Supporting Information

for

Molecular signatures of fusion proteins in cancer

Natasha S. Latysheva and M. Madan Babu

MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, United Kingdom

10.1021/acsptsci.9b00019

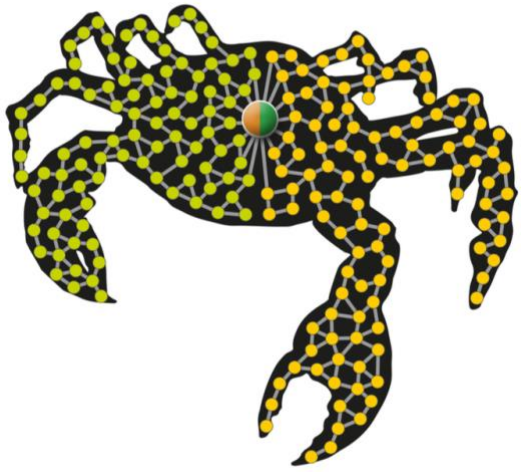


Table of contents

Online Methods

Data acquisition and integration

119 features covering a broad range of structural, functional, regulatory, expression and interaction were acquired from 25 data sources (**Table S1**). Using custom R scripts, datasets were cleaned, processed, and integrated into the Ensembl [1] framework.

Gene fusion dataset

A list of 2371 unique in-frame fusion events detected in a recent, complete transcriptomic screen of 675 human cancer cell lines [2] was acquired (of which 2358 fusion events were successfully integrated); fusion candidate quality checks included requiring the presence of multiple breakpoint-spanning reads in the sequencing experiments, absence of the fusion event in normal (non-diseased) tissues, and the potential of the fusion transcript to undergo in-frame translation [2]. In this study, we designate those genes that form gene fusions as "parent genes" and the corresponding proteins they encode as "parent proteins". Due to the protein-level focus of this study, we occasionally use "parent proteins" interchangeably with "parent genes". In total, 3161 distinct parent gene symbols were mapped to 3151 protein-coding Ensembl gene IDs (retaining only the longest isoform per gene). Non-protein coding parent genes were identified as Ensembl gene IDs which did not correspond to any Ensembl protein entries (as catalogued by the BioMart portal [3]) and were excluded from further analysis.

Parent genes and fusion events were classified into several categories. Parent genes were classed as "recurrent" if they formed at least 2 different gene fusions, and "non-recurrent" otherwise. Different levels of recurrence (i.e. parent genes are only classed as "recurrent" if they form ≥ 3, ≥ 4, or ≥ 5 gene fusions), together with their associated properties, were also investigated at the level of feature values and classification accuracies. Fusion events were labeled as "metastatic" if they occurred in at least 1 cell line of metastatic tumour origin [2], and "primary" if they only occurred in cell lines derived from primary tumours. Cancer types for fusion events were taken from the tissue annotation of cell lines, and cancer types with $n<100$ (n=486) were reassigned to the "Other" cancer type category and omitted from further analysis. The remaining 8 cancer type categories, together with the total counts of fusion events by category, were: lung, n=679; breast, n=328; ovary, n=183; lymphoid, n=166; colo-rectal, n=145; brain, n=142; head-neck, n=123; skin, n=108. To ensure approximately similarly sized samples for the purposes of classification, the "lung", "breast", and "ovary" fusion events were randomly downsampled without replacement to reach n=166, and the test set size was increased to 40% from 30% to improve confusion matrix resolution.

Gene and protein datasets

Human genome and proteome data (genome build GRCh38.p5) was acquired from the Ensembl database – specifically, human genes in GTF format (60675 unique genes) and protein FASTA sequences were downloaded from the Ensembl ftp page. Sequences were processed using the seqinR package [4] (version 3.1-3) for R. In total, 20295 protein-coding Ensembl genes, corresponding to 93892 Ensembl protein IDs, were identified and used as the base information for integrating other protein-level data and gene-level functional features. Protein lengths were calculated by processing FASTA sequences, and average exon counts and protein-coding transcript counts per Ensembl protein isoform were retrieved using the BioMart portal [3].

Cancer gene, oncogene, and tumour suppressor gene datasets

A collection of 1571 cancer genes was acquired from the Network of Cancer Genes resource [5] (version 5.0), which manually curated 175 studies to identify genes associated with cancer. Additionally, a more stringent list of 572 cancer genes was retrieved from the Cancer Gene Census provided by the COSMIC resource [6] (accessed March 2016), with accompanying information concerning the mutation type (somatic or germline) and tissue

types (leukemia/lymphoma, epithelial, mesenchymal, other) affected by the mutations. A list of 1217 tumour suppressor genes (TSGs) was acquired from the TSGene 2.0 database [7], which also contained information on which TSGs were downregulated and across cancer types. A total of 985 TSGs were labeled as being downregulated in 1 or more of 11 available cancer types. Finally, the identity of 239 human oncogenes (OGs) was retrieved from the Tumour Associated Gene (TAG) database [8], of which 216 were protein-coding.

Gene multifunctionality

Gene ontology terms for all human Ensembl genes were acquired from the Gene Ontology (GO) Consortium [9] via BioMart. GOSlim terms, which offer a broader annotation of gene function, were also acquired. GO and GOSlim accession counts per gene were tallied as simple measures of gene multifunctionality.

Gene essentiality

A list of 2750 essential human genes was extracted from the Online Gene Essentiality database [10], which derived its essentiality annotation from the collection and analysis of large-scale, genome-wide experimental data, supplemented with text mining.

Dosage sensitivity and loss of function upon haploinsufficiency data

373 haploinsufficient genes in human, as well as curated annotation on whether gene loss results in a loss of function, were downloaded from the Clinical Genome Resource consortium [11] ftp site (ftp://ftp.ncbi.nlm.nih.gov/pub/dbVar/clingen/).

Protein kinase dataset

Human kinase genes and kinase classifications were retrieved from the Uniprot [12] kinase resource (http://www.uniprot.org/docs/pkinfam; March 2016 release), based on the Kinase Database [13] and previous kinase classification work [14]. A total of 509 kinase Uniprot accessions were mapped to 515 Ensembl gene IDs. All "atypical" subgroups of kinases were grouped together for a total of 11 kinase classes (**Table S1**).

Transcription factors and associated regulatory interactions

A set of 748 human transcription factors (TFs) and 8,015 TF-mediated regulatory links (i.e. activating or repressive relationship between a TF and target gene) was acquired from the TRRUST database [15]. The TF-target regulatory interactions in TRRUST were identified via manual curation of Medline abstracts [15]. Counts of the total number of unique TF-mediated regulatory links (including links of an "Unknown" regulatory nature) were tallied per Ensembl gene, as were the number of activating and repressive links.

Epigenetic complexes and chromatin modifying genes

A list of epigenetic modifier genes (i.e. genes involved in DNA modification, histone modification, or chromatin remodelling) and subunits of epigenetic complexes were downloaded from the EpiFactors database [16] (version 1.7.1). Data was cleaned, and a total of 770 Ensemble gene IDs with epigenetic gene functions were identified. To simplify gene function labels, the number of categories was reduced from 56 to 3 (chromatin modifying, n=151; histone modification, n=523; other, n=112, largely RNA and DNA processing and histone chaperones) using a combination of regular expressions and manual curation. Further, using the same resource, 302 Ensemble gene IDs were identified as participating in 69 epigenetic complexes.

Subcellular localization of proteins

Subcellular localization data for all human proteins was extracted from the COMPARTMENTS database [17]. The "Knowledge" and "Experiments" data channels were downloaded and combined to yield 356,960 localization tags for 16,189 Ensembl protein IDs. The number of localization categories (1,656) was reduced by choosing 10 categories from the 221 categories containing 100 or more proteins on the basis of their breadth and biological interest: "Cytoplasm" (n=10508 protein IDs), "Nucleus" (8114), "Plasma membrane" (4482),

"Organelle lumen" (4135), "Vesicle" (3731), "Endomembrane system" (3563), "Extracellular region" (3367), "Cytoskeleton" (2273), "Nucleolus" (1587) and "Cell junction" (1092). We find that 1 or more of these categories applied to 14,843 Ensembl gene IDs.

Protein domain dataset

A collection of all domains and domain families in the human proteome was acquired from the Pfam database [18] using BioMart. The number of Pfam domains per Ensembl protein, the average length of Pfam domain per protein, and the density of Pfam domains (domain count over protein length) were calculated. These values were averaged over isoforms to obtain values per gene.

Intrinsic structural disorder

Residue by residue predictions for intrinsic structural disorder in all human proteins were acquired by running the IUPred program [19] on FASTA sequences (using the "long" option). Intrinsic disorder scores were calculated for genes as an average over isoforms.

Interaction-mediating protein segments and domains

Interaction-mediating segments of proteins were acquired from Interactome3D [20] (May 2015 release). Only the "representative" set of interactions and associated segments, comprising the top ranking structures and models, were used. Per Uniprot protein, unique stretches were identified using protein accessions and start and stop coordinates, the number of interaction-mediating stretches counted, and the length of the longest stretch identified. For the 35 Uniprot accessions that mapped to multiple Ensembl gene IDs, the maximum values for these quantities was taken and assigned to the gene. Additionally, the INstruct database of curated protein-protein interactions resolved to the level of protein domains was acquired [21], and the number and density of these interaction-mediating domains per protein were calculated.

Structural interface-forming residues

Protein residues involved in the formation of macromolecular interfaces were acquired from the Protein Interfaces, Surfaces and Assemblies (PISA) database [22], which collects and details macromolecular interfaces in the Protein Data Bank [23], were acquired as in [24]. The number and density of PISA residues were summarized per protein.

Linear motif data

A list of 1548 human linear motifs (LMs) was extracted from the Eukaryotic Linear Motif (ELM) database [25] and filtered down to 1538 by finding entries with unique combinations of Uniprot accessions and start and end coordinates. The LM analysis was expanded by taking a further 1,036,282 putative short protein-binding regions, previously identified by the ANCHOR program [26] (as in [24]).

