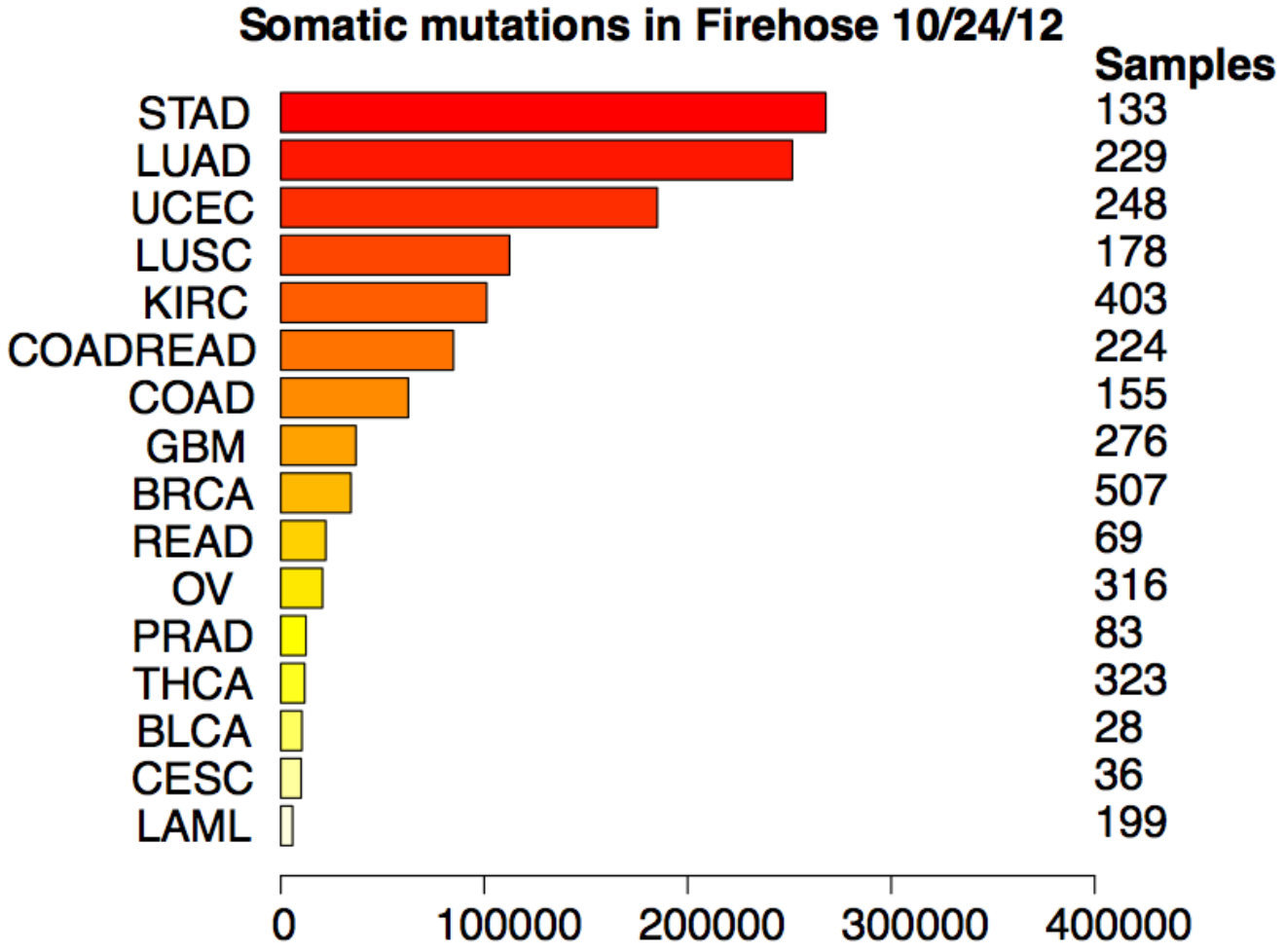
CRAVAT and MuPIT Interactive: Web Tools for Cancer Mutation Analysis

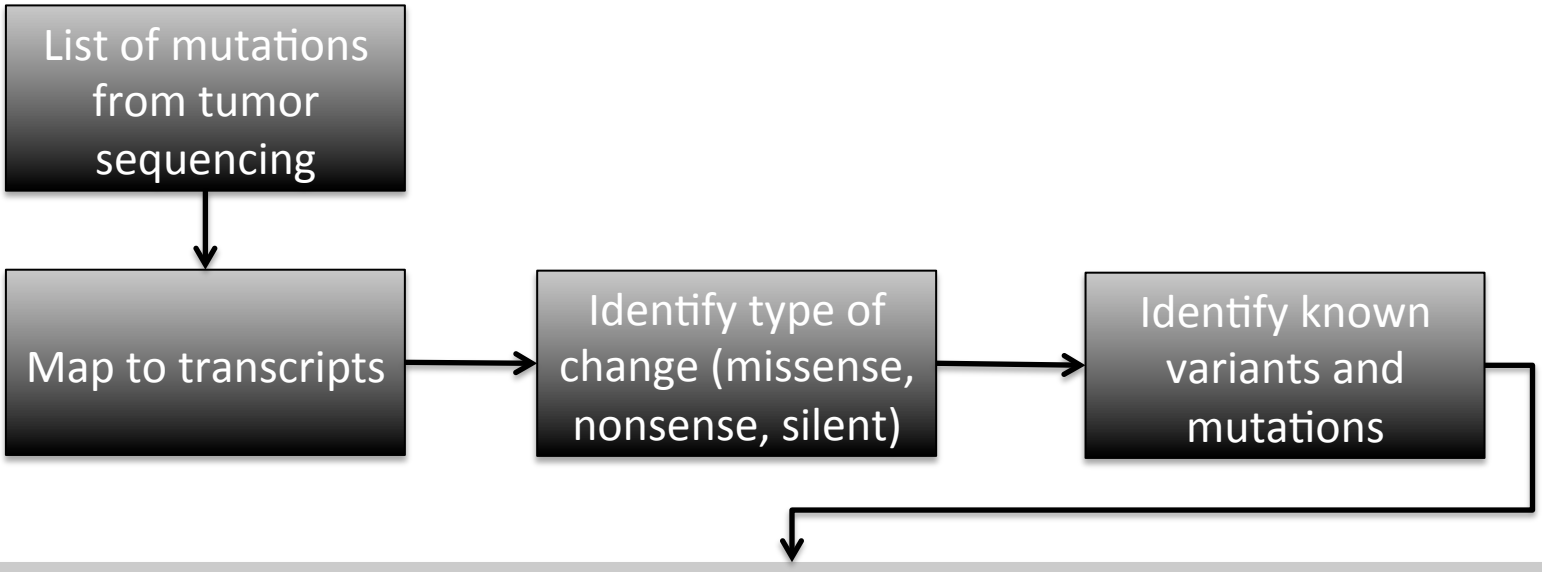
Rachel Karchin, Ph.D.
Department of Biomedical Engineering
Institute for Computational Medicine
Johns Hopkins University

The Cancer Genome Atlas' 2nd Annual Scientific Symposium
November 27-28, 2012

Need for computational tools to analyze large-scale cancer mutation data



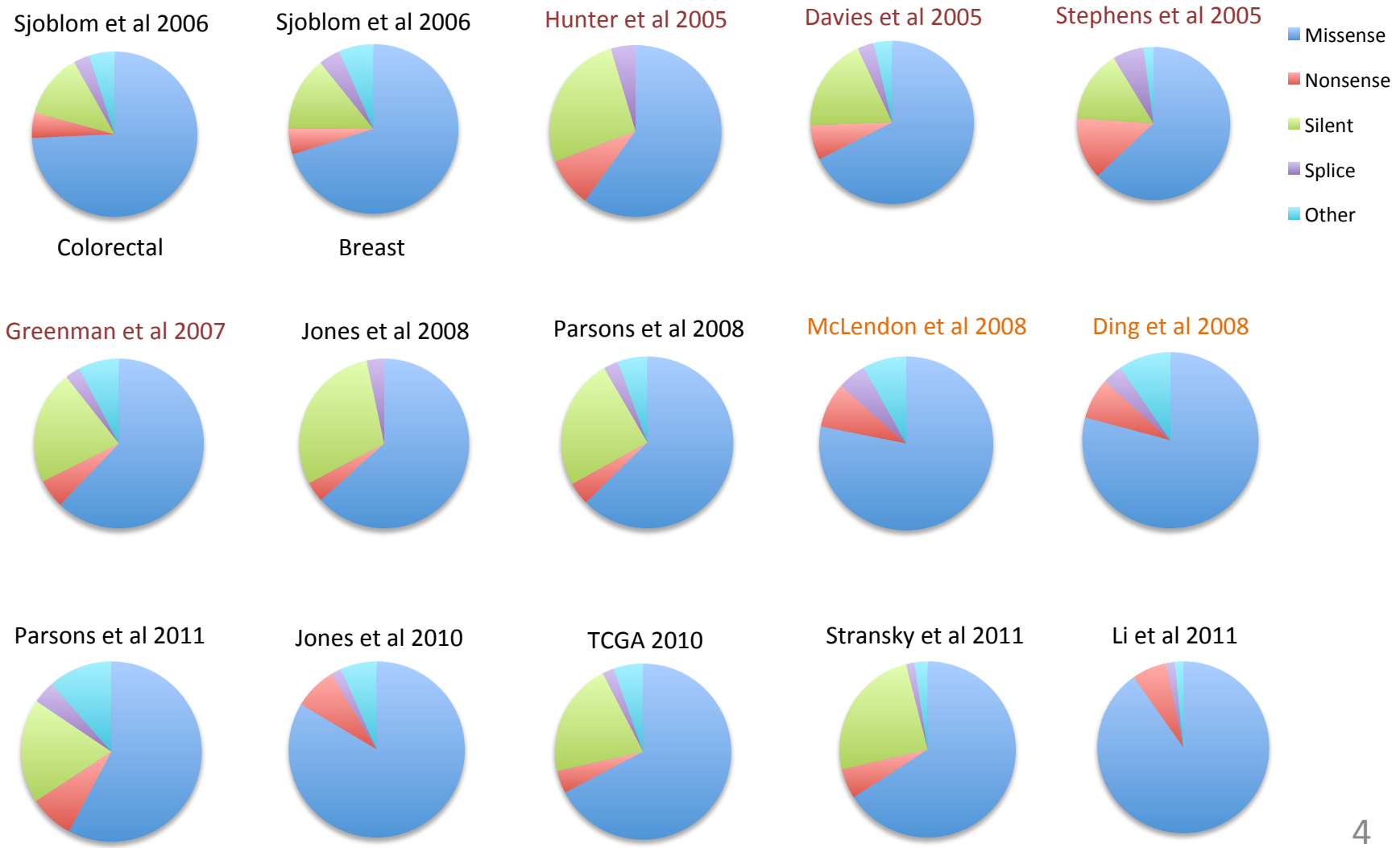
Goal is to provide an end-to-end mutation analysis workflow



