Personalized Medicine:

Redefining Cancer Treatment

Matt Shaffer · W207 Final Project · 16 August 2017

Memorial Sloan Kettering Cancer Center

Workflow

1. A molecular pathologist selects a list of genetic variations of interest that he/she want to analyze

2. The molecular pathologist searches for evidence in the medical literature that somehow are relevant to the genetic variations of interest

3. Finally, this molecular pathologist spends a huge amount of time analyzing the evidence related to each of the variations to classify them

Goal

Replace step 3 by a machine learning model.

kaggle

Features

1. **Gene**

   (the gene where this genetic mutation is located)

2. **Variation**

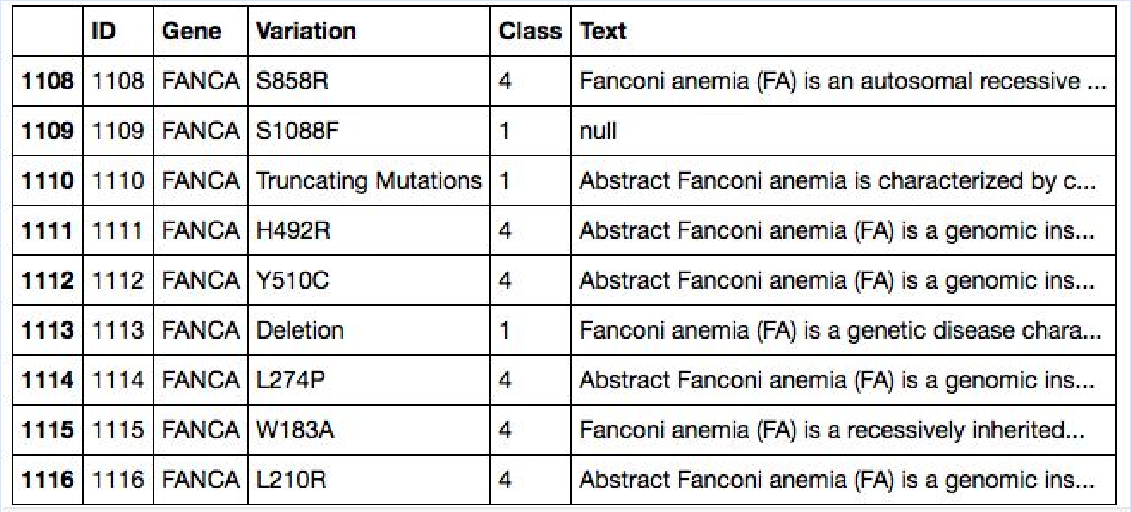
   (the aminoacid change for this mutation)

3. **Class**

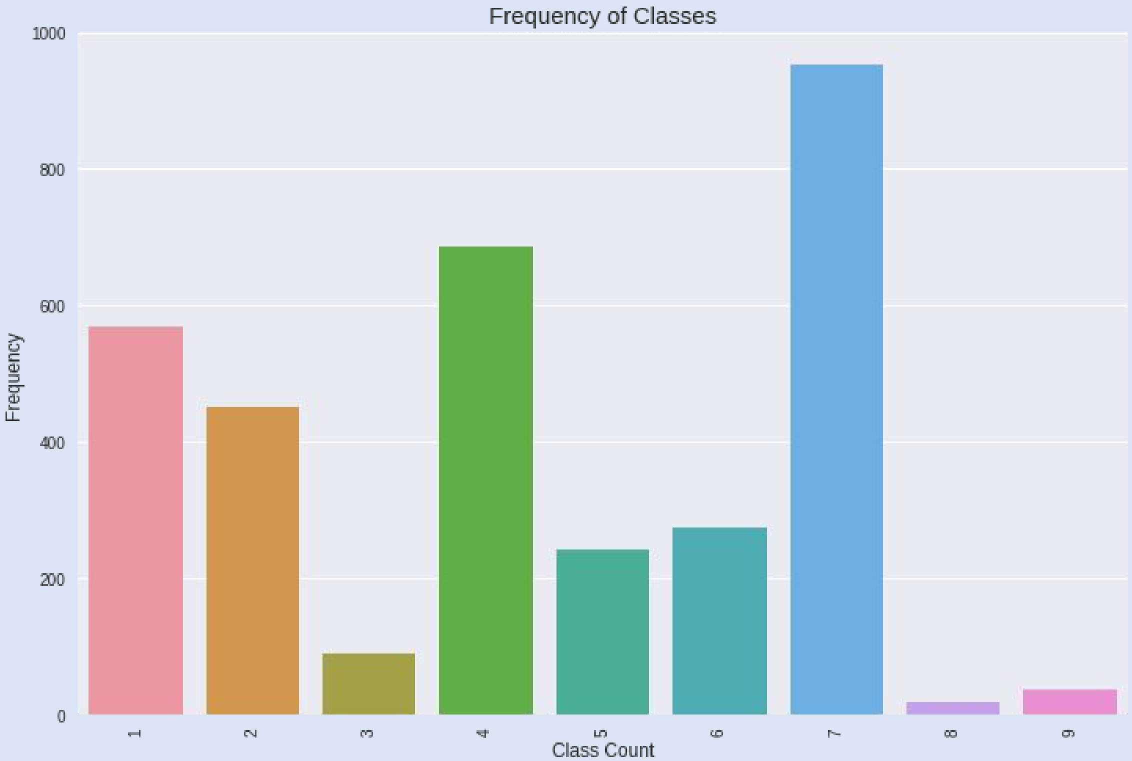
   (1-9 the class this genetic mutation has been classified on)

4. **Text corpus**

   (the clinical evidence used to classify the genetic mutation)



	ID	Gene	Variation	Class	Text
1108	1108	FANCA	S858R	4	Fanconi anemia (FA) is an autosomal recessive ...
1109	1109	FANCA	S1088F	1	null
1110	1110	FANCA	Truncating Mutations	1	Abstract Fanconi anemia is characterized by c...
1111	1111	FANCA	H492R	4	Abstract Fanconi anemia (FA) is a genomic ins...
1112	1112	FANCA	Y510C	4	Abstract Fanconi anemia (FA) is a genomic ins...
1113	1113	FANCA	Deletion	1	Fanconi anemia (FA) is a genetic disease chara...
1114	1114	FANCA	L274P	4	Abstract Fanconi anemia (FA) is a genomic ins...
1115	1115	FANCA	W183A	4	Fanconi anemia (FA) is a recessively inherited...
1116	1116	FANCA	L210R	4	Abstract Fanconi anemia (FA) is a genomic ins...



