Efficient Representations of Tumor Diversity with Paired DNA-RNA Anomalies

Qian Ke[1,*], Wikum Dinalankara[2,*], Laurent Younes[1,a], Donald Geman[1,a], and Luigi Marchionni[2,a]

[1]Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD, USA

[2]Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[*] Equally contributing authors

[a] Correspondence to: Luigi Marchionni (marchion@jhu.edu), Laurent Younes (laurent.younes@jhu.edu), and Donald Geman (geman@jhu.edu)

May 25, 2020

Abstract. Cancer cells display massive dysregulation of key regulatory pathways due to now well-catalogued mutations and other DNA-related aberrations. Moreover, enormous heterogeneity has been commonly observed in the identity, frequency, and location of these aberrations across individuals with the same cancer type or subtype, and this variation naturally propagates to the transcriptome, resulting in myriad types of dysregulated gene expression programs. Many have argued that a more integrative and quantitative analysis of heterogeneity of DNA and RNA molecular profiles may be necessary for designing more systematic explorations of alternative therapies and improving predictive accuracy.

We introduce a representation of multi-omics profiles which is sufficiently rich to account for observed heterogeneity and support the construction of quantitative, integrated metrics of variation. Starting from the network of interactions existing in Reactome, we build a library of "paired DNA-RNA anomalies" that represent prototypical and recurrent patterns of dysregulation in cancer; each two-gene "motif" consists of a "source" regulatory gene and a "target" gene whose expression is "controlled" by the source gene. The pair motif is then "active" or "realized" in a joint DNA-RNA profile if the source gene is DNA-aberrant (*e.g.*, mutated, deleted, or duplicated), and the downstream target gene is "RNA-divergent", meaning its expression level is outside the normal, baseline range. With M motifs, each sample profile has exactly one of the 2^M possible configurations.

We focus our analyses on reduced configurations by selecting tissue-dependent minimal coverings that we define as the smallest family of motifs with the property that every sample in the considered population displays at least one active motif; these minimal coverings can be computed with integer programming. Given such a covering, a natural measure of cross-sample diversity is the extent to which the particular active motifs vary from sample to sample; this variability is captured by the entropy of the distribution over configurations.

We apply this program to data from TCGA for six distinct tumor types (breast, prostate, lung, colon, liver, and kidney cancer). This enables an efficient simplification of the complex landscape observed in cancer populations, resulting in the identification of novel signatures of molecular alterations which are not detected with frequency-based criteria. Estimates of cancer heterogeneity across tumor phenotypes reveals a stable pattern: entropy increases with disease severity. This framework is then well-suited to accommodate the expanding complexity of cancer genomes and epigenomes emerging from large consortia projects.

Author Summary. A large variety of genomic and transcriptomic anomalies are observed in cancer cells, and their identity, location, and frequency can be highly indicative of the particular subtype or molecular phenotype, and thereby inform treatment options. However, elucidating this association between sets of anomalies and subtypes of cancer is severely impeded by considerable diversity in the set of anomalies across samples from the same population. Most attempts at analyzing tumor heterogeneity have dealt with either the genome or transcriptome in isolation. Here we present a novel, multi-omics approach for quantifying heterogeneity by determining a small set of paired DNA-RNA anomalies that incorporates potential downstream effects on gene expression. We apply integer programming to identify a small set of paired anomalies such that at least one anomaly is present in every sample of a given cancer population. The resulting "coverings" are analyzed for six cancer cohorts from the Cancer Genome Atlas, and facilitates introducing an information-theoretic measure of heterogeneity. Our results identify many known facets of tumorigenesis as well as suggest potential novel genes of interest.

Data Availability Statement. RNA-Seq data, somatic mutation data and copy number data for The Cancer Genome Atlas were obtained through the Xena database from individual cancer type cohorts. Selected tissue level coverings are also provided in excel format as additional supplementary material.

1 Introduction

Cancer cells evade the normal mechanisms controlling cellular growth and tissue homeostasis through the disruption of key regulatory pathways controlling these processes. Such dysregulation results from genetic and epigenetic aberrations, encompassing mutations, copy number alterations, and changes in chromatin states, which affect the genes participating in such regulatory networks.

Over the past several decades, the list of known genetic and genomic aberrations in cancer has greatly expanded, thanks to large-scale projects such as the The Cancer Genome Atlas (TCGA, (Cancer Genome Atlas Research Network et al., 2013), the Catalogue Of Somatic Mutations In Cancer (COSMIC, Tate et al. (2019)), the MSK/IMPACT study (Zehir et al., 2017), and recents efforts from the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020).

Whereas the number of aberrations which suffice for progression to an advanced cancer is thought to be rather small, at least for solid tumors (Vogelstein and Kinzler, 2015; Tomasetti et al., 2015) and at the pathway level (Sever and Brugge, 2015), the number of ways (combinations of aberrations) for which this can be actualized is very large. In particular, the landscape collectively emerging from these studies exhibits a high degree of variation in the identity, frequency, and location of these aberrations, as well as tissue- and expression-dependency (Haigis et al., 2019; Lawrence et al., 2013). These differences – collectively referred to as *tumor heterogeneity* – are "context-specific", differing among tissue types and epigenetic conditions (Haigis et al., 2019), across different cells within a lesion (*intra-tumor heterogeneity*), between tumor lesions within the same individual (*inter-tumor heterogeneity*), and across distinct individuals with the same cancer type or sub-type (*across-sample or population-level heterogeneity*).

In addition, such DNA defects, in order to be "functional" (*i.e.*, manifest themselves) and ultimately alter the cellular phenotype, must propagate through the signaling and regulatory network and alter the downstream gene expression programs (Osmanbeyoglu et al., 2017; PCAWG Transcriptome Core Group et al., 2020). These downstream transcriptional changes are in fact also context-specific, varying within and among cancers, local environments, and individuals. Most importantly, it has been speculated that transcriptionally heterogeneous tumors may be more adaptable to changes in the tumor microenvironment and therefore more likely to acquire new properties such as metastatic potential and resistance to treatments, leading to dismal patient outcomes; in addition, for predicting the response to targeted therapies, gene expression profiles may be more discriminating than mutational status (Costello et al., 2014). The analysis of heterogeneity of molecular profiles, both DNA and RNA, is therefore of paramount importance. Consequently, a deeper, integrative and quantitative analysis of tumor heterogeneity is necessary for achieving a better understanding of the underlying biology, for designing more systematic explorations of candidate therapies, and for improving the accuracy of prognosis and treatment response predictions.

Unsurprisingly, even *representing* such high-dimensional variability poses great challenges, especially if a major goal is to find suitable metrics to quantify the level of tumor heterogeneity. We assume that large-scale projects (see above) and studies (*e.g.*, Bailey et al. (2018)) have already provided reasonably comprehensive lists of the most important recurrent molecular alterations driving cancer initiation and progression. But merely counting or cataloging aberrations will not suffice to precisely measure heterogeneity in a tumor population, and to quantify how this differs across diverse contexts (*e.g.*, between cancer arising in distinct organs, or between tumor sub-types). In order

to identify functional aberrations potentially exploitable as biomarkers and therapeutic targets, it is necessary to go well beyond frequency estimates to more powerful representations rooted in biological mechanism and accounting for statistical dependency among aberrations.

We introduce a representation of omics profiles which is sufficiently rich to account for observed heterogeneity and to support the construction of quantitative, integrated metrics. Our framework is centered on the joint analysis of "paired DNA-RNA aberrations" that represent prototypical and recurrent patterns of dysregulation in cancer. Specifically, we represent the space of gene alterations that result in network perturbations and downstream changes of gene expression levels as a catalogue of mechanistic, two-gene "motifs," each consisting of a "source" gene and a "target" gene for which the mRNA expression is controlled either directly by the source gene or indirectly by a close descendant of the source.

Other approaches for coupling DNA and RNA alterations and filtering out "non-functional" events have been described (Curtis et al., 2012; Ciriello et al., 2013; Osmanbeyoglu et al., 2014). Here, we extend motifs from a network property to a sample property (like the existence of individual aberrations) by declaring a motif to be "active" or "realized" in a joint DNA-RNA profile if the source gene is DNA-aberrant (*e.g.*, mutated, deleted, or duplicated), and the target gene is RNA-aberrant, meaning its expression level is "divergent" (*i.e.*, outside the normal, baseline range (Dinalankara et al., 2018)). This defines one binary random variable per motif, of which there are typically hundreds of thousands, most of which have a very small probability to be realized in a sample (see Figure 1). Examples of motifs which are active in multiple TCGA samples are "*APC* controls *AXIN2*, *APC* is mutated and *AXIN2* is divergent" in colon cancer; and "*PTEN* controls *TWIST1*, *PTEN* is deleted and *TWIST1* is divergent" in prostate cancer.

Samples are then characterized by their entire set of paired DNA-RNA anomalies. Therefore, given there are M motifs, exactly one of the 2^M possible configurations is assigned to each sample. The extent to which these subsets vary from sample to sample is then a natural measure of heterogeneity in the population from which the samples are drawn. More formally, cross-sample heterogeneity might then be *defined* as the entropy of P, the probability distribution over configurations, *i.e.*, the random vector of the binary states of each motif. However, M is far too large to estimate anything from real data except low-dimensional properties of P corresponding to simple functions of the bit string, such as the marginal probabilities of each motif separately or the distribution of the total number of active motifs. Due to the difficulty of estimating rare events with the modest sample sizes available in cancer genomics today, any multivariate property of P, for example, its entropy, cannot be accurately approximated without a substantial further reduction of complexity.

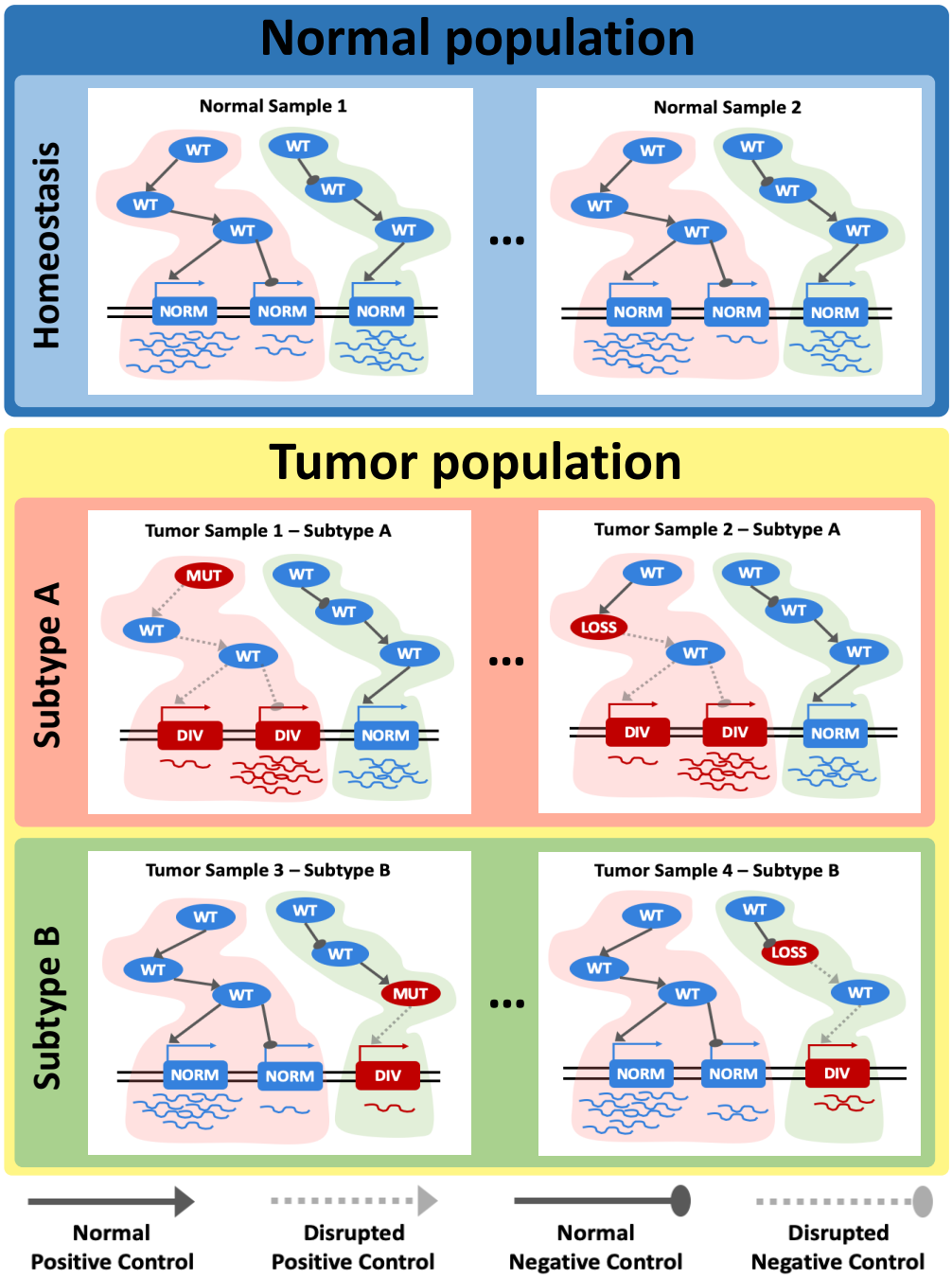

Figure 1: DNA-RNA paired anomalies in normal and tumor populations. In homeostatic conditions, normal cells in each tissue exert their biological functions by running specific gene expression programs, which are tightly regulated by upstream networks and signaling pathways. In cancer cells, as the consequence of driving genomic and epigenomic events (*e.g.*, mutations and copy number alterations), key regulatory proteins in such networks cease to function properly, with consequences that propagate downstream, resulting in profound modifications of the controlled gene expression programs (*i.e.*, expression levels depart from the normal range, or become " divergent"). Importantly, distinct sets of alterations might be present in different tumor samples, affecting distinct portions of the network, which in turn result in specific aberrant expression profiles driving different tumor phenotypes. In the figure, sub-type A tumors are characterized by paired DNA-RNA anomalies affecting the signaling pathway on the left, while sub-type B tumors are hit on the signaling pathway on the right. Notably, different subsets of paired anomalies are "realized" in the different tumors.

