

CHASM:

Cancer-specific high-throughput annotation of somatic mutations

Rachel Karchin, Ph.D.

Department of Biomedical Engineering

Institute for Computational Medicine

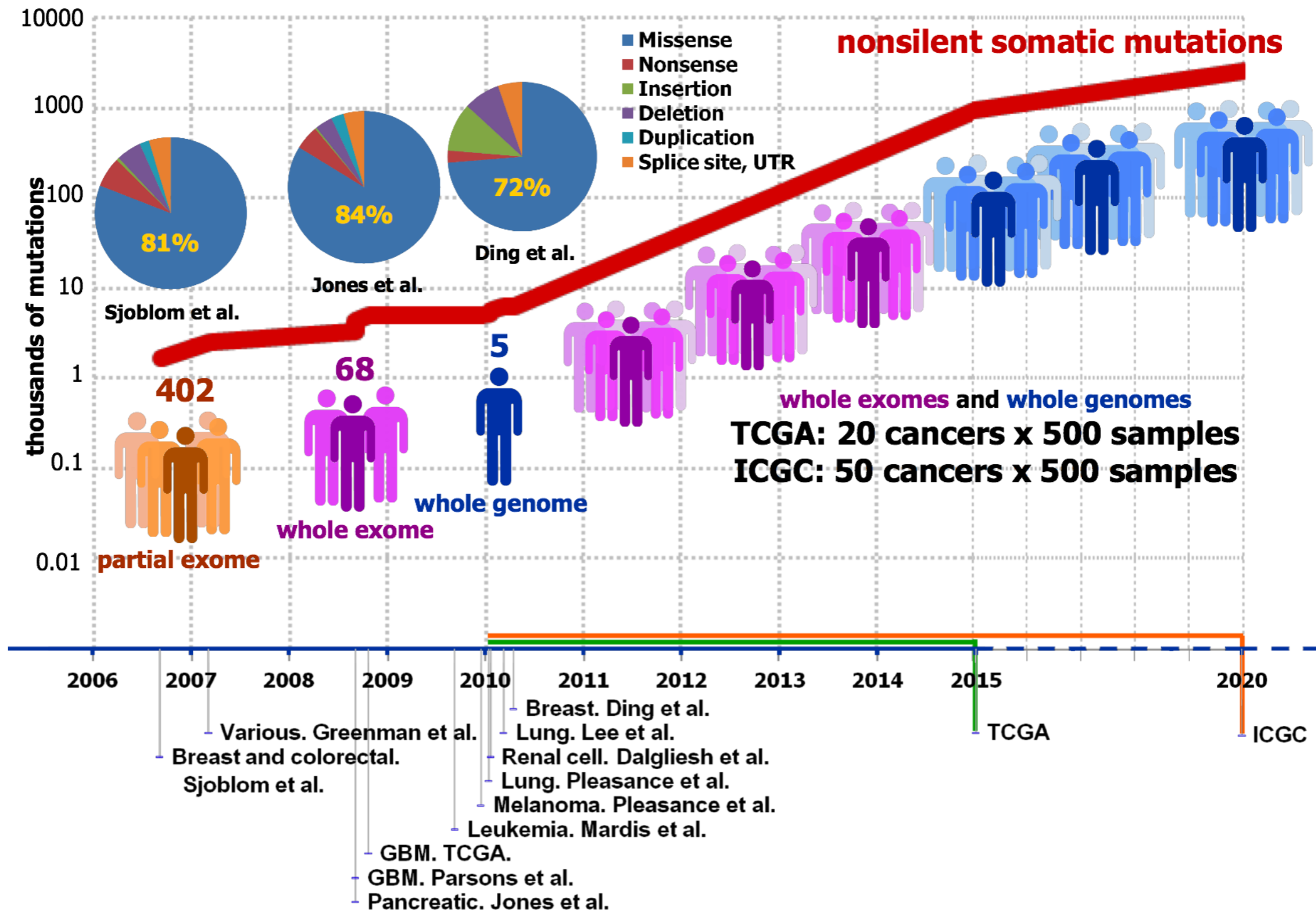
Johns Hopkins University



CHASM Overview

- Machine learning method to identify “driver” missense mutations
- High-throughput
 - Automated
 - No time-consuming calculations
- High coverage
 - Classifications based on sequence only
 - Protein structure, interactions, pathways not used
- Prioritization for functional studies

Data deluge from whole-exome sequencing



Classic approach to finding driver genes

Significantly mutated genes



Patterns of somatic mutation in human cancer genomes

Greenman *et al.* 2007

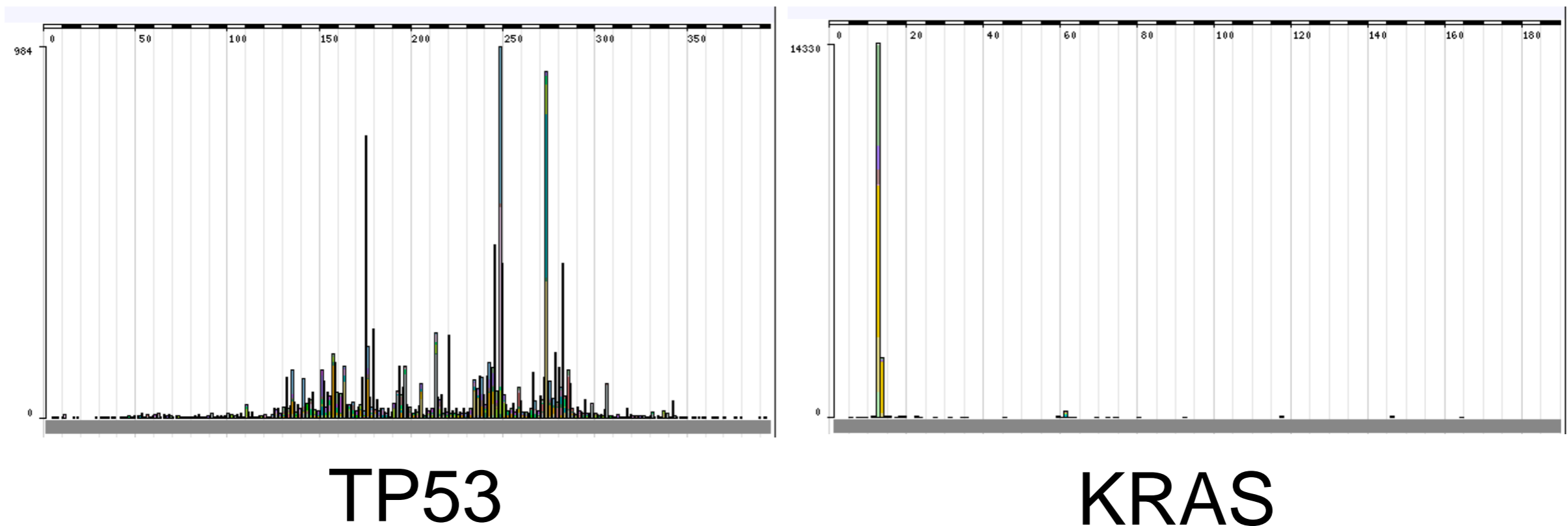
The Genomic Landscapes of Human Breast and Colorectal Cancers