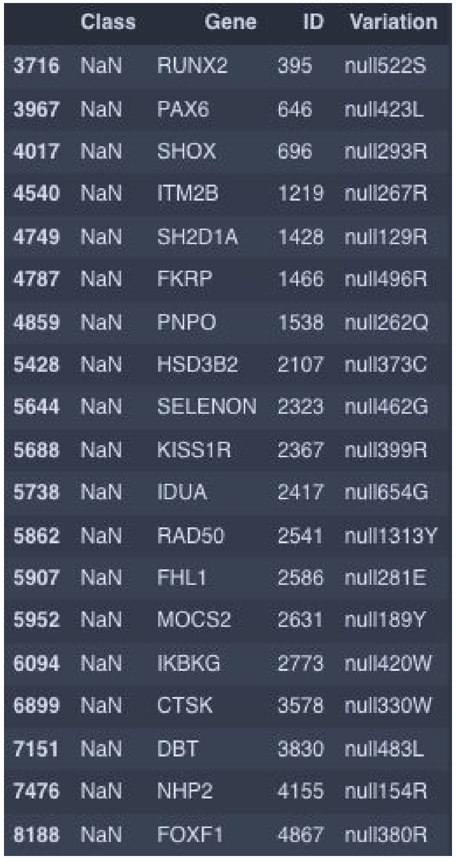
https://www.kaggle.com/sudalairajkumar/simple-exploration-notebook-personalized-medicine

CLASSES

1. Likely Loss-of-function
2. Likely Gain-of-function
3. Neutral
4. Loss-of-function
5. Likely Neutral
6. Inconclusive
7. Gain-of-function
8. Likely Switch-of-function
9. Switch-of-function

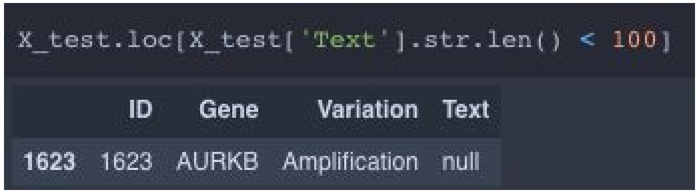
EDA

Inconsistent Data



	Class	Gene	ID	Variation
3716	NaN	RUNX2	395	null522S
3967	NaN	PAX6	646	null423L
4017	NaN	SHOX	696	null293R
4540	NaN	ITM2B	1219	null267R
4749	NaN	SH2D1A	1428	null129R
4787	NaN	FKRP	1466	null496R
4859	NaN	PNPO	1538	null262Q
5428	NaN	HSD3B2	2107	null373C
5644	NaN	SELENON	2323	null462G
5688	NaN	KISS1R	2367	null399R
5738	NaN	IDUA	2417	null654G
5862	NaN	RAD50	2541	null1313Y
5907	NaN	FHL1	2586	null281E
5952	NaN	MOCS2	2631	null189Y
6094	NaN	IKBKG	2773	null420W
6899	NaN	CTSK	3578	null330W
7151	NaN	DBT	3830	null483L
7476	NaN	NHP2	4155	null154R
8188	NaN	FOXF1	4867	null380R

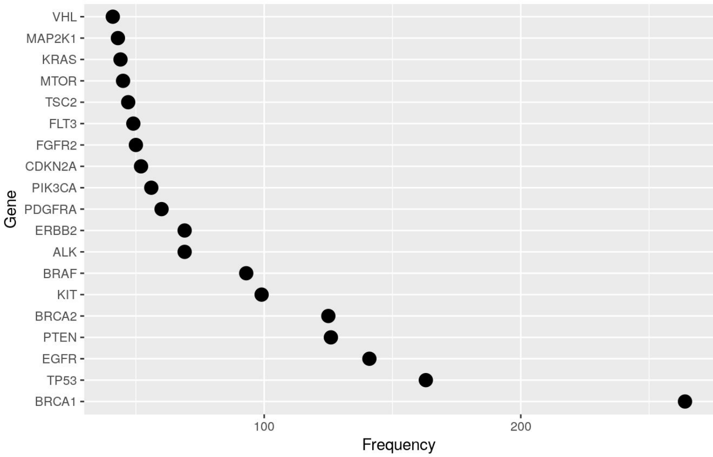
Missing Values



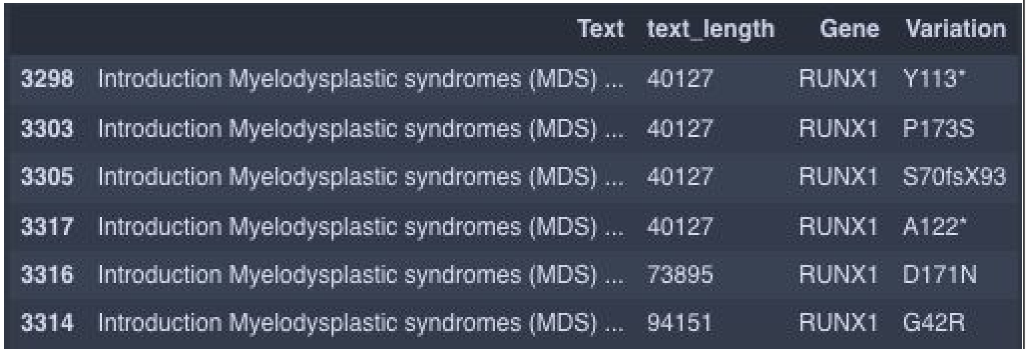
```
X_test.loc[X_test['Text'].str.len() < 100]
```

	ID	Gene	Variation	Text
1623	1623	AURKB	Amplification	null

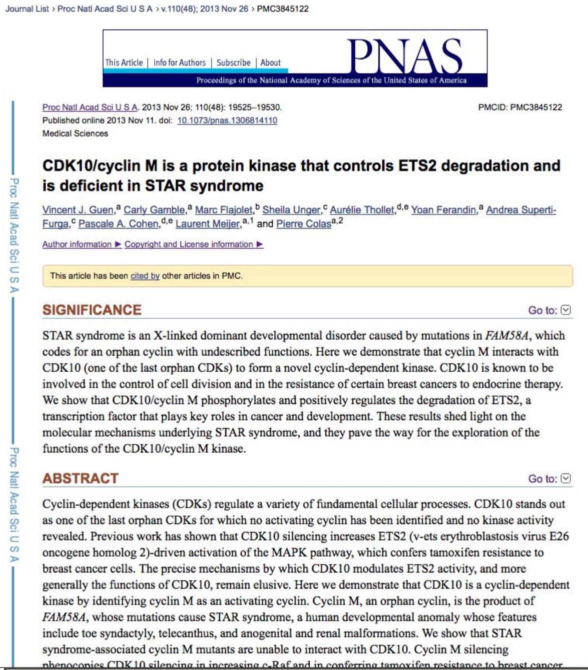
Observations Disproportionately Represented



Shared Text Corpus for Multiple Variations



	Text	text_length	Gene	Variation
3298	Introduction Myelodysplastic syndromes (MDS) ...	40127	RUNX1	Y113*
3303	Introduction Myelodysplastic syndromes (MDS) ...	40127	RUNX1	P173S
3305	Introduction Myelodysplastic syndromes (MDS) ...	40127	RUNX1	S70fsX93
3317	Introduction Myelodysplastic syndromes (MDS) ...	40127	RUNX1	A122*
3316	Introduction Myelodysplastic syndromes (MDS) ...	73895	RUNX1	D171N
3314	Introduction Myelodysplastic syndromes (MDS) ...	94151	RUNX1	G42R



Journal List › Proc Natl Acad Sci U S A › v.110(48); 2013 Nov 26 › PMC3845122

This Article | Info for Authors | Subscribe | About

PNAS

Proceedings of the National Academy of Sciences of the United States of America

Proc Natl Acad Sci U S A

Proc Natl Acad Sci U S A. 2013 Nov 26; 110(48): 19525–19530.
Published online 2013 Nov 11. doi: 10.1073/pnas.1306814110
Medical Sciences

PMCID: PMC3845122

CDK10/cyclin M is a protein kinase that controls ETS2 degradation and is deficient in STAR syndrome

Vincent J. Guen,[a] Carly Gamble,[a] Marc Flajolet,[b] Sheila Unger,[c] Aurélie Thollet,[d,e] Yoan Ferandin,[a] Andrea Superti-Furga,[c] Pascale A. Cohen,[d,e] Laurent Meijer,[a,1] and Pierre Colas[a,2]

Author information ► Copyright and License information ►

This article has been cited by other articles in PMC.

