Bioinformatics, 2017, 1–8
doi: 10.1093/bioinformatics/btx562
Advance Access Publication Date: 8 September 2017
Original Paper

OXFORD

Systems biology

Integrating splice-isoform expression into genome-scale models characterizes breast cancer metabolism

Claudio Angione

Department of Computer Science and Information Systems, Teesside University, Middlesbrough, UK

Associate Editor: Oliver Stegle

Received on March 6, 2017; revised on July 29, 2017; editorial decision on September 3, 2017; accepted on September 6, 2017

Abstract

Motivation: Despite being often perceived as the main contributors to cell fate and physiology, genes alone cannot predict cellular phenotype. During the process of gene expression, 95% of human genes can code for multiple proteins due to alternative splicing. While most splice variants of a gene carry the same function, variants within some key genes can have remarkably different roles. To bridge the gap between genotype and phenotype, condition- and tissue-specific models of metabolism have been constructed. However, current metabolic models only include information at the gene level. Consequently, as recently acknowledged by the scientific community, common situations where changes in splice-isoform expression levels alter the metabolic outcome cannot be modeled.
Results: We here propose GEMsplice, the first method for the incorporation of splice-isoform expression data into genome-scale metabolic models. Using GEMsplice, we make full use of RNA-Seq quantitative expression profiles to predict, for the first time, the effects of splice isoform-level changes in the metabolism of 1455 patients with 31 different breast cancer types. We validate GEMsplice by generating cancer-versus-normal predictions on metabolic pathways, and by comparing with gene-level approaches and available literature on pathways affected by breast cancer. GEMsplice is freely available for academic use at https://github.com/GEMsplice/GEMsplice_code. Compared to state-of-the-art methods, we anticipate that GEMsplice will enable for the first time computational analyses at transcript level with splice-isoform resolution.
Availability and implementation: https://github.com/GEMsplice/GEMsplice_code
Contact: c.angione@tees.ac.uk
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The increasing availability of multi-omic datasets has rapidly moved the focus of many research efforts from data collection to finding effective methods for data interpretation and analysis. In terms of biomedical results, this step is therefore currently considered more limiting than data collection itself (Stephens *et al.*, 2015).

Genes and their expression have been the subject of the vast majority of research studies in computational biology and bioinformatics. However, it is the interaction of genes, proteins, reactions and metabolites (different *omics*) that shapes the behavior of a cell. When analyzing biological models, the classical pathway-based perspective has been replaced, in the last 20 years, by a network-based approach, therefore bypassing parametrization and need for kinetic data, and leading to the generation of genome-scale metabolic models. These models include thousands of biochemical reactions, often the full set of reactions known for a given organism, allowing for prediction of cell phenotype. In this regard, metabolism is the only biological system that can be fully modeled at genome-scale, and also the closest to the phenotype.

Metabolism is now considered as a driver, rather than a marker, of cancer onset and proliferation. In breast cancer, metabolic heterogeneity is one of the causes of poor clinical outcome: although being classified as a single disease, more than 20 types of breast cancer

exist. A genome-scale analysis of cancer metabolism captures many effects that could not be identified using standard data and gene expression analysis. Furthermore, a wide set of tools for the incorporation of cancer omic data into these models have been developed, making them suitable for integrating and interpreting the great amount of data that is being gathered on cancer metabolic alterations (Qi *et al.*, 2017).

As mentioned above, reduced cost for data gathering and analysis has recently yielded a rapid increase in the amount of available omics data. RNA-Seq is now widely used to produce high-throughput data in more detail compared to microarrays, with a more accurate estimation of transcript levels. As a result, splice isoform expression levels are now becoming available in cancer studies, where the same gene can code for multiple proteins due to alternative RNA splicing before translation. Alterations of specific splice-isoform expression in some genes can constitute a biomarker for cancer metabolism. In humans, alternative splicing affects 95% of multiexon genes (Pan *et al.*, 2008).

However, to date, functional information included in most metabolic models refers to genes or proteins only. The idea exploited here is that splice isoform data can be readily integrated and used in conjunction with annotated genome-scale models in order to constrain and refine existing models. Although Recon1 (Duarte *et al.*, 2007) was the only human metabolic model that introduced isoform-level annotations for some genes through bibliographic research, it adopted a custom annotation for isoforms, and therefore did not allow any mapping to known identifiers from public databases. For these reasons, such annotations have been subsequently lost and then simply ignored, mapping transcriptomic data with gene-level resolution only (Ryu *et al.*, 2015). Consequently, expression data at the splice-isoform level has been neglected or simply averaged within the same gene to approximate the expression at the gene level.

Nevertheless, especially in human metabolic models, the incorporation of splice isoforms is key to understanding complex diseases like cancer. In this regard, pyruvate kinase (PK) is a striking example. In fact, the switch to the second isoform of pyruvate kinase (PKM2) is considered essential for cancer growth (Wong *et al.*, 2015). Conversely, the switch from PKM2 to the main pyruvate kinase isoform (PKM1) is able to reverse the Warburg effect and could therefore constitute a therapeutic target. In general, these crucial isoform-level events, e.g. the switch to minor isoforms, which get overexpressed compared to the major isoform of a gene, cannot be captured by current metabolic models, as highlighted in a number of recent reviews (Geng *et al.*, 2017; Pfau *et al.*, 2016; Ryu *et al.*, 2015; Yizhak *et al.*, 2015).

Here we propose GEMsplice (genome-scale metabolic modeling with splice-isoform integration), the first method for incorporating RNA-Seq data at the splice-isoform level into a metabolic model. The GEMsplice pipeline exploits, for the first time, the full potential of the next-generation sequencing technology in the context of genome-scale metabolic reconstructions. As a result, it enables more accurate predictions of human metabolic behavior and cancer metabolism. We show that GEMsplice compares favorably to existing tools for integration of transcriptomics data at the gene level only. We also validate GEMsplice by building breast cancer-versus-normal genome-scale models and by comparing predictions with available results on metabolic pathways affected by breast cancer. The full GEMsplice pipeline is illustrated in Figure 1. GEMsplice is made freely available in Matlab/Octave at https://github.com/GEMsplice/GEMsplice_code, and is fully compatible with the COBRA 2.0 Toolbox (Schellenberger *et al.*, 2011).