Such a reduction is provided by the concept of *minimal coverings* of a population (previously employed for modeling networks (Hristov and Singh, 2017)). Here, we focus on the smallest collection C of paired anomalies with the property that (nearly) every tumor sample has at least one active motif in C. Indeed, since nearly all tumor samples exhibit at least one or more paired anomaly, a *minimal covering* necessarily exists (perhaps not unique), which can be found using well-known algorithms for formulating "optimal set covering" as the solution of an integer-programming problem (see Methods). Importantly, the motifs in these sets can be indexed at the "source level" or "target level" meaning, respectively, that the conditions for a realized pair anomaly are satisfied by *some* target gene linked to a given source gene (source level index) or *some* source gene linked to a given target gene (target level index). The sizes of the resulting coverings for common tissues (breast, colon, kidney, lung and prostate) are of the order 10–100.

Using coverings, we are able represent each sample in a concise and interpretable way. Moreover, we are able to analyze heterogeneity in tumor populations for given tissues and subtypes, which was not feasible in the original configuration space. Our analyses include the computation of entropy and their comparison across sub-types, and summarized representations of configurations through conjunctions of genomic events. While minimal set coverings are typically not unique, we were able to exhibit "core motifs," representing anomalies that appear in every solution and turn out to provide an interesting list of the most salient events associated with tumors of a given type.

Our methods are described in more details in the next sections, followed by a presentation of our results. We conclude this paper with a discussion and provide additional results in supplementary material.

2 Methods

2.1 Overall Strategy

Identifying and quantifying the cross-sample heterogeneity of datasets consisting of *omics* profiles with tens of thousands of random variables representing the genotypes or mRNA expression of tumor samples is a daunting task. In particular, it can only be achieved by making simplifying assumptions and approximations on the joint distribution of the considered variables. This requirement is reflected in our multi-step analysis (see Figure 2), which relies on the principles described below.

We performed our analyses using matched DNA mutation, copy number alterations, and RNA expression data, pre-processed as previously described (Dinalankara et al., 2018). In the present study we focused on six distinct tumor types (in parenthesis the TCGA code): breast invasive carcinoma (BRCA), prostate adenocarcinoma (PRAD), lung adenocarcinoma (LUAD), liver hepatocellular carcinoma (LIHC), kidney renal clear cell carcinoma (KIRC), and colon adenocarcinoma (COAD). For simplicity, hereafter, we will refer to these tumor types according to the organ of origin (breast, prostate, lung, kidney, liver, and colon cancer).

Anomaly detection. In order to focus the analysis on the heterogeneity resulting from cancer and disregard variation in the normal population, we reduce the data to binary variables indicating deviations from normal behavior. The resulting indicators are furthermore filtered using a motif-based analysis requiring plausible mechanisms leading to the anomaly. This operation is described in Section 2.2

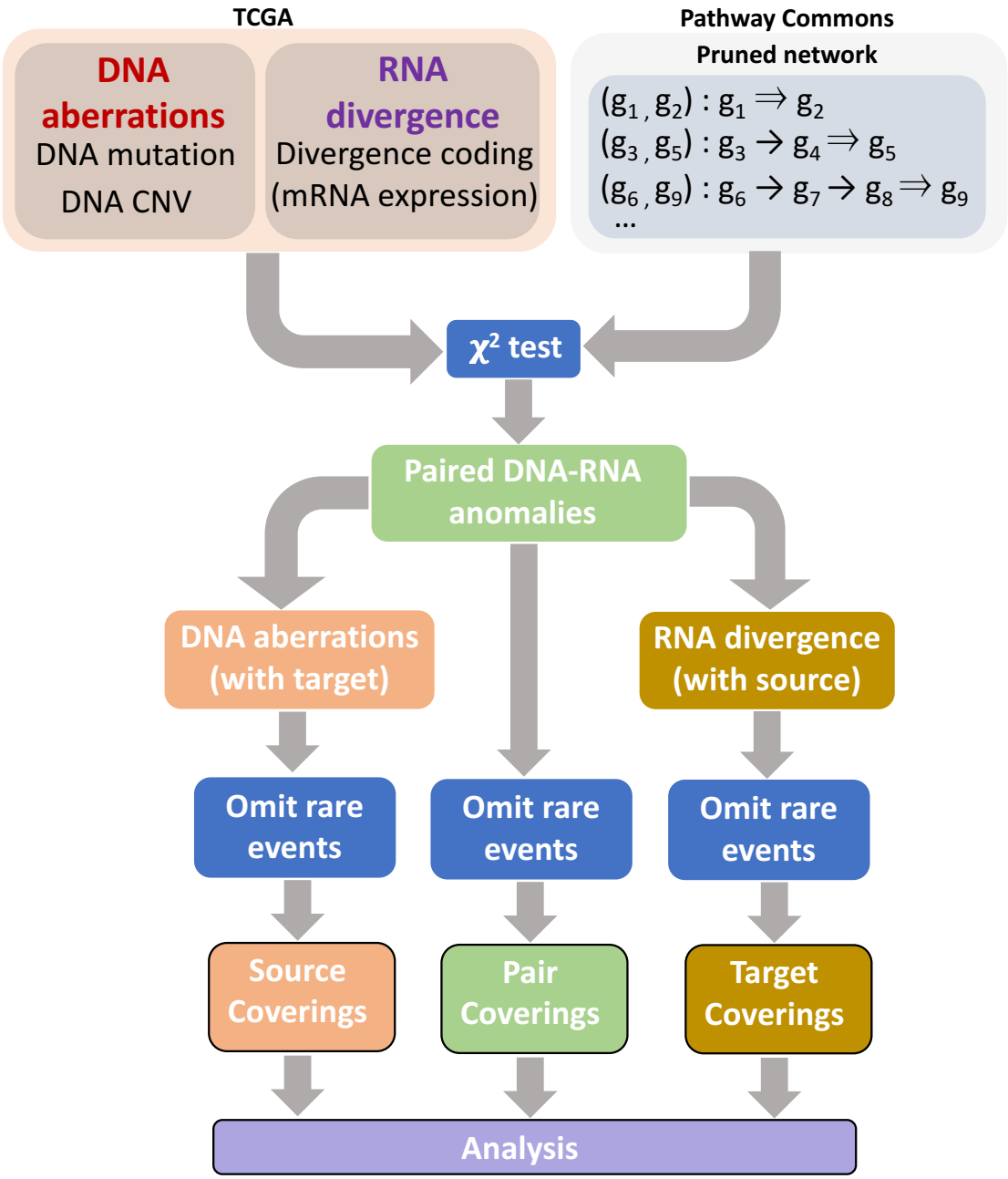

Figure 2: Overall analytical workflow. Pair motifs are constructed using the links available in Reactome (Jassal et al., 2020). In the TCGA cancer cohorts, the mutation and copy number variation data are used to construct binary DNA aberration profiles; the presence of either a mutation or high/low copy number variation at a given gene is treated as an aberration for the given gene for that sample *omics* profile. The gene expression data are used to construct binary RNA divergence profiles based on falling outside the "normal" expression range (in quantiles) for each gene based on TCGA normal tissue expression data, as previously described (Dinalankara et al., 2018). The binary profiles are combined to produce paired DNA-RNA anomalies, following which filtering is performed by selecting pairs that are determined to be significant (two-sided χ^2 test). These selected pair motifs are then used to construct source (DNA) and target (RNA) motifs, providing binary *omics* profiles at the level of source, target, and pairs. Motifs that are present in less than 2% of samples for a given tissue are omitted. Then coverings are computed at the pair, source and target levels and subtype analysis and heterogeneity analysis carried out.

Covering estimation. In order to reduce the number of variables under consideration, we estimate subsets of "important variables," called coverings, defined as minimal sets of variables from which, with high probability, cancer samples have at least one abnormal deviant observation. Our approach to computing these coverings is described in section 2.3. Because such optimal coverings are generally not unique, this representation may take various flavors, including the consideration of variables that are present in at least one of them (union), or the restriction to variables that appear in all of them (intersection), that we refer to as "core" variables, or the use of a single covering, for example the one maximizing the sum of frequencies of anomalies within its variables.

Entropy estimation. We assess the heterogeneity of a population of samples by computing the entropy of the reduced profiles computed within this population. This computation is not straightforward; even though reduced profiles involve a relatively small number k of binary variables (typically a few dozen), the observed sample size remains insufficient to allow for the estimation of the probabilities of the 2^k joint configurations of these variables. Some approximations are necessary and are described in section 2.4.

Code-based reduction. Using a tree-based decomposition, we decompose tumor samples resulting from a given cancer type into cells, or bins, associated with a small number of conjunctions and disjunctions of anomalies. It is them possible to visualize and compare the resulting histograms in sub-populations defined by specific subtypes or phenotypes. This is described in section 2.5.

2.2 Anomalies

2.2.1 Univariate deviation from normality in *omics* data

We transform the original data into sparse binary vectors indicating whether each variable deviates from a reference state or normal range when observed on a given sample. Our pre-processing of DNA data, which is already providing deviation from wild type, is quite simple and we consider that a gene is aberrant at the DNA level if it includes a mutation that differs from the wild type or if its copy number variation is equal to -2, or is larger than 2. We will write Z_g^{dna} for the corresponding binary random variable, so $Z_g^{dna} = 1$ when g is DNA-aberrant.

The binarization of RNA data is more involved, and it based on the notion of "divergence" we previously developed (Dinalankara et al., 2018). Roughly speaking, a gene is declared as RNA-divergent if its ranking among other genes in the same sample falls outside of its normal range estimated from baseline data. Let Z_g^{rna} be the corresponding binary random variable. This dichotomization requires a training step, solely based on normal tissue data, in order to estimate these normal intervals of variation. This being done, the decision for a gene to be RNA-divergent in a tumor sample only involves the RNA profile of this sample and is in particular independent of other tumor observations in the dataset.

2.2.2 Building motifs

These binary *omics* variables are filtered by requiring that the deviations they represent have a plausible causal explanation as parts of two-genes "motifs." Such motifs, denoted $g^{(source)} \Rightarrow g^{(target)}$, are built using the Reactome pathway database (Jassal et al., 2020) as retrieved from Pathway Commons (Cerami et al., 2011). We use gene interaction links provided by this database, focusing on links $g \longrightarrow g'$ for which "g controls the state of g'" (notation: $g \xrightarrow{state} g'$) or "g controls the

expression of g'" (notation $g \xrightarrow{expr} g'$). We say that two genes $g^{(source)}, g^{(target)}$ form a "pair motif," with notation $g^{(source)} \Rightarrow g^{(target)}$ is there exists a sequence $g_1, \ldots, g_l$ of intermediate genes such that

$$g^{(source)} \longrightarrow g_1 \longrightarrow \cdots \longrightarrow g_l \xrightarrow{expr} g^{(target)}.$$

where the intermediate links are either $\xrightarrow{state}$ or $\xrightarrow{expr}$. Such a sequence has $k = l + 1$ links, and the minimal number of links required to achieve the motif is called the length of $g^{(source)} \Rightarrow g^{(target)}$.

Let $\mathcal{M}_k^*$ denote the set of motifs of length k or less. This operation, which is tissue independent, results in a large number of pairs (more than 200,000 for $k = 3$). Since we take $k = 3$ in our experiments, let $\mathcal{M}^* = \mathcal{M}_3^*$ from here on. To reduce the computational burden of subsequent computations, and because not all such pairs are relevant for a given tissue, this set is reduced by applying a χ^2 test for independence, only keeping motifs $g^{(source)} \Rightarrow g^{(target)}$ for which the independence between the events "$g^{(source)}$ aberrant" and "$g^{(target)}$ divergent" is rejected at a 5% level by the test (without correction for multiple hypotheses) using a dataset of tumor samples. Let $\mathcal{M}$ denote the set of remaining pairs (typically 5,000-10,000), which is therefore tissue dependent (see Figure 2).

We then let $\mathcal{S}$ denote the set of "sources" in $\mathcal{M}$, *i.e.*, the set of genes g such that there exists g' with $(g \Rightarrow g') \in \mathcal{M}$ and, similarly, let $\mathcal{T}$ be the set of all possible targets. We let $T(g)$ denote the set of all the targets of $g \in \mathcal{S}$, that is, the set of all g' such that $(g \Rightarrow g') \in \mathcal{M}$ and $S(g')$ the set of all sources pointing to $g' \in \mathcal{T}$.

2.2.3 Motif-supported anomalies

We can now define a family of binary random variables $(Z_m^{(p)}, m \in \mathcal{M})$ of "pair anomalies" with $Z_m^{(p)}$, for $m = (g^{(source)} \Rightarrow g^{(target)})$, equal to 1 if and only if $g^{(source)}$ is aberrant at the DNA level (either due to mutation or copy-number variation) and $g^{(target)}$ is RNA-divergent.

From this, we also define binary variables $(Z_g^{(s)}, g \in \mathcal{S})$ indicating anomalies at the source level letting $Z_g^{(s)} = 1$ if and only if g participates to an aberrant motif as a source gene. Therefore,

$$Z_g^{(s)} = \max\{Z_m^{(p)} : m = (g \Rightarrow g'), g' \in T(g)\}.$$

Similarly, we consider anomalies at the target level letting

$$Z_g^{(t)} = \max\{Z_m^{(p)} : m = (g' \Rightarrow g), g' \in S(g)\}.$$

for $g \in \mathcal{T}$. We will refer to the event $Z_g^{(s)} = 1$ as g *aberrant with target* and the event $Z_g^{(t)} = 1$ as g *divergent with source.*

2.3 Coverings

2.3.1 Definition

Let $(Z_i, i \in \mathcal{I})$ denote one of three types of anomalies (pair, source, target) that were just defined, so that $\mathcal{I}$ is one of the index sets $\mathcal{M}$, $\mathcal{S}$ or $\mathcal{T}$ and $Z = Z^{(p)}$, $Z^{(s)}$ or $Z^{(t)}$. Our goal is to identify relatively small subsets of $\mathcal{I}$ that contain "important variables" for describing the stochastic behavior of Z. These subsets will be designed as the smallest ones within which at least one anomaly occurs with high probability, leading to the notion of covering that we now formally describe.