Post-translational modification sites

Post-translational modification (PTM) sites were acquired from the dbPTM resource [27] (version 3.0; accessed March 2016). Only unique PTM locations were considered, yielding 118047 unique PTMs in human, and PTM counts and densities per gene were calculated. Ubiquitination sites were acquired by subsetting dbPTM. Phosphorylation site subtypes were grouped into one category using regular expressions. Furthermore, a set of PTM sites which putatively function in the regulation of protein-protein interactions was acquired from the PTMcode database [28] (v2). Gene-wise counts and densities of ubiquitination, phosphorylation, and PTMcode PTM sites were calculated.

Cancer pathway involvement

A list of genes involved in human cancer pathways (KEGG pathway id: hsa05200) was acquired from the KEGG [29,30] database (http://www.kegg.jp/). A total of 402 Ensemble genes were identified as participating in cancer signaling pathways.

General and tissue-specific protein interactions

Genome-wide physical protein-protein interactions (PPIs) were acquired from the BioGRID database [31] (release 3.4.135, March 2016), converted to Ensembl accessions, and unique interactions taken. The centrality of all genes (nodes) in the network was summarized using several common metrics (degree, betweenness, closeness, and eigenvector centrality), which capture different notions of a node's importance in a graph (**Table S1**). For each gene, the number of interactions with known oncogenes, tumour suppressor genes, and cancer genes was identified through the (non tissue-specific) BioGRID PPI set, using cancer gene set annotation (see above).

Gene expression levels and tissue-specificity of expression

Averaged, log-transformed tissue-specific gene expression values in Reads Per Kilobase per Million (RPKM) covering 27 human tissues (**Table S1**) from a study analyzing several RNA-seq data sources was obtained [32]. Mean and maximum expression levels per gene across all tissues were also acquired. One popular metric for calculating expression heterogeneity is tau [33], which has been shown to be superior to several other methods due to its robustness and ability to separate ubiquitously-expressed (sometimes called "housekeeping") and tissue-specific genes [32]. Tau (τ) ranges from 0 (broadly expressed) to 1 (entirely tissue-specifically expressed) and is defined for each gene by:

$$\tau = \frac{\sum_{i=1}^{n}(1 - \hat{x}_i)}{n - 1}$$

where n is the total number of tissues and $\hat{x}_i$ is defined to be:

$$\hat{x}_i = \frac{x_i}{\max\limits_{1 \le i \le n} x_i}$$

where x_i is a gene's expression in tissue i.

For simplicity, we omit the tissue-specific (TS) expression values and retain only the two derived summary values derived (i.e. average and maximum expression over all tissues), were retained.

Missing data handling and imputation

The compiled dataset (**Table S1**), 10 variables possessed missing values, namely: degree_centrality, betweenness_centrality, closeness_centrality and eigenvector_centrality (n missing/total = 4978/20295); tau, mean_expression and max_expression (2377/20295); num_GO_terms and num_GOSlim_terms (1576); and avg_disorder (1278). Gene set label features (e.g. is_oncogene, is_cancer_gene, is_kinase) have no missing values, since all genes without an explicit gene set labels are deemed to be non-members of the gene set (i.e. genes which are not known kinases are assumed to be known non-kinases). Missing values were imputed using multiple imputation by chained equations [34], also called MICE, which generates native-like, model-based imputations for each variable with missing data [35]. The R MICE package [36] was employed for the imputation procedure, using predictive mean matching, with parameters *m* (number of multiple imputations) and *maxit* (maximum number of iterations) set to 5.

Identifying parent protein and fusion event functional groups using PCA and hierarchical clustering

To improve cluster interpretability and stability, a dimensionality reduction method (principal components analysis; PCA) was employed as a preprocessing step before hierarchical clustering. PCA involves a linear transformation of the data, specifically the projection of a set of potentially correlated variables into orthogonal (linearly uncorrelated) axes called principal components (PCs). As a result of the PCA procedure, the dataset is transformed into

a new coordinate system, whereby the most variance by projection occurs in the direction of the first principal component, then the second, and so on up to *p* PCs, the dimensionality of the original dataset. Formally, PCA proceeds by finding the eigenvalue decomposition of the symmetric covariance matrix, where the PCs are the eigenvectors and the corresponding eigenvalues capture the amount of variance present in the data in that direction. The eigenvector with the highest eigenvalue is the first PC, and the eigenvalues of the different PCs can be visualized by creating a Scree plot. These principal components are by definition uncorrelated. Hence, PCA acts as a reframing of data using new axes that better capture variance. PCA is frequently used for dimensionality reduction, since considering fewer than p principal components generates a lower dimensional space that still captures a large amount of the variance observed in the original data. This is useful for visualization and for revealing dominant trends in complex data using only a few interpretable latent variables. PCs can be interpreted by examining which of the original variables correlate most heavily with the PC values. The first portion of PCs (e.g. the first 10 PCs) can be used in place of the original dataset as input into a clustering procedure.

Two separate clustering tasks were performed in this study: the identification of clusters of similar parent proteins, and the identification of clusters of similar fusion events (i.e. pairs of parent protein features). PCA was first performed on the datasets using the R package FactoMineR [37]. Variables that were >90% sparse were omitted from the PCA and clustering analysis. The first 10 PCs were clustered using agglomerative hierarchical clustering with Ward's method on Euclidean distances. The optimal number of clusters (based on relative inertia loss levels across different tree cuts) suggested by FactoMineR was accepted and individual genes were plotted on PC1 and PC2 axes and colored by cluster. Outlier parent proteins (n=7 parent proteins, n=6 fusion events) were identified as instances assigned to clusters containing <10 members when n_clust was set to 10, and these instances were omitted. PCs were characterized by describing feature correlations with PC values, where the correlation coefficient $|r| \geq 0.4$. Distributions of the top correlated features in the first 4 PCs were visualized by cluster. A set of 5 "paragon" examples, defined to be instances closest to cluster centroids were generated for each cluster.

A chi-squared test of independence was performed on the frequency matrix of parent protein clusters by 5' and 3' parent cluster membership. As an additional analysis of whether certain parent protein clusters fuse disproportionately often, a randomization analysis was conducted by generating 1000 random gene fusions by bootstrap sampling two sets of parent genes, and calculating the proportion of fusions by parent cluster in each batch. 10000 batches were generated to construct a null distribution of same cluster fusion proportions, and empirical p values were calculated. Enrichment for cancer types by fusion event cluster was examined with a chi-squared test on a contingency test for cancers in which all cell counts were $n \geq 20$.

Functional enrichment analysis

Gene sets corresponding to the three clusters of parent proteins were tested for enrichments of GOSlim molecular functions using PantherDB [38], using the set of all human genes as background, with p values corrected for multiple testing using Bonferroni's correction. For each gene set, the log of fractional difference (observed versus expected number of genes) was calculated. Categories with <10 genes were disregarded.

Fusion of metabolic pathways

Gene to REACTOME pathway and cellular process mappings for 8088 Ensembl genes were obtained using Ensembl BioMart, and REACTOME pathway IDs were mapped to descriptions using annotation from the REACTOME website (http://www.reactome.org/). Of the 2,371 gene fusions in our dataset, 577 possessed available pathway annotations for both parent genes (572 with fusion event outliers excluded); in these cases, the frequencies with which each pair of pathways was joined via fusion was calculated (pathway counts per gene ranged from 1 to 334, IQR=4-9, median=5, mean=11.01). To estimate null expected

frequencies of pathway fusions (i.e. the frequencies with which any given pair of pathways would be expected to fuse if fusion events were randomly generated from any protein-coding genes), 1,000,000 random fusions with pathway annotation were simulated and pathway fusion frequencies calculated. A repeated set of calculations was performed whereby random fusions were generated from only the set of parent proteins. Enrichments were calculated as the $\log_{10}$ of observed/random pathway fusion frequencies (log chosen to normalize enrichment distribution), and to increase stability, only where pairs were fused ≥10 times. In the cancer-type specific analyses, pathway pairs were required to be fused ≥5 times and only in lung, lymphoid, head-neck, breast, and ovary due a lack of annotation availability in other cancer types. By-cluster trends in pathway fusion were investigated and enrichments calculated (cluster 1 n=184, cluster 2 n=232, cluster 3 n=156). In the main text figures, stroke weights under 1 were increased to 1 for improved visibility. We note that fusion does not necessitate that pathways necessarily come into contact (e.g. both parent proteins could be too severely truncated), but instead provides a broad overview of affected pathways; furthermore, since just under a third of protein-coding genes have REACTOME pathway participation annotations, the findings of the enrichment calculations may change as coverage and annotation improve.

Class balancing

Given the unbalanced nature of the datasets (e.g. most genes are not parent genes, most parent genes are not recurrent), and the issues inherent to building models on unbalanced data (i.e. signal loss and an inflated sense of model performance due to many algorithms tending towards universally classifying data points as the majority case), datasets were balanced with respect to the target class frequencies. Outcome class balancing was done by undersampling the majority classes without replacement until minority class frequencies were matched.

Feature normalization, categorical variable handling, and train/test splitting

Random forest models are scale invariant and were fed raw feature data, whereas numerical features input to logistic regression models were scaled and centered to have a mean of 0 and variance of 1. To ensure symmetrical treatment, all categorical features were encoded into as many binary dummy variables as there were factor levels. Unless stated otherwise, in each classification task, 70% of the data was used as the training set and 30% was held out for testing.

Predictive modeling and feature importance ranking

Random forest (RF) and penalized/regularized logistic regression (RLR) models generate, as part of their model structure, intuitive measures of the relative predictive importance of variables (see below). The two models have highly different mathematical formulations and may therefore capture different predictive signals within the datasets, and together generate a diversified view of feature importance. The 10 highest importance features from the RF and RLR models were acquired for each prediction task and assigned integer ranks from 10 (most important) to 1 (least important). These ranks were then added and displayed by feature, and distributions of the most informative features were visualised with violin and boxplots. Statistical significance of differences between continuous distributions was assessed using Wilcoxon rank sum tests with continuity corrections. For categorical variables, contingency tables were constructed and Pearson's Chi-squared test with Yates' continuity correction was used to assess variable dependence; if counts of any cell were less than 30, Fisher's exact test for count data (H_0=two sided) was used instead. In the few subplots where boxplots and violin plots were not appropriate or visible (e.g. although the count_tissues_cancer_mutation variable is continuous, 97.3% of its values are 0, which renders continuous distribution geometries inappropriate), means were plotted by category instead.

