|X)P(Z|Y, X), \quad (3)$$

where  $X$ ,  $Y$  and  $Z$  correspond to the copy number profile, transcription factor activity and immune phenotype measurements respectively. We assume that  $Y$  and  $Z$  are normally distributed with a constant variance such that the likelihood of each model can be described by multivariate Gaussian density functions. Individual component likelihoods are summed over all copy number states ( $J = \{Amplified, Deleted, Neutral\}$ ) and the likelihood of each joint distribution given the parameterisation is computed by multiplying over the entire sample space  $N$  as such:

$$L(\theta_{M_1}; M_1) = \prod_{i=1}^N \sum_{j \in J} P(X_j)P(Y_i|X_j)P(Z_i|Y_i), \quad (4)$$

$$L(\theta_{M_2}; M_2) = \prod_{i=1}^N \sum_{j \in J} P(X_j)P(X_j|Y_i)P(Y_i|Z_i), \quad (5)$$

$$L(\theta_{M_3}; M_3) = \prod_{i=1}^N \sum_{j \in J} P(X_j)P(Y_i|X_j)P(Z_i|Y_i, X_j). \quad (6)$$

Individual component definitions are described in the Additional file 1.

*Maximum Likelihood Estimation and Model Selection* The joint distribution likelihoods are maximised over the parameter space using the maximum likelihood estimation algorithm as implemented by the *optim* function in R. We then compute the Akaike information criterion (AIC) as

$$AIC_j = -2 * \log(L(\theta_{M(j)}; M(j))) * |\theta_{M(j)}|, \forall j \in J. \quad (7)$$

The model with the smallest AIC is the strongest causal candidate as given by the data.

### Validation Approaches

The robustness of our analysis is evaluated with several validation strategies that use orthogonal datasets such as H&E images and flow cytometry data.

*T Cell Score Validation using Flow Cytometry* Gene expression profiles for 20 peripheral blood mononuclear cell admixture samples and their corresponding flow cytometry profiles as measured by Newman *et al* [9] were downloaded from (<http://cibersort.stanford.edu>). Pearson's test was then used to infer correlations between flow cytometry measurements for 9 subsets of leukocytes and the TCS measurements for all 20 samples.

*H&E Section Validation* We made use of the image dataset published by Yuan and colleagues [13], comprising the segmented H&E stained primary tissue sections of 564 patients sampled from the METABRIC cohort. Segmented objects were classified using a SVM trained by an expert pathologist and metrics pertaining to the absolute number of lymphocytes and the lymphocyte density relative to the number of overall objects were measured. The absolute number of lymphocytes were then log-transformed to enable more robust comparison across the patient space. Furthermore, this transformation ensures input to the MLE process exhibits a similar range of values and ensure that the algorithm can be initialised with the same parameters for all cohorts. Finally, CMIF was run with the lymphocyte statistics, their paired copy number profiles and TF activities.

### Additional analyses

*GO Term Enrichment Analysis* GO term enrichment analysis is used to characterise groups of TFs causal for our observed phenotypes. Causal TF regulators of immune infiltration were split into two groups given by the directionality of their association with the TCS phenotype. Their regulons were aggregated and GO term enrichment was performed using the *clusterProfiler* package in R [38]. Further details can be found in the Supplementary Statistical Analysis (Additional file 1).

### Code & Data Availability

The R script [14] implementation for the CMIF is provided in the Supplementary R Sweave file (Additional file 1). Additional file 1 can also be used to reproduce all figures and results of the described case studies.

Data accession details are provided in the Materials and Methods section. Furthermore, the datasets used can be downloaded directly using either code available in the R Sweave file (Additional file 1) or from the github repository <https://github.com/databro/Chlon2017>.

## Ethics approval

Ethics approval was not needed for this study.

## Acknowledgements

We thank Dr. Federico Giorgi (University of Cambridge) for fruitful discussions and project reproducibility testing.

## Competing interests

The authors declare that they have no competing interests.

## Author's contributions

LC contributed to project conceptualisation, data curation, formal analysis, investigation, methodology, validation and visualisation. LC & FM contributed to writing, review and editing. FM contributed to funding acquisition, project administration and supervision.