Analysis

- Predict driver vs. random mutations
- Predict functional impact of mutations
- Visualize mutations on tertiary structure
- Find significantly mutated genes and pathways

The majority of somatic mutations in tumor exomes are missense



http://www.cravat.us

Tools for evaluating missense mutations

- CHASM: cancer driver analysis
- VEST: Functional effect analysis
- Annotations (1000g, ESP6500, COSMIC, GeneCards, PubMed)

http://mupit.icm.jhu.edu

Interactive visualization on 3D protein structure

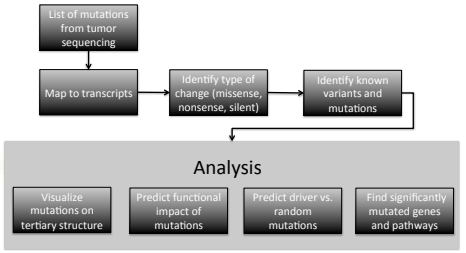
- Automatic mapping onto available structures
- Simple interactive interface
- UniProtKB feature table annotations provided
- Publication quality figures

Outline

- Introduction to CRAVAT
- Introduction to MuPIT
- Future plans
- Your input

CRAVAT Web Server

http://www.cravat.us



CRAVAT

Help | Release note

1 Input

Check for input example.

Enter variants below:

or upload a variant file:

Choose File No file chosen Clear

hg18

2 Analysis

Choose analysis type

- Cancer driver analysis
- Functional effect analysis
- Annotation only

Choose analysis program

- CHASM
- SnvGet

Breast

Include gene annotation

3 Results

Send the analysis report to:

Enter your email

Include text reports

Include MuPIT Interactive input file

SUBMIT

What will I get?

Input

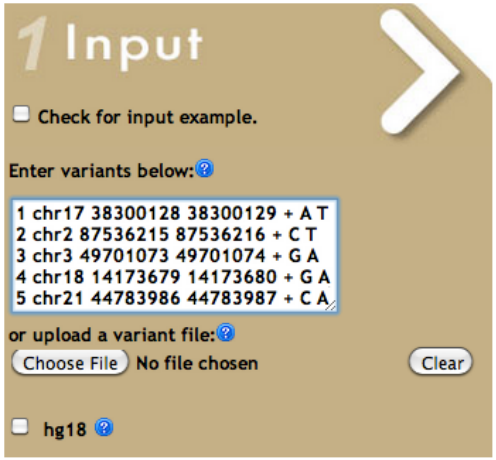
List of mutations from tumor sequencing

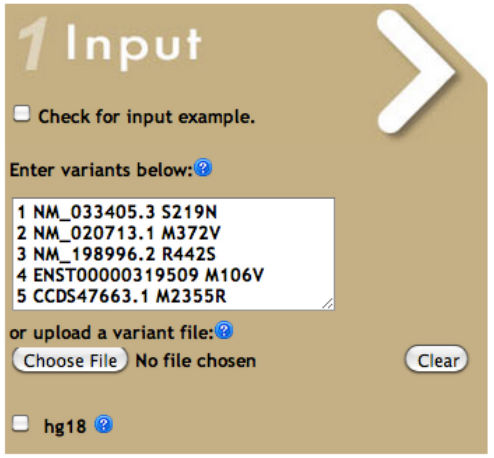
- Paste into text box or upload text file

RECOMMENDED

- Genomic Coordinates
  - Unique id for variant
  - Chromosome
  - 0-based start position
  - 1-based end position
  - Strand on which bases are reported
  - Reference base
  - Alternate base
  - sample ID (optional)

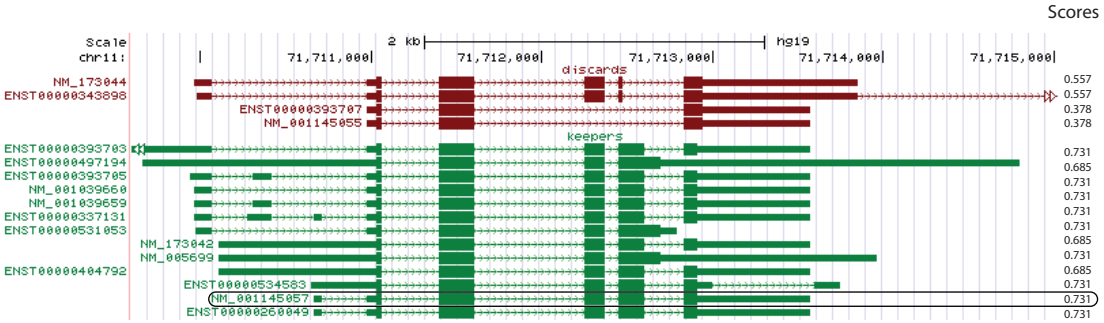
- Transcript Coordinates
  - Unique id for variant
  - Transcript
  - Amino Acid Substitution
  - sample ID (optional)





Map to transcripts

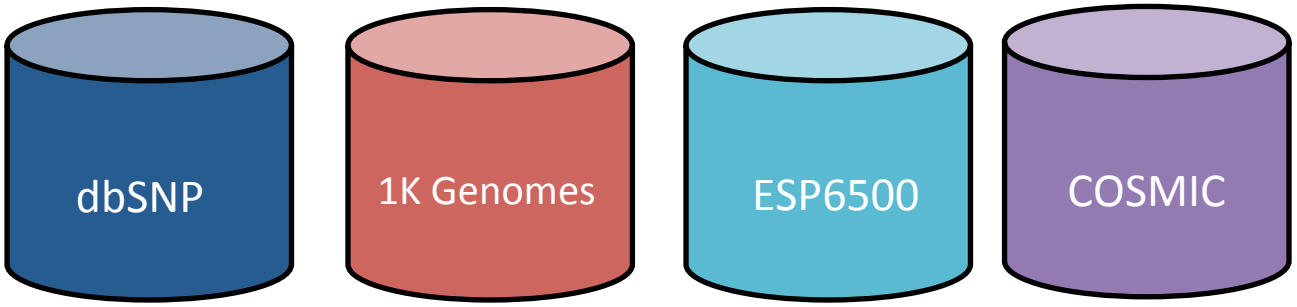
- Transcripts are scored by coverage of coding bases and agreement between RefSeq and Ensembl transcript definitions.
- A greedy algorithm selects the “best transcript” for each mutated position



Douville *et al.* CRAVAT: Cancer-Related Analysis of VAriants Toolkit. 2012 (submitted)