Modeling with random forests

Random forests, a type of recursive partitioning model [39], are powerful and robust tools for classification and regression [40,41]. Random forests ensemble together base learner classification trees, each of which recursively bifurcates bootstrapped data by splitting on

features that maximize class purity at the terminal nodes (**Figure 7**). Each bifurcation considers a sample of the available features, and optimal splits are identified by greedily minimizing a purity criterion, which measures the homogeneity of the target variable. Commonly, the purity criterion used is either Gini impurity or entropy loss. The Gini index is given by:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

which measures the total variance across K classes, where $\hat{p}_{mk}$ is the proportion of training examples in the *m*th region from the *k*th class. The Gini index approaches 1 if all $\hat{p}_{mk}$ are close to zero or 1 (i.e. the tree terminal regions are highly pure). An alternative popular measure of purity is information gain (termed "entropy" in the scikit-learn implementation of a random forest classifier), defined as:

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log_2 \hat{p}_{mk}$$

Random forest models are characterized by several hyperparameters, the most important of which are the choice of purity criterion, the maximum tree depth, and the maximum number of features considered at each split. These values were optimized for each classification task using grid search and 10-fold cross-validation. Cross-validation (CV) is a method for estimating model performance using only the training data, and is a powerful tool for hyperparameter optimization: individual models (i.e. models using different values for the hyperparameters) can be trained on some portion of the data (e.g. $1/10^{th}$ in the case of 10-fold CV) and tested on the remaining training data. The procedure is then repeated for an appropriate number of times until the entire training data has been used for model training (e.g. 10 times in the case of 10-fold CV, using each $1/10^{th}$ of the training data exactly once). The optimal hyperparameter setting from the grid of possible options is the one that maximizes the average performance metric of interest across the CV rounds. Scikit-learn implements Gini importance [42,43] for measuring feature importance in random forest models, defined as the decrease in node impurity when splitting on the feature, weighted by the probability of data points reaching the node, and averaged over the base learner trees in the ensemble. The ten most important features in each classification task were identified from these feature importance rankings.

Modeling with regularized in regression

L1 regularized multiple logistic regression is a common and efficient binary classification method [44], even in the presence of many irrelevant features. The algorithm estimates the natural logarithm of the odds ratio that a data point will belong to class 1 based on the values of predictors $X_1, X_2, \ldots, X_p$. The probability of being in class 1 *(*$p(X = 1)$, or simply $p(X)$ is given by the sigmoidal logistic function:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X}}$$

where the probability of being in class 0 is 1 minus this quantity. The logistic regression model has a logit that is linear in X:

$$\log \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

where the maximum likelihood fitted coefficients $\beta_1, \beta_2, \ldots, \beta_p$ represent the change in log odds of being class 1 per unit change in the corresponding X_1, $X_2, \ldots, X_p$. Model training involves estimating the values of the coefficients. As with other regression methods, logistic regression is sensitive to overfitting when used on high dimensional feature spaces. There is an additional issue of potential instability in coefficient values where non-negligible levels of correlation are present. Regularization techniques alter the standard regression loss function (such as root mean squared error) in order to penalize coefficient size by shrinking the maximum likelihood estimates towards 0. Penalization promotes model stability and increases generalization, and has the general regularization form of:

$$\min\left\{\frac{1}{n}\sum_{i=1}^{n} l\big(y_i, f(x_i)\big) + \lambda||f||_k\right\}$$

in which $(x_i, y_i)_{i=1}^{n}$ is the dataset, $f(x_i)$ are the predicted values, $l\big(y_i, f(x_i)\big)$ is the loss function, and λ is the regularization parameter. k specifies the regularization type: k=0 denotes the Akaike information criterion or Bayes information criterion; k=1 results in L_1 regularization or (in the case of linear regression) the Lasso; k=2 results in L_2 regularization (or ridge regression). An L1 penalty, which attaches as penalty proportional to the sum of the absolute sizes of the model coefficients, generally results in many 0 coefficients, with the remainder of the coefficients being characterised by relatively little shrinkage (cf. an L2 penalty, which is proportional to the squares of model coefficients, and produces many small but non-zero coefficients). The amount of shrinkage is controlled by the regularization parameters λ_1 and λ_2, where a value of 0 indicates no shrinkage and generates the normal maximum likelihood estimate, and a value of infinity leads to infinite shrinkage and coefficients of 0. Since the L1 penalty leads to increasingly sparse solutions as the regularization strength increases, it can be used as a form of feature selection. In the context of L1 penalized logistic regression [45], the algorithm uses maximum likelihood to minimize the penalized negative binomial log-likelihood function. We use L1 regularized logistic regression models that were build using the penalized R package [46]. Response variables were coded as 0 (non-parent, non-recurrent, etc.) or 1 (parent, recurrent, etc.), and all numerical features were normalized to have mean 0 and variance 1. Models were trained on 70% of the data, and tested on the remaining 30%. A series of penalized logistic regression models were built over a range of regularization strengths, and the model which resulted in 10 non-zero coefficients (or the closest value greater than 10 – in practice, 11) was extracted and absolute values of coefficients plotted in descending order as a measure of feature importance.

Generating predictions with trained logistic regression models produces probabilities. These continuous values between 0 and 1 were binarized, such that test instances were classified as the "1" class when $p \geq 0.5$, and as "0" when $p < 0.5$. For final predictions, optimal regularization strengths λ_1 were found using 10-fold cross-validation on training data. Values of λ_1 between 0.5 to 500 were tested, with the optimal value being the one that maximized the log-likelihood, as suggested by the penalized R package [46].

All human proteins were ranked by their "similarity" to known parent proteins on the basis of the RF and RLR model predictions (i.e. by ordering proteins by their RF class and RLR probability of being parent proteins).

Predicting fusion event pairing patterns

To assess if randomly-generated fusion events are distinguishable from observed fusion events, a mostly sparse (99.92%) 1849x1876 product matrix of all possible fusion events was first constructed, composed of 1849 distinct 5' genes and 1876 distinct 3' genes. A paired feature space consisting of the features of the 5' and 3' partners of known gene fusion pairings, where both parents had available feature vectors, was created. To identify potential

differences between known pairings and random pairings, an equal number (n=2189) of hypothetical parent gene pairing was sampled from the matrix of possible fusion events, and a corresponding paired feature space was generated. Randomly generated fusion pairs were constrained to be novel combinations of known 5' and 3' parents. As before, RF and RLR logistic models were trained to distinguish between the two classes. The procedure was repeated by generating randomized fusions where a fusion could contain any parent (i.e. 5' and 3' parents were pooled).

Software

All data processing, visualization and feature engineering was conducted in the R environment for statistical computing [47], with substantial use of the dplyr [48], tidyr [49], and RMySQL [50] R packages for data processing and the ggplot2 [51] package for data visualization. Network analyses were conducted using the igraph R package version 1.0.0 [52]. Random forests were built with the scikit-learn [53] Python library and regularized logistic regression models were built with the penalized R package [46].

Online Methods References

[1] Yates, A., Akanni, W., Amode, M.R., Barrell, D., et al., Ensembl 2016. *Nucleic Acids Res.* 2016, 44, D710–D716.

[2] Klijn, C., Durinck, S., Stawiski, E.W., Haverty, P.M., et al., A comprehensive transcriptional portrait of human cancer cell lines. *Nat. Biotechnol.* 2015, 33, 306–12.

[3] Smedley, D., Haider, S., Durinck, S., Pandini, L., et al., The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.* 2015, 43, W589–W598.

[4] Charif, D., Lobry, J.R., in:, Bastolla U, Porto M, Roman HE, Vendruscolo M (Eds.), *Struct. approaches to Seq. Evol. Mol. networks, Popul.*, Springer Verlag, 2007, pp. 207–232.

[5] An, O., Dall'Olio, G.M., Mourikis, T.P., Ciccarelli, F.D., NCG 5.0: updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings. *Nucleic Acids Res.* 2016, 44, D992–D999.

[6] Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., et al., COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* 2015, 43, D805–D811.

[7] Zhao, M., Kim, P., Mitra, R., Zhao, J., Zhao, Z., TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. *Nucleic Acids Res.* 2016, 44, D1023–D1031.

[8] Chen, J.-S., Hung, W.-S., Chan, H.-H., Tsai, S.-J., Sun, H.S., In silico identification of oncogenic potential of fyn-related kinase in hepatocellular carcinoma. *Bioinformatics* 2013, 29, 420–7.

[9] The Gene Ontology Consortium, Gene Ontology Consortium: going forward. *Nucleic Acids Res.* 2015, 43, D1049–D1056.

[10] Chen, W.-H., Minguez, P., Lercher, M.J., Bork, P., OGEE: an online gene essentiality database. *Nucleic Acids Res.* 2012, 40, D901–D906.

[11] Rehm, H.L., Berg, J.S., Brooks, L.D., Bustamante, C.D., et al., ClinGen — The Clinical Genome Resource. *N. Engl. J. Med.* 2015, 372, 2235–2242.

[12] The UniProt Consortium, UniProt: a hub for protein information. *Nucleic Acids Res.* 2015, 43, D204–D212.

[13] Manning, G., Whyte, D.B., Martinez, R., Hunter, T., Sudarsanam, S., The protein kinase complement of the human genome. *Science* 2002, 298, 1912–34.

[14] Miranda-Saavedra, D., Barton, G.J., Classification and functional annotation of eukaryotic protein kinases. *Proteins* 2007, 68, 893–914.

[15] Han, H., Shim, H., Shin, D., Shim, J.E., et al., TRRUST: a reference database of human transcriptional regulatory interactions. *Sci. Rep.* 2015, 5, 11432.

