CHASM

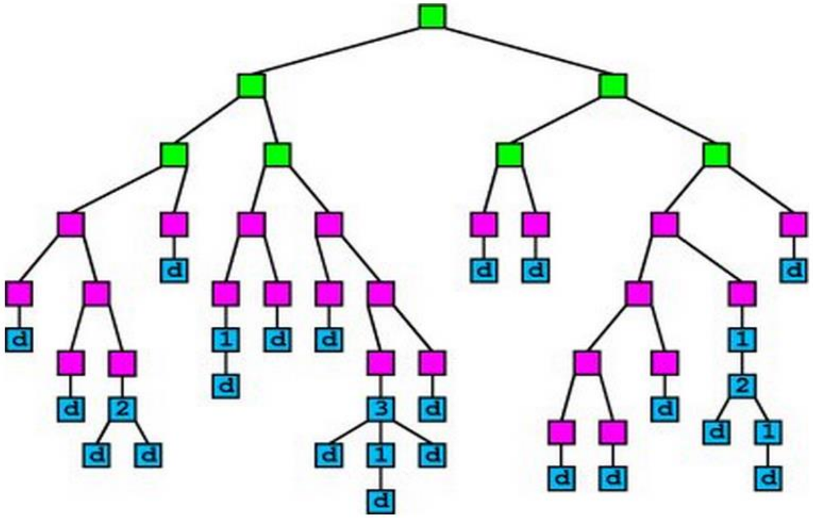
Taylor Jaraczewski

Background

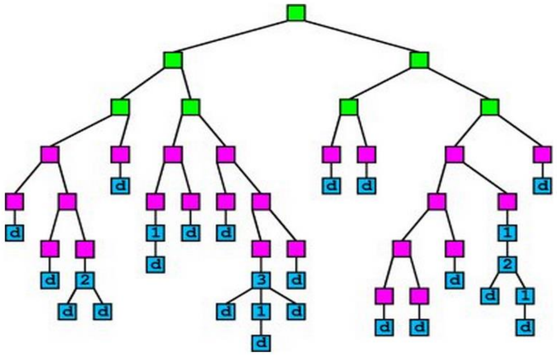
- Yet again….. Drivers vs. passengers
- Only a very small fraction of tumors drives proliferation (hill vs. mountains)
- Need ways to determine drivers NOT based on frequency
- CHASM focuses on missense mutations
  - Make up majority of mutations