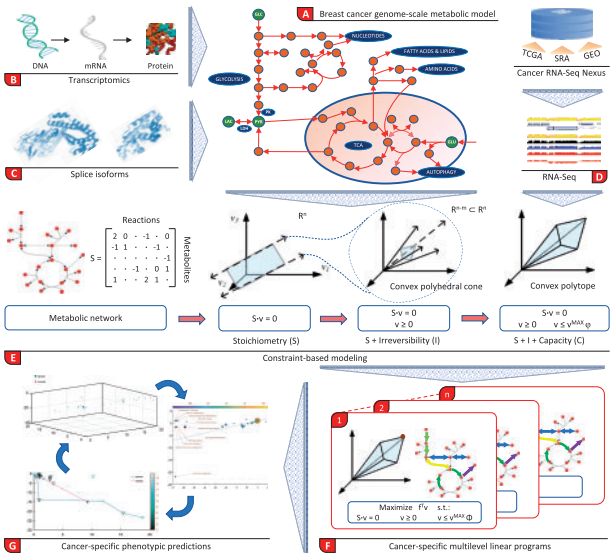


Fig. 1. GEMsplice incorporates RNA-Seq data into genome-scale metabolic models at the splice-isoform level. Starting from a metabolic model of breast cancer metabolism (Jerby *et al.*, 2012) (**A**), gene expression (**B**) and transcript level information (**C**) are incorporated into the model. As a result, for the first time, we can exploit the full potential of next-generation sequencing in the context of genome-scale metabolic reconstructions. A set of phenotype-specific RNA-Seq transcript expression levels in a variety of breast cancer types and stages from the Cancer RNA-Seq Nexus dataset (Li *et al.*, 2016), including data from TCGA, GEO and SRA (**D**), are then mapped onto the model using constraint-based modeling (**E**). Cancer-specific metabolic models are finally generated and investigated using multilevel linear programming (**F**), leading to phenotype prediction for different types of breast cancer (**G**)

1.1 Beyond the Warburg effect: metabolism is a key player in cancer formation and progression

Known molecular markers of cancer can be associated with two main classes: (i) cell proliferation and (ii) tissue remodeling (Markert *et al.*, 2015). Cancer cells show higher rate of proliferation than normal cells in the tissue from which they originated. The main goal of metabolism in cancer cells is to keep cell viability and ensure new biomass production by acquiring nutrients from an environment where nutrients are often scarce. To this end, cancer cells exhibit a modified metabolism and an increased need for proteins, energy, nucleotides and lipids. The two main nutrients used by cancer cells are glucose and glutamine; both support cell growth and are used to build carbon intermediates, which are employed in a number of processes that build macromolecules.

The first difference between normal and cancer metabolism was observed in the 1920s by Otto Warburg (Warburg *et al.*, 1927). Normal cells take up glucose and perform glycolysis to obtain pyruvate, which is then transferred to the mitochondrion and used by oxidative phosphorylation (OXPHOS) in the TCA cycle along with oxygen to efficiently produce ATP. Conversely, proliferating cancer cells require higher amount of glucose, but after obtaining pyruvate they preferentially use it to secrete lactate in the cytoplasm. Strikingly, this rather inefficient energy production pipeline is used also when the cell is exposed to oxygen. Preferential use of glycolysis allows faster proliferation but maximizes the secretion of lactic acid, which damages surrounding cells. This behavior, called *Warburg effect* or *aerobic glycolysis*, facilitates proliferation and migration of cancer cells.

However, a misinterpretation of this phenomenon by Warburg himself led to the enduring misconception that cancer cells do not use the TCA cycle to produce ATP. Indeed, it must be noted that the vast majority of cancer cells still use the mitochondrion and its TCA cycle to produce a fraction of the required ATP. In fact, in contrast to quiescent cells, glycolysis does not fuel directly the TCA cycle, but is essentially decoupled from it (Richardson *et al.*, 2008). The advantage of converting excess pyruvate into lactate instead of transferring it into the mitochondrion is that high glycolytic activity can continue without leading to excessive flux through the electron transport system and consequent overproduction of reactive oxygen species (ROS), or excess ATP and NADH generation, which would in turn repress glycolysis. As a result of this decoupling, useful glycolytic intermediates (e.g. precursors of serine biosynthesis) can be generated at higher rates.

Pyruvate kinase (PK), the last step in the glycolytic pathway, is responsible for keeping the balance between the production of pyruvate for the mitochondrion (high PK activity) and the production of glycolytic intermediates for biosynthesis (low PK activity). This is achieved through the preferential expression of PKM1 or PKM2, two splice isoforms of PK whose activity respectively supports pyruvate or glycolytic intermediates for biosynthetic processes. In cancer phenotypes, being mostly controlled by the isoform PKM2, PK and its ATP product are independent of oxygen, therefore enabling ATP generation and growth during hypoxia (unlike mitochondrial respiration, which needs oxygen).

Glucose is not the only key nutrient for cancer cells. Proliferating cells can consume up to ten times more glutamine than any other amino acid. Glutamine is used as a carbon source for fatty acid synthesis, in biosynthetic pathways (especially the biosynthesis of nucleotides), and for TCA cycle intermediates (although as a secondary source, less than glucose). While in normal cells glucose/glutamine intake depends mainly on extracellular stimuli, mutations undergone by cancer cells confer the ability to proliferate with a high degree of independence, by constantly importing amino acids and glucose from the extracellular environment (Yang *et al.*, 2017). Taken together, these studies show that metabolism, once believed to be mainly a passive indicator of the state of a cell, is now widely recognized as a key player in cancer formation and progression.