[16] Medvedeva, Y.A., Lennartsson, A., Ehsani, R., Kulakovskiy, I. V., et al., EpiFactors: a comprehensive database of human epigenetic factors and complexes. *Database* 2015, 2015, bav067.
[17] Binder, J.X., Pletscher-Frankild, S., Tsafou, K., Stolte, C., et al., COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database* 2014, 2014, bau012.
[18] Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., et al., The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 2016, 44, D279–D285.
[19] Dosztányi, Z., Csizmok, V., Tompa, P., Simon, I., IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 2005, 21, 3433–4.
[20] Mosca, R., Céol, A., Aloy, P., Interactome3D: adding structural details to protein networks. *Nat. Methods* 2012, 10, 47–53.
[21] Meyer, M.J., Das, J., Wang, X., Yu, H., INstruct: a database of high-quality 3D structurally resolved protein interactome networks. *Bioinformatics* 2013, 29, 1577–9.
[22] Krissinel, E., Henrick, K., Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.* 2007, 372, 774–797.
[23] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., et al., The Protein Data Bank. *Nucleic Acids Res.* 2000, 28, 235–242.
[24] Latysheva, N.S., Oates, M., Maddox, L., Flock, T., et al., Molecular principles of gene fusion mediated rewiring of protein interaction networks in cancer. *Mol. Cell* 2016, 63, 1–14.
[25] Dinkel, H., Van Roey, K., Michael, S., Kumar, M., et al., ELM 2016—data update and new functionality of the eukaryotic linear motif resource. *Nucleic Acids Res.* 2016, 44, D294–D300.
[26] Dosztányi, Z., Mészáros, B., Simon, I., ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics* 2009, 25, 2745–6.
[27] Huang, K.-Y., Su, M.-G., Kao, H.-J., Hsieh, Y.-C., et al., dbPTM 2016: 10-year anniversary of a resource for post-translational modification of proteins. *Nucleic Acids Res.* 2016, 44, D435-46.
[28] Minguez, P., Letunic, I., Parca, L., Garcia-Alonso, L., et al., PTMcode v2: a resource for functional associations of post-translational modifications within and between proteins. *Nucleic Acids Res.* 2014, 43, D494–D502.
[29] Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M., KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 2016, 44, D457-62.
[30] Kanehisa, M., KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 2000, 28, 27–30.
[31] Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., et al., BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* 2006, 34, D535-9.
[32] Kryuchkova-Mostacci, N., Robinson-Rechavi, M., A benchmark of gene expression tissue-specificity metrics. *Brief. Bioinform.* 2016.
[33] Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., et al., Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 2005, 21, 650–659.
[34] White, I.R., Royston, P., Wood, A.M., Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med.* 2011, 30, 377–399.
[35] Azur, M.J., Stuart, E.A., Frangakis, C., Leaf, P.J., Multiple imputation by chained equations: what is it and how does it work? *Int. J. Methods Psychiatr. Res.* 2011, 20, 40–49.
[36] Buuren, S. van, Groothuis-Oudshoorn, K., mice : Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* 2011, 45.
[37] Lê, S., Josse, J., Husson, F., FactoMineR: An R Package for Multivariate Analysis. *J. Stat. Softw.* 2008, 25, 1–18.

[38] Mi, H., Poudel, S., Muruganujan, A., Casagrande, J.T., Thomas, P.D., PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res.* 2016, 44, D336–D342.
[39] Strobl, C., Malley, J., Tutz, G., An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol. Methods* 2009, 14, 323–348.
[40] Boulesteix, A.-L., Janitza, S., Kruppa, J., König, I.R., Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 2012, 2, 493–507.
[41] Siroky, D.S., Navigating Random Forests and related advances in algorithmic modeling. *Stat. Surv.* 2009, 3, 147–163.
[42] Brieman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., *Classification and Regression Trees*, Chapman and Hall/CRC, 1984.
[43] Louppe, G., Wehenkel, L., Sutera, A., Geurts, P., in:, Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (Eds.), *Adv. Neural Inf. Process. Syst.*, vol. 26, Curran Associates, Inc., 2013, pp. 431--439.
[44] Lee, S.-I., Lee, H., Abbeel, P., Ng, A.Y., in:, *Twenty-First Natl. Conf. Artif. Intell. Eighteenth Innov. Appl. Artif. Intell. Conf.*, 2006, pp. 16–20.
[45] Vidaurre, D., Bielza, C., Larrañaga, P., A Survey of L 1 Regression. *Int. Stat. Rev.* 2013, 81, 361–387.
[46] Goeman, J.J., Meijer, R.J., Chaturvedi, N., Penalized: L1 (lasso and fused lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model 2016.
[47] R Core Team, R: A language and environment for statistical computing 2014.
[48] Wickham, H., Francois, R., dplyr: A Grammar of Data Manipulation 2015.
[49] Wickham, H., tidyr: Easily Tidy Data with `spread()` and `gather()` Functions 2016.
[50] Ooms, J., James, D., DebRoy, S., Wickham, H., Horner, J., RMySQL: Database Interface and "MySQL" Driver for R 2015.
[51] Wickham, H., *ggplot2: elegant graphics for data analysis*, Springer-Verlag, New York 2009.
[52] Csardi, G., Tamas, N., The igraph software package for complex network research. *InterJournal* 2006, Complex Sy, 1695.
[53] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., et al., Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 2011, 12, 2825–2830.

Legend for Supporting Figures

Figure S1. Biological function enrichments by parent protein cluster. Parent gene sets by cluster were tested for GOSlim molecular function enrichments using PantherDB, with the set of all human genes as background. P values associated with enrichments, shown to the right of each subplot, were corrected for multiple testing using Bonferroni's correction. For each gene set, the log of fractional difference (observed versus expected number of genes) was calculated. Biological process categories that did not have 10 or more member genes in at least one parent cluster were excluded from the visualization.

Figure S2. Cellular pathway fusions in all gene fusions. Comparison of pathway fusion enrichments when using **(a)** all human genes or **(b)** the set of parent genes only as the gene set from which random fusions were generated and enrichments calculated. The 50 most enriched pathway pairs are displayed.

Figure S3. Features of fusion proteins associated with metastatic tumours. Fusion events were labelled as "metastatic" if they occurred in at least 1 cell line of metastatic tumour origin, and "primary" if they only occurred in cell lines derived from primary tumours (see **Methods**). RF and RLR models were trained to distinguish between the two categories as before. **(a)** Feature importance rankings for distinguishing metastatic from non-metastatic parent proteins. **(b)** Distributions of key features identified in (a).

Figure S4. Distinguishing 5' and 3' parent proteins. Gene fusions involve two parent genes, i.e. a 5' and 3' parent. 5' and 3' parent genes have previously been suggested to be functionally distinct. RF and RLR models were trained to classify genes into 5' or 3' parent genes on a balanced dataset. **(a)** Counts of 5' parent genes, 3' parent genes, and both 5' and 3' parent genes. **(b)** Features of highest importance when distinguishing between 5' and 3' genes. **(c)** Distributions of key features from (b).

Figure S5. Analysis of the randomness of 5'-3' pairing patterns in fusion events. (a) Comparison of observed versus random pairings between 5' and 3' fusion partners. Gene fusions were randomly generated by sampling once from the 5' parent set and once from the 3' parent set. RF and RLR classifiers were trained to attempt to distinguish between the features of randomly generated fusions and observed fusions. Note that gene symbols are used for illustration, while Ensembl gene accessions were used for randomization. Random forest predictive performance when **(b)** 5' and 3' gene ordering within fusions was conserved (i.e. in randomly generated fusion events, the 5' parent must be sampled from the set of known 5' parents and the 3' parent from the set of known 3' parents) and when **(c)** 5' and 3' genes are pooled (i.e. a randomly generated fusion can be composed of any combination of two known parent genes). **(d)** Model performance of the trained random forest model on test data, showing a confusion matrix and classification report.

Figure S6. Features of recurrent fusion parent proteins. Recurrent parent genes/proteins are those that participate in more than one fusion event. RF and RLR classifiers were trained to classify recurrent and non-recurrent parent proteins using their features. **(a)** Parent gene counts by category. **(b)** Association between recurrent and metastatic parent genes. **(c)** Feature importance rankings for distinguishing recurrent from non-recurrent parent genes (where a "recurrent" parent forms at least 2 fusions. **(d)** Distributions of key predictive features identified in (c).

Figure S7. **Properties of parent proteins with progressively higher levels of recurrence. (a)** In parent proteins, the values of several features correlate with the number of fusions the proteins form (i.e. their level of recurrence). **(b)** Classification accuracies of RF and RLR models trained to distinguish between recurrent and non-recurrent parents, where the definition of what constitutes "recurrent" was altered to mean forming ≥ 2, ≥ 3, ≥ 4, or ≥ 5 gene fusions. RLR models were constrained to using the 10 most highly predictive features as before. **(c)** The number of gene fusions formed by parent genes.

