Training set design and description

As far as the design of the training set is concerned, we hereby address two important points. First is the issue of why we augment the negative training set with deliberately frame-shifted fusions. The second regards the issue of whether we can quantify the importance of the "in-frame" feature given that we have altered the training set composition.

Regarding the first issue, we began our study by partitioning all the available data into positive training data from ChimerDB2.0 (POS), negative training data from reactive lymph node tissue (NEG), and a small validation set of transcripts not encountered during classifier training. At this early stage we had not yet augmented the negative training examples with a set of deliberately frame-shifted ChimerDB2.0 transcripts (NEGFS). We were concerned with the performance of Pegasus when trained with only POS and NEG. Scoring of the 39 non-oncogenic transcripts from the validation set can be seen in the fourth columns in the boxplots below, and we have also tabulated them for easy reference. Discriminating POS vs. NEG ignored the scenario where the 3' sequence has maintained important functional domains but they are out of frame. This scenario is biologically plausible and indeed is observed in the **bolded** rows of the table. The 3' gene contains domains with oncogenic potential, but they are out of frame. After re-training the classifier, though, using an augmented negative training set of NEG+NEGFS we can clearly see the improvement in specificity.

Left: trained with NEG, Right: trained with NEG+NEGFS

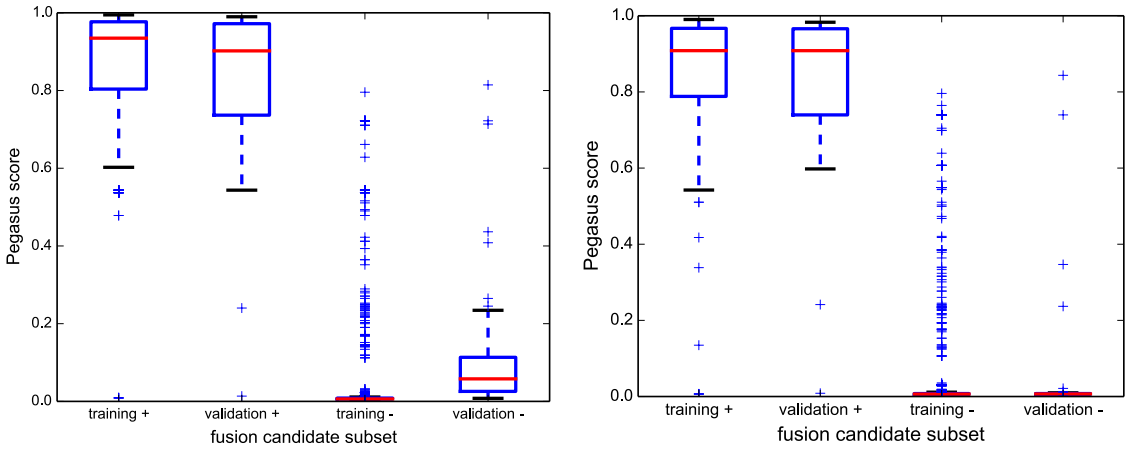
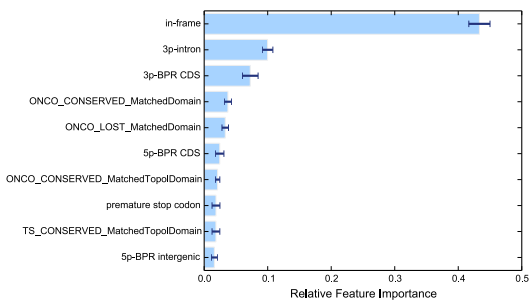


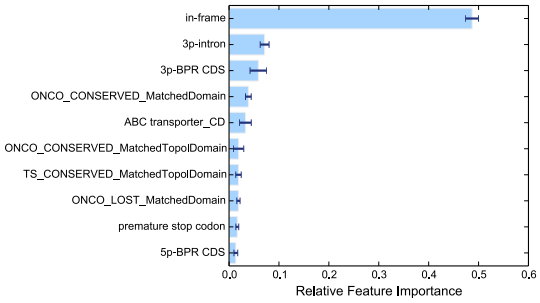
Table of 39 non-oncogenic transcripts from validation set