2 Materials and methods

2.1 Incorporating RNA-Seq data into a breast cancer model at the splice-isoform level

To model splice-isoform data in the breast cancer metabolic model, 31 RNA-Seq expression profiles of breast cancer were considered from Cancer RNA-Seq Nexus (Li *et al.*, 2016), which provides phenotype-specific transcript expression levels in a variety of cancer types and stages. In Cancer RNA-Seq Nexus, data were processed as follows. For GEO/SRA samples, Bowtie (Langmead *et al.*, 2009) was used to align RNA-Seq reads to the reference human transcriptome from GENCODE (Harrow *et al.*, 2012); eXpress (Roberts *et al.*, 2013) was then used to obtain FPKM isoform abundances. For TCGA samples, starting from Level 3 RNA-Seq v2 'tau values' quantified through RSEM (Li *et al.*, 2011), TPM values were obtained through multiplication by 10^6.

The expression values, at this point, are measured in FPKM for GEO/SRA, and in TPM for TCGA samples. FPKMs were then converted to TPMs using the following formula for all the expression values j in each sample i: $\mathrm{TPM}_{ij} = 10^6 \frac{\mathrm{FPKM}_{ij}}{\lambda_i \sum_j \mathrm{FPKM}_{ij}}$, where $\sum_j \mathrm{FPKM}_{ij}$ is the sum of the FPKM values of all the transcripts in the ith sample, and λ_i a sample-specific scaling factor accounting for the fact that the number of transcripts varies across the different samples, $\lambda_i =$ (number of transcripts in the largest sample)/(number of transcripts in sample i). We remark that TPM (and not FPKM) expression values were used as a starting point to generate the metabolic outcome across samples because they are proportional to the abundance and independent of the mean expressed transcript length, therefore making them more suitable to comparisons across samples (Li *et al.*, 2011).

Transcript annotations in Cancer RNA-Seq Nexus consist of Ensembl and UCSC IDs. These were converted to RefSeq annotations using BioMart (Smedley *et al.*, 2015). However, the breast cancer model is annotated only with gene-level Entrez IDs. To expand these and generate transcript-level IDs with splice-isoform resolution, RefSeq IDs for 141 transcripts were retrieved from the SBML source of Recon1 using a custom script, as they were not included in the final Matlab version of the model (see Supplementary File S1). These newly generated transcript-level RefSeq IDs and the gene-level default Entrez IDs of the remaining genes in the model were used to associate each transcript in the model with the corresponding expression value in the 31 RNA-Seq cancer profiles. The profiles were finally mapped onto the breast cancer metabolic model (see Fig. 1, and the following subsection for details on how the mapping was achieved).

2.2 Phenotype predictions from RNA-Seq data

Breast cancer is a multifactorial disease and, as a result, breast cancer cells are intrinsically multi-target. To model this behavior and generate cancer-specific models from Cancer RNA-Seq Nexus, an extended flux balance analysis (FBA) framework was used, coupled with a multi-omic integration method and multi-level linear programming. Several approaches have been proposed and reviewed for the integration of gene expression data into FBA models (Machado *et al.*, 2014; Vijayakumar *et al.*, 2017). Each reaction in a FBA model is controlled by an associated gene set, defined through AND/OR Boolean operators between genes. In this work, METRADE (Angione *et al.*, 2015a) was used to constrain the upper- and lower-limits of each reaction as a function of the expression level of the gene set controlling the reaction.

Multiple cellular targets were modeled as a trilevel linear program:

$$\begin{array}{lll} \max & h^\top v & \\ \text{such that} & \max g^\top v & \\ & \text{such that} & \max f^\top v, \quad Sv = 0, \\ & & v^{\min}\varphi(\Theta) \leq v \leq v^{\max}\varphi(\Theta), \end{array} \tag{1}$$

where S is the stoichiometry matrix of the metabolic reactions in the cell, v is the vector of reaction flux rates, while f, g, h are Boolean vectors of weights selecting the three reactions in v whose flux rate will be considered as objective. Lower- and upper-limits for the flux rates in v in the unconstrained model are given by the vectors $v^{\min}$ and $v^{\max}$. The vector Θ represents the gene set expression of the reactions associated with the fluxes in v. The expression level Θ of a gene set is defined from the expression levels $\theta(g)$ of its genes. According to the type of gene set, we define $\Theta(g) = \theta(g)$ for single genes, $\Theta(g_1 \wedge g_2) = \min\{\theta(g_1), \theta(g_2)\}$ for enzymatic complexes, and $\Theta(g_1 \vee g_2) = \max\{\theta(g_1), \theta(g_2)\}$ for isozymes. These rules were applied recursively in case of nested gene sets. To enable transcript-specificity, METRADE was also applied at the splice-isoform level, with annotations retrieved as above (Section 2.1). The function φ maps the expression level of each gene set to a coefficient for the

lower- and upper-limits of the corresponding reaction, and is defined as

$$\varphi(\Theta) = [1 + \gamma|\log(\Theta)|]^{\mathrm{sgn}(\Theta-1)}. \quad (2)$$

Note that the vector notation was adopted, with the convention $0^0 = 1$ when some element of Θ is 1. The sign operator returns a vector of ± 1 (signs of $\Theta - 1$). γ models the reliability of the gene set expression level as an indicator of the rate of production of the associated enzyme.

One may argue that gene expression data is not a good proxy for protein abundance and ultimately for flux rates. However, gene expression data is certainly the omic data with better quality and coverage (almost always genome-scale). Furthermore, in mammals, mRNA levels can be considered the main contributors to the overall protein expression level (Jovanovic *et al.*, 2015; Li *et al.*, 2014). Positive correlation between mRNA and protein levels was also recently found in most normal and cancer cell lines (Kosti *et al.*, 2016). We remark that, although GEMsplice natively handles absolute expression values, it can also handle fold-change values directly, by replacing the logarithmic map (2) with $\varphi(\Theta) = \Theta^{\gamma}$ (see README file in the source code). Furthermore, our method is general and not dependent on a particular proxy. For instance, one may use protein abundances within the same approach proposed here. All simulations were carried out in Matlab R2016b and Octave 4.0.3, with the GLPK solver.