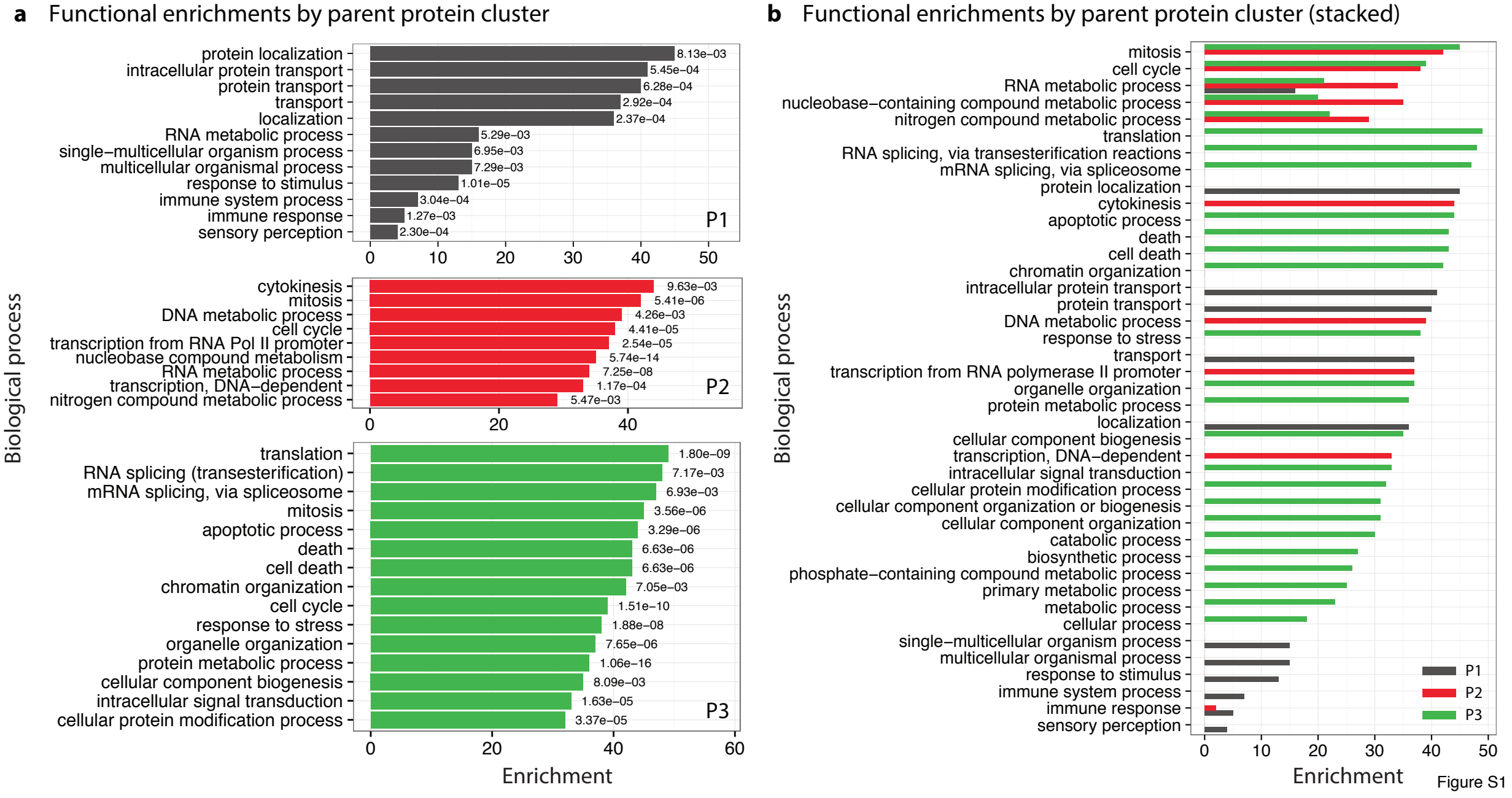
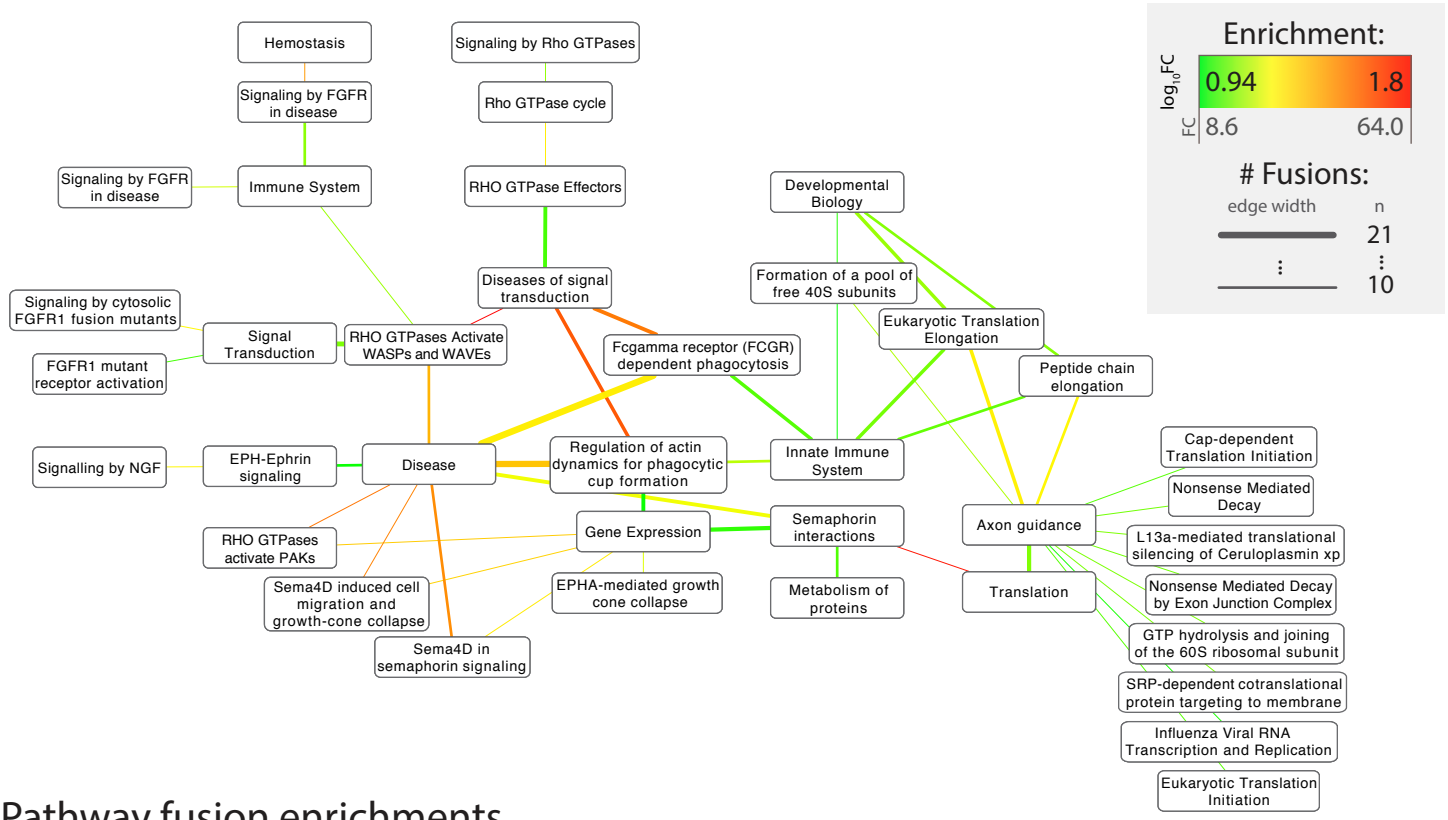

a
Functional enrichments by parent protein cluster
Biological process
protein localization
intracellular protein transport
protein transport
transport
localization
RNA metabolic process
single−multicellular organism process
multicellular organismal process
response to stimulus
immune system process
immune response
sensory perception
8.13e−03
5.45e−04
6.28e−04
2.92e−04
2.37e−04
5.29e−03
6.95e−03
7.29e−03
1.01e−05
3.04e−04
1.27e−03
2.30e−04
P1
0
10
20
30
40
50
cytokinesis
mitosis
DNA metabolic process
cell cycle
transcription from RNA Pol II promoter
nucleobase compound metabolism
RNA metabolic process
transcription, DNA−dependent
nitrogen compound metabolic process
9.63e−03
5.41e−06
4.26e−03
4.41e−05
2.54e−05
5.74e−14
7.25e−08
1.17e−04
5.47e−03
P2
0
20
40
translation
RNA splicing (transesterification)
mRNA splicing, via spliceosome
mitosis
apoptotic process
death
cell death
chromatin organization
cell cycle
response to stress
organelle organization
protein metabolic process
cellular component biogenesis
intracellular signal transduction
cellular protein modification process
1.80e−09
7.17e−03
6.93e−03
3.56e−06
3.29e−06
6.63e−06
6.63e−06
7.05e−03
1.51e−10
1.88e−08
7.65e−06
1.06e−16
8.09e−03
1.63e−05
3.37e−05
P3
0
20
40
60
Enrichment
b
Functional enrichments by parent protein cluster (stacked)
Biological process
mitosis
cell cycle
RNA metabolic process
nucleobase−containing compound metabolic process
nitrogen compound metabolic process
translation
RNA splicing, via transesterification reactions
mRNA splicing, via spliceosome
protein localization
cytokinesis
apoptotic process
death
cell death
chromatin organization
intracellular protein transport
protein transport
DNA metabolic process
response to stress
transport
transcription from RNA polymerase II promoter
organelle organization
protein metabolic process
localization
cellular component biogenesis
transcription, DNA−dependent
intracellular signal transduction
cellular protein modification process
cellular component organization or biogenesis
cellular component organization
catabolic process
biosynthetic process
phosphate−containing compound metabolic process
primary metabolic process
metabolic process
cellular process
single−multicellular organism process
multicellular organismal process
response to stimulus
immune system process
immune response
sensory perception
P1
P2
P3
0
10
20
30
40
50
Enrichment

Figure S1

a Pathway fusion enrichments
(all proteins as background)



b Pathway fusion enrichments
(parent proteins as background)