## Funding

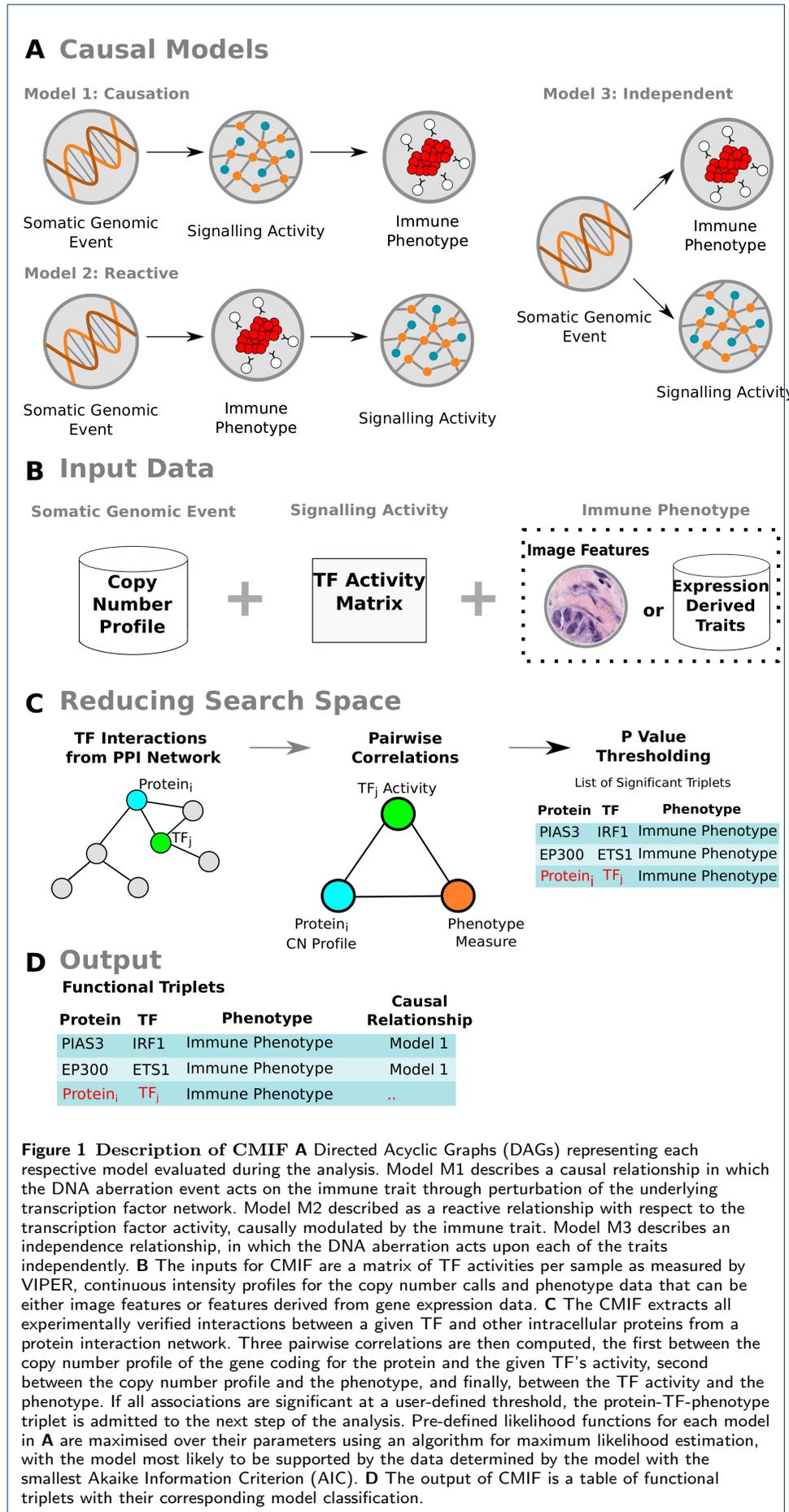
This work was funded by the Cancer Research UK and Engineering and Physical Sciences Research Council Imaging Centre in Cambridge and Manchester grant (C197/A16465).