Wood *et al.* 2007

Baudot *et al.* 2009

Classic approach to finding driver genes

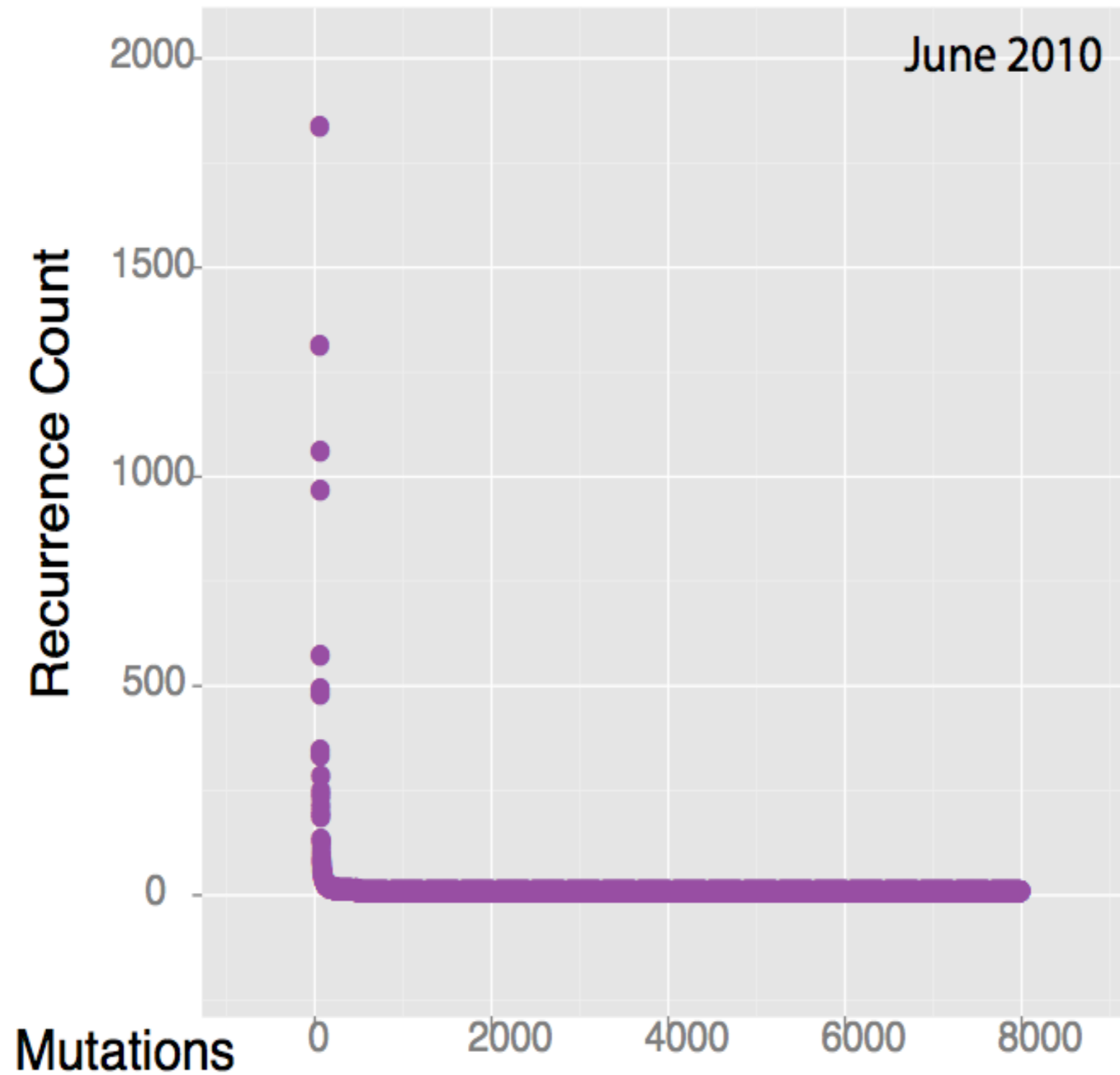
Recurrent mutations



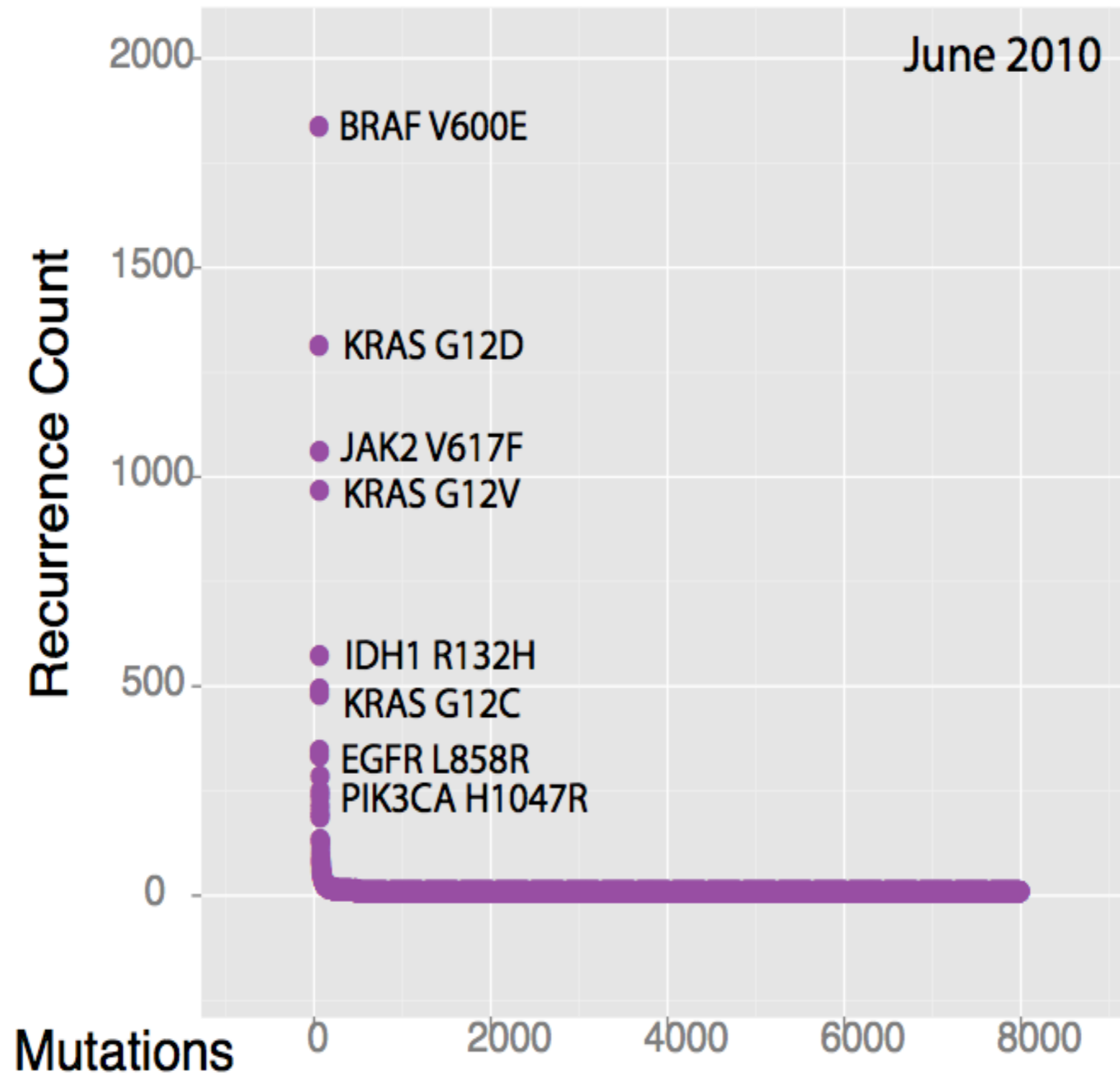
Limitations to the classic approach

- Driver genes may not be significantly mutated
- Driver mutations may not be recurrent

Confirmed somatic mutations in COSMIC



Confirmed somatic mutations in COSMIC



- Methods that predict driving mutations and genes **independent of frequency** needed

Bioinformatics methods?

Nucleic Acids Res. 2007 Jul;35(Web Server issue):W595-8. Epub 2007 May 30.

CanPredict: a computational tool for predicting cancer-associated missense mutations.

Kaminker JS, Zhang Y, Watanabe C, Zhang Z.

Cancer Res. 2008 Mar 15;68(6):1675-82.

Prediction of cancer driver mutations in protein kinases.

Torkamani A, Schork NJ.

Nucleic Acids Res. 2003 Jul 1;31(13):3812-4.

SIFT: Predicting amino acid changes that affect protein function.

Ng PC, Henikoff S.

Nat Methods. 2010 Apr;7(4):248-9.

A method and server for predicting damaging missense mutations.

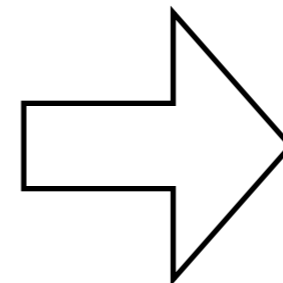
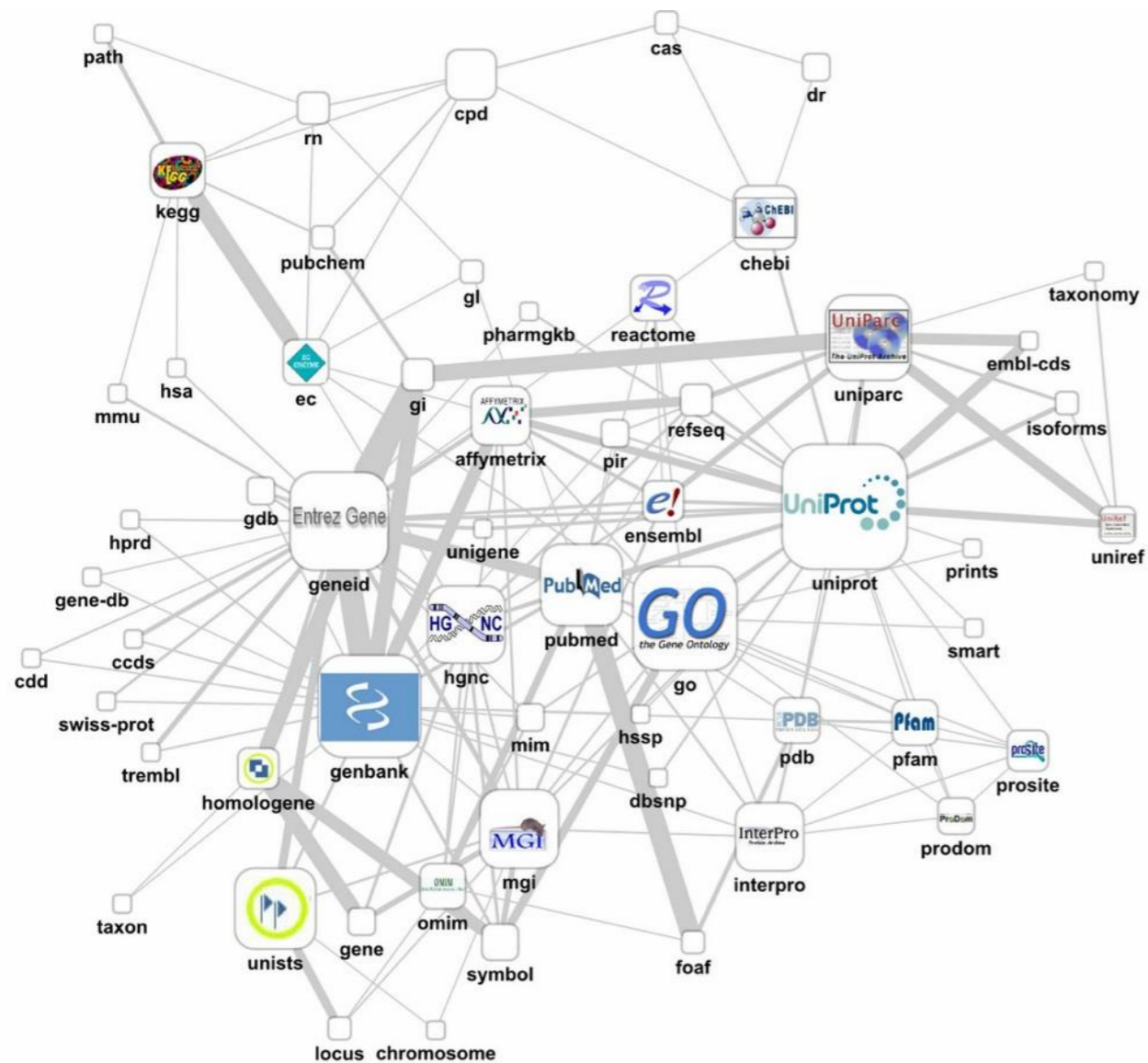
Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR.



Bioinformatics analysis of mutation is indirect

- Consider ways that a single amino acid substitution can impact protein function
 - Protein aggregates and does not fold
 - Protein is destabilized and unfolds partially
 - Binding interfaces are disrupted
 - Active sites are disrupted
 - PTM sites are disrupted
- Infer these events using data mining and/or **simplifying proxies**



