Homework 5

made by Ilya German

Part 1.

1.

Choose the gene: FAT1.

Find the gene on IntOGen.

On screenshot you can see the types of cancer which corresponds with mutations.

Method signals per Cancer Type

Show 25 entries

Search:

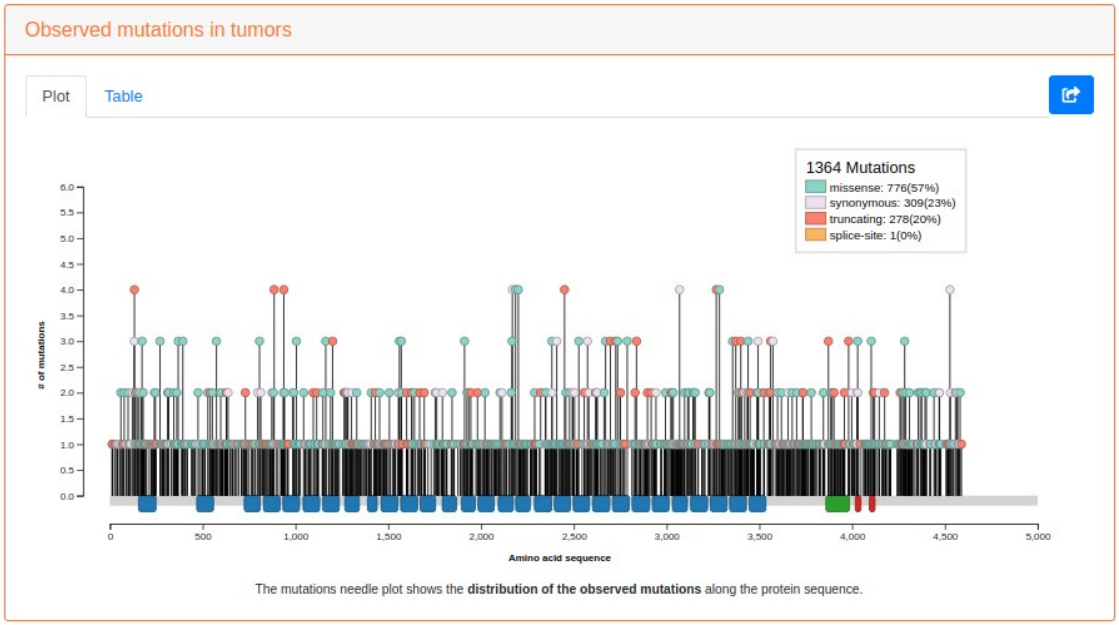
Cancer type	Methods	Samples	Samples (%)
Head and neck squamous cell carcinoma	CBaSE dNdScv MutPanning FML smRegions	131	18.96
Bladder cancer	CBaSE combination dNdScv MutPanning FML	72	8.76
Lung squamous cell carcinoma	CBaSE combination dNdScv MutPanning FML	67	11.47
Endometrial cancer	CBaSE dNdScv	52	8.97
Skin basal cell carcinoma	combination	41	42.71
Cervix squamous cancer	CBaSE dNdScv MutPanning FML	21	6.16
Skin squamous cell carcinoma	CBaSE dNdScv MutPanning	16	32.65
Esophageal cancer	CBaSE dNdScv MutPanning FML	15	1.59
Chronic lymphoblastic leukemia	combination	12	1.65
Ovary cancer	combination	8	1.16
Vulva Cancer	combination	7	31.82
Pancreas adenocarcinoma	combination	5	0.56
Breast adenocarcinoma	smRegions	4	0.15
Prostate adenocarcinoma	combination	3	0.11
Anus cancer	FML	2	11.76
Small intestine cancer neuroendocrine	combination	2	4.44
Hepatic cancer	combination	2	0.12
Medulloblastoma	smRegions	2	0.31
Stomach adenocarcinoma	combination	2	0.28
Uveal melanoma	combination	2	2.5

Showing 1 to 20 of 20 entries

Previous 1 Next

ClustL HotMAPS smRegions Clustered Mutations
CBaSE dNdScv Recurrent Mutations
FML Functional Mutations
MutPanning Tri-nucleotide specific bias
combination Combination

Another screenshot reflects the mutations. There we faces with lots of missence mutations, which refers to point mutations, synonymous mutations, that also refers to point mutations, and truncating mutations that might be both - point or structured mutation. The majority of mutations refers to point-type.