SIGNIFICANCE

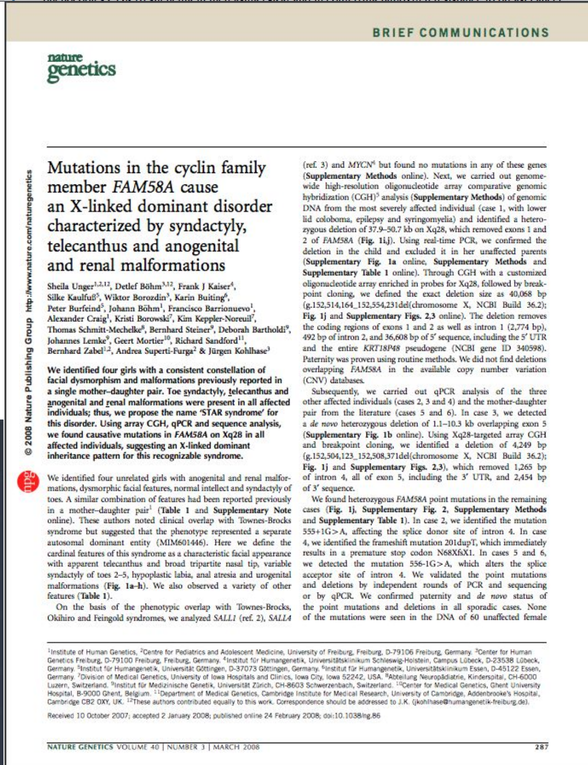
Go to: ⌄

STAR syndrome is an X-linked dominant developmental disorder caused by mutations in *FAM58A*, which codes for an orphan cyclin with undescribed functions. Here we demonstrate that cyclin M interacts with CDK10 (one of the last orphan CDKs) to form a novel cyclin-dependent kinase. CDK10 is known to be involved in the control of cell division and in the resistance of certain breast cancers to endocrine therapy. We show that CDK10/cyclin M phosphorylates and positively regulates the degradation of ETS2, a transcription factor that plays key roles in cancer and development. These results shed light on the molecular mechanisms underlying STAR syndrome, and they pave the way for the exploration of the functions of the CDK10/cyclin M kinase.

ABSTRACT

Go to: ⌄

Cyclin-dependent kinases (CDKs) regulate a variety of fundamental cellular processes. CDK10 stands out as one of the last orphan CDKs for which no activating cyclin has been identified and no kinase activity revealed. Previous work has shown that CDK10 silencing increases ETS2 (v-ets erythroblastosis virus E26 oncogene homolog 2)-driven activation of the MAPK pathway, which confers tamoxifen resistance to breast cancer cells. The precise mechanisms by which CDK10 modulates ETS2 activity, and more generally the functions of CDK10, remain elusive. Here we demonstrate that CDK10 is a cyclin-dependent kinase by identifying cyclin M as an activating cyclin. Cyclin M, an orphan cyclin, is the product of *FAM58A*, whose mutations cause STAR syndrome, a human developmental anomaly whose features include toe syndactyly, telecanthus, and anogenital and renal malformations. We show that STAR syndrome-associated cyclin M mutants are unable to interact with CDK10. Cyclin M silencing phenocopies CDK10 silencing in increasing c-Raf and in conferring tamoxifen resistance to breast cancer



BRIEF COMMUNICATIONS

nature genetics

Mutations in the cyclin family member *FAM58A* cause an X-linked dominant disorder characterized by syndactyly, telecanthus and anogenital and renal malformations

Sheila Unger[1,2,12], Detlef Böhm[3,12], Frank J Kaiser[4], Silke Kaulfuß[5], Wiktor Borozdin[5], Karin Buiting[6], Peter Burfeind[5], Johann Böhm[1], Francisco Barrionuevo[1], Alexander Craig[1], Kristi Borowski[7], Kim Keppler-Noreuil[7], Thomas Schmitt-Mechelke[8], Bernhard Steiner[9], Deborah Bartholdi[9], Johannes Lemke[9], Geert Mortier[10], Richard Sandford[11], Bernhard Zabel[1,2], Andrea Superti-Furga[2] & Jürgen Kohlhase[3]

We identified four girls with a consistent constellation of facial dysmorphism and malformations previously reported in a single mother–daughter pair. Toe syndactyly, telecanthus and anogenital and renal malformations were present in all affected individuals; thus, we propose the name 'STAR syndrome' for this disorder. Using array CGH, qPCR and sequence analysis, we found causative mutations in *FAM58A* on Xq28 in all affected individuals, suggesting an X-linked dominant inheritance pattern for this recognizable syndrome.

We identified four unrelated girls with anogenital and renal malformations, dysmorphic facial features, normal intellect and syndactyly of toes. A similar combination of features had been reported previously in a mother–daughter pair[1] (**Table 1** and **Supplementary Note** online). These authors noted clinical overlap with Townes-Brocks syndrome but suggested that the phenotype represented a separate autosomal dominant entity (MIM601446). Here we define the cardinal features of this syndrome as a characteristic facial appearance with apparent telecanthus and broad tripartite nasal tip, variable syndactyly of toes 2–5, hypoplastic labia, anal atresia and urogenital malformations (**Fig. 1a–h**). We also observed a variety of other features (**Table 1**).

On the basis of the phenotypic overlap with Townes-Brocks, Okihiro and Feingold syndromes, we analyzed *SALL1* (ref. 2), *SALL4* (ref. 3) and *MYCN*[4] but found no mutations in any of these genes (**Supplementary Methods** online). Next, we carried out genome-wide high-resolution oligonucleotide array comparative genomic hybridization (CGH)[5] analysis (**Supplementary Methods**) of genomic DNA from the most severely affected individual (case 1, with lower lid coloboma, epilepsy and syringomyelia) and identified a heterozygous deletion of 37.9–50.7 kb on Xq28, which removed exons 1 and 2 of *FAM58A* (**Fig. 1i,j**). Using real-time PCR, we confirmed the deletion in the child and excluded it in her unaffected parents (**Supplementary Fig. 1a** online, **Supplementary Methods** and **Supplementary Table 1** online). Through CGH with a customized oligonucleotide array enriched in probes for Xq28, followed by breakpoint cloning, we defined the exact deletion size as 40,068 bp (g.152,514,164_152,554,231del(chromosome X, NCBI Build 36.2); **Fig. 1j** and **Supplementary Figs. 2,3** online). The deletion removes the coding regions of exons 1 and 2 as well as intron 1 (2,774 bp), 492 bp of intron 2, and 36,608 bp of 5′ sequence, including the 5′ UTR and the entire *KRT18P48* pseudogene (NCBI gene ID 340598). Paternity was proven using routine methods. We did not find deletions overlapping *FAM58A* in the available copy number variation (CNV) databases.

Subsequently, we carried out qPCR analysis of the three other affected individuals (cases 2, 3 and 4) and the mother-daughter pair from the literature (cases 5 and 6). In case 3, we detected a *de novo* heterozygous deletion of 1.1–10.3 kb overlapping exon 5 (**Supplementary Fig. 1b** online). Using Xq28-targeted array CGH and breakpoint cloning, we identified a deletion of 4,249 bp (g.152,504,123_152,508,371del(chromosome X, NCBI Build 36.2); **Fig. 1j** and **Supplementary Figs. 2,3**), which removed 1,265 bp of intron 4, all of exon 5, including the 3′ UTR, and 2,454 bp of 3′ sequence.

