Splitpea: quantifying protein interaction network rewiring changes due to alternative splicing in cancer

Ruth Dannenfelser and Vicky Yao[†]

Department of Computer Science, Rice University,
Houston, TX 77005, USA
[†]*E-mail: vy@rice.edu*

Protein-protein interactions play an essential role in nearly all biological processes, and it has become increasingly clear that in order to better understand the fundamental processes that underlie disease, we must develop a strong understanding of both their context specificity (e.g., tissue-specificity) as well as their dynamic nature (e.g., how they respond to environmental changes). While network-based approaches have found much initial success in the application of protein-protein interactions (PPIs) towards systems-level explorations of biology, they often overlook the fact that large numbers of proteins undergo alternative splicing. Alternative splicing has not only been shown to diversify protein function through the generation of multiple protein isoforms, but also remodel PPIs and affect a wide range diseases, including cancer. Isoform-specific interactions are not well characterized, so we develop a computational approach that uses domain-domain interactions in concert with differential exon usage data from The Cancer Genome Atlas (TCGA) and the Genotype-Tissue Expression project (GTEx). Using this approach, we can characterize PPIs likely disrupted or possibly even increased due to splicing events for individual TCGA cancer patient samples relative to a matched GTEx normal tissue background.

Keywords: alternative splicing; protein-protein interaction networks; protein network rewiring

1. Introduction

Alternative splicing is a crucial mechanism that underlies the increased complexity of higher eukaryotes. It is now estimated that ∼95% of human genes[1,2] undergo splicing changes, and the increase in protein diversity that results from splicing has been put forth as one of the primary explanations for the apparent mismatch between species complexity and their genome size.[3,4] Importantly, alternative isoforms of the same gene can exhibit highly different interaction profiles and thus affect the dynamics of protein interaction networks.[5] Splicing has been shown to be a key regulator of tissue specificity (especially in the brain),[2,6] and dysregulation has been increasingly implicated in a wide array of diseases,[7] from cancer[8,9] to neurodegenerative diseases.[10] Thus, it is critical to understand the changes in protein interactions due to splicing that underlie cellular function and dysfunction.

However, a systematic study of splicing-related protein network dynamics is hampered by multiple challenges. Although emergent experimental approaches to directly screen for isoform-level protein-protein interactions are promising,[5] they are very early in development and highly restricted in resolution. Furthermore, all such screens are naturally bounded by not only a combination of technical and cost constraints, but also the inherent complexity of the underlying networks and the vast number of potential cell types and conditions of interest. Fortunately, the now standard use of RNA-sequencing provides a window into the exploration of splicing patterns across varied conditions. While RNA-seq data alone is still insufficient to chart out the entirety of any particular splicing interaction network, it can be used to understand condition-specific splicing dynamics.

Here, we present Splitpea (SPLicing InTeractions PErsonAlized), a method for detecting sample-specific PPI network rewiring events. Splitpea takes advantage of the key insight that splicing can disrupt critical *protein domains* that mediate PPIs through domain-domain interactions (DDIs), which have been derived based on a mix of structural, evolutionary, and computational approaches.[11–14] Splitpea integrates PPI and DDI information with sample-specific differential splicing events, and can be used easily in concert with existing, established computational approaches for the identification and quantification of differential splicing.[15] In the scenario where only an individual sample is available or a different background context is preferable (versus existing control samples), Splitpea provides functionality to use a separate reference database of background splice events; for example, one can choose to use normal GTEx data as background for individual TCGA cancer samples (matched by tissue type). Furthermore, as part of Splitpea's characterization of the potential downstream interaction network changes, Splitpea indicates likely direction: gain, loss, or chaos (mixed / unclear).

Thus, to our knowledge, Splitpea is the first general tool to characterize potential direction of protein interaction rewiring due to splicing for individual samples. We demonstrate the utility of Splitpea on breast and pancreatic cancer samples from TCGA, using matched normal tissue samples (breast and pancreas) from GTEx. All source code for Splitpea and the corresponding analyses are available via Github (`https://github.com/ylaboratory/splitpea`), with additional links to download all data and associated networks.

1.1. *Prior work*

Prior work considering domain-domain interactions in the context of splicing have mostly focused on query-based or visualization interfaces. Many consider interactions at the isoform level, aiming to provide a context-specific isoform interaction graph.[16–18] There has been relatively less work focusing on characterizing network rewiring events. Recently, the first tool to characterize the mechanistic effects of splicing on downstream PPIs was proposed,[19] but this tool is unable to differentiate between the potential directionality of interaction rewiring (likely gain or loss events). Specifically for the study of cancer, there has also been large-scale analysis efforts to characterize the impact of splicing on PPIs across patients.[8] Though this work was not patient-specific, it provided strong evidence to demonstrate that there exists a large catalog of isoform changes (with potential downstream impacts on PPIs and regulatory networks) that exist independently of expression changes in cancer. Beyond using PPI networks, there have also been exciting efforts integrating cancer RNA-seq together with somatic mutation data and using functional networks to interpret the downstream impact of splicing.[20]

2. Methods

2.1. *Protein interaction and domain interaction data*

Human protein-protein interactions were downloaded from BioGRID (v4.4.207),[21] DIP (2017-02-05),[22] HIPPIE (v2.2),[23] HPRD (Release 9),[24] Human Interactome (HI-II),[25] IntAct (2022-04-18),[26] iRefIndex (v18.0),[27] and MIPS (Nov 2014).[28] All proteins were mapped to Entrez Gene IDs.[29]

Known and predicted domain-domain interactions were downloaded from 3did (v2017_06),[11] DOMINE (v2.0),[12] IDDI (2011.05.16),[13] and iPFAM (v1.0).[14] For predicted DDIs, only interactions with confidence > 0.5 were used in downstream analyses.

Protein domain locations were translated to genomic locations using the Ensembl BioMart API and the biomaRt R package[30] and indexed using tabix[31] to facilitate fast retrieval given a set of genomic coordinates.

