Study Case: panel 2

Tumor type: lung metastasis – gastrointestinal stromal tumor

Sequencing: Illumina HiSeq2500

Panel: Ion Ampliseq Comprehensive Cancer Panel (409 genes)

Reference Genome: HG19

VCF file with variants detected with MuTect2

tumor_passlable_filtered.vcf.gz

OBJECTIVES:

1. Process the VCF file to obtain the set of annotations

2. Identify clinically relevant variants involved in the disease

3. Review variants in IGV

1. Process the vcf file

Remove variants without PASS label in FILTER

```
zcat panel2/out/mutect_filter/
tumor_passlable_filtered.vcf.gz | awk -F '\t' '{if($0 ~
/#/) print; else if($7 == "PASS") print}' > panel2/out/
tumor_passlable_filtered_onlyPASS.vcf
```

How to obtain the extended set of annotations in cancer disease

https://www.pandrugs.org/

Query PanDrugs

Genes Drugs Gene Ranking **Genomic Variants**

Upload a **VCF file** to execute this query.

WARNING: genomic coordinates **MUST BE** expressed in the human genome
HG19 assembly.

New variants analysis...

Query with affected genes

New variants analysis ✕

Select a variants file (in **VCF format**). Note: The genome positions must be
accordiyng to the GRCh37/hg19 human genome assembly

[   ] 📂 Browse ...

Download a VCF example file (Melanoma patient BRAF mutant from TCGA)
Computation name

My Computation

Close   Submit VCF

VEP annotations → KEGG
ClinVar
APPRIS
Pfam
Interpro
Uniprot ID + Bibliography → Cancer disease specific
annotations → Score
computation
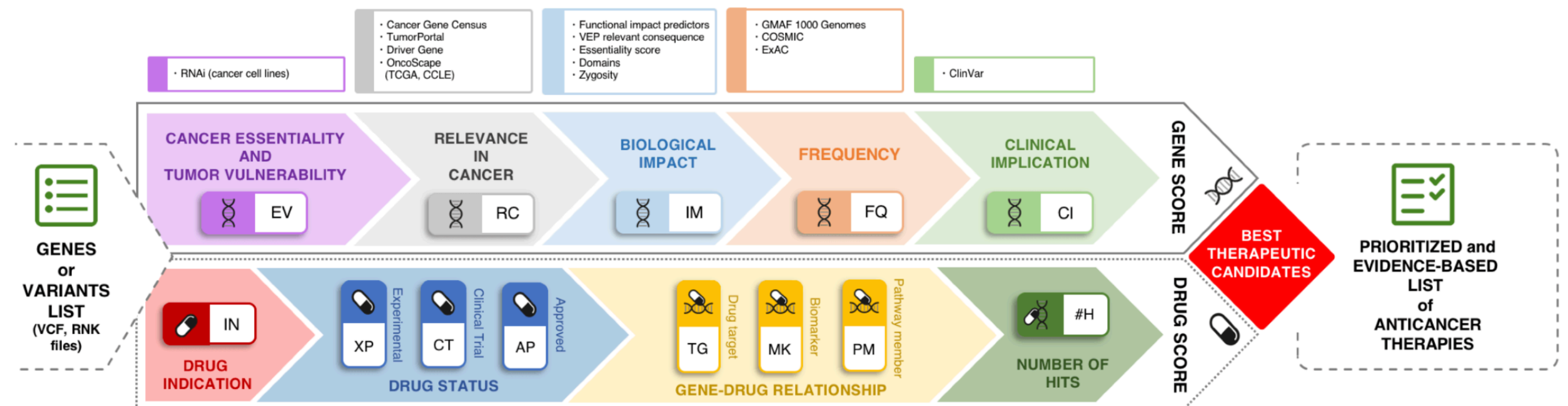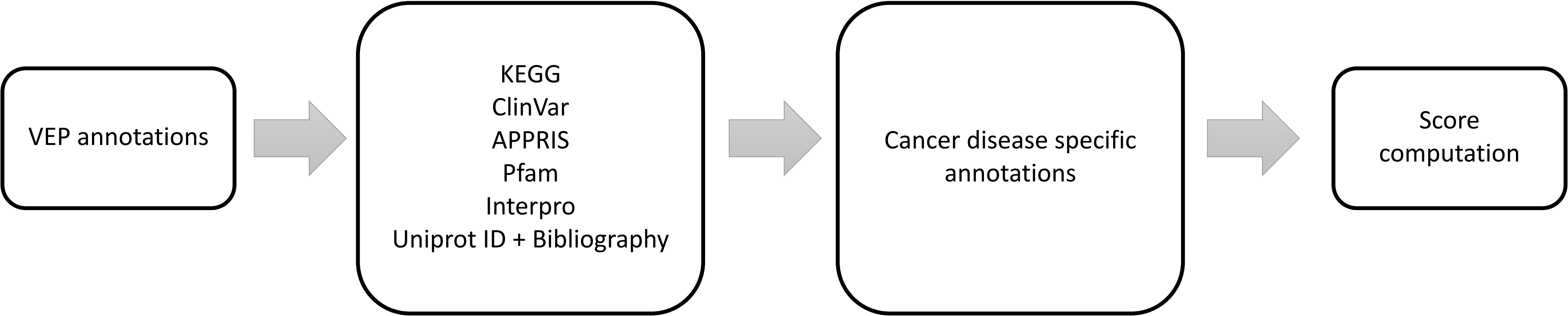
Query PanDrugs