2.3 Modeling cancer biomarkers in multi-level linear programming

The three objectives of the linear program in Equation (1) are chosen to model known mechanisms and key players in cancer metabolism. Here we consider: (i) the biomass (growth rate) reaction, (ii) pyruvate kinase (PK) and (iii) lactate dehydrogenase (LDH). (Note however that this is fully customizable in GEMsplice.) In our pipeline, we minimize the maximum allowed flux rate of PK as the second level of the trilevel linear program, selected through the vector g. As a third (outer) level of our linear program, selected by the vector h, we model the Warburg effect by minimizing the flux through LDH. In fact, a negative flux rate of this reversible reaction models the production of lactate from pyruvate. The first level (inner level), governed by the vector f, is assigned to represent the maximization of the flux through the biomass reaction, modeling the observation that cancer cells grow and divide faster than normal cells in order to achieve their main goal, proliferation.

3 Results

3.1 GEMsplice maps RNA-Seq expression levels and splice isoforms onto a genome-scale model of breast cancer

GEMsplice maps RNA-Seq expression profiles to the metabolic model of breast cancer, using for the first time splice isoform annotations. A cohort of 1455 patients from Cancer RNA-Seq Nexus was taken into account with 31 different breast cancer subsets, 13 of which are invasive carcinomas at different stages. The subset classification is built from observation of genotype and phenotype, and can for instance refer to a specific type of stage of cancer, a disease state, or a particular cell line. To assess differences in the metabolic response among the different types and stages of breast cancer, we consider the average of RNA-Seq expression levels within the same cancer subset, and we finally map these to the phenotypic space of flux rates in the metabolic model using trilevel optimization (Fig. 2,

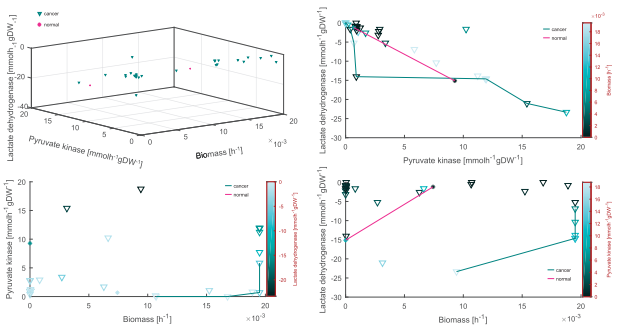


Fig. 2. The 31 RNA-Seq profiles representing different breast cancer subsets are mapped to the tridimensional space of biomass, pyruvate kinase and lactate dehydrogenase. The trilevel linear program (1) is solved after constraining the breast cancer metabolic model using each RNA-Seq expression profile from Cancer RNA-Seq Nexus, including splice-isoform expression levels. The bottom left and both right panels show the three projections onto each pair of axes. In each of these, the optimal breast cancer profiles are highlighted and connected for both cancer and normal profiles to identify the trade-off Pareto optimality

Supplementary Fig. S1 and Supplementary File S2; see Section 2 for details on how phenotype predictions are achieved from RNA-Seq data). GEMsplice correctly predicts breast cancer cells to grow faster than normal healthy cells, while keeping comparable levels of PK activity. Some breast cancer cells are also predicted to use less PK compared to the healthy counterpart, leading to accumulation of intermediates in the glycolytic pathway, therefore fueling the serine biosynthesis pathway (Locasale *et al.*, 2011). The trade-off between high and low PK flux might also explain the balance between PKM2 (limiting step, controlling PK flux) and PKM1 (promoting OXPHOS in the mitochondrion instead). The trade-off between LDH and PK in cancer cells is also lower than in normal cells in terms of negative flux rate values, consistent with the widely accepted evidence that cancer cells use the reverse direction of LDH to produce lactate from pyruvate (Warburg effect). When all the 31 breast cancer subsets are considered, the bottom right panel shows a weak negative correlation between biomass and negative flux of lactate dehydrogenase (Pearson's $r = -0.34$, $P - \text{value} = 7.60 \times 10^{-2}$, Spearman's $\rho = -0.37$, $P - \text{value} = 5.58 \times 10^{-2}$). These predictions are in keeping with the widely accepted fact that the amount of lactate production is positively correlated with tumor growth. More specifically, in breast cancer, a positive correlation has been highlighted between degree of malignancy, degree of mitochondrial structural abnormality (Elliott *et al.*, 2012), and the most common biomarkers of malignant tumor, e.g. intensive use of glycolysis and increased production of lactate (Gonzalez *et al.*, 2012).

3.2 Splice isoform expression-based flux control analysis

To further analyze the cancer-specific models, and to assess the role of the expression of splice isoforms in the model, we propose a steady-state control analysis based on transcript specificity. This method builds on the technique named *metabolic control analysis* (Kacser *et al.*, 1995), which considers the relationship between enzyme activity and flux rates. We adapt this analysis to assess the contribution of each splice isoform to the cellular goals.

As described above, flux rates in our model also depend on the expression level of splice isoforms. In each sample, a control analysis on flux rates v_i with respect to these expression levels x_j involves the estimation of the relation between fractional changes in the flux

rates and fractional changes in the splice isoform expression. This can be written as a scaled partial derivative of the form $C = \frac{x}{v}\frac{\partial v}{\partial x}$.

