Supplementary Tables

Supplementary Table 1: Candidate predictive features that were initially evaluated for use in the CHASM classifier. The most informative features that are currently being used in CHASM are highlighted in yellow in Supplementary Table 3.

	Feature	Description
1	Net residue charge change	The change in formal charge resulting from the mutation. Histidine is assumed protonated (formal charge of +1)
2	Net residue volume change	The change in residue volume resulting from the mutation (27).
3	Net residue hydrophobicity change	The change in hydrophobicity resulting from the substitution (28).
4	Positional Hidden Markov model (HMM) conservation score	This feature is calculated based on the degree of conservation of the residue estimated from a multiple sequence alignment built with SAM-T2K software (29), using the protein in which the mutation occurred as the seed sequence (30). The SAM-T2K alignments are large, superfamily-level alignments that include distantly related homologs (as well as close homologs and orthologs) of the protein of interest.
5	Entropy of HMM alignment	The Shannon entropy calculated for the column of the SAM-T2K multiple sequence alignment, corresponding to the location of the mutation (31).
6	Relative entropy of HMM alignment	Difference in Shannon entropy calculated for the column of the SAM-T2K multiple sequence alignment (corresponding to the location of the mutation) and that of a background distribution of amino acid residues computed from a large sample of multiple sequence alignments (31).
7	Compatibility score for amino acid substitution in the column of a multiple sequence alignment of orthologs	These multiple sequence alignments are calculated using groups of orthologous proteins from the OMA database (32), which are aligned with T-Coffee software(33). The compatibility score for the mutation in the column of interest is computed from a large sample of multiple sequence alignments (31).
8	Grantham Score	The Grantham substitution score for the wild type to mutant transition (34).
9-11	Predicted residue solvent accessibility	These features consist of the probability of the wild type residue being buried, intermediate or exposed as predicted by a neural network trained with Predict-2nd software (29) on a set of 1763 proteins with high-resolution X-ray crystal structures sharing less than 30% homology (35).
12-14	Predicted contribution to protein stability	These features consist of the probability that the wild type residue contributes to overall protein stability in a manner that is highly stabilizing, average or destabilizing, as predicted by a neural network trained with Predict-2nd software (29) on a set of 1763 proteins with less than 30% homology. Stability estimates for the neural network training data were calculated using the FoldX force field (36, 37).
15-17	Predicted flexibility (Bfactor)	These features consist of the probability that the wild

		type residue backbone is stiff, intermediate or flexible as predicted by a neural network trained with Predict-2nd software (29) on a set of 1763 proteins with less than 30% homology. Flexibilities for the neural net training data were estimated based on normalized temperature factors, computed using the method of (38) from the X-ray crystal structure files.
18-20	Predicted secondary structure	These features consist of the probability that the secondary structure of the region in which the wild type residue exists is helix, loop or strand as predicted by a neural net trained with Predict-2nd software (29) on a set of 1763 proteins with crystal structures and with less than 30% homology.
21	Change in hydrophobicity	Change in residue hydrophobicity due to the wild type to mutant transition.
22	Change in volume	Change in residue volume due to the wild type to mutant transition.
23	Change in charge	Change in residue charge due to the wild type to mutant transition. Histidine is assumed neutral.
24	Change in polarity	Change in residue polarity due to the wildtype to mutant transition calculated in (34)
25	EX substitution score	Amino acid substitution score from the EX matrix (37).
26	PAM250 substitution score	Amino acid substitution score from the PAM250 matrix (39).
27	BLOSUM 62 substitution score	Amino acid substitution score from the BLOSUM 62 matrix (40).
28	MJ substitution score	Amino acid substitution score from the Miyazawa-Jernigan contact energy matrix (37, 41).
29	HGMD2003 mutation count	Number of times that the wild type to mutant substitution occurs in the Human Gene Mutation Database, 2003 version (25, 30, 31).
30	VB mutation count	Amino acid substitution score from the VB (Venkatarajan and Braun) matrix (37, 42).
31-33	Probability of seeing the wild type residue in the first, middle, or last position of an amino acid triple	Calculated by joint frequencies of amino acid triples in human proteins found in UniProtKB* (10).
34-36	Probability of seeing the mutant residue in the first, middle, or last position of an amino acid triple	Calculated by joint frequencies of amino acid triples in human proteins found in UniProtKB* (10).
37-39	Difference in probability of seeing the wildtype vs. the mutant reside in the first, middle, or last position of an amino acid triple	Calculated by joint frequencies of amino acid triples in human proteins found in UniProtKB* (10).
40	Background probability of wildtype residue in UniProtKB* human proteins	Estimated as frequency of amino acid residue type occurrence.
41	Background probability of mutant residue in UniProtKB* human proteins	Estimated as frequency of amino acid residue type occurrence.
42	Probability of seeing the wild type at the center of a window of 5 amino acid residues	Calculated by a Markov chain of amino acid quintuples in human proteins found in UniProtKB* (10).
43	Probability of seeing the mutant at the center of a window of 5 amino acid residues	Calculated by a Markov chain of amino acid quintuples in human proteins found in UniProtKB* (10).