From the next picture we can get, that since 17th page (from 120) the mutations affects only 1 nucleotide, so the majority of truncating mutations are point-type too.

This is only one page with stop_gained consequence, but all of them you can find here.

Observed mutations in tumors

Plot Table

Show 25 entries

Search:

Mutation (GRCh38)	Protein Position	Samples	Consequence
4:186600297:G>A	3902	2	stop_gained
4:186603323:G>A	3735	2	stop_gained
4:186603849:G>C	3559	2	stop_gained
4:186603956:G>A	3524	2	stop_gained
4:186606072:G>A	3450	2	stop_gained
4:186606209:G>C	3404	2	stop_gained
4:186614334:G>C	3029	2	stop_gained
4:186617837:G>A	2917	2	stop_gained
4:186617897:G>A	2897	2	stop_gained
4:186618073:G>C	2838	2	stop_gained
4:186618098:G>A	2830	2	stop_gained
4:186618920:G>A	2556	2	stop_gained
4:186619625:G>A	2321	2	stop_gained
4:186620645:G>A	1981	2	stop_gained
4:186620753:T>A	1945	2	stop_gained
4:186621506:G>A	1694	2	stop_gained
4:186621580:G>C	1669	2	stop_gained
4:186621707:G>A	1627	2	stop_gained
4:186628167:G>C	1599	2	stop_gained
4:186633719:C>A	1430	2	stop_gained
4:186636752:C>A	1269	2	stop_gained
4:186636773:G>A	1262	2	stop_gained
4:186639765:G>C	1200	2	stop_gained
4:186663539:G>A	1114	2	stop_gained
4:186663593:G>A	1096	2	stop_gained