We here adopt an adjusted control coefficient by shrinking the sample-specific denominator v towards its average $\bar{v}$ across all the samples. As in the standard deviation correction for differential expression tests (Tusher *et al.*, 2001), our correction prevents gross overestimation of control coefficients. Without this correction, 'false positives' would arise when $v_i(x_j)$ is a very small value, irrespective of the numerator. More formally, to approximate this partial derivative, we separately take into account each splice isoform $j = 1, \ldots, N$ included in the breast cancer metabolic model, and we evaluate positive and a negative intermediate control coefficients C_{ij}^{+} and C_{ij}^{-}, defined as

$$C_{ij}^{\pm} = \frac{v_i(x_j \pm \delta) - v_i(x_j)}{(v_i(x_j) + \bar{v}_i)/2} \Big/ \frac{\delta}{x_j},\ i = 1, \ldots, M,\ j = 1, \ldots, N, \qquad (3)$$

where δ is small enough so that the ratio approximates the derivative; v_i is the flux of interest (in our analysis, we focus on the three objectives of the linear program, namely biomass, PK and LDH). Then, for each flux and splice isoform of interest, we finally calculate the overall *flux control coefficient* C_{ij} as the maximum variation caused by a positive or negative perturbation of the isoform expression level:

$$C_{ij} = \max(|C_{ij}^{+}|, |C_{ij}^{-}|). \qquad (4)$$

These transcript- and cancer-specific coefficients evaluate the relative steady-state change in three pivotal flux rates in breast cancer cells, with respect to the relative change in the expression level of the transcript (see Supplementary File S3). For our simulations, we set the denominator $x_j = 1$ to avoid proportionality between the final control coefficient and the expression value itself, and $\delta = 10^{-3}$. In general, the smaller the value chosen for δ, the more effective C in approximating the scaled derivative of v_i. Note that calculating a flux control coefficient for each splice isoform does not explicitly identify controlling transcripts, but rather provides an effective way to quantify their influence on the key flux rates in breast cancer cells.

The ten most influential transcripts were selected independently for biomass, PK and LDH. The union of these three sets of ten transcripts, composed of 14 transcripts, was chosen to perform further enrichment analysis through PANTHER (Mi *et al.*, 2016). A functional classification of each transcript was obtained using the 'protein class' ontology, which is adapted from the PANTHER/X molecular function ontology and also includes Gene Ontology (GO) annotations. Interestingly, the selected transcripts are highly enriched for transmembrane transport and respiratory electron transport chain.

As shown in Figure 3a, the most effective transcript at controlling the value of LDH with a less stringent but non-negligible control on PK and biomass is ENST00000330775, a transcript of glucose-6-phosphate translocase (G6PT). In brain cancer, G6PT is a key player in transducing intracellular signaling events; modulating its expression has been proposed as an anti-cancer strategy (Belkaid *et al.*, 2006). Furthermore, our results suggest that PK and LDH are maximally and simultaneously controlled by ENST00000591899 and ENST00000378667, transcripts of ubiquinol-cytochrome c reductase, whose amplification was suggested to correlate with more aggressive breast cancer (Ohashi *et al.*, 2004). ENST00000507754 and ENST00000327772, transcripts of Complex I [whose activity is known to regulate breast cancer progression (Santidrian *et al.*, 2013)], can control both PK and LDH without disrupting the growth rate. Although further experimental investigation is needed on these key transcripts, our genome-scale method may suggest for the first time transcript-specific targets for anti-cancer strategies.

3.3 Pathway-based flux analysis

A pathway-based perspective has been often taken in genome-scale models with the aim of investigating sensitivity analysis (Costanza *et al.*, 2012; Conway *et al.*, 2016; Kent *et al.*, 2013), and coupled with Bayesian techniques to detect pathway crosstalks and temporal activation profiles (Angione *et al.*, 2015b). To assess the variation in the average flux of each pathway with respect to the unconstrained breast cancer model, we here compute a normalized average pathway flux

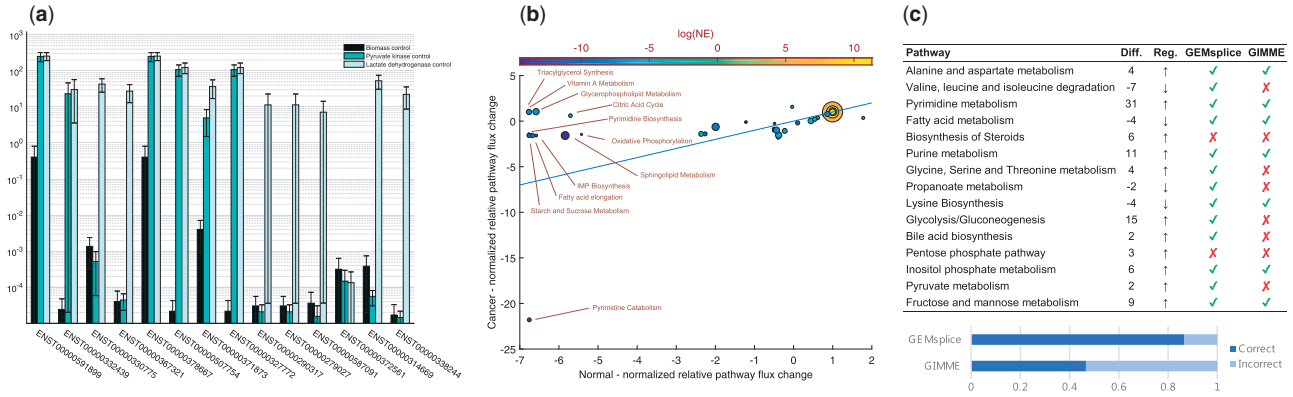
$$d_i = (\bar{w}^{(i)} - w_U^{(i)})/w_U^{(i)}, \quad i = 1, \ldots, P, \qquad (5)$$

where $\bar{w}^{(i)}$ is the average flux in the ith pathway across the different cancer subsets, while $w_U^{(i)}$ indicates the flux of the ith pathway in the unconstrained breast cancer model. To account for the tolerance of the linear solver, average pathway fluxes of less than 10^{-10} were assumed to be zero. Pathways, defined using the KEGG LIGAND database, were inherited from Recon1.

Figure 3b shows the indicator d plotted for each pathway when computed across breast cancer cells and normal cells (results for invasive and unlabeled breast cells are shown in Supplementary Fig. S2). The least-square linear regression reveals a positive pathway flux correlation (Pearson's $r = 0.43$, $P - \text{value} = 1.82 \times 10^{-3}$, Spearman's $\rho = 0.80$, $P - \text{value} = 2.99 \times 10^{-12}$). The eleven pathways indicated in the figure were detected as outliers, with significantly different behavior between cancer and normal breast cells.