Identify known variants and mutations

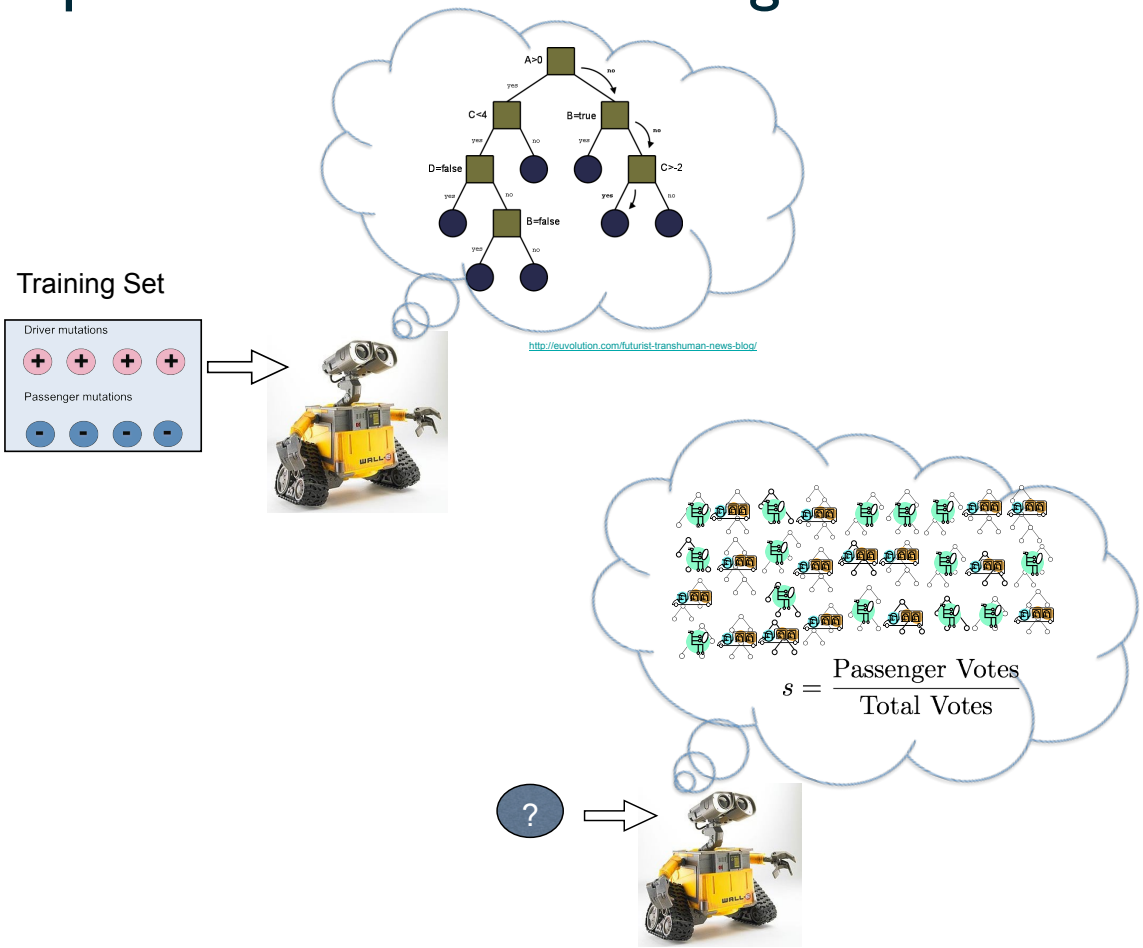
HUGO symbol	Transcript	Amino acid position	Amino acid change	dbSNP	1000 Genomes allele frequency	ESP6500 allele frequency (European American)	ESP6500 allele frequency (African American)	Occurrences in COSMIC [exact nucleotide change]	Occurrences in COSMIC by primary sites [exact nucleotide change]
RALYL	NM_173848.5	270	M270I		0	0	0	1	upper_aerodigestive_tract(1)
TRIM29	NM_012101.3	216	P216H		0	0	0	1	breast(1)
ZNF317	NM_001190791.1	280	G280V		0	0	0	1	large_intestine(1)
CD248	NM_020404.2	115	T115N	rs140616335	0	0	0.0007	0	
SRPX2	NM_014467.2	393	R393Q		0	0	0.0007	0	
KDR	NM_002253.2	539	G539R	rs55716939	0	0.0022	0.0005	0	
LGI2	NM_018176.3	322	R322H	rs149130054	0.0005	0.0001	0	0	
LILRB1	NM_006669.3	189	P189L		0	0.0001	0	0	

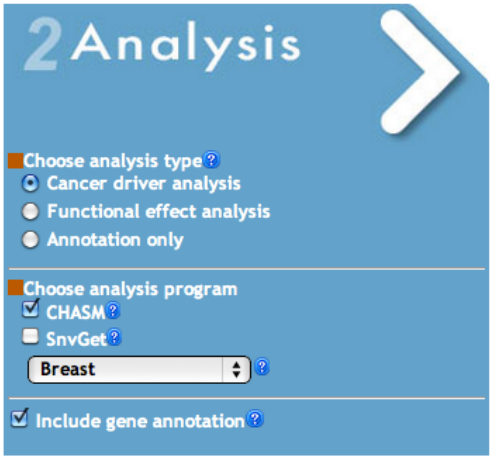


Analysis

Predict driver vs. random mutations

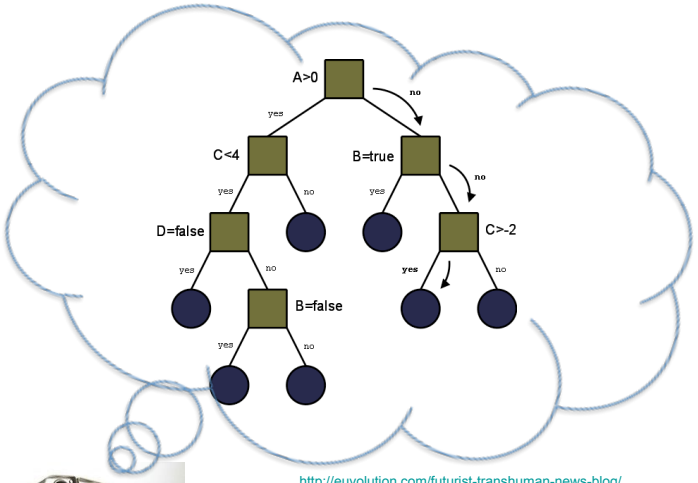
Supervised machine learning





Carter *et al.*. Cancer-specific high-throughput annotation of somatic mutations. Cancer Research. 2009

Where do the decision rules come from?



http://euvolution.com/futurist-transhuman-news-blog/

amino acid substitution properties

human polymorphism properties

predicted local protein structure

UniprotKB features

regional amino acid composition

vertebrate ortholog conservation (genome alignments)

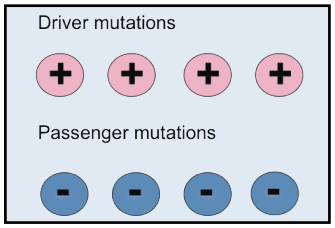
superfamily homolog conservation (deep protein alignments)

amino acid compatibility with orthologs and homologs

86 features pre-computed exome-wide in SNVBox

Wong *et al.*. CHASM and SNVBox toolkit. Bioinformatics. 2011

Where does the training set come from ?



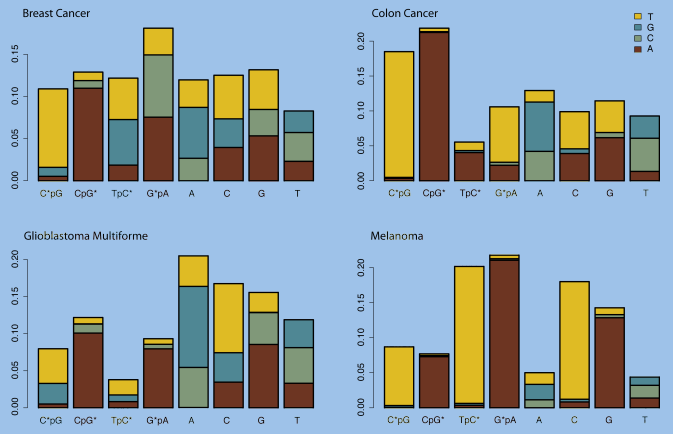
Driver missense mutations:

- Gene harbors at least 5 mutations
- Oncogene: >15% of NS mutations occur at the same position
- Tumor suppressor: > 15% of NS mutations are inactivating
- All missense mutations in these genes are included

Bozic et al PNAS 2010
Jones et al Science 2010
Vogelstein et al. submitted

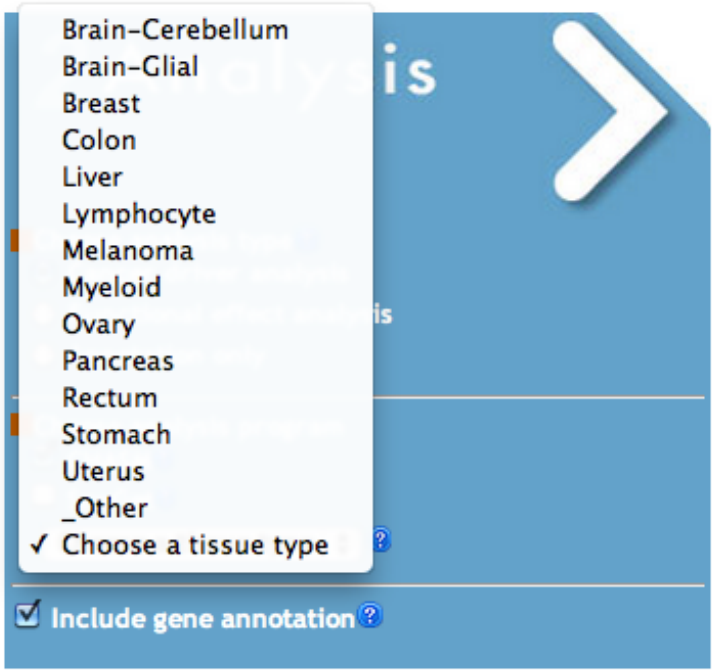
Random passenger missense mutations:

- Generated in silico with a generative model
- Designed to match tumor tissue di-nucleotide mutation spectrum
- In CHASM feature space, they do not look like high allele frequency SNPs



Tissue-specific classifiers in CRAVAT

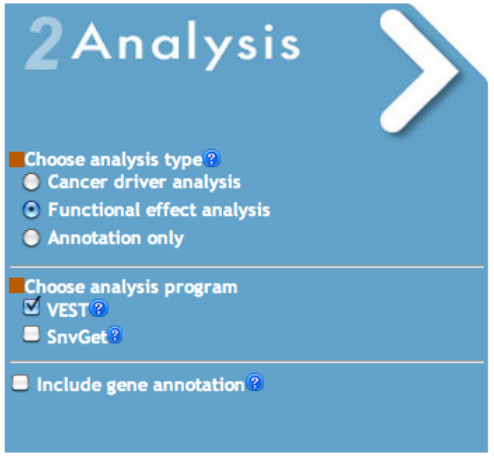
- Select from pull down list
- Pre-constructed classifiers for tumor types under analysis by TCGA and ICGC
- 'Other' offers a generic classifier if tissue under study is not yet supported