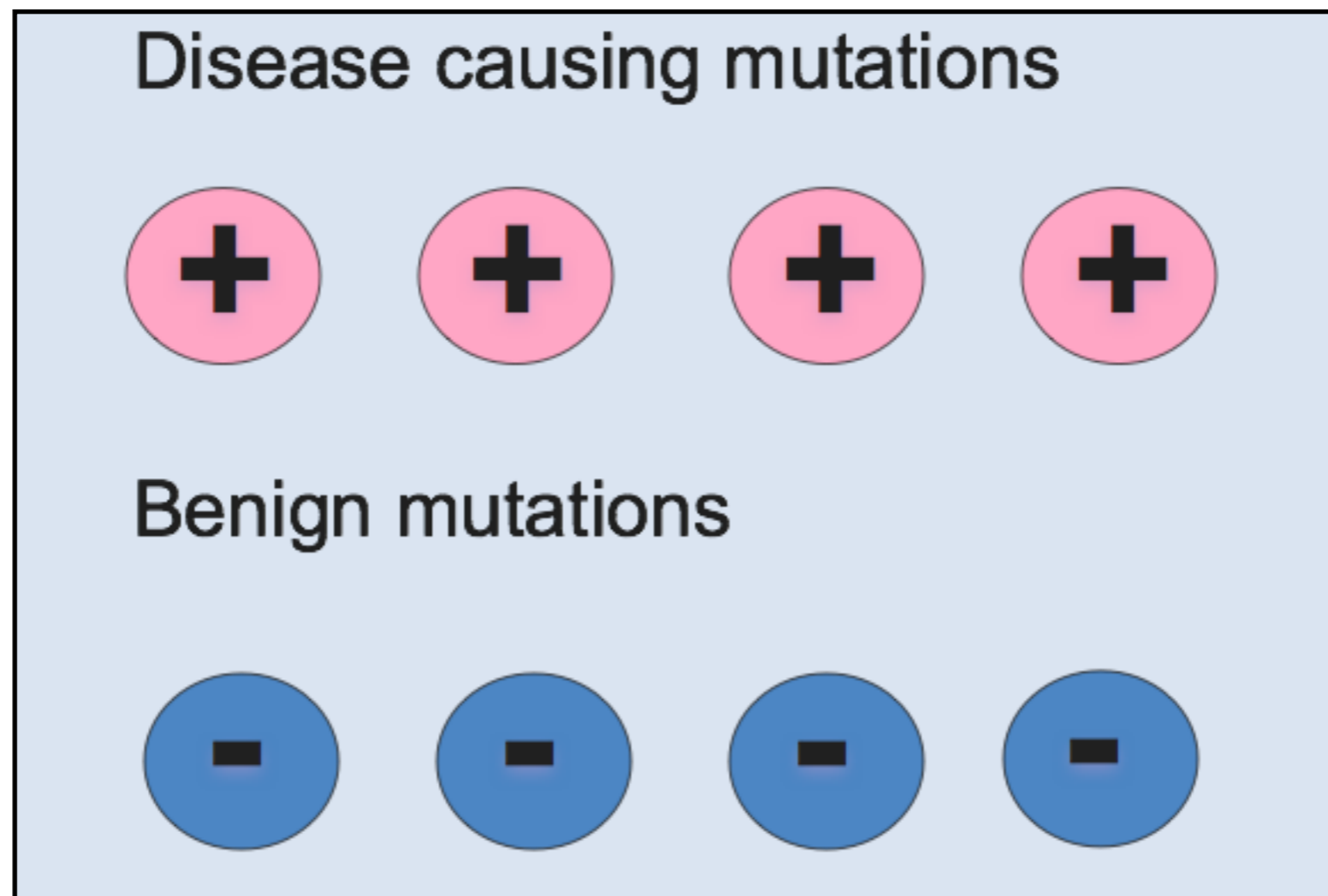


http://www.mquter.qut.edu.au/bio/bio2rdf_default.aspx

Mutation	Evolutionarily conserved	Charge Change	Binding Site
TP53 S362A	Yes	0	Yes
PIK3CA P539R	Yes	1	No
PIK3CA E545K	No	2	No

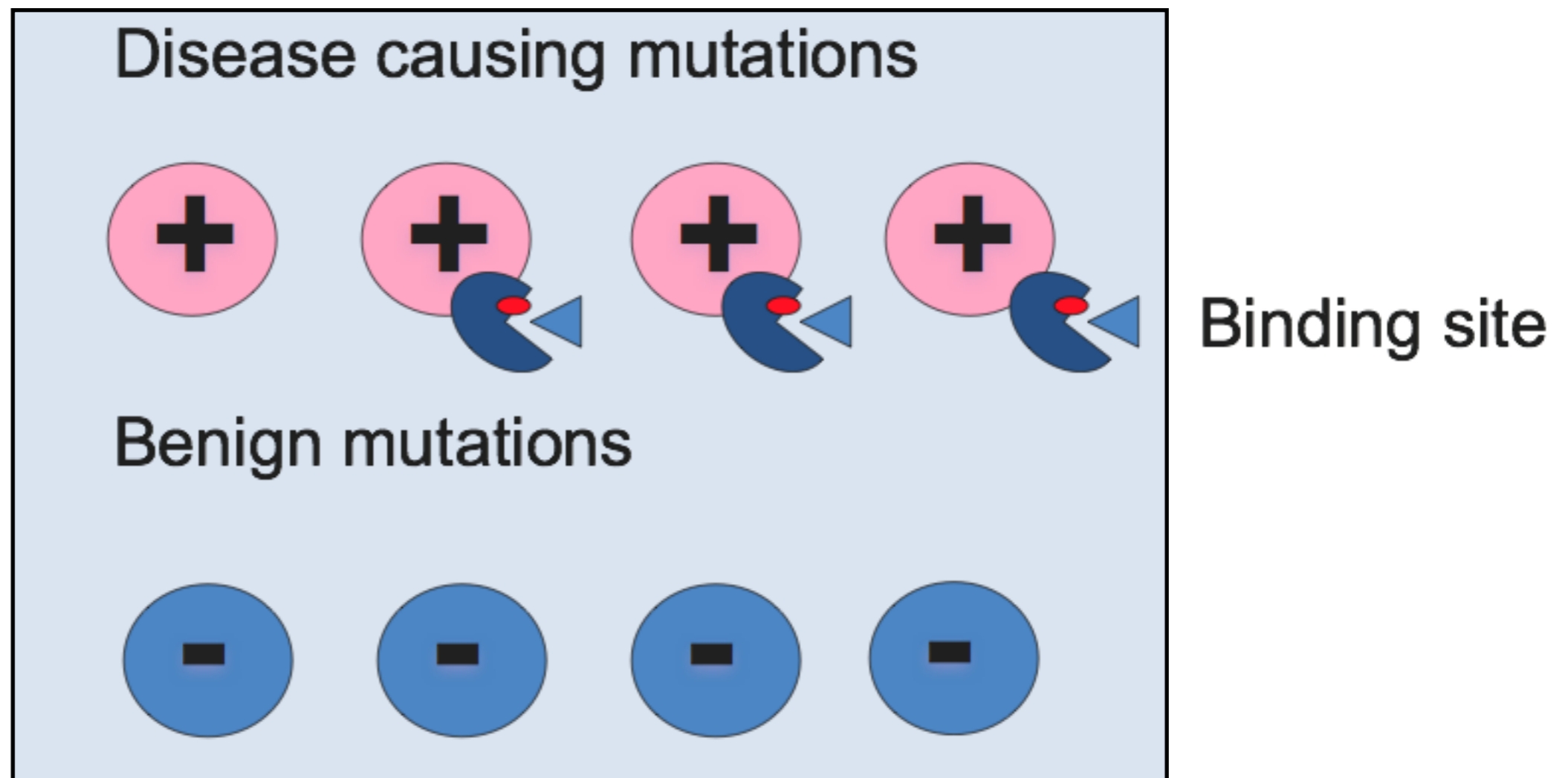
Training / validation set

- An empirical approach to assess if a feature is relevant
 - Collect examples from two classes



Training / validation set

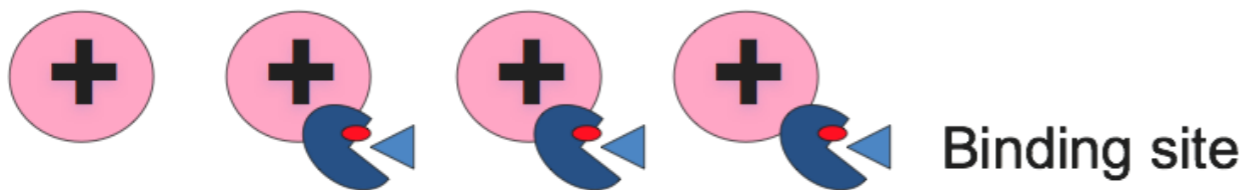
- An empirical approach to assess if a feature is relevant
 - Collect examples from two classes
 - Compute feature for all examples



Training / validation set



- An empirical approach to assess if a feature is relevant
 - Collect examples from two classes
 - Compute feature for all examples
 - Compute association between feature and class membership

Disease causing mutations

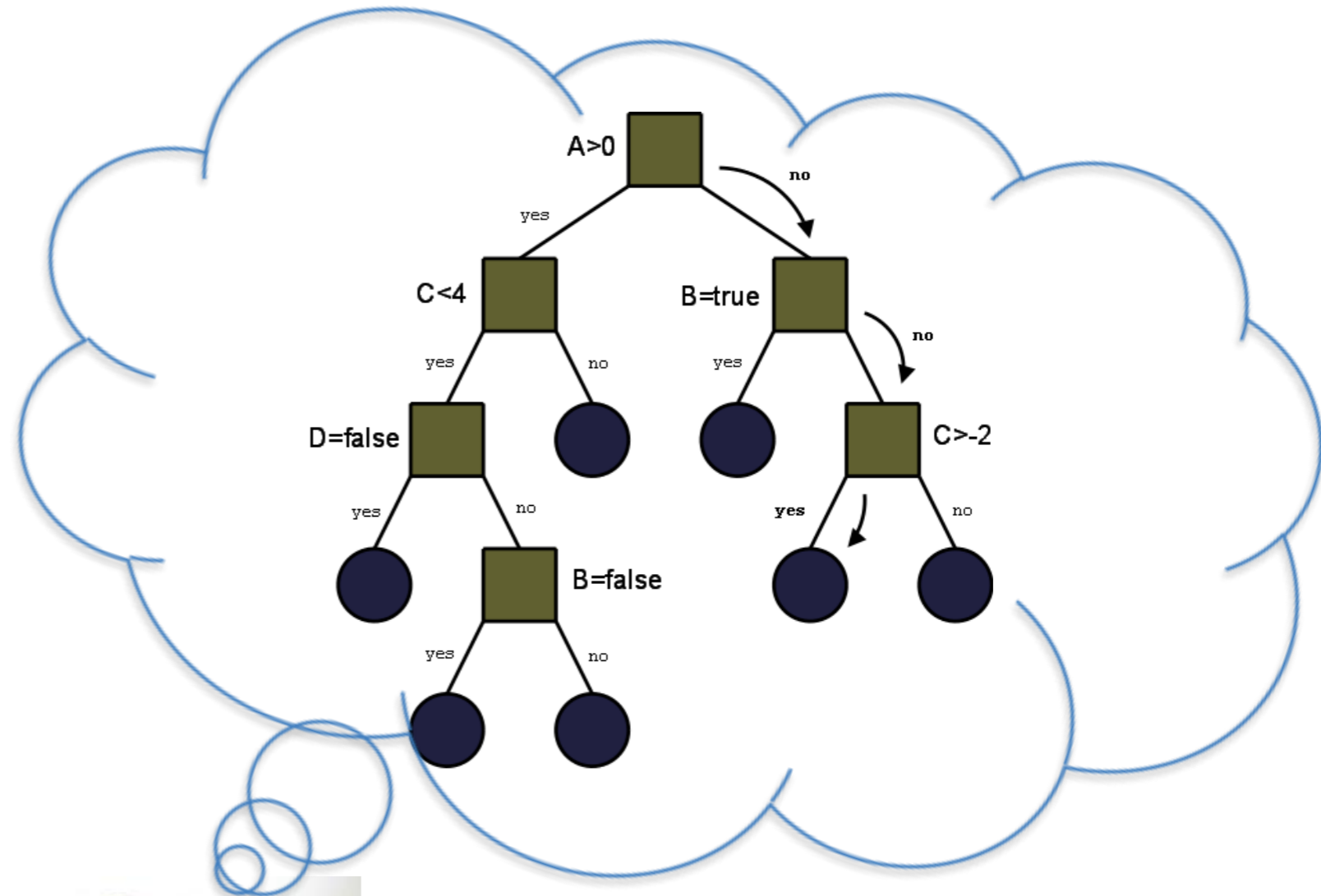


Benign mutations

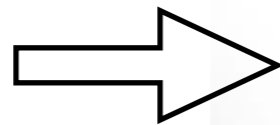
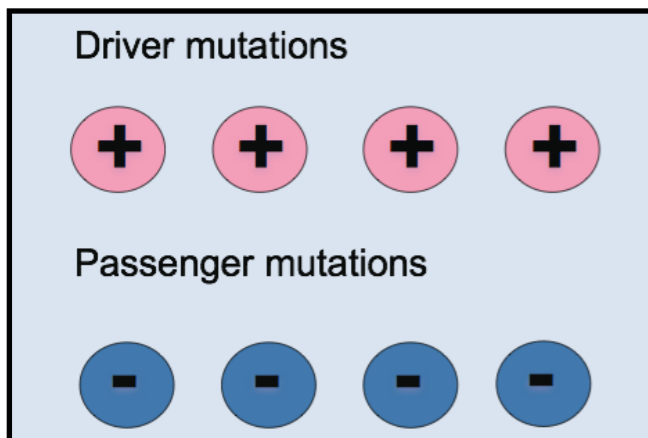