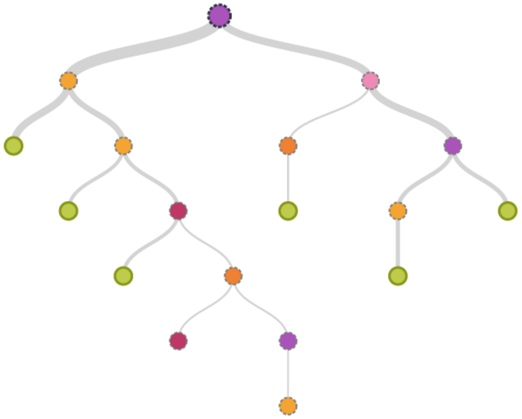
Random Forest Classification

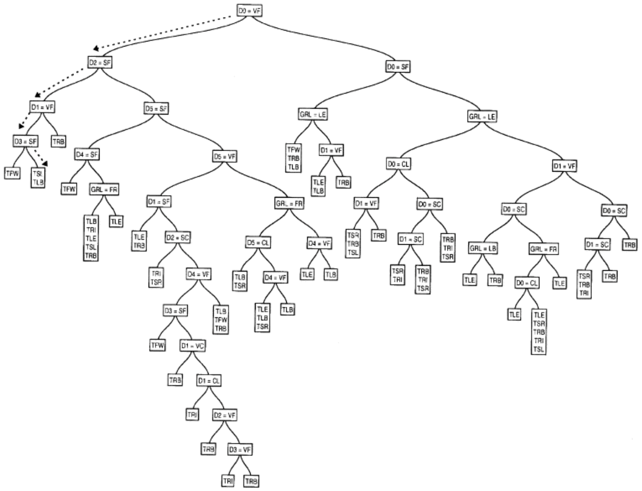
1) Decision Trees

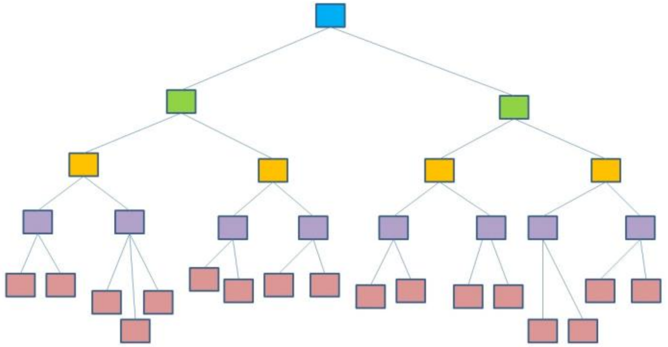


Random Forrest Classifier





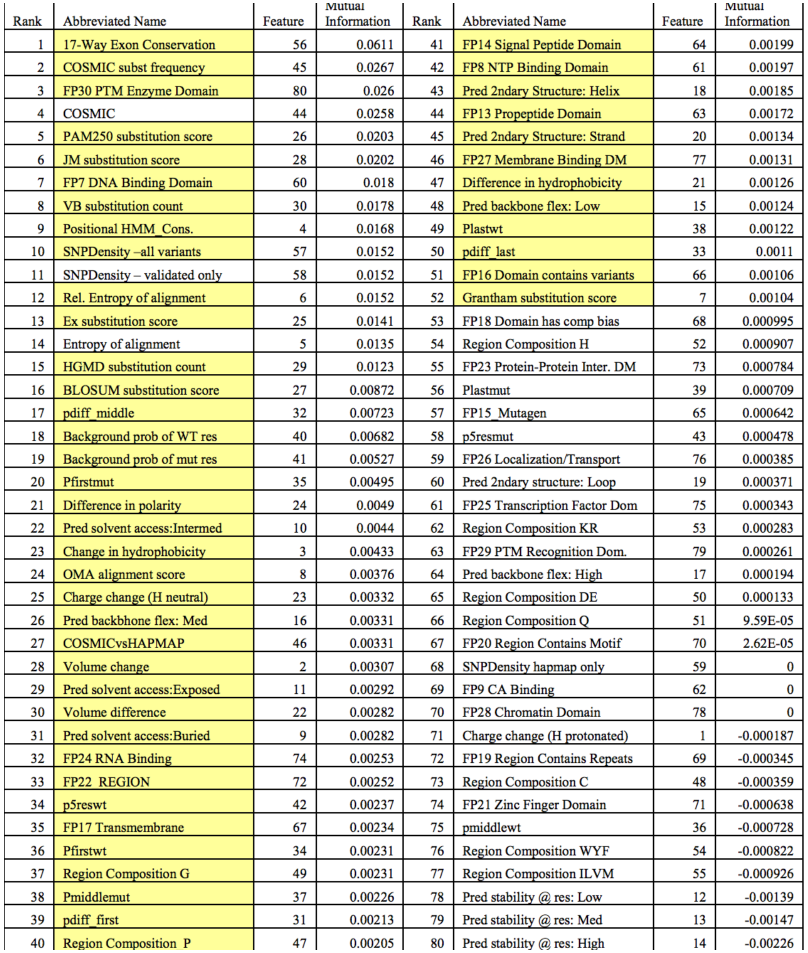




Feature Selection

- - Feature capable of correct classification would require 2.05 bits of info. Top had 0.37