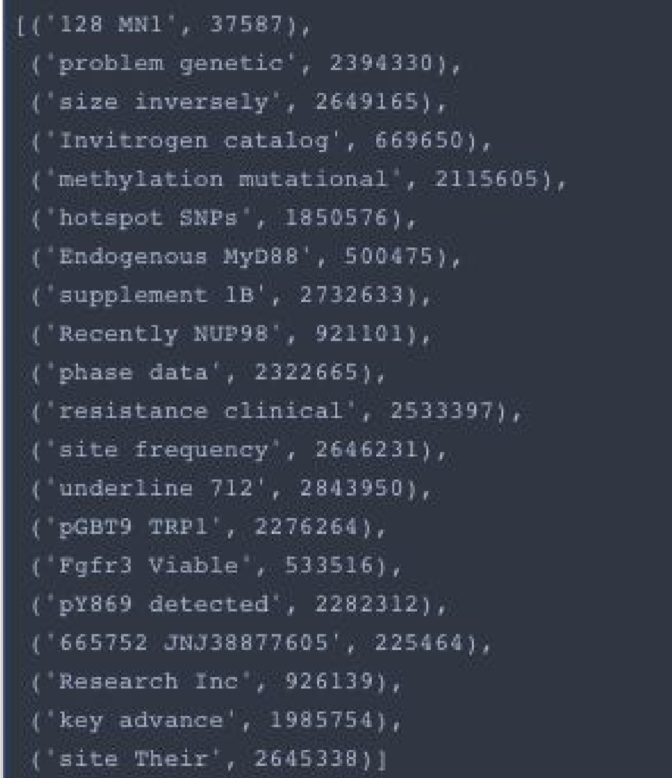
We found heterozygous *FAM58A* point mutations in the remaining cases (**Fig. 1j**, **Supplementary Fig. 2**, **Supplementary Methods** and **Supplementary Table 1**). In case 2, we identified the mutation 555+1G>A, affecting the splice donor site of intron 4. In case 4, we identified the frameshift mutation 201dupT, which immediately results in a premature stop codon N68XfsX1. In cases 5 and 6, we detected the mutation 556-1G>A, which alters the splice acceptor site of intron 4. We validated the point mutations and deletions by independent rounds of PCR and sequencing or by qPCR. We confirmed paternity and *de novo* status of the point mutations and deletions in all sporadic cases. None of the mutations were seen in the DNA of 60 unaffected female

[1] Institute of Human Genetics, [2] Centre for Pediatrics and Adolescent Medicine, University of Freiburg, Freiburg, D-79106 Freiburg, Germany. [3] Center for Human Genetics Freiburg, D-79100 Freiburg, Freiburg, Germany. [4] Institut für Humangenetik, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, D-23538 Lübeck, Germany. [5] Institut für Humangenetik, Universität Göttingen, D-37073 Göttingen, Germany. [6] Institut für Humangenetik, Universitätsklinikum Essen, D-45122 Essen, Germany. [7] Division of Medical Genetics, University of Iowa Hospitals and Clinics, Iowa City, Iowa 52242, USA. [8] Abteilung Neuropädiatrie, Kinderspital, CH-6000 Luzern, Switzerland. [9] Institut für Medizinische Genetik, Universität Zürich, CH-8603 Schwerzenbach, Switzerland. [10] Center for Medical Genetics, Ghent University Hospital, B-9000 Ghent, Belgium. [11] Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Addenbrooke's Hospital, Cambridge CB2 0XY, UK. [12] These authors contributed equally to this work. Correspondence should be addressed to J.K. (jkohlhase@humangenetik-freiburg.de).

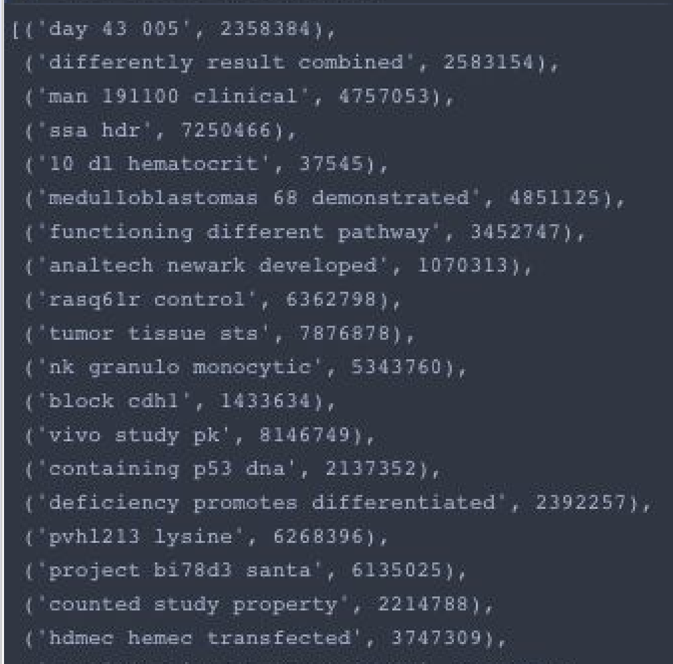
Received 10 October 2007; accepted 2 January 2008; published online 24 February 2008; doi:10.1038/ng.86

TF-IDF

- Bigram- 2690998 tokens
- Trigrams – 33,126,986 tokens
- 10% of Vocabulary

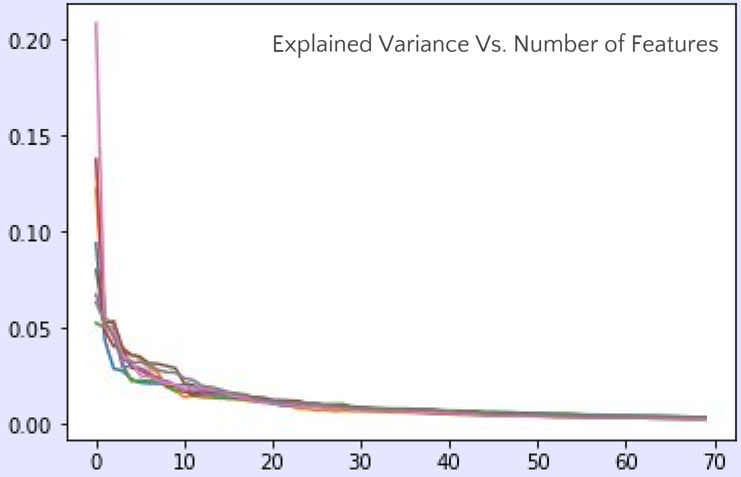


```
[('128 MN1', 37587),
 ('problem genetic', 2394330),
 ('size inversely', 2649165),
 ('Invitrogen catalog', 669650),
 ('methylation mutational', 2115605),
 ('hotspot SNPs', 1850576),
 ('Endogenous MyD88', 500475),
 ('supplement 1B', 2732633),
 ('Recently NUP98', 921101),
 ('phase data', 2322665),
 ('resistance clinical', 2533397),
 ('site frequency', 2646231),
 ('underline 712', 2843950),
 ('pGBT9 TRP1', 2276264),
 ('Fgfr3 Viable', 533516),
 ('pY869 detected', 2282312),
 ('665752 JNJ38877605', 225464),
 ('Research Inc', 926139),
 ('key advance', 1985754),
 ('site Their', 2645338)]
```