Showing 776 to 800 of 1,196 entries

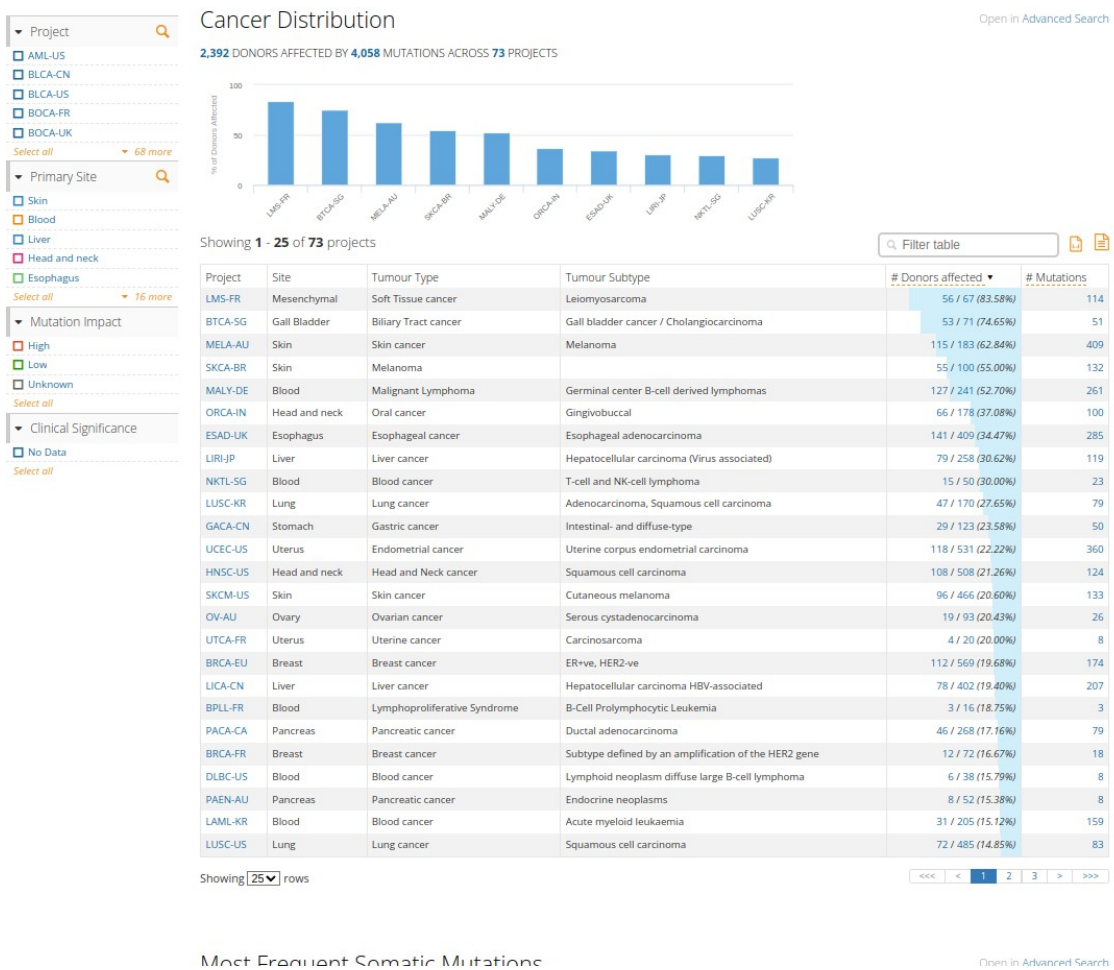
Previous 1 ... 31 32 33 ... 48 Next

2.

Search the ICGC portal for the selected gene ICGC (need to have VPN).

On screenshot we can see the types of cancer, in which the gene FAT1 mutates. (In the column `Tumor Type`.)

Again, only one screen is here, all cancer types on github-page.



Project	Site	Tumour Type	Tumour Subtype	# Donors affected	# Mutations
LMS-FR	Mesenchymal	Soft Tissue cancer	Leiomyosarcoma	56 / 67 (83.58%)	114
BTCA-SG	Gall Bladder	Biliary Tract cancer	Gall bladder cancer / Cholangiocarcinoma	53 / 71 (74.65%)	51
MELA-AU	Skin	Skin cancer	Melanoma	115 / 183 (62.84%)	409
SKCA-BR	Skin	Melanoma		55 / 100 (55.00%)	132
MALY-DE	Blood	Malignant Lymphoma	Germinal center B-cell derived lymphomas	127 / 241 (52.70%)	261
ORCA-IN	Head and neck	Oral cancer	Gingivobuccal	66 / 178 (37.08%)	100
ESAD-UK	Esophagus	Esophageal cancer	Esophageal adenocarcinoma	141 / 409 (34.47%)	285
LIRI-JP	Liver	Liver cancer	Hepatocellular carcinoma (Virus associated)	79 / 258 (30.62%)	119
NKTL-SG	Blood	Blood cancer	T-cell and NK-cell lymphoma	15 / 50 (30.00%)	23
LUSC-KR	Lung	Lung cancer	Adenocarcinoma, Squamous cell carcinoma	47 / 170 (27.65%)	79
GACA-CN	Stomach	Gastric cancer	Intestinal- and diffuse-type	29 / 123 (23.58%)	50
UCEC-US	Uterus	Endometrial cancer	Uterine corpus endometrial carcinoma	118 / 531 (22.22%)	360
HNSC-US	Head and neck	Head and Neck cancer	Squamous cell carcinoma	108 / 508 (21.26%)	124
SKCM-US	Skin	Skin cancer	Cutaneous melanoma	96 / 466 (20.60%)	133
OV-AU	Ovary	Ovarian cancer	Serous cystadenocarcinoma	19 / 93 (20.43%)	26
UTCA-FR	Uterus	Uterine cancer	Carcinosarcoma	4 / 20 (20.00%)	8
BRCA-EU	Breast	Breast cancer	ER+ve, HER2-ve	112 / 569 (19.68%)	174
LICA-CN	Liver	Liver cancer	Hepatocellular carcinoma HBV-associated	78 / 402 (19.40%)	207
BPLL-FR	Blood	Lymphoproliferative Syndrome	B-Cell Prolymphocytic Leukemia	3 / 16 (18.75%)	3
PACA-CA	Pancreas	Pancreatic cancer	Ductal adenocarcinoma	46 / 268 (17.16%)	79
BRCA-FR	Breast	Breast cancer	Subtype defined by an amplification of the HER2 gene	12 / 72 (16.67%)	18
DLBC-US	Blood	Blood cancer	Lymphoid neoplasm diffuse large B-cell lymphoma	6 / 38 (15.79%)	8
PAEN-AU	Pancreas	Pancreatic cancer	Endocrine neoplasms	8 / 52 (15.38%)	8
LAML-KR	Blood	Blood cancer	Acute myeloid leukaemia	31 / 205 (15.12%)	159
LUSC-US	Lung	Lung cancer	Squamous cell carcinoma	72 / 485 (14.85%)	83