Gene Name1	Gene Name2	Reading Frame	DriverScore (NEG)	DriverScore (NEG+NEGFS)
PTCRA	CNPY3	InFrame	0.828078171	0.840731419
HSPE1	MOBKL3	InFrame	0.720768052	0.739180545
HMGCS1	**LCK**	**FrameShift**	**0.616223438**	**0.006046606**
KDSR	STAT6	InFrame	0.337369454	0.294550384
KLHL24	**TNK2**	**FrameShift**	**0.321294446**	**0.005866983**

CDK11A	**PTPRS**	**FrameShift**	**0.315252193**	**0.005641046**
UNC45B	**DLG2**	**FrameShift**	**0.3060631**	**0.01055812**
HLA-DRB1	**AKT3**	**FrameShift**	**0.247467869**	**0.005169276**
CKLF	CKLFSF1	InFrame	0.244274047	0.229173079
PPP2R1B	SIK2	FrameShift	0.180824126	0.007822775
EPN2	MAPK7	FrameShift	0.164994389	0.010663401
MKRN2	RAF1	FrameShift	0.1449181	0.00663566
L2HGDH	TAF1	FrameShift	0.144017721	0.00723047
STX16	RPS6KB1	FrameShift	0.135060143	0.006046606
ACSL6	CAMK4	FrameShift	0.135060143	0.006046606
MGRN1	CSNK1A1	FrameShift	0.135060143	0.006046606
ITSN1	PIM2	FrameShift	0.135060143	0.006046606
LIMD2	MAP3K3	FrameShift	0.101303656	0.00663566
EEF1A1	PAN3	FrameShift	0.096906335	0.00663566
SFRS8	ULK1	FrameShift	0.089613688	0.007933935
PTGIS	PRKCQ	FrameShift	0.0850081	0.007933935
PPIL3	CLK1	FrameShift	0.0850081	0.007933935
OAZ1	GAK	FrameShift	0.0850081	0.007933935
OAZ1	CSNK1G2	FrameShift	0.0850081	0.007933935
IL12RB1	MAST3	FrameShift	0.079465309	0.005383732
TMED5	NEK9	FrameShift	0.077757011	0.009479786
TLN1	GRK4	FrameShift	0.070029196	0.001644421
PIK3IP1	LIMK2	FrameShift	0.056493145	0.006395999
KLHL22	SCARF2	FrameShift	0.045875245	0.013841911
AMDHD2	PDPK1	FrameShift	0.043102962	0.015918359
PTBP1	DDX5	FrameShift	0.029069924	0.009069422
ANP32A	PKM2	FrameShift	0.027411948	0.008605665
NCOR2	CDK2AP1	FrameShift	0.021583033	0.008715146
NOTCH2NL	NOTCH2	FrameShift	0.019193874	0.006939551
PIGL	JAK3	FrameShift	0.010442874	0.006307432
HBS1L	STAT3	FrameShift	0.009569203	0.005747346
STAT5B	STAT5A	FrameShift	0.008533391	0.006307432

Regarding the second issue, a precise quantification of the “in-frame” feature importance may be clouded by our deliberate introduction of NEGFS to the training data. Below we show a side-by-side comparison of the feature rankings when trained using POS vs. NEG [left] and POS vs. (NEG+NEGFS) [right]:





The “in-frame” feature does increase slightly in its importance score, but the overall distribution of feature importances is largely unchanged. The ChimerDB2.0 database contains largely in-frame fusions, so it comes as no surprise that this feature would be highly discriminative.

In summary, there is no evidence that the inclusion of NEGFS in the training data significantly changes the relative contributions of the features to the classification function. On the other hand, training with NEGFS *does* provide superior specificity on held out validation data and directly addresses a very real source of false positives.