Denote by (Ω, P) the probability space on which $Z = (Z_i, i \in \mathcal{I})$ is defined. (An element $\omega \in \Omega$ can be interpreted, for example, as one instance of a tissue sample.) If $\alpha \in [0,1]$ and J is a subset of $\mathcal{I}$, we will say that $Z_J = (Z_j, j \in J)$ is a covering (or 1-covering) of Z at level α if, with probability larger than $1-\alpha$, at least one of its variables is active, *i.e.*,

$$P(\exists j \in J : Z_j = 1) = 1 - \alpha\,.$$

In other terms, Ω is covered (with high probability if α is small) by the union of events $M_j = [Z_j = 1]$, $j \in J$. With a small abuse, we will also call the index set, J, a covering rather than the set of variables it indexes.

More generally, we can define an r-covering at level α as a set J with at least r active variables with probability one, *i.e.*,

$$P\left(|\{j \in J : Z_j = 1\}| \geq r\right) = 1 - \alpha\,.$$

2.3.2 Optimal coverings

We assume that a family of weights $(w_i, i \in \mathcal{I})$ is given and consider the function

$$\sigma(J) = \sum_{j \in J} w_j$$

representing the weighted size of J. For example, if $w_j = 1$ for all j, then $\sigma(J)$ is just the number of elements in J. We define a minimal covering as any covering minimizing σ among all other coverings.

Assume that $\hat{\Omega}$ is a finite subset of Ω, considered as a finite sample from the distribution P (*i.e.*, $\hat{\Omega}$ represents a finite multi-*omics* dataset). Approximating the probability P by normalized frequencies based on $\hat{\Omega}$, the search for an optimal covering can be formulated as an integer linear program, using the fact that it is equivalent to consider subsets J of $\mathcal{I}$ and families of binary configurations $x = (x_j, j \in \mathcal{I})$, with the correspondence $J \longleftrightarrow x$ if and only if $J = \{j \in \mathcal{I} : x_j = 1\}$. Given this, a minimal r-covering at level α is a solution of the problem minimizing

$$F(x, y) = \sum_{j \in \mathcal{I}} w_j x_j$$

subject to the constraints

$$\begin{cases} \forall \omega \in \hat{\Omega} : \displaystyle\sum_{j \in \mathcal{I}} x_j Z_j(\omega) \geq r y(\omega) \\ \qquad\qquad \displaystyle\sum_{\omega \in \hat{\Omega}} y(\omega) \geq |\hat{\Omega}|(1-\alpha) \end{cases}$$

where $(y(\omega), \omega \in \hat{\Omega})$ is a set of auxiliary binary variables such that $y(\omega) = 0$ indicates that ω is excluded from the covering constraint (*i.e.*, it is considered as an outlier).

When all weights w_j are equal, the covering problem will typically have a large number of optimal solutions. Assume that one obtains several sets $J^{(1)}, \ldots, J^{(N)}$, all with same cardinality, and all providing coverings at level α of the considered population. (While it may be computationally prohibitive to compute all solutions, it is often possible to collect a large number of them.) These sets can be combined in at least two obvious way, namely via their union

$$J_{\text{all}} = \bigcup_{i=1}^{N} J^{(i)}$$

or their intersection

$$J_{\text{core}} = \bigcap_{i=1}^{N} J^{(i)} .$$

Many of the lists described in the result section focus on the latter one, that we found particularly interesting in capturing the necessary abnormalities observed in the dataset (even though it does not constitute, by itself, a covering). Another option is to use an auxiliary criterion for the selection of one specific optimal covering, for example, the one maximizing the sum of probabilities that each of its elements is aberrant.

Once a subset J of variables is chosen (a covering, or a core), we obtain a representation of each sample ω as a binary vector $z = (Z_j(\omega), j \in J)$. The obtained reduced representation will, hopefully, retain essential information from the whole *omics* profile associated to the sample. It has, in addition, a mechanistic interpretation, since each variable Z_j is associated to one or a group of motifs $g \Rightarrow g'$. For example, if $Z_j = Z_j^{(s)}$ is computed at the source level, the event $Z_j = 1$ results from a DNA anomaly at a fixed gene, say $g_j \in \mathcal{S}_k$, together with an mRNA anomaly in at least one of its target $g' \in T_k(g_j)$. Because of the relatively small number of variables involved, all these events can be rendered together, using, for example, the visualization provided in Figures 3 or 5.

2.4 Measuring heterogeneity

We want evaluate the heterogeneity of a family of binary random variables $Z = (Z_j, j \in J)$, defined on the probability space Ω, where J is a subset of $\mathcal{I}$ (*e.g.*, a covering). A natural measure for this purpose is the Shannon entropy (Cover and Thomas, 2012), that we need to estimate based on the finite random sample $(Z(\omega), \omega \in \hat{\Omega} \subset \Omega)$. More precisely, we will focus on the entropy of Z conditional to a specific cancer condition, phenotype or subtype, but we ignore this point in this discussion, since it does not bring any special difficulty and makes the notation more cumbersome.

Let $\mathfrak{S} = \mathfrak{S}_J$ denote the set of all binary configurations $z = (z_j, j \in J)$, which has $2^{|J|}$ elements. Let $\pi(z) = P(Z = z)$, so that the Shannon entropy of Z is

$$H(\pi) = -\sum_{z \in \mathfrak{S}} \pi(z) \log_2 \pi(z) .$$

The sample probability mass function of Z is then given by

$$\hat{\pi}(z) = \frac{N(z)}{N}$$

where $N(z) = |\{\omega \in \hat{\Omega} : Z(\omega) = z\}|$ and $N = |\hat{\Omega}|$. One can plug these relative frequencies in the definition of the entropy to obtain the estimator

$$H(\hat{\pi}) = -\sum_{z \in \mathfrak{S}} \hat{\pi}(z) \log_2 \hat{\pi}(z) .$$

This estimator, however, significantly under-estimates the entropy for small and even moderate sample sizes, and several bias-correction methods have been introduced in the literature. We refer the reader to the short review provided in Schürmann (2004), from which we have used the formula

provided in equation (7), itself derived from a computation made in Grassberger (1988, 2003). This estimator computes the entropy using the expression (in which ψ denotes the digamma function)

$$\hat{H}_c = \log_2(e) \sum_{z \in \mathfrak{S}} \frac{N(z)}{N} \left(\log N - \psi(N(z)) - \frac{(-1)^{N(z)}}{N(z)(N(z)+1)} \right) .$$

Still, this estimator is only accurate when the number of variables, $|J|$, is small, because the ratio $(2^{|J|} - 1)/N$ is the first order term in the expansion of the entropy bias (Miller, 1955; Schürmann, 2004) in powers of $1/N$. In our experiments, we would like M to be as small as 30, and can reasonably estimate the entropy for, say, 4 or 5 variables. Since sets J of interest are typically larger, we estimate an upper-bound to the entropy, along the following lines.

Given two random variables X and Y, one always has $H(X, Y) \leq H(X) + H(Y)$. This implies that, if the set J is partitioned into subsets $J_1, \ldots, J_\ell$, then

$$H(Z) \leq H(Z_{J_1}) + \cdots + H(Z_{J_\ell})$$

where $Z_{J_h} = (Z_j, j \in J_h)$, $h = 1, \ldots, \ell$. We use the right-hand side as an upper-bound, determining the partition $J_1, \ldots, J_\ell$ using the following greedy aggregating procedure:

(i) Initialize the partition with singletons, *i.e.*, $J_j = \{j\}$, $j \in J$, computing the estimated entropy $\hat{H}_c(Z_j)$ of the binary variable Z_j. Fix a maximal subset size, L.

(ii) Given a current decomposition $J_1, \ldots, J_\ell$, compute, for all pairs h, h' such that $|J_h \cup J_{h'}| \leq L$, the difference $\hat{H}_c(Z_{J_h}) + \hat{H}_c(Z_{J_{h'}}) - \hat{H}_c(Z_{J_h \cup J_{h'}})$, and replace the pair of sets for which this difference is largest by their union (setting $\ell \leftarrow \ell + 1$).

(iii) If no pair h, h' satisfies $|J_h \cup J_{h'}| \leq L$, stop the procedure.

We used $L = 4$ in our experiments.

The obtained decomposition also provides a statistical model (denoted $\hat{\pi}_*$) approximating the distribution of Z, namely the one for which $Z_{J_1}, \ldots, Z_{J_\ell}$ are independent and the distribution of Z_{J_i} is estimated using relative frequencies. To allow for comparisons between entropies evaluated for different sub-populations, we used this model within a Monte-Carlo simulation to estimate confidence intervals for $H(\pi)$. We generated M =1,000 new N-samples of Z (recall that N is the size of the original sample of Z used to estimate the entropy), using the distribution $\hat{\pi}_*$, resulting in M new empirical distributions $\hat{\pi}_*^{(1)}, \ldots, \hat{\pi}_*^{(M)}$ with associated corrected entropies $\hat{H}_c^{(1)}, \ldots, \hat{H}_c^{(M)}$. Fixing a probability $\beta > 0$, we let $\hat{H}_c^{(\beta)}$ and $\hat{H}_c^{(1-\beta)}$ denote the β and $1 - \beta$ quantiles of the sample $\hat{H}_c^{(1)}, \ldots, \hat{H}_c^{(M)}$ so that $\hat{H}_c^{(j)} - H(\hat{\pi}_*)$ belongs to $[\hat{H}_c^{(\beta)} - H(\hat{\pi}_*), \hat{H}_c^{(1-\beta)} - H(\hat{\pi}_*)]$ with probability $1 - 2\beta$. We use the same interval for the difference $\hat{H}_c - H(\pi)$, yielding the confidence interval for $H(\pi)$:

$$[\hat{H}_c + H(\hat{\pi}_*) - \hat{H}_c^{(1-\beta)}, \hat{H}_c + H(\hat{\pi}_*) - \hat{H}_c^{(\beta)}].$$

2.5 Subtype analysis through partitioning

Assuming here again a family of binary random variables $Z = (Z_j, j \in J)$, where J is a tissue-dependent covering, we build a partition $(S_1, \ldots, S_\ell)$ of the sample space where each S_j is specified by a small number of events involving conjunctions or disjunctions of anomalies. This partition will be associated with the terminal nodes of a binary "coding tree" of limited depth d (*e.g.*, $d = 5$), so that $\ell = 2^d$. To each node in the tree we associate a subset S of $\hat{\Omega}$, and unless S is a terminal node, its children form a partition $S = S' \cup S''$, where $z \in S'$ is based on a certain splitting criterion. There are many ways to build such a tree-structured code. Since our objective is to compare the distributions between cancer subtypes (for a given tissue) with respect to a fixed partition, independent of subtype labels, we will learn a decomposition for the tissue population which is as balanced as possible and unsupervised. In order to balance the sample sizes among subtypes, each sample is weighted inversely proportional to the size of the subtype to which it belongs. Then, at each node, the event which most equally balances the total weight of the two daughter nodes is selected. The standard choice for a binary tree are individual binary features, so events of the form $\{Z_{j(S)} = 1\}$, for a suitably chosen $j(S) \in J$. One could also use more complex ones, such as $\{Z_{j_1(S)} = 1 \text{ or } Z_{j_2(S)} = 1\}$, $\{Z_{j_1(S)} = 1 \text{ and } Z_{j_2(S)} = 1\}$ with $j_1(S), j_2(S) \in J$. (We have used both types of events in our experiments: two-gene disjunctions for tree based on DNA aberrations with targets and two-gene conjunctions for trees based on RNA divergence with sources.) The stopping criterion is that either all samples at the node have identical configuration or a maximum depth has been reached.

3 Results

3.1 Motifs

The only parameter in the process of building motifs ($g^{(source)} \Rightarrow g^{(target)}$) is the maximum length of the directed chain from the source to the target. Based on the genes and interactions found in the Reactome (Jassal et al., 2020), the statistics for different values of $k = 1, 2, 3$ are in Table S1 of the Supplement. We set $k = 3$ in all our experiments. There are then 272,237 valid motifs with 3,124 distinct source genes and 598 distinct target genes. Recall that a pair motif is "active" in a sample DNA-RNA profile if the source gene is DNA-aberrant and the target gene is RNA-divergent. (this event was denoted $\{Z_m^{(p)} = 1\}$ in Methods).

Our samples are those in TCGA with available matched mutation, extreme copy number variation (deleting or amplifying both copies), and mRNA expression data; the conversion of expression counts to divergence states was described in section 2.2.1. For these data nearly all samples exhibit at least one active ($g^{(source)} \Rightarrow g^{(target)}$). In fact, many thousands of such instances can be inferred from TCGA data for many different tissues and spanning at least two orders of magnitude in their frequencies of occurrences.

3.2 Filtering

There are too many motifs to meaningfully analyze. One first filter is based on rejecting the hypothesis that the existence of source and target aberrations are independent (see Section 2.2.2). The statistics of the pair motifs remaining after this filtering procedure for different tissue types are shown in

Supplementary Table S2. For example, for the 953 TGCA breast cancer samples, there are 17,261 valid motifs after the test for independence, with 2,130 source genes and 421 target genes. Another natural filter is to remove very rare events, in particular those which only appear in a very small number of samples; therefore, for each tissue at one of three levels, we require each motif to be realized (active) in at least 2% of the samples for that tissue. For each of our six tissue types and each of our three motif types, the number of qualifying motifs after 2% filter were applied is given in Supplementary Table S3. For instance, for breast cancer sample, there are 4,026 pair motifs, 690 distinct source, and 256 distinct target genes after adding the 2% filter separately to three levels.