- Chose 49 features determined by mutual information



Rank	Abbreviated Name	Feature	Mutual Information	Rank	Abbreviated Name	Feature	Mutual Information
1	17-Way Exon Conservation	56	0.0611	41	FP14 Signal Peptide Domain	64	0.00199
2	COSMIC subst frequency	45	0.0267	42	FP8 NTP Binding Domain	61	0.00197
3	FP30 PTM Enzyme Domain	80	0.026	43	Pred 2ndary Structure: Helix	18	0.00185
4	COSMIC	44	0.0258	44	FP13 Propeptide Domain	63	0.00172
5	PAM250 substitution score	26	0.0203	45	Pred 2ndary Structure: Strand	20	0.00134
6	JM substitution score	28	0.0202	46	FP27 Membrane Binding DM	77	0.00131
7	FP7 DNA Binding Domain	60	0.018	47	Difference in hydrophobicity	21	0.00126
8	VB substitution count	30	0.0178	48	Pred backbone flex: Low	15	0.00124
9	Positional HMM_Cons.	4	0.0168	49	Plastwt	38	0.00122
10	SNPDensity –all variants	57	0.0152	50	pdiff_last	33	0.0011
11	SNPDensity – validated only	58	0.0152	51	FP16 Domain contains variants	66	0.00106
12	Rel. Entropy of alignment	6	0.0152	52	Grantham substitution score	7	0.00104
13	Ex substitution score	25	0.0141	53	FP18 Domain has comp bias	68	0.000995
14	Entropy of alignment	5	0.0135	54	Region Composition H	52	0.000907
15	HGMD substitution count	29	0.0123	55	FP23 Protein-Protein Inter. DM	73	0.000784
16	BLOSUM substitution score	27	0.00872	56	Plastmut	39	0.000709
17	pdiff_middle	32	0.00723	57	FP15_Mutagen	65	0.000642
18	Background prob of WT res	40	0.00682	58	p5resmut	43	0.000478
19	Background prob of mut res	41	0.00527	59	FP26 Localization/Transport	76	0.000385
20	Pfirstmut	35	0.00495	60	Pred 2ndary structure: Loop	19	0.000371
21	Difference in polarity	24	0.0049	61	FP25 Transcription Factor Dom	75	0.000343
22	Pred solvent access:Intermed	10	0.0044	62	Region Composition KR	53	0.000283
23	Change in hydrophobicity	3	0.00433	63	FP29 PTM Recognition Dom.	79	0.000261
24	OMA alignment score	8	0.00376	64	Pred backbone flex: High	17	0.000194
25	Charge change (H neutral)	23	0.00332	65	Region Composition DE	50	0.000133
26	Pred backbhone flex: Med	16	0.00331	66	Region Composition Q	51	9.59E-05
27	COSMICvsHAPMAP	46	0.00331	67	FP20 Region Contains Motif	70	2.62E-05
28	Volume change	2	0.00307	68	SNPDensity hapmap only	59	0
29	Pred solvent access:Exposed	11	0.00292	69	FP9 CA Binding	62	0
30	Volume difference	22	0.00282	70	FP28 Chromatin Domain	78	0
31	Pred solvent access:Buried	9	0.00282	71	Charge change (H protonated)	1	-0.000187
32	FP24 RNA Binding	74	0.00253	72	FP19 Region Contains Repeats	69	-0.000345
33	FP22_REGION	72	0.00252	73	Region Composition C	48	-0.000359
34	p5reswt	42	0.00237	74	FP21 Zinc Finger Domain	71	-0.000638
35	FP17 Transmembrane	67	0.00234	75	pmiddlewt	36	-0.000728
36	Pfirstwt	34	0.00231	76	Region Composition WYF	54	-0.000822
37	Region Composition G	49	0.00231	77	Region Composition ILVM	55	-0.000926
38	Pmiddlemut	37	0.00226	78	Pred stability @ res: Low	12	-0.00139
39	pdiff_first	31	0.00213	79	Pred stability @ res: Med	13	-0.00147
40	Region Composition_P	47	0.00205	80	Pred stability @ res: High	14	-0.00226

General Random Forest Info

- Used 500 trees

- Used known drivers and synthetic passengers for feature selection and classifier training

- Mtry = 7
  - Number of variables available for splitting at each node

Comparison to Other Methods

Receiver Operator Characteristic (ROC)

- Points that reperesent trade-off between sensitivity (fraction of drivers correctly classified) and specificity (“ “ passengers)

Precision Recall

- Points that represent the trade-off between precision (fraction of true drivers out of all predicted drivers) and recall (sensitivity)