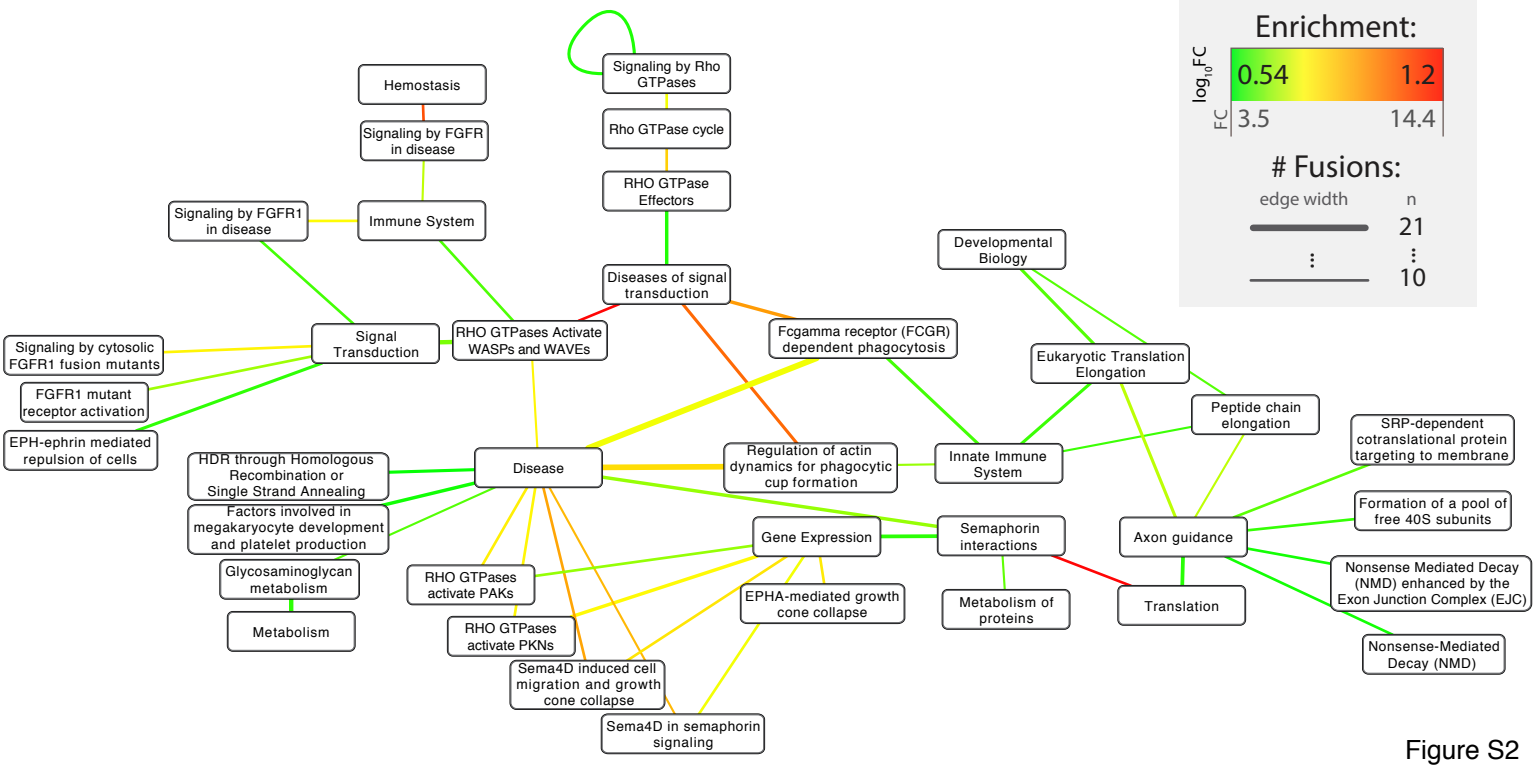
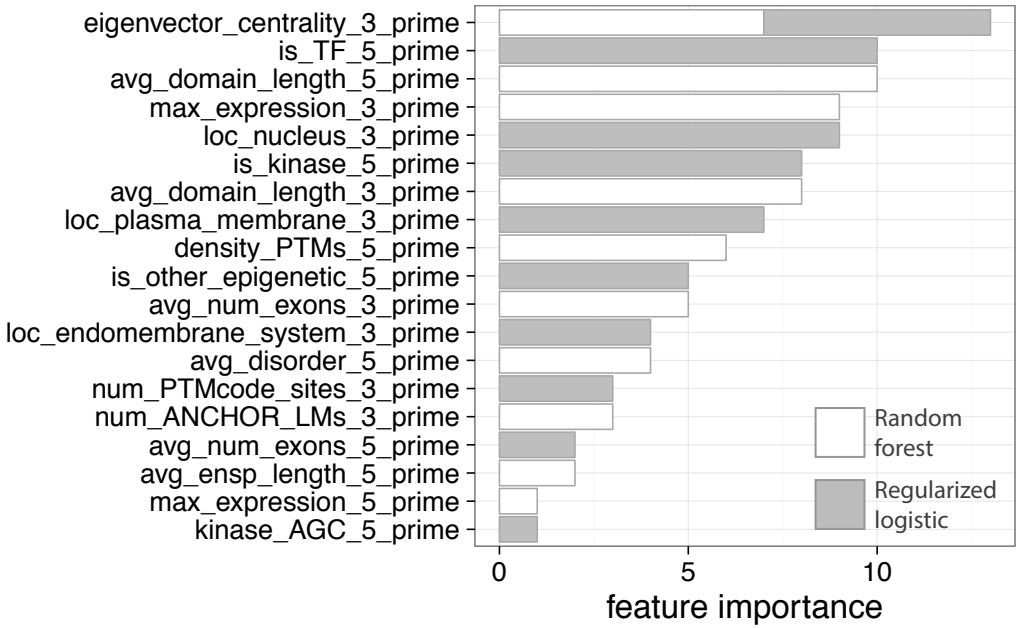