3.3 Pair anomalies

Table 1 shows a sampling of motifs $m = (g \Rightarrow g')$ and their associated probabilities of realization in the indicated tissue. For example, in the motif shown for colon cancer in Table 1, *APC* is the source gene, *AXIN2* is the target gene, and there must be a directed signaling path from *APC* to *AXIN2* of length at most three links (two intermediate genes) in Reactome such that the second-to-last link, namely the direct parent of *AXIN2*, controls the mRNA expression of *AXIN2*. In fact, from Supplementary Figure S1, upper left corner, we know that the gene controlling *AXIN2* is in fact *CTNNB1*. This motif is active in a sample ω if *APC* is either mutated, deleted, or amplified *and* the mRNA expression of *AXIN2* is divergent (with respect to baseline mRNA expression for *AXIN2*). In the case of *APC*, the DNA anomaly is nearly always a mutation and *AXIN2* is over-expressed.

The probability P(DNA&RNA) is the sample estimate, namely the fraction of colon samples for which *APC* is mutated *and* *AXIN2* is divergent. Similarly, P(DNA) and P(RNA) stand for the marginal probabilities that the source is DNA-aberrant and the target RNA-aberrant, respectively. The conditional probabilities are then self-explanatory. For example, *APC* is mutated in 73.9% of our samples and in 79.1% of those samples *AXIN2* is divergent. Multiplying these two probabilities gives the frequency of the joint occurrence (58.5%). Other motifs commonly found in colon samples include the four core motifs described in Table 3.

Tissue	Pair	P(DNA&RNA)	P(DNA)	P(RNA)	P(RNA|DNA)	P(DNA|RNA)
Breast	*PIK3CA* $\Rightarrow$ *S100B*	0.316	0.356	0.838	0.888	0.377
Colon	*APC* $\Rightarrow$ *AXIN2*	0.585	0.739	0.676	0.791	0.864
Kidney	*VHL* $\Rightarrow$ *CA9*	0.482	0.485	0.967	0.994	0.498
Liver	*TP53* $\Rightarrow$ *MYBL2*	0.308	0.319	0.814	0.965	0.379
Lung	*TP53* $\Rightarrow$ *TOP2A*	0.529	0.535	0.923	0.988	0.573
Prostate	*PTEN* $\Rightarrow$ *TWIST1*	0.161	0.216	0.654	0.745	0.246

Table 1: Examples of pair motifs. For each of the six tissues, one example of a common pair motif $m = (g \Rightarrow g')$ is shown. P(DNA&RNA) is our sample-based estimate of the probability that m is active, namely, the fraction of samples of the indicated tissue for which the source gene g is DNA-aberrant and the target gene g' is RNA-divergent. Similarly, P(DNA) (respectively, P(RNA)) is the fraction of samples for which g is DNA-aberrant (resp., g' is RNA-divergent), and P(RNA|DNA) is the (estimated) conditional probability that g' is RNA-divergent given g is DNA-aberrant.

The probabilities for *APC* $\Rightarrow$ *AXIN2* are atypically large. In particular, most pair probabilities are smaller than .575, generally of order 0.01–0.10 with a few above 0.3, usually involving main tumor drivers such as *PIK3CA* in breast cancer, and *TP53* and *KRAS* in lung cancer. Moreover,

DNA aberrations tend to be considerably rarer than RNA aberrations, *i.e.*, the marginal source probabilities are generally far smaller than the marginal target probabilities. It is noteworthy that the conditional probability of a particular RNA aberration given a particular DNA aberration (as those in Table 1) is usually in the range 0.5–1, whereas the reverse is not the case: given a target gene is RNA-aberrant the probability of any particular gene serving as a source rarely exceeds 0.2 (see Supplementary Tables S4–S8).

We have also defined separate source-level and target-level events in the sense of partially realized motifs; see Section 2.2.3 of Methods. Recall that $\{Z_g^{(s)} = 1\}$ represents the event that a given source gene g is DNA-aberrant *and that there exists some* target of g which is RNA-aberrant, denoted as "aberration with target" or "source aberration with downstream target divergence". The probability of this event is denoted by $P(\text{DNA} \,\&\, \text{downstreamRNA})$; see Table 4 for some examples in Colon. Similarly, for the other direction, $\{Z_{g'}^{(t)} = 1\}$ is the event that *some* source gene renders $g \Rightarrow g'$ an active motif. Table 4 and Table 5 provide the probabilities and conditional probabilities for selected core genes at the source and target levels in colon; many other examples appear in Supplementary Tables S4 to S18.

A general pattern is that, given a source gene g is aberrant, there is a strong likelihood that at least one of its targets g' is RNA-aberrant. These targets represent plausible downstream consequences of g having a DNA-aberration. The converse, however, is not valid; in particular, there are many targets g' for which there is no upstream DNA-aberrant source linked to g'. This makes sense since a gene can be RNA-aberrant for many reasons other than an upstream genetic anomaly. In particular, the event driving the divergence of g' might be some perturbation not considered here, for example be epigenetic or fusion-related or as yet unrecognized.

3.4 Coverings

Recall that indexing a covering by source genes refers to leaving the particular target gene and the two modes of aberration unspecified. Indexing by targets is the opposite: only $g^{(target)}$ is specified and the motif can be completed with any $g^{(source)}$ and any DNA and RNA aberrations. The corresponding events were denoted in Methods by $\{Z_g^{(s)} = 1\}$ for a source gene and $\{Z_g^{(t)} = 1\}$ for a target gene.

As described in Section 2.3, minimal coverings composed of pairs, sources, or targets are all found with the same optimization program. For the pair and source levels, we calculate the optimal covering with $\alpha = 0$ and $r = 1$. At the pair level, this means that we seek the smallest set of pair motifs such that, for every sample ω, at least one pair motif from this set is active in ω. At the source level, it means we find the smallest set of source genes such that, for every sample ω, at least one source gene from this set is DNA-aberrant with target in ω. At the target level, however, we select $\alpha = 0$ but $r = 3$; that is, we attempt to cover every tumor sample with at least three RNA-aberrations (with source). This choice is justified by the higher frequency of RNA-aberrations in tumor samples.

Table 2 shows the optimal covering statistics at all levels for the six tissue of origin. For example, selecting $\alpha = 0$ in the optimization algorithm, the minimum number of pair motifs (resp., source, target) necessary to cover of the 953 breast cancer samples is 67 (resp., 60, 53); the realized rates are 95% (resp., 96%, 96%). In contrast, virtually all colon samples can be covered with many fewer motifs as 11 (resp., 10, 15). In addition, the minimal covering size (size of solution) is usually largely determined by the incidence of anomalies in any given population, e.g., mutation rates. In particular, given two phenotypes A and B, if the samples of B are consistently more aberrant than those of A,

then the minimal B covering will be smaller. More comprehensive statistics for all tissues are given in Supplementary Table S3; as seen in the column labeled "No. of solutions", there are in general a great many instances of minimal coverings.

Tissue	Samples	Covering Type	Candidate motifs	Size of solution	Fraction of samples covered	Core motifs in solution set
Breast	953	Pair Motif	4,026	67	0.954	21
		Source Level	690	60	0.964	34
		Target Level	256	53	0.955	35
Colon	207	Pair Motif	1,195	11	1.000	4
		Source Level	525	10	1.000	5
		Target Level	226	15	0.995	6
Kidney	336	Pair Motif	347	26	0.827	12
		Source Level	133	28	0.854	21
		Target Level	176	60	0.890	45
Liver	360	Pair Motif	1,198	32	0.931	11
		Source Level	460	34	0.958	20
		Target Level	287	41	0.942	26
Lung	465	Pair Motif	3,154	27	0.985	10
		Source Level	908	25	0.989	19
		Target Level	350	29	0.985	26
Prostate	491	Pair Motif	430	53	0.686	32
		Source Level	211	53	0.743	42
		Target Level	160	72	0.699	66

Table 2: Statistics of optimal coverings. For each of the six tissues, this table provides basic information about the optimal coverings at all levels: pair motif, source with target, target with source. For instance, for breast cancer, there are 4,026 candidate pair motifs after both filters (rejecting source-target independence and 2% tissue sample frequency); the minimal covering size is 67 pair motifs; at least one of these 67 pair motifs is active in 95.4% of the breast cancer samples; and there are 21 pair motifs which appear in *every* minimal covering.

Figure 3 shows one such tissue level covering obtained for breast cancer as a graphical network with nodes representing genes forming the pair motifs. The source and target genes are shown in orange and blue respectively, while genes representing the intermediary links are shown in green. Due to the union of the coverings containing many more motifs and therefore being difficult to visualize, we show here only individual coverings. Supplementary Figures S1– S5 depict the networks associated with the coverings obtained for the other types of cancer.

These visual representations allow us to go beyond lists of names and numbers and begin to interpret coverings in biological terms and incorporate mechanism (see Discussion). For instance, in the breast network shown in Figure 3, important breast cancer genes *STAT3, TP53, BRCA1* and *ERBB4* all form important hubs through which multiple sources and targets in the covering link according to Reactome. Similarly, the network figures for the remaining networks show similar positioning for many important cancer genes: *NOTCH1* in liver, *NOTCH3* and *EGFR* in lung, and *STAT4* in prostate are some other examples. *KRAS, TP53* and *STAT3* make an appearance in multiple cancers. See Supplementary Figures S1–S8.

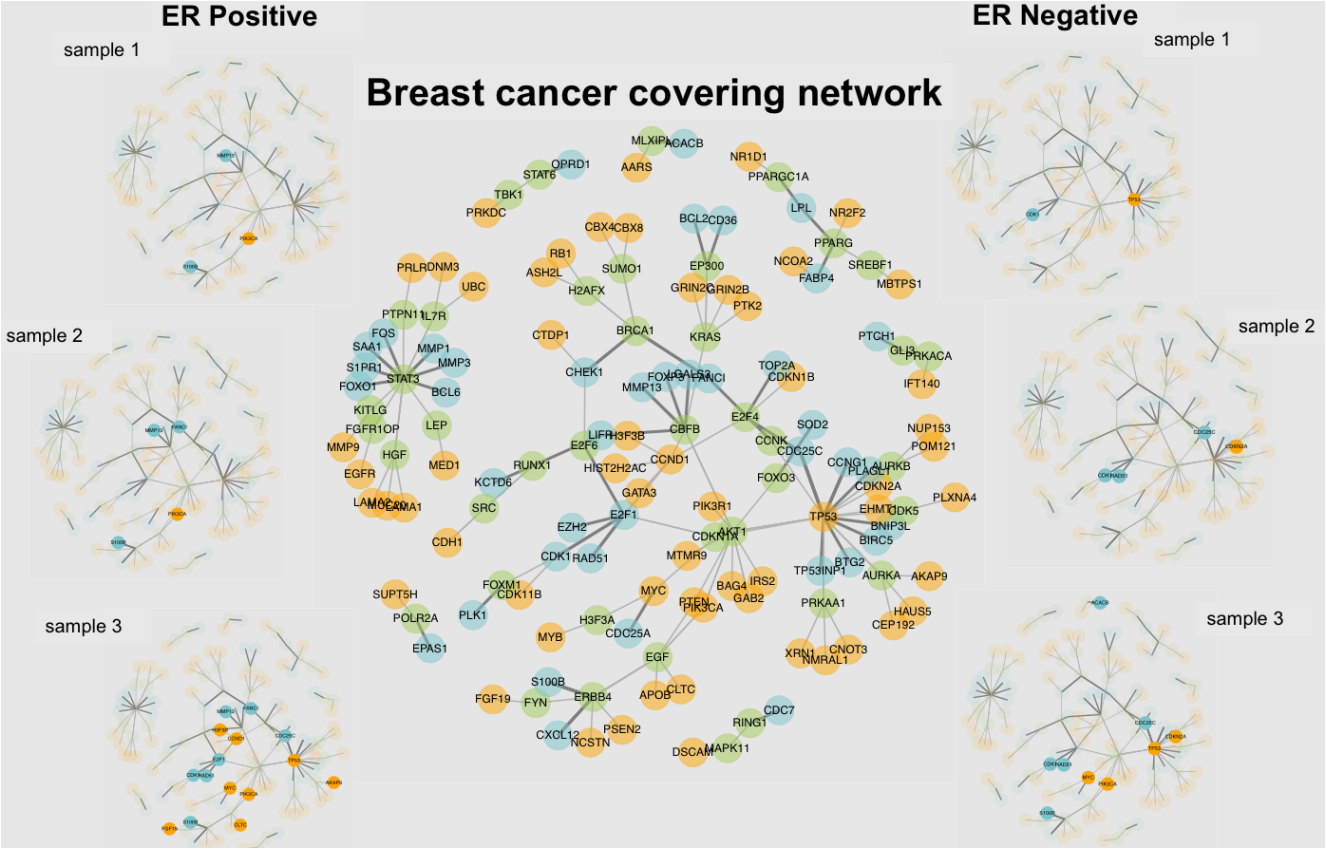

Figure 3: Networks of pair coverings in breast cancer. The network shown in the center depicts one covering of breast cancer samples by pair motifs, with source genes in orange, target genes in blue, and intermediary link genes in green. The thin and thick edges represent, respectively, the two types of relationships: "controls state change of" and "controls expression of" as designated in Reactome (Jassal et al., 2020). On the left are presented a selection of covering realizations for three ER-positive samples, where active (*i.e.*, aberrant) pair motifs are highlighted, while and on the right, three ER-negative samples samples are shown. The samples have different realizations over the covering network, and are ranked (top to bottom) by the number of events they exhibit. The sample networks demonstrate the inter-sample heterogeneity among the source and target realizations.

