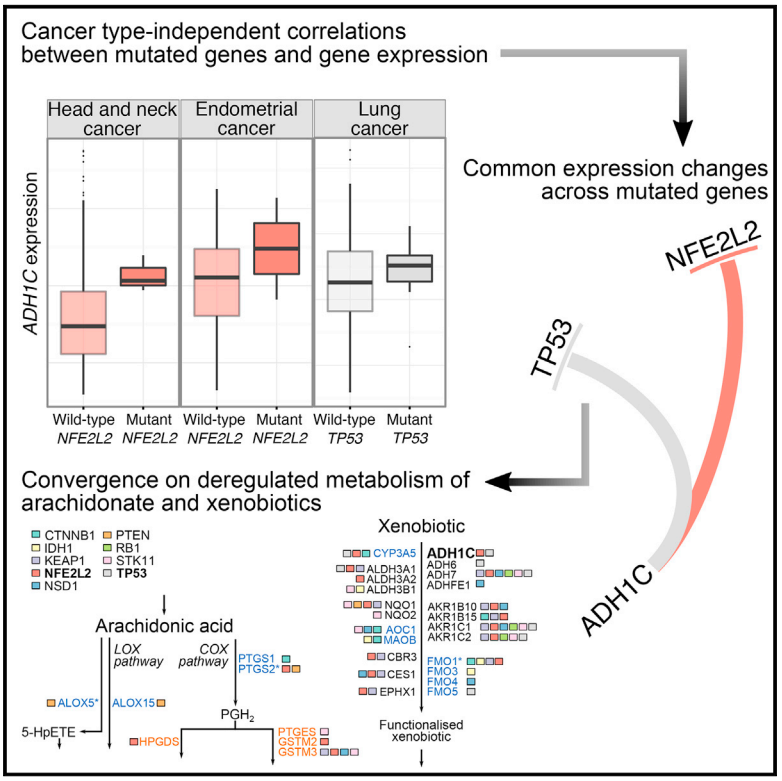
Resource

Cell Reports

Systematic Analysis Reveals that Cancer Mutations Converge on Deregulated Metabolism of Arachidonate and Xenobiotics

Graphical Abstract



Highlights

- Cancer-type independent correlations exist between mutated genes and gene expression
- Nine of 158 mutated genes correlated with expression changes, chiefly in metabolism
- These changes converged on a sub-network centered on arachidonate and xenobiotics
- The correlations were validated in 4,462 independent samples

Authors

Francesco Gatto, Almut Schulze, Jens Nielsen

Correspondence

nielsenj@chalmers.se

In Brief

Gatto et al. systematically analyze cancer-type-independent correlations between mutations in key cancer genes and gene-expression changes and find convergence on a sub-network of reactions involved in the metabolism of xenobiotics and arachidonic acid.

Gatto et al., 2016, Cell Reports *16*, 878–895
July 19, 2016
http://dx.doi.org/10.1016/j.celrep.2016.06.038

CellPress

OPEN ACCESS
CellPress

Cell Reports
Resource

Systematic Analysis Reveals that Cancer Mutations Converge on Deregulated Metabolism of Arachidonate and Xenobiotics

Francesco Gatto,[1] Almut Schulze,[2,3] and Jens Nielsen[1,*]
[1]Department of Biology and Biological Engineering, Chalmers University of Technology, 41296 Göteborg, Sweden
[2]Theodor-Boveri-Institute, Biocenter, 97074 Würzburg, Germany
[3]Comprehensive Cancer Center Mainfranken, 97080 Würzburg, Germany
*Correspondence: nielsenj@chalmers.se
http://dx.doi.org/10.1016/j.celrep.2016.06.038

SUMMARY

Mutations are the basis of the clonal evolution of most cancers. Nevertheless, a systematic analysis of whether mutations are selected in cancer because they lead to the deregulation of specific biological processes independent of the type of cancer is still lacking. In this study, we correlated the genome and transcriptome of 1,082 tumors. We found that nine commonly mutated genes correlated with substantial changes in gene expression, which primarily converged on metabolism. Further network analyses circumscribed the convergence to a network of reactions, termed AraX, that involves the glutathione- and oxygen-mediated metabolism of arachidonic acid and xenobiotics. In an independent cohort of 4,462 samples, all nine mutated genes were consistently correlated with the deregulation of AraX. Among all of the metabolic pathways, AraX deregulation represented the strongest predictor of patient survival. These findings suggest that oncogenic mutations drive a selection process that converges on the deregulation of the AraX network.

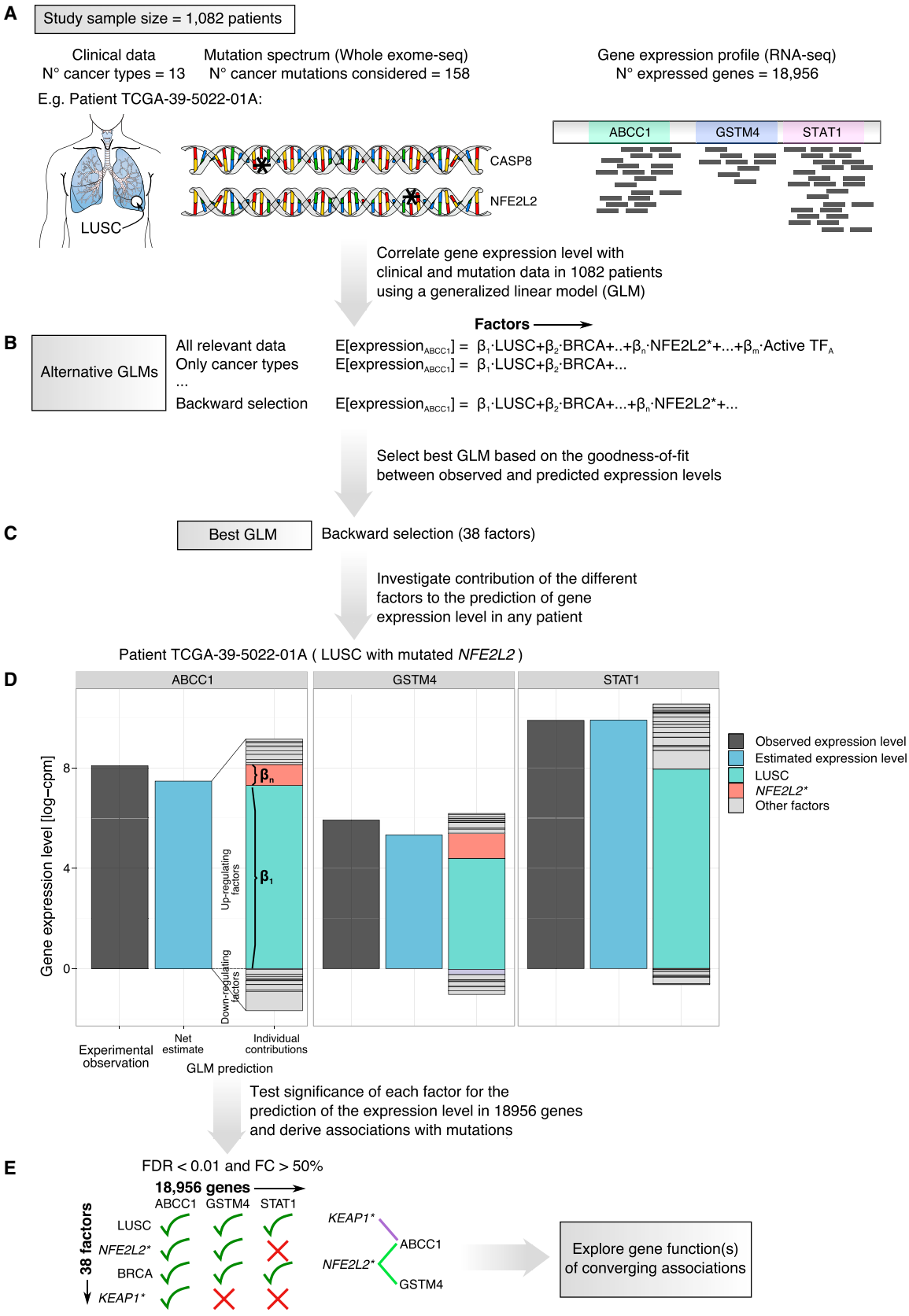
INTRODUCTION

The sequencing of an increasing number of cancer genomes has revealed the extent of the genomic heterogeneity of the disease, which stems from a complex interplay of mutations and natural selection of clones (Yates and Campbell, 2012). The complexity of the cancer genome is a daunting challenge for the rational treatment of the disease. While progress has been made in the attempt to tailor treatments to the defined molecular features of individual tumors, the need for more precise patient stratification provides a rational limit to these strategies (Chin et al., 2011). Moreover, the concept of convergent evolution in cancer could explain the acquisition of the cancer phenotype through multiple routes (Gerlinger et al., 2014; Hanahan and Weinberg, 2011; Weinberg, 2014).

Mutations are central in the evolution of most cancers, and, once acquired, they are liabilities that cancers carry throughout their progression. In addition to the direct effects on cellular signaling networks and the reprogramming of gene expression, cancer mutations also initiate a process of natural selection, which results in the emergence of cell lineages that exhibit the transformed characteristics of cancer (Vogelstein et al., 2013). It is therefore likely that the aggregate of the molecular features of a given tumor, including the presence of a given mutation, is represented in its gene-expression profile. Thus, it is conceivable to factorize the expression level of each gene as the contribution of different tumor features and extract the contribution due to the occurrence of a cancer mutation. In turn, common transcriptional changes attributable to different mutations, such as convergence toward a common set of deregulated genes, should correspond to the deregulation of biological processes crucial for cancer evolution. These key processes are then selected via mutagenesis and natural selection and define the phenotype of the cancer.

Many studies have characterized the gene-expression changes that occur due to prominent cancer-associated mutations in cell lines and animal models (DeNicola et al., 2011; Fodde et al., 1994; Johnson et al., 2001; Podsypanina et al., 1999; Sasaki et al., 2012). However, these mechanistic studies are technologically limited by focusing on one or a few cancer mutations in one or a few cancer types, questioning whether the observed effects of mutations are model or context dependent. On the contrary, a systematic analysis can identify meaningful correlations, but it requires simultaneous knowledge of the presence of a cancer mutation and the levels of all of the transcripts in the same sample in a sufficiently large number of samples that span distinct cancer types. Examples of such pan-cancer studies have so far concentrated on the identification of biological processes putatively affected by cancer mutations and/or epigenetic alterations, without taking into account the underlying changes in gene expression (Ciriello et al., 2013; Hofree et al., 2013; Kandoth et al., 2013).

Here, we used genomic and transcriptomic data from 1,082 human tumor samples across 13 cancer types to derive genome-wide correlations between cancer mutations and transcript levels in human primary tumors. In the first part of this study, we describe the technical details behind the identification of a statistical model to derive meaningful correlations. In the second part of this study, these correlations were used to



(legend on next page)

investigate whether different mutations converge in the transcriptional regulation of defined biological processes. These processes are likely to represent cellular functions that are critical for positive selection during cancer evolution.

RESULTS

Identification of Relevant Factors that Correlate with Changes in Gene Expression in Cancer Using Generalized Linear Models

We first sought to test the existence of a statistical association between changes in gene expression and the presence of a mutation in a cancer-associated gene in the tumor, i.e., if the occurrence of a mutated gene correlates with an increase or decrease in the expression of other genes. RNA sequencing (RNA-seq) profiles for 1,082 primary tumor samples were retrieved from the Cancer Genome Atlas for 13 distinct cancer types (range of 21–199 samples per type; Figure S1) for which a validated mutation spectrum was available (Cerami et al., 2012) (Figure 1A). In this cohort, we focused on the 158 genes mutated at a moderate frequency (>2% samples), of which 12 are mutated at a high frequency (>10% samples; Figure S2). We hypothesized that the level of gene expression could be factorized as the contribution of four sample features: the histopathological cancer type; the expression level of transcription factors; the presence or absence of a mutated gene; and the synergy induced by the occurrence of a mutated gene in a particular cancer type. We therefore employed the established statistical framework of generalized linear models (GLMs) to perform a linear regression of gene expression on the following factors: the 13 cancer types (CTs); the activation status of 119 well-characterized transcription factors (TFs) (Zambelli et al., 2012); the presence or absence of a mutation in one of the 158 genes mutated at a moderate frequency (Muts); and the interaction between the presence of a mutated gene and the cancer type where it occurred (Ints) (Figure 1B). This generated an initial GLM (All), which comprised 316 non-collinear factors, with at least 20 samples per factor.