To support our predictions, literature-based evidence of breast cancer alterations in such pathways was aggregated. Alanine and aspartate are byproducts of amino acid and glutamine fermentation, an important source for protein synthesis, especially when the cell lacks oxygen. Aspartate is also a main contributor to nucleotide and protein synthesis; the malate/aspartate shuttle pathway translocates electrons for the mitochondrial electron transport chain to produce ATP (Pecqueur *et al.*, 2013). Perturbations in the citric acid cycle (TCA) were largely expected because of the known 'Warburg' reduction of glucose metabolism through the mitochondrion. Likewise, the prediction of altered fatty acid elongation metabolism is in keeping with recent studies reporting elongation of fatty acids as a marker in breast cancer (Feng *et al.*, 2016).

GEMsplice also correctly predicts that pyrimidine biosynthesis pathways are downregulated in normal cells when compared to breast cancer cells (Sigoillot *et al.*, 2004). Vitamin A and carbohydrate altered metabolism are also associated with breast cancer (Doldo *et al.*, 2015; Nagarajan *et al.*, 2016). In breast cancer cells, Inosine monophosphate (IMP) dehydrogenase inhibitors cause growth reduction and phenotypic alterations (Sidi *et al.*, 1988). The prediction of starch and sucrose metabolism confirms previous genome-wide association studies, which reported significant relation between altered starch and sucrose pathway and the risk of developing ER-negative breast cancer (Li *et al.*, 2010). Sphingolipids are responsible for oestrogen signaling and can control differentiation and proliferation of breast cancer cells (Sukocheva *et al.*, 2014). The glycerophospholipid pathway, also flagged as biomarker by GEMsplice, was previously found altered in the metabolic phenotype of breast (Cadenas *et al.*, 2012). Finally, triacylglycerol was found to be a biomarker and a prognostic predictor of poor clinical outcome in triple negative breast cancer (Dai *et al.*, 2016).



Pathway	Diff.	Reg.	GEMsplice	GIMME
Alanine and aspartate metabolism	4	↑	✓	✓
Valine, leucine and isoleucine degradation	-7	↓	✓	✗
Pyrimidine metabolism	31	↑	✓	✓
Fatty acid metabolism	-4	↓	✓	✓
Biosynthesis of Steroids	6	↑	✗	✗
Purine metabolism	11	↑	✓	✓
Glycine, Serine and Threonine metabolism	4	↑	✓	✗
Propanoate metabolism	-2	↓	✓	✗
Lysine Biosynthesis	-4	↓	✓	✓
Glycolysis/Gluconeogenesis	15	↑	✓	✗
Bile acid biosynthesis	2	↑	✓	✗
Pentose phosphate pathway	3	↑	✗	✗
Inositol phosphate metabolism	6	↑	✓	✓
Pyruvate metabolism	2	↑	✓	✗
Fructose and mannose metabolism	9	↑	✓	✓

Fig. 3. Transcript-based metabolic flux and pathway analysis. (**a**) The control analysis is applied to the set of 14 transcripts obtained as a union of the three sets of ten most influential transcripts with respect to biomass, pyruvate kinase and lactate dehydrogenase flux rate. Error bars represent the standard error of the mean $SE = \sigma/\sqrt{n}$, where σ is the standard deviation of the measured effect, computed across all breast cancer subsets, and n represents the number of breast cancer subsets. (**b**) The indicator d plotted for each pathway in breast cancer versus normal breast cells (adjacent to tumor cells). The size of each point represents the size of the corresponding pathway, quantified as the number of reactions belonging to it. The $y = x$ line is shown on the plots to highlight outliers. By definition of d, if a pathway lies outside the line, its perturbation in cancer samples is different from its perturbation in healthy samples. (Both perturbations are computed with respect to the default breast cell model run without omic-derived constraints). Colors are given according to a normalized error of the computed mean flux in each pathway, namely $NE = \sigma/\sqrt{p}$, where σ is the standard deviation of the flux rate in the pathway, computed across all breast cancer subsets, and p represents the size of the pathway. Note that the average flux in a pathway is positive or negative depending on the direction in which its reactions take place. (**c**) GEMsplice correctly predicts 13 out of 15 pathway increased/decreased activity (difference 'Diff' between upregulated and downregulated reactions), while GIMME only predicts 7 out of 15. More accurate predictions are achieved for key cancer biomarkers: glycolysis, glycine, serine, and threonine metabolism, and valine, leucine and isoleucine degradation

3.4 Comparison with gene-level approaches

We here compare the results obtained by GEMsplice with GIMME (Becker *et al.*, 2008), where expression levels are mapped to the model with gene-level resolution only. Present/absent calls for GIMME were obtained by applying zFPKM normalization (Hart *et al.*, 2013). To compare the pathway activity predicted by GEMsplice and GIMME in cancer conditions, we computed the average absolute value of flux rates in all nonzero biomass samples. As a benchmark for comparison and validation, we used the set of metabolic pathways found to be significantly upregulated and downregulated in unfavorable breast cancer conditions (Schramm *et al.*, 2010). Pathways with equal number of upregulated and downregulated reactions were excluded from the analysis.

GEMsplice correctly predicts 13 out of 15 pathway overexpression/underexpression patterns, while GIMME only predicts 7 out of 15 (Fig. 3c, and Supplementary File S5). Three key pathways widely accepted to be dysregulated in breast cancer are correctly identified by GEMsplice, but incorrectly by GIMME: (i) glycolysis, highly upregulated in cancer, shows 3.44 cancer/normal fold change in GEMsplice, while 0.55 in GIMME; (ii) glycine, serine and threonine metabolism, upregulated in cancer: 2.54 fold change in GEMsplice, 0.41 in GIMME; (iii) valine, leucine and isoleucine degradation, downregulated in cancer, 0.30 fold change in GEMsplice, 2.85 in GIMME.