For a given tissue and fixed covering level, about 30%–60% of the genes appearing in any covering in fact appear in all coverings, referred to as the *core set*. Full core motifs at all three levels are presented in Tables 3, 4 and 5 for colon, and in Supplementary Tables S4–S18 for the other tissues. We also provide the relevant frequencies of occurrence in all three cases, as described in the previous section. For example, the motif *TP53* ⇒ *PTPN12* appears in 39.6% of the colon samples (see Table 3), the source gene *KRAS* in 42.5% of samples (see Table 4), and the target gene *PERP* in 80.7% of samples (see Table 5). From Table 4 we see that there is some aberrant target for *every sample* for which *KRAS* is DNA-aberrant in colon; hence the probability that *KRAS* is DNA-aberrant *and* there is a matching target gene is again 42.5%. Finally, from the target covering we see that targets gene *PDX1* is divergent in 95.7% of colon samples (see Table 5) but only 70.2% of samples for which *PDX1* is divergent have some corresponding corresponding upstream DNA-aberrant source gene.

	P(DNA & RNA)	P(DNA)	P(RNA)	P(RNA|DNA)	P(DNA|RNA)
APC ⇒ *AXIN2*	0.585	0.739	0.676	0.791	0.864
TP53 ⇒ *PTPN12*	0.396	0.560	0.604	0.707	0.656
PIK3CA ⇒ *TNFRSF10B*	0.198	0.271	0.589	0.732	0.336
MAML1 ⇒ *PBX1*	0.034	0.039	0.401	0.875	0.084

Table 3: Colon core pair motifs. There are four "core" pair motifs which appear in every minimal covering of the colon samples. P(DNA & RNA) is the fraction of samples for which the source gene g is DNA-aberrant and target gene g' is RNA-divergent; P(DNA) is the fraction of samples satisfying the source gene g is DNA-aberrant; P(RNA) is the fraction of samples with g' RNA-divergent; P(RNA|DNA) is the fraction of DNA-aberrant samples for which g' is RNA-divergent.

	P(DNA & downstream RNA)	P(DNA)	P(downstream RNA|DNA)
APC	0.585	0.739	0.791
TP53	0.560	0.560	1.000
KRAS	0.425	0.425	1.000
LAMA5	0.217	0.217	1.000
MAML1	0.034	0.039	0.875

Table 4: Colon core source genes. There are five "core" source genes which appear in every minimal source covering of the colon samples. P(DNA) is the fraction of samples for which the indicated source gene is DNA-aberrant; P(DNA & downstream RNA) is the fraction of samples for which the indicated source gene is DNA-aberrant and there exists an RNA-divergent gene among its targets. P(downstream RNA|DNA) is the fraction of the samples with the indicated source gene DNA-aberrant for which there exists some RNA-divergent gene among its targets.

Target	P(RNA & upstream DNA)	P(RNA)	P(upstream DNA|RNA)
PERP	0.710	0.807	0.880
PDX1	0.671	0.957	0.702
AXIN2	0.662	0.676	0.979
SALL4	0.638	0.918	0.695
TNFRSF10B	0.565	0.589	0.959
MYBL2	0.261	0.300	0.871

Table 5: Colon core target genes. There are six "core" target genes which appear in every minimal target covering of the colon samples. P(RNA) is the fraction of samples for which the indicated target gene is RNA-divergent; P(RNA & upstream DNA) is the fraction of samples for which the indicated target gene is RNA-divergent and there exists an DNA-aberrant gene among its sources. P(upstream DNA|RNA) is the fraction of the samples with the indicated gene RNA-divergent for which at least one of its sources is DNA-aberrant.

We have also analyzed the core set across multiple tissues. Figure 4 shows the core source genes which appear in at least two tissues. Notice that there are 18 source genes which appear in the core set of at least 2 tissues. For instance, *TP53* is a core source gene shared by all 6 tissues, and is DNA-aberrant in more than 60% of colon cancers, and also a large percentage in other cancers. And a full figure of all core source genes across multiple tissues is showed in Figure S9. In addition, in Supplementary Figure S10, we show all core target genes across multiple tissues. For instance, *CDC25C* appears in core set of breast, lung and prostate cancer, and the probability that *CDC25* is RNA divergent and there exist an upstream aberrant source is nearly 0.8.

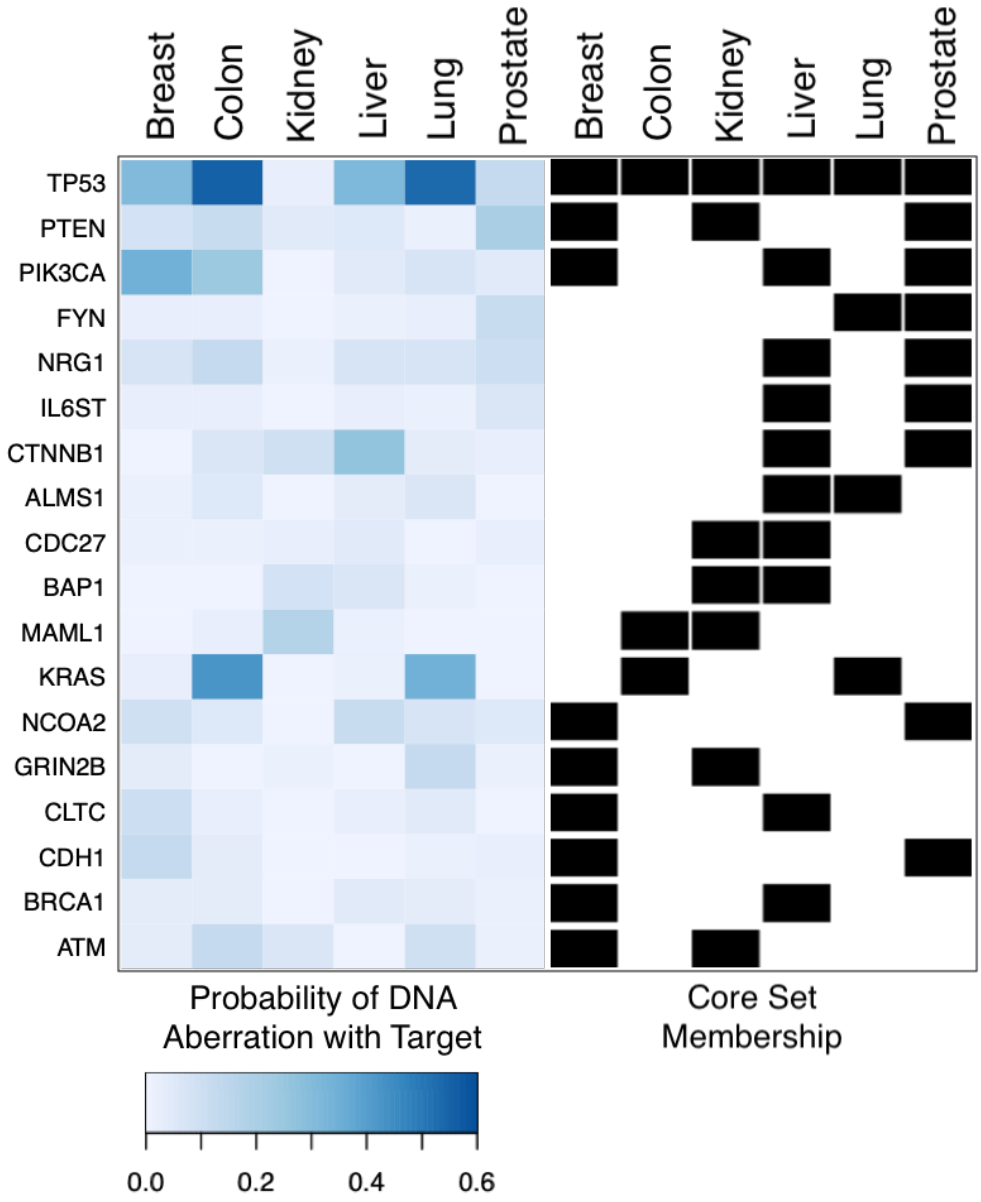

Figure 4: Core set across tissues at source level. There are 18 source genes which appear in the core set of at least two tissues. For instance, gene *TP53* is a core gene for all six tissues, and genes *PTEN* and *PIK3CA* are core genes for three tissues. The color in the heatmap on the left represents the probability that the corresponding source gene is DNA-aberrant and there exists an RNA-divergent target gene (thereby forming an active motif). On the right, black marks indicate the membership of each gene to the corresponding core set for each tumor type.

3.5 Subtype coverings

Having computed the tissue-level coverings, we then proceeded to examine them with respect to certain phenotypes of interest, including the PAM50 subtypes in breast cancer, smoking history in lung cancer, Gleason grade in prostate cancer and and the CRIS-classes in colon cancer (Isella et al., 2017). Simply, we can observe the range of aberration frequencies among subtypes. Table 6A shows the probabilities of DNA-aberration (with targets) for PAM50 subtypes. The genes presented here

are selected from the core set of source breast cancer coverings. The other panels show similar selections of sources from the core set for other cancers. A simple visualization such as this is useful for any observer keen on discovering potential discriminating motifs between subtypes. For example, *TP53* shows much lower probabilities for the luminal subtypes in comparison to the basal-like and HER2-enriched subtypes. Similar observations, also presented in Table 6, can be made among the subtypings of other cancers. Similar observations persist for target motifs and are presented in supplementary tables S19 to S23.

(A)

	Luminal A	Luminal B	HER2-enriched	Basal-like
TP53	0.123	0.273	0.655	0.770
MED1	0.055	0.116	0.545	0.046
PIK3CA	0.438	0.314	0.418	0.172
CLTC	0.068	0.190	0.255	0.057
PTEN	0.050	0.074	0.073	0.253
PTK2	0.105	0.190	0.182	0.310
GATA3	0.151	0.190	0.055	0.253
NCSTN	0.068	0.132	0.127	0.264
NCOA2	0.064	0.149	0.218	0.126

PAM 50 breast cancer subtypes

(B)

	Smoker	Recently reformed	Reformed	Non smoker
TP53	0.581	0.452	0.500	0.259
EGFR	0.093	0.129	0.233	0.370
KRAS	0.395	0.371	0.350	0.148
ANK2	0.140	0.274	0.117	0.037
STK11	0.140	0.290	0.183	0.074
SPTA1	0.372	0.355	0.317	0.185

Smoking status

(C)

	CRIS-A	CRIS-B	CRIS-C	CRIS-D	CRIS-E
APC	0.348	0.217	0.686	0.750	0.710
TP53	0.261	0.478	0.771	0.571	0.774
KRAS	0.500	0.478	0.057	0.393	0.581
LAMA5	0.370	0.261	0.229	0.107	0.194

Colon CRIS subtypes

(D)

	3	4	5
TP53	0.071	0.143	0.286
MYC	0.036	0.094	0.245
PTEN	0.133	0.245	0.347
ZFHX3	0.056	0.102	0.184
FGF17	0.102	0.139	0.163

Prostate primary Gleason grade

Table 6: Probabilities of source aberration with downstream target divergence. For various subtypes of breast, lung, colon and prostate cancer, the heatmaps represent the probabilities that the indicated gene is a DNA-aberrant source gene with some downstream RNA-divergent target. The sources are selected from the set of core genes for coverings of the given tissue; the selection criterion is that the probability of a DNA-aberration is high for at least one of the subtypes for that tissue. Core sources with varying probabilities present interesting candidates for discrimination between subtypes. For example, among the PAM50 subtypes, the DNA-aberration frequency of *TP53* is much higher in the HER2-enriched and Basal-like subtypes than in Luminal A and Luminal B, whereas an aberration in *PIK3CA* is less frequent among basal-like samples than among the other subtypes. In the case of smoking history in lung cancer samples, *TP53* and *KRAS* are both more frequently DNA-aberrant (with some downstream RNA-divergent target) among smokers than non-smokers whereas *EGFR* is a more aberrant source among non-smokers.

A visual comparison between the subtypes can also be obtained as a graphical network, as shown in Figure 5. Supplementary Figure S6 presents the breast covering with the size of the nodes representing the source (with target) aberration probabilities and target (with source) divergence probabilities for the subtype considered. Similar networks for lung cancer with respect to smoking history are presented in Supplementary Figure S7 and Figure S8 presents primary Gleason grade for prostate cancer.

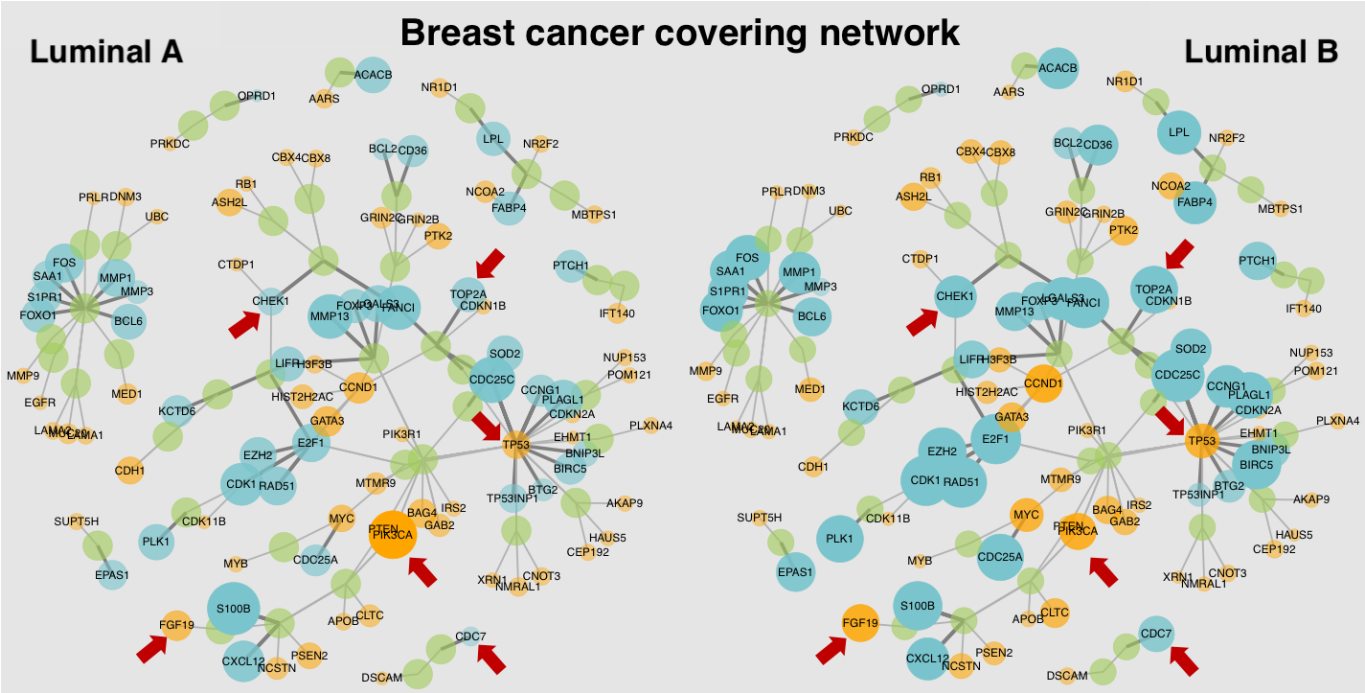

Figure 5: Comparison of one covering network for luminal breast cancer subtypes. The probabilities of DNA aberration (with targets) and RNA divergence (with sources) over the Luminal A and Luminal B populations of breast cancer are depicted by the size of each node in the network, which corresponds to one possible breast cancer covering. The red arrows indicate some sources and target genes that have noticeable differences in the respective probabilities between the two luminal subtypes (*e.g.*, *TP53*, *CHEK1*, *PIK3CA*, and *TOP2A*, also see Table 6).