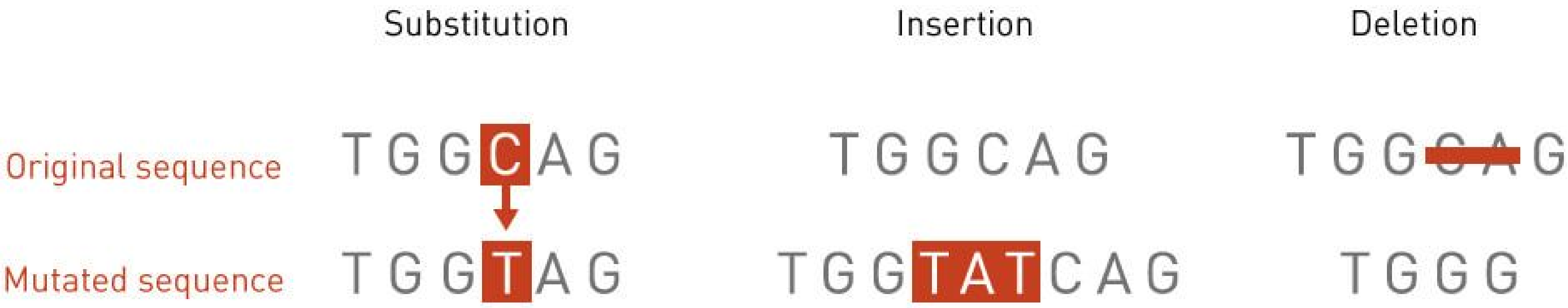


```
[('day 43 005', 2358384),
 ('differently result combined', 2583154),
 ('man 191100 clinical', 4757053),
 ('ssa hdr', 7250466),
 ('10 dl hematocrit', 37545),
 ('medulloblastomas 68 demonstrated', 4851125),
 ('functioning different pathway', 3452747),
 ('analtech newark developed', 1070313),
 ('rasq6lr control', 6362798),
 ('tumor tissue sts', 7876878),
 ('nk granulo monocytic', 5343760),
 ('block cdhl', 1433634),
 ('vivo study pk', 8146749),
 ('containing p53 dna', 2137352),
 ('deficiency promotes differentiated', 2392257),
 ('pvhl213 lysine', 6268396),
 ('project bi78d3 santa', 6135025),
 ('counted study property', 2214788),
 ('hdmec hemec transfected', 3747309),
```

SVD



- 20 Features for dataset scaled to 10% of original
- 200 Final Model



tmVar normalization format:

Substitution:

<Sequence type>|SUB|<wild type>|<mutation position>|<mutant>

e.g., "c.435C>G" --> "c|SUB|C|435|G"

Deletion:

<Sequence type>|DEL|<mutation position>|<mutant>

e.g., "c.104delT" --> "c|DEL|104|T"

e.g., "c.1544-?_2916+?" --> "c|DEL|1544-?_2916+?|"

Insertion:

<Sequence type>|INS|<mutation position>|<mutant>

e.g., "c.104insT" --> "c|INS|104|T"

Insertion+Deletion:

<Sequence type>|INDEL|<mutation position>|<mutant>

e.g., "c.2153_2155delinsTCCTGGTTTA" --> "c|INDEL|2153_2155|TCCTGGTTTA"

Duplication:

<Sequence type>|DUP|<mutation position>|<mutant>|<duplication times>

e.g., "c.1285-1301dup" --> "c|DUP|1285_1301||"

e.g., "c.1978(TATC)(1-2)" --> "c|DUP|1978|TATC|1-2"

Frame shift:

<Sequence type>|FS|<wild type>|<mutation position>|<mutant>|<frame shift position>

e.g., "p.Val35AlafsX25" --> "p|FS|V|35|A|25"

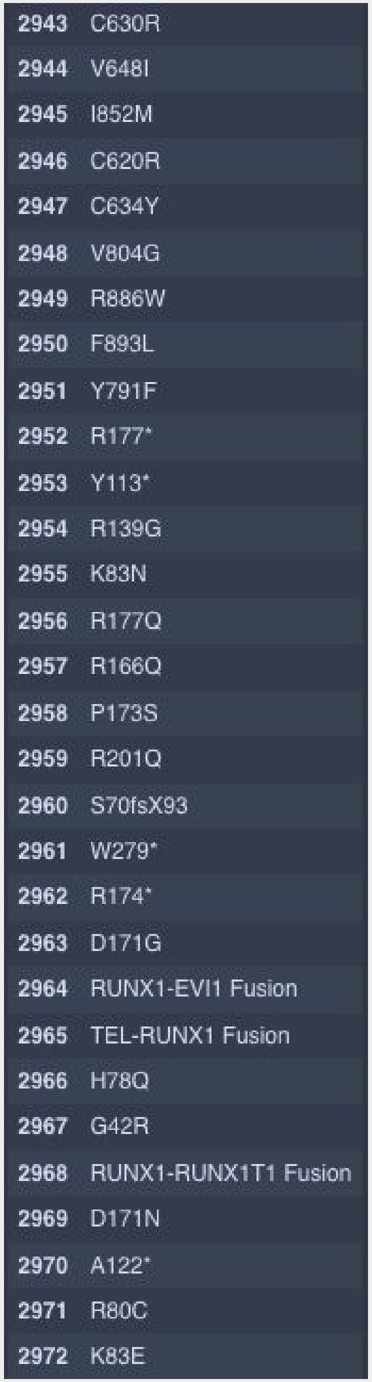
e.g., "p.Ser119fsX" --> "p|FS|S|119||"

<Sequence type>:

c: DNA sequence
r: RNA sequence
g: Genome sequence
p: Protein sequence
m: Mitochondrial sequence

<wild type> / <mutant>:

A,T,C,G: DNA nucleotide
C,I,S,Q,M,N,P,K,D,T,F,A,G,H,L,R,W,V,E,Y,X: Amino acid



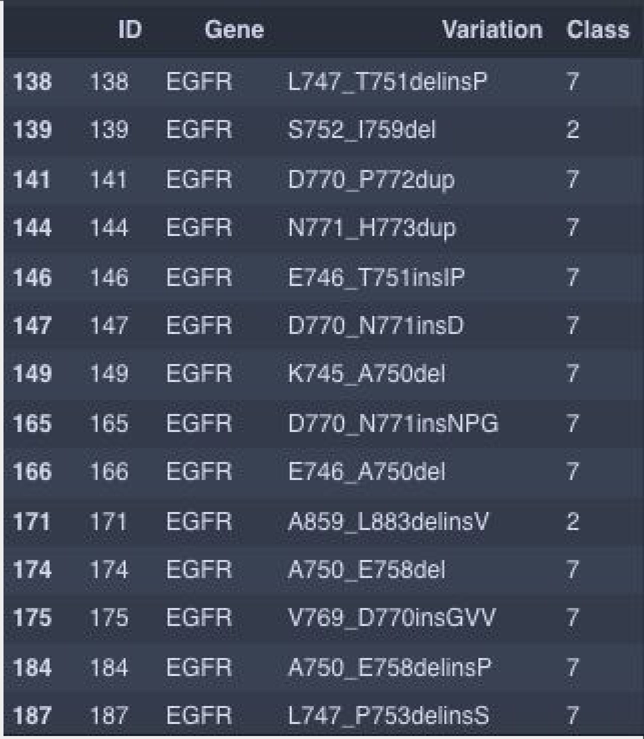
2943	C630R
2944	V648I
2945	I852M
2946	C620R
2947	C634Y
2948	V804G
2949	R886W
2950	F893L
2951	Y791F
2952	R177*
2953	Y113*
2954	R139G
2955	K83N
2956	R177Q
2957	R166Q
2958	P173S
2959	R201Q
2960	S70fsX93
2961	W279*
2962	R174*
2963	D171G
2964	RUNX1-EVI1 Fusion
2965	TEL-RUNX1 Fusion
2966	H78Q
2967	G42R
2968	RUNX1-RUNX1T1 Fusion
2969	D171N
2970	A122*
2971	R80C
2972	K83E