2.2. *Tissue and tumor splicing data processing*

Spliced exon values in the form of percent spliced in (PSI or ψ) were obtained for both normal pancreas and breast tissue samples from the Genotype-Tissue Expression (GTEx) project and pancreatic cancer and breast cancer samples from The Cancer Genome Atlas (TCGA) using the IRIS database.[32] IRIS uses rMATS[33] to tabulate ψ values for skipped exon events (the most abundant splicing event). Though we use rMATS ψ values in this study, Splitpea is agnostic to the choice of upstream differential splicing analysis tool and can easily be applied in concert with other tools that use a form of ψ as their quantification metric.[34–37]

Specifically, we delineate ψ_i as the ψ value for exon $i = 1, ..., n_E$, where there are n_E total exons that had a reported exon skipping event. Note that the precise exons captured in the sample of interest and the background samples are typically non-identical. We are only able to estimate ψ for exons that are captured in both, and thus, n_E represents the number of exons that lie at the intersection of the two larger sets of exons. In the scenario where a background reference distribution of ψ values are provided, we calculate $\Delta\psi_i$ as the following:

$$\Delta\psi_i^{(\mathrm{s})} = \psi_i^{(s)} - \frac{1}{n_B}\sum_{b=1}^{n_B}\psi_i^{(b)} \tag{1}$$

where $\psi_i^{(s)}$ is the ψ for exon i in our sample of interest s (e.g., a cancerous pancreatic sample from TCGA), while $\psi_i^{(b)}$ is the ψ for the same exon i in an individual background sample b (e.g., a normal pancreatic sample from GTEx), and n_B is total number of background samples. Intuitively, larger n_B will provide better estimates of the background distribution, especially if there is large variability in splicing patterns. We recommend assembling backgrounds with at least $n_B \geq 30$ for the empirical cumulative density function estimate below.

ψ values lie in the range $[0, 1]$; thus $\Delta\psi \in [-1, 1]$, and we are naturally primarily interested in significant events for large $|\Delta\psi|$ values (cases where exons are significantly skipped or significantly retained relative to reference). To calculate an estimated significance level for $|\Delta\psi_i^{(\mathrm{s})}|$, we rely on similar intuition as used in previous studies,[8,38] that the normal reference samples can be used to construct an empirical cumulative density function for each exon:

$$\hat{F}_{n_E}(t_i) = \frac{1}{n_B}\sum_{b=1}^{n_B} \mathbf{1}_{|\psi_i^{(b)}| \leq t_i} \tag{2}$$

where $\mathbf{1_A}$ is the indicator function for event A. Given this exon-specific $\hat{F}_{n_E}(t_i)$, we can estimate an empirical p-value for each exon i in sample s

$$\hat{p}_i^{(s)} = \frac{1}{2}(1 - \hat{F}_{n_E}(|\Delta\psi_i^{(\mathrm{s})}|)) \tag{3}$$

Finally, as input to Splitpea, we filtered exons to only those that are significantly different from background ($\hat{p}_i^{(s)} < 0.05$) and those with a $\Delta\psi$ change bigger than 0.05 ($|\Delta\psi| > 0.05$), defined as ψ below. We chose to use a p-value cutoff here as opposed to a multiple hypothesis corrected value to reduce false negatives, because we are interested in any possible rewiring events. We hope that this will better enable Splitpea's use for hypothesis generation tasks. In general, these thresholds can be easily varied depending on the downstream purpose.

2.3. *Clustering $\Delta\psi$ values*

For each cancer type, we remove any exons that had missing values in any of the samples, then filtered the exons by variance, keeping only those with variance greater than 0.01. The final set of $\Delta\psi$ for each cancer type were clustered using the complete hiearchical clustering algorithm and plotted with the heatmap.2 function in the gplots R package.[39] Clinical annotations for TCGA samples were obtained from the Genomic Data Commons portal with Pam50 calls from Netanely et al.[40]

2.4. *Network rewiring algorithm*

There is inherent complexity in considering the impact of exon changes on protein domains, and finally, proteins, as there are several many-to-many relationships. A single exon can include multiple protein domains, but a single protein domain can also span multiple exons; proteins can thus consist of multiple exons as well as multiple protein domains. Splitpea hones in on potentially domain-mediated protein interactions by first overlaying DDIs on the aggregated PPI network based on the presence of each of the domains that constitute the pair of interactors in the protein. In other words, for a pair of proteins g_1 and g_2, we consider protein domain d_1 in g_1 and domain d_2 in g_2 as potentially mediating a known PPI between g_1 and g_2 if a DDI has been reported between d_1 and d_2. Fig. 1A depicts an example interaction where several DDIs potentially mediate the same PPI.

In the event that there are multiple exons within the same protein domain, we attribute the *minimum* $\Delta\psi$ value to the entire protein domain. The underlying assumption here is that loss of any portion of a particular protein domain may potentially negatively impact the protein domain's downstream capacity to interact with other domains. Splitpea then determines the directionality of change based on whether or not there is consistency across the changing domains. In the event that there are mixed exon changes, the directionality is labeled as "chaos," or undetermined (Fig. 1B). The weight of the edge is calculated as

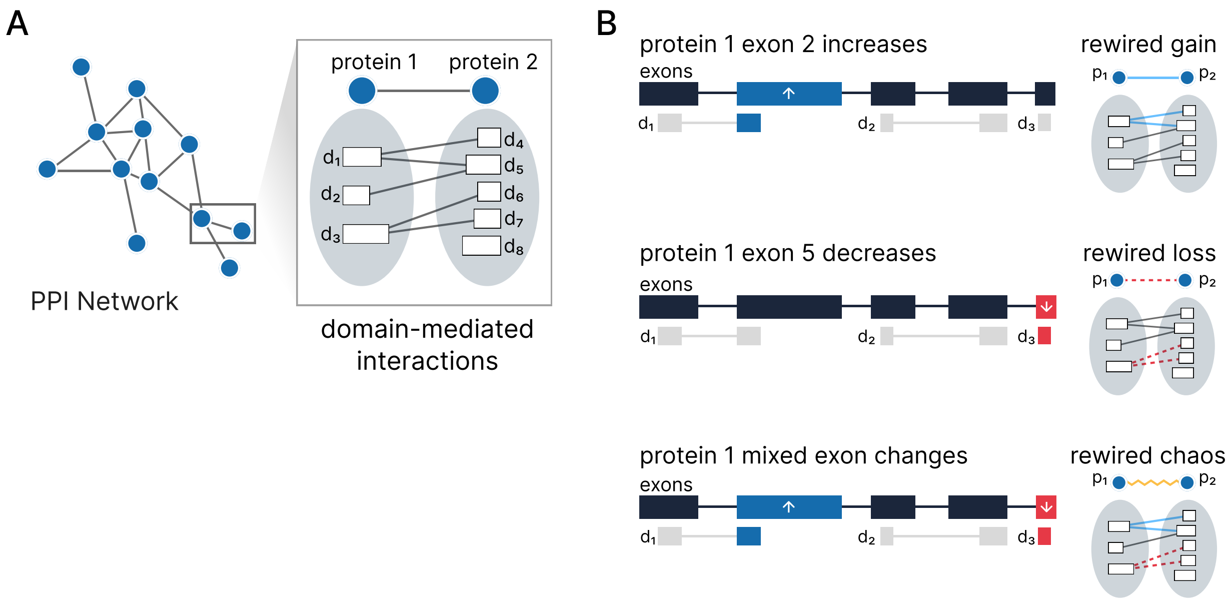

Fig. 1. *Overview of Splitpea.* Splitpea combines prior knowledge in the form of protein-protein and domain-domain interactions with splicing changes to provide a view of a rewired network for a given experimental context. Splitpea defines a rewiring event when exon changes affect an underlying domain-domain interaction. Toy scenarios that would result in the three possible rewiring events predicted by Splitpea are illustrated in B.