	Yes	No
	3	1
	0	4

Supervised machine learning



Training Set



<http://euolution.com/futurist-transhuman-news-blog/>

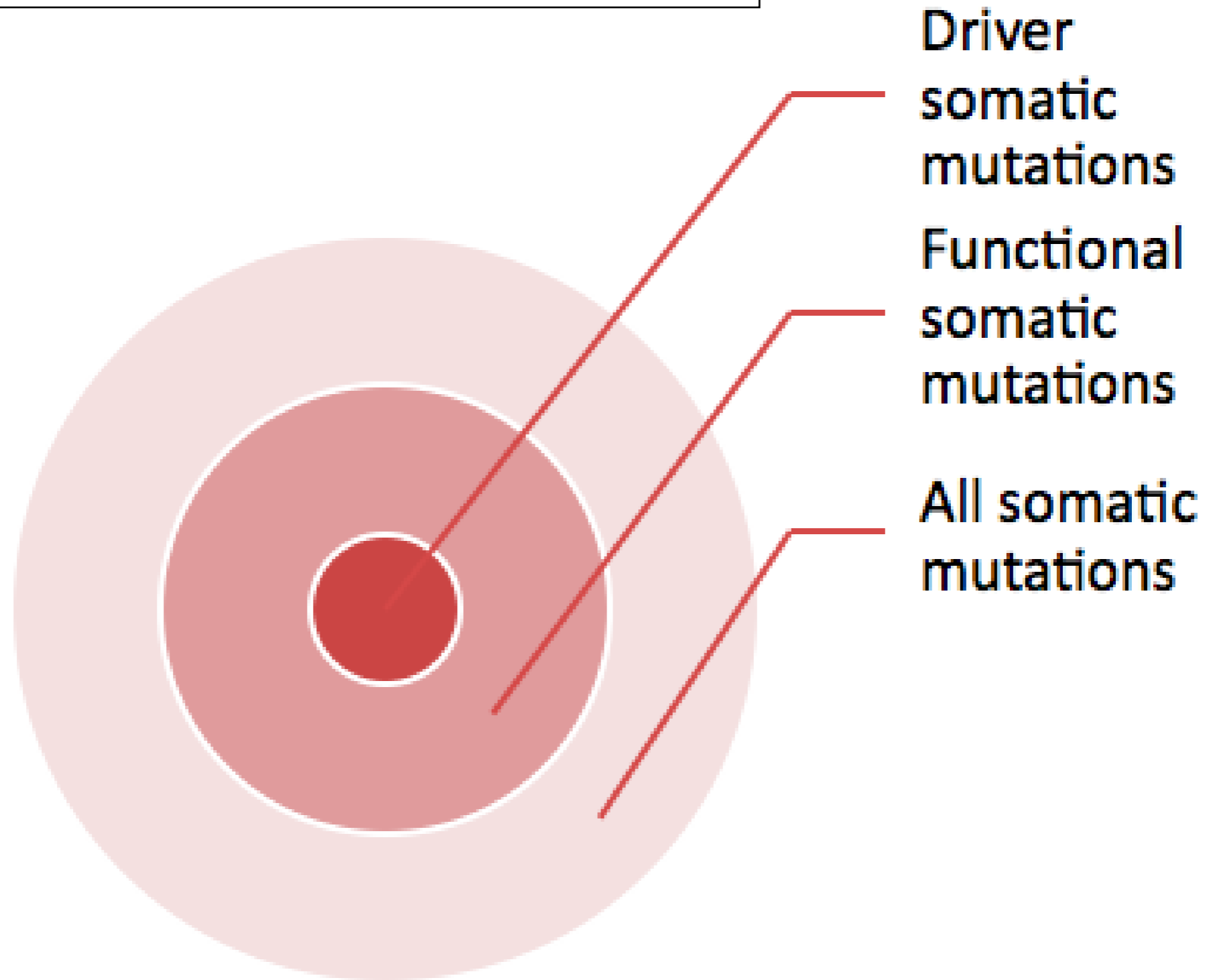
Does it work?

- Features capture the relevant differences between the classes
- Training set is representative of the actual population

Cancer Res. 2009 Aug 15;69(16):6660-7. Epub 2009 Aug 4.

**Cancer-specific high-throughput annotation of somatic mutations:
computational prediction of driver missense mutations.**

Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R.



CHASM Training Set



3300 missense mutations
(75) genes



Synthetic passenger missense mutations

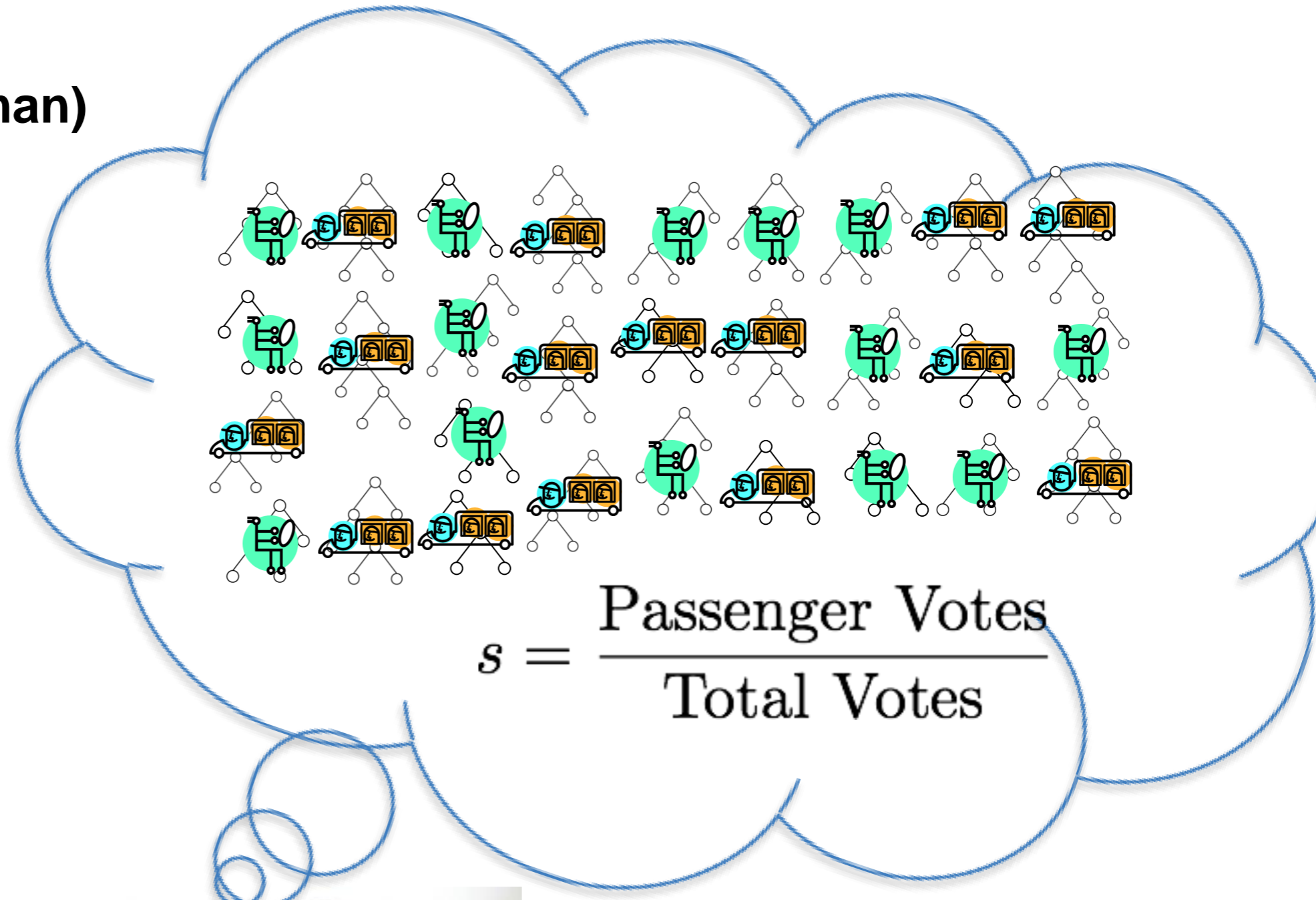
- Generate using mutation rates by di-nucleotide context for a particular tumor type

8 contexts: C*_pG, CpG*, TpC*, G*_pA, A, C, T, G

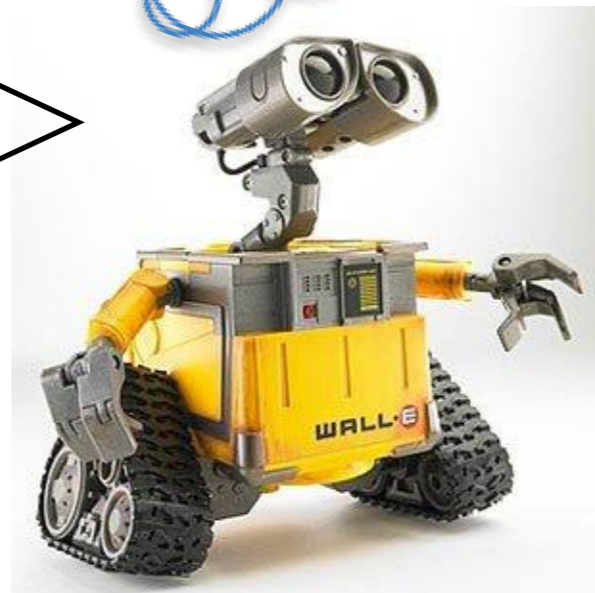
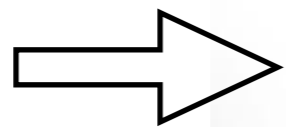
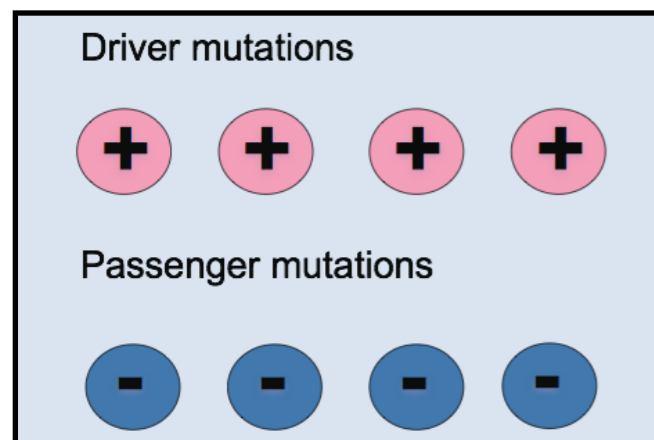
	C in CpG	G in CpG	C in TpC	G in GpA	A	C	G	T
A	0.05	0.97	0.31	0.44	0.00	0.29	0.50	0.39
C	0.00	0.02	0.00	0.22	0.13	0.00	0.13	0.39
G	0.02	0.00	0.21	0.00	0.62	0.20	0.00	0.22
T	0.93	0.01	0.48	0.33	0.25	0.51	0.37	0.00

