GULBENKIAN CIÊNCIA

PO: Precision Oncology Course

Variant annotation, filtering and prioritization

BU CNIO BIOINFORMATICS UNIT

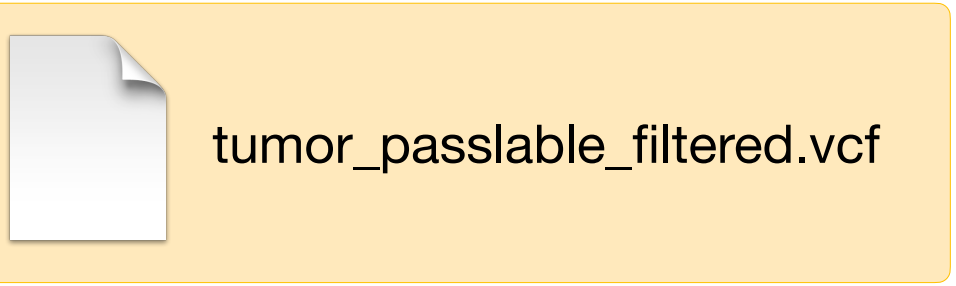
cnio stop cancer

Exercise

Annotation of the panel 1 using PanDrugs

Study case

Panel 1



Tumor type: Patient with Colon Adenocarcinoma

Sequencing platform: Illumina HiSeq2500

Type of data: Sequencing panel (paired). Ion Ampliseq Cancer Hotspot Panel v2 (46 genes)

Samples: Tumor with matched healthy tissue

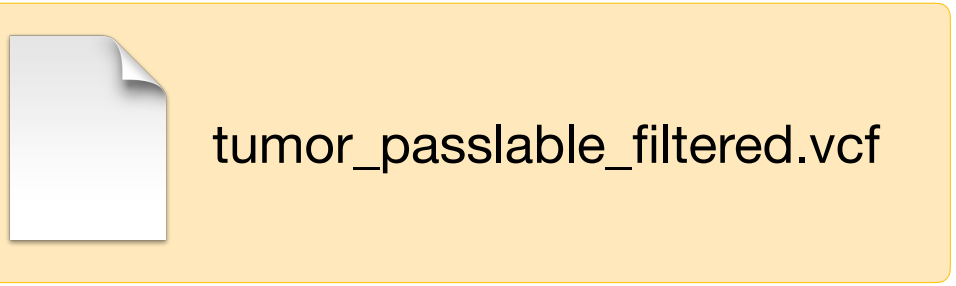
File with somatic variants from Mutect2: Variants detected in tumor sample but not in the corresponding control

Data: https://drive.google.com/drive/folders/1d1lRuevln_rL9yx2oYUppt9ylJRCtnJ7?usp=sharing

Reference genome: hg19

Study case

Panel 1



The VCF has only **5 somatic mutations** and all of them have a **PASS label** in the column FILTER.