A further experiment we carried out was designed to compare the covering of subtypes with control for population sizes. For all the phenotypes under a given sub-typing, an equal number of samples were selected and coverings that covered all these samples simultaneously were obtained. Then we examine the proportion from each subtype that was covered. This process was repeated over multiple sampling iterations, and the results obtained are shown in Figure 6. A general pattern of more pathological phenotypes having higher coverage proportions can be observed throughout these results.

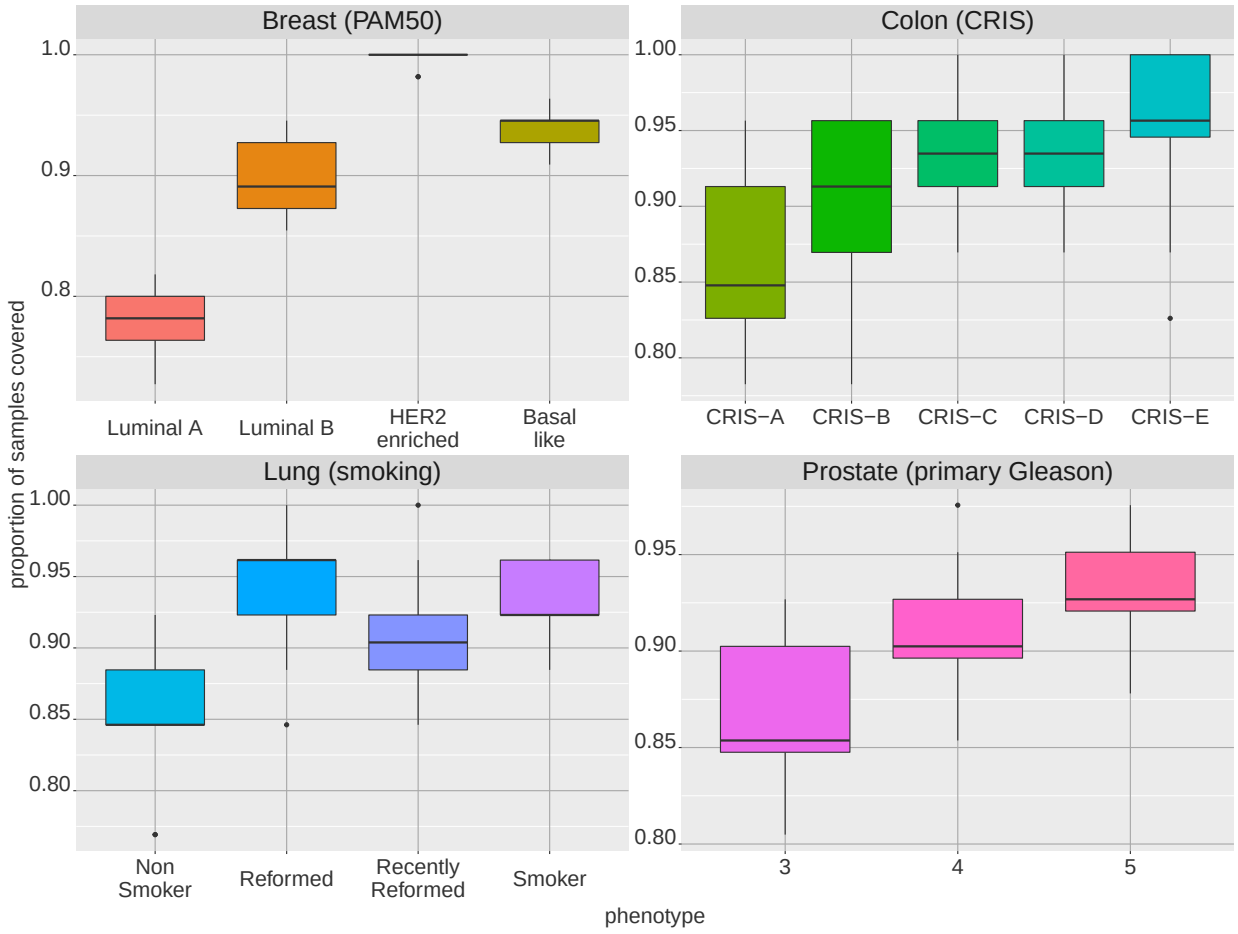

Figure 6: Rates of covering assembly. For each of four tissues (breast, colon, lung and prostate), several phenotypes are compared based on the proportion of samples actually covered when requesting 90% coverage or more for the given tissue by the optimization procedure. The boxplots represent the results of 20 iterations of normalizing for sample size among the phenotypes by random sampling. In general, coverings for more aggressive phenotypes assemble faster.

3.6 Measuring heterogeneity

We applied the approach described in sections 2.4 to assess the relative heterogeneity of different cancer phenotypes and subtypes in the analyzed tissues. Note that we are analyzing heterogeneity within phenotypes at a population level, so that the measurement that is primarily made is about the variability across tumors in this population. Without single-cell data, one cannot evaluate variability within a tumor, although it is likely that a higher variability at this level should also trigger larger heterogeneity at the population level. It is worth noting, however, that the analytical framework we propose here can be easily extended also to single cell data once paired molecular measurements will become available in the future.

We base our analysis on coverings estimated on source aberration with targets and on target divergence with sources. Detailed results can be found in Supplemental Tables S25 and S26. In all cases, coverings are obtained for each tissue of origin separately, and entropy estimates are computed after restricting the data to samples exhibiting each considered cancer phenotype (*e.g.*, the breast cancer molecular subtypes, smoking status in lung cancer, and so on...).

The following observations can be made at the source level (*i.e.*, DNA aberrations with targets). In the comparisons for which differences between patients groups and tumor phenotypes can be observed, the general trend is that heterogeneity estimates increase with increasing disease severity. In prostate cancer, for instance, entropy grows with Gleason sum, primary Gleason grade, and with tumor stage, while no clear ordering is observable for lymph node status (see Table 7). Similar observations can be also made for tumors originating in other tissues (see Supplementary Table S25). In breast cancer, the entropy for ER positive tumors is less than that for ER negative ones, and it also increases with tumor size, and with more aggressive molecular subtypes (*i.e.*, Luminal A < Luminal B < HER 2 < Basal, with a small overlap between confidence intervals for Luminal B and HER 2). For lung, samples from patients with recent smoking history (reformed for less than 15 years or current smokers) have a higher entropy than those with either ancient or no history.

Subtype	Value	N	Entropy	Conf. Interval
All		491	12.16	[11.67, 12.68]
Gleason	6	45	8.05	[6.82, 8.98]
Gleason	7	244	10.80	[10.10, 11.45]
Gleason	8	63	11.33	[10.07, 12.41]
Gleason	9	135	14.21	[13.24, 15.06]
Primary	3	196	9.45	[8.68, 10.12]
Primary	4	245	12.33	[11.65, 13.01]
Primary	5	49	16.60	[15.07, 17.88]
Tumor Status	T2	186	10.73	[9.92, 11.45]
Tumor Status	T3-T4	298	12.88	[12.26, 13.52]
Lymph Node Status	Negative	342	12.14	[11.55, 12.71]
Lymph Node Status	Positive	77	12.82	[11.66, 13.76]

Table 7: Entropy Estimation on DNA aberrations. Entropy estimation (upper bound) on DNA aberrations for prostate Gleason sum, Gleason grade, tumor stage, and lymph-node status. N is the total number of samples available in the given subtype.

When we analyze the entropy across tumor phenotypes at the target level (*i.e.*, mRNA divergence with source), a similar trend of increasing heterogeneity with increasing disease severity is observed in prostate cancer for all variables considered (Gleason sum, primary Gleason grade, tumor stage, size, and lymph node status), in Kidney cancer for tumor stage and tumor size, and in breast cancer for the molecular subtypes (with Luminal A samples exhibiting the lowest heterogeneity while Luminal B the highest). Finally, in the tumor types originating in the other tissues, we observe large overlaps between confidence intervals, and no obvious and clear trends emerged across cancer subtypes. Complete summaries for this analysis can be found in Supplementary Table 7.

3.7 Partitioning

We applied the approach described in section 2.5 to all six cancer types. Let T denote a coding tree with terminal nodes $\{t_1, t_2, \ldots, t_l\}$. Recall that for each subtype, the resulting histogram (number of samples per bin) is, by construction, as balanced as possible for the whole population, which

corresponds, in our experiments, to samples collected for a given tumor type. It is then easy to visualize the histograms conditional on sub-populations (samples for a given subtype), and assess differences across subtypes or phenotypes in this representation. An example is provided in Figure 7, which displays T for ER status in breast cancer based on RNA Divergence with sources.

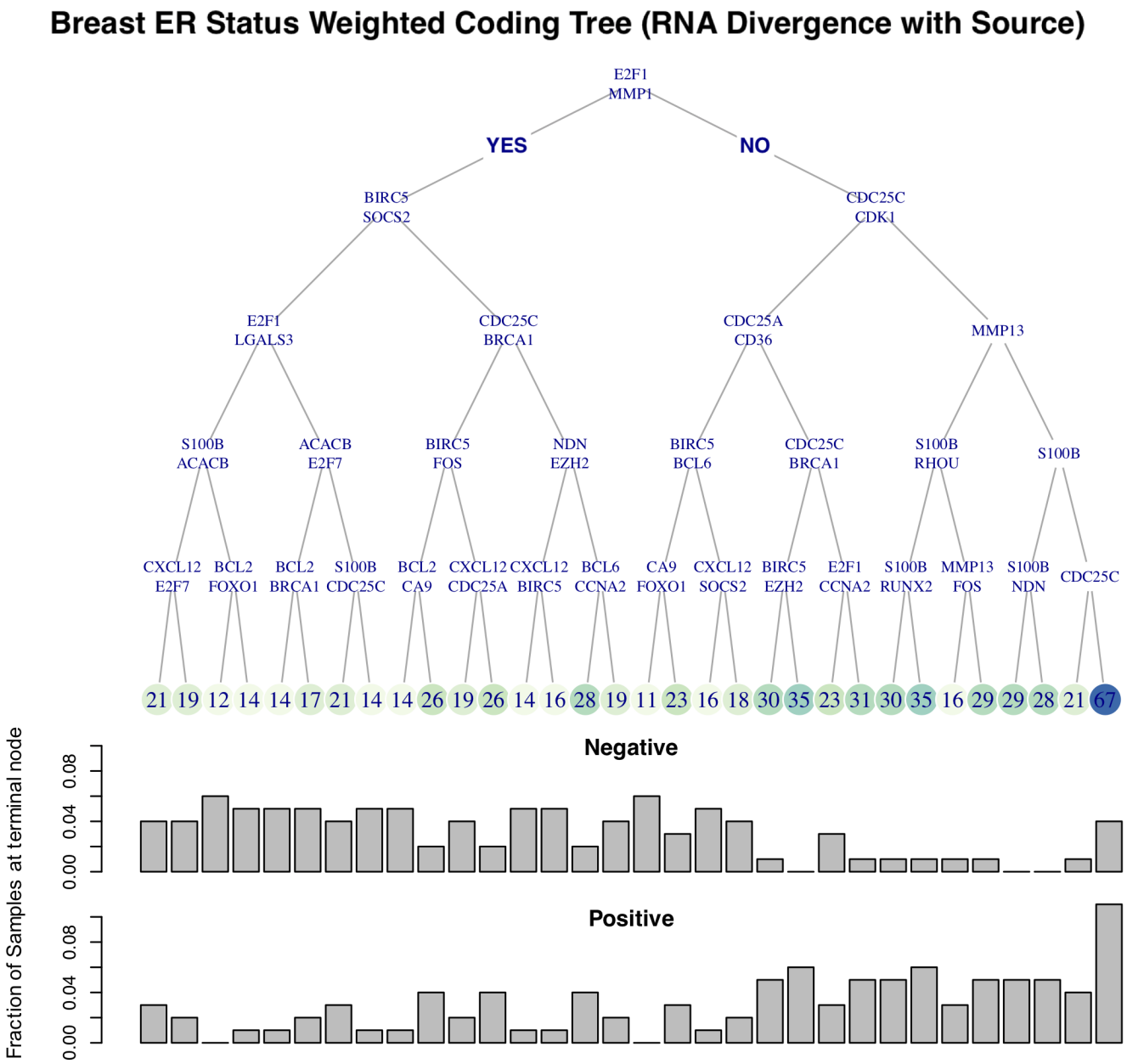

Figure 7: Coding tree for breast ER status using RNA divergence with sources. For Breast tumor samples, a weighted coding tree with depth $d = 5$ is constructed using one covering at target level. At each internal node, a sample is sent to the left if the indicated gene at the node is RNA divergent with some aberrant source, whereas it is sent to the right otherwise. Sample counts for each terminal nodes are indicated and highlighted using a blue palette. The two histograms at the bottom show the sample distribution at the 32 terminal nodes for the ER negative and ER positive sub-populations.