Glioblastoma multiforme (GBM)

Random Forest (Amit/Geman, Breiman)



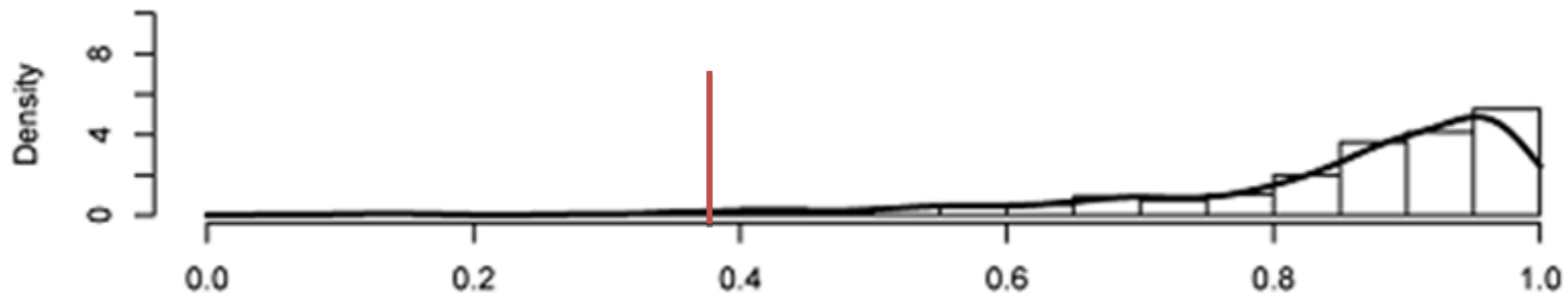
Training Set



CHASM P-value

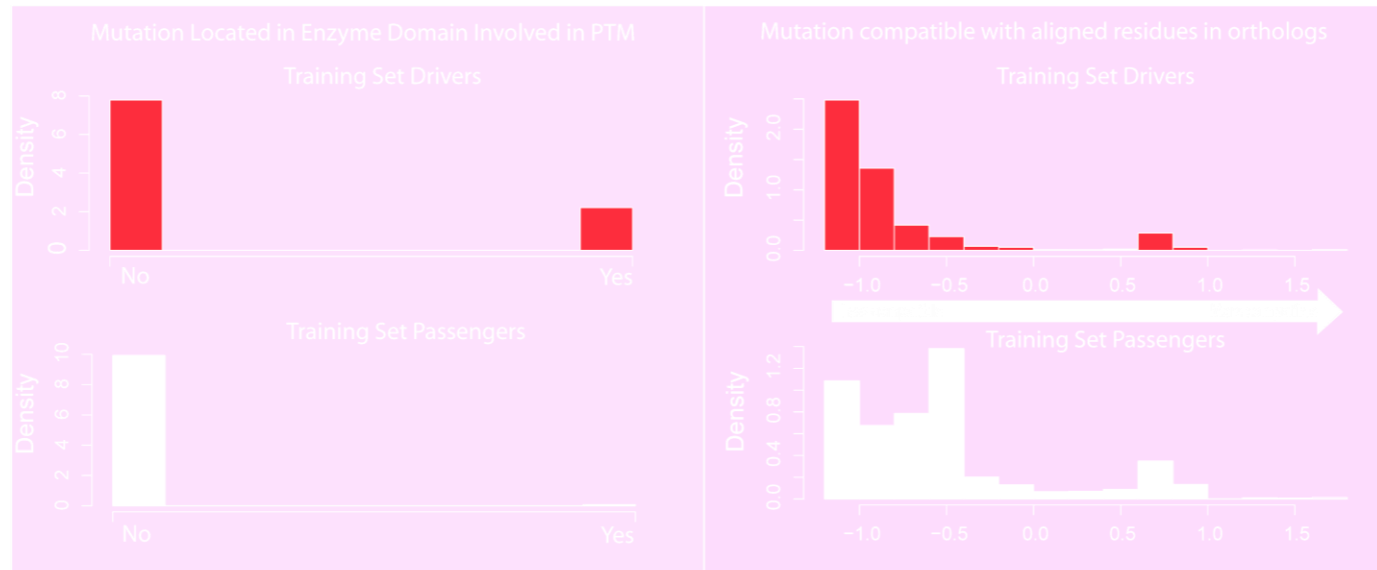
$$P(S \leq s | \text{Null hypothesis is true})$$

- Null hypothesis: mutation is a passenger
- Empirical null: scores of mutations that have (almost) no possibility of being drivers