Analysis

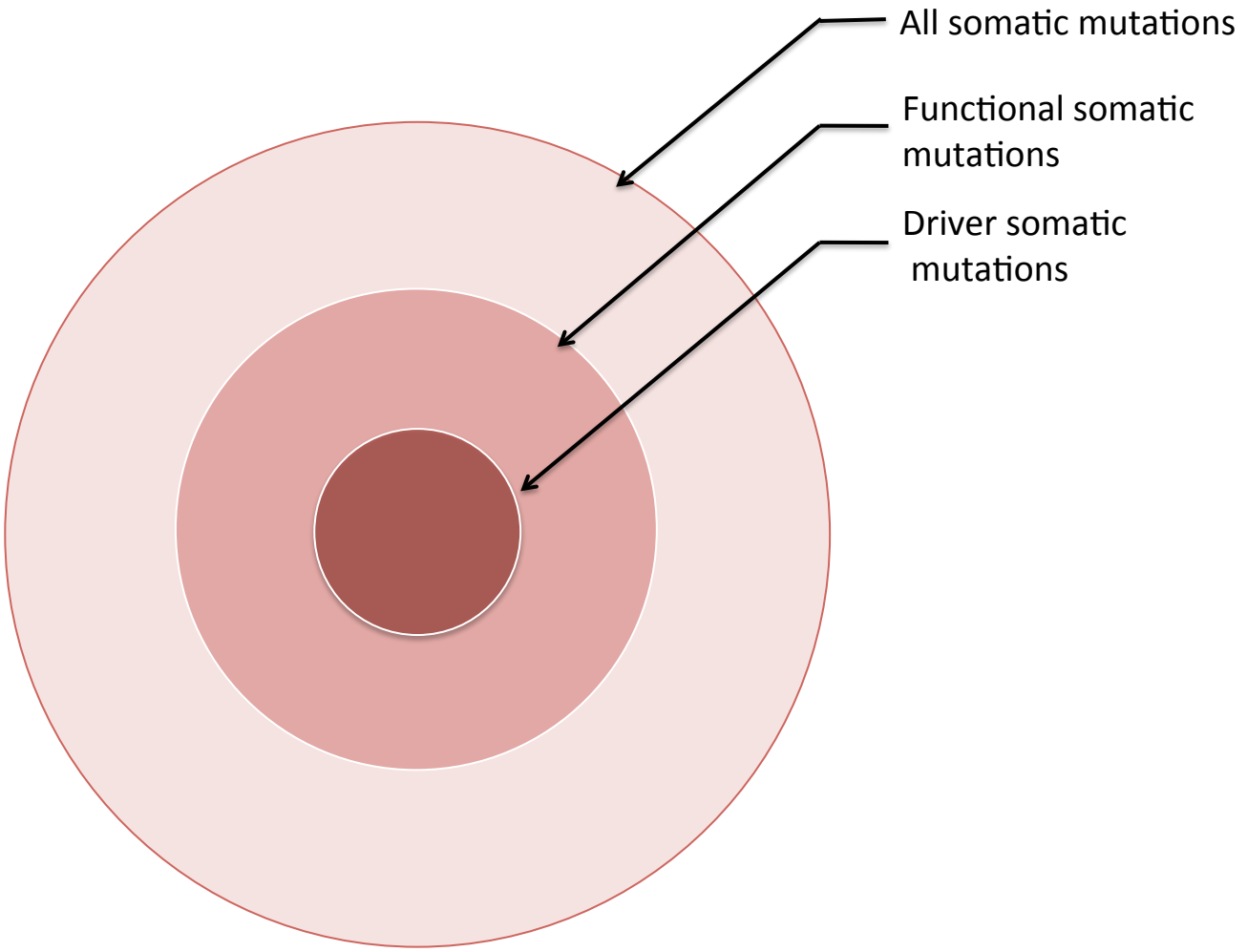
Predict functional impact of mutations

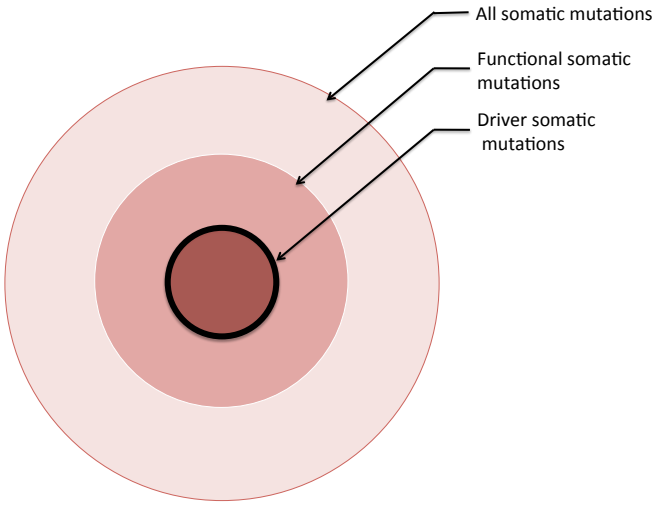
- Variant Effect Scoring Tool (VEST)
  - Identify functional mutations
  - Same supervised learning algorithm and features as CHASM
  - Different training set
    - 47000 disease missense mutations from HGMD[1] 2012v2
    - 45000 neutral missense variants from the ESP6500[2] (AF>1%)

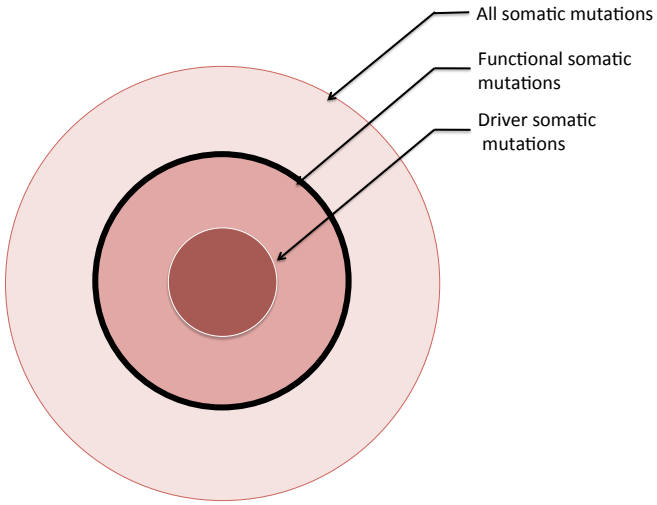


1 Stenson P, Mort M, Ball E, Howells K, Phillips A, Thomas N, Cooper D, et al.: The human gene mutation database: 2008 update. Genome Med 2009, 1:13.
2 Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA [http://evs.gs.washington.edu/EVS/].

Why VEST?


All somatic mutations
Functional somatic mutations
Driver somatic mutations

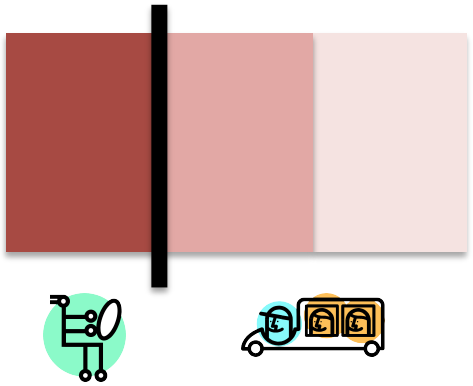


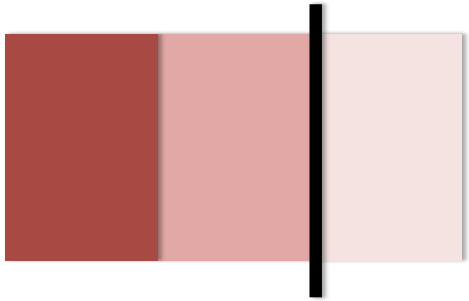


CHASM

VEST

mutationassessor.org
functional impact of protein mutations





S I F T

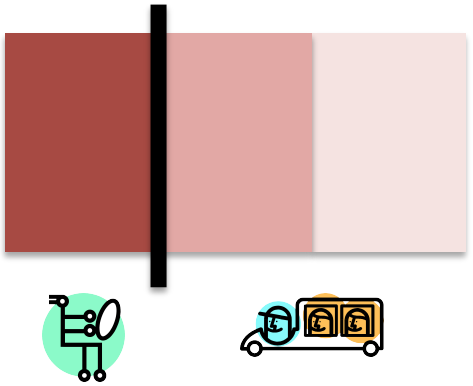
Functional (Damaging)

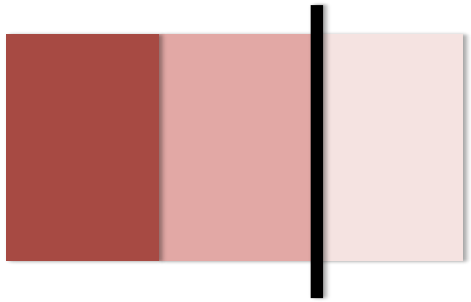
Benign

PolyPhen-2 Submitter

...

Why VEST?



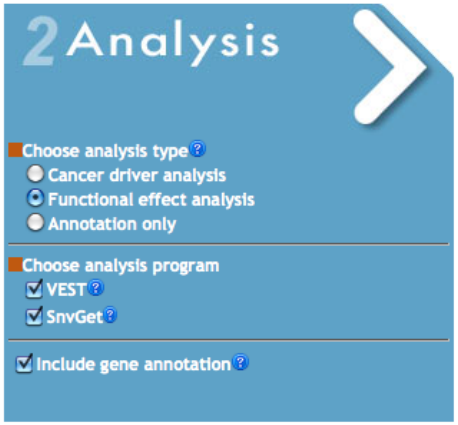


Driver vs. Passenger and Damaging vs. Benign classifiers provide complementary benefits for mutation analysis.