Genes   Drugs   Gene Ranking   **Genomic Variants**

Upload a **VCF file** to execute this query.

WARNING: genomic coordinates **MUST BE** expressed in the human genome HG19 assembly.

New variants analysis...

somatic_caseCHP (id: fc04f27d-50ac-453c-ae40-ac42f0a70fc7)

[6 affected genes]   Download VScores

Annotation Process Finished

Query with affected genes

Chr	Loc	ID	REF	ALT	QUAL	FILTER	INFO
2	29432629		G	-	11505.73	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=-15.167;DP=1549;FS=171.142;MQ0=0;MQRankSum=1.542;QD=7.46;RPA=4,3;RU=G;ReadPosRankSum=-2
2	29432629		G	-	11505.73	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=-15.167;DP=1549;FS=171.142;MQ0=0;MQRankSum=1.542;QD=7.46;RPA=4,3;RU=G;ReadPosRankSum=-2
3	178952085	rs121913279	A	G	28044.77	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=-20.123;DB;DP=1959;Dels=0.00;FS=0.000;HaplotypeScore=632.2851;MQRankSum=1.129;QD=14.32;Readl
3	178952085	rs121913279	A	G	28044.77	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=-20.123;DB;DP=1959;Dels=0.00;FS=0.000;HaplotypeScore=632.2851;MQRankSum=1.129;QD=14.32;Readl
4	153245446		G	A	16147.77	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=18.653;DP=2800;Dels=0.00;FS=0.000;HaplotypeScore=196.2637;MQ0=0;MQRankSum=0.422;QD=5.77;Rea
4	153245446		G	A	16147.77	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=18.653;DP=2800;Dels=0.00;FS=0.000;HaplotypeScore=196.2637;MQ0=0;MQRankSum=0.422;QD=5.77;Rea
4	153245446		G	A	16147.77	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=18.653;DP=2800;Dels=0.00;FS=0.000;HaplotypeScore=196.2637;MQ0=0;MQRankSum=0.422;QD=5.77;Rea
4	153245446		G	A	16147.77	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=18.653;DP=2800;Dels=0.00;FS=0.000;HaplotypeScore=196.2637;MQ0=0;MQRankSum=0.422;QD=5.77;Rea
4	153245446		G	A	16147.77	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=18.653;DP=2800;Dels=0.00;FS=0.000;HaplotypeScore=196.2637;MQ0=0;MQRankSum=0.422;QD=5.77;Rea
4	153245446		G	A	16147.77	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=18.653;DP=2800;Dels=0.00;FS=0.000;HaplotypeScore=196.2637;MQ0=0;MQRankSum=0.422;QD=5.77;Rea
4	153245446		G	A	16147.77	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=18.653;DP=2800;Dels=0.00;FS=0.000;HaplotypeScore=196.2637;MQ0=0;MQRankSum=0.422;QD=5.77;Rea
4	153245446		G	A	16147.77	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=18.653;DP=2800;Dels=0.00;FS=0.000;HaplotypeScore=196.2637;MQ0=0;MQRankSum=0.422;QD=5.77;Rea
5	112175423	rs121913329	C	T	37006.77	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=-37.564;DB;DP=2583;Dels=0.01;FS=1.821;HaplotypeScore=207.0235;MQRankSum=-1.046;QD=14.34;Rea
5	112175423	rs121913329	C	T	37006.77	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=-37.564;DB;DP=2583;Dels=0.01;FS=1.821;HaplotypeScore=207.0235;MQRankSum=-1.046;QD=14.34;Rea
5	112175423	rs121913329	C	T	37006.77	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=-37.564;DB;DP=2583;Dels=0.01;FS=1.821;HaplotypeScore=207.0235;MQRankSum=-1.046;QD=14.34;Rea
5	112175423	rs121913329	C	T	37006.77	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=-37.564;DB;DP=2583;Dels=0.01;FS=1.821;HaplotypeScore=207.0235;MQRankSum=-1.046;QD=14.34;Rea
5	112175423	rs121913329	C	T	37006.77	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=-37.564;DB;DP=2583;Dels=0.01;FS=1.821;HaplotypeScore=207.0235;MQRankSum=-1.046;QD=14.34;Rea
5	112175423	rs121913329	C	T	37006.77	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=-37.564;DB;DP=2583;Dels=0.01;FS=1.821;HaplotypeScore=207.0235;MQRankSum=-1.046;QD=14.34;Rea
5	112175423	rs121913329	C	T	37006.77	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=-37.564;DB;DP=2583;Dels=0.01;FS=1.821;HaplotypeScore=207.0235;MQRankSum=-1.046;QD=14.34;Rea
5	112175423	rs121913329	C	T	37006.77	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=-37.564;DB;DP=2583;Dels=0.01;FS=1.821;HaplotypeScore=207.0235;MQRankSum=-1.046;QD=14.34;Rea