C630R

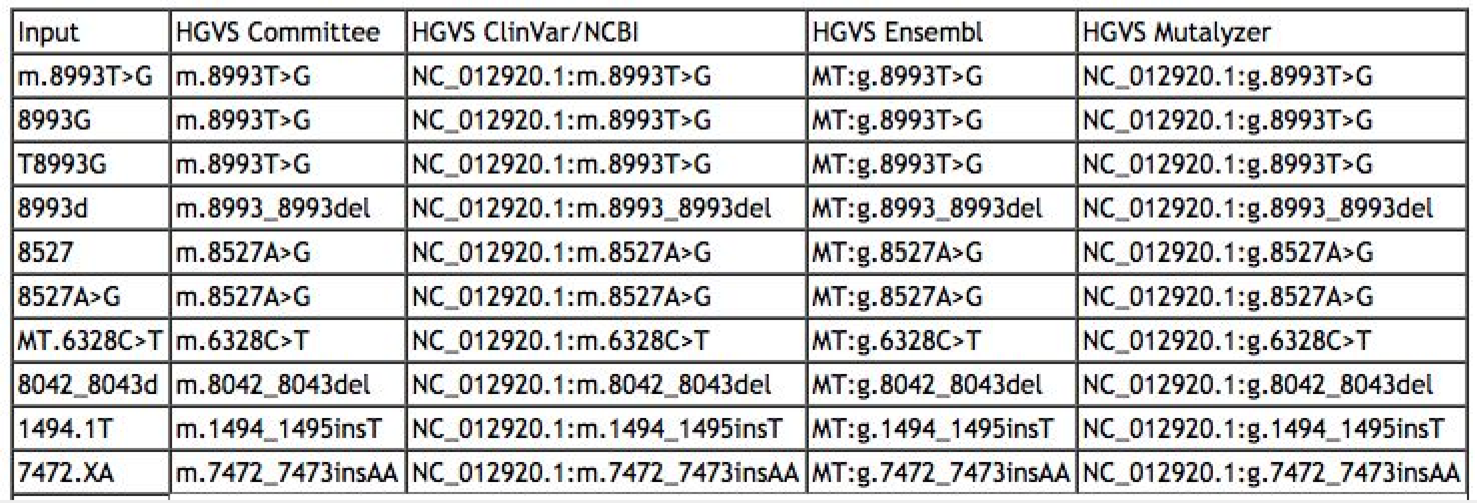
C 630 **R**

Cys 630 **Arg**

Cysteine 630 **Arginine**

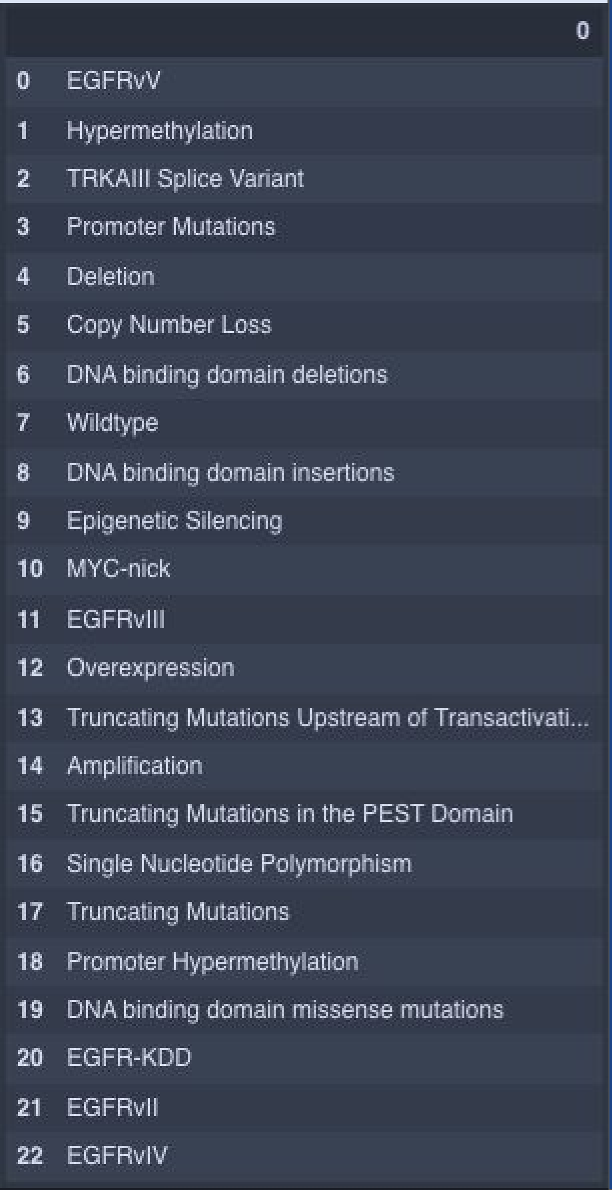


	ID	Gene	Variation	Class
138	138	EGFR	L747_T751delinsP	7
139	139	EGFR	S752_I759del	2
141	141	EGFR	D770_P772dup	7
144	144	EGFR	N771_H773dup	7
146	146	EGFR	E746_T751insIP	7
147	147	EGFR	D770_N771insD	7
149	149	EGFR	K745_A750del	7
165	165	EGFR	D770_N771insNPG	7
166	166	EGFR	E746_A750del	7
171	171	EGFR	A859_L883delinsV	2
174	174	EGFR	A750_E758del	7
175	175	EGFR	V769_D770insGVV	7
184	184	EGFR	A750_E758delinsP	7
187	187	EGFR	L747_P753delinsS	7