the mean domain-level $\Delta\psi$ values. Essentially, the following pseudocode describes the crux of Splitpea's algorithm for a given sample with a set of exons with associated $\Delta\psi$ values:

```
for each PPI between $g_u, g_v$ do
  $\Psi^{(u)}$ := significant exons for gene u
  $\Psi^{(v)}$ := significant exons for gene v
  $D^{(u)} := \{d_u | d_u \in g_u, \exists \text{ exon}_i \text{ s.t. exon}_i \in \Psi^{(u)} \ \& \ \text{exon}_i \in d_u\}$
  $D^{(v)} := \{d_v | d_v \in g_v, \exists \text{ exon}_i \text{ s.t. exon}_i \in \Psi^{(v)} \ \& \ \text{exon}_i \in d_v\}$
  $w_{uv}$ := network rewiring edge weight between $g_u, g_v$
  $\delta_{uv}$ := direction classification of network rewiring between $g_u, g_v$
  for each DDI between $d_u \in D^{(u)}, d_v \in D^{(v)}$ do
    $\Delta\psi_{d_u} := \min(\{\Delta\psi_i | \text{ exon}_i \text{ s.t. exon}_i \in \Psi^{(u)} \ \& \ \text{exon}_i \in d_u\})$
    $\Delta\psi_{d_v} := \min(\{\Delta\psi_i | \text{ exon}_i \text{ s.t. exon}_i \in \Psi^{(v)} \ \& \ \text{exon}_i \in d_v\})$

  if $\forall d_u, d_v \in \text{DDI}(d_u, d_v), \Delta\psi_{d_u} > 0, \Delta\psi_{d_v} > 0$ then
    $\delta_{uv}$ = positive
  else if $\forall d_u, d_v \in \text{DDI}(d_u, d_v), \Delta\psi_{d_u} < 0, \Delta\psi_{d_v} < 0$ then
    $\delta_{uv}$ = negative
  else
    $\delta_{uv}$ = chaos
  $w_{uv} = \frac{1}{|D^{(u)}|+|D^{(v)}|}(\sum_{d \in D^{(u)}} \Psi^{(u)} + \sum_{d \in D^{(v)}} \Psi^{(v)})$

  return $w_{uv}, \delta_{uv}$
```

w_{uv} and δ_{uv} are reported as long as $\Psi^{(u)}$ *or* $\Psi^{(v)}$ is non-empty. Please note that the w_{uv} calculation only includes domains that have a DDI that is considered to be mediating the PPI between g_u, g_v. For readability, the equation above omits the removal of non-DDI pairs.

2.5. *Consensus network*

The main factor to consider when aggregating several sample-specific Splitpea networks into a consensus network is whether the directionality of edges agree. Thus, a "positive" consensus network and "negative" consensus network are built separately. "Chaos" edges are ignored since they are of ambiguous state. For each consensus network, two factors are considered for the edge weight: the sum of the original edge weights w_{uv} and how many networks support the same directionality δ_{uv}. The downstream analysis with each consensus network focuses on the largest connected component. As is common in biological networks, we found that the largest connected component covers the majority of the edges of the complete consensus network (breast cancer: 96.4% edges retained in negative consensus, 89.5% edges retained in positive consensus; pancreatic cancer: 96.1% edges retained in negative consensus, 88.8% edges retained in positive consensus).

2.6. *Network embedding and clustering*

To enable network clustering and other downstream uses of the Splitpea patient-specific networks, we created whole graph level embeddings. Here, we chose to focus only on potential gain-of-interaction edges and first filtered each patient-specific network accordingly. Taking the largest connected component, we applied the FEATHER[41] algorithm from the KarateClub NetworkX extension library[42] to generate an embedding for each network.

We clustered the resulting embeddings for each cancer type using hierarchical density-based clustering (HDBSCAN)[43] with minimum cluster sizes of 10. Clustering results were generally robust to the choice of the minimum cluster size parameter; 10 was chosen for downstream interpretability (and we would consider samples with fewer neighbors as outliers). Final plots were produced using principal component analysis (PCA), plotting all embeddings by their first two components.

3. Results

3.1. *Quantifying splicing changes in pancreatic and breast tumors*

In total, we collected data from TCGA covering 177 pancreatic primary tumors and 1,088 breast primary tumors, together with 192 normal pancreatic tissue and 218 normal breast tissue samples from GTEx that were used as a reference distribution of normal splicing variation for each respective cancer type. With these data, we calculated a $\Delta\psi$ value corresponding to the change in exon splicing in each tumor sample relative to its normal tissue background, resulting in $\Delta\psi$ estimates for a total of 139,661 unique exons across all breast cancer samples and 98,761 unique exons across the pancreatic cancer samples. Furthermore, we calculated an accompanying p-value that compares how extreme the observed ψ value for each exon in each cancer sample is relative to the corresponding background distribution of ψ values for normal tissue samples (see Methods).

3.2. $\Delta\psi$ values primarily reflect primary diagnoses

We then clustered the $\Delta\psi$ matrices for each tumor type and checked whether they corresponded to relevant clinical and pathological tumor features for both breast cancer (Fig. 2A) (pam50 subtypes, diagnosed type, pathologic stage, and age) and pancreatic cancer (Fig. 2B) (site of origin, diagnosed type, pathologic stage, age, and sex). While the majority of clinical features are not meaningfully clustered with $\Delta\psi$ values, we do observe that the most unique patient cluster for pancreatic cancer (far right columns in Fig. 2B) are all pancreatic neuroendocrine tumors. Neuroendocrine tumors are a rare subset of pancreatic cancers that originate not in the cells of the pancreas but in neuroendocrine cells. Interestingly, this cell type has commonality with neurons which are known to undergo more splicing changes.[44] For breast cancer, we see some clustering of lobular carcinomas (red cluster in "type" bar Fig. 2A), but otherwise do not see obvious patterns of clinical or pathological separation with $\Delta\psi$ values alone.