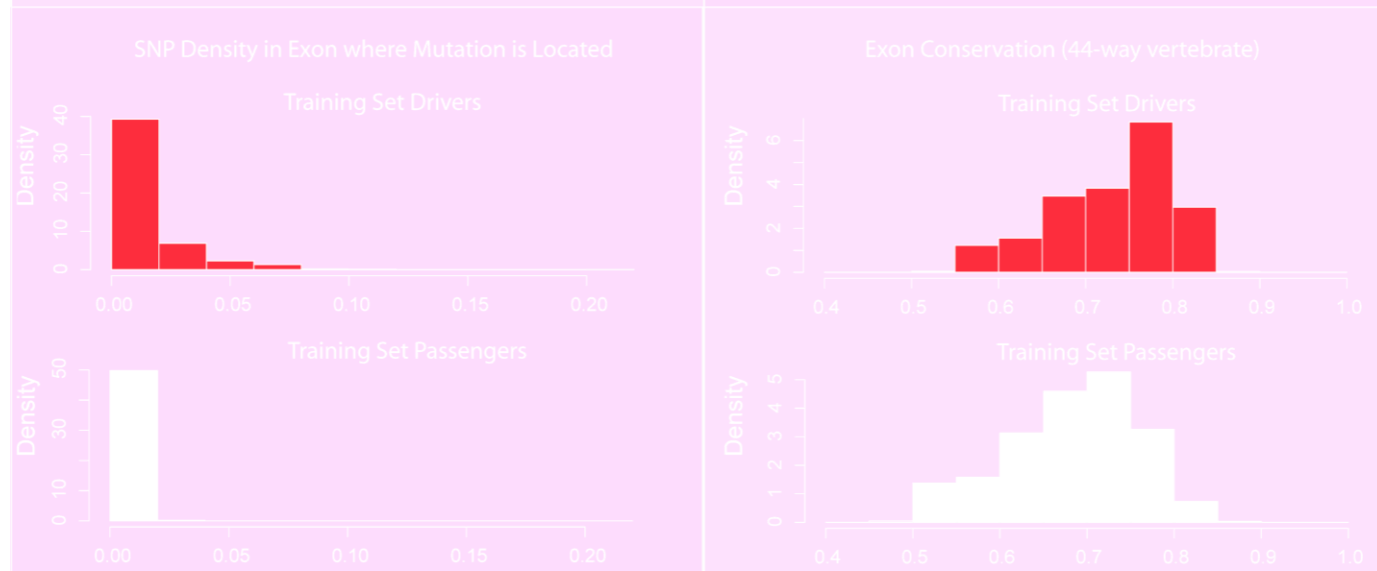
CHASM scores

PTM enzyme



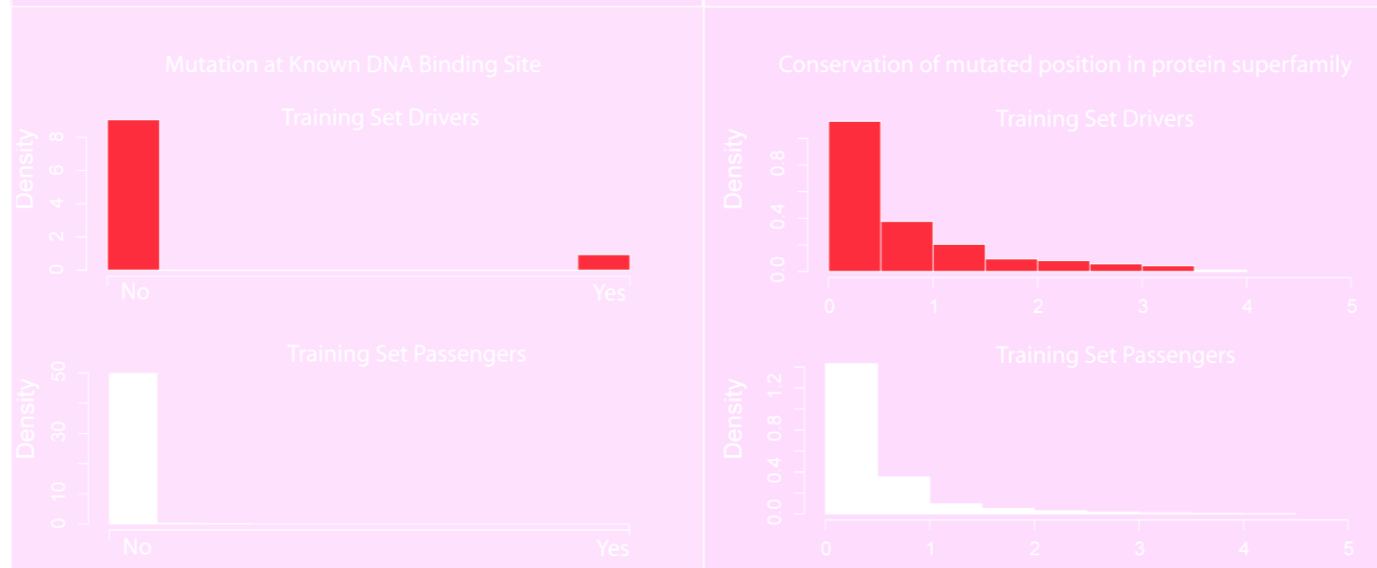
Ortholog
compatible
amino acid

SNP density

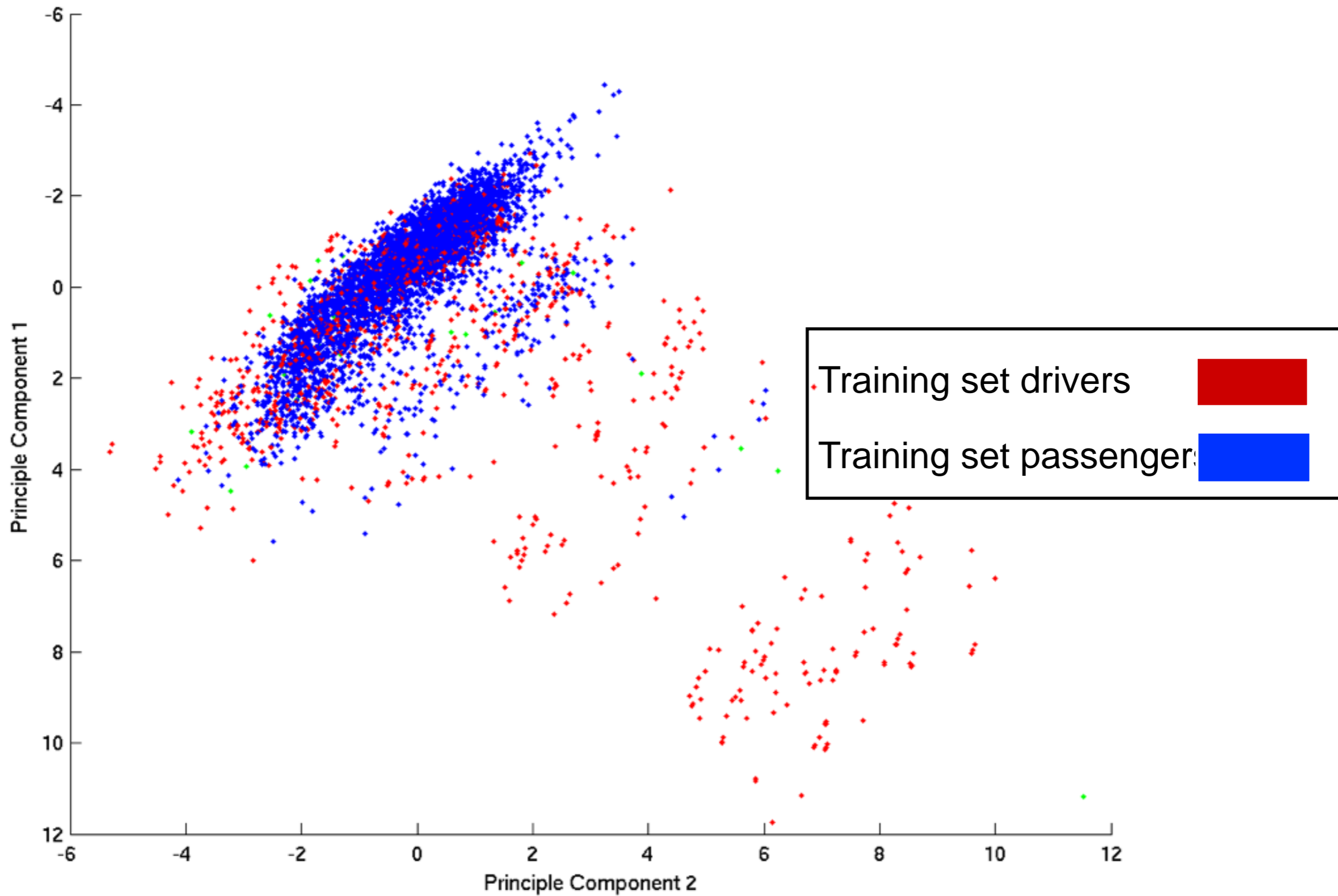


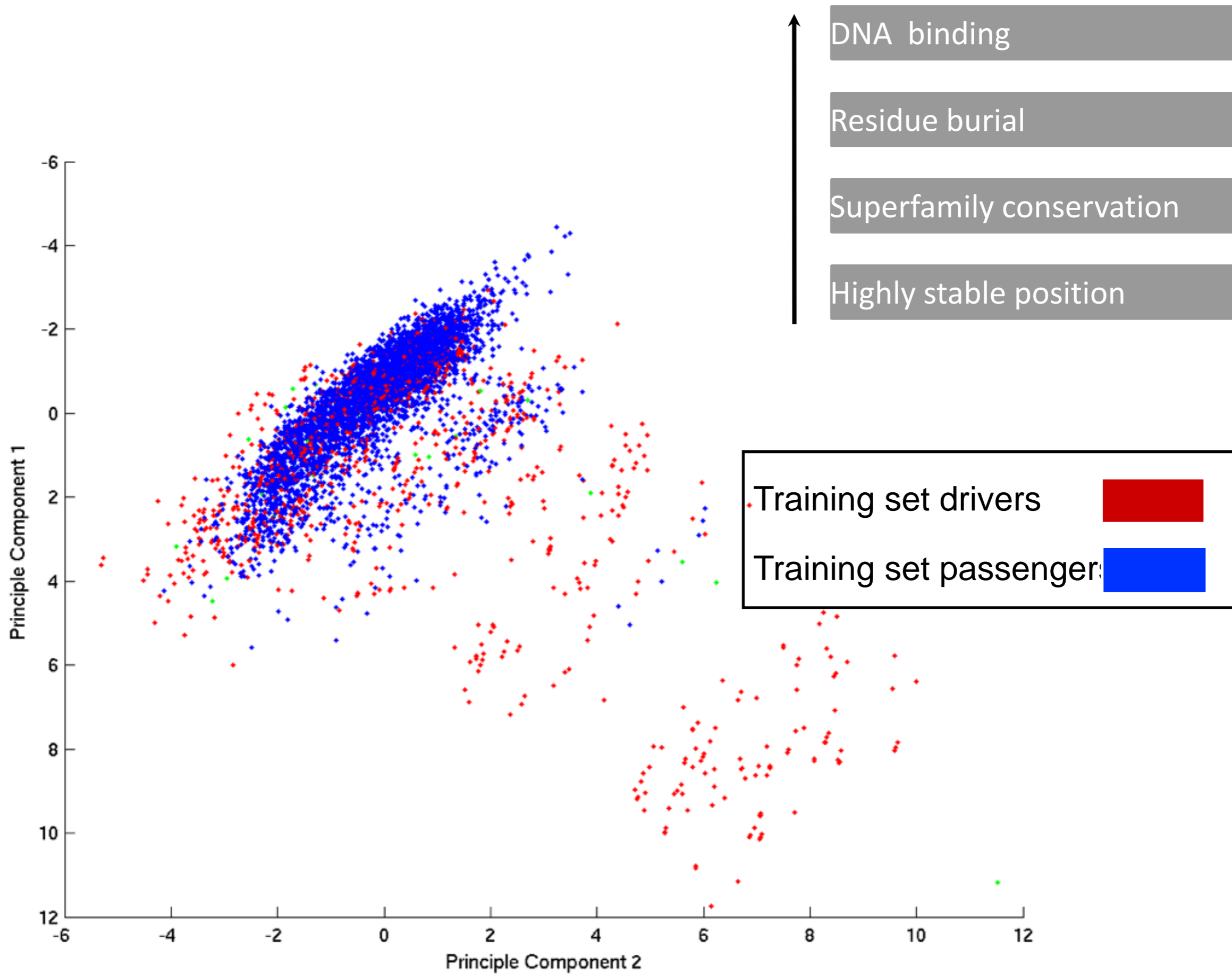
Exon
conservation

DNA binding



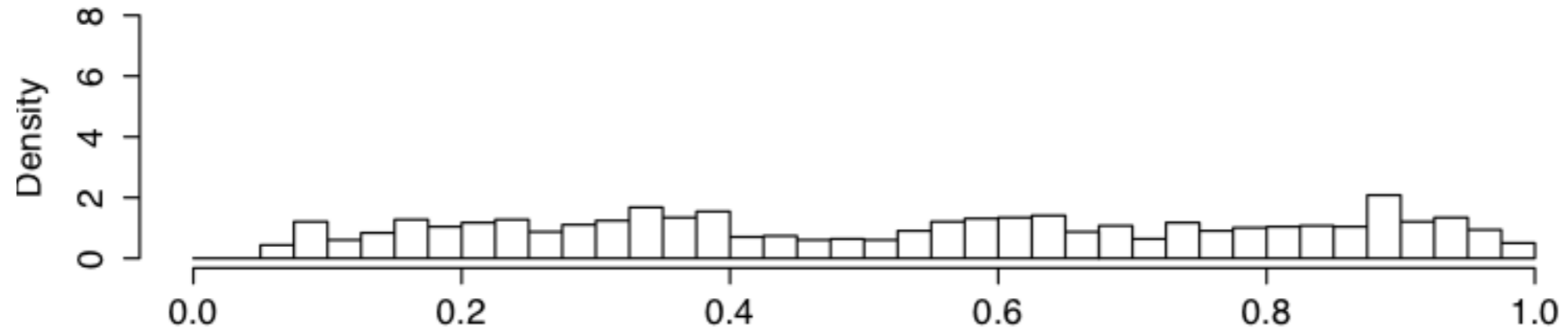
Superfamily
conservation



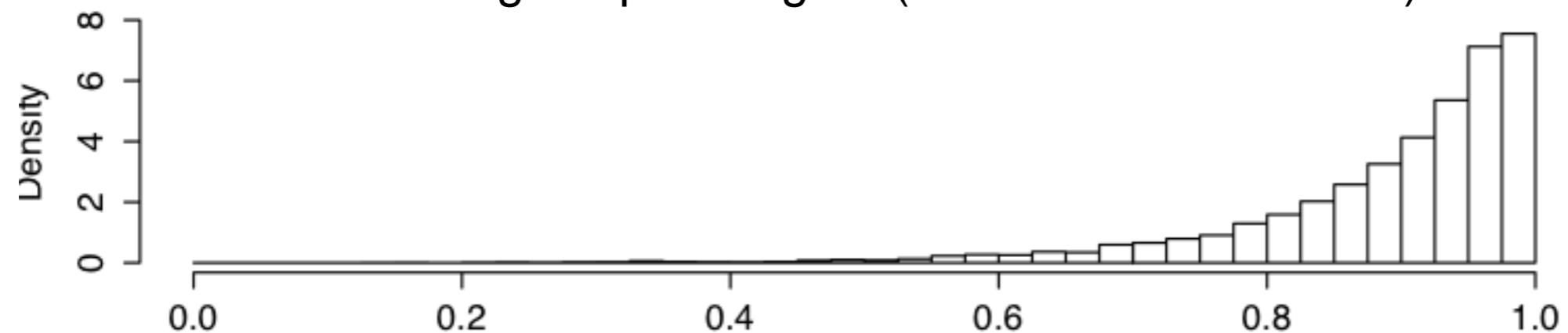


CHASM and false negatives

Training set drivers



Training set passengers (serous ovarian cancer)