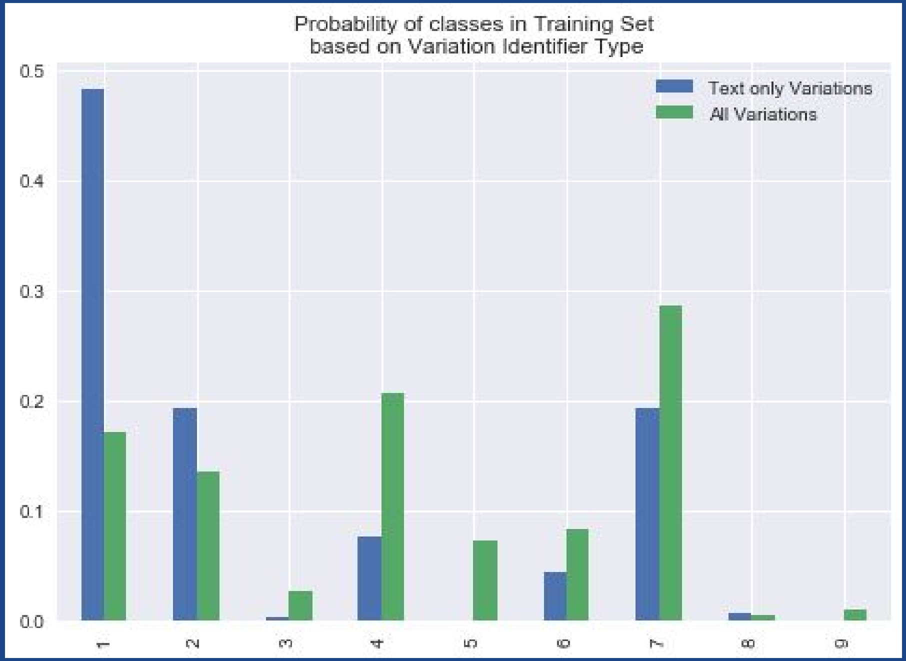


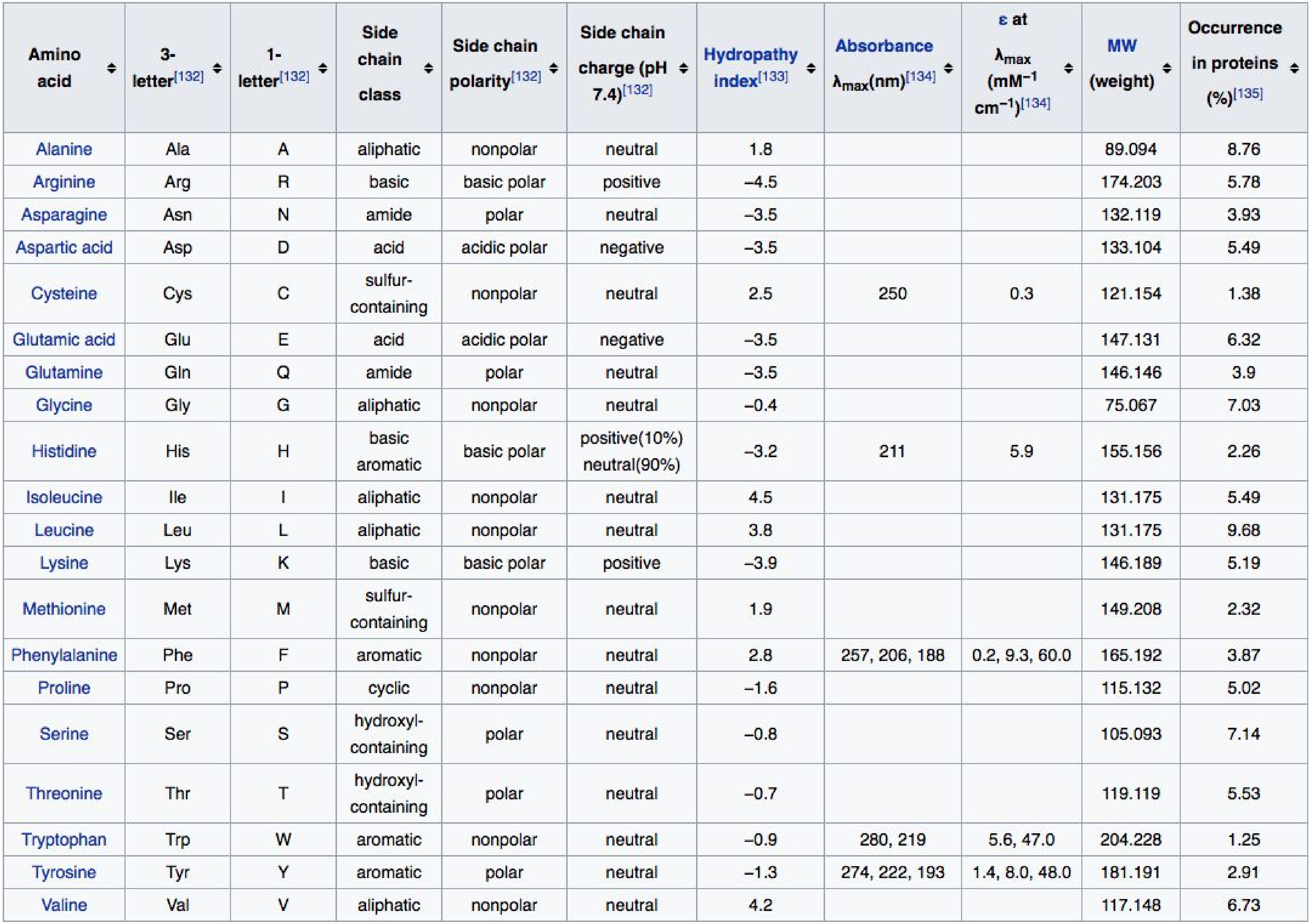
Input	HGVS Committee	HGVS ClinVar/NCBI	HGVS Ensembl	HGVS Mutalyzer
m.8993T>G	m.8993T>G	NC_012920.1:m.8993T>G	MT:g.8993T>G	NC_012920.1:g.8993T>G
8993G	m.8993T>G	NC_012920.1:m.8993T>G	MT:g.8993T>G	NC_012920.1:g.8993T>G
T8993G	m.8993T>G	NC_012920.1:m.8993T>G	MT:g.8993T>G	NC_012920.1:g.8993T>G
8993d	m.8993_8993del	NC_012920.1:m.8993_8993del	MT:g.8993_8993del	NC_012920.1:g.8993_8993del
8527	m.8527A>G	NC_012920.1:m.8527A>G	MT:g.8527A>G	NC_012920.1:g.8527A>G
8527A>G	m.8527A>G	NC_012920.1:m.8527A>G	MT:g.8527A>G	NC_012920.1:g.8527A>G
MT.6328C>T	m.6328C>T	NC_012920.1:m.6328C>T	MT:g.6328C>T	NC_012920.1:g.6328C>T
8042_8043d	m.8042_8043del	NC_012920.1:m.8042_8043del	MT:g.8042_8043del	NC_012920.1:g.8042_8043del
1494.1T	m.1494_1495insT	NC_012920.1:m.1494_1495insT	MT:g.1494_1495insT	NC_012920.1:g.1494_1495insT
7472.XA	m.7472_7473insAA	NC_012920.1:m.7472_7473insAA	MT:g.7472_7473insAA	NC_012920.1:g.7472_7473insAA

Variant Types



	0
0	EGFRvV
1	Hypermethylation
2	TRKAIII Splice Variant
3	Promoter Mutations
4	Deletion
5	Copy Number Loss
6	DNA binding domain deletions
7	Wildtype
8	DNA binding domain insertions
9	Epigenetic Silencing
10	MYC-nick
11	EGFRvIII
12	Overexpression
13	Truncating Mutations Upstream of Transactivati...
14	Amplification
15	Truncating Mutations in the PEST Domain
16	Single Nucleotide Polymorphism
17	Truncating Mutations
18	Promoter Hypermethylation
19	DNA binding domain missense mutations
20	EGFR-KDD
21	EGFRvII
22	EGFRvIV