Many of these factors do not contribute significantly to explaining the expression level of a gene. Therefore, we employed different methods for model selection, including backward selection and regularized regression via the Lasso algorithm (Tibshirani, 1996). These methods identify a minimal number of relevant factors while maintaining an acceptable prediction of the observed gene-expression levels. Each method returned a set of relevant factors that constitute an alternative GLM to the initial All model (Figure 1B). In total, we generated the following 11 GLMs: a backward selection (BS) model (yielding 38 factors); a Lasso model (Lasso, 29 factors); and nine models solely based on a subset of the sample features (i.e., only CT, only TFs, only Muts, or any other combination of these). The best GLM was selected based on the goodness of fit between the observed and predicted expression levels for each gene and the number of factors on which the GLM leverages. A quality measure of this trade-off is the Bayesian information criterion (BIC), which tends to penalize models with too many factors (i.e., higher BIC values), thereby reducing over-fitting. Using each GLM, we calculated the BIC values for each gene (Figure 2A). The Lasso, BS, and onlyCT models performed equally well compared to the other GLMs (Figure 2A). To choose among these three GLMs, we resorted to calculating the Akaike information criterion (AIC), which tends to penalize models with poorer goodness of fit. The conditional probability that a particular GLM performs better in the prediction of the expression level of a given gene can be derived by directly comparing the AIC values of the three GLMs in the form of AIC weights (Wagenmakers and Farrell, 2004). This analysis revealed that, for 15,040 of the genes (79%), the BS model had the highest probability of predicting the expression more accurately than the Lasso model or the GLM in which only cancer type factors were used (onlyCT) (Figure 2B). We noticed that the cancer type still represents the strongest factor in the prediction of gene-expression changes, as exemplified by a principal component analysis on the 1,082 gene-expression profiles (Figure S3) and the reasonable goodness of fit achieved by the onlyCT model (Figure 2A). The major role of the cancer type in defining the tumor phenotype at the transcript level is consistent with a previous pan-cancer study (Hoadley et al., 2014). Nevertheless, a comparison of the gene-wise BIC value using either the onlyCT model or the BS model revealed a shift toward lower BIC values when employing the BS model, which suggested that the additional factors in the BS model contribute to the expression level of many genes (Figure 2C). Overall, the goodness of fit between observed versus predicted gene-expression levels across all 1,082 samples using the BS model generated a Pearson correlation coefficient of $R = 0.963$ (Figure 2D). Considering these results, we adopted the BS model to test for associations between gene expression and mutated genes.

Figure 1. Workflow Used to Derive Statistical Associations between Mutated Cancer Genes and Changes in Gene Expression

(A) Input data for the study were collected from 1,082 patients for which clinical, mutation, and gene-expression level data were simultaneously generated. See also Figures S1 and S2. LUSC, lung squamous cell carcinoma.

(B) The observed level of gene expression was correlated to clinical and mutation data by constructing alternative generalized linear models (GLMs). Each GLM factorized the contribution of predefined factors to the expression level of a given gene (e.g., *ABCC1*) as a linear regression, where coefficients were estimated by fitting the observed gene-expression level in the 1,082 samples. Each GLM predicted an expected value for the expression level of a gene in a sample given the factor values for that sample (e.g., if the sample was LUSC, the GLM added a contribution equal to its estimated coefficient, β_1).

(C) Model selection was performed to decide which GLM returned the best predictions while using a minimal number of factors.

(D) The predicted expression was the net sum of positive and negative factors as determined by the model. As example, expression of *ABCC1* was positively affected by a cancer type factor (LUSC, green bar) and a mutation in *NFE2L2* (red bar).

(E) The significance of each factor could be tested using a threshold for the moderated *t*-statistics and for the minimum expression fold change. The factors representing mutated genes could hereby be associated with gene-expression changes; see also Figure S4. For example, a mutation in *NFE2L2* showed a significant statistical association with expression changes in *ABCC1* (green line). Associations identified in this manner were used to derive networks of deregulated biological processes that were independently associated with mutated genes; see also Figure S5.

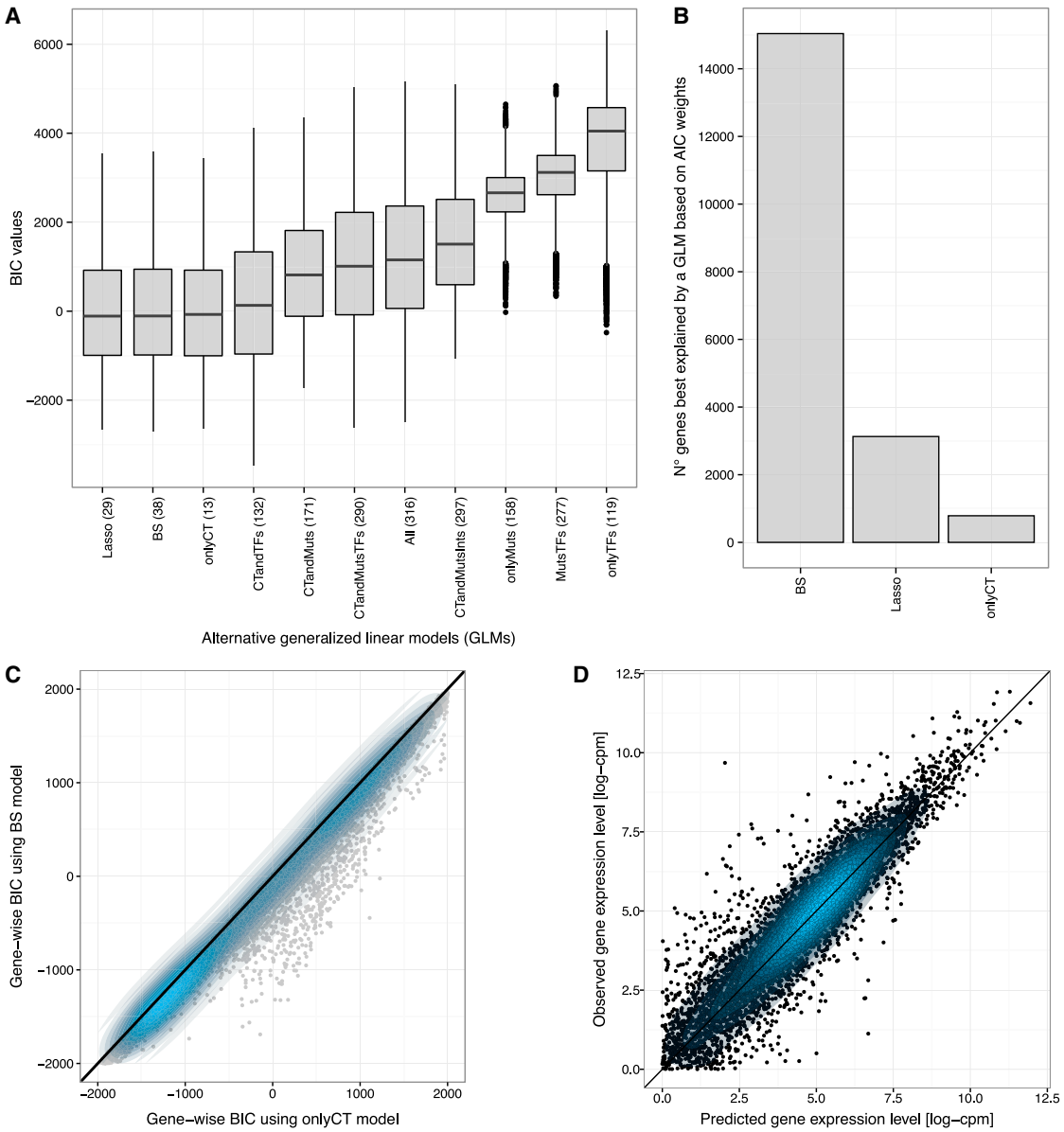


Figure 2. Model Selection According to the Minimum Bayesian or Akaike Information Criterion Revealed that the Backward Selection Model Was Better at Fitting Gene Expression across Samples Than the Alternative GLMs

(A) Boxplot of Bayesian information criterion (BIC) values (one for each gene) using alternative GLMs. Key: Lasso, Lasso non-null factors in >0.5% of all genes (29 factors); BS, backward selection model (38 factors); CT, cancer type factors (13 factors); TFs, transcription factor expression level factors (119 factors); Muts, presence of a mutation in cancer genes (158 factors); Ints, interaction term between presence of a mutated gene and cancer type (126 factors); All, all factors (316 factors).

(B) Number of genes whose expression was best explained by one of the alternative GLMs based on Akaike information criterion (AIC) weights.

(C) Comparison of the BIC value for the regression of expression of each individual gene using either the onlyCT or the BS model. Bluer contours define areas with increasing density of points.

(D) Correlation between observed and predicted gene-expression levels using the *BS* model. Bluer contours define areas with increasing density of points.

Mapping Gene-Expression Changes to Mutated Genes in Cancer

Because the model selection revealed that factors other than the cancer type could contribute to the observed gene-expression levels, we investigated whether mutations in cancer-associated genes represent relevant factors (Figure 1D). Interestingly, mutations in nine genes (out of the initial 158 genes mutated at moderate frequency) were featured as factors in the BS model. These mutated genes are *CTNNB1* (also known as β-catenin), *IDH1*, *KEAP1*, *NFE2L2* (Nrf2), *NSD1*, *PTEN*, *RB1*, *STK11* (LKB1), and *TP53*. The second best performing GLM, the *Lasso* model, also featured six mutated genes as factors, all of which were

among the nine mutated genes identified by the BS model. The contribution of each mutated gene to gene expression was independent of the cancer type and the activation of a given transcription factor, as these contributions were already accounted for by their respective factors. Thus, we sought to determine which genes changed their expression in association with the occurrence of each of these mutated genes by applying differential gene-expression analysis, performed using the voom algorithm (Law et al., 2014) (Figure 1E). We found that on average, the occurrence of a mutated gene was correlated with expression changes in 495 genes (range of 302–764 genes per mutated gene, false discovery rate (FDR) <1% and minimum absolute fold change (FC) >50%; Figure S4). In total, 2,750 genes were associated with at least one mutated gene (1,075 genes [39%] were associated with at least two mutated genes).

We next sought to validate whether the genes found to be associated with one of the nine mutated genes changed their expression in the data derived from independent experiments. To this end, we used 189 experimentally derived gene sets, each representing genes whose expression is altered in response to a perturbation in a key cancer-associated gene (Subramanian et al., 2005). We then performed a gene-set analysis to evaluate whether the genes found to be associated with a given mutated gene are also enriched in any of these 189 gene sets. We observed an overall high consistency between the direction of the regulation of the genes found to be associated with a given mutated gene and the corresponding experimentally derived gene sets (Figure S5). For example, genes found to be upregulated when *RB1* was mutated also significantly enriched the RB_P107_DN.V1_UP gene set, which features genes upregulated in primary keratinocytes from RB1 and RBL1 skin-specific knockout mice (Lara et al., 2008). As a second example, genes associated with *NFE2L2* mutations were also exquisitely overrepresented in the NFE2L2.V2 gene set, which contains genes upregulated in embryonic fibroblasts with a knockout of *NFE2L2* (Malhotra et al., 2010). As a final example, genes found to be upregulated with *CTNNB1* mutations specifically enriched the BCAT_GDS748_UP gene set, which includes genes upregulated in kidney fibroblasts expressing the constitutively active form of *CTNNB1* (Chamorro et al., 2005).

Taken together, these results suggest that differential gene-expression analysis based on the BS model uncovered associations between gene expression and the nine mutated genes that recapitulate the experimentally observed findings. These expression changes are likely to be context independent and not attributable to a specific cancer type. This results from the fact that the BS model accounted for the cancer type as a distinct factor and precluded collinearity among cancer types and mutated genes.

Convergence of Mutation-Associated Gene-Expression Changes in the Regulation of Metabolism

Next, we aimed to elucidate whether the genes whose expression is associated with each mutated gene are involved in specific biological processes. Particularly, we expected that the nine mutated genes were independently associated with processes linked to important cancer-relevant phenotypes, known as the hallmarks of cancer (Hanahan and Weinberg, 2011). Convergence on any of these processes would provide strong evidence that cancer mutations drive the selection of clones that feature properties that reflect these hallmarks. Therefore, we checked whether the genes associated with the nine mutated genes were enriched in any particular biological process, each represented by a distinct Gene Ontology (GO) term. We employed consensus gene-set analysis using Piano (Väremo et al., 2013), which revealed a diverse number of GO biological processes that were significantly associated with each of the examined mutated genes (FDR < 0.01). However, contrary to the premise, only a small number of GO biological processes were simultaneously associated with more than one mutated gene (Figure 3). We therefore further classified those processes that displayed a significant convergence compared to 10,000 random permutations ($p < 0.01$) according to the 24 ancestor categories they are assigned to within the GO hierarchy. We thus observed an overrepresentation of the GO ancestor category of metabolic processes (Figure S6). Intriguingly, metabolism was the GO ancestor category with the most stable overrepresentation when more stringent criteria for convergence are enforced (Figure S6). Collectively, these results suggest that the presence of each of these nine mutated genes entails a diverse spectrum of gene-expression changes in terms of affected biological processes, but that the reprogramming induced by these mutations primarily converges on the regulation of metabolism.

Mutation-Associated Gene-Expression Changes Converge on a Sub-network of Metabolic Reactions

Metabolism appeared to be the biological process that displayed the largest extent of regulation associated with the nine mutated genes. Indeed, mutations in cancer genes have been recognized to regulate metabolism to meet the metabolic requirements of rapid proliferation and allow cancer cells to adapt to the microenvironment (Cairns et al., 2011; Schulze and Harris, 2013). We and others previously found that distinct cancer types featured few common gene-expression changes in metabolism from their respective non-cancerous tissues, which were primarily ascribed to altered nucleotide biosynthesis (Gatto et al., 2014; Hu et al., 2013; Nilsson et al., 2014). However, these studies could not distinguish whether the observed changes in gene expression are attributable to a common adaptation process during cancer progression or are rather the consequence of a specific mutation event (Gatto and Nielsen, 2016). To determine this, we selected among the genes found to be associated with nine mutated genes those that overlap with the 3,765 genes that participate in the human metabolic network (Mardinoglu et al., 2014). This set corresponds to 499 metabolic genes, each associated with the presence of at least one of the nine mutated genes, for a total of 852 associations.