The exact flux rates predicted for lactate dehydrogenase activity were also evaluated against the widely accepted cancer 'Warburg' metabotype. Ten cancer samples are incorrectly predicted by GIMME with zero or negligible ($< 10^{-12}$) 'Warburg' LDH flux (therefore not switching to aerobic glycolysis), while only three by GEMsplice. Furthermore, all non-malignant and normal cells with nonzero growth are correctly predicted by GEMsplice with lowest activity of LDH, while two out of four nonzero-growth normal or non-malignant tissues are incorrectly predicted by GIMME with a typical cancer metabotype, consisting of very high LDH activity. These examples suggest that integrating splice isoform information with metabolic models improves the characterization of breast cancer metabolism. With the current fast-paced improvements in speed and costs of omic profiling, and therefore with increasing availability of isoform-level model annotations and RNA-Seq data, we expect such specific differences to increase even further.

4 Conclusion

While the rate of acquisition of omics data is rapidly increasing, analyzing and extracting information through computational tools arguably remains the main bottleneck in biology. Most studies focus on statistical analysis on gene expression values to evaluate how they vary across different samples. However, modeling how the gene expression alterations change metabolic processes at genome scale provides greater understanding of the phenotypic outcome with respect to studies involving only transcriptomic data (Angione *et al.*, 2016). Given that the last decade in cancer research has repeatedly shown that cancer is a complex disease that cannot be studied by narrowing it down to a single gene or enzyme, a genome-wide metabolic approach seems therefore the right direction to assess and predict the phenotype of a cancer cell.

We here proposed GEMsplice, a method for linking gene expression and splice isoform data to genome-scale metabolic models. GEMsplice is the first attempt at solving one of the main issues of metabolic modeling: as outlined in recent reviews (Geng *et al.*, 2017; Pfau *et al.*, 2016; Ryu *et al.*, 2015; Yizhak *et al.*, 2015), current methods only allow integration of omics data up to the gene level, but not with splice-isoform resolution.

The idea is that every single profile can be used to create a profile-specific model of metabolism that includes splice-isoform annotations. This integrated model can be readily used to predict the flux rate of any biochemical reaction included in the metabolic model, in a range of cancer subtypes. Individual profiles, patients or cells related to the same cancer can be mapped onto this model, for instance to cross-compare their metabolic behavior within the same

cancer type, instead of cross-comparing cells using transcriptomics data only.

Although it allows mapping cancer-specific transcript expression levels onto any metabolic network, GEMsplice comes with limitations and room for improvement. For instance, if reaction-specific information is available, the map from genes to reaction flux bounds could be chosen in a reaction-specific manner and tailored to specific environmental, growth or physiological conditions. Likewise, post-transcriptional regulation of expression levels could be readily included in the model if available. On the other hand, inconsistencies can be used to improve the model and the associated biological knowledge on metabolic enzymes. For instance, incorrect predictions of specific flux rates may shed light on where post-transcriptional regulation takes place (Markert *et al.*, 2015). A further development of cancer metabolic studies may involve a model that takes into account interactions between cancer cells and their environment, e.g. interactions with supporting cells. This approach will likely capture features that cannot be observed with models of single cancer cells.

Breast cancer is manifested through a variety of effects that cannot be reduced to a single feature. The difference in behavior of different cancer cells shows that reconstructing a generic model of a cancer cell is not a viable approach (Ghaffari *et al.*, 2015). Models should therefore always be created in a tissue- and stage-specific fashion. We anticipate that GEMsplice will allow for the first time the generation of such models harnessing the full potential of RNA-Seq, and will facilitate in silico combinatorial experiments with RNA-Seq data. In fact, such omic-informed models can now effectively integrate information on the expression level of splice isoforms, to date largely ignored.

Acknowledgements

The author would like to thank Dr Jim Liu for discussions on Cancer RNA-Seq Nexus, and Dr Syed Haider for discussions on splice isoforms.

Conflict of Interest: none declared.

References

Angione,C. *et al.* (2016) Multiplex methods provide effective integration of multi-omic data in genome-scale models. *BMC Bioinformatics*, **17**, 257.

Angione,C. *et al.* (2015a) Predictive analytics of environmental adaptability in multi-omic network models. *Sci. Rep.*, **5**, 15147.

Angione,C. *et al.* (2015b) A hybrid of metabolic flux analysis and Bayesian factor modeling for multi-omics temporal pathway activation. *ACS Synth. Biol.*, **4**, 880–889.

Becker,S.A. *et al.* (2008) Context-specific metabolic networks are consistent with experiments. *PLoS Comput. Biol.*, **4**, e1000082.

Belkaid,A. *et al.* (2006) Silencing of the human microsomal glucose-6-phosphate translocase induces glioma cell death: potential new anticancer target for curcumin. *FEBS Lett.*, **580**, 3746–3752.

Cadenas,C. *et al.* (2012) Glycerophospholipid profile in oncogene-induced senescence. *Biochim. Biophys. Acta (BBA) Mol. Cell Biol. Lipids*, **1821**, 1256–1268.

Conway,M. *et al.* (2016) Iterative multi level calibration of metabolic networks. *Curr. Bioinf.*, **11**, 93–105.

Costanza,J. *et al.* (2012) Robust design of microbial strains. *Bioinformatics*, **28**, 3097–3104.

Dai,D. *et al.* (2016) Pretreatment tg/hdl-c ratio is superior to triacylglycerol level as an independent prognostic factor for the survival of triple negative breast cancer patients. *J. Cancer*, **7**, 1747.

Doldo,E. *et al.* (2015) Vitamin a, cancer treatment and prevention: the new role of cellular retinol binding proteins. *BioMed Res. Int.*, **2015**, 1.

Duarte,N.C. *et al.* (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. USA*, **104**, 1777–1782.

Elliott,R. *et al.* (2012) Mitochondria organelle transplantation: introduction of normal epithelial mitochondria into human cancer cells inhibits proliferation and increases drug sensitivity. *Breast Cancer Res. Treat.*, **136**, 347–354.

Feng,Y.H. *et al.* (2016) Elovl6 is a poor prognostic predictor in breast cancer. *Oncol. Lett.*, **12**, 207–212.