Figure S2

a Features distinguishing metastatic from primary tumour gene fusions



b Distributions of most informative features by class

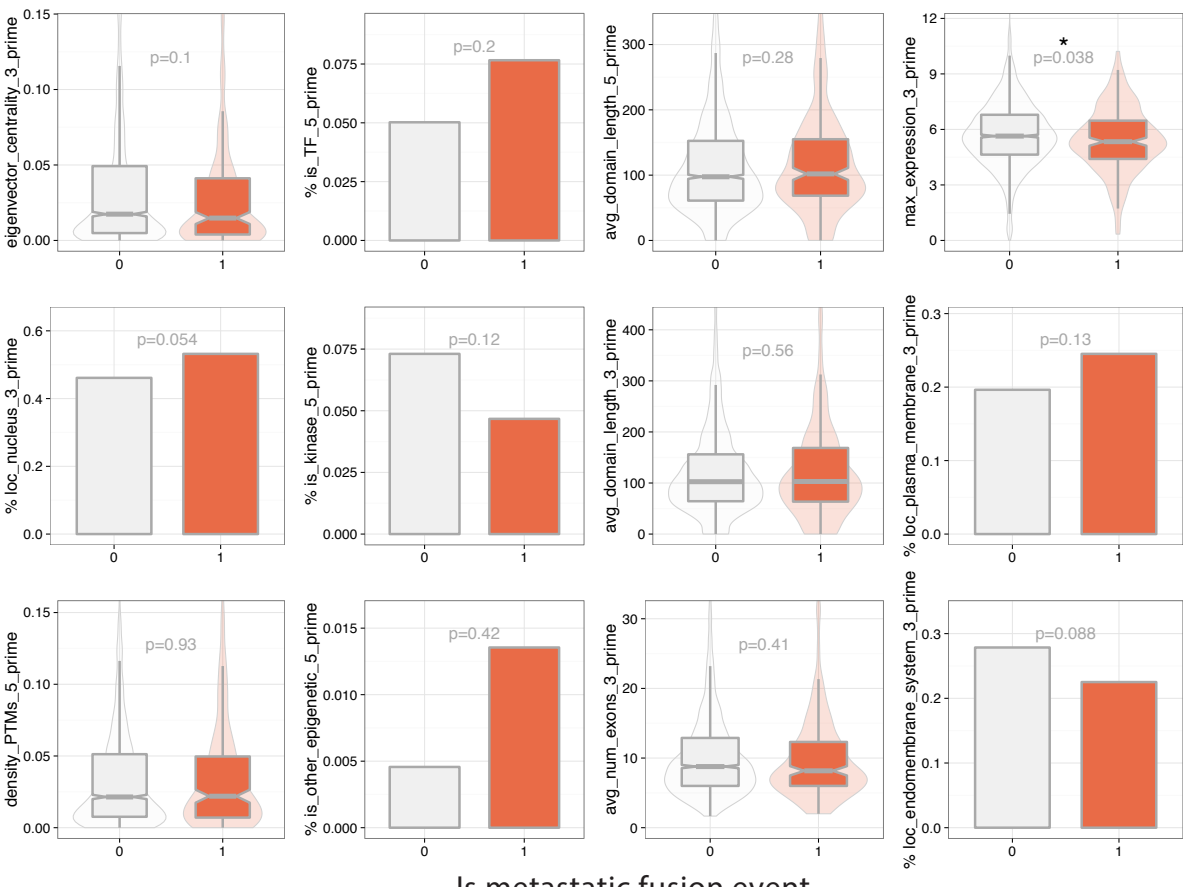


Figure S3

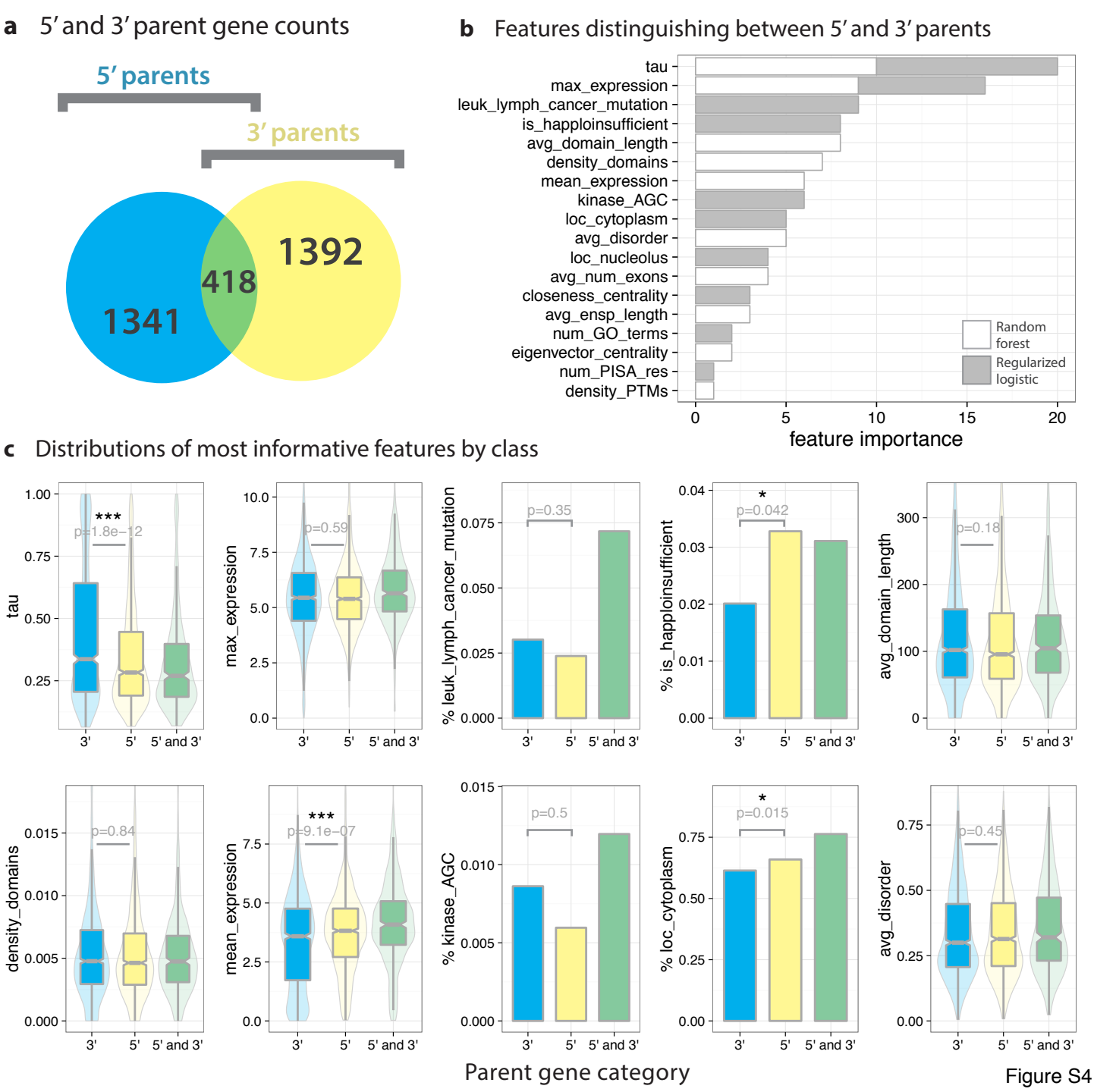

a 5′ and 3′ parent gene counts
5′ parents
3′ parents
1341
418
1392
b Features distinguishing between 5′ and 3′ parents
tau
max_expression
leuk_lymph_cancer_mutation
is_haploinsufficient
avg_domain_length
density_domains
mean_expression
kinase_AGC
loc_cytoplasm
avg_disorder
loc_nucleolus
avg_num_exons
closeness_centrality
avg_ensp_length
num_GO_terms
eigenvector_centrality
num_PISA_res
density_PTMs
Random forest
Regularized logistic
0
5
10
15
20
feature importance
c Distributions of most informative features by class
tau
p=1.8e−12
max_expression
p=0.59
% leuk_lymph_cancer_mutation
p=0.35
% is_haploinsufficient
p=0.042
avg_domain_length
p=0.18
density_domains
p=0.84
mean_expression
p=9.1e−07
% kinase_AGC
p=0.5
% loc_cytoplasm
p=0.015
avg_disorder
p=0.45
3′
5′
5′ and 3′
Parent gene category

Figure S4

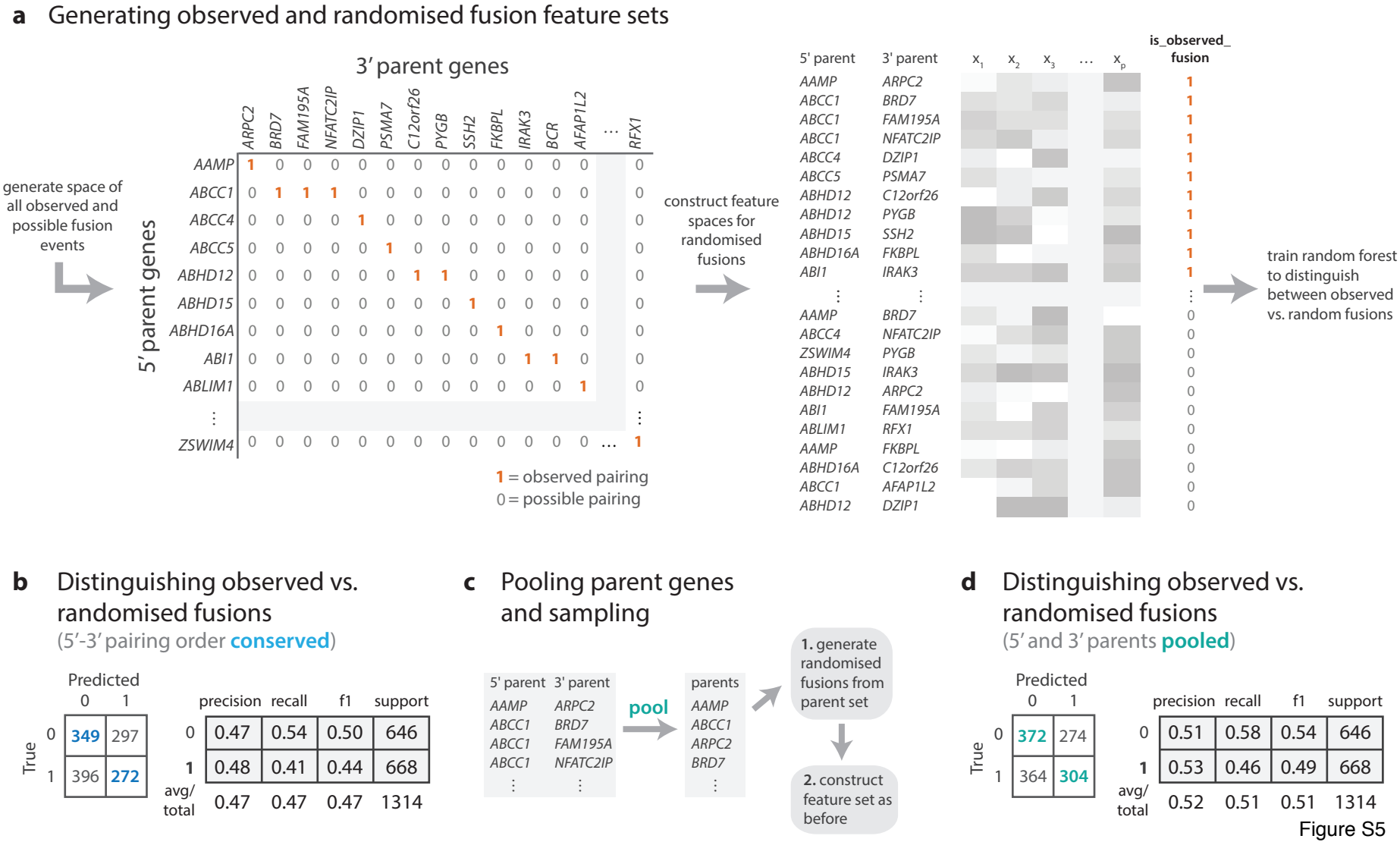
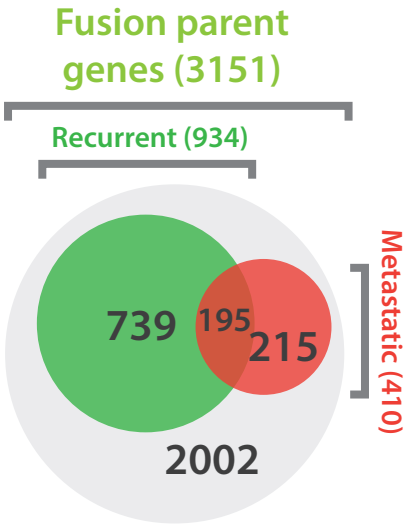

a Generating observed and randomised fusion feature sets
3′ parent genes
5′ parent genes
ARPC2 BRD7 FAM195A NFATC2IP DZIP1 PSMA7 C12orf26 PYGB SSH2 FKBPL IRAK3 BCR AFAP1L2 … RFX1
AAMP 1 0 0 0 0 0 0 0 0 0 0 0 0 0
ABCC1 0 1 1 1 0 0 0 0 0 0 0 0 0 0
ABCC4 0 0 0 0 1 0 0 0 0 0 0 0 0 0
ABCC5 0 0 0 0 0 1 0 0 0 0 0 0 0 0
ABHD12 0 0 0 0 0 0 1 1 0 0 0 0 0 0
ABHD15 0 0 0 0 0 0 0 0 1 0 0 0 0 0
ABHD16A 0 0 0 0 0 0 0 0 0 1 0 0 0 0
ABI1 0 0 0 0 0 0 0 0 0 0 1 1 0 0
ABLIM1 0 0 0 0 0 0 0 0 0 0 0 0 1 0
ZSWIM4 0 0 0 0 0 0 0 0 0 0 0 0 0 … 1
1 = observed pairing
0 = possible pairing
generate space of all observed and possible fusion events
construct feature spaces for randomised fusions
5' parent 3' parent x1 x2 x3 … xp is_observed_fusion
AAMP ARPC2 1
ABCC1 BRD7 1
ABCC1 FAM195A 1
ABCC1 NFATC2IP 1
ABCC4 DZIP1 1
ABCC5 PSMA7 1
ABHD12 C12orf26 1
ABHD12 PYGB 1
ABHD15 SSH2 1
ABHD16A FKBPL 1
ABI1 IRAK3 1
AAMP BRD7 0
ABCC4 NFATC2IP 0
ZSWIM4 PYGB 0
ABHD15 IRAK3 0
ABHD12 ARPC2 0
ABI1 FAM195A 0
ABLIM1 RFX1 0
AAMP FKBPL 0
ABHD16A C12orf26 0
ABCC1 AFAP1L2 0
ABHD12 DZIP1 0
train random forest to distinguish between observed vs. random fusions
b Distinguishing observed vs. randomised fusions
(5′-3′ pairing order conserved)
Predicted 0 1
True 0 349 297
1 396 272
precision recall f1 support
0 0.47 0.54 0.50 646
1 0.48 0.41 0.44 668
avg/ total 0.47 0.47 0.47 1314
c Pooling parent genes and sampling
5' parent 3' parent
AAMP ARPC2
ABCC1 BRD7
ABCC1 FAM195A
ABCC1 NFATC2IP
pool
parents
AAMP
ABCC1
ARPC2
BRD7
1. generate randomised fusions from parent set
2. construct feature set as before
d Distinguishing observed vs. randomised fusions
(5′ and 3′ parents pooled)
Predicted 0 1
True 0 372 274
1 364 304
precision recall f1 support
0 0.51 0.58 0.54 646
1 0.53 0.46 0.49 668
avg/ total 0.52 0.51 0.51 1314