Analysis

Find significantly mutated genes and pathways

- Gene annotations
  - PubMed
  - GeneCards
- Gene scores
  - Combined P-values of VEST or CHASM scores (Stouffer's)



Submit


3 Results
Send the analysis report to:
Include text reports
SUBMIT

Job completed: Email Notification

CRAVAT_DoNotReply@karchinlab.org 6:56 PM (0 minutes ago)

to me

Your CRAVAT analysis has been completed in 10 seconds.

Please download your CRAVAT analysis result at

http://www.cravat.us/results/hcarte10_20120905_185632/hcarte10_20120905_185632.zip

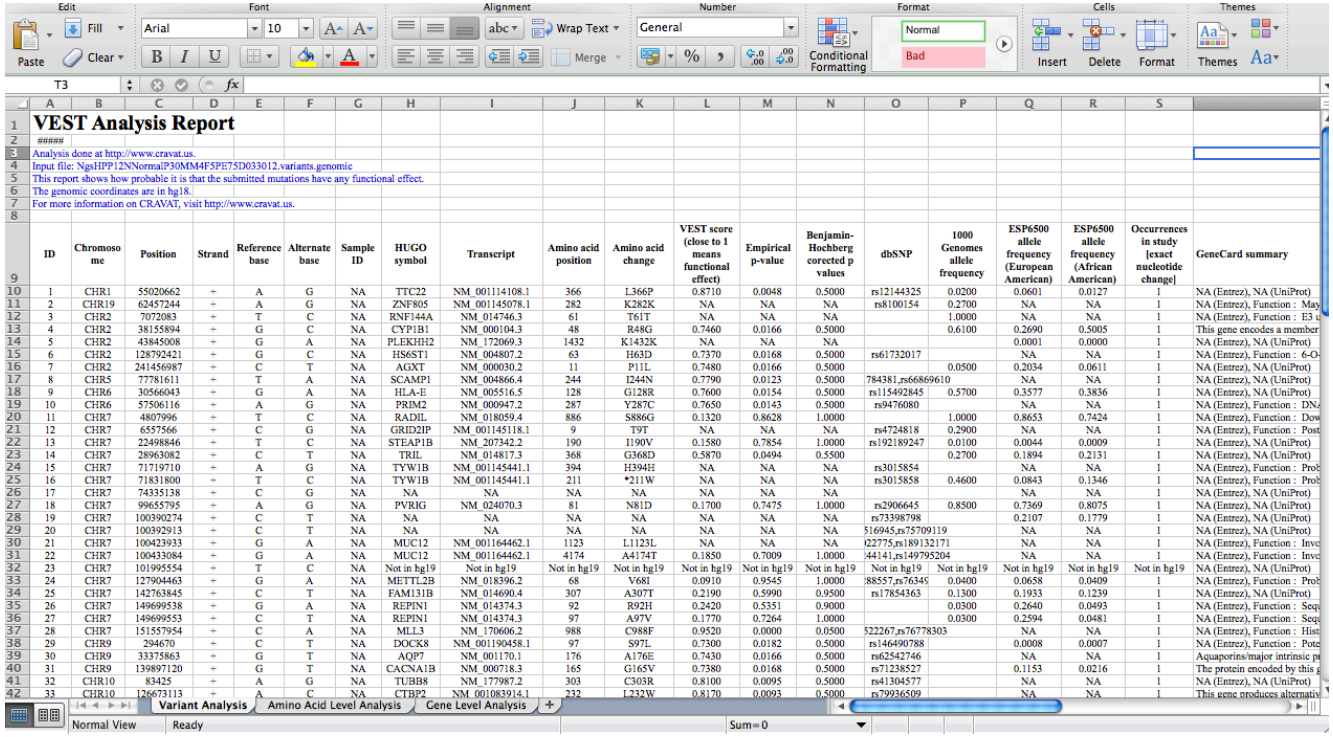
Uncompressing the file will create the directory hcarte10_20120905_185632, which will contain all the analysis result files.

XLS format file can be opened by Excel or any other spreadsheet program, and TSV format files are tab-separated-values text files.

To do more CRAVAT analysis or for more information on CRAVAT, please visit http://www.cravat.us.

CRAVAT Team

CRAVAT Results



VEST Analysis Report

#####

Analysis done at http://www.cravat.us.

Input file: NgsHPP12NNormalP30MM4F5PE75D033012.variants.genomic

This report shows how probable it is that the submitted mutations have any functional effect.

The genomic coordinates are in hg18.

For more information on CRAVAT, visit http://www.cravat.us.