Geng,J. *et al.* (2017) In silico analysis of human metabolism—reconstruction, contextualization and application of genome-scale models. *Curr. Opin. Syst. Biol.*, **2**, 28–37.

Ghaffari,P. *et al.* (2015) Identifying anti-growth factors for human cancer cell lines through genome-scale metabolic modeling. *Sci. Rep.*, **5**, Article number 8183.

Gonzalez,M.J. *et al.* (2012) The bio-energetic theory of carcinogenesis. *Med. Hypotheses*, **79**, 433–439.

Harrow,J. *et al.* (2012) Gencode: the reference human genome annotation for the encode project. *Genome Res.*, **22**, 1760–1774.

Hart,T. *et al.* (2013) Finding the active genes in deep rna-seq gene expression studies. *BMC Genomics*, **14**, 778.

Jerby,L. *et al.* (2012) Metabolic associations of reduced proliferation and oxidative stress in advanced breast cancer. *Cancer Res.*, **72**, 5712–5720.

Jovanovic,M. *et al.* (2015) Dynamic profiling of the protein life cycle in response to pathogens. *Science*, **347**, 1259038.

Kacser,H. *et al.* (1995) The control of flux. *Biochem. Soc. Trans.*, **23**, 341–366.

Kent,E. *et al.* (2013) What can we learn from global sensitivity analysis of biochemical systems? *PLoS One*, **8**, e79244.

Kosti,I. *et al.* (2016) Cross-tissue analysis of gene and protein expression in normal and cancer tissues. *Sci. Rep.*, **6**, Article number 24799.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Li,B. *et al.* (2011) RSEM: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

Li,J. *et al.* (2010) A genome-wide association scan on estrogen receptor-negative breast cancer. *Breast Cancer Res.*, **12**, R93.

Li,J. *et al.* (2016) Cancer RNA-Seq Nexus: a database of phenotype-specific transcriptome profiling in cancer cells. *Nucleic Acids Res.*, **44**, D944–D951.

Li,J.J. *et al.* (2014) System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ*, **2**, e270.

Locasale,J.W. *et al.* (2011) Phosphoglycerate dehydrogenase diverts glycolytic flux and contributes to oncogenesis. *Nat. Genet.*, **43**, 869–874.

Machado,D. *et al.* (2014) Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS Comput. Biol.*, **10**, e1003580.

Markert,E.K. *et al.* (2015) Mathematical models of cancer metabolism. *Cancer Metab.*, **3**, 1–13.

Mi,H. *et al.* (2016) Panther version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res.*, **44**, D336–D342.

Nagarajan,A. *et al.* (2016) Oncogene-directed alterations in cancer cell metabolism. *Trends Cancer*, **2**, 365–377.

Ohashi,Y. *et al.* (2004) Ubiquinol cytochrome c reductase (uqcrfs1) gene amplification in primary breast cancer core biopsy samples. *Gynecol. Oncol.*, **93**, 54–58.

Pan,Q. *et al.* (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.

Pecqueur,C. *et al.* (2013) Targeting metabolism to induce cell death in cancer cells and cancer stem cells. *Int. J. Cell Biol.*, **2013**, 1.

Pfau,T. *et al.* (2016) Towards improved genome-scale metabolic network reconstructions: unification, transcript specificity and beyond. *Brief. Bioinf.*, **17**, 1060–1069.

Qi,Z. *et al.* (2017) Inference of cancer mechanisms through computational systems analysis. *Mol. BioSyst.*, **13**, 489–497.

Richardson,A.D. *et al.* (2008) Central carbon metabolism in the progression of mammary carcinoma. *Breast Cancer Res. Treat.*, **110**, 297–307.

Roberts,A. *et al.* (2013) Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat. Methods*, **10**, 71–73.

Ryu,J.Y. *et al*. (2015) Reconstruction of genome-scale human metabolic models using omics data. *Integrative Biol*., **7**, 859–868.

Santidrian,A.F. *et al*. (2013) Mitochondrial complex i activity and nad+/nadh balance regulate breast cancer progression. *J. Clin. Investig*., **123**, 1068.

Schellenberger,J. *et al*. (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat. Protoc*., **6**, 1290–1307.

Schramm,G. *et al*. (2010) Analyzing the regulation of metabolic pathways in human breast cancer. *BMC Med. Genomics*, **3**, 39.

Sidi,Y. *et al*. (1988) Growth inhibition and induction of phenotypic alterations in mcf-7 breast cancer cells by an imp dehydrogenase inhibitor. *Br. J. Cancer*, **58**, 61.

Sigoillot,F.D. *et al*. (2004) Breakdown of the regulatory control of pyrimidine biosynthesis in human breast cancer cells. *Int. J. Cancer*, **109**, 491–498.

Smedley,D. *et al*. (2015) The biomart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res*., **43**, W589–W598.

Stephens,Z.D. *et al*. (2015) Big data: astronomical or genomical? *PLoS Biol*., **13**, e1002195.

Sukocheva,O. *et al*. (2014) Role of sphingolipids in oestrogen signalling in breast cancer cells: an update. *J. Endocrinol*., **220**, R25–R35.

Tusher,V.G. *et al*. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, **98**, 5116–5121.

Vijayakumar,S. *et al*. (2017) Seeing the wood for the trees: a forest of methods for optimization and omic-network integration in metabolic modelling. *Brief. Bioinf*., bbx053.

Warburg,O. *et al*. (1927) The metabolism of tumors in the body. *J. Gen. Physiol*., **8**, 519–530.

Wong,N. *et al*. (2015) Pkm2 contributes to cancer metabolism. *Cancer Lett*., **356**, 184–191.

Yang,L. *et al*. (2017) Glutaminolysis: a hallmark of cancer metabolism. *Annual Review of Biomed. Eng*., **19**, 163–194.

Yizhak,K. *et al*. (2015) Modeling cancer metabolism on a genome scale. *Mol. Syst. Biol*., **11**, 817.