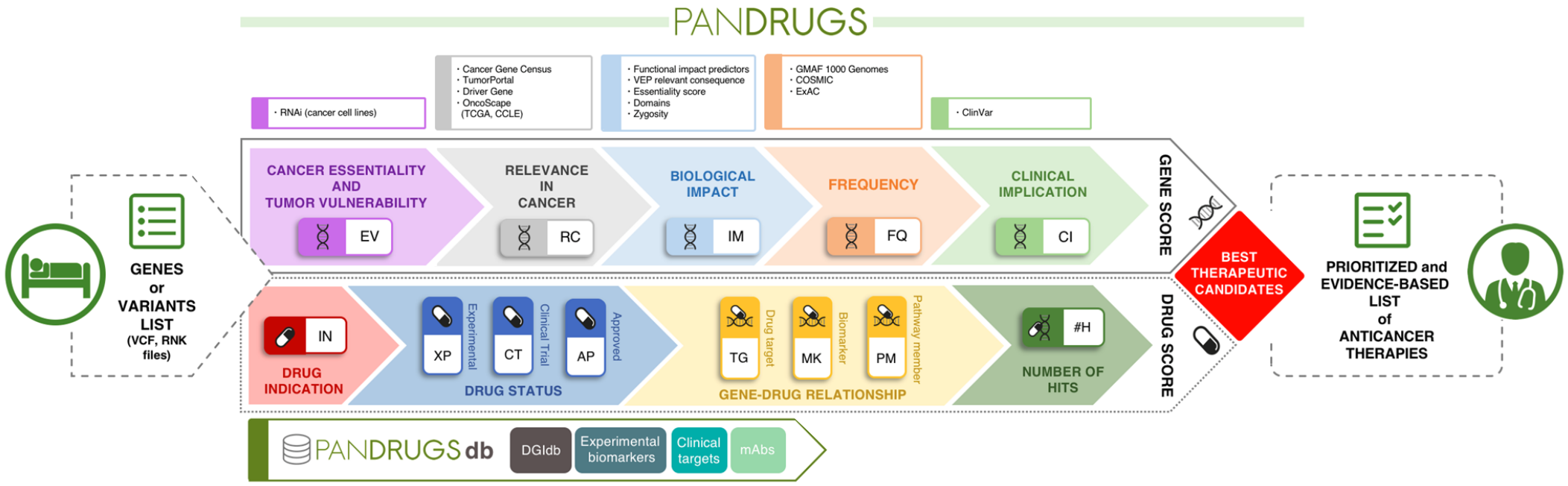
```
##normal_sample=normal
##source=FilterMutectCalls
##source=Mutect2
##source=VariantFiltration
##tumor_sample=tumor
#CHROM   POS      ID       REF      ALT      QUAL     FILTER   INFO     FORMAT   normal   tumor
chr3     178952085        .        A        G        .        PASS
CONTQ=93;DP=2498;ECNT=1;GERMQ=93;MBQ=20,20;MFRL=187,193;MMQ=60,60;MPOS=31;NALOD=2.04;NLOD=358.52;POPAF=6.00;SEQQ=93;STRANDQ=93;TLOD=876.40
GT:AD:AF:DP:F1R2:F2R1:SB  0/0:1219,2:1.536e-03:1221:1145,1:0,0:802,417,0,2   0/1:745,448:0.373:1193:694,418:0,0:499,246,300,148
chr4     153245446        .        G        A        .        PASS
CONTQ=93;DP=1949;ECNT=1;GERMQ=93;MBQ=23,31;MFRL=191,180;MMQ=60,60;MPOS=36;NALOD=2.89;NLOD=229.94;POPAF=6.00;SEQQ=93;STRANDQ=93;TLOD=2194.74
GT:AD:AF:DP:F1R2:F2R1:SB  0/0:776,1:1.282e-03:777:725,0:0,0:616,160,1,0      0/1:317,799:0.720:1116:288,747:0,0:249,68,621,178
chr5     112175423        .        C        T        .        PASS
CONTQ=93;DP=5354;ECNT=1;GERMQ=93;MBQ=32,31;MFRL=203,198;MMQ=60,60;MPOS=30;NALOD=2.88;NLOD=713.02;POPAF=6.00;SEQQ=93;STRANDQ=93;TLOD=3699.09
GT:AD:AF:DP:F1R2:F2R1:SB  0/0:2410,4:7.317e-04:2414:2250,2:0,0:995,1415,3,1  0/1:1274,1484:0.536:2758:1200,1390:0,0:509,765,773,711
chr11    108117798        .        C        T        .        PASS
CONTQ=93;DP=1223;ECNT=1;GERMQ=93;MBQ=32,33;MFRL=186,189;MMQ=60,60;MPOS=35;NALOD=2.78;NLOD=177.83;POPAF=6.00;SEQQ=93;STRANDQ=93;TLOD=721.32
GT:AD:AF:DP:F1R2:F2R1:SB  0/0:597,0:1.663e-03:597:566,0:0,0:478,119,0,0      0/1:320,281:0.468:601:301,275:0,0:252,68,224,57
chr12    25398281 .       C        T        .        PASS
CONTQ=93;DP=2926;ECNT=1;GERMQ=93;MBQ=20,20;MFRL=189,188;MMQ=60,60;MPOS=31;NALOD=2.56;NLOD=106.16;POPAF=6.00;SEQQ=93;STRANDQ=93;TLOD=3583.42
GT:AD:AF:DP:F1R2:F2R1:SB  0/0:358,1:2.746e-03:359:334,1:0,0:238,120,0,1      0/1:1015,1446:0.590:2461:950,1359:0,0:658,357,944,502
```