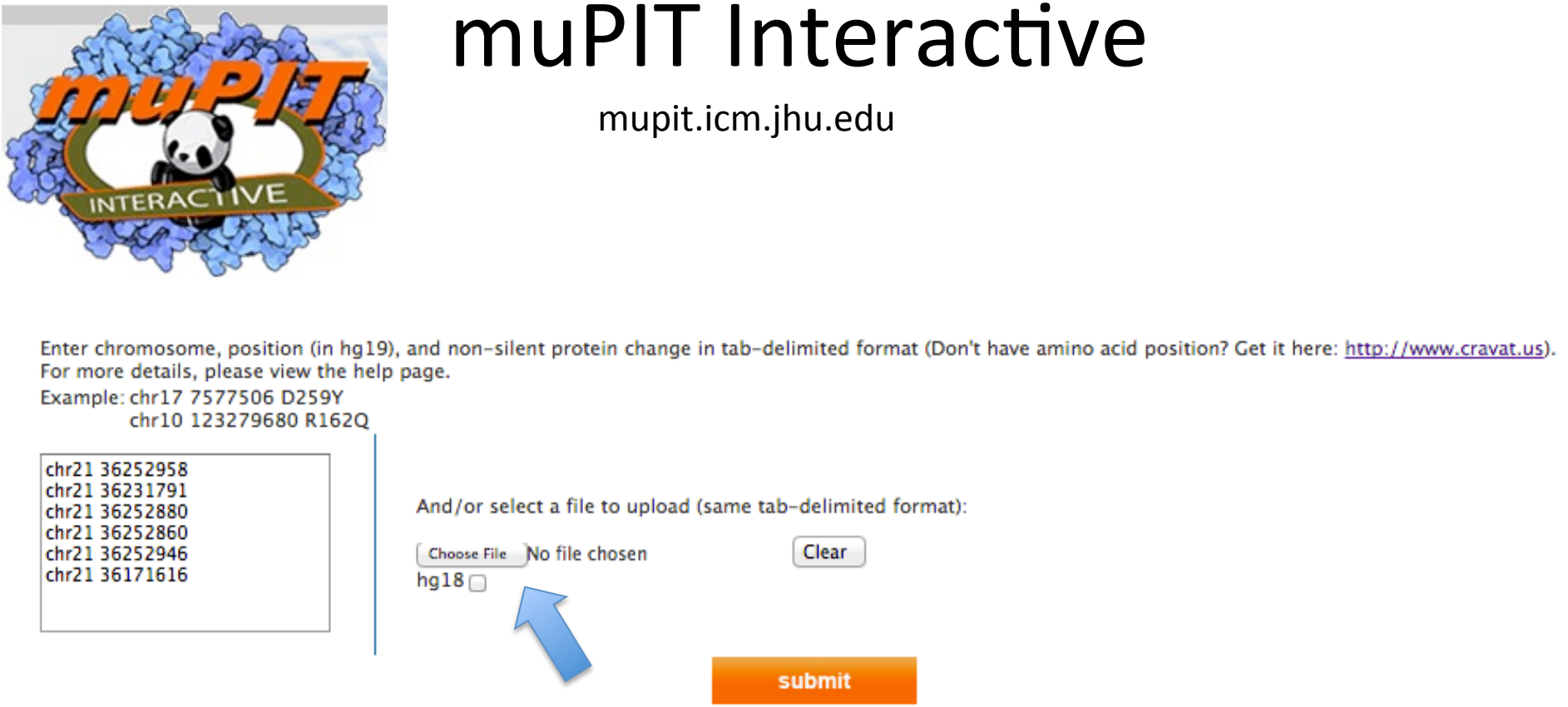
ID	Chromosome	Position	Strand	Reference base	Alternate base	Sample ID	HUGO symbol	Transcript	Amino acid position	Amino acid change	VEST score (close to 1 means functional effect)	Empirical p-value	Benjamin-Hochberg corected p values	dbSNP	1000 Genomes allele frequency	ESP6500 allele frequency (European American)	ESP6500 allele frequency (African American)	Occurrences in study [exact nucleotide change]	GeneCard summary
1	CHR1	55020662	+	A	G	NA	TTC22	NM_001114108.1	366	L366P	0.8710	0.0048	0.5000	rs12144325	0.0200	0.0601	0.0127	1	NA (Entrez), NA (UniProt)
2	CHR19	62457244	+	A	G	NA	ZNF805	NM_001145078.1	282	K282K	NA	NA	NA	rs8100154	0.2700	NA	NA	1	NA (Entrez), Function : May
3	CHR2	7072083	+	T	C	NA	RNF144A	NM_014746.3	61	T61T	NA	NA	NA		1.0000	NA	NA	1	NA (Entrez), Function : E3
4	CHR2	38155894	+	G	C	NA	CYP1B1	NM_000104.3	48	R48G	0.7460	0.0166	0.5000		0.6100	0.2690	0.5005	1	This gene encodes a member
5	CHR2	43845008	+	G	A	NA	PLEKHH2	NM_172069.3	1432	K1432K	NA	NA	NA			0.0001	0.0000	1	NA (Entrez), NA (UniProt)
6	CHR2	128792421	+	G	C	NA	HS6ST1	NM_004807.2	63	H63D	0.7370	0.0168	0.5000	rs61732017		NA	NA	1	NA (Entrez), Function : 6-O-
7	CHR2	241456987	+	C	T	NA	AGXT	NM_000030.2	11	P11L	0.7480	0.0166	0.5000		0.0500	0.2034	0.0611	1	NA (Entrez), NA (UniProt)
8	CHR5	77781611	+	T	A	NA	SCAMP1	NM_004866.4	244	I244N	0.7790	0.0123	0.5000	784381,rs66869610		NA	NA	1	NA (Entrez), NA (UniProt)
9	CHR6	30566043	+	G	A	NA	HLA-E	NM_005516.5	128	G128R	0.7600	0.0154	0.5000	rs115492845	0.5700	0.3577	0.3836	1	NA (Entrez), NA (UniProt)
10	CHR6	57506116	+	A	G	NA	PRIM2	NM_000947.2	287	Y287C	0.7650	0.0143	0.5000	rs9476080		NA	NA	1	NA (Entrez), Function : DNA
11	CHR7	4807996	+	T	C	NA	RADIL	NM_018059.4	886	S886G	0.1320	0.8628	1.0000		1.0000	0.8653	0.7424	1	NA (Entrez), Function : Dow
12	CHR7	6557566	+	C	G	NA	GRID2IP	NM_001145118.1	9	T9T	NA	NA	NA	rs4724818	0.2900	NA	NA	1	NA (Entrez), Function : Post
13	CHR7	22498846	+	T	C	NA	STEAP1B	NM_207342.2	190	I190V	0.1580	0.7854	1.0000	rs192189247	0.0100	0.0044	0.0009	1	NA (Entrez), NA (UniProt)
14	CHR7	28963082	+	C	T	NA	TRIL	NM_014817.3	368	G368D	0.5870	0.0494	0.5500		0.2700	0.1894	0.2131	1	NA (Entrez), NA (UniProt)
15	CHR7	71719710	+	A	G	NA	TYW1B	NM_001145441.1	394	H394H	NA	NA	NA	rs3015854		NA	NA	1	NA (Entrez), Function : Prob
16	CHR7	71831800	+	T	C	NA	TYW1B	NM_001145441.1	211	*211W	NA	NA	NA	rs3015858	0.4600	0.0843	0.1346	1	NA (Entrez), Function : Prob
17	CHR7	74335138	+	C	G	NA	NA	NA	NA	NA	NA	NA	NA			NA	NA	1	NA (Entrez), NA (UniProt)
18	CHR7	99655795	+	A	G	NA	PVRIG	NM_024070.3	81	N81D	0.1700	0.7475	1.0000	rs2906645	0.8500	0.7369	0.8075	1	NA (Entrez), NA (UniProt)
19	CHR7	100390274	+	C	T	NA	NA	NA	NA	NA	NA	NA	NA	rs73398798		0.2107	0.1779	1	NA (Entrez), NA (UniProt)
20	CHR7	100392913	+	C	T	NA	NA	NA	NA	NA	NA	NA	NA	516945,rs75709119		NA	NA	1	NA (Entrez), NA (UniProt)
21	CHR7	100423933	+	G	A	NA	MUC12	NM_001164462.1	1123	L1123L	NA	NA	NA	22775,rs189132171		NA	NA	1	NA (Entrez), Function : Invo
22	CHR7	100433084	+	G	A	NA	MUC12	NM_001164462.1	4174	A4174T	0.1850	0.7009	1.0000	44141,rs149795204		NA	NA	1	NA (Entrez), Function : Invo
23	CHR7	101995554	+	T	C	NA	Not in hg19	Not in hg19	Not in hg19	Not in hg19	Not in hg19	Not in hg19	Not in hg19	Not in hg19	Not in hg19	Not in hg19	Not in hg19	Not in hg19	NA (Entrez), NA (UniProt)
24	CHR7	127904463	+	G	A	NA	METTL2B	NM_018396.2	68	V68I	0.0910	0.9545	1.0000	88557,rs76345	0.0400	0.0658	0.0409	1	NA (Entrez), Function : Prob
25	CHR7	142763845	+	C	T	NA	FAM131B	NM_014690.4	307	A307T	0.2190	0.5990	0.9500	rs17854363	0.1300	0.1933	0.1239	1	NA (Entrez), NA (UniProt)
26	CHR7	149699538	+	G	A	NA	REPIN1	NM_014374.3	92	R92H	0.2420	0.5351	0.9000		0.0300	0.2640	0.0493	1	NA (Entrez), Function : Seq
27	CHR7	149699553	+	C	T	NA	REPIN1	NM_014374.3	97	A97V	0.1770	0.7264	1.0000		0.0300	0.2594	0.0481	1	NA (Entrez), Function : Seq
28	CHR7	151557954	+	C	A	NA	MLL3	NM_170606.2	988	C988F	0.9520	0.0000	0.0500	522267,rs76778303		NA	NA	1	NA (Entrez), Function : Hist
29	CHR9	294670	+	C	T	NA	DOCK8	NM_001190458.1	97	S97L	0.7300	0.0182	0.5000	rs146490788		0.0008	0.0007	1	NA (Entrez), Function : Pote
30	CHR9	33375863	+	G	T	NA	AQP7	NM_001170.1	176	A176E	0.7430	0.0166	0.5000	rs62542746		NA	NA	1	Aquaporins/major intrinsic p
31	CHR9	139897120	+	G	T	NA	CACNA1B	NM_000718.3	165	G165V	0.7380	0.0168	0.5000	rs71238527		0.1153	0.0216	1	The protein encoded by this g
32	CHR10	83425	+	A	G	NA	TUBB8	NM_177987.2	303	C303R	0.8100	0.0095	0.5000	rs41304577		NA	NA	1	NA (Entrez), NA (UniProt)
33	CHR10	126673113	+	A	C	NA	CTBP2	NM_001083914.1	232	L232W	0.8170	0.0093	0.5000	rs79936509		NA	NA	1	This gene produces alternativ