CHASM score

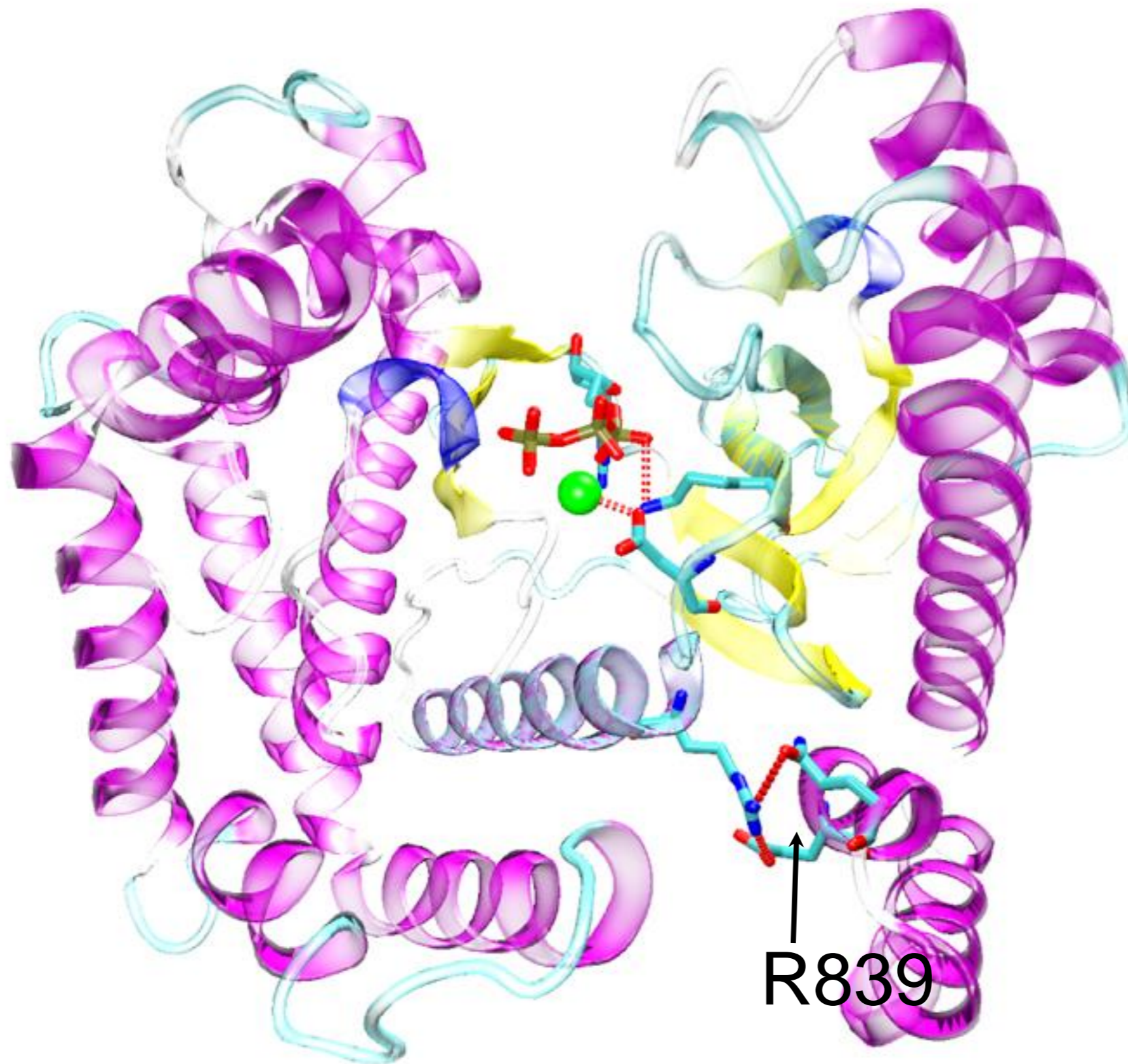
Prioritization of driver mutations in pancreatic cancer using cancer-specific high-throughput annotation of somatic mutations (CHASM).

Carter H, Samayoa J, Hruban RH, Karchin R.



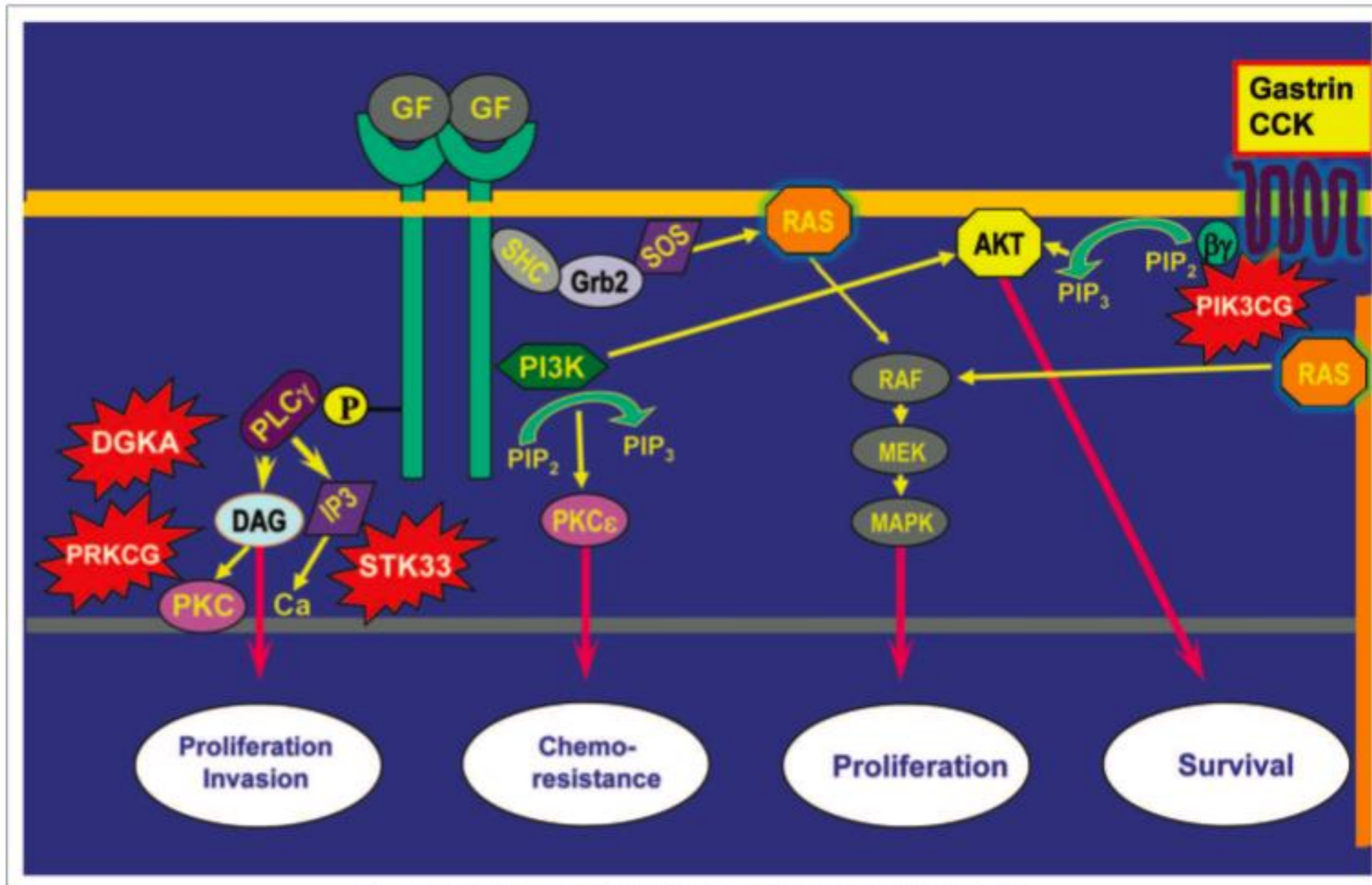
Hugo	Transcript	Mut	Zygoty	CHASM score	p-value	q-value<=	In training set?
TP53	CCDS11118.1	Y234C	Homo	0.034	0.0004	0.05	Yes
CDKN2A	CCDS6510.1	H98P	Homo	0.052	0.0004	0.05	Position Only
TP53	CCDS11118.1	I255N	Homo	0.062	0.0004	0.05	Yes
TP53	CCDS11118.1	S241F	Homo	0.066	0.0004	0.05	Yes
CDKN2A	CCDS6510.1	L63V	Het	0.068	0.0004	0.05	Yes
TP53	CCDS11118.1	L257P	Homo	0.072	0.0004	0.05	No
TP53	CCDS11118.1	C275Y	Homo	0.078	0.0004	0.05	Yes
TP53	CCDS11118.1	G266V	Homo	0.078	0.0004	0.05	Yes
TP53	CCDS11118.1	R248W	Homo	0.134	0.0004	0.05	Yes
NEK8	NP_835464	A197P	Het	0.144	0.0004	0.05	
PIK3CG	CCDS5739.1	R839C	Homo	0.166	0.0008	0.05	
TP53	CCDS11118.1	H179R	Homo	0.180	0.0013	0.05	Yes
SMAD4	CCDS11950.1	C363R	Homo	0.184	0.0013	0.10	Position Only
TP53	CCDS11118.1	R282W	Homo	0.198	0.0013	0.10	Yes
KRAS	CCDS8702.1	G12D	Het	0.202	0.0013	0.10	Yes

PIK3CG



Driver mutations: A roadmap for getting close and personal in pancreatic cancer.