So we can proceed to **annotate** these 5 variants **using PanDrugs.**

PanDrugs

Annotation and prioritization of variants

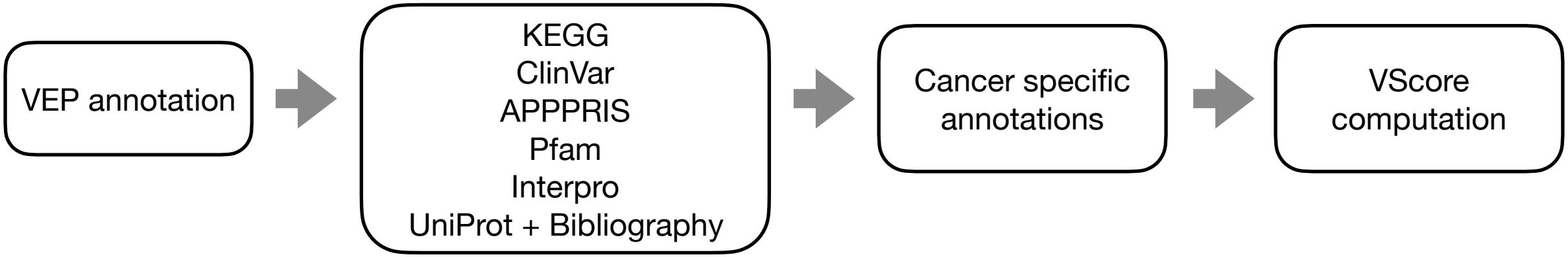
PanDrugs database unifies several resources for variant annotation. When given a VCF, PanDrugs annotates the variants and provides a Variant Score or **VScore (0 to 1)** which reflects their implication in cancer.



PanDrugs

Annotation and prioritization of variants

PanDrugs database unifies several resources for variant annotation. When given a VCF, PanDrugs annotates the variants and provides a Variant Score or **VScore (0-1)** which reflects their implication in cancer.



PanDrugs

Annotation and prioritization of variants

Databases versions

Cosmic Release v82 - hg19

Pfam 31.0 (Mar 2017)

UniProt release 2017_07 (28/08/2017)

InterPro 64.0 (28/08/2017)

Clinvar 1.49 (26/08/2017)

CGC (Cosmic v82) → The corresponding assembly is GRCh38 (but we search at gene level)

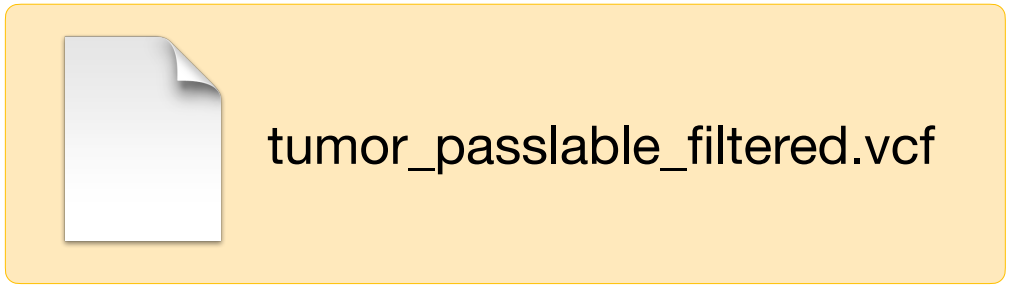
APPRIS (gen19.ensembl74 29/08/2017)

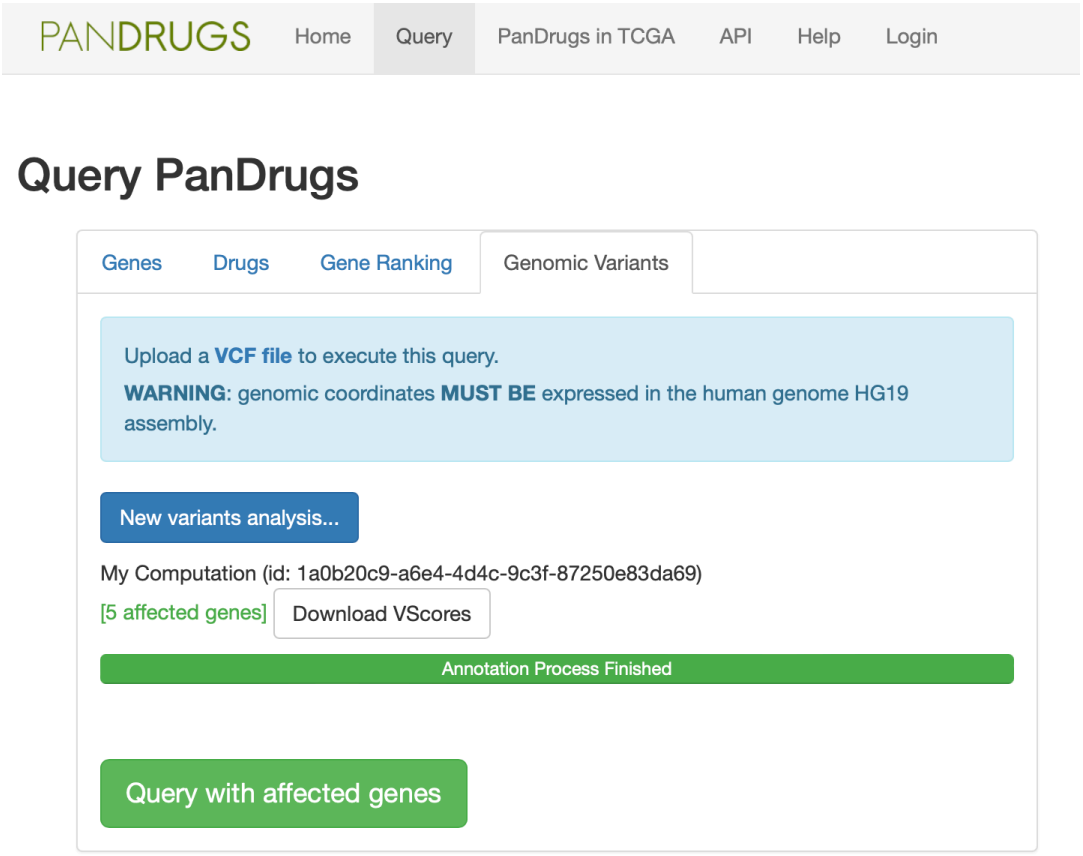
KEGG (25/08/2017)