44-46	Frequency of missense change type in the Catalog of Somatic Mutations in Cancer (COSMIC) database	Frequency that missense change type (amino acid type X to amino acid type Y, e.g. ALANINE to GLYCINE) is seen in COSMIC. These frequencies were calculated during the week of August 14, 2008, using COSMIC release 38 (43) and normalized by the occurrences of the wild type residue in human proteins found in UniProtKB* (10), the occurrences of the wild type reside in cosmic or the number of times the change type is observed in the HapMap SNPs database (44).
47-55	Regional AA composition	The percentage of amino acids in a 15 residue window surrounding the mutation that fall into one of the following categories (P,C,G,DE,Q,H,KR,WYF,ILVM).
56	17way exon conservation	The conservation score for the entire exon calculated from a 17-species phylogenetic alignment using the UCSC Genome Browser (45). Scores are given for windows of nucleotides. We retrieve the scores for each region that overlaps the exon in which the base substitution occurred and calculated a weighted average of the conservation scores where the weight is the number of bases with a particular score.
57-59	SNP Density	The number of genetic variants, polymorphisms or verified HapMap SNPs (44) in the exon where the mutation is located
60-80	UniProt Annotations (fingerprints)	These features give annotations, curated from the literature, of general binding sites, general active sites, lipid, metal, carbohydrate, DNA, phosphate and calcium binding sites, disulfides, modified residues, propeptide residues, signal peptide residues, known mutagenic sites, transmembrane regions, compositionally biased regions, repeat regions, known motifs, and zinc fingers. The integer 1 indicates that a feature is present and the integer 0 indicates that it is absent at a mutated position

Supplementary Table. 2:

Synthetic Mutations were generated from eight multinomial distributions that depend on both tumor type and DNA context. The columns of the table show the eight contexts for each wild type DNA base and the rows show the (multinomial) probability distributions of base substitutions in GBM, based on (46).

Glioblastoma Multiforme (GBM)

	C in CpG	G in CpG	C in TpC	G in GpA	A	C	G	T
A	0.05	0.97	0.31	0.44	0.00	0.29	0.50	0.39
C	0.00	0.02	0.00	0.22	0.13	0.00	0.13	0.39
G	0.02	0.00	0.21	0.00	0.62	0.20	0.00	0.22
T	0.93	0.01	0.48	0.33	0.25	0.51	0.37	0.00

Supplementary Table 3. 80 candidate predictive features ranked according to their mutual information (in units of bits) with respect to driver and passenger classes. Detailed feature descriptions are in Supplementary Table 1. FP = fingerprint (a binary feature that takes on values of either 0 or 1).

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