The network of associations between a mutated gene and regulated metabolic genes revealed a number of genes on which multiple mutated genes converged (Figures 4A and S7). However, no metabolic gene showed a convergent association with all mutated genes, nor was there a canonical metabolic process (as defined by GO) to which all mutated genes were associated (refer to Figure 3). We therefore tested the hypothesis that mutations collectively associate with metabolic genes encoding a

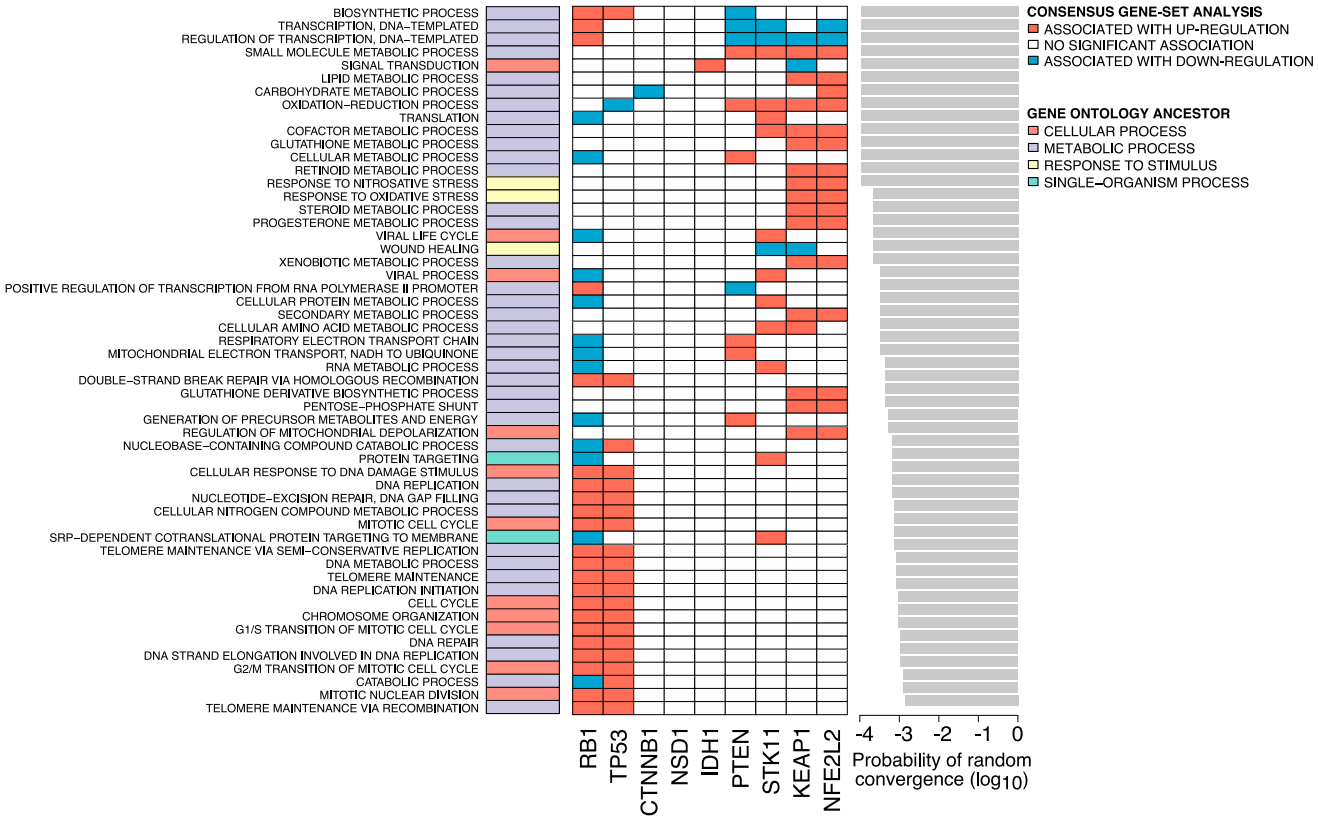


Figure 3. Mutated Genes Converged on the Regulation of GO Biological Processes Primarily Related to Metabolism

Each row indicates a GO term enriched with up- (red) or down- (blue) regulated genes associated with each mutated gene (column) in the consensus gene-set analysis. GO terms were classified according to the ancestor GO category and sorted by the significance of the convergence (bar plot on the right, see also Figure S6).

common yet non-canonical sub-network of reactions. We first mapped the number of mutated genes that converged on each reaction in the human metabolic reaction network (where two reactions are linked if they share a common metabolite) through the association with the underlying reaction-coding gene(s) (Figure 4B). This highlighted distinct clusters of reactions within the human metabolic network. To extract the largest functional cluster, we searched for a connected sub-network of reactions in which the number of converging mutated genes was maximized by using the jActiveNetworks algorithm (Ideker et al., 2002). This approach returned a single high-convergence reaction sub-network (Figure 4C). We characterized this sub-network by determining whether its nodes significantly enrich any pathway and/or metabolite compared to the background human metabolic network. We found that the sub-network featured an overrepresentation of the metabolism of xenobiotics, estrogen, and arachidonic acid (Figure 4D). In addition, individual metabolites such as hydrochloride (a byproduct of xenobiotic metabolism), glutathione, arachidonic acid, and oxygen were also overrepresented within the sub-network (Figure 4D). Collectively, these findings suggest that the regulation of a sub-network of reactions that connects arachidonic acid and xenobiotics via glutathione and oxygen correlates independently with nine frequently mutated genes in cancer.

Curation of the High-Convergence Sub-network of Metabolic Reactions: AraX

Starting from the high-convergence reaction sub-network, we manually curated a representation of the candidate pathway that best represents these reactions according to the literature. We termed this pathway AraX (Figure 5), for arachidonic acid and xenobiotic metabolism. The AraX pathway contains 20% of all mutation-metabolic gene associations found in our study (166 of the 852 links in Figure 4A). One branch of the AraX pathway comprises reactions that control the availability of arachidonic acid and catalyze its conversion to eicosanoids. The second branch facilitates the detoxification of xenobiotics. Importantly, nine enzymes encoded by the genes associated with this pathway are involved in both branches (e.g., CYP4F11). In addition, transporters that can secrete the end products of the pathway are also included (Figure 5). The main co-substrates for arachidonic acid and xenobiotic metabolism are oxygen and glutathione, whose levels are controlled by the remaining genes in the pathway. The overrepresentation of xenobiotic metabolism with genes mutated in cancer was unexpected, considering that the samples used for this study were derived from untreated tumors. The importance of AraX in cancer may reside in its individual components, some of which have established roles in cancer initiation and progression.

Aberrant arachidonic acid metabolism regulates processes critical for cancer progression, mainly by establishing a tumor-supporting microenvironment where immune cells and endothelial cells are recruited to produce mitogens, pro-inflammatory cytokines, and angiogenic factors (Wang and Dubois, 2010). Enzymes within xenobiotic metabolism form reactive intermediates from exogenous and endogenous substrates that can cause cancer initiation, potentially by promoting genotoxicity (Nebert and Dalton, 2006). Both pathways are a primary source of cytosolic reactive oxygen species, which exhibit a characteristically abnormal concentration in many types of cancer cells (Trachootham et al., 2009). Finally, a number of xenobiotic-metabolizing enzymes and transporters in AraX confer cancer cells with mechanisms of detoxification and drug resistance (Fletcher et al., 2010). Taken together, this suggests that AraX is implicated in a number of host-cancer interactions that result in pro-tumorigenic functions.

We next confirmed that, compared to all 186 KEGG pathways, AraX is, on average, the pathway most significantly overrepresented by the genes associated with any of the nine mutated genes (odds ratio, 17.07; 95% 10,000 bootstraps confidence interval [CI], 4.62–26.70); Figure 5B), followed by xenobiotic metabolism by cytochrome P450 (odds ratio, 5.91; 95% CI, 1.73–9.44). AraX is also the most overrepresented pathway when compared to the 674 Reactome pathways, followed by the termination of O-glycan biosynthesis (Figure 5C). Notably, the KEGG and Reactome pathways also include non-metabolic genes, particularly signaling pathways, which are more commonly dysregulated in cancer. In contrast, AraX was constructed based on metabolic genes alone. Overall, this finding suggests that the regulation of a network of metabolic reactions connected to arachidonic acid and xenobiotic metabolism and mediated by glutathione and oxygen is advantageous in cancer, as nine frequently mutated genes independently entail transcriptional changes that converge on this pathway.

Convergence on AraX Regulation Is Validated in an Independent Cohort

We also sought to validate whether the expression changes correlated here with mutations in a cancer-associated gene could be confirmed in an independent cohort and, in particular, whether these correlations indeed converged primarily in the regulation of AraX. We therefore retrieved genomic and transcriptomic data from 4,462 additional primary tumor samples spanning the same 13 cancer types (ranging between 94 and 978 samples per type; Figure S8). First, we checked the validity of the BS model or whether it was over-fitted to the samples in the discovery cohort. To this end, we compared the BIC values in the regression of the expression of each gene using either the BS model or the onlyCT model. The BS model outperformed the onlyCT model in the prediction of expression of most genes, as proven by a substantial shift toward lower BIC values (Figure 6A). This suggests not only that additional factors other than the cancer type are important to explain the expression level of many genes, but also that those factors previously included in the BS model provide a noticeable contribution. In particular, we checked whether gene-expression changes that we associated with the presence of a mutated gene in the discovery cohort were consistent with the changes associated with the same mutated gene in the validation cohort (FDR <1% and minimum absolute fold change >50%). In the validation cohort, the occurrence of a mutated gene correlated on average with expression changes in 796 genes (range 169–2,235 per mutation; Figure S9), for a total of 4,810 genes (note that 1,455 genes [30%] were associated with more than one mutated gene). For each of the nine mutated genes, we found highly significant linear correlations between the fold changes in the expression of associated genes estimated using either the discovery or the validation cohort, with Pearson correlation coefficients ranging from 0.26 for *CTNNB1* to 0.66 for *NFE2L2* ($p = 5 \times 10^{-34}$ to 7×10^{-297}; Figure 6B).

Next, we verified whether expression changes correlated to each of the mutated genes in the validation cohort also converged preferably on AraX compared to other metabolic processes. Compared to the KEGG and Reactome pathways, AraX is the second most significantly overrepresented pathway (average odds ratio across mutated genes, 6.98; 95% bootstrap CI, 2.95–13.24; Figures 6C and 6D), and the only pathway for which we observed a consistent overrepresentation of all nine mutated genes. The most overrepresented pathway was aldarate and ascorbate metabolism in KEGG and glucuronidation in the Reactome, both functionally related to AraX. Furthermore, three of the 12 genes that were associated with at least six mutated genes in the validation cohort belonged to AraX: *HGD*, a dioxygenase in tyrosine and phenylalanine catabolism; *ADH7*, a dehydrogenase that metabolizes hydroxysteroids and lipid peroxidation products; and *ALDH3A1*, which oxidizes aldehyde substrates. Consistently, multiple mutated genes converged in the association with these three genes already in the discovery cohort, particularly for *HGD* and *ADH7*, as indicated by their expression profiles in mutated and non-mutated samples (Figures S10A and S10B). The increased statistical power in the validation cohort allowed us to discover nine additional mutation-associated genes that encode reactions in or connected to AraX. These include *PTGS2* (also known as *COX-2*), an enzyme in the prostaglandin synthesis pathway, and the monooxygenase *FMO1*. With these additional genes, the AraX pathway could be expanded to a total of 84 genes (Figure 5). Taken together, these findings indicate that our analysis yielded reproducible correlations between gene expression and the occurrence of mutations in a cancer-associated gene. Importantly, these correlations primarily converge on the regulation of AraX over any other metabolic process, highlighting the potential importance of this pathway during cancer evolution.

Deregulation of AraX in Cancer Is the Strongest Predictor of Survival among Metabolic Pathways

We sought to investigate the implications of the convergence on AraX in cancer. We observed no obvious pattern in the direction of the regulation of AraX by the different mutated genes, even though we noticed similar effects on AraX in cases of mutated *KEAP1*, *NFE2L2*, *STK11*, and *PTEN*, which tended to be the opposite in cases of mutated *CTNNB1*, *IDH1*, *NSD1*, *RB1*, and *TP53* (Figure S11). Nevertheless, there was an evident mutation-specific modulation in the expression of AraX genes, with varying degrees of overlap. This poses a challenge when