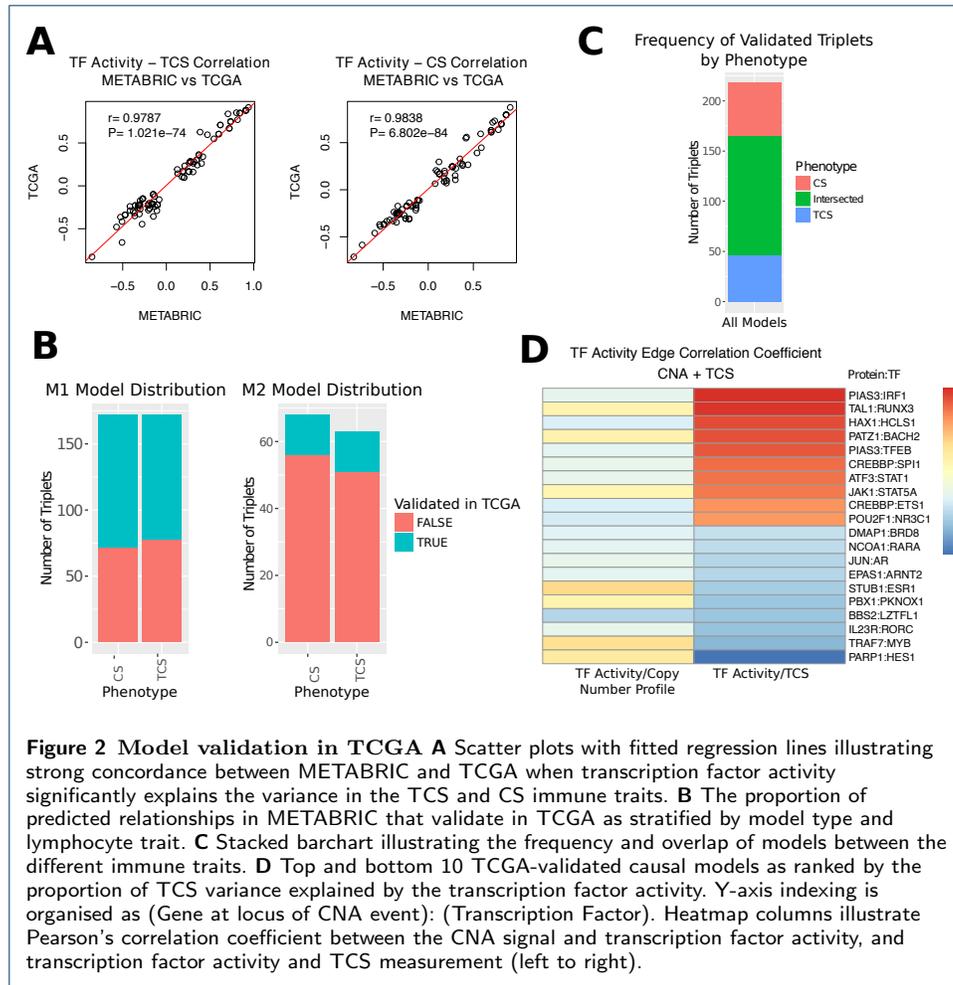
## References

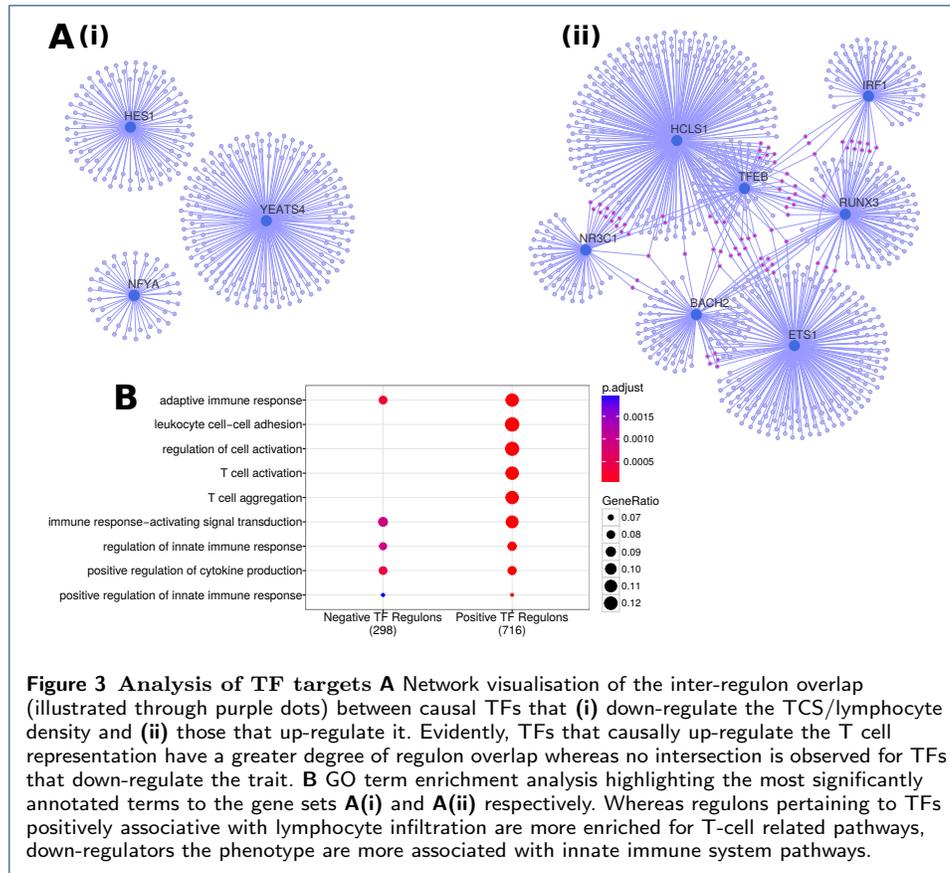
1. Nowell, P.C.: The clonal evolution of tumor cell populations. *Science* **194**(4260), 23–28 (1976)
2. Balkwill, F.R., Capasso, M., Hagemann, T.: The tumor microenvironment at a glance. *J. Cell Sci.* **125**(Pt 23), 5591–5596 (2012)
3. Mbeunkui, F., Johann, D.J. Jr: Cancer and the tumor microenvironment: a review of an essential relationship. *Cancer Chemother. Pharmacol.* **63**(4), 571–582 (2009)
4. Sang, L., Roberts, J.M., Collier, H.A.: Hijacking HES1: how tumors co-opt the anti-differentiation strategies of quiescent cells. *Trends Mol. Med.* **16**(1), 17–26 (2010)
5. Rizzo, P., Osipo, C., Foreman, K., Golde, T., Osborne, B., Miele, L.: Rational targeting of notch signaling in cancer. *Oncogene* **27**(38), 5124–5131 (2008)
6. Bindea, G., Mlecnik, B., Tosolini, M., Kirilovsky, A., Waldner, M., Obenauf, A.C., Angell, H., Fredriksen, T., Lafontaine, L., Berger, A., Bruneval, P., Fridman, W.H., Becker, C., Pagès, F., Speicher, M.R., Trajanoski, Z., Galon, J.: Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* **39**(4), 782–795 (2013)
7. Ali, H.R., Chlon, L., Pharoah, P.D.P., Markowetz, F., Caldas, C.: Patterns of immune infiltration in breast cancer and their clinical implications: A Gene-Expression-Based retrospective study. *PLoS Med.* **13**(12), 1002194 (2016)
8. Oh, E.-Y., Christensen, S.M., Ghanta, S., Jeong, J.C., Bucur, O., Glass, B., Montaser-Kouhsari, L., Knoblauch, N.W., Bertos, N., Saleh, S.M., Haibe-Kains, B., Park, M., Beck, A.H.: Extensive rewiring of epithelial-stromal co-expression networks in breast cancer. *Genome Biol.* **16**, 128 (2015)
9. Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., Alizadeh, A.A.: Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**(5), 453–457 (2015)
10. Schadt, E.E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S.K., Monks, S., Reitman, M., Zhang, C., Lum, P.Y., Leonardson, A., Thieringer, R., Metzger, J.M., Yang, L., Castle, J., Zhu, H., Kash, S.F., Drake, T.A., Sachs, A., Lusk, A.J.: An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* **37**(7), 710–717 (2005)
11. Curtis, C., Shah, S.P., Chin, S.-F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., METABRIC Group, Langerød, A., Green, A., Provenzano, E., Wishart, G., Pinder, S., Watson, P., Markowetz, F., Murphy, L., Ellis, I., Purushotham, A., Børresen-Dale, A.-L., Brenton, J.D., Tavaré, S., Caldas, C., Aparicio, S.: The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**(7403), 346–352 (2012)
12. Cancer Genome Atlas Network: Comprehensive molecular portraits of human breast tumours. *Nature* **490**(7418), 61–70 (2012)
13. Yuan, Y., Failmezger, H., Rueda, O.M., Ali, H.R., Gräf, S., Chin, S.-F., Schwarz, R.F., Curtis, C., Dunning, M.J., Bardwell, H., Johnson, N., Doyle, S., Turashvili, G., Provenzano, E., Aparicio, S., Caldas, C., Markowetz, F.: Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Sci. Transl. Med.* **4**(157), 157–143 (2012)
14. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2008). R Foundation for Statistical Computing. ISBN 3-900051-07-0. <http://www.R-project.org>
15. Carro, M.S., Lim, W.K., Alvarez, M.J., Bollo, R.J., Zhao, X., Snyder, E.Y., Sulman, E.P., Anne, S.L., Doetsch, F., Colman, H., et al.: The transcriptional network for mesenchymal transformation of brain tumours. *U.S. National Library of Medicine* (2010). <https://www.ncbi.nlm.nih.gov/pubmed/20032975>