Steps

Run PanDrugs from the web

1. Go to: https://www.pandrugs.org/#!/
2. Click in “Query” and “Genomic Variants”
3. Upload the VCF
4. Download the results clicking in “Download VScores”





Steps

Summarize PanDrugs annotation

Execute SummaryGenerator.py

```
$ python2 SummaryGenerator.py <input> <output>
```

The <input> is the output of PanDrugs.
This scripts filters the rows of <input> based on **APPRIS** annotation.

The script was created to work with the specific data used in these exercises.

APPRIS annotation

Annotation of splice isoforms

Selection of the principal isoform:

PRINCIPAL:1 - Transcript(s) expected to code for the main functional isoform based solely on the core modules in the APPRIS database.
PRINCIPAL:2 - Where the APPRIS core modules are unable to choose a clear principal variant (approximately 25% of human protein coding genes), the database chooses two or more of the CDS variants as "candidates" to be the principal variant.
PRINCIPAL:3 - Where the APPRIS core modules are unable to choose a clear principal variant and more than one of the variants have distinct CCDS identifiers, APPRIS selects the variant with lowest CCDS identifier as the principal variant.
PRINCIPAL:4 - Where the APPRIS core modules are unable to choose a clear principal CDS and there is more than one variant with a distinct (but consecutive) CCDS identifiers, APPRIS selects the longest CCDS isoform as the principal variant.
PRINCIPAL:5 - Where the APPRIS core modules are unable to choose a clear principal variant and none of the candidate variants are annotated by CCDS, APPRIS selects the longest of the candidate isoforms as the principal variant.
REST (ALTERNATIVE:1 (Candidate transcript(s) models that are conserved in at least three tested non-primate species), ALTERNATIVE:2 (Candidate transcript(s) models that appear to be conserved in fewer than three tested non-primate species), NO LABEL (Non-candidate transcripts are not flagged and are considered as "MINOR" transcripts)).

APPRIS annotation

Keep the annotations with the most reliable isoforms

Selection of the principal isoform:

PRINCIPAL:1 - Transcript(s) expected to code for the main functional isoform based solely on the core modules in the APPRIS database.
PRINCIPAL:2 - Where the APPRIS core modules are unable to choose a clear principal variant (approximately 25% of human protein coding genes), the database chooses two or more of the CDS variants as "candidates" to be the principal variant.
PRINCIPAL:3 - Where the APPRIS core modules are unable to choose a clear principal variant and more than one of the variants have distinct CCDS identifiers, APPRIS selects the variant with lowest CCDS identifier as the principal variant.
PRINCIPAL:4 - Where the APPRIS core modules are unable to choose a clear principal CDS and there is more than one variant with a distinct (but consecutive) CCDS identifiers, APPRIS selects the longest CCDS isoform as the principal variant.
PRINCIPAL:5 - Where the APPRIS core modules are unable to choose a clear principal variant and none of the candidate variants are annotated by CCDS, APPRIS selects the longest of the candidate isoforms as the principal variant.
REST (ALTERNATIVE:1 (Candidate transcript(s) models that are conserved in at least three tested non-primate species), ALTERNATIVE:2 (Candidate transcript(s) models that appear to be conserved in fewer than three tested non-primate species), NO LABEL (Non-candidate transcripts are not flagged and are considered as "MINOR" transcripts)).