2. Part 2.

1.

Choose the genome: DO7777.

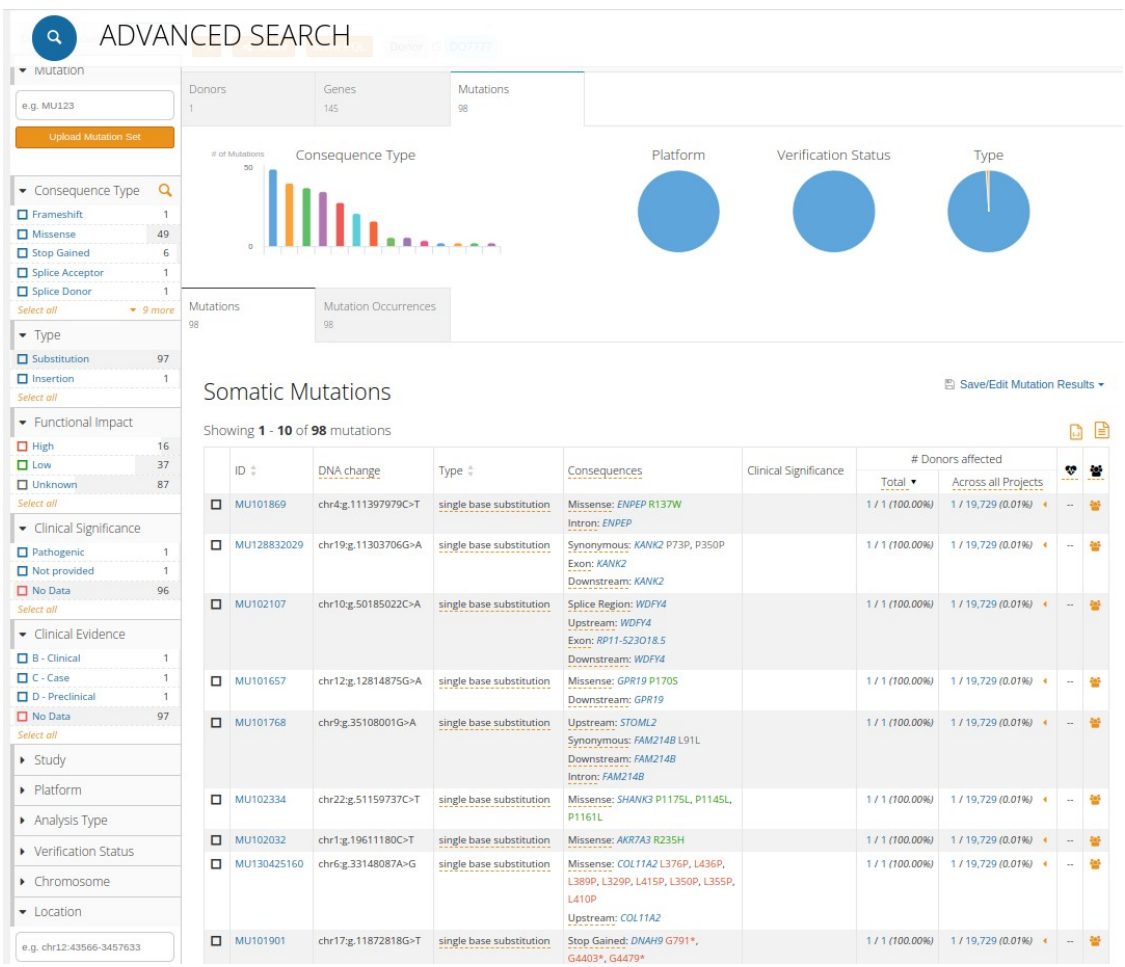
Download it from ICGC. (I will use my own copy of archive, tha placed on github, because it is easier to code on colab with github instead of ICGC, because there we need to pass info-fields for downloading files.)

```
!wget
https://github.com/ilyagerman52/bioinformatics/raw/main/homework5/
simple_somatic_mutation.open.tsv
```