16. Alvarez, M.J., Shen, Y., Giorgi, F.M., Lachmann, A., Ding, B.B., Ye, B.H., Califano, A.: Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* **48**(8), 838–847 (2016)
17. Rooney, M.S., Shukla, S.A., Wu, C.J., Getz, G., Hacohen, N.: Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160**(1-2), 48–61 (2015)
18. Pearl, J.: *Causality: Models, Reasoning, and Inference*. Cambridge University Press, ??? (2000)
19. Akaike, H.: A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **19**(6), 716–723 (1974)
20. Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., Kuhn, M., Bork, P., Jensen, L.J., von Mering, C.: STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**(Database issue), 447–52 (2015)
21. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**(1), 289–300 (1995)
22. Dunn, G.P., Bruce, A.T., Ikeda, H., Old, L.J., Schreiber, R.D.: Emerging landscape of oncogenic signatures across human cancers. *Nat. Immunol.* **3**(11), 991–998 (2002)
23. Nakagawa, K., Yokosawa, H.: PIAS3 induces SUMO-1 modification and transcriptional repression of IRF-1. *FEBS Lett.* **530**(1-3), 204–208 (2002)
24. Tamura, T., Yanai, H., Savitsky, D., Taniguchi, T.: The IRF family transcription factors in immunity and oncogenesis. *Annu. Rev. Immunol.* **26**, 535–584 (2008)
25. Bae, S.-C., Choi, J.-K.: Tumor suppressor activity of RUNX3. *Oncogene* **23**(24), 4336–4340 (2004)
26. Yano, T., Ito, K., Fukamachi, H., Chi, X.-Z., Wee, H.-J., Inoue, K.-I., Ida, H., Bouillet, P., Strasser, A., Bae, S.-C., Ito, Y.: The RUNX3 tumor suppressor upregulates bim in gastric epithelial cells undergoing transforming growth factor $\beta$ -induced apoptosis. *Mol. Cell. Biol.* **26**(12), 4474–4488 (2006)
27. Li, M.O., Wan, Y.Y., Sanjabi, S., Robertson, A.-K.L., Flavell, R.A.: Transforming growth factor-beta regulation of immune responses. *Annu. Rev. Immunol.* **24**, 99–146 (2006)
28. Landry, J.-R., Kinston, S., Knezevic, K., de Bruijn, M.F.T.R., Wilson, N., Nottingham, W.T., Peitz, M., Edenhofer, F., Pimanda, J.E., Ottersbach, K., Götting, B.: Runx genes are direct targets of Scl/Tal1 in the yolk sac and fetal liver. *Blood* **111**(6), 3005–3014 (2008)
29. Russell, L., Garrett-Sinha, L.A.: Transcription factor ets-1 in cytokine and chemokine gene regulation. *Cytokine* **51**(3), 217–226 (2010)
30. Foulds, C.E., Nelson, M.L., Blaszczyk, A.G., Graves, B.J.: Ras/mitogen-activated protein kinase signaling activates ets-1 and ets-2 by CBP/p300 recruitment. *Mol. Cell. Biol.* **24**(24), 10954–10964 (2004)
31. Pan, D., Kocherginsky, M., Conzen, S.D.: Activation of the glucocorticoid receptor is associated with poor prognosis in estrogen receptor-negative breast cancer. *Cancer Res.* **71**(20), 6360–6370 (2011)
32. Ciriello, G., Miller, M.L., Aksoy, B.A., Senbabaoglu, Y., Schultz, N., Sander, C.: Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45**(10), 1127–1133 (2013)
33. Kumar, D., Hosse, J., Toerne, C.V., Noessner, E., Nelson, P.J.: Jnk mapk pathway regulates constitutive transcription of ccl5 by human nk cells through sp1. *The Journal of Immunology* **182**(2), 1011–1020 (2009). doi:10.4049/jimmunol.182.2.1011
34. Barton, K., Muthusamy, N., Fischer, C., Ting, C.-N., Walunas, T.L., Lanier, L.L., Leiden, J.M.: The ets-1 transcription factor is required for the development of natural killer cells in mice. *Immunity* **9**(4), 555–563 (1998). doi:10.1016/s1074-7613(00)80638-x
35. Liu, S., Lachapelle, J., Leung, S., Gao, D., Foulkes, W.D., Nielsen, T.O.: CD8+ lymphocyte infiltration is an independent favorable prognostic indicator in basal-like breast cancer. *Breast Cancer Res.* **14**(2), 48 (2012)
36. Mao, Y., Qu, Q., Chen, X., Huang, O., Wu, J., Shen, K.: The prognostic value of Tumor-Infiltrating lymphocytes in breast cancer: A systematic review and Meta-Analysis. *PLoS One* **11**(4), 0152500 (2016)
37. Fletcher, M.N.C., Castro, M.A.A., Wang, X., de Santiago, I., O'Reilly, M., Chin, S.-F., Rueda, O.M., Caldas, C., Ponder, B.A.J., Markowetz, F., Meyer, K.B.: Master regulators of FGFR2 signalling and breast cancer risk. *Nat. Commun.* **4**, 2464 (2013)
38. Yu, G., Wang, L.-G., Han, Y., He, Q.-Y.: clusterprofiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**(5), 284–287 (2012)