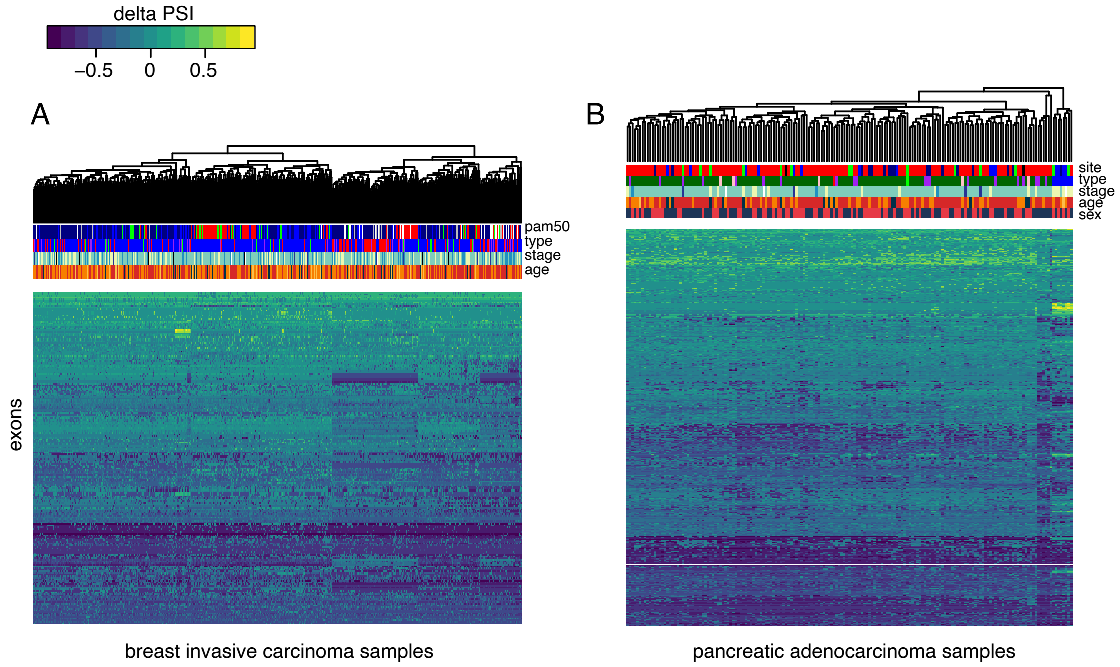

Fig. 2. *Clustering on $\Delta\psi$ values.* We cluster the $\Delta\psi$ values showing different sample groups for different spliced exons. Heatmaps depict splicing changes relative to average normal tissue background. Bar columns show known clinical information about each sample. In general, there are more subgroup level exon changes for breast cancer, (A) but these are not strongly correlated with any clinical variable. In pancreatic cancer, a small subset of neuroendocrine samples (B, dark blue) share similar splicing patterns. All other samples do not have obvious meaningful structure.

3.3. *Quantifying rewired protein-protein interactions for pancreatic and breast tumors*

We applied Splitpea to build patient-specific rewired PPI networks for 177 pancreatic and 1,088 breast primary tumor samples. Each PPI network contains three types of edges (gain, loss, or chaotic (mixed)) based on how underlying splicing changes may affect the individual protein-protein interaction (Fig. 3). In general, most splicing changes cause potential loss of protein interactions, though breast cancer had relatively fewer loss of edges proportionally on average (76% edges) than pancreatic cancer (84% of edges). Chaos (mixed) edges, where domain interactions have inconsistent directions per protein are relatively uncommon and comprise on average less than 2% of total edges for pancreatic and breast cancer. Between the two cancer types, breast cancer has more potential gain-of-interaction edges and a lower proportion of potential lost edges relative to pancreatic cancer. Interestingly, there is also more variability across edge types per sample in breast cancer samples.

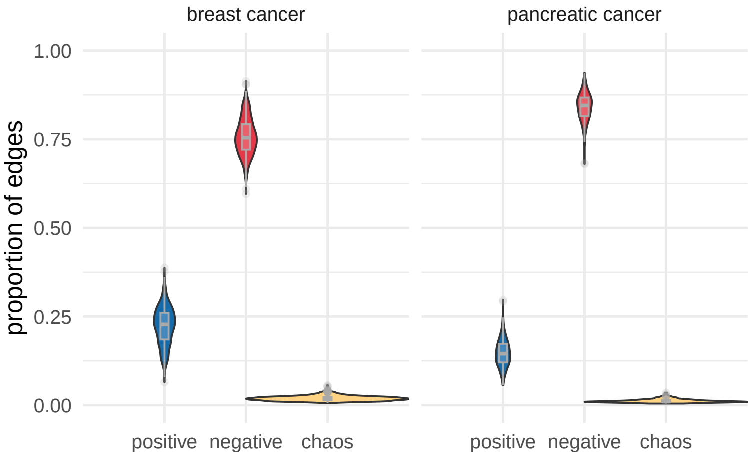

Fig. 3. *Proportion of relative gain and loss in edges across breast cancer and pancreatic cancer samples.* Breast cancer samples have proportionally more “gain of interactions” than pancreatic cancer samples, but in both cancer types, interaction loss is much more prevalent. For each TCGA cancer sample, the proportion of edges gained versus lost is calculated using the total number of edges in the largest connected component of the entire Splitpea rewired network (both directions) as the denominator. To be conservative, the number of edges retained in the largest connected components for the gain-only subnetwork and loss-only subnetworks are used as numerators.

Looking at individual patient networks (Fig. 4), we can see potential hubs and protein clusters that undergo extensive remodeling. In Fig. 4A, we show an example of one pancreatic tumor network with the most remodeling changes in the oncogene, RAB35, proto-oncogenes, HRAS and FYN, the signaling protein, MAPK3, the cell cycle and growth genes, NEDD8 and PRKAA1, among others. Breast cancer patient-specific networks have a different topology (Fig. 4C), though there is also overlap of proto-oncogenes HRAS and FYN.