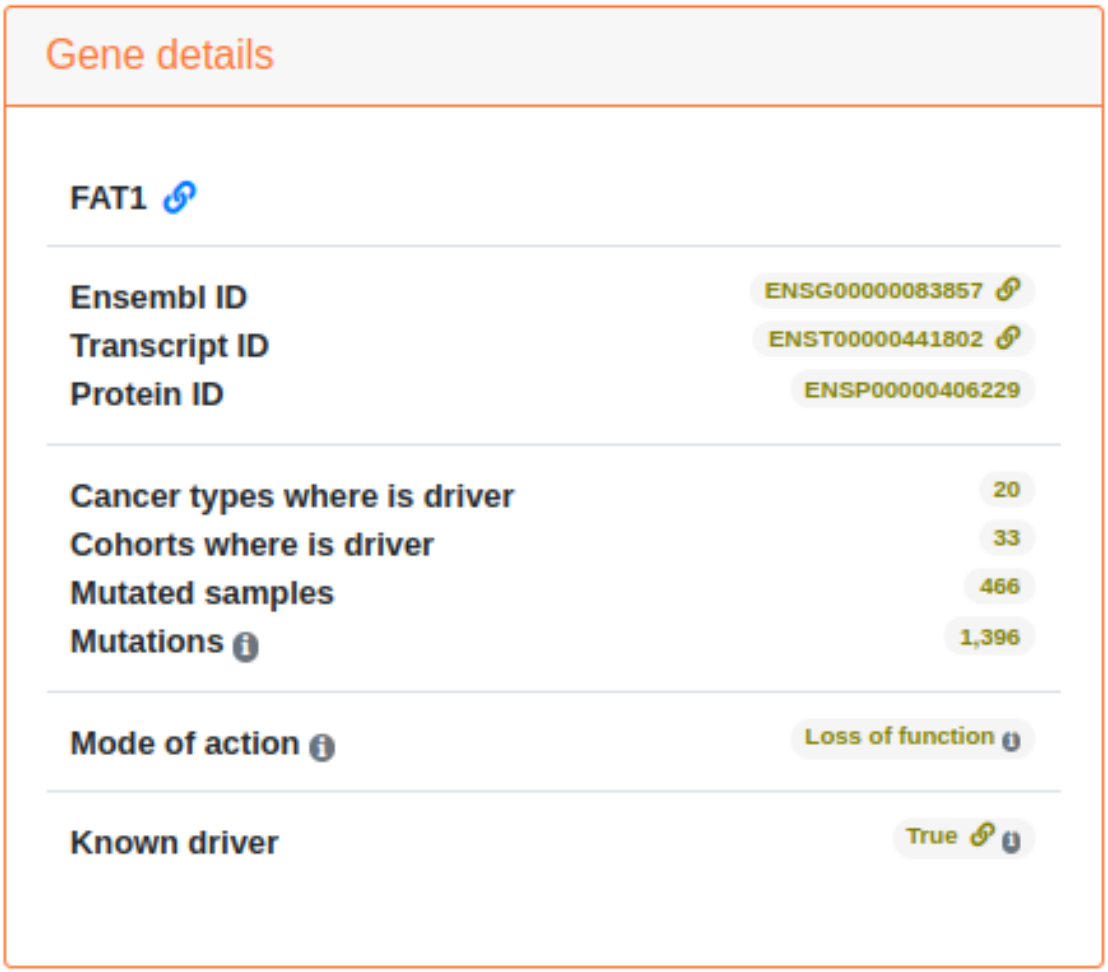
```
--2023-03-23 19:50:58--
https://github.com/ilyagerman52/bioinformatics/raw/main/homework5/
simple_somatic_mutation.open.tsv
Resolving github.com (github.com)... 140.82.112.3
Connecting to github.com (github.com)|140.82.112.3|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location:
https://raw.githubusercontent.com/ilyagerman52/bioinformatics/main/
```

```
homework5/simple_somatic_mutation.open.tsv [following]
--2023-03-23 19:50:58--
https://raw.githubusercontent.com/ilyagerman52/bioinformatics/main/
homework5/simple_somatic_mutation.open.tsv
Resolving raw.githubusercontent.com (raw.githubusercontent.com)...
185.199.108.133, 185.199.109.133, 185.199.110.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|
185.199.108.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 265136 (259K) [text/plain]
Saving to: ‘simple_somatic_mutation.open.tsv’

simple_somatic_muta 100%[===================>] 258.92K  --.-KB/s    in
0.03s

2023-03-23 19:50:58 (7.86 MB/s) - ‘simple_somatic_mutation.open.tsv’
saved [265136/265136]
```