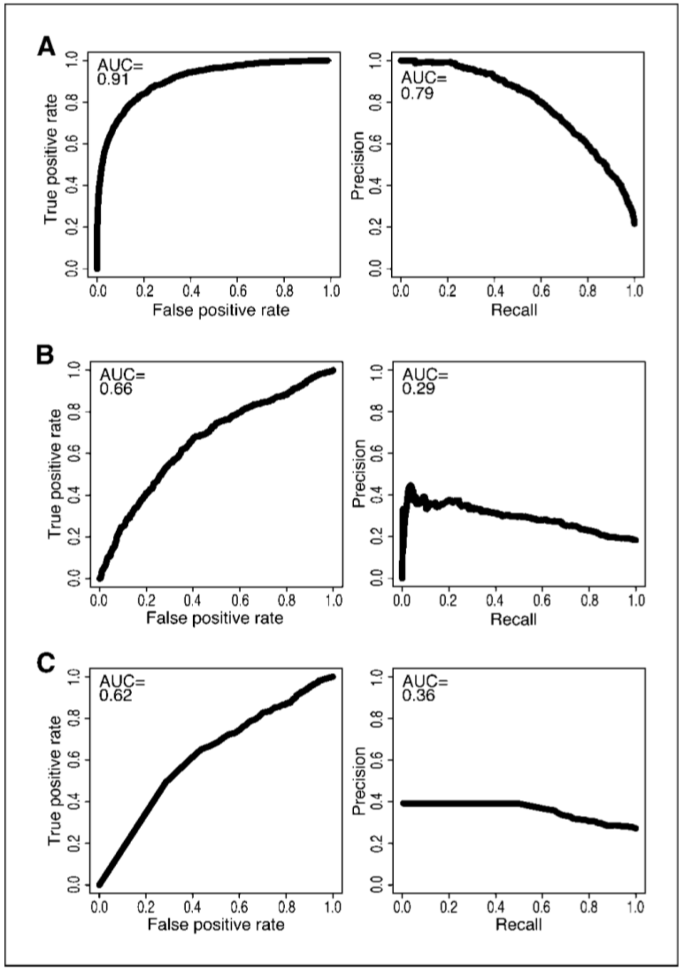


Figure 2. ROC and PR curves calculated for (*A*) CHASM, (*B*) PolyPhen PSIC, and (*C*) SIFT on the training set mutations. CHASM training out-of-bag scores were used to generate the ROC and PR curves in *A*. A color version is available as Supplementary Fig. S6.

Other Models

PolyPhen - Uses Bayes classification; queries BLAST data base to predict impact of amino acid substitution on the structure/function of proteins

SIFT – Provides score for probability that a missense mutation will be tolerated.

CanPredict – Combination of SIFT score, LogRE score, and GOSS score to train a random forest classifier

KinaseSVM – Uses protein kinases

GBM

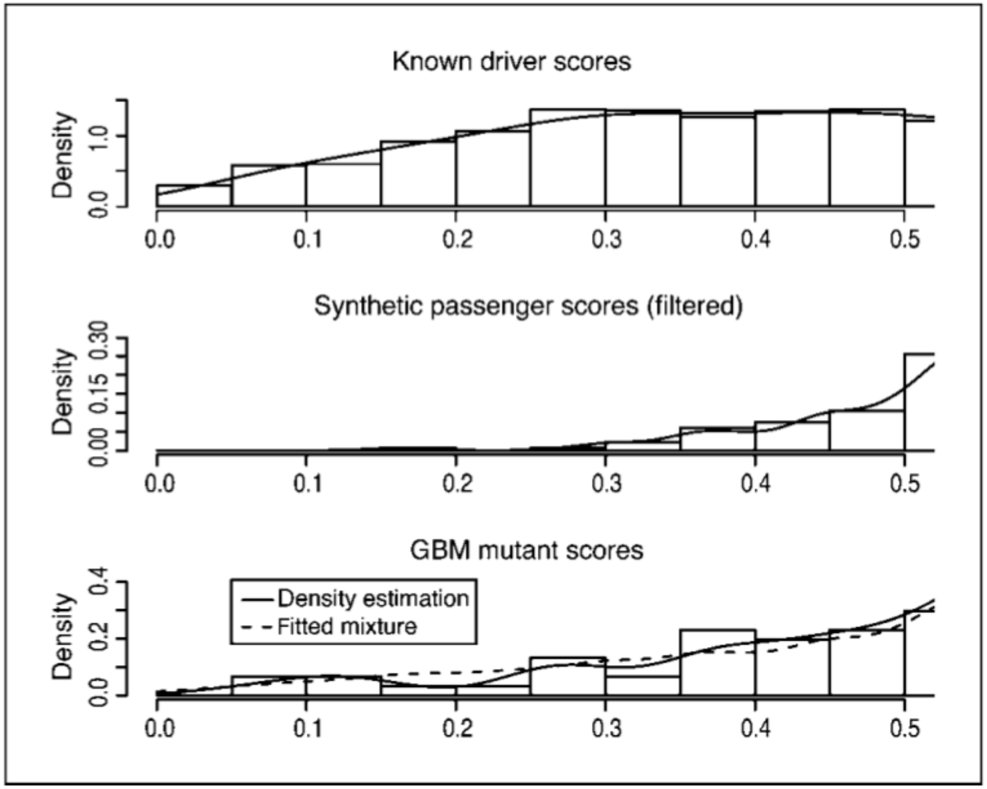


Figure 5. Histograms of CHASM scores for driver mutations and passenger mutations held out from the training set, and 607 mutations experimentally identified in GBM. Estimated kernel density for each set of scores (*solid line*) and fitted mixture of the driver and passenger score densities (*dashed line*) are shown superimposed on the histograms.