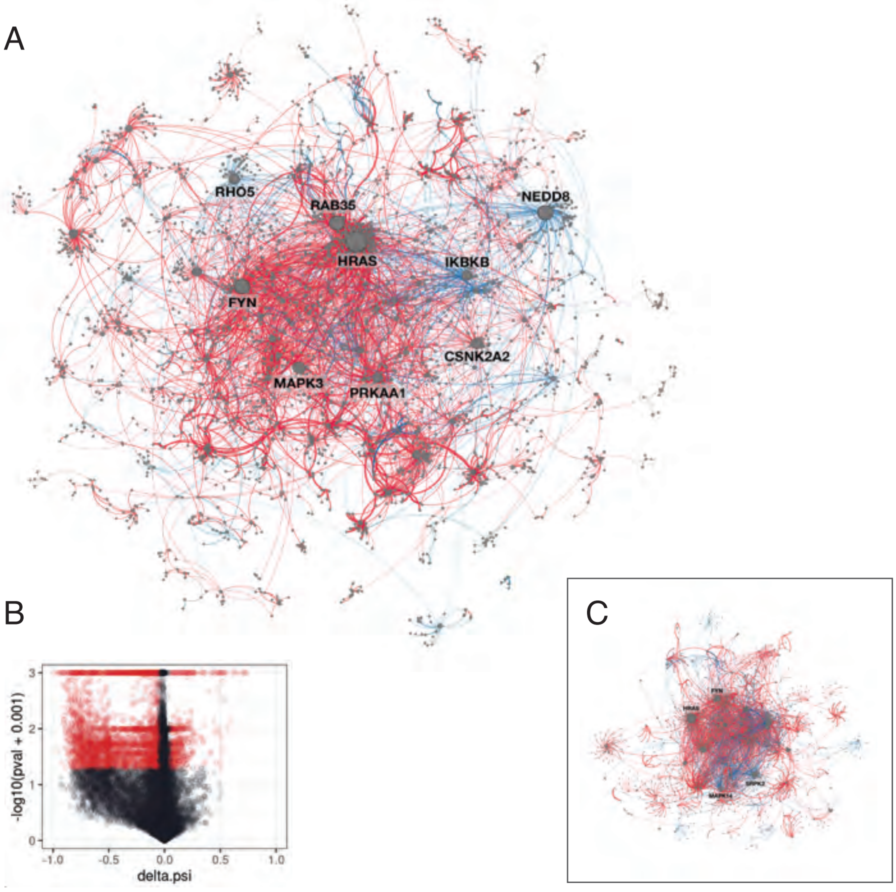

Fig. 4. *Patient specific rewired networks.* Here, we show two sample network outputs from Splitpea and the accompanying exon value cutoff. The large network (A) depicts pancreatic patient sample (TCGA-HZ-7918-01A-11R-2156-07), with edge losses in red and gains in blue. The corresponding volcano plot is shown in (B), where exons with significant $\Delta\psi$ ($\hat{p} < 0.05$) as well as absolute change ($|\Delta\psi| > 0.05$) are shown in red. Box (C) shows a patient-specific network for an example breast cancer sample, TCGA-BH-A0BG-01A-11R-A115-07, which exhibits a very different topology from the pancreatic sample in A.

3.4. *A consensus network of changes across breast cancer patients*

While patient-specific networks highlight network rewiring at the level of individual tumor samples, we also sought to look for more general cancer level patterns of PPI rewiring. Towards this end, we assembled a consensus rewiring network for breast cancer by taking splicing rewiring events conserved across 80% of patient samples and assembling a meta-network of these events. Edges were only preserved when their type (gain, loss) was consistent. Chaos edges were not included in the consensus network. Naturally, as the threshold increases, the number of genes preserved in the network decreases (Fig. 5A). Interestingly, up through the 80% threshold, gained edges are relatively more consistently preserved (Fig. 5B). Visualizing the breast cancer consensus network (Fig. 5C) revealed that the most gained interaction

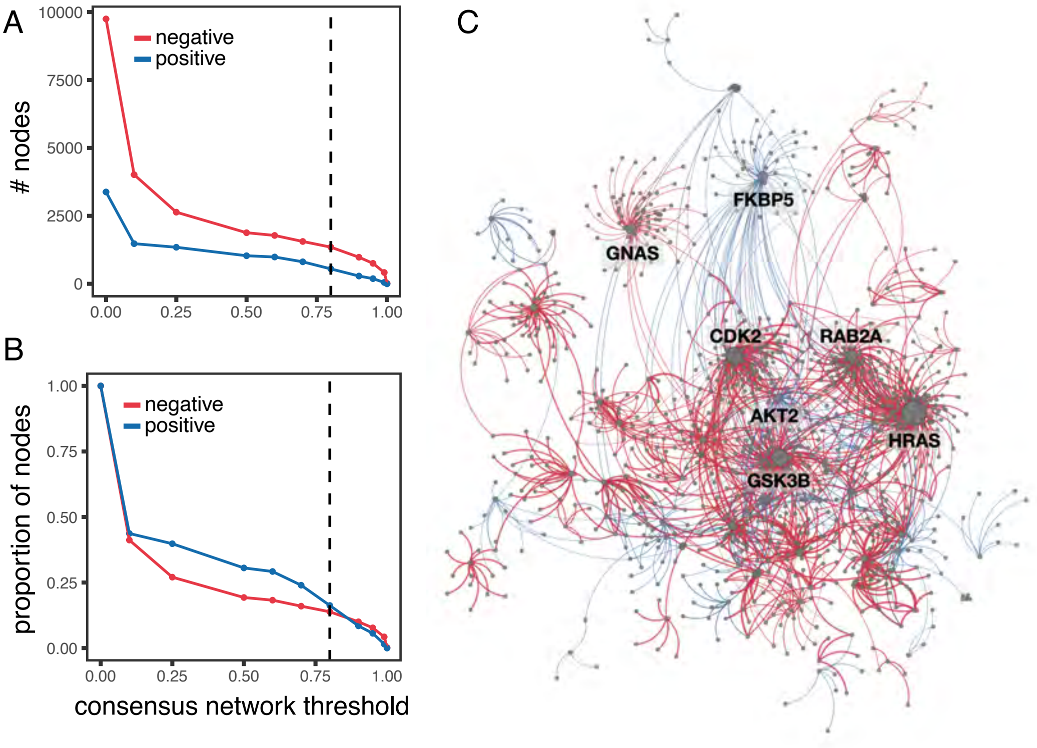

Fig. 5. *Meta-network of breast cancer patients.* The line graphs show the number of nodes preserved for different consensus thresholds (A) or the proportion of nodes relative to the non-thresholded consensus network (B) for edge loss (negative, red) and edge gain (positive, blue) events. The dashed line in both graphs denotes a threshold of 80%, corresponding to the visualization of the consensus network of splicing rewiring events conserved across 80% of breast cancer patient samples (C, red: edge loss; blue: edge gain).

involved the gene, FKBP5, which is an immune regulator responsible for protein trafficking and folding. This protein has been studied in breast cancer for its various hormone receptor signaling functions.[45]