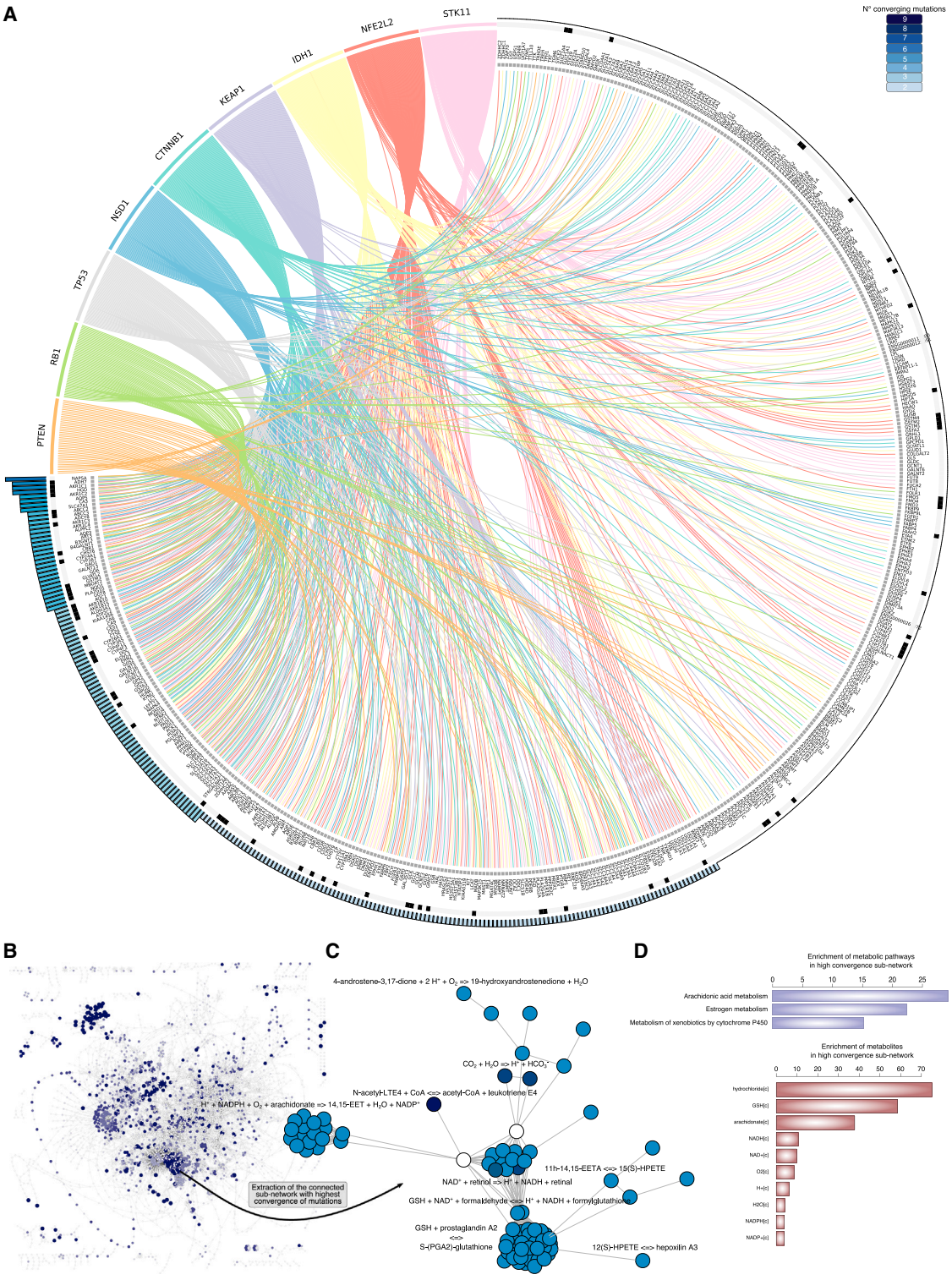


Figure 4. The Network of Associations between Mutated Cancer Genes and Metabolic Genes Highlighted a Region of High Convergence in which Genes Encoded for a Metabolic Sub-network Revolving around Arachidonic Acid and Xenobiotics

(A) Circos plot in which mutated genes were connected to metabolic genes if a statistical association was found (see also Figure S7). Metabolic genes were sorted counter-clockwise according to the number of links (i.e., the number of mutation-metabolic gene associations). Bars indicate the number of mutated genes converging to a particular gene (see also Figure S10). Black entries in the outer circle indicate genes belonging to AraX (introduced in Figure 5).

(legend continued on next page)

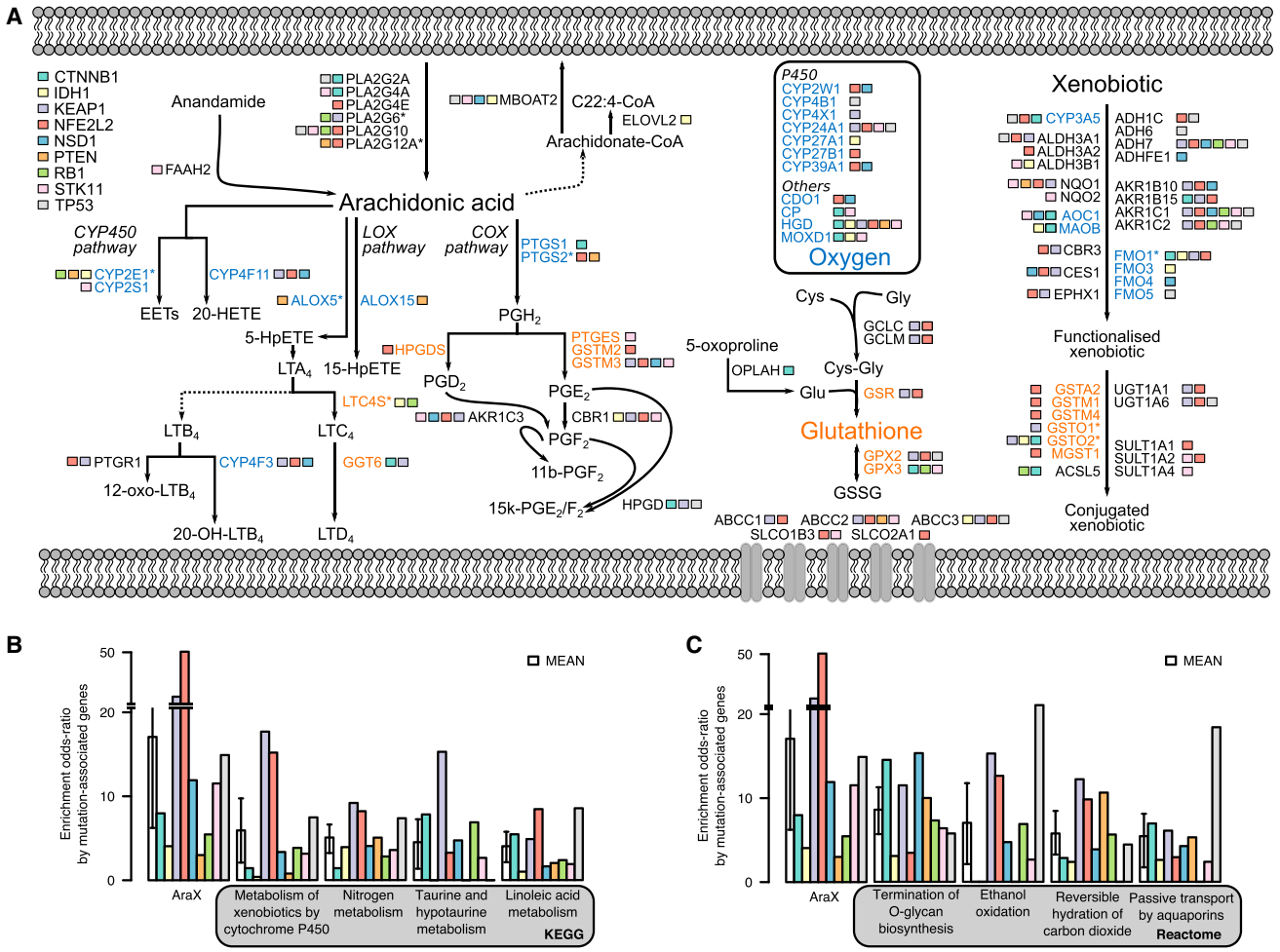


Figure 5. A Literature Curated Sub-network of Reactions that revolved around Arachidonic Acid and Xenobiotic Metabolism, Termed AraX, Showed Convergence by Multiple Mutated Genes in Cancer

(A) The boxes next to each gene indicate which mutated genes were associated with it (see also Figure S11).

(B and C) Overrepresentation of AraX compared to KEGG (B) or Reactome (C) metabolic pathways by genes associated with a mutated gene. Each bar indicates the odds ratio for the corresponding mutation. The top five ranked pathways were sorted according to mean overrepresentation (gray bar), where the error bars span the 95% bootstrap confidence interval.

devising an intervention strategy to normalize the expression or activity of the AraX pathway aimed at halting cancer progression. However, this also suggests that a generic deviance (i.e., deregulation) in the expression of AraX is likely to confer a context-independent selective advantage in cancer. Therefore, we speculated that the extent of AraX deregulation in the tumor should be predictive of an independent measure of selective advantage, such as the patient's survival. Thus, we first estimated a deregulation score for the AraX pathway in each tumor sample using Pathifier (Drier et al., 2013). This score captures the extent to which the expression of a pathway in a tumor sample deviates from its expression in the normal tissue of origin. Then, we performed survival analysis for a subset of the discovery cohort consisting of 718 samples, selected because they encompass six cancer types for which the reference normal samples were available. We regressed overall survival on the AraX deregulation score using a Cox proportional hazards model and we observed a significant increase in hazard with higher AraX deregulation ($p = 6 \times 10^{-8}$). We tested whether a similar trend could be observed concomitantly with a high deregulation of any other metabolic pathway or metabolism in general. However, compared to the 70 KEGG metabolic pathways and a gene

(B) The human metabolic reaction network where each node is a reaction and the blue gradient indicates the number of mutated genes converging to it via association with any reaction-encoding gene.

(C) Extraction of the sub-network in which the number of converging mutation-driven transcriptional changes was maximized.

(D) Characterization of the sub-network in terms of overrepresented pathways (top) and metabolites (bottom) compared to the background human metabolic network.

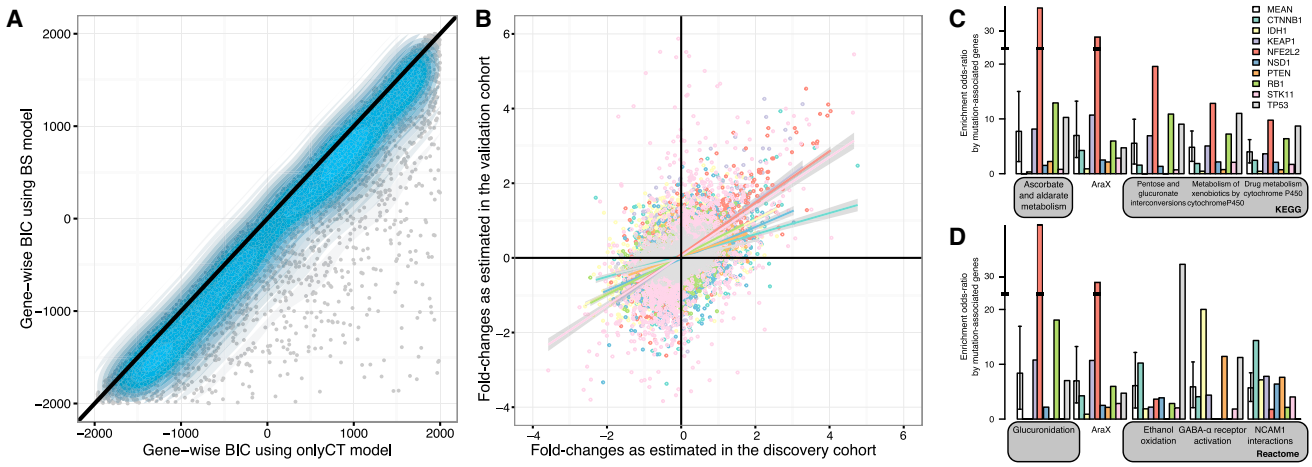


Figure 6. Validation of the Associations between Mutated Genes and Gene Expressions and Their Convergence on AraX Deregulation in an Independent Cohort of 4,462 Samples

(A) Comparison of the BIC value for the regression of expression of each individual gene using either the onlyCT or the BS model. Bluer contours define areas with increasing density of points.

(B) Correlation of expression fold changes for mutation-associated genes as estimated using either the discovery or the validation cohort (each color defines genes associated with a given mutated gene).

(C and D) Overrepresentation of AraX compared to KEGG (C) or Reactome (D) metabolic pathways by genes associated with a mutated gene in the validation cohort. Each bar indicates the odds ratio for the corresponding mutation. The top five ranked pathways were sorted according to mean overrepresentation (gray bar), where the error bars span the 95% bootstrap confidence interval.

set comprising 3,714 metabolic genes, the deregulation of AraX ranked as the best and most robust predictor for survival as estimated by a Lasso penalized Cox proportional hazard model (Figure 7A). At the cross-validated penalty value (log-$\lambda = -2.5$), only two other KEGG metabolic pathways were predictive of survival, oxidative phosphorylation, and the pentose phosphate pathway. Nevertheless, the AraX deregulation score resulted in the highest hazard. To further corroborate this, we could not achieve a comparably significant increase in hazard when we performed a univariate Cox regression of survival on the deregulation score of pathways larger than AraX, such as purine metabolism (159 genes) or the cell cycle (128 genes), despite their established role in malignant transformation (Figures 7B and 7C). These results suggest that AraX deregulation is predictive of survival likely because it confers an evolutionary advantage, and not due to the generic deregulation attributable to heterogeneity in advanced stage tumor samples.

We investigated whether poor prognosis could be attributed to the fact that advanced tumors select for clones with high rather than low AraX deregulation. To this end, we gathered a subset of samples from both cohorts consisting of 1,908 samples, selected to represent the same six cancer types as described above (range of 184–778 per type) and randomly split them into two sub-cohorts (954 sample each). We first verified whether there was an optimal threshold score for AraX deregulation that maximized the difference in prognosis between patients in the discovery sub-cohort using maximally selected rank statistics (Hothorn and Lausen, 2003). This returned a statistically significant threshold score for AraX deregulation equal to 0.764 ($p = 7 \times 10^{-3}$, 1,000 bootstraps 95% CI: 0.731–0.802), above which patients had indeed substantially worse clinical outcome (log-rank test $p = 8 \times 10^{-6}$; Figure 7D). This correlation was independently confirmed when we applied the threshold to classify samples in the validation sub-cohort as having either low or high Arax deregulation ($p = 1 \times 10^{-5}$; Figure 7E). When leveraging on all samples, there was an evident correlation between sample classification into low versus high AraX deregulation and survival (Wald test $p = 6 \times 10^{-10}$). The increased hazard was robust to sub-sampling (hazard ratio = 2.26, 10,000 bootstraps 95% CI: 1.72–2.93) and was not attributable to a bias in the score distribution, as verified by randomly shuffling the sample labels 10,000 times (permutation test $p < 10^{-5}$).