On page we can find that chosen genome has 98 mutations



For answer the question if any mutations in gene FAT1, we must get the Ensembl ID (for FAT1 it is ENSG00000083857). This is also from IntOGen, the section `Gene Details`.




```
!grep 'ENSG00000083857' simple_somatic_mutation.open.tsv
```

The empty output implies no mutations in FAT1.

Choose another genes. For choosing interesting genes let's see all of ENSG-values of mutated genes. Then find the gene name.

```
!grep -o 'ENSG[0-9]*' simple_somatic_mutation.open.tsv | sort -u
```

```
ENSG00000005238
ENSG00000005302
ENSG00000007174
ENSG00000015479
ENSG00000025293
ENSG00000037280
ENSG00000046889
ENSG00000050730
ENSG00000060656
ENSG00000060718
```

ENSG00000065135
ENSG00000075035
ENSG00000080608
ENSG00000089123
ENSG00000091656
ENSG00000095066
ENSG00000096696
ENSG00000099260
ENSG00000099812
ENSG00000100393
ENSG00000100429
ENSG00000101230
ENSG00000101405
ENSG00000101605
ENSG00000103852
ENSG00000104970
ENSG00000105229
ENSG00000108576
ENSG00000111684
ENSG00000112033
ENSG00000112964
ENSG00000114544
ENSG00000115232
ENSG00000115267
ENSG00000117501
ENSG00000118762
ENSG00000119681
ENSG00000120211
ENSG00000121742
ENSG00000122644
ENSG00000124721
ENSG00000125207
ENSG00000125676
ENSG00000125910
ENSG00000126749
ENSG00000128815
ENSG00000131061
ENSG00000132026
ENSG00000132256
ENSG00000132274
ENSG00000132321
ENSG00000133703
ENSG00000135951
ENSG00000138792
ENSG00000139865
ENSG00000140675
ENSG00000140688
ENSG00000141258
ENSG00000141646
ENSG00000147162

ENSG00000147174
ENSG00000148671
ENSG00000155657
ENSG00000156395
ENSG00000160050
ENSG00000160051
ENSG00000160505
ENSG00000161091
ENSG00000161381
ENSG00000162482
ENSG00000163060
ENSG00000163909
ENSG00000165283
ENSG00000166147
ENSG00000168566
ENSG00000168779
ENSG00000169551
ENSG00000171222
ENSG00000172037
ENSG00000172046
ENSG00000172403
ENSG00000172995
ENSG00000173039
ENSG00000173269
ENSG00000174891
ENSG00000175175
ENSG00000177103
ENSG00000181023
ENSG00000181143
ENSG00000182253
ENSG00000182463
ENSG00000182841
ENSG00000183150
ENSG00000183287
ENSG00000183397
ENSG00000183569
ENSG00000184634
ENSG00000185010
ENSG00000185070
ENSG00000185313
ENSG00000186113
ENSG00000187800
ENSG00000188130
ENSG00000188133
ENSG00000188687
ENSG00000189056
ENSG00000196159
ENSG00000197256
ENSG00000197712
ENSG00000198286

```
ENSG00000204248
ENSG00000204790
ENSG00000204793
ENSG00000207625
ENSG00000213445
ENSG00000215021
ENSG00000222046
ENSG00000222386
ENSG00000223656
ENSG00000224066
ENSG00000230837
ENSG00000232119
ENSG00000233665
ENSG00000234715
ENSG00000237298
ENSG00000251322
ENSG00000259848
ENSG00000260740
ENSG00000261054
ENSG00000263345
ENSG00000263818
ENSG00000263938
ENSG00000264324
ENSG00000265106
ENSG00000265911
ENSG00000266368
ENSG00000267424
ENSG00000267436
ENSG00000270574
ENSG00000271253
ENSG00000271774
ENSG00000271880
ENSG00000272201
ENSG00000272734
ENSG00000273413
```