3.5. *Network clusters reveal novel patient subgroups*

The patient-specific networks generated by Splitpea have many downstream applications, especially when the networks are used as features for other machine learning tasks. Here, we demonstrate their utility by finding patient subgroups across both breast and pancreatic cancer when the networks are clustered (Fig. 6). Specifically, we use a state-of-the-art graph embedding method, FEATHER,[41] which calculates characteristic functions using different random walk weights for node features, but any graph embedding method could be used for this type of analysis. For each cancer type, we clustered the network embeddings using HDBSCAN (see Methods). Interestingly, three distinct groups emerged across the cancer types (Fig. 6A). The dominant source of variation across the networks is the gain or loss of PPIs involving KRAS (Fig. 6B). Mutations in KRAS are known to affect subgroups of both pancreatic and breast cancer[46] with ties to prognosis. It is possible that splicing changes in interacting partner genes also induce changes to KRAS that may have yet unknown interaction effects with these somatic mutations, highlighting the potential of Splitpea to find additional disease subtypes. Furthermore, other interesting cancer drivers have distinct patterns of gains and losses, including RAB5A, which appears to have PPI gains in the BRCA outliers, and IKBKB, which is enriched for gains in the predominantly pancreatic cancer cluster 3.

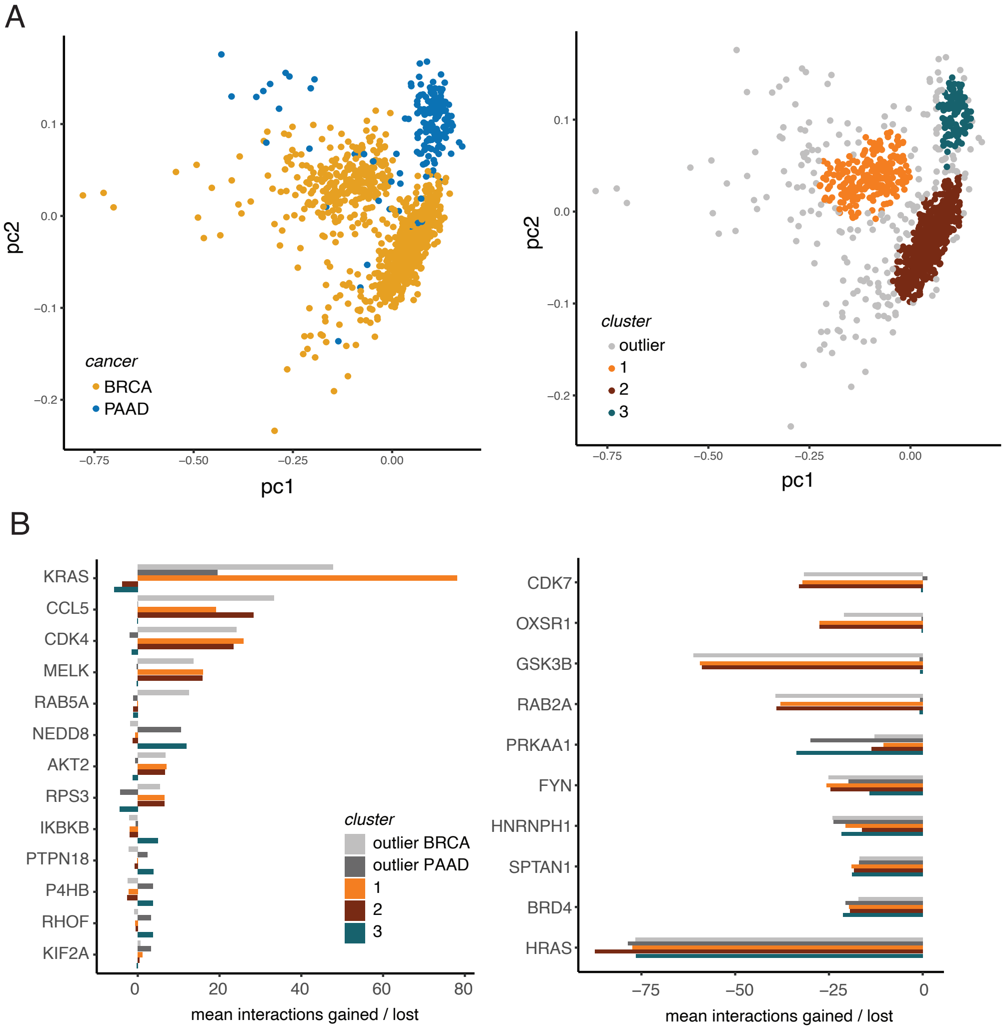

Fig. 6. *Splitpea networks cluster into distinct subgroups.* (A) PCA plots of graph embeddings of each patient-specific Splitpea network, with samples colored by either cancer type (left) or cluster (right). Clusters were assigned using HDBSCAN, with outliers colored in grey. (B) For each cluster, the top nodes undergoing the most changes (mean interactions gained or lost) were also identified. The bar graphs are roughly separated by genes that have the most gain of interactions (left) versus those that have primarily losses (right). Interestingly, the main variation captured in PC1 seems to be defined by networks that change in KRAS. Other cancer driver genes also undergo distinct patterns of gains and losses that drive clustering patterns.

4. Discussion and conclusion

We present a new method, Splitpea, for characterizing protein-protein network rewiring events. Splitpea is flexible and can be applied with different background contexts to highlight splicing changes between a disease and relevant background context of interest. We applied Splitpea to breast and pancreatic cancer samples to highlight the potential of Splitpea to find new and relevant cancer biology, both on an individual patient sample level and more broadly across samples of a single tumor type. To our knowledge, Splitpea is the first systematic method for identifying both potential gains in addition to PPIs lost for individual experimental samples.

Splitpea makes heavy use of existing knowledge of protein-protein interactions. Because of this, our method is inherently limited by the availability of known PPIs (which are largely incomplete), as well as DDIs, which are even less complete. As more of these are experimentally characterized, Splitpea will continue to improve, capturing more accurate and comprehensive sets of network rewiring events. Since we wrote Splitpea to be modular, updates to PPIs and DDIs can be easily integrated once they become available. Specifically, study bias is a well-reported issue in PPIs, and thus there is a large amount of overlap between well-studied nodes (including many cancer driver genes) with nodes of high degree in PPI networks, and given the dependency of Splitpea on reported PPIs, this also affects our results. As more systematic experimental PPI screens and more reliable PPI predictions become available, we can also readily adapt Splitpea.

We have only scratched the surface of cancer biology here. In our initial exploration of breast and pancreatic cancer, we have discovered subgroups and outliers within each cancer type that can be characterized by different network hubs. We believe this merits more thorough exploration, as it may carry important implications for precision medicine efforts. Beyond this, it will also be interesting to apply Splitpea to more cancer types and look for pan-cancer conservation patterns.