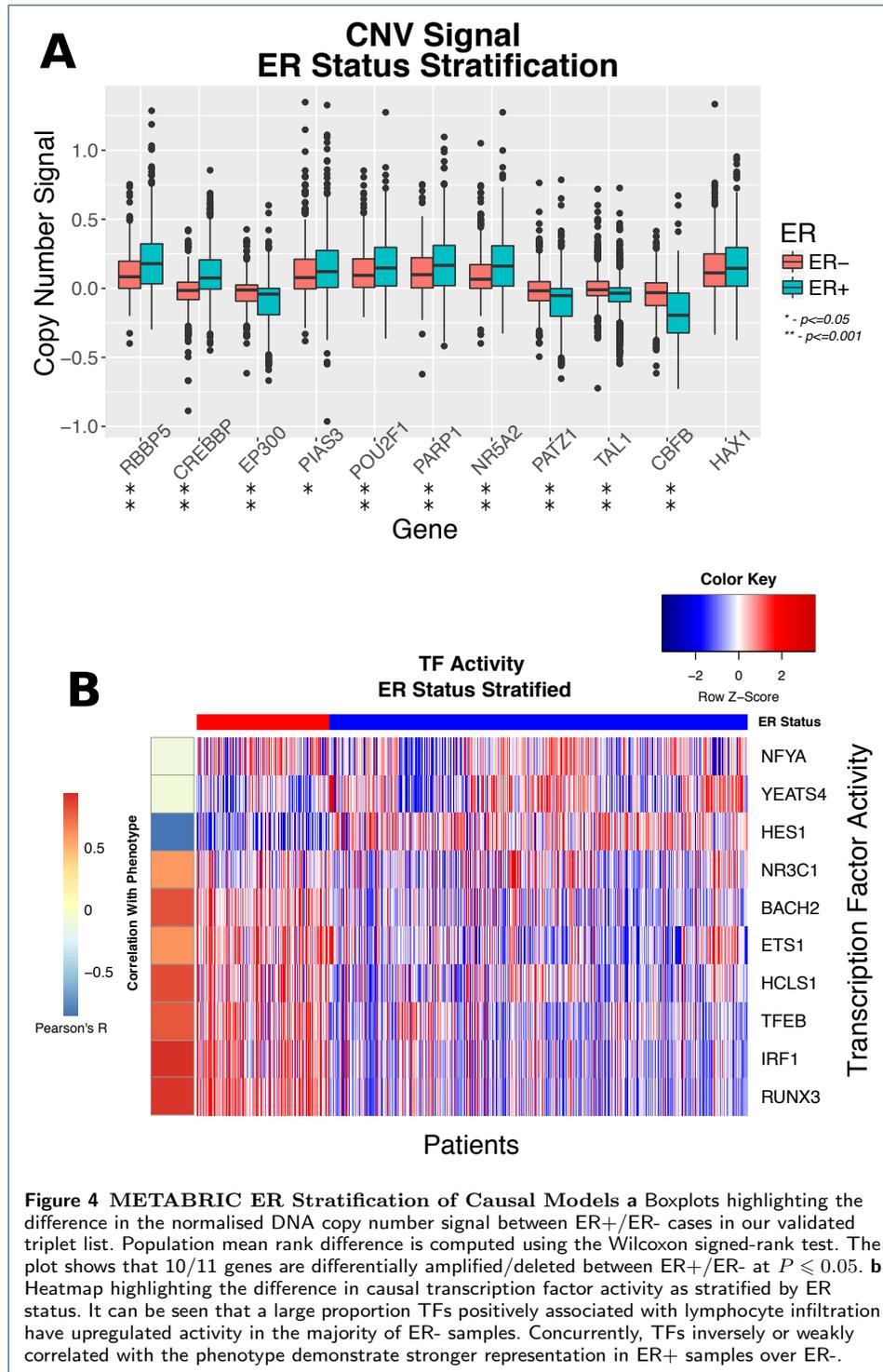
## Figures







**Figure 3 Analysis of TF targets** **A** Network visualisation of the inter-regulon overlap (illustrated through purple dots) between causal TFs that **(i)** down-regulate the TCS/lymphocyte density and **(ii)** those that up-regulate it. Evidently, TFs that causally up-regulate the T cell representation have a greater degree of regulon overlap whereas no intersection is observed for TFs that down-regulate the trait. **B** GO term enrichment analysis highlighting the most significantly annotated terms to the gene sets **A(i)** and **A(ii)** respectively. Whereas regulons pertaining to TFs positively associative with lymphocyte infiltration are more enriched for T-cell related pathways, down-regulators the phenotype are more associated with innate immune system pathways.



Tables

CNA	TF	TF-TCS Correlation	TF-TCS P value	Model	Type
RBBP5	YEATS4	-0.10	0.00	1	TCS
CREBBP	ETS1	0.62	0.00	1	TCS
EP300	ETS1	0.62	0.00	1	TCS
PIAS3	IRF1	0.94	0.00	1	TCS
PIAS3	TFEB	0.80	0.00	1	TCS
POU2F1	NR3C1	0.60	0.00	1	TCS
PARP1	HES1	-0.85	0.00	1	TCS
NR5A2	NFYA	-0.08	0.00	1	TCS
PATZ1	BACH2	0.81	0.00	1	TCS
TAL1	RUNX3	0.91	0.00	1	TCS
CBFB	RUNX3	0.91	0.00	1	TCS
HAX1	HCLS1	0.84	0.00	1	TCS

**Table 1** CMIF output of genomic drivers and TF perturbations causal for the TCS trait that also predict lymphocyte density in stained tumour sections.

**Additional Files**

Additional file 1

Zipped file containing a set of R code, results, and datasets to fully reproduce the results and analysis outlined in this manuscript.