Finally, we sought to characterize the prognostic relevance of AraX deregulation. Across cancer types, we did not detect any dependency between low or high AraX deregulation and other relevant clinical features, in that we found no correlation with age (Wilcoxon rank-sum test $p = 0.745$), with metastatic status (Fisher's exact test $p = 0.199$) nor with cancer-type-specific tumor stages (likelihood ratio test $p = 0.488$–0.782, excluding endometrial cancer due to missing information). This confirms that the association between AraX deregulation and survival is independent of other clinical features represented in the dataset, most notably tumor stage and metastatic status. Within individual cancer types, we recovered a positive trend between AraX deregulation and low survival for invasive breast carcinoma (age-adjusted hazard ratio = 3.468, 95% CI: 1.03–11.7, $p = 0.044$), but it was not significant in any of the other cancer types. The lack of this correlation might be attributable to the fact that we observed an association between a cancer type and its expected AraX deregulation score (ranging from 0.28 in endometrial cancer to 0.67 in head and neck squamous cell carcinoma, likelihood ratio test $p < 10^{-16}$; Figure S12A). This displayed an inverse albeit low correlation with the corresponding 5-year

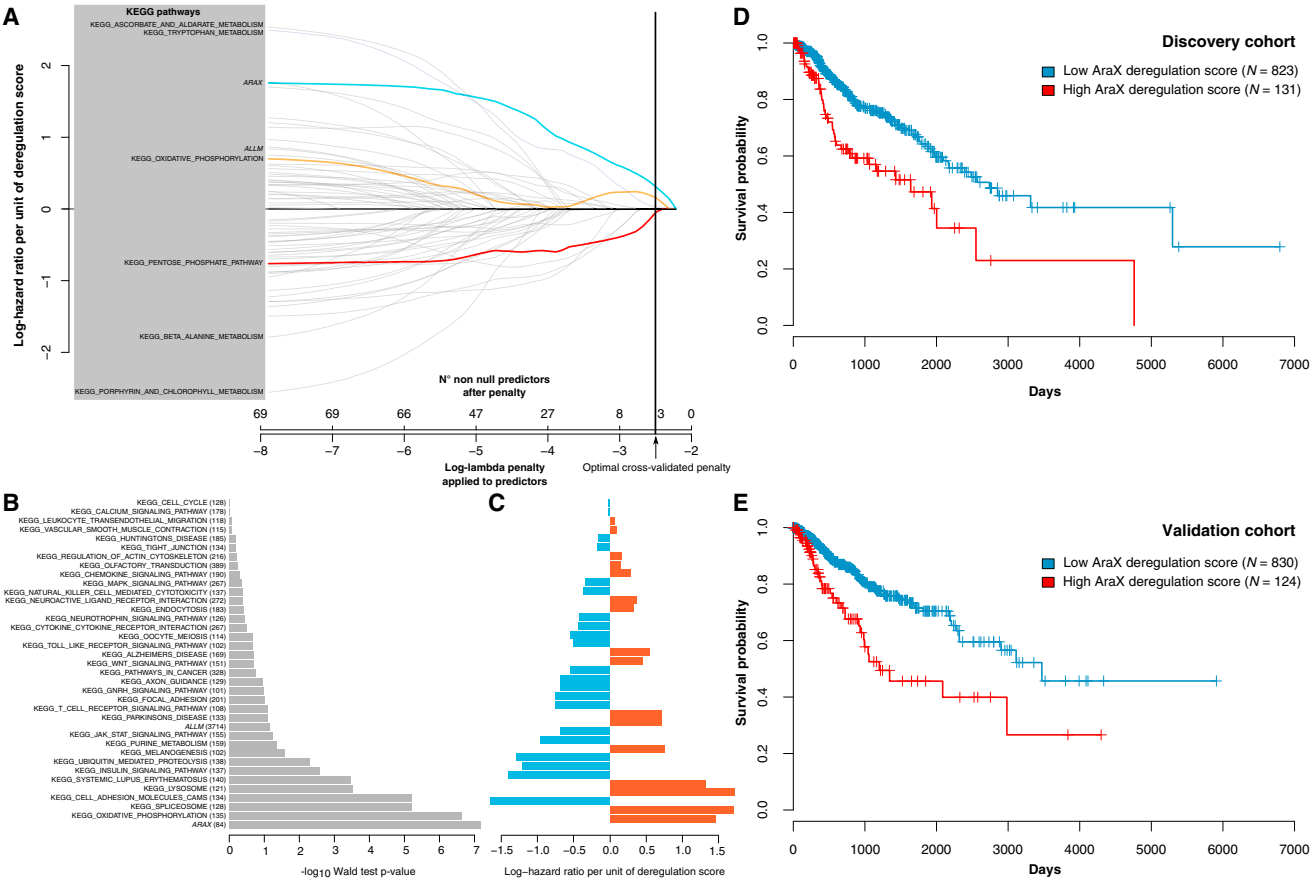


Figure 7. Survival Analysis of Patients Stratified upon Metabolic Pathway Deregulation Revealed that AraX Was the Strongest Predictor of Survival

(A) Log-hazard ratio per unit of deregulation score for AraX, 186 KEGG metabolic pathways, and a gene set including 3,714 metabolic genes at different Lasso penalties (log-λ) in the multivariate prediction of overall survival for 718 tumors. Each path represents a different pathway. Only the paths relative to pathways that were predictive of survival at the optimal lasso penalty, log-$\lambda = -2.5$ (vertical line) were colored. The graph shows that AraX was the strongest predictor of survival at the optimal lasso penalty, followed by oxidative phosphorylation and the pentose phosphate pathway and that its predictive strength is robust to different choices of lasso penalties.

(B) Wald test statistic in the univariate Cox regression of survival using deregulation of the pathways in (A) that contained at least 100 genes.

(C) Log-hazard ratio per unit of deregulation score for the pathways in (B).

(D and E) Kaplan-Meier survival plots for 1,908 tumor samples equally split in a discovery (D) and validation (E) cohort and stratified upon low (blue) versus high (red) AraX deregulation score according to a threshold derived in the discovery cohort (see also Figures S12 and S13).

survival for cancers of the same tissue (likelihood ratio test $p < 4 \times 10^{-12}$; Figure S12B), which suggests that more aggressive cancer types tend to feature higher AraX deregulation. An analysis of statistical power indicated that larger sample sizes are needed to further discern a correlation between cancer-type-specific overall survival and the corresponding AraX deregulation scores (Figure S13). This was the case with breast invasive carcinoma, for which the highest number of samples was available (n = 778).

In conclusion, the strong association of AraX deregulation with poor prognosis as opposed to other metabolic pathways underscores the biological significance of this pathway in cancer, which indicates that advanced tumors select for AraX deregulation. This conclusion corroborates that our study has correctly identified a relevant cancer process as a node of convergent evolution and suggests that the aberrant expression of AraX confers a selective advantage for cancer progression, potentially more than any other metabolic process.

DISCUSSION

Cancer cells exhibit heterogeneous combinations of genetic alterations that are the result of a process of natural selection. Through this process, cancer cells deregulate critical biological functions to establish the hallmarks of the transformed phenotype (Vogelstein et al., 2013). The concept of convergent evolution in cancer implies that different genetic alterations can result in functionally similar outputs, which are likely to reflect an evolutionary advantage for the cancer cells with respect to their microenvironment (Gerlinger et al., 2014).

Probing convergent evolution in molecular studies is technically challenging in that typically few mutations can be induced

in defined tumor models, raising the possibility that the observed effects are context dependent. Here, we resorted to a systematic analysis that extracted the regulation of gene expression concomitant with mutations in major cancer genes. Unexpectedly, we found that mutations in only nine of 158 cancer genes were associated with substantial and recurrent changes in gene expression, and these were largely heterogeneous. Within this complexity, we could uncover a single node of convergence, a metabolic pathway that we termed AraX. Consistently, a parallel and complementary systematic analysis to ours suggested that also copy-number alterations seem to provide an evolutionary advantage in cancer if they deregulate cell metabolism (Sharma et al., 2016). AraX is a network of metabolic reactions that revolve around the metabolism of arachidonic acid and xenobiotics mediated by oxygen and glutathione, which is consistent with the importance of regulating these cofactors in different tumor models. In concordance with this finding, previous network analyses of cancer metabolism revealed elements of AraX as the only sub-networks recurrently deregulated across several different tumor types compared to matched normal tissues (Agren et al., 2012; Wang et al., 2012). Our results showed that nine frequently mutated genes in cancer converge in a significant association with the transcriptional deregulation of AraX, more than with any other metabolic or biological pathway. This convergence is striking in that it seemingly occurs regardless of the cancer type and independent of the expression of a number of transcription factors, at least to the extent that our generalized linear model could adjust for the effect of these confounding factors. A limitation is that this approach cannot disentangle severe collinearities for some genes typically mutated in specific cancer types (e.g., *VHL* in renal cell carcinoma), so these were excluded from further analysis.

Survival analysis further corroborated that the deregulation of AraX likely confers a context-independent selective advantage in cancer evolution. However, AraX deregulation was prominent in tumors belonging to known aggressive cancer types, e.g., head and neck squamous cell carcinoma. This suggests that tumor aggressiveness correlates with AraX deregulation but cannot validate the extent to which AraX deregulation is responsible for a poorer prognosis in an individual cancer type. The limited sample sizes in our cohort allowed us to test and confirm this cancer-type-specific correlation only in the case of breast cancer.

Furthermore, our analyses also unveiled other aspects regarding the convergence between mutations in different cancer genes. First, only two genes showed expression changes that were independently associated with at least six mutated genes both in the discovery and the validation cohort. These genes are *HGD* and *ADH7*. Remarkably, both genes code for proteins with metabolic functions and are linked to AraX. To our knowledge, *HGD* has never been implicated in cancer, but polymorphisms in *ADH7* have been associated with susceptibility to upper aero-digestive tract cancers (Hashibe et al., 2008; McKay et al., 2011) and head and neck squamous cell carcinoma (Wang et al., 2014; Wei et al., 2010). Second, other metabolic processes showed patterns of convergence, although not as pronounced as for AraX. Prominently, many mutation-associated genes were related to protein glycosylation (e.g., O-glycan biosynthesis).

Intriguingly, the fact that AraX is a transcriptionally regulated pathway of oxygen-consuming reactions could reflect a strategy by which cancer cells adapt to tumor hypoxia by regulating oxygen-dependent enzymes in an attempt to compensate for reduced oxygen availability. Mutations in cancer genes select independently for the deregulation of this pathway, potentially under the selective pressure of hypoxia. Moreover, the direct link between glutathione metabolism and the processes within AraX implicate a central role for oxidative stress in cancer development.

Collectively, our analysis suggests that, in cancer, convergent evolution results in the transcriptional deregulation of metabolic processes, primarily the AraX pathway. We speculate that an effective strategy to arrest cancer evolution could be represented by either modulating the activity of components of the AraX pathway or by impeding the major regulatory axis associated with it, the Keap1-Nrf3 pathway, using a multi-targeted approach, a strategy also advocated by network pharmacology (Hopkins, 2008).

EXPERIMENTAL PROCEDURES

Data and scripts for the computational workflow described in Figure 1 are deposited under Synapse ID: syn3163200.

Data Retrieval

RNA-seq gene-expression profiles and clinical data for 1,082 primary tumor samples encompassing 13 cancer types (BLCA, bladder adenocarcinoma; BRCA, breast carcinoma; COAD, colon adenocarcinoma; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell carcinoma; KIRC, clear cell renal cell carcinoma; LGG, low-grade glioma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; OV, ovarian carcinoma; READ, rectum adenocarcinoma; PAAD, pancreatic adenocarcinoma; UCEC, uterine corpus endometrial carcinoma) were downloaded from the Cancer Genome Atlas (TCGA) in November 2013. A second group of 4,462 primary tumor samples encompassing the same 13 cancer types were also downloaded from TCGA in August 2015. Mutation profiles for all samples in this study were obtained from the cBioPortal (Gao et al., 2013).

Differential Gene-Expression Analysis

RNA-seq-generated read count tables were used to estimate gene expression in each sample in the pan-cancer cohort. To this end, we adopted voom, an approach that extends the generalized linear model (GLM) for microarray gene-expression signals to analyze count-based expression data (Law et al., 2013). The gene-wise count variance is calculated from the linear regression of gene-wise observed log counts across all samples in the cohort according to a number of factors (to be decided), and it is defined as the gene-wise residual SD of the regression. If a lowess curve is fitted to square-root residual SD as a function of mean log counts, it is possible to predict the square-root SD of each observation (i.e., log counts for a given gene in a given sample) from this mean-variance trend. Differential gene-expression analysis for each factor is then performed using the standard linear modeling procedure proposed by limma (Smyth, 2004), with the addition that the log counts per million of each observation are corrected using the predicted variance as an inverse weight. Even if voom assumes that each observation is normally distributed, this method proved to outperform count-based approaches in differential expression analysis comparison studies (Rapaport et al., 2013). The significance of each factor in the regression of the expression of each gene was then tested using moderated *t*-statistics. So generated p values were corrected for multiple testing by controlling the false discovery rate (FDR) across genes using the Benjamini and Hochberg correction and by adopting the nestedF correction across contrasts. A factor was deemed significant in the regression of the expression of a gene if it was associated to at least 50% fold change ($|\log_2 FC| > 1.5$) with a FDR <0.01.