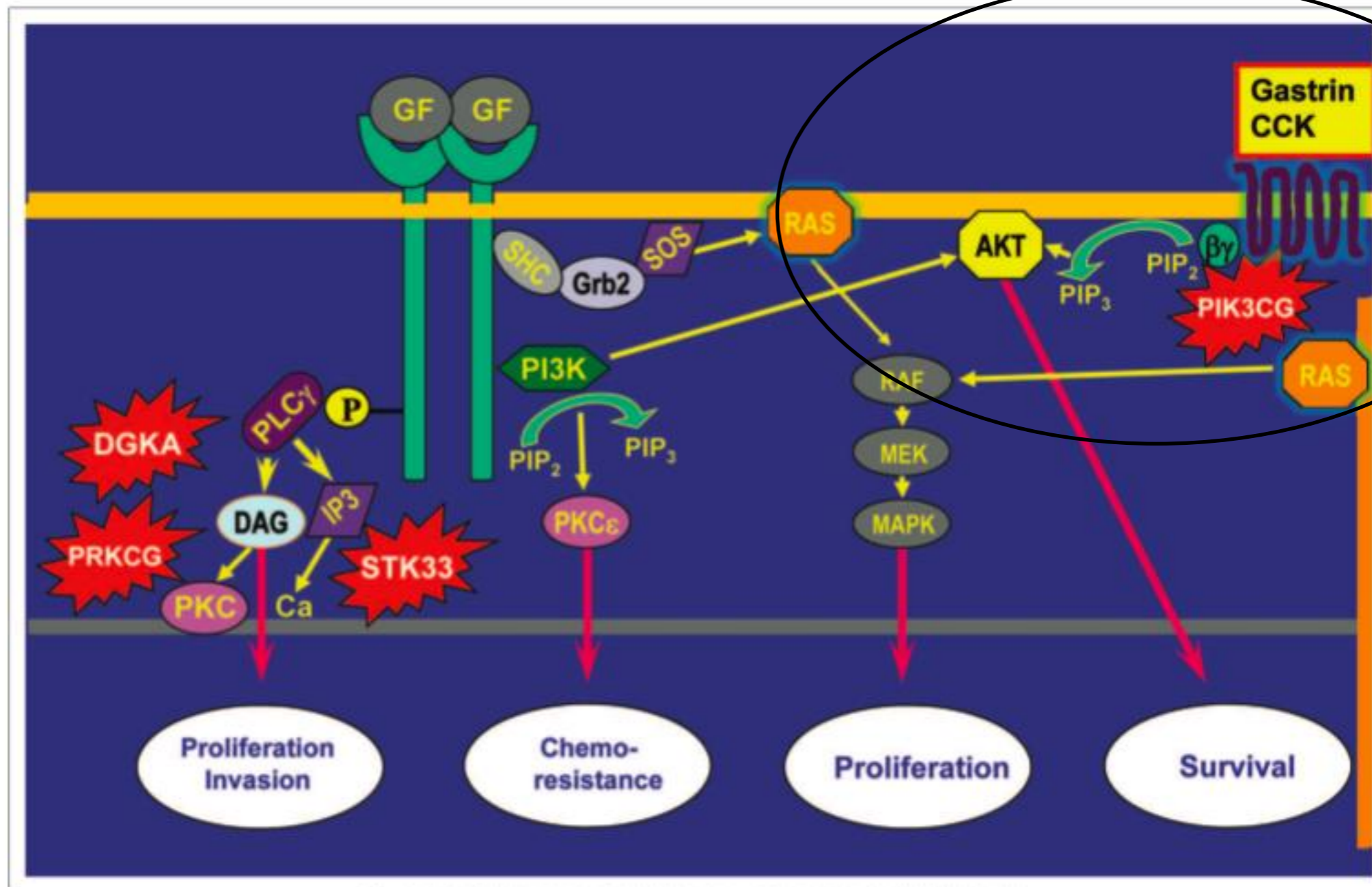
Korc M.

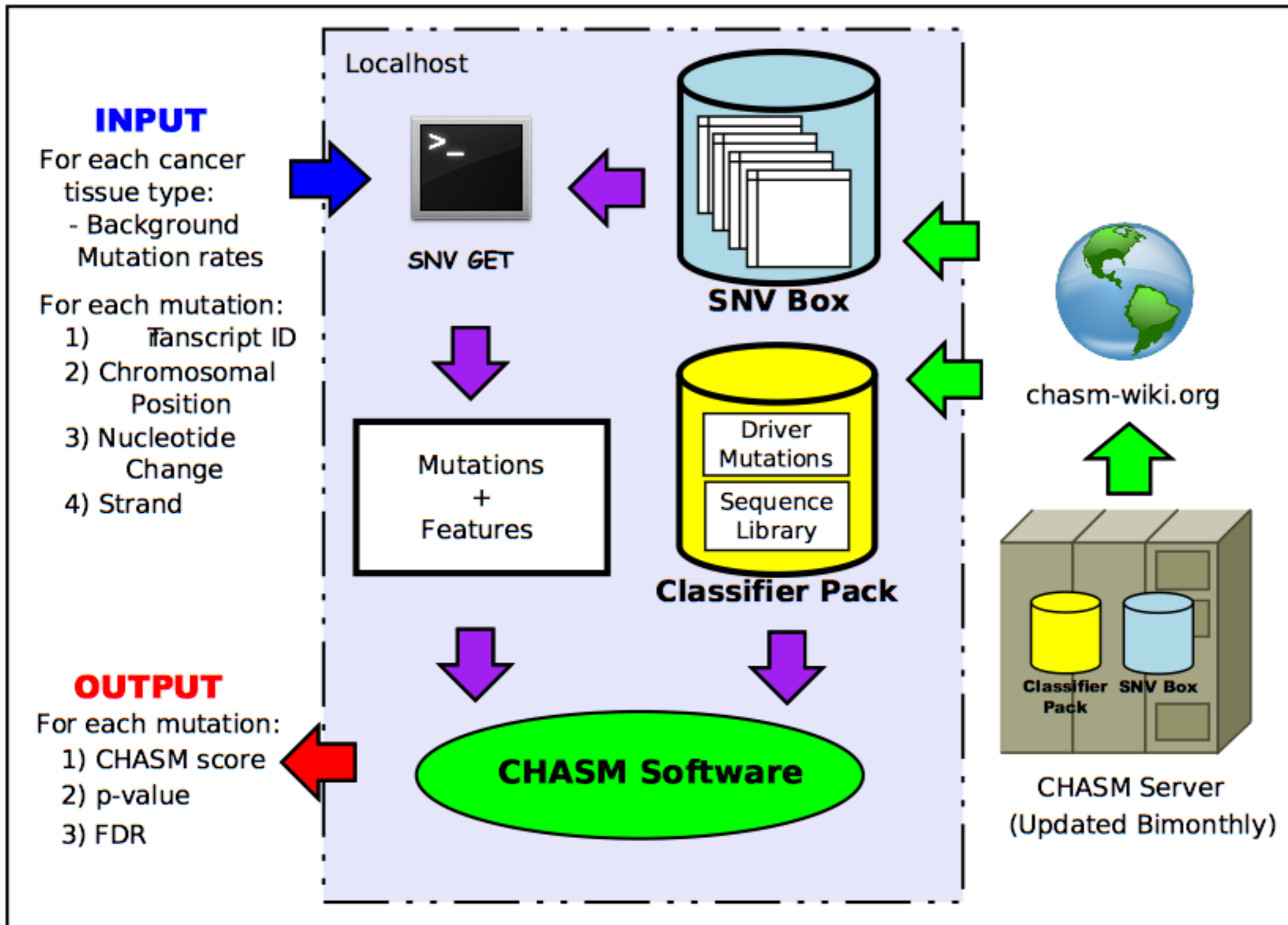


©2010 Landes Bioscience.

Potential to cross-talk with Kras-driven pathways?

PIK3CG may be a tumor suppressor in pancreatic cancer

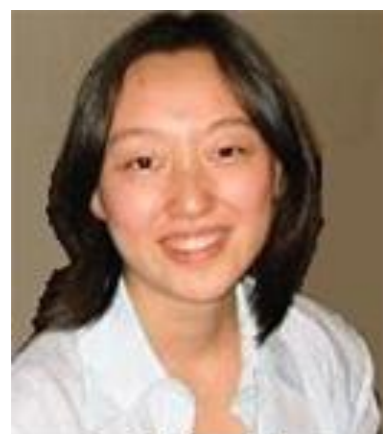
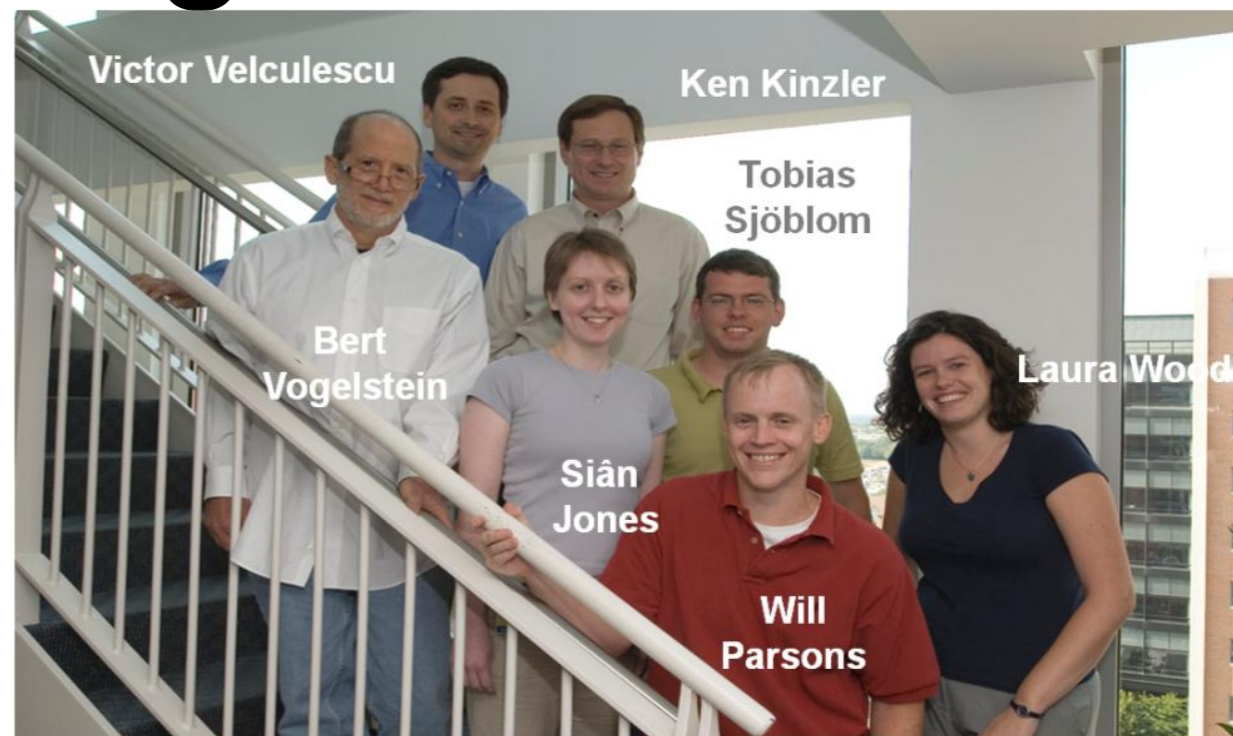




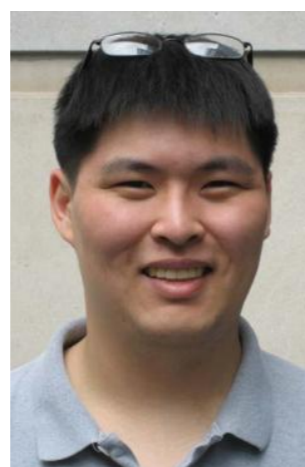
Acknowledgments



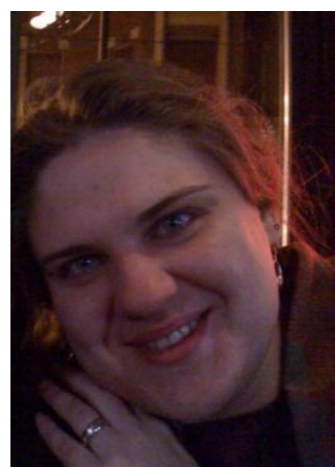
Hannah Carter



Dr. Sining Chen



Dewey Kim



Dr. Svitlana Tyekucheva



Andy Wong



Mark Diekhans



Dr. Kideok Jin



Dr. Saraswati Sukumar



NIH R21 CA135866
NSF DBI 0845275
Susan G. Komen KG080137
DoD NDSEG graduate fellowship 32 CFR 168a