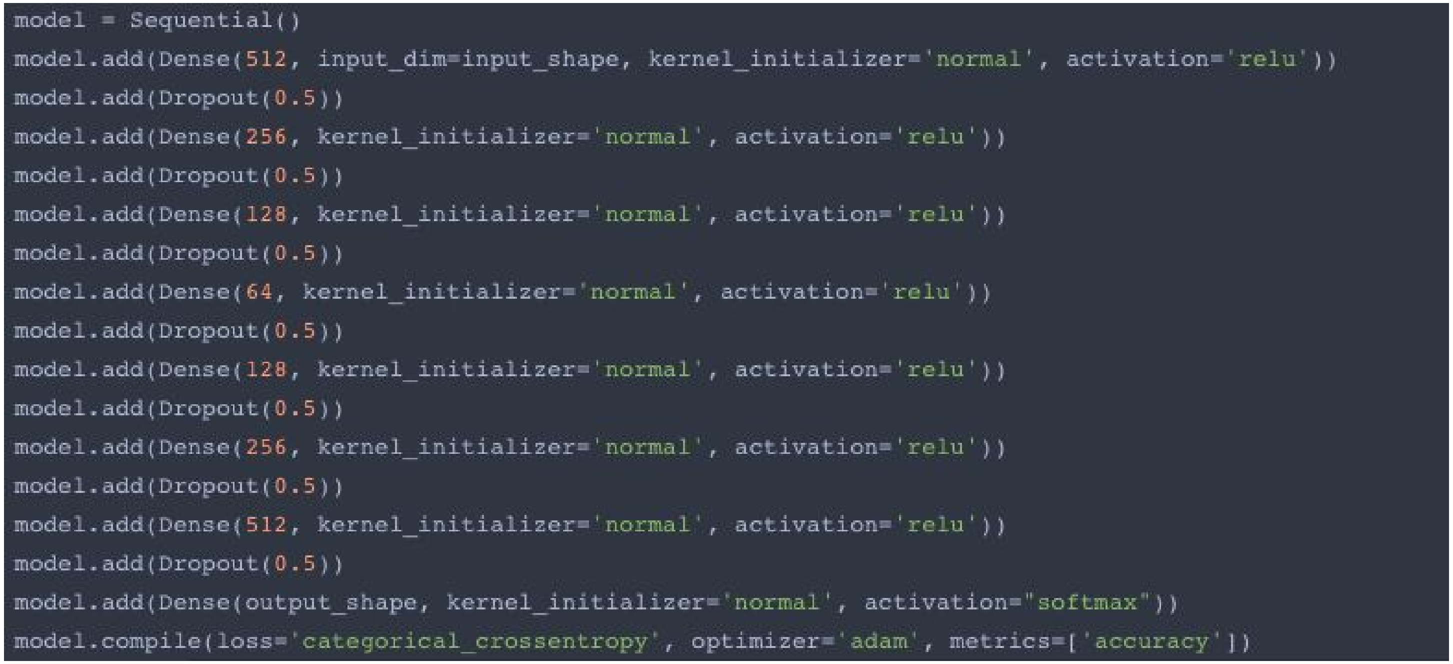
Amino acid	3-letter[132]	1-letter[132]	Side chain class	Side chain polarity[132]	Side chain charge (pH 7.4)[132]	Hydropathy index[133]	Absorbance λ_{max}(nm)[134]	ε at λ_{max} (mM^{-1} cm^{-1})[134]	MW (weight)	Occurrence in proteins (%)[135]
Alanine	Ala	A	aliphatic	nonpolar	neutral	1.8			89.094	8.76
Arginine	Arg	R	basic	basic polar	positive	−4.5			174.203	5.78
Asparagine	Asn	N	amide	polar	neutral	−3.5			132.119	3.93
Aspartic acid	Asp	D	acid	acidic polar	negative	−3.5			133.104	5.49
Cysteine	Cys	C	sulfur-containing	nonpolar	neutral	2.5	250	0.3	121.154	1.38
Glutamic acid	Glu	E	acid	acidic polar	negative	−3.5			147.131	6.32
Glutamine	Gln	Q	amide	polar	neutral	−3.5			146.146	3.9
Glycine	Gly	G	aliphatic	nonpolar	neutral	−0.4			75.067	7.03

Histidine	His	H	basic aromatic	basic polar	positive(10%)
					neutral(90%)
						−3.2	211	5.9	155.156	2.26

Isoleucine	Ile	I	aliphatic	nonpolar	neutral	4.5			131.175	5.49
Leucine	Leu	L	aliphatic	nonpolar	neutral	3.8			131.175	9.68
Lysine	Lys	K	basic	basic polar	positive	−3.9			146.189	5.19
Methionine	Met	M	sulfur-containing	nonpolar	neutral	1.9			149.208	2.32
Phenylalanine	Phe	F	aromatic	nonpolar	neutral	2.8	257, 206, 188	0.2, 9.3, 60.0	165.192	3.87
Proline	Pro	P	cyclic	nonpolar	neutral	−1.6			115.132	5.02
Serine	Ser	S	hydroxyl-containing	polar	neutral	−0.8			105.093	7.14
Threonine	Thr	T	hydroxyl-containing	polar	neutral	−0.7			119.119	5.53
Tryptophan	Trp	W	aromatic	nonpolar	neutral	−0.9	280, 219	5.6, 47.0	204.228	1.25
Tyrosine	Tyr	Y	aromatic	polar	neutral	−1.3	274, 222, 193	1.4, 8.0, 48.0	181.191	2.91
Valine	Val	V	aliphatic	nonpolar	neutral	4.2			117.148	6.73

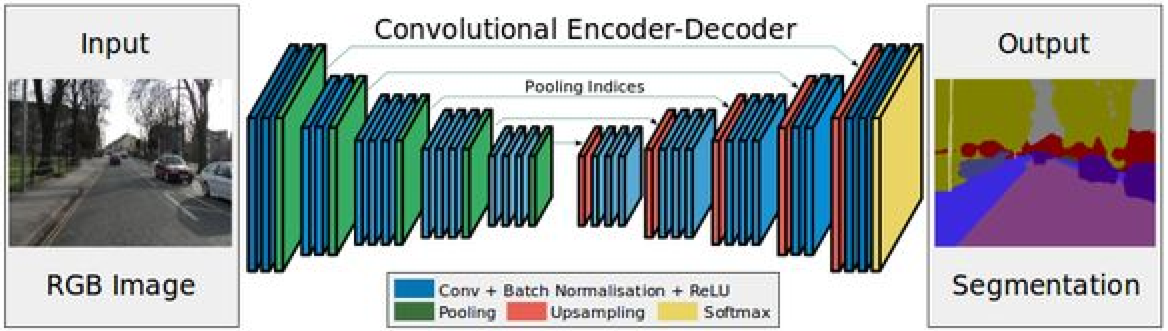
Model Architecture

Dense Network



```
model = Sequential()
model.add(Dense(512, input_dim=input_shape, kernel_initializer='normal', activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(256, kernel_initializer='normal', activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(128, kernel_initializer='normal', activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(64, kernel_initializer='normal', activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(128, kernel_initializer='normal', activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(256, kernel_initializer='normal', activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(512, kernel_initializer='normal', activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(output_shape, kernel_initializer='normal', activation="softmax"))
model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
```

(Similar Idea)



Lessons

Feature engineering takes a long time.

Genetics is complicated

Text mining is hard with limited data

Still to Try

Further exploration with models

Parsing external data sources:

Collect more text data using APIs

Collect more data on genes using APIs

Sampling methods to overcome data imbalance

Current best: 443 of 790
Score 0.82386