Generalized Linear Model Selection

In order to perform the differential gene-expression analysis above, it is required to define the factors for the regression. These factors are devoted to explain the biological variability of gene-wise counts across the samples in the pan-cancer cohort. They should capture the main contributions and some smaller contributions interesting to our investigation. Hence we tentatively selected the following factors for an initial design (All):

(1) The cancer types, i.e., the belonging to a histopathologically defined cancer type among the 13 types in the cohort.
(2) The mutation status of 158 cancer-associated genes. An initial list of 260 genes was generated by merging the Cancer5000 and Cancer5000-S lists in Lawrence et al. (2014). We excluded *HIST1H3B*, *HIST1H4E*, and *MLL4*, which could not be uniquely mapped using the Ensembl v.73 annotation. Furthermore, 102 genes that were not mutated at moderate frequency in the cohort (>2%) were also excluded. For the purpose of this study, any type of mutation in these genes was sufficient to qualify the gene as mutated in the sample.
(3) The activation status of 119 well-characterized transcription factors (Zambelli et al., 2012), which was defined by the belonging to a certain quintile of expression in the pan-cancer cohort.
(4) The interaction terms between a cancer type and a cancer-associated gene mutated at high frequency. These were defined as the 12 mutations with a frequency >10% across the pan-cancer cohort. There were 126 such interaction terms, excluding those linearly dependent on the other factors. These factors took into account cancer-type-dependent contributions of mutated genes.

We applied the following filters to exclude factors from the initial design that may confound the regression:

(1) At least 20 samples in the cohort belonged to each factor (e.g., at least 20 samples belonged to a certain cancer type).
(2) Each factor had a maximum variance inflation factor (VIF) equal to 4, excluding interaction terms. This filter attempted to minimize collinearity, which may occur in this cohort due to cancer-type-specific mutations (e.g., *VHL* in clear cell renal cell carcinoma). In this case, the gene-expression signal could not be properly factorized in the contribution of the collinear factors, and only the main factor was retained (in our case the cancer type).

Using the same notation (where appropriate) as in voom, the GLM (1) was

$$E\left(y_{g,i}\right) = \mu_{g,i} = I_g + x_i^T \beta_g, \qquad \text{(Equation 1)}$$

where $y_{g,i}$ is the log counts per million (log-cpm) value for gene g in sample i, $\mu_{g,i}$ is the expected value, x_i is the vector of covariate values in sample i, β_g is the (unknown) vector of coefficients representing the contribution of each covariate on the expected value, and I_g is the explicitly formulated intercept of the GLM. In our formulation, the All model (2) becomes

$$\mu_{g,i} = I_g + \left(\sum_{m=1}^{nCancerMutations} \beta_m x_m + \sum_{t=1}^{nCancerTypes} \beta_t x_t + \sum_{f=1}^{nTranscriptionFactors} \beta_f x_f + \sum_{l=1}^{nInteractions} \beta_l x_l \right)^T, \qquad \text{(Equation 2)}$$

where x_m is a binary value {0,1} indicating the absence or presence of a mutation in gene m in the sample i; x_t is a binary value {0,1} indicating the belonging of sample i to the cancer type t; x_f is a ternary value {–1,0,1} indicating whether the expression of transcription factor f in sample i is in the bottom quintile, second to fourth quintile, or top quintile with respect to the distribution of its expression values in the pan-cancer cohort; and x_l is a binary value {0,1} indicating whether there is the interaction l between the cancer type to which sample i belongs and a frequently mutated gene.

We excluded the following observations from this study:

1. All genes that had ambiguous annotation in Ensembl v.73. This set corresponded to 565 genes.
2. All genes that were not detected in any sample. A gene was considered detected if at least ten counts were reported in 10% of the samples. Although the opposite may have occured due to an actual repression of the gene, this signal could not be distinguished from genes that were misannotated or, more likely, from genes whose transcripts could not be detected due to technical limitation in the sensitivity of the sequencing instrument. These observations did not add any information on the expression status of the (presumptive) gene, and thus their removal did not alter the result of downstream analyses. This set corresponded to 1,075 genes.

Overall, 1,575 genes were excluded from the initial set of 20,531 genes (65 overlapped between the above mentioned filtered sets), yielding a total of 18,956 genes analyzed.

Many factors in the All model are unlikely to contribute in explaining the expression of most genes, thereby increasing the risk of over-fitting. We adopted two different model selection methods to derive the most relevant factors while using a minimal number of factors. First, backward selection (Yan and Su, 2009) was used to exclude, at each iteration, the factor that was associated with the least number of differentially expressed genes. The procedure was stopped once the number of differentially expressed genes (defined as FDR <0.01 and $|\log_2 FC| > 1.5$) was greater than 0.5% of all genes (i.e., 90 genes). The resulting GLM contained 38 factors (BS model). Second, we used L1-constrained regression shrinkage using the Lasso algorithm (Tibshirani, 1996) to compute, for each gene, the factors in the All model with a non-null coefficient. The penalty value used for the Lasso regression was calculated such that the mean 10-fold cross-validated error was minimum. The Lasso method was implemented using the R-package glmnet (Friedman et al., 2010). We constructed a GLM based on the factors with a significant coefficient ($|\beta| > \log_2(1.5)$) in at least 0.5%, of all genes (Lasso model), resulting in 29 factors. Finally, we constructed alternative GLMs that feature either only the cancer type (CT) or the transcription factor levels (TFs) or the mutation statuses (Muts) or any other meaningful combination of these classes with interactions, if appropriate.

The best GLM was evaluated by first calculating the Bayesian information criterion (BIC) values for the goodness of fit of all genes by each GLM. This criterion was chosen for its ability to capture the trade-off between the goodness of fit and the stringent penalty on the number of factors utilized in the regression of the expression of a gene (for each GLM, there is a BIC value per gene), thus minimizing over-fitting. Given that the Lasso, BS, and onlyCT performed equally well, we compared the goodness of fit of these models in terms of Akaike information criterion (AIC) values, which, compared to BIC values, penalize a poorer goodness of fit over the number of factors. To this end, we computed, for each gene, the difference between the AIC value returned by the current GLM and the minimum AIC value observed using any of the three GLMs. From this, we calculated the AIC weight of the alternative GLMs in the regression of each gene. The AIC weights were transformed into probabilities that a certain GLM was the most likely to explain the expression of that gene. Finally, we counted for each GLM the number of genes whose expression was best explained by that GLM. If the onlyCT model was considered as a positive control for the regression of gene expression, the comparison of gene-wise BIC value between the onlyCT model and an alternative GLM was used to determine whether the additional factors in the alternative GLM provided a better goodness of fit while controlling for over-fitting (a positive comparison means that the gene-wise BIC values are skewed toward more negative values when using the alternative GLM). The model selection was implemented in R 3.1.2.

Gene-Set Analyses

The gene-set analyses were performed using the R-package Piano (Väremo et al., 2013). Compared to other gene-set analyses methods, this package both distinguishes the direction of gene-set expression regulation and leverages on the consensus of different statistical tests. In all analyses, we evaluated the significance of a gene set using the genes found here to be associated with a mutated gene (here on mutation-associated genes). For each mutated gene, the list of mutation-associated genes was generated using the differentially

gene-expression analysis based on the BS model (see Differential gene-expression analysis). In the case of enrichment of the 189 gene sets representing each a genetic perturbation in a key cancer-associated gene (retrieved from the Molecular Signatures Database (MSigDB) [Subramanian et al., 2005]), the significance of a gene set was tested using the Stouffer's test, and the p values were controlled for multiple testing by transformation to FDR using the Benjamini and Hochberg correction. To check for consistency between the genetic perturbation represented by a gene set and the expected effect on gene-expression by a mutation, we compared separately the gene sets (if significant, i.e., gene-set FDR <0.01) mostly associated with upregulated or downregulated genes (in Piano, so called "mixed directional" classes). For example, genes here found upregulated when *CTNNB1* (β-catenin) was mutated were significantly associated with the BCAT_BILD_ET_AL_UP gene set, in which β-catenin (BCAT) was overexpressed in primary epithelial breast cancer cell.

In the case of enrichment of GO biological processes, 8,255 gene sets were retrieved using the R-package biomaRt (Durinck et al., 2009). The significance of a gene set was tested using the consensus between six tests (Fisher's test, Stouffer's test, Reporter test, Tail strength test, mean, and median), and the p values were controlled for multiple testing by transformation to FDR using the Benjamini and Hochberg correction. If gene-set FDR was <0.01, the underlying biological process was deemed significantly associated with the mutated gene. To compute the probability that multiple mutated genes were simultaneously associated with a gene set, we designed a permutation test in which the gene sets significantly associated with a mutated gene were randomly permuted 10,000 times. Then, we calculated a p value as the frequency at which a gene set is randomly associated with a number of mutated genes greater or equal to that observed prior randomization. Next, we computed using the Fisher's exact test which ancestor GO category (defined as the children of the GO term "biological process") was overrepresented by the GO terms that showed significant convergence. Finally, we estimated the robustness of the supposed overrepresentation of an ancestor GO category repeating this above operation using only those GO terms that showed convergence by an increasing number of mutated genes (i.e., given *n* GO terms associated with at least x mutated genes, we computed which GO ancestors were overrepresented by the *n* GO terms).

Extraction of the High-Convergence Reaction Sub-network

The human genome-scale metabolic model HMR2 was downloaded from http://www.metabolicatlas.com/. We generated a reaction network from the model where reactions were nodes, and an edge linked two nodes if there was at least one metabolite shared by the two reactions. We excluded 18 metabolites with exceptionally high degree (>200) to prevent a combinatorial explosion of reaction-reaction edges. Then, we used the jActiveNetwork algorithm (Ideker et al., 2002) to extract from this reaction network a connected sub-network that maximizes the number of mutations converging to it. To this end, we counted for each reaction the number of times that any mutation was found associated with a gene encoding that reaction. Each reaction of the network was then scored using this count. We subtracted a penalty equal to 5 to the score to ensure that the extracted sub-network was reasonably small yet comprised as many reactions with at least four mutated genes converging to them. This prevented that biologically related mutated genes (like *KEAP1* and *NFE2L2*) could significantly bias the emerging sub-network. Artificial reactions introduced in HMR2 for modeling purposes (defined by the HMR2 sub-systems "Isolated," "Artificial reactions," "Exchange reactions," "Pool reactions") were further penalized with a score of −100. The search was implemented using the R-package BioNet (Beisser et al., 2010). The returned high-convergence reaction sub-network contained 90 reactions (nodes) out of the 8,184 reactions that were present in the reaction network.

Analysis of the High-Convergence Reaction Sub-network

We characterized the high-convergence reaction sub-network by comparing the frequency of metabolites and pathways represented by the reactions in the sub-network to the background frequency in HMR2. The overrepresentation of metabolites and pathways was calculated using the Fisher's exact test. To further aid the interpretation of the reactions part of the high-convergence reaction sub-network, this was broken down in reaction clusters, defined as sets of reactions that share the same gene-reaction association. These were returned by applying unsupervised hierarchical clustering to the gene-reaction association matrix in HMR2 limited to include the reactions in the high-convergence reaction sub-network and the genes associated with at least one mutated gene. This operation reduced the complexity of the high-convergence reaction sub-network to 14 reaction clusters.

Curation of the High-Convergence Reaction Sub-network

Starting from the above analysis, we consulted the literature to frame the high-convergence reaction sub-network in the context of well-defined metabolic functions and reconstruct a comprehensive pathway. Also, we manually reviewed every metabolic gene associated with at least one mutated gene (Figure 4A) and verified whether there exists a relation with the emerging pathway. We discarded a candidate gene if its pan-cancer expression level was not appreciable in a reasonable number of samples (minimum library size-adjusted log-cpm in the top 20% equal to 1).

We initially focused on arachidonic acid and its metabolism given its prominent enrichment in the high-convergence reaction sub-network compared to HMR2. The reaction clusters 3 and 4 indicate inclusion of reactions belonging to the cytochrome P450 pathways of arachidonic acid. These include reactions in the hydroxylase pathway, catalyzed by *CYP4F11* (Arnold et al., 2010; Chuang et al., 2004; Kroetz and Zeldin, 2002). Other mutation-associated genes belong to the epoxygenase pathway, specifically *CYP2S1* (Bui et al., 2011). *CYP4X1* is also a likely member of this pathway, but evidence for specificity to arachidonic acid is still inconclusive (Kumar, 2015; Stark et al., 2008). The reaction clusters 5 and 7 implicate another major route of arachidonic acid, the cyclooxygenase (COX) pathway to produce prostaglandins. In total, eight mutation-associated genes participated in the metabolism of prostaglandin H_2, the first product of arachidonic acid conversion in the COX pathway. Among these is *PTGS1* (also known as *COX-1*), which catalyzes the first common step in the COX pathway from arachidonic acid to prostaglandin H_2 (Schneider and Pozzi, 2011). *PTGES*, *GSTM2*, and *GSTM3* can convert prostaglandin H_2 to prostaglandin E_2 (Hayes et al., 2005; Schneider and Pozzi, 2011), which, in turn, can be converted to prostaglandin $F_{2\alpha}$ by *CBR1* (2015; Malátková et al., 2010). *HPGDS* is responsible for the conversion of prostaglandin H_2 to prostaglandin D_2. *AKR1C3* can reduce prostaglandin H_2 and D_2 to prostaglandin $F_{2\alpha}$ and 11β-prostaglandin $F_{2\alpha}$, respectively (Penning, 2014). Finally, *HPGD* inactivates prostaglandin D_2, E_2, and $F_{2\alpha}$ by conversion to their respective dehydrogenated forms (Schneider and Pozzi, 2011). The third pathway of arachidonic acid metabolism is the lipoxygenase (LOX) pathway. Manual review of mutation-associated genes revealed that four genes encode for reactions downstream of arachidonic acid. On one hand, three genes are involved in the metabolism of two compounds derived from leukotriene A_4, which is itself derived from arachidonic acid, namely leukotriene B_4 and C_4. *CYP4F3* and *PTGR1* catalyze the inactivation of leukotriene B_4 either by ω-oxidation or via the 12HDH/15oPGR pathway respectively (Murphy and Gijón, 2007). GGT6 is involved in the conversion of leukotriene C_4 to leukotriene D_4 (Murphy and Gijón, 2007). On the other hand, one gene, *ALOX15*, catalyzes the direct synthesis from arachidonic acid of yet another class of LOX products, lipoxilins (Schneider and Pozzi, 2011). The reaction cluster 10 implicates reactions upstream of arachidonic acid. Manual review revealed a significant number of enzymes responsible for the cleavage of arachidonic acid from cellular lipids among the mutation-associated genes. *PLA2G2A*, *PLA2G4A*, *PLA2G4E*, and *PLA2G10* all belong to the class of phospholipases A_2 and function to release free fatty acids from the *sn*-2 position of phospholipids (Astudillo et al., 2012). Noteworthy, *PLA2G2A* shows an exquisite preference toward phospholipids containing arachidonic acid at the *sn*-2 position (Murphy and Gijón, 2007). *FAAH2* also affects arachidonic acid availability. Specifically, *FAAH2* degrades endogenous cannabinoid anandamide to release arachidonic acid (Wei et al., 2006). Finally, *ELOVL2* elongates selectively activated arachidonic acid (Ohno et al., 2010) and *MBOAT2* is involved in the Land's cycle to reincorporate activated arachidonic acid in the membrane lipids (Astudillo et al., 2012).