For instance, the event chosen at the root node is whether both *E2F1* and *MMP1* are RNA

divergent, each with some upstream DNA aberrant source. Such samples take the left branch whereas all others go to the right. The number in the terminal node t is the overall number of tumor samples reaching t (regardless of subtypes). The two histograms show the numbers of samples collected at terminal nodes for the ER positive and the ER Negative sub-populations. In particular, the two distributions are significantly different (permutation test $p \leq 0.0001$). Other coding trees for other cancer subtypes are showed in Figures S11–S13.

4 Discussion

We have introduced a new representation of individual cancer profiles and a corresponding computational framework for measuring dysregulation and heterogeneity. The approach is rooted in the biological tenant that cancer is driven by genetic and genomic alterations that alter normal cell behavior through the modification of downstream gene expression programs governing cell proliferation, cell identity, and cell differentiation. The representation is binary and integrative, centered on two-gene motifs which encode paired DNA and RNA anomalies. The set of possible pairings is fixed, based on signaling pathways and regulatory network topology; the motif is active in an individual profile if the first, or "source", gene carries the specified DNA aberration and the gene expression of the second, or "target", gene diverges from the normal baseline. Whereas it may be tempting to interpret anomalous pairs as *functional driver events*, we are not assuming the DNA alterations are "drivers" of cancer or the source-target links are necessarily causative; rather, given an active motif, we regard the divergent target as a putative effect of the upstream source aberration, which itself provides a putative explanation for the divergence of the target.

We have used this representation, together with integer programming, to extract and distill parsimonious sets of important anomalies in the tumor populations. Specifically, we find the smallest set of motifs which cover the cancer population, i.e., (nearly) every individual profile exhibits at least one active motif. In general these sets are not unique; of particular interest are the core motifs which are those belonging to *every* solution. In addition to minimal coverings with paired anomalies, the same algorithm can be applied to find the minimal coverings by DNA aberrations alone for which there is some downstream effect (the source gene appears in some active motif), and vice-versa for coverings by RNA-divergent target genes (i.e., gene expression alterations resulting from some upstream DNA aberrations). In this way we have characterized six distinct tumor types originating in different tissues and organs, observing shared sub-signatures across tumor types as well as subtype-specific ones. In addition, the restriction to coverings massively reduces the number of considered aberrations in a given tumor population, making it mathematically feasible to quantify and measure tumor heterogeneity at the population level in a natural, information-theoretic way.

Overall, we were able to identify well-known cancer anomalies, as well as to uncover novel potential molecular circuits involved in this disease. Briefly, at the source level –*i.e.*, when we recover DNA aberrations with downstream divergent transcriptional effects– among the canonical cancer-drivers, the tumor suppressor gene *TP53* was the only one present in the core signature for all tissues. This finding confirms the well-known notion that this DNA alteration is the most frequent and (possibly) important one across multiple tumor types of different lineage. Among other anomalies involving key cancer drivers, a number played prominent roles in some tumor types, but not in others, following expected patterns. For instance, *PTEN* (inactivation or deletion) appears in the core signatures for

breast, kidney, and prostate cancers,*PIK3CA* (mostly by mutation) was found in breast and colon cancer, while *BRCA1* was inactivated in breast and liver cancer. In addition, a number of anomalies affecting important cancer genes proved to be restricted to specific tissue of origin according to cell lineage and identity, and to canonical pathways of pathogenesis. For instance, *GATA3* – a transcription factor required for normal mammary gland development, known to be altered in breast tumors (Eeckhoute et al., 2007) – was only found in the core signature for breast cancer. Similarly, *APC*, the tumor suppressor gene mutated in familial colon cancer and most sporadic cases of the disease (Goss and Groden, 2000), was present only in the colon cancer core set. Finally, among other known cancer drivers, for instance, *VHL* was part of the core set only in kidney cancer,*EGFR* and *BRAF* were recovered in lung cancer, and *CDKNB1* and *FOXA1* in prostate cancer, according to expectations in line with known cancer biology and recent extensive analyses (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020; Martincorena et al., 2017).

No single gene appears in the core target set of all tissue types. In fact, only one gene, *FABP4*, lies in at least four out of six tissue-specific core signatures; *FABP4* is a fatty acid binding protein playing an important role in glucose and lipid homeostasis (Cao et al., 2008) and has divergent expression (with an upstream DNA alteration) in all cancer types except colon and lung, suggesting that the expression of this gene could serve as cancer biomarker. Similarly, and unexpectedly, several well-known cancer drivers – including *BRCA1, MSH2, TOP2A* and *EZH2* – were recovered among the core target genes. For instance, *EZH2* is a core target in prostate cancer (Varambally et al., 2002) and *S100B* is one in breast tumors (Charmsaz et al., 2017). Finally, irrespective of source, target, or pair level, all core genes belong to signaling pathways commonly disrupted in the analyzed cancer types. For instance, paradigmatic examples of affected cancer pathways emerging from our analyses are *Ras* and *Wnt* in colon cancer and the *PI3K* and *mTor* pathways in breast and prostate cancers.

We also encountered a number of unexpected and presumably novel findings by identifying genes that could play an important role in cancer pathogenesis but and would not otherwise attract attention solely based on the frequency of their DNA and RNA aberrations. Such unexpected genes, and the pathways and regulatory modules in which they participate, can possibly be regarded as potential novel cancer drivers. For instance, *GRIN2B* – a gene encoding for a subunit of a N-methyl-D-aspartate (NMDA) receptor family member– is a source-level core gene for both breast and kidney cancer.Despite a relatively low incidence of DNA aberration (2.1% and 3.6% respectively in kidney and breast cancer), this gene was always associated with the divergent expression of a downstream RNA in both tumor types. Our findings, along with the previously reported promoter hyper-methylation observed in gastric (Liu et al., 2007), esophageal (Kim et al., 2006), and lung cancer (Tamura et al., 2011), collectively suggest this gene might play a role as tumor suppressor.

Another interesting example is *PTK6* which encodes a cytoplasmic protein kinase also known as breast cancer kinase. This is a core target gene in breast, lung, and prostate cancer, with high probabilities of RNA divergence and upstream DNA aberration in breast and lung (17.2% and 50.1% respectively). This suggests, at least in these two tumor types, that different upstream DNA alterations can propagate their effects downstream through the signaling network, “converging” on a particular gene and possibly "explaining" its aberrant expression. Based on this observation, one could speculate that, given such convergence, the protein encoded by *PTK6* could represent a suitably “unifying” target for treatment, despite the heterogeneous set of mutations observed in the patient population. Interestingly, inhibition of *PTK6* has been proposed for treatment in triple negative

breast cancer (Ito et al., 2016) and *PTEN*-null prostate cancer (Wozniak et al., 2017).

In addition, interesting differences between cancer subtypes and phenotypes emerged when we analyzed paired anomaly coverings and tumor heterogeneity at population level. For instance, in lung cancer source level paired anomalies involving *KRAS* were most strongly associated with smoking, whereas those involving *EGRF* showed an opposite trend, consistent with well-established patterns (Herbst et al., 2008). Similarly, *KRAS* aberrations were virtually absent in the CRIS-C colon cancer subtype, which was in turn enriched for anomalies involving *TP53*, as previously described (Isella et al., 2017). Another interesting example is the identification in breast cancer of anomalies involving *CHEK1*, an important regulator of the response to DNA damage, which is over-expressed in triple-negative breast cancer (TNBC) and has therefore been proposed as a potential target for treatment (Marzio et al., 2019). Notably, we were not only able to confirm *CHEK1* divergence (with over-expression) in basal-like tumors (which are enriched for TNBC), but also reveal this anomaly in the luminal B subtype, suggesting a possible vulnerability of this more aggressive type of breast cancer.

In regard to heterogeneity, the general emergent trend was as follows: whenever we observed significant differences in the computed entropy (of the joint distribution of anomalies) between tumor groups, larger entropy estimates were associated with more severe disease phenotypes. Roughly speaking, this suggests that, in general, there is more diversity or variation in the DNA and RNA profiles of sub-population of patients with more aggressive disease phenotypes. Such heterogeneity observed at the population level probably reflects the variability present at the individual level –*i.e.*, the intra-tumoral heterogeneity, stemming from genetic, epigenetic, and cellular variation– which is a well-known factor impacting clinical outcome and therapy response (Jamal-Hanjani et al., 2015; Marusyk et al., 2020). In this regard, another interesting observation is that the profiles of samples of aggressive phenotypes are covered more efficiently than are less aggressive one, implying that, despite more heterogeneity, paired anomalies shared by many samples are more prevalent in more advanced disease states with promising implications for treatment development.

Many studies have reported surprisingly large variations in the frequency and identity of genetic aberrations found in sequenced tumor samples from different individuals with the same diagnosis. Due to this heterogeneity combined with small sample sizes, it is difficult to associate specific changes in gene expression with specific aberrations in cancer genomes. In a recent pan-cancer study (PCAWG Transcriptome Core Group et al., 2020), the authors use matched whole-genome DNA and RNA sequencing data for 1188 patients in order to identify common associations between DNA and RNA aberrations, focusing on fusions, copy number changes, and mutation-driven aberrant splicing. Genetic and transcriptional aberrations are linked based on co-occurrence; and putative causal or mechanistic explanations enter the analysis in a second step in order to annotate and interpret the findings. In contrast, our approach leverages mechanistic constraints at the beginning and by design, focusing on network motifs involving two usually distinct genes paired based on prior, independent, biological information. One common finding is the significant correlation between DNA and RNA alterations, estimated in PCAWG Transcriptome Core Group et al. (2020) to be fifty percent among RNA changes. As can be seen from the values of $P(\text{upstream DNA}|\text{RNA})$ (see Tables 5, S14–S18) our estimates range from fifty to one hundred percent. The higher value we have observed is simply because, in our framework, "RNA" refers to RNA-divergence in a fixed target gene and "upstream DNA" means that the target gene is linked (forms a motif) with *some* DNA-aberrant gene.

Another approach to associating somatic DNA aberrations with downstream transcriptional

changes is multivariate statistical modeling; for example, in Osmanbeyoglu et al. (2017), the authors first predict gene expression from (phospho)protein expression and gene-specific TF binding sites using affinity regression, then predict TF and protein activities from somatic changes. Biological analysis centers on specific genes and pathways, notably the dysregulating effect on TFs of activating mutations in the *PIK3CA* pathway. This pathway also emerges as pivotal in our results: indeed, the most common pair motifs in breast cancer are *PIK3CA* ⇒ *S100B* and *PIK3CA* ⇒ *MMP13* (see Table 1); *PIK3CA* is one of only three core source genes appearing in at least three tissues (see Figure 4); and in breast cancer, *PIK3CA* is virtually certain to have a downstream RNA divergent target (see Table S9). Whereas the methods here and in Osmanbeyoglu et al. (2017) are largely non-overlapping, the spotlight falls on many of the same DNA-RNA associations.

The major objective in (Cai et al., 2019) is to identify driver genes by estimating likely causal relationships between somatic genomic alterations and genes which are differentially expressed relative to normal. Bayesian inference is applied to each individual tumor sample to associate differential expression with a set of genomic alterations in the sample. Characterizations of cancer types are then obtained by summarizing the discovered relationships at the sample and the population levels. Whereas this program bears similarities with ours the objectives and methodology are quite different: our analysis is top-down, based on applying a known network to directly characterize a tumor population with a relatively concise set of paired genomic-transcriptomic relationships, and designed for quantification of inter-tumor heterogeneity. In contrast, the approach in (Cai et al., 2019) is model-driven and the networks are learned.

We have identified "coverings" of cancer populations using combinatorial optimization algorithms and integer programming. Other recent work utilizing similar computational methods includes UNCOVER by (Basso et al., 2019), where the goal is to distinguish between driver and passenger somatic events by detecting sets of genomic alterations associated with an end-point variable, under the constraint of a high degree of mutual exclusivity among the alterations. As posed in (Basso et al., 2019), the problem is NP-hard and only small sets of genes, usually of size three or four, can be found. In contrast, we seek to cover a tumor population with sets of paired anomalies, the number of which are necessary usually ranging from 10-100. This is possible in our case due to the binarization of the targets using divergence, and the utilization of a given known network of interactions, both of which massively reduce the computational burden.

At the DNA level, we have only considered non-synonymous somatic mutations and extreme copy number variations, each of which can be present or absent in any gene for a given sample profile (from the TCGA in our case). Specifically, we did not consider any annotation for mutations (*e.g.*, specific base changes) beyond population frequencies and of course the identity of the harboring gene. We have also limited our annotation of downstream effects to the presence or absence of deviation of RNA expression from a baseline (normal) population, accounting for the direction of the divergence from baseline (*i.e.*, up-regulation or down-regulation). In this way, all variables are binary. The simplicity of this representation has enabled new findings and allows for some analysis of mechanism. For example, dissecting the different types of aberration observed at the source level for important paired anomalies can in principle also facilitate the process of annotating mutations with an unknown effect (so-called variants of uncertain significance). Mutations that are recovered at the population level together with copy number losses can be presumed to be inactivating, and vice-versa. In addition, our representation supports the application of powerful optimization algorithms to identify

minimal coverings of tumor populations, which then serve as fundamental molecular signatures for that population.