Figure S5

a Fusion parent gene categories

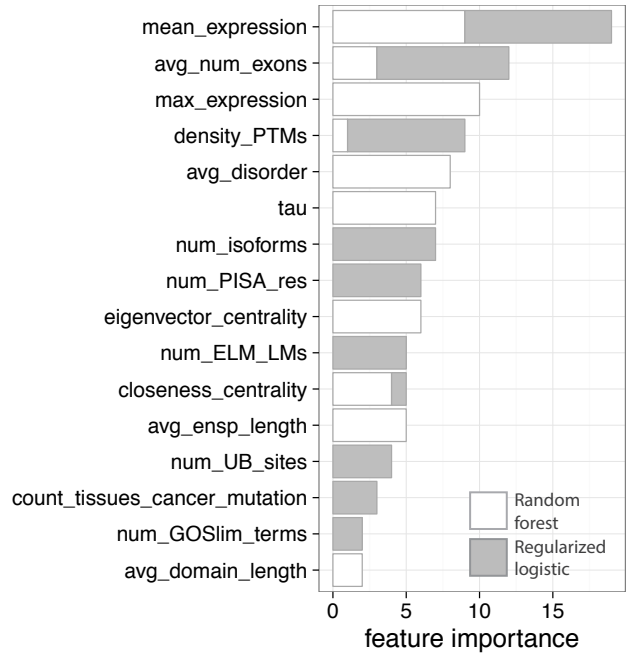


b Parent gene recurrency and cancer progression

	Metastatic	Non-metastatic (primary)
Recurrent	195	739
Non-recurrent	215	2002

χ^2=71.6, df=1, p<2.2e^{-16}
Odds ratio = 2.46, Fisher's p<2.2e^{-16}

c Features predictive of recurrent parents



d Distributions of most informative features by class

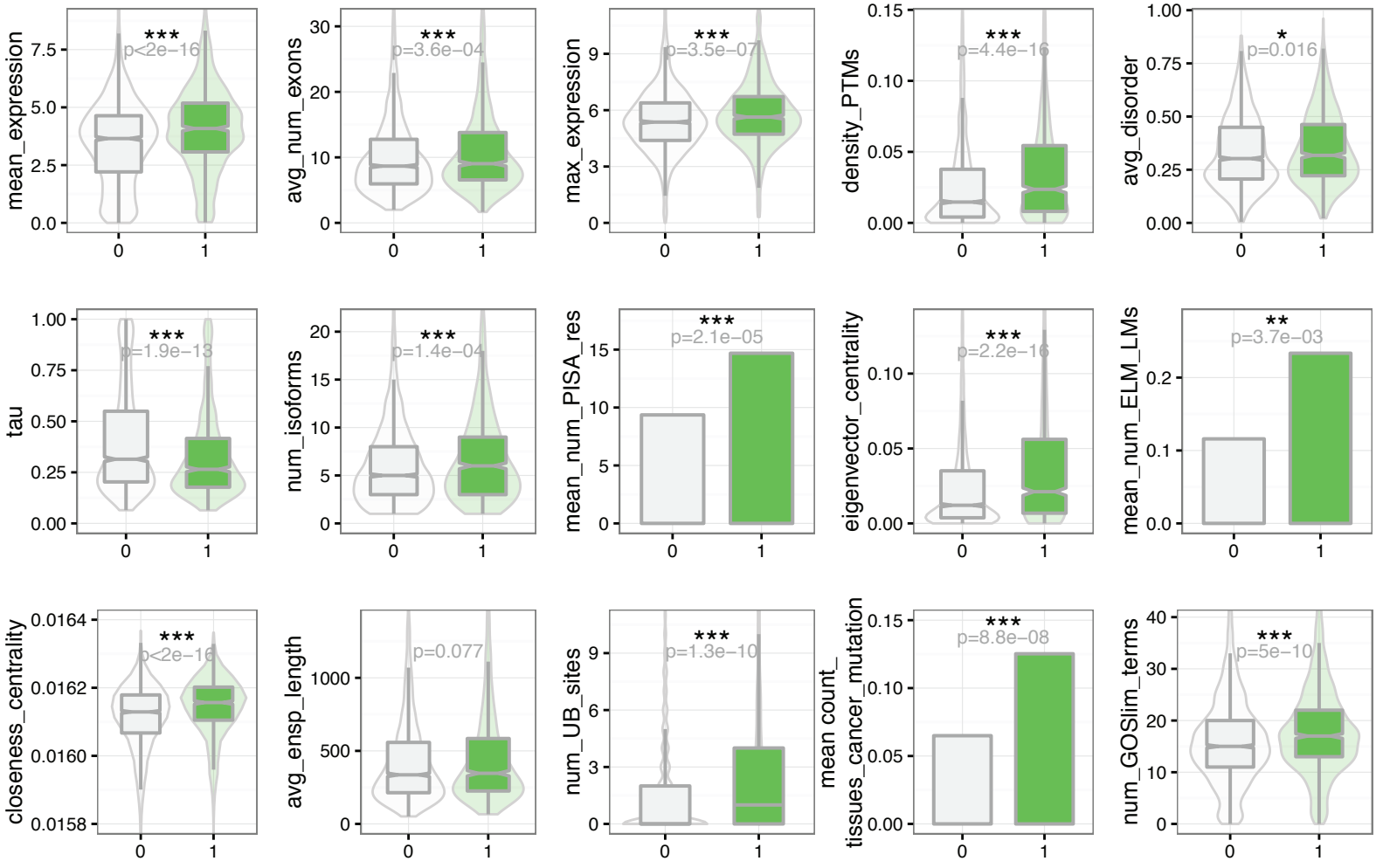


Figure S6

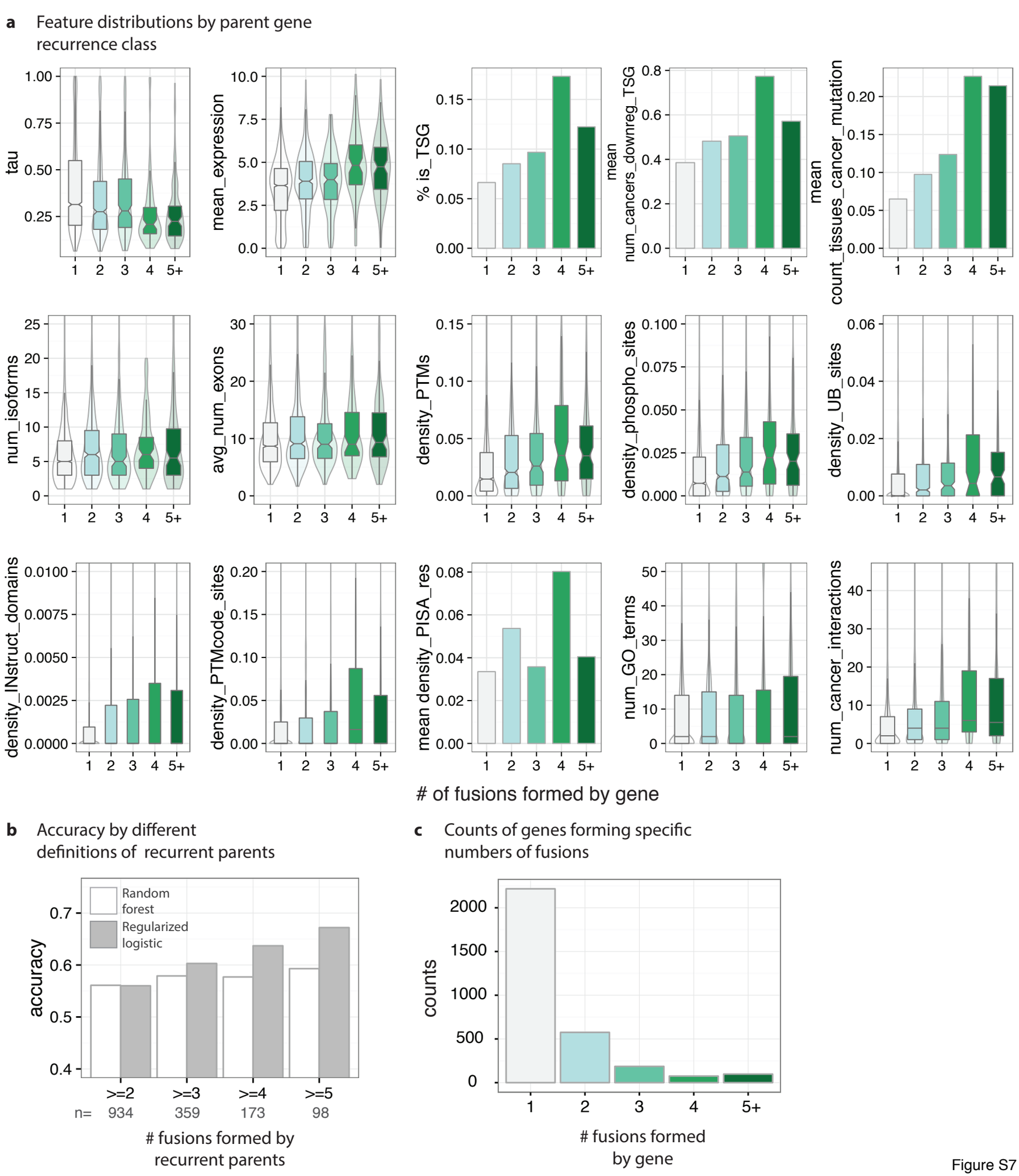

a
Feature distributions by parent gene recurrence class
tau
mean_expression
% is_TSG
mean num_cancers_downreg_TSG
mean count_tissues_cancer_mutation
num_isoforms
avg_num_exons
density_PTMs
density_phospho_sites
density_UB_sites
density_INstruct_domains
density_PTMcode_sites
mean density_PISA_res
num_GO_terms
num_cancer_interactions
1 2 3 4 5+
# of fusions formed by gene
b
Accuracy by different definitions of recurrent parents
Random forest
Regularized logistic
accuracy
>=2 >=3 >=4 >=5
n= 934 359 173 98
# fusions formed by recurrent parents
c
Counts of genes forming specific numbers of fusions
counts
fusions formed by gene

Figure S7