Next, we focused on xenobiotics metabolism, among the most enriched pathways in the high-convergence reaction sub-network. We first noticed that four genes overlap with the metabolism of arachidonic acid. *AKR1C3* (Penning, 2014), *CBR1* (Malátková et al., 2010), and *GSTM2* and *GSTM3* (Hayes et al., 2005) have also reported activity in the detoxification of

electrophilic xenobiotics. Reaction clusters 2, 9, and 14 implicate phase I of xenobiotics metabolism (also called functionalization). After manual review, we gathered a total of 22 genes involved in the functionalization phase. The great majority (20) are oxidoreductases in the family of cytochrome P450 (*CYP3A5*), alcohol dehydrogenases (*ADH1C*, *ADH6*, *ADH7*, *ADHFE1*), flavin-containing monoxygenases (*FMO3*, *FMO4*, *FMO5*), aldo-keto reductases (*AKR1B10*, *AKR1B15*, *AKR1C1*, *AKR1C2*), quinone reductases (*NQO1*, *NQO2*), carbonyl reductases (*CBR3*), aldehyde dehydrogenases (*ALDH3A1*, *ALDH3A2*, *ALDH3B1*), and amine oxidases (*AOC1*, *MAOB*) (Brožič et al., 2011; Quinn et al., 2008; Wermuth, 2003). The two remaining genes, *CES1* and *EPHX1*, belong instead to the class of hydrolases (Wermuth, 2003). Reaction cluster 1 implicates phase II of xenobiotics metabolism, also known as conjugation. Collectively, we found ten genes that can catalyze conjugation reactions among the mutation-associated genes. *UGT1A1* and *UGT1A6* are UDPGA transferases that carry glucuronidation reactions on xenobiotics (Wermuth, 2003). *GSTA2*, *GSTM1*, *GSTM4*, and *MGST1* catalyze the conjugation of glutathione (Hayes et al., 2005). *SULT1A1*, *SULT1A2*, and *SULT1A4* belong to the family of sulfotransferases and are responsible for sulfonation reactions on xenobiotics using 3′-phospho-5′-adenylyl sulfate (PAPS) as cofactor (Wermuth, 2003). *ACSL5* is a acyl-CoA synthetase that conjugates xenobiotic carboxylic acid by forming acyl-CoA thioesters (Wermuth, 2003).

Finally, we also observed five transporters for both arachidonic-acid-derived products and solubilized xenobiotics in the list of mutation-associated genes. The organic anion transporters *SLCO2A1* and *SLCO1B3* show affinity for prostaglandin D_2 and leukotriene C_4, respectively (Thiriet, 2012). The ABC transporters *ABCC1*, *ABCC2*, and *ABCC3* are renowned for their ability to move a variety of xenobiotics, but other substrates include prostaglandin A_1, A_2, D_2, E_2, 15d J_2, and leukotriene C_4 (Fletcher et al., 2010).

The enrichment for the occurrence of oxygen- and glutathione-consuming reactions in the high-convergence reaction sub-network persuaded us to investigate which other genes support their metabolism. Reaction clusters 6 and 13 feature two genes in glutathione metabolism, *GPX2* and *GPX3*. In addition, there are four more enzymes among the mutation-associated genes that are involved in glutathione biosynthesis, *GCLC*, *GCLM*, *GSR*, and *OPLAH* (Pompella et al., 2002). These expanded the list of glutathione-utilizing enzymes in the candidate pathway to a total of 15 members. In addition, several mutation-associated genes encode for reactions that use oxygen, most notably seven members of the cytochrome P450 (*CYP2W1*, *CYP4B1*, *CYP4X1*, *CYP24A1*, *CYP27A1*, *CYP27B1*, *CYP39A1*) and four others associated with at least two mutations: *HGD* participates in the metabolism of tyrosine; *CDO1* catabolizes cysteine and controls its cellular concentration; *CP* is a glycoprotein involved in iron ion homeostasis; and *MOXD1* is a monooxygenase of unknown substrate. These expanded the list of oxygen-utilizing reactions in the candidate pathway to a total of 21 members.

We neglected the result on the enrichment for the estrogen metabolism pathway because the associated genes were best explained by xenobiotics metabolism.

During the validation of our findings (see below), the increased statistical power allowed us to discover nine new mutation-associated genes that encode for reaction in or related to AraX. Six of these genes belong to arachidonic acid metabolism: *ALOX5* and *LTC4S* belong to the LOX pathway (Murphy and Gijón, 2007); *CYP2E1* belongs to the epoxygenase branch of the cytochrome P450 pathway Kroetz and Zeldin, 2002; *PLA2G6* and *PLA2G12A* are phospholipases A_2 involved in the release of arachidonic acid from the plasma membrane (Astudillo et al., 2012); and *PTGS2* encodes for the first step in the conversion of arachidonic acid to prostaglandins together with *PTGS1* (Schneider and Pozzi, 2011). The remaining three belong to xenobiotics metabolism: *FMO1* is a flavin-containing monoxygenase in the functionalization phase thioesters (Wermuth, 2003), while *GSTO1* and *GSTO2* belong to the conjugation phase (Wermuth, 2003).

The so-reconstructed candidate pathway features 27 genes attributable to arachidonic acid metabolism, 35 genes attributable to xenobiotics metabolism, 17 genes that mediate glutathione and oxygen metabolism, and five genes in the transport system. We reviewed each protein in this pathway in UniProt and/or Reactome to validate the gene annotation provided by literature (UniProt, 2015; Croft et al., 2014). In total, 84 metabolic genes are represented in this pathway. We termed this pathway AraX.

Enrichment of Pathways by Mutation-Associated Genes

We calculated the overrepresentation of AraX by each group of mutation-associated genes compared to any other KEGG pathway (186) or Reactome pathway (674), as retrieved in MSigDB, using the Fisher's exact test. The mean enrichment of a pathway across all mutations was subject to bootstrapping (10,000 replicates) in order to calculate the 95% confidence interval for the mean enrichment. This operation allowed evaluation of the robustness of a pathway mean enrichment to outliers (i.e., mutated genes strongly associated with a pathway).

Validation of the Generalized Linear Model and Mutation-Associated Genes

We performed differential gene-expression analysis, as described above, using the BS model on the validation cohort, consisting of 4,462 samples. The samples encompassed the same 13 cancer types as in the discovery cohort (range: 94–978 samples). We verified that the factors in the BS model featured at least 20 samples also in the validation cohort. As described above, the comparison of gene-wise BIC value between the onlyCT model and the BS model was used to determine whether the additional factors in the BS model provided a better goodness of fit also in the validation cohort. We sought to validate the list of genes associated with a mutated gene in the discovery cohort and their corresponding fold changes by linearly correlating them to the fold changes estimated using the BS model on the validation cohort. Finally, to prove that the expression changes associated with multiple mutated genes in the validation cohort indeed converge in the deregulation of AraX, we computed the overrepresentation of this pathway compared to any other KEGG or Reactome pathway as described earlier (see Enrichment of pathways by mutation-associated genes).

Survival Analysis

The deregulation at the level of gene expression for a metabolic pathway in a sample was estimated using Pathifier (Drier et al., 2013). This algorithm returns a score between 0 and 1 that represents the extent to which the expression of a pathway in a sample is deviating from the centroid pathway expression in normal samples. Hence, we calculated the score for all tumor samples in this study belonging to six cancer types for which matched normal samples were available in TCGA. These cancer types were breast invasive carcinoma, colon adenocarcinoma, head and neck squamous cell carcinoma, lung adenocarcinoma, lung squamous cell carcinoma, and uterine corpus endometrial carcinoma. The normal samples were used to provide the reference expression level of the pathway in a tissue.

We regressed the survival time until censoring or death to the AraX deregulation score for each sample in the discovery cohort (718 samples) to estimate whether AraX deregulation conferred a selective advantage to cancer evolution. We adopted as controls the same regression to the deregulation scores for other KEGG metabolic pathways (70) or for a gene set including 3,714 metabolic genes (*ALLM*). Then, we used a multivariate lasso penalized Cox regression model to calculate which metabolic pathway deregulation had the foremost effect in the prediction of survival, using as variables the deregulation score of the 70 KEGG metabolic pathways, AraX, and *ALLM*. The selection of variables relevant to predict survival was performed using increasing values for the lasso penalty (log-λ) used in the regression. The optimal penalty value was calculated such that the mean 10-fold cross-validated error was minimum. Out of 72 initial variables, only three variables were predictive of survival at the optimal penalty. To further rule out that a simple pathway deregulation is sufficient to predict poor prognosis, we performed univariate Cox regression of survival on the deregulation scores of any KEGG pathway (also non metabolic ones) with more than 100 genes and compared the Wald test statistic and log-hazard ratio per unit of deregulation score to the regression on AraX deregulation scores.

We determined whether poor prognosis could be predicted by the level of deregulation of AraX by equally splitting all samples in this study belonging to the six cancer types used above and with complete survival information (1,908 samples) into a discovery and validation sub-cohorts and stratifying the samples into low or high deregulation. The threshold score upon which a sample is classified as highly deregulated was computed in the discovery sub-cohort by using maximally selected rank statistics, which identifies a threshold score that maximizes the difference in survival between the two

groups and tests its statistical significance (Hothorn and Lausen, 2003). A robust threshold score was finally selected by repeating this computation using 1,000 bootstraps. Kaplan-Meier curves were generated for the two groups, and the significance of survival difference was estimated using the Wald test. The validity of the threshold score and the difference in survival between the two groups were verified in the validation sub-cohort. The difference in survival according to the low versus high stratification was finally computed using the Wald test leveraging on all samples, and the corresponding statistic was tested against sub-sampling using 10,000 bootstraps and random sample label permutation using 10,000 permutations.

Due to missing clinical information in a non-negligible number of samples, we verified the independency of AraX deregulation from other prognostic clinical features individually, by performing a statistical test of dependency in the subset of samples where the information was reported. We tested a correlation between low versus high AraX deregulation and age using the Wilcoxon rank-sum test in 1,343 samples and with metastatic status using the Fisher's exact test in 351 samples. We tested an association between the AraX deregulation scores and the tumor stages within each cancer type using the likelihood ratio test, in a number of samples ranging 48 to 132 depending on the cancer type (endometrial cancer was excluded because no samples were annotated with tumor stage information). We tested whether the distribution of AraX deregulation scores are cancer type dependent using a likelihood ratio test. The correlation between the cancer-type-specific distribution of AraX deregulation scores and the 5-year survival for cancers of the corresponding tissue (as retrieved from https://nccd.cdc.gov/uscs/Survival/Relative_Survival_Tables.pdf) was tested using a likelihood ratio test. The significance of the univariate regression of survival in a given cancer type and low versus high AraX deregulation (according to the threshold score identified earlier in the pan-cancer cohort) was tested using the Wald test. A power analysis for this test at a confidence level $\alpha = 0.01$ was conducted by sub-sampling the pan-cancer cohort into sizes ranging from 100 to 1,900 samples 1,000 times, and by counting the percent of times that a significant association between survival and AraX deregulation was found.

SUPPLEMENTAL INFORMATION

Supplemental Information includes 13 figures and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2016.06.038.

AUTHOR CONTRIBUTIONS

Methodology, Validation, Formal Analysis, Data Curation, Writing – Original Draft, and Visualization, F.G.; Conceptualization, F.G. and J.N.; Supervision, A.S. and J.N.; Writing – Review and Editing, F.G., A.S. and J.N.; Funding Acquisition, J.N.

ACKNOWLEDGMENTS

The authors would like to acknowledge the following people for their contributions to this study: Erik Kristiansson's lab and Rebecka Jörnsten in Chalmers University of Technology for support in biostatistics; Pontus Hjortskog, David Jensen, Kenny Nilsson, and Luuk van Egeraat for support in the data retrieval and storage; Martin Eilers in University of Würzburg for discussion; Adil Mardinoglu, Ivan Mijakovic, and Leif Väremo for a critical review of the manuscript. The computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at C3SE. F.G. and J.N. acknowledge the Knut and Alice Wallenberg Foundation for financing this work.

Received: May 11, 2016
Revised: May 13, 2016
Accepted: June 5, 2016
Published: July 7, 2016

REFERENCES

Agren, R., Bordel, S., Mardinoglu, A., Pornputtapong, N., Nookaew, I., and Nielsen, J. (2012). Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. PLoS Comput. Biol. *8*, e1002518.