SummaryGenerator.py

How it works

In the original PanDrugs output you can have several rows for the same variant:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Chr	Loc	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	sample	mut	location	allele
28	12	25398281		C	T	.	PASS	CONTQ=93;D	GT:AD:AF:DP	0/0:358,1:2.7	C/T	12:25398281	T
29	12	25398281		C	T	.	PASS	CONTQ=93;D	GT:AD:AF:DP	0/0:358,1:2.7	C/T	12:25398281	T
30	12	25398281		C	T	.	PASS	CONTQ=93;D	GT:AD:AF:DP	0/0:358,1:2.7	C/T	12:25398281	T
31	12	25398281		C	T	.	PASS	CONTQ=93;D	GT:AD:AF:DP	0/0:358,1:2.7	C/T	12:25398281	T

In column 26 you have the APPRIS isoform annotation:

	A	B	C	D	E	F–Y	Z	AA	AB	AC	AD	AE
1	Chr	Loc	ID	REF	ALT		principal	poly_effe	poly_scor	condel_ef	condel_sc	sift_effec
28	12	25398281		C	T	. P/ C(G 0/ C/ 1.7 T El El Tr mis # # # G g(CM1	PRINCIPAL:4	benign	0.44	deleterious	0.492	deleterious
29	12	25398281		C	T	. P/ C(G 0/ C/ 1.7 T El El Tr mis # # # G g(CM1	ALTERNATIVE:1	benign	0.14	neutral	0.361	deleterious
30	12	25398281		C	T	. P/ C(G 0/ C/ 1.7 T El El Tr mis # # # G/D		benign	0.018	neutral	0.349	deleterious
31	12	25398281		C	T	. P/ C(G 0/ C/ 1.7 T El El Tr mis # # # G/D		probably_dar	0.969	deleterious	0.776	deleterious_

SummaryGenerator.py

How it works

The program will **keep the row with the most reliable annotation** (PRINCIPAL:4). So, in the output of SummaryGenerator.py there is only one row for this variant:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	chr	loc	mut	gene_hgn	tumorpor	role_drive	gene	feature	conseque	functiona	cosmic_id	cosmic_v	kegg_data
9	12	25398281	C/T	KRAS	AML:Highly s	CGC:oncoger	ENSG000001	ENST000002	missense_va	benign/delet	COSM532:PA	COSM11401	EGFR tyrosin

If there are several rows with the same APPRIS annotation level, SummaryGenerator.py keeps them all.

Also, this script collapses some PanDrugs annotations (i.e. protein impact prediction by PolyPhen, SIFT and CONDEL) into a single column. **The number of columns is reduced from 62 to 33.**

Questions

30 min

- **Which fields indicate polymorphisms?**
- **Which fields have information about the effect in the sequence?**
- **Which fields have information about the effect in the protein?**
- **Which fields give specific information about the pathology under study?**

Answers

30 min

- **Which fields indicate polymorphisms?** GMAF and GMAF_freq from 1000 Genomes; gnomAD and gnomAD_NFE
- **Which fields have information about the effect in the sequence?** consequence
- **Which fields have information about the effect in the protein?** functional_impact_prediction; pfam and interpro (domains)
- **Which fields give specific information about the pathology under study?** tumorportal, role_driver, cosmic_id and clinvar annotations

Questions

30 min

- **In which processes are involved *APC* and *FBXW7* genes?**
- **Is the gene *KRAS* frequently mutated within the same tumor types?**
- **Which variant has been reported more times in tumors?**
- **Should *ATM* gene be inhibited?**
- **Name 3 candidates as relevant variants in the disease.**

Answers

30 min

- **In which processes are involved *APC* and *FBXW7* genes?** Check kegg_data
- **Is the gene *KRAS* frequently mutated within the same tumor types?** Yes (check tumorportal)
- **Which variant has been reported more times in tumors?** *KRAS* c.38G>A (check mut_cosmic_freq)
- **Should *ATM* gene be inhibited?** No, because it is a TSG (check role_driver)
- **Name 3 candidates as relevant variants in the disease:** *KRAS*, *PIK3CA* and *APC* (highest VSCore)


GULBENKIAN
CIÊNCIA

Thanks!


BU
CNIO BIOINFORMATICS UNIT


cnio
stop cancer