Script <mark>SummaryGenerator.py</mark>

```
$ python2 SummaryGenerator.py input_file output_file
```

- Column selection

- Selection of principal transcripts (if PRINCIPAL label)

Script created to work with the specific data used in the exercise

Selection of 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))

Most reliable

Reduced to relevant isoforms if PRINCIPAL

2. Identify clinically relevant variants and likely benign variants

1. Identify benign variants (such as those common in the population, low impact)

2. Find the variants with evidence of clinical relevance (driver variants, risk factors, drug response)

3. Check in IGV that relevant variants are well supported

You can set tiers according to the number of evidences that support the clinical relevance

Column legend in the file

chr	Chromosome
loc	Variant genomic position (hg19)
mut	Reference allele/Variant allele
gene_hgnc	Gene official name
tumorportal	tumor types where the gene is significantly mutated as indicated in TumorPortal
role_driver	oncogene or tumor suppressor gene as indicated in Cancer Genome Census (CGC) and OncodriveROLE and driver gene as referenced in Nature, 2013. Tamborero D, et al.
gene	Ensembl Gene ID
feature	Ensembl TranscrIpt ID
consequence	Variant effect on the sequence
functional_impact_prediction_(PolyPhen/SIFT/CONDEL)	Predicitve functional effect based on different algorithms
cosmic_id	COSMIC id (if any)
cosmic_vep_id	COSMIC id (if any)
kegg_data	KEGG pathway
protein_position	Variant position on protein sequence
amino_acids	aminoacid affected by the variant
dbSNP	dbSNP id (if any)
clinvar_acc	ClinVar id (if any), semicolon separated
clinvar_disease	Associated conditions (in same order as ClinVar_acc)
clinvar_clinical_significance	Clinical significance (in same order as ClinVar_acc)
variation_type	Type of mutation
HGVS_cDNA	HGVS DNA
HGVS_protein	HGVS protein
GMAF	Global minor allele frequency (MAF) in 1000 Genomes Project: dbSNP is reporting the minor allele frequency for each rs included in a default global population
GMAF_freq	GMAF frequency in percentage in 1000 Genomes Project (<1% is considered rare variant)
gnomAD	Global Allele Frequency for the variant in gnomAD (exome data) in percentage
gnomAD_NEF	Allele Frequency for the variant in Non-Finnish European population in gnomAD (exome data) in percentage
pfam	PFAM domain affected
uniprot	Uniprot bibliographic references for the variant
interpro	Interpro domain affected
gene_cosmic_freq	Gene frequency in COSMIC
mut_cosmic_freq	Variant frequency in COSMIC
vscore	CNIO score based on relevance (0-1): 1 max and 0 min