Arnold, C., Konkel, A., Fischer, R., and Schunck, W.H. (2010). Cytochrome P450-dependent metabolism of omega-6 and omega-3 long-chain polyunsaturated fatty acids. Pharmacol. Rep. *62*, 536–547.

Astudillo, A.M., Balgoma, D., Balboa, M.A., and Balsinde, J. (2012). Dynamics of arachidonic acid mobilization by inflammatory cells. Biochim. Biophys. Acta *1821*, 249–256.

Beisser, D., Klau, G.W., Dandekar, T., Müller, T., and Dittrich, M.T. (2010). BioNet: an R-Package for the functional analysis of biological networks. Bioinformatics *26*, 1129–1130.

Brožič, P., Turk, S., Rižner, T.L., and Gobec, S. (2011). Inhibitors of aldo-keto reductases AKR1C1-AKR1C4. Curr. Med. Chem. *18*, 2554–2565.

Bui, P., Imaizumi, S., Beedanagari, S.R., Reddy, S.T., and Hankinson, O. (2011). Human CYP2S1 metabolizes cyclooxygenase- and lipoxygenase-derived eicosanoids. Drug Metab. Dispos. *39*, 180–190.

Cairns, R.A., Harris, I.S., and Mak, T.W. (2011). Regulation of cancer cell metabolism. Nat. Rev. Cancer *11*, 85–95.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. *2*, 401–404.

Chamorro, M.N., Schwartz, D.R., Vonica, A., Brivanlou, A.H., Cho, K.R., and Varmus, H.E. (2005). FGF-20 and DKK1 are transcriptional targets of beta-catenin and FGF-20 is implicated in cancer and development. EMBO J. *24*, 73–84.

Chin, L., Andersen, J.N., and Futreal, P.A. (2011). Cancer genomics: from discovery science to personalized medicine. Nat. Med. *17*, 297–303.

Chuang, S.S., Helvig, C., Taimi, M., Ramshaw, H.A., Collop, A.H., Amad, M., White, J.A., Petkovich, M., Jones, G., and Korczak, B. (2004). CYP2U1, a novel human thymus- and brain-specific cytochrome P450, catalyzes omega- and (omega-1)-hydroxylation of fatty acids. J. Biol. Chem. *279*, 6305–6314.

Ciriello, G., Miller, M.L., Aksoy, B.A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. Nat. Genet. *45*, 1127–1133.

Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., et al. (2014). The Reactome pathway knowledgebase. Nucleic Acids Res. *42*, D472–D477.

DeNicola, G.M., Karreth, F.A., Humpton, T.J., Gopinathan, A., Wei, C., Frese, K., Mangal, D., Yu, K.H., Yeo, C.J., Calhoun, E.S., et al. (2011). Oncogene-induced Nrf2 transcription promotes ROS detoxification and tumorigenesis. Nature *475*, 106–109.

Drier, Y., Sheffer, M., and Domany, E. (2013). Pathway-based personalized analysis of cancer. Proc. Natl. Acad. Sci. USA *110*, 6388–6393.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191.

Fletcher, J.I., Haber, M., Henderson, M.J., and Norris, M.D. (2010). ABC transporters in cancer: more than just drug efflux pumps. Nat. Rev. Cancer *10*, 147–156.

Fodde, R., Edelmann, W., Yang, K., van Leeuwen, C., Carlson, C., Renault, B., Breukel, C., Alt, E., Lipkin, M., Khan, P.M., et al. (1994). A targeted chain-termination mutation in the mouse Apc gene results in multiple intestinal tumors. Proc. Natl. Acad. Sci. USA *91*, 8969–8973.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. *33*, 1–22.

Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. *6*, pl1.

Gatto, F., and Nielsen, J. (2016). In search for symmetries in the metabolism of cancer. Wiley Interdiscip. Rev. Syst. Biol. Med. *8*, 23–35.

Gatto, F., Nookaew, I., and Nielsen, J. (2014). Chromosome 3p loss of heterozygosity is associated with a unique metabolic network in clear cell renal carcinoma. Proc. Natl. Acad. Sci. USA *111*, E866–E875.

Gerlinger, M., McGranahan, N., Dewhurst, S.M., Burrell, R.A., Tomlinson, I., and Swanton, C. (2014). Cancer: evolution within a lifetime. Annu. Rev. Genet. *48*, 215–236.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

Hashibe, M., McKay, J.D., Curado, M.P., Oliveira, J.C., Koifman, S., Koifman, R., Zaridze, D., Shangina, O., Wünsch-Filho, V., Eluf-Neto, J., et al. (2008). Multiple ADH genes are associated with upper aerodigestive cancers. Nat. Genet. *40*, 707–709.

Hayes, J.D., Flanagan, J.U., and Jowsey, I.R. (2005). Glutathione transferases. Annu. Rev. Pharmacol. Toxicol. *45*, 51–88.

Hoadley, K.A., Yau, C., Wolf, D.M., Cherniack, A.D., Tamborero, D., Ng, S., Leiserson, M.D., Niu, B., McLellan, M.D., Uzunangelov, V., et al.; Cancer Genome Atlas Research Network (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell *158*, 929–944.

Hofree, M., Shen, J.P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. Nat. Methods *10*, 1108–1115.

Hopkins, A.L. (2008). Network pharmacology: the next paradigm in drug discovery. Nat. Chem. Biol. *4*, 682–690.

Hothorn, T., and Lausen, B. (2003). On the exact distribution of maximally selected rank statistics. Comput. Stat. Data Anal. *43*, 121–137.

Hu, J., Locasale, J.W., Bielas, J.H., O'Sullivan, J., Sheahan, K., Cantley, L.C., Vander Heiden, M.G., and Vitkup, D. (2013). Heterogeneity of tumor-induced gene expression changes in the human metabolic network. Nat. Biotechnol. *31*, 522–529.

Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A.F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics *18* (*Suppl 1*), S233–S240.

Johnson, L., Mercer, K., Greenbaum, D., Bronson, R.T., Crowley, D., Tuveson, D.A., and Jacks, T. (2001). Somatic activation of the K-ras oncogene causes early onset lung cancer in mice. Nature *410*, 1111–1116.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333–339.

Kroetz, D.L., and Zeldin, D.C. (2002). Cytochrome P450 pathways of arachidonic acid metabolism. Curr. Opin. Lipidol. *13*, 273–283.

Kumar, S. (2015). Computational identification and binding analysis of orphan human cytochrome P450 4X1 enzyme with substrates. BMC Res. Notes *8*, 9.

Lara, M.F., García-Escudero, R., Ruiz, S., Santos, M., Moral, M., Martínez-Cruz, A.B., Segrelles, C., Lorz, C., and Paramio, J.M. (2008). Gene profiling approaches help to define the specific functions of retinoblastoma family in epidermis. Mol. Carcinog. *47*, 209–221.

Law, C.W., Chen, C., Shi, W., and Smyth, G.K. (2013). Voom! precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. *15*, R29.

Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. *15*, R29.

Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495–501.

Malátková, P., Maser, E., and Wsól, V. (2010). Human carbonyl reductases. Curr. Drug Metab. *11*, 639–658.

Malhotra, D., Portales-Casamar, E., Singh, A., Srivastava, S., Arenillas, D., Happel, C., Shyr, C., Wakabayashi, N., Kensler, T.W., Wasserman, W.W., and Biswal, S. (2010). Global mapping of binding sites for Nrf2 identifies novel targets in cell survival response through ChIP-Seq profiling and network analysis. Nucleic Acids Res. *38*, 5718–5734.

Mardinoglu, A., Agren, R., Kampf, C., Asplund, A., Uhlen, M., and Nielsen, J. (2014). Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. Nat. Commun. *5*, 3083.

McKay, J.D., Truong, T., Gaborieau, V., Chabrier, A., Chuang, S.C., Byrnes, G., Zaridze, D., Shangina, O., Szeszenia-Dabrowska, N., Lissowska, J., et al. (2011). A genome-wide association study of upper aerodigestive tract cancers conducted within the INHANCE consortium. PLoS Genet. *7*, e1001333.

Murphy, R.C., and Gijón, M.A. (2007). Biosynthesis and metabolism of leukotrienes. Biochem. J. *405*, 379–395.

Nebert, D.W., and Dalton, T.P. (2006). The role of cytochrome P450 enzymes in endogenous signalling pathways and environmental carcinogenesis. Nat. Rev. Cancer *6*, 947–960.

Nilsson, R., Jain, M., Madhusudhan, N., Sheppard, N.G., Strittmatter, L., Kampf, C., Huang, J., Asplund, A., and Mootha, V.K. (2014). Metabolic enzyme expression highlights a key role for MTHFD2 and the mitochondrial folate pathway in cancer. Nat. Commun. *5*, 3128.

Ohno, Y., Suto, S., Yamanaka, M., Mizutani, Y., Mitsutake, S., Igarashi, Y., Sassa, T., and Kihara, A. (2010). ELOVL1 production of C24 acyl-CoAs is linked to C24 sphingolipid synthesis. Proc. Natl. Acad. Sci. USA *107*, 18439–18444.

Penning, T.M. (2014). The aldo-keto reductases (AKRs): Overview. Chem Biol Interact.

Podsypanina, K., Ellenson, L.H., Nemes, A., Gu, J., Tamura, M., Yamada, K.M., Cordon-Cardo, C., Catoretti, G., Fisher, P.E., and Parsons, R. (1999). Mutation of Pten/Mmac1 in mice causes neoplasia in multiple organ systems. Proc. Natl. Acad. Sci. USA *96*, 1563–1568.

Pompella, A., Bánhegyi, G.b., and Wellman-Rousseau, M. (2002). Thiol Metabolism and Redox Regulation of Cellular Functions (IOS Press).

Quinn, A.M., Harvey, R.G., and Penning, T.M. (2008). Oxidation of PAH transdihydrodiols by human aldo-keto reductase AKR1B10. Chem. Res. Toxicol. *21*, 2207–2215.

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C.E., Socci, N.D., and Betel, D. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol. *14*, R95.

Sasaki, M., Knobbe, C.B., Munger, J.C., Lind, E.F., Brenner, D., Brüstle, A., Harris, I.S., Holmes, R., Wakeham, A., Haight, J., et al. (2012). IDH1(R132H) mutation increases murine haematopoietic progenitors and alters epigenetics. Nature *488*, 656–659.

Schneider, C., and Pozzi, A. (2011). Cyclooxygenases and lipoxygenases in cancer. Cancer Metastasis Rev. *30*, 277–294.

Schulze, A., and Harris, A.L. (2013). How cancer metabolism is tuned for proliferation and vulnerable to disruption (vol 491, pg 364, 2012). Nature *494*, 130–130.

Sharma, A.K., Eils, R., and König, R. (2016). Copy number alterations in enzyme-coding and cancer-causing genes reprogram tumor metabolism. Cancer Res. Published online May 23, 2016. http://dx.doi.org/10.1158/0008-5472.CAN-15-2350.

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. *3*, Article3.

Stark, K., Dostalek, M., and Guengerich, F.P. (2008). Expression and purification of orphan cytochrome P450 4X1 and oxidation of anandamide. FEBS J. *275*, 3706–3717.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

Thiriet, M. (2012). Signaling at the Cell Surface in the Circulatory and Ventilatory Systems (Springer).

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. B *58*, 267–288.

Trachootham, D., Alexandre, J., and Huang, P. (2009). Targeting cancer cells by ROS-mediated mechanisms: a radical therapeutic approach? Nat. Rev. Drug Discov. *8*, 579–591.

UniProt Consortium (2015). UniProt: a hub for protein information. Nucleic Acids Res. *43*, D204–D212.

Väremo, L., Nielsen, J., and Nookaew, I. (2013). Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucleic Acids Res. *41*, 4378–4391.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546–1558.

Wagenmakers, E.-J., and Farrell, S. (2004). AIC model selection using Akaike weights. Psychon. Bull. Rev. *11*, 192–196.

Wang, D., and Dubois, R.N. (2010). Eicosanoids and cancer. Nat. Rev. Cancer *10*, 181–193.

Wang, Y., Eddy, J.A., and Price, N.D. (2012). Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. BMC Syst. Biol. *6*, 153.

Wang, J., Wei, J., Xu, X., Pan, W., Ge, Y., Zhou, C., Liu, C., Gao, J., Yang, M., and Mao, W. (2014). Replication study of ESCC susceptibility genetic polymorphisms locating in the ADH1B-ADH1C-ADH7 cluster identified by GWAS. PLoS ONE *9*, e94096.

Wei, B.Q., Mikkelsen, T.S., McKinney, M.K., Lander, E.S., and Cravatt, B.F. (2006). A second fatty acid amide hydrolase with variable distribution among placental mammals. J. Biol. Chem. *281*, 36569–36578.

Wei, S., Liu, Z., Zhao, H., Niu, J., Wang, L.E., El-Naggar, A.K., Sturgis, E.M., and Wei, Q. (2010). A single nucleotide polymorphism in the alcohol dehydrogenase 7 gene (alanine to glycine substitution at amino acid 92) is associated with the risk of squamous cell carcinoma of the head and neck. Cancer *116*, 2984–2992.

Weinberg, R.A. (2014). Coming full circle-from endless complexity to simplicity and back again. Cell *157*, 267–271.

Wermuth, C.G. (2003). The Practice of Medicinal Chemistry (Academic).

Yan, X., and Su, X. (2009). Linear Regression Analysis: Theory and Computing (World Scientific).

Yates, L.R., and Campbell, P.J. (2012). Evolution of the cancer genome. Nat. Rev. Genet. *13*, 795–806.

Zambelli, F., Prazzoli, G.M., Pesole, G., and Pavesi, G. (2012). Cscan: finding common regulators of a set of genes by using a collection of genome-wide ChIP-seq datasets. Nucleic Acids Res. *40*, W510–W515.