However, in order to carry out a deeper analysis of the underlying biology, for example ascertaining the biological role of motifs and their interactions, one would need to impose consistency constraints at the filtering or covering stages. For example, a motif may appear to be "inconsistent" if the source gene is duplicated or has an activating mutation and yet the target gene is down-divergent, assuming the intermediate genes do not further modulate signaling propagation through the network. Such situations should clearly not be excluded. Needless to say there are many other explanations for such observations, e.g., methylation to take but one example. Indeed, there are many cases of divergent target genes which do not appear in any motif, *i.e.*, for which there is no putative explanation in terms of upstream mutations and copy number variations. Uncovering a mechanistically coherent picture of the upstream-downstream synergy would evidently require incorporating additional types of data (e.g., gene fusions, histone modifications and changes in methylation), other sources of transcriptional dysregulation (e.g., expression of microRNAs) and other downstream effects, such as post-transcriptional changes in regulation and aberrant protein structure and concentration. Without such data, making assumptions about consistency among the catalogued and detected anomalies would result in damaging over-simplifications.

The signatures associated with minimal coverings events vary considerably from one tissue or subtype to another. For instance, among our six tissues and at any level (source, target, or pair) the core set of features (those shared by all minimal coverings) is the smallest in colon and accounts for all colon samples in TCGA, whereas substantially larger signatures were necessary in other tumors (*e.g.*, in breast), and some populations could not even be largely covered (*e.g.*, in prostate) regardless of the number of features. The most plausible reason for these differences are a) the network information used to build motifs (source-target pairings) is biased; or b) the *omics* data types do not account for some key molecular mechanisms altered in specific cancer cell types. For the first point, it is plausible, indeed, that better and more refined network information is available for some cancer types than others. For instance, colon cancer has served for years as a model of tumorigenesis, and a wealth of data is available to derive "realistic" signaling pathways and regulatory networks compared with cancer types studied to a lesser extent. For the second point, other DNA alterations, beside mutations and copy number changes, can drive tumorigenesis and may be necessary to efficiently "cover" a cancer population. A prototypical example is that the exceptionally large coverings in prostate cancer may be due to the absence of data on gene fusions; in fact, over a half of the tumors could be accounted for by a small subset of such alterations (*e.g.*, the fusion between *TMPRSS2* and *ERG*, or other *ETS* family genes (Tomlins et al., 2009)).

Finally, the theoretical framework we have developed is based on the paradigm "*Regulators -> Targets*" and would support the incorporation of additional *omics*information. In fact, gene fusion, epigenetic measurements, epigenomic states, enhancer expression, and so forth, could be simply integrated to generate an expanded repertoire of motifs, as could proteomics or metabolomics serve as additional downstream "targets".

5 Conclusions

We have described an integrated analysis of DNA and RNA anomalies, which is grounded in cancer biology and enabled by a highly simplified summary representation of the complex and heterogeneous landscape of anomalies in cancer populations. The summary is a signature of two-gene motifs, each linking a particular DNA aberration with a downstream RNA expression change, and derived automatically from a "covering" algorithm in combinatorial optimization. Beside recapitulating many known alterations, our signature flags potentially important aberrations and interactions which might go unrecognized using simple frequency criteria, being somewhat "diluted" in the vast collection of low frequency events observed at the population level. This distillation can in turn facilitate discriminating cancer drivers from passenger aberrations as well as the identify of potential novel therapeutic targets and better prioritize genes for further functional studies. Furthermore, this representation allows for a rigorous quantitative estimate of cancer heterogeneity in a cancer population and across distinct tumor phenotypes, which would not be otherwise feasible. The heterogeneity differences observed between distinct cancer phenotypes, along with the interactions among paired anomalies, suggest that our approach can represent an alternative to standard statistical filtering to identify important features for predictive model building and machine learning application in cancer. Finally, these motif-covering representations provide a highly efficient and innovative computational tool for harnessing the expanding knowledge about cancer genomes and epigenomes that is emerging from the efforts of large genomics consortia projects, along with powerful algorithms to analyze the resulting compendium of molecular alterations.

Acknowledgements

We thank Drs. Giovanni Parmigiani, Nathan Price, Diego Fernando Sanchez and Eddie Luidy-Imada for helpful discussions. This research was supported by NIH National Cancer Institute Grant R01CA200859.

References

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M. C., Kim, J., Reardon, B., et al. (2018). Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173(2):371–385.

Basso, R. S., Hochbaum, D. S., and Vandin, F. (2019). Efficient algorithms to discover alterations with complementary functional association in cancer. *PLoS computational biology*, 15(5):e1006802.

Cai, C., Cooper, G. F., Lu, K. N., Ma, X., Xu, S., Zhao, Z., Chen, X., Xue, Y., Lee, A. V., Clark, N., et al. (2019). Systematic discovery of the functional impact of somatic genome alterations in individual tumors through tumor-specific causal inference. *PLoS computational biology*, 15(7):e1007088.

Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. (2013). The cancer genome atlas pan-cancer analysis project. *Nat Genet*, 45(10):1113–20.

Cao, H., Gerhold, K., Mayers, J. R., Wiest, M. M., Watkins, S. M., and Hotamisligil, G. S. (2008). Identification of a lipokine, a lipid hormone linking adipose tissue to systemic metabolism. *Cell*, 134(6):933–44.

Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G. D., and Sander, C. (2011). Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res*, 39(Database issue):D685–90.

Charmsaz, S., Hughes, É., Bane, F. T., Tibbitts, P., McIlroy, M., Byrne, C., Cocchiglia, S., McBryan, J., Hennessy, B. T., Dwyer, R. M., Kerin, M. J., Hill, A. D., and Young, L. S. (2017). S100β as a serum marker in endocrine resistant breast cancer. *BMC Med*, 15(1):79.

Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nature genetics*, 45(10):1127.

Costello, J. C., Heiser, L. M., Georgii, E., Gönen, M., Menden, M. P., Wang, N. J., Bansal, M., Hintsanen, P., Khan, S. A., Mpindi, J.-P., et al. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. *Nature biotechnology*, 32(12):1202.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346.

Dinalankara, W., Ke, Q., Xu, Y., Ji, L., Pagane, N., Lien, A., Matam, T., Fertig, E. J., Price, N. D., Younes, L., et al. (2018). Digitizing omics profiles by divergence from a baseline. *Proceedings of the National Academy of Sciences*, 115(18):4545–4552.

Eeckhoute, J., Keeton, E. K., Lupien, M., Krum, S. A., Carroll, J. S., and Brown, M. (2007). Positive cross-regulatory loop ties gata-3 to estrogen receptor alpha expression in breast cancer. *Cancer Res*, 67(13):6477–83.

Goss, K. H. and Groden, J. (2000). Biology of the adenomatous polyposis coli tumor suppressor. *J Clin Oncol*, 18(9):1967–79.

Grassberger, P. (1988). Finite sample corrections to entropy and dimension estimates. *Physics Letters A*, 128(6-7):369–373.

Grassberger, P. (2003). Entropy estimates from insufficient samplings. *arXiv preprint physics/0307138*.

Haigis, K. M., Cichowski, K., and Elledge, S. J. (2019). Tissue-specificity in cancer: The rule, not the exception. *Science*, 363(6432):1150–1151.

Herbst, R. S., Heymach, J. V., and Lippman, S. M. (2008). Lung cancer. *N Engl J Med*, 359(13):1367–80.

Hristov, B. H. and Singh, M. (2017). Network-based coverage of mutational profiles reveals cancer genes. *Cell systems*, 5(3):221–229.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82–93.

Isella, C., Brundu, F., Bellomo, S. E., Galimi, F., Zanella, E., Porporato, R., Petti, C., Fiori, A., Orzan, F., Senetta, R., et al. (2017). Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer. *Nature communications*, 8(1):1–16.

Ito, K., Park, S. H., Nayak, A., Byerly, J. H., and Irie, H. Y. (2016). Ptk6 inhibition suppresses metastases of triple-negative breast cancer via snail-dependent e-cadherin regulation. *Cancer Res*, 76(15):4406–17.

Jamal-Hanjani, M., Quezada, S. A., Larkin, J., and Swanton, C. (2015). Translational implications of tumor heterogeneity. *Clin Cancer Res*, 21(6):1258–66.

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., Loney, F., May, B., Milacic, M., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Weiser, J., Wu, G., Stein, L., Hermjakob, H., and D'Eustachio, P. (2020). The reactome pathway knowledgebase. *Nucleic Acids Res*, 48(D1):D498–D503.

Kim, M. S., Yamashita, K., Baek, J. H., Park, H. L., Carvalho, A. L., Osada, M., Hoque, M. O., Upadhyay, S., Mori, M., Moon, C., and Sidransky, D. (2006). N-methyl-d-aspartate receptor type 2b is epigenetically inactivated and exhibits tumor-suppressive activity in human esophageal cancer. *Cancer Res*, 66(7):3409–18.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214.

Liu, J.-W., Kim, M. S., Nagpal, J., Yamashita, K., Poeta, L., Chang, X., Lee, J., Park, H. L., Jeronimo, C., Westra, W. H., Mori, M., Moon, C., Trink, B., and Sidransky, D. (2007). Quantitative hypermethylation of nmdar2b in human gastric cancer. *Int J Cancer*, 121(9):1994–2000.

Martincorena, I., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Van Loo, P., Davies, H., Stratton, M. R., and Campbell, P. J. (2017). Universal patterns of selection in cancer and somatic tissues. *Cell*, 171(5):1029–1041.e21.

Marusyk, A., Janiszewska, M., and Polyak, K. (2020). Intratumor heterogeneity: The rosetta stone of therapy resistance. *Cancer Cell*, 37(4):471–484.

Marzio, A., Puccini, J., Kwon, Y., Maverakis, N. K., Arbini, A., Sung, P., Bar-Sagi, D., and Pagano, M. (2019). The f-box domain-dependent activity of emi1 regulates parpi sensitivity in triple-negative breast cancers. *Mol Cell*, 73(2):224–237.e6.

Miller, G. (1955). Note on the bias of information estimates. *Information theory in psychology: Problems and methods*, pages 95–100.

Osmanbeyoglu, H. U., Pelossof, R., Bromberg, J. F., and Leslie, C. S. (2014). Linking signaling pathways to transcriptional programs in breast cancer. *Genome research*, 24(11):1869–1880.

Osmanbeyoglu, H. U., Toska, E., Chan, C., Baselga, J., and Leslie, C. S. (2017). Pancancer modelling predicts the context-specific impact of somatic mutations on transcriptional programs. *Nature communications*, 8:14249.

PCAWG Transcriptome Core Group, Calabrese, C., Davidson, N. R., Demircioğlu, D., Fonseca, N. A., He, Y., Kahles, A., Lehmann, K.-V., Liu, F., Shiraishi, Y., Soulette, C. M., Urban, L., Greger, L., Li, S., Liu, D., Perry, M. D., Xiang, Q., Zhang, F., Zhang, J., Bailey, P., Erkek, S., Hoadley, K. A., Hou, Y., Huska, M. R., Kilpinen, H., Korbel, J. O., Marin, M. G., Markowski, J., Nandi, T., Pan-Hammarström, Q., Pedamallu, C. S., Siebert, R., Stark, S. G., Su, H., Tan, P., Waszak, S. M., Yung, C., Zhu, S., Awadalla, P., Creighton, C. J., Meyerson, M., Ouellette, B. F. F., Wu, K., Yang, H., PCAWG Transcriptome Working Group, Brazma, A., Brooks, A. N., Göke, J., Rätsch, G., Schwarz, R. F., Stegle, O., Zhang, Z., and PCAWG Consortium (2020). Genomic basis for rna alterations in cancer. *Nature*, 578(7793):129–136.

Schürmann, T. (2004). Bias analysis in entropy estimation. *Journal of Physics A: Mathematical and General*, 37(27):L295.

Sever, R. and Brugge, J. S. (2015). Signal transduction in cancer. *Cold Spring Harbor perspectives in medicine*, 5(4):a006098.

Tamura, H., Suzuki, M., Moriya, Y., Hoshino, H., Okamoto, T., Yoshida, S., and Yoshino, I. (2011). Aberrant methylation of n-methyl-d-aspartate receptor type 2b (nmdar2b) in non-small cell carcinoma. *BMC Cancer*, 11:220.

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., Speedy, H. E., Stefancsik, R., Thompson,

S. L., Wang, S., Ward, S., Campbell, P. J., and Forbes, S. A. (2019). Cosmic: the catalogue of somatic mutations in cancer. *Nucleic Acids Res*, 47(D1):D941–D947.

Tomasetti, C., Marchionni, L., Nowak, M. A., Parmigiani, G., and Vogelstein, B. (2015). Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proc Natl Acad Sci U S A*, 112(1):118–23.

Tomlins, S. A., Bjartell, A., Chinnaiyan, A. M., Jenster, G., Nam, R. K., Rubin, M. A., and Schalken, J. A. (2009). Ets gene fusions in prostate cancer: from discovery to daily clinical practice. *Eur Urol*, 56(2):275–86.

Varambally, S., Dhanasekaran, S. M., Zhou, M., Barrette, T. R., Kumar-Sinha, C., Sanda, M. G., Ghosh, D., Pienta, K. J., Sewalt, R. G. A. B., Otte, A. P., Rubin, M. A., and Chinnaiyan, A. M. (2002). The polycomb group protein ezh2 is involved in progression of prostate cancer. *Nature*, 419(6907):624–9.

Vogelstein, B. and Kinzler, K. W. (2015). The path to cancer—three strikes and you're out. *N Engl J Med*, 373(20):1895–1898.

Wozniak, D. J., Kajdacsy-Balla, A., Macias, V., Ball-Kell, S., Zenner, M. L., Bie, W., and Tyner, A. L. (2017). Pten is a protein phosphatase that targets active ptk6 and inhibits ptk6 oncogenic signaling in prostate cancer. *Nat Commun*, 8(1):1508.

Zehir, A., Benayed, R., Shah, R. H., Syed, A., Middha, S., Kim, H. R., Srinivasan, P., Gao, J., Chakravarty, D., Devlin, S. M., et al. (2017). Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nature medicine*, 23(6):703.