If MuPIT Interactive input file is requested

- File with mutations formatted for input to MuPIT Interactive

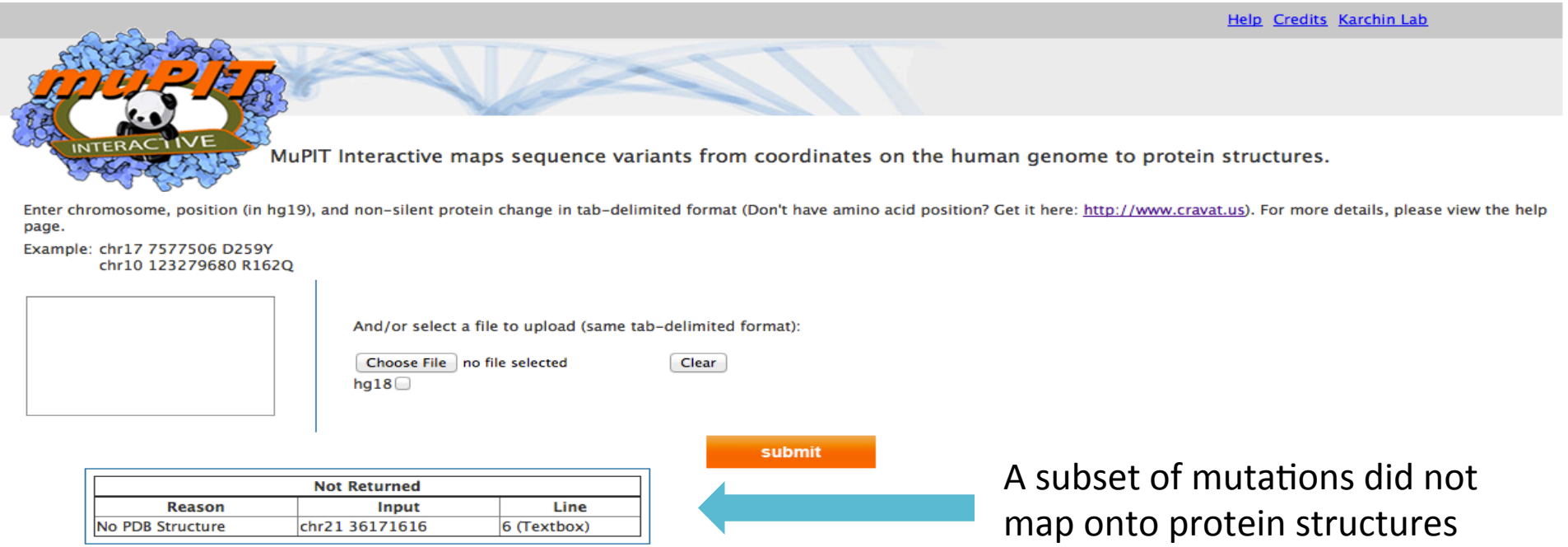
muPIT Interactive

mupit.icm.jhu.edu



Input genomic coordinates of mutations of interest

Table of protein structures onto which your mutations can be mapped



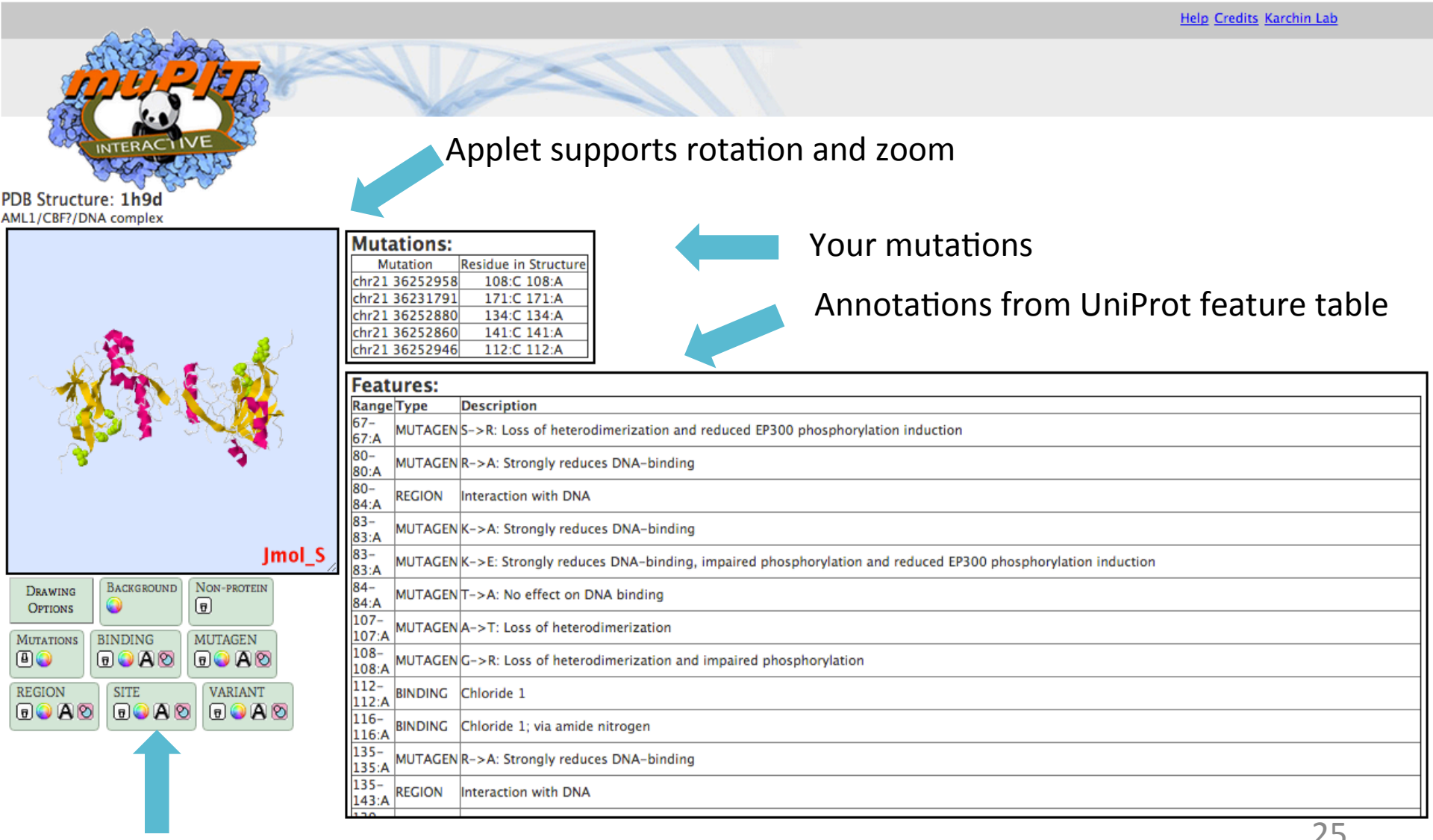
Not Returned		
Reason	**Input**	**Line**
No PDB Structure	chr21 36171616	6 (Textbox)

A subset of mutations did not map onto protein structures

Gene	PDBId	Chromosome_Position	Description	Annotations
RUNX1	1e50	chr21_36252860, chr21_36252946, chr21_36252880, chr21_36231791, chr21_36252958	AML1/CBF? complex	MUTAGEN,BINDING,REGION,VARIANT,SITE
RUNX1	1h9d	chr21_36252860, chr21_36252946, chr21_36252880, chr21_36231791, chr21_36252958	AML1/CBF?/DNA complex	MUTAGEN,REGION,BINDING,VARIANT,SITE
RUNX1	1ljm	chr21_36252860, chr21_36252946, chr21_36252880, chr21_36231791, chr21_36252958	DNA recognition is mediated by conformational transition and by DNA bending	REGION,MUTAGEN,BINDING,VARIANT

Position-specific annotations available for each structure can be used to select most interesting structures.

Selecting a structure brings you to a details page for interactive viewing



Firehose breast cancer mutations

Help Credits Karchin Lab

PDB Structure: **1h9d**
AML1/CBF?/DNA complex

Drawing Options | Background | Non-protein
Mutations | BINDING | MUTAGEN
REGION | SITE | VARIANT

Mutations:

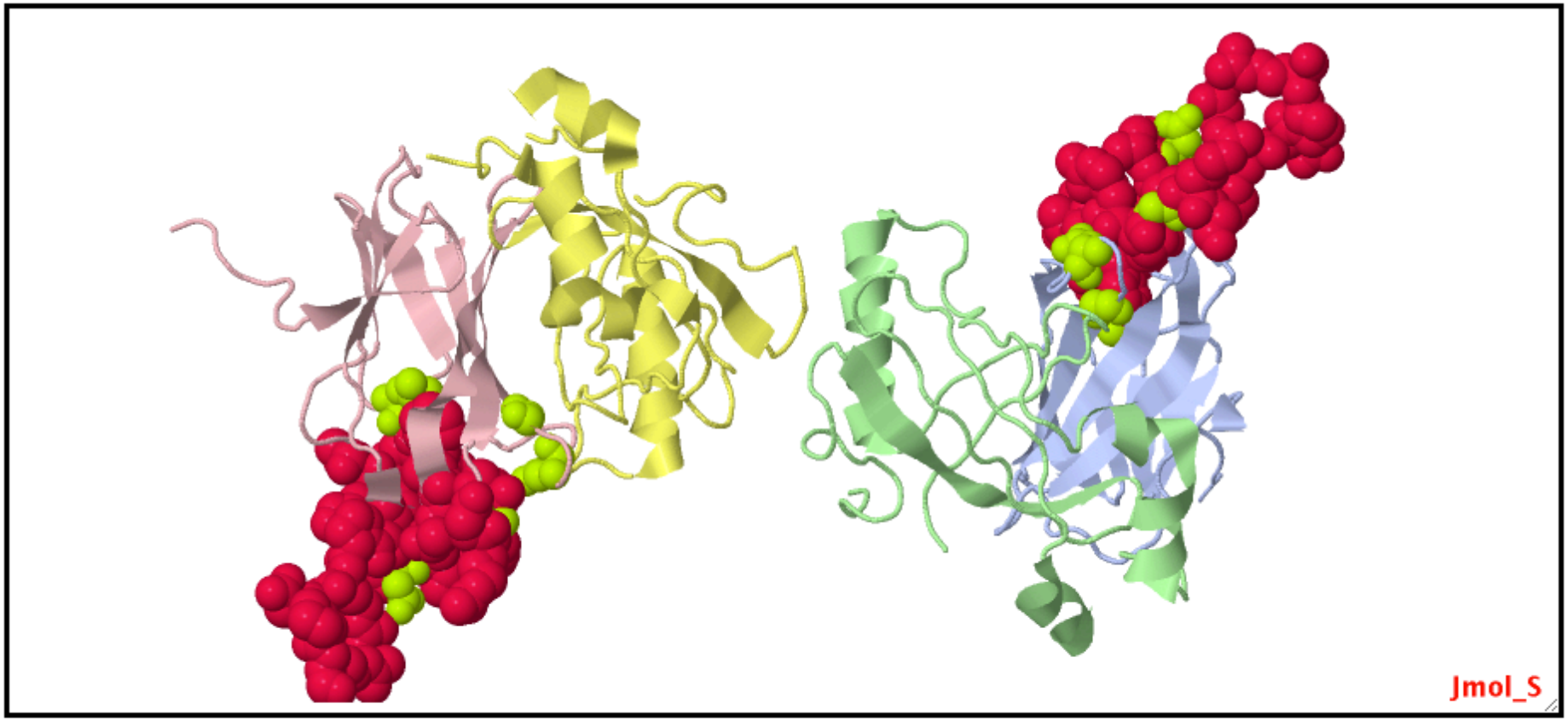
Mutation	Residue in Structure
chr21 36252958	108:C 108:A
chr21 36231791	171:C 171:A
chr21 36252880	134:C 134:A
chr21 36252860	141:C 141:A
chr21 36252946	112:C 112:A

Five mutations cluster together on a structure of RUNX1 (in complex with CBFbeta)

Features:

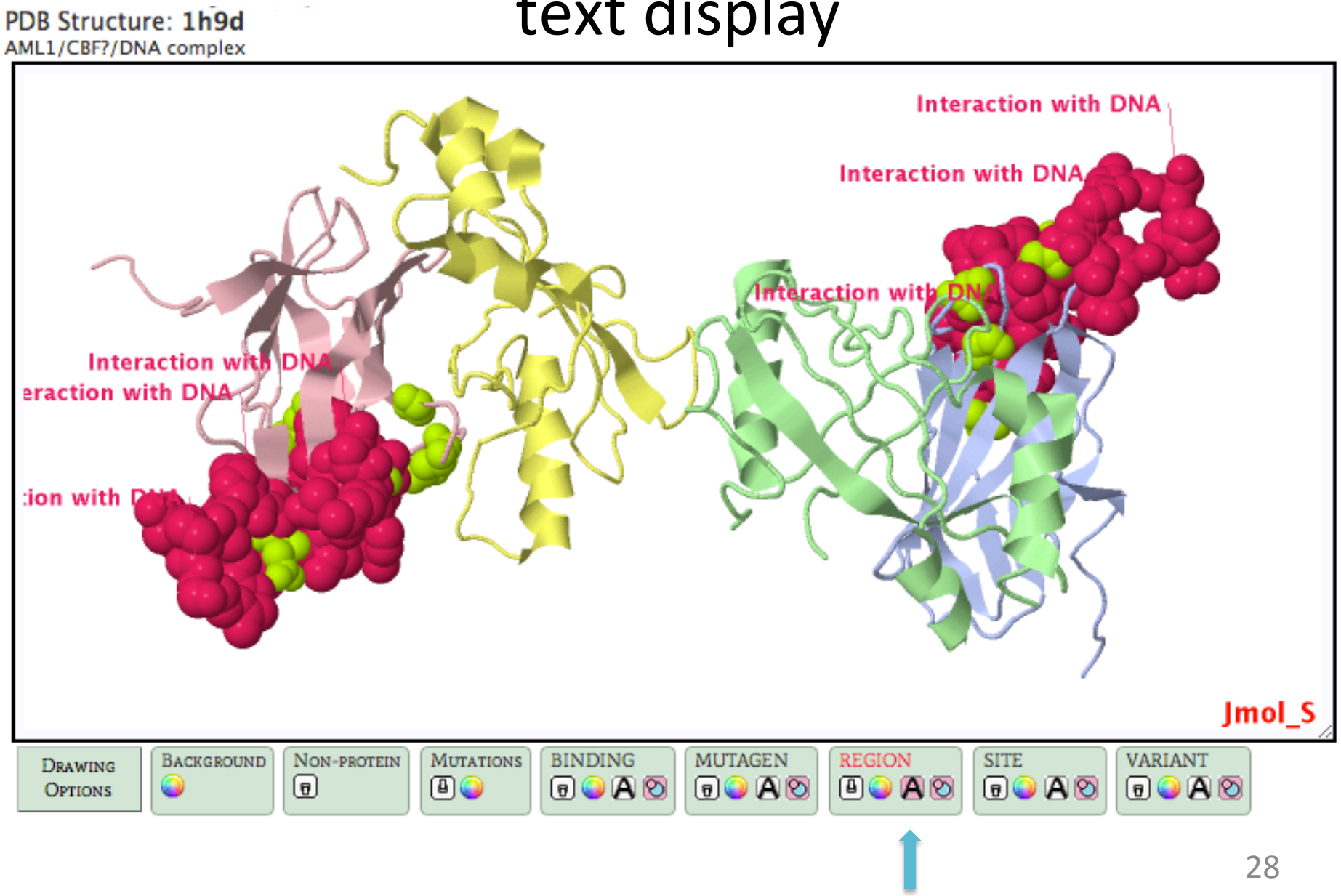
Range	Type	Description
67–67:A	MUTAGEN	S->R: Loss of heterodimerization and reduced EP300 phosphorylation induction
80–80:A	MUTAGEN	R->A: Strongly reduces DNA-binding
80–84:A	REGION	Interaction with DNA
83–83:A	MUTAGEN	K->A: Strongly reduces DNA-binding
83–83:A	MUTAGEN	K->E: Strongly reduces DNA-binding, impaired phosphorylation and reduced EP300 phosphorylation induction
84–84:A	MUTAGEN	T->A: No effect on DNA binding
107–107:A	MUTAGEN	A->T: Loss of heterodimerization
108–108:A	MUTAGEN	G->R: Loss of heterodimerization and impaired phosphorylation
112–112:A	BINDING	Chloride 1
116–116:A	BINDING	Chloride 1; via amide nitrogen
135–135:A	MUTAGEN	R->A: Strongly reduces DNA-binding
135–143:A	REGION	Interaction with DNA

Zoom-in for view of the mutations (green) and a functionally annotated region (red)



Displayed Controls

Functionally annotated region with text display



Oncogene. 2012 Oct 8. doi: 10.1038/onc.2012.328. [Epub ahead of print]

The RUNX family in breast cancer: relationships with estrogen signaling.

Chimge NO, Frenkel B.

Department of Biochemistry and Molecular Biology, Institute for Genetic Medicine, Keck School of Medicine of the University of Southern California, Los Angeles, CA, USA.

Nature. 2012 Jun 20;486(7403):405-9. doi: 10.1038/nature11154.

Sequence analysis of mutations and translocations across breast cancer subtypes.

Banerji S, Cibulskis K, Rangel-Escareno C, Brown KK, Carter SL, Frederick AM, Lawrence MS, Sivachenko AY, Sougnez C, Zou L, Cortes ML, Fernandez-Lopez JC, Peng S, Ardlie KG, Auclair D, Bautista-Piña V, Duke F, Francis J, Jung J, Maffuz-Aziz A, Onofrio RC, Parkin M, Pho NH, Quintanar-Jurado V, Ramos AH, Rebollar-Vega R, Rodriguez-Cuevas S, Romero-Cordoba SL, Schumacher SE, Stransky N, Thompson KM, Uribe-Figueroa L, Baselga J, Beroukhim R, Polyak K, Sgroi DC, Richardson AL, Jimenez-Sanchez G, Lander ES, Gabriel SB, Garraway LA, Golub TR, Melendez-Zajgla J, Toker A, Getz G, Hidalgo-Miranda A, Meyerson M.

The Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA.

Cell Cycle. 2011 Oct 15;10(20):3461-5. doi: 10.4161/cc.10.20.18029. Epub 2011 Oct 15.

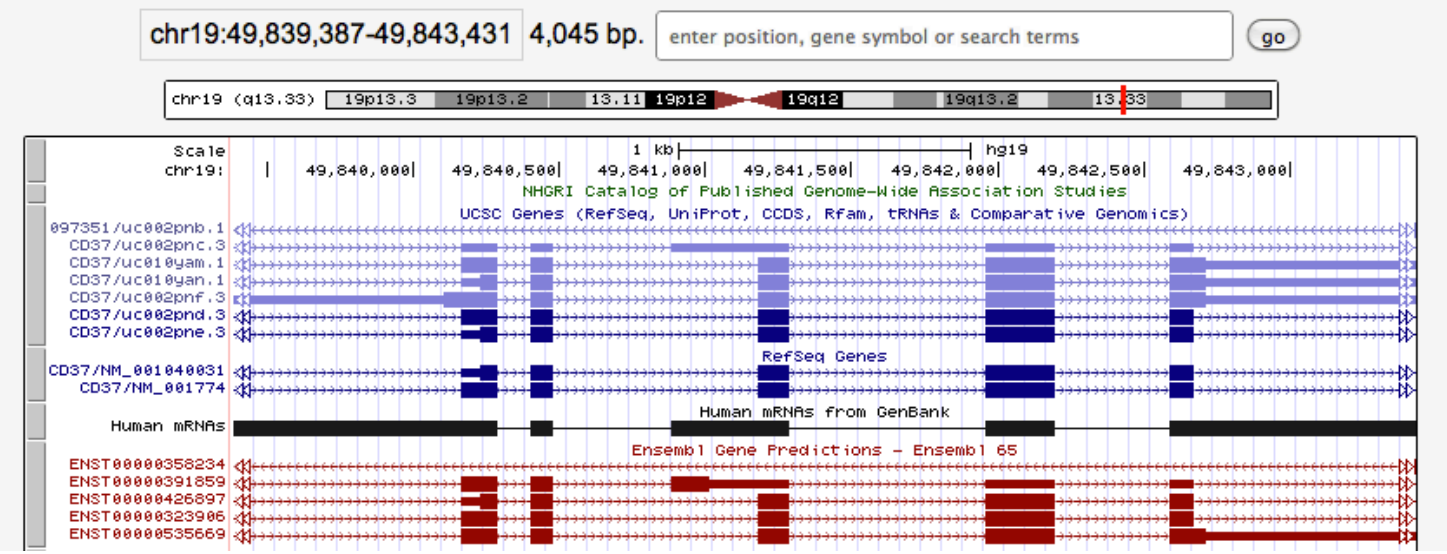
RUNX1 and its understudied role in breast cancer.

Janes KA.

Department of Biomedical Engineering, University of Virginia, Charlottesville, VA, USA. kjanes@virginia.edu

Future Functionality

- Integration of CRAVAT and MuPIT into a single end-to-end service
- GANT - a new statistic to find significantly mutated genes and pathways
- Exhaustive mapping of genomic coordinates onto all protein coding transcripts defined at a locus



External integrations

Coming in 2013

HGMD®
The Human Gene Mutation Database
Cardiff

Cancer Genomics Browser

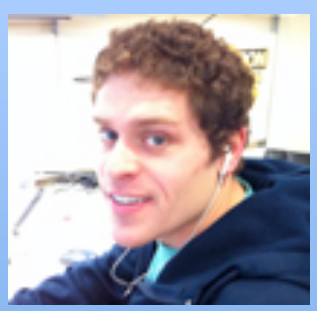
UC SANTA CRUZ

Thank you for listening!

Please visit Poster #75 to talk with us about CRAVAT and MuPIT

Please visit Poster #80 to learn about the Intogen-SM analysis pipeline from Lopez-Bigas Lab (Spain)

Acknowledgments


JOHNS HOPKINS
BIOMEDICAL ENGINEERING

JOHNS HOPKINS
MEDICINE
THE SIDNEY KIMMEL COMPREHENSIVE CANCER CENTER



In Silico
Solutions



CENTER FOR BIOMOLECULAR SCIENCE & ENGINEERING
promoting discovery and invention for human health and well-being

NIH R21 CA135866
NIH R21 CA152432
NIH 3U24CA143858---2S1
NSF CAREER DBI 0845275

INSTITUTE for COMPUTATIONAL MEDICINE

HGMD®
The Human Gene Mutation Database
Cardiff