Acknowledgments

This work was supported by the Cancer Prevention & Research Institute of Texas (RR190065). VY is a CPRIT Scholar in Cancer Research.

References

1. Q. Pan, O. Shai, L. J. Lee, B. J. Frey and B. J. Blencowe, Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing, *Nature Genetics* **40**, 1413 (December 2008).
2. E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth and C. B. Burge, Alternative isoform regulation in human tissue transcriptomes, *Nature* **456**, 470 (November 2008).
3. D. L. Black, Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology, *Cell* **103**, 367 (October 2000).
4. T. W. Nilsen and B. R. Graveley, Expansion of the eukaryotic proteome by alternative splicing, *Nature* **463**, 457 (January 2010).
5. X. Yang, J. Coulombe-Huntington, S. Kang, G. M. Sheynkman, T. Hao, A. Richardson, S. Sun, F. Yang, Y. A. Shen, R. R. Murray, K. Spirohn, B. E. Begg, M. Duran-Frigola, A. MacWilliams,

S. J. Pevzner, Q. Zhong, S. A. Trigg, S. Tam, L. Ghamsari, N. Sahni, S. Yi, M. D. Rodriguez, D. Balcha, G. Tan, M. Costanzo, B. Andrews, C. Boone, X. J. Zhou, K. Salehi-Ashtiani, B. Charloteaux, A. A. Chen, M. A. Calderwood, P. Aloy, F. P. Roth, D. E. Hill, L. M. Iakoucheva, Y. Xia and M. Vidal, Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing, *Cell* **164**, 805 (February 2016).

6. F. E. Baralle and J. Giudice, Alternative splicing as a regulator of development and tissue identity, *Nature Reviews. Molecular Cell Biology* **18**, 437 (2017).
7. M. M. Scotti and M. S. Swanson, RNA mis-splicing in disease, *Nature Reviews. Genetics* **17**, 19 (January 2016).
8. H. Climente-González, E. Porta-Pardo, A. Godzik and E. Eyras, The Functional Impact of Alternative Splicing in Cancer, *Cell Reports* **20**, 2215 (August 2017).
9. A. Kahles, K.-V. Lehmann, N. C. Toussaint, M. Hüser, S. G. Stark, T. Sachsenberg, O. Stegle, O. Kohlbacher, C. Sander, S. J. Caesar-Johnson *et al.*, Comprehensive analysis of alternative splicing across tumors from 8,705 patients, *Cancer cell* **34**, 211 (2018).
10. J. E. Love, E. J. Hayden and T. T. Rohn, Alternative Splicing in Alzheimer's Disease, *Journal of Parkinson's Disease and Alzheimer's Disease* **2** (August 2015).
11. R. Mosca, A. Céol, A. Stein, R. Olivella and P. Aloy, 3did: a catalog of domain-based interactions of known three-dimensional structure, *Nucleic Acids Research* **42**, D374 (January 2014).
12. S. Yellaboina, A. Tasneem, D. V. Zaykin, B. Raghavachari and R. Jothi, DOMINE: a comprehensive collection of known and predicted domain-domain interactions, *Nucleic Acids Research* **39**, D730 (January 2011).
13. Y. Kim, B. Min and G.-S. Yi, IDDI: integrated domain-domain interaction and protein interaction analysis system, *Proteome Science* **10 Suppl 1**, p. S9 (June 2012).
14. R. D. Finn, B. L. Miller, J. Clements and A. Bateman, iPfam: a database of protein family and domain interactions found in the Protein Data Bank, *Nucleic Acids Research* **42**, D364 (January 2014).
15. A. Mehmood, A. Laiho, M. S. Venäläinen, A. J. McGlinchey, N. Wang and L. L. Elo, Systematic evaluation of differential splicing tools for rna-seq studies, *Briefings in bioinformatics* **21**, 2052 (2020).
16. M. A. Ghadie, L. Lambourne, M. Vidal and Y. Xia, Domain-based prediction of the human isoform interactome provides insights into the functional impact of alternative splicing, *PLoS computational biology* **13**, p. e1005717 (2017).
17. T. Will and V. Helms, Ppixpress: construction of condition-specific protein interaction networks based on transcript expression, *Bioinformatics* **32**, 571 (2016).
18. Z. Louadi, K. Yuan, A. Gress, O. Tsoy, O. V. Kalinina, J. Baumbach, T. Kacprowski and M. List, Digger: exploring the functional role of alternative splicing in protein interactions, *Nucleic acids research* **49**, D309 (2021).
19. E. Gjerga, I. S. Naarmann-de Vries and C. Dieterich, Characterizing alternative splicing effects on protein interaction networks with linda, *Bioinformatics* **39**, i458 (2023).
20. Y. Li, N. Sahni, R. Pancsa, D. J. McGrail, J. Xu, X. Hua, J. Coulombe-Huntington, M. Ryan, B. Tychhon, D. Sudhakar *et al.*, Revealing the determinants of widespread alternative splicing perturbation in cancer, *Cell reports* **21**, 798 (2017).
21. A. Chatr-aryamontri, R. Oughtred, L. Boucher, J. Rust, C. Chang, N. K. Kolas, L. O'Donnell, S. Oster, C. Theesfeld, A. Sellam, C. Stark, B.-J. Breitkreutz, K. Dolinski and M. Tyers, The BioGRID interaction database: 2017 update, *Nucleic Acids Research* **45**, D369 (January 2017).
22. L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie and D. Eisenberg, The Database of Interacting Proteins: 2004 update, *Nucleic Acids Research* **32**, D449 (January 2004).
23. G. Alanis-Lobato, M. A. Andrade-Navarro and M. H. Schaefer, HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks, *Nucleic Acids Research* **45**,

D408 (January 2017).