This is Ensembl ID of genes, which are mutated in case of our donor.

I can find some of them, for example let's choose this four:

- ENSG00000075035 - gene WSCD2
- ENSG00000132321 - gene IQCA1
- ENSG00000177103 - gene DSCAML1
- ENSG00000251322 - gene SHANK3

And two extra genes:

- ENSG00000114544 - gene SLC41A3
- ENSG00000005238 - gene ATOSB

Using that grep we find all mutations with `ENSG00000075035`, then greps by ID (first column with values that start with `MU...`). Sorting with flag -u effords to count unique IDs. So, all of these genes have only 1 mutation.

```
print('main 4:\n')
!grep 'ENSG00000075035' simple_somatic_mutation.open.tsv | grep -o
'MU[0-9]*' | sort -u
print('-' * 100)
!grep 'ENSG00000132321' simple_somatic_mutation.open.tsv | grep -o
'MU[0-9]*' | sort -u
print('-' * 100)
!grep 'ENSG00000177103' simple_somatic_mutation.open.tsv | grep -o
'MU[0-9]*' | sort -u
print('-' * 100)
!grep 'ENSG00000251322' simple_somatic_mutation.open.tsv | grep -o
'MU[0-9]*' | sort -u
print('-' * 100)

print('\n\nextras:\n')
!grep 'ENSG00000114544' simple_somatic_mutation.open.tsv | grep -o
'MU[0-9]*' | sort -u
print('-' * 100)
!grep 'ENSG00000005238' simple_somatic_mutation.open.tsv | grep -o
'MU[0-9]*' | sort -u
```

```
main 4:

MU130337701
----------------------------------------------------------------------------------------------------
MU101543
----------------------------------------------------------------------------------------------------
MU130328708
----------------------------------------------------------------------------------------------------
MU102334
----------------------------------------------------------------------------------------------------


extras:

MU102303
----------------------------------------------------------------------------------------------------
MU101768
```

The very interesting information is here. For example we can find the mutations `MU1330337701` and `MU130328708`, the corresponds with our mini-research based on greps :)

☐	MU130337701	chr12:g.108634268G>A	single base substitution	Missense: *WSCD2* R431H		1 / 1 *(100.00%)*	1 / 19,729 *(0.01%)*	--
☐	MU130328708	chr11:g.117389195C>T	single base substitution	Missense: *DSCAML1* R559Q, R289Q		1 / 1 *(100.00%)*	1 / 19,729 *(0.01%)*	--

Finally, we can see that the mutation `MU130337701` connected with Colon cancer, subtype Adenocarcinoma.

M MU130337701 SOMATIC

Summary | Clinical Evidence | Protein

Summary

ID	MU130337701
DNA change	chr12:g.108634268G>A
Type	single base substitution
Reference genome assembly	GRCh37
Allele in the reference assembly	G
Functional Impact	Low

Consequences

Open in Advanced Search

Showing **1** Consequences

Filter table

Gene	AA Change	Consequence	Coding DNA Change	Strand	Transcript(s)
WSCD2	R431H	Missense variant	1292G>A	+	WSCD2-201 WSCD2-002 WSCD2-001 WSCD2-003

Cancer Distribution

Open in Advanced Search

MU130337701 AFFECTS **1** DISTINCT DONORS ACROSS **1** CANCER PROJECTS

Showing **1** projects

Filter table

Project	Site	Tumour Type	Tumour Subtype	# Donors affected
COAD-US	Colorectal	Colon cancer	Adenocarcinoma	1 / 402 *(0.25%)*