24. T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady and A. Pandey, Human Protein Reference Database–2009 update, *Nucleic Acids Research* **37**, D767 (January 2009).
25. T. Rolland, M. Taşan, B. Charloteaux, S. J. Pevzner, Q. Zhong, N. Sahni, S. Yi, I. Lemmens, C. Fontanillo, R. Mosca, A. Kamburov, S. D. Ghiassian, X. Yang, L. Ghamsari, D. Balcha, B. E. Begg, P. Braun, M. Brehme, M. P. Broly, A.-R. Carvunis, D. Convery-Zupan, R. Corominas, J. Coulombe-Huntington, E. Dann, M. Dreze, A. Dricot, C. Fan, E. Franzosa, F. Gebreab, B. J. Gutierrez, M. F. Hardy, M. Jin, S. Kang, R. Kiros, G. N. Lin, K. Luck, A. MacWilliams, J. Menche, R. R. Murray, A. Palagi, M. M. Poulin, X. Rambout, J. Rasla, P. Reichert, V. Romero, E. Ruyssinck, J. M. Sahalie, A. Scholz, A. A. Shah, A. Sharma, Y. Shen, K. Spirohn, S. Tam, A. O. Tejeda, S. A. Trigg, J.-C. Twizere, K. Vega, J. Walsh, M. E. Cusick, Y. Xia, A.-L. Barabási, L. M. Iakoucheva, P. Aloy, J. De Las Rivas, J. Tavernier, M. A. Calderwood, D. E. Hill, T. Hao, F. P. Roth and M. Vidal, A proteome-scale map of the human interactome network, *Cell* **159**, 1212 (November 2014).
26. S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, G. Chavali, C. Chen, N. del Toro, M. Duesbury, M. Dumousseau, E. Galeota, U. Hinz, M. Iannuccelli, S. Jagannathan, R. Jimenez, J. Khadake, A. Lagreid, L. Licata, R. C. Lovering, B. Meldal, A. N. Melidoni, M. Milagros, D. Peluso, L. Perfetto, P. Porras, A. Raghunath, S. Ricard-Blum, B. Roechert, A. Stutz, M. Tognolli, K. van Roey, G. Cesareni and H. Hermjakob, The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases, *Nucleic Acids Research* **42**, D358 (January 2014).
27. S. Razick, G. Magklaras and I. M. Donaldson, iRefIndex: a consolidated protein interaction database with provenance, *BMC bioinformatics* **9**, p. 405 (September 2008).
28. P. Pagel, S. Kovac, M. Oesterheld, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, P. Mark, V. Stümpflen, H.-W. Mewes, A. Ruepp and D. Frishman, The MIPS mammalian protein-protein interaction database, *Bioinformatics (Oxford, England)* **21**, 832 (March 2005).
29. G. R. Brown, V. Hem, K. S. Katz, M. Ovetsky, C. Wallin, O. Ermolaeva, I. Tolstoy, T. Tatusova, K. D. Pruitt, D. R. Maglott and T. D. Murphy, Gene: a gene-centered information resource at NCBI, *Nucleic Acids Research* **43**, D36 (January 2015).
30. S. Durinck, P. Spellman, E. Birney and W. Huber, Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart., *Nature Protocols* **4**, 1184 (2009).
31. H. Li, Tabix: fast retrieval of sequence features from generic TAB-delimited files, *Bioinformatics (Oxford, England)* **27**, 718 (March 2011).
32. Y. Pan, J. W. Phillips, B. D. Zhang, M. Noguchi, E. Kutschera, J. McLaughlin, P. A. Nesterenko, Z. Mao, N. J. Bangayan, R. Wang, W. Tran, H. T. Yang, Y. Wang, Y. Xu, M. B. Obusan, D. Cheng, A. H. Lee, K. E. Kadash-Edmondson, A. Champhekar, C. Puig-Saus, A. Ribas, R. M. Prins, C. S. Seet, G. M. Crooks, O. N. Witte and Y. Xing, IRIS: Discovery of cancer immunotherapy targets arising from pre-mRNA alternative splicing, *Proceedings of the National Academy of Sciences* **120**, p. e2221116120 (May 2023), Publisher: Proceedings of the National Academy of Sciences.
33. S. Shen, J. W. Park, Z.-x. Lu, L. Lin, M. D. Henry, Y. N. Wu, Q. Zhou and Y. Xing, rmats: robust and flexible detection of differential alternative splicing from replicate rna-seq data, *Proceedings of the National Academy of Sciences* **111**, E5593 (2014).
34. A. Kahles, C. S. Ong, Y. Zhong and G. Rätsch, Spladder: identification, quantification and testing of alternative splicing events from rna-seq data, *Bioinformatics* **32**, 1840 (2016).

35. J. Vaquero-Garcia, A. Barrera, M. R. Gazzara, J. Gonzalez-Vallinas, N. F. Lahens, J. B. Hogenesch, K. W. Lynch and Y. Barash, A new view of transcriptome complexity and regulation through the lens of local splicing variations, *elife* **5**, p. e11752 (2016).
36. Y. I. Li, D. A. Knowles, J. Humphrey, A. N. Barbeira, S. P. Dickinson, H. K. Im and J. K. Pritchard, Annotation-free quantification of rna splicing using leafcutter, *Nature genetics* **50**, 151 (2018).
37. J. L. Trincado, J. C. Entizne, G. Hysenaj, B. Singh, M. Skalic, D. J. Elliott and E. Eyras, SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions, *Genome Biology* **19**, p. 40 (March 2018).
38. J. L. Trincado, J. C. Entizne, G. Hysenaj, B. Singh, M. Skalic, D. J. Elliott and E. Eyras, Suppa2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions, *Genome biology* **19**, 1 (2018).
39. G. R. Warnes, B. Bolker, L. Bonebakker, R. Gentleman, W. Huber, A. Liaw, T. Lumley, M. Maechler, A. Magnusson, S. Moeller, M. Schwartz and B. Venables, gplots: Various r programming tools for plotting data (2015).
40. D. Netanely, A. Avraham, A. Ben-Baruch, E. Evron and R. Shamir, Expression and methylation patterns partition luminal-A breast tumors into distinct prognostic subgroups, *Breast Cancer Research* **18**, p. 74 (July 2016).
41. B. Rozemberczki and R. Sarkar, Characteristic functions on graphs: Birds of a feather, from statistical descriptors to parametric models, in *Proceedings of the 29th ACM international conference on information & knowledge management*, 2020.
42. B. Rozemberczki, O. Kiss and R. Sarkar, Karate club: An api oriented open-source python framework for unsupervised learning on graphs, in *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20)*, 2020.
43. R. J. Campello, D. Moulavi and J. Sander, Density-based clustering based on hierarchical density estimates, in *Pacific-Asia conference on knowledge discovery and data mining*, 2013.
44. G. Yeo, D. Holste, G. Kreiman and C. B. Burge, Variation in alternative splicing across human tissues, *Genome Biology* **5**, p. R74 (September 2004).
45. L. Li, Z. Lou and L. Wang, The role of FKBP5 in cancer aetiology and chemoresistance, *British Journal of Cancer* **104**, 19 (January 2011).
46. I. A. Prior, F. E. Hood and J. L. Hartley, The frequency of Ras mutations in cancer, *Cancer research* **80**, 2969 (July 2020).