Documentation

PTV_prioritization.py

Running PTV_prioritization.py

python /path/to/PTV_prioritization.py –s step –I input.txt

Input

PTV_prioritization requires 1 command line argument:

-s step:

0 - run steps 1, 2, 3 consecutively

1 - Sample Summary PTV

2 - Variant Summary PTV

3 - Gene Summary PTV

Steps 0 and 1 require another command line argument:

-I input.txt:

Tab-separated list of samples with annotated_calls.txt files generated by OpEx.

Additional columns after sample_ID are optional.

If step 0 is being run there are another 2 optional arguments:

-o: directory where output files will be stored [default: cwd]

--name: add a representative name to all out files

When running steps 2 or 3 separately make sure that the required file (obtained from the previous step) exists in the current directory and is named in the default format i.e. for step 2, SampleSummaryPTV.txt must exist in the cwd and for step 3, VariantSummaryPTV.txt must exist in the cwd

Output

Step 1: SampleSummaryPTV.txt

Column	Field	Description	Filter
1	CHROM	Chromosome of variant	
2	POS	Genomic position of variant	
3	REF	Reference allele of variant	<=11
4	ALT	Alternative allele of variant	<=11
5	QUAL	QUAL value in the VCF record (see Platypus documentation)	
6	QUALFLAG	Value of “high” if the variant is a base substitution with QUAL score of 100 or higher, or the variant is an indel with a variant allele proportion (as defined by the TR value divided by the TC value) greater than 0.2 and the variant has a FILTER value of PASS. value of "low" otherwise.	high
7	FILTER	Variant calling FILTER value in the VCF record	
8	TR	Total number of reads containing the variant (see Platypus documentation)	>=10
9	TC	Total coverage at this locus	>=20
10	SAMPLE	Sample name	
11	GT	Genotype called in the sample (see Platypus documentation)	
12	TYPE	Variant type (SUBSTITUTION, INSERTION, DELETION, COMPLEX)	
13	ENST	Ensembl transcript ID	
14	GENE	Gene symbol	
15	TRINFO	Transcript information	
16	LOC	Within-transcript location of variant	
17	CSN	Clinical Sequence Nomeclature (see CAVA documentation)	
18	CLASS	Variant class annotation (see CAVA documentation)	
19	SO	Sequence ontology annotation (see CAVA documentation)	
20	IMPACT	Variant impact (see CAVA documentation)	1
21	ALTANN	Alternative annotation (see CAVA documentation)	
22	ALTCLASS	Alternative CLASS annotation (see CAVA documentation)	
23	ALTSO	Alternative SO annotation (see CAVA documentation)	
24	CountICR1000	Number of times ENST+CSN is seen in ICR1000 series, i.e variant count	<=1
25	CountInHouse419	Number of times ENST+CSN is seen in InHouse419 series, i.e variant count	
26	CountExACNFE	Number of times ENST+CSN is seen in ExAC NFE series, i.e variant count	<=10
27	CountExACTotal	Number of times ENST+CSN is seen in the full ExAC series, i.e variant count	
	C1…CN	Additional optional columns, taken from the input file	

Step 2: VariantSummaryPTV.txt

Column	Field	Description
1	CHROM	Chromosome of variant
2	POS	Genomic position of variant
3	REF	Reference allele of variant
4	ALT	Alternative allele of variant
5	TYPE	Variant type (SUBSTITUTION, INSERTION, DELETION, COMPLEX)
6	ENST	Ensembl transcript ID
7	GENE	Gene symbol
8	TRINFO	Transcript information
9	LOC	Within-transcript location of variant
10	CSN	Clinical Sequence Nomeclature (see CAVA documentation)
11	CLASS	Variant class annotation (see CAVA documentation)
12	SO	Sequence ontology annotation (see CAVA documentation)
13	IMPACT	Variant impact (see CAVA documentation)
14	ALTANN	Alternative annotation (see CAVA documentation)
15	ALTCLASS	Alternative CLASS annotation (see CAVA documentation)
16	ALTSO	Alternative SO annotation (see CAVA documentation)
17	CountCase	number of times ENST+CSN is seen in SampleSummaryPTV.txt i.e. variant count
18	CountICR1000	Number of times ENST+CSN is seen in ICR1000 series, i.e variant count
19	CountInHouse419	Number of times ENST+CSN is seen in InHouse419 series, i.e variant count
20	CountExACNFE	Number of times ENST+CSN is seen in ExAC NFE series, i.e variant count
21	CountExACTotal	Number of times ENST+CSN is seen in the full ExAC series, i.e variant count

Step_3: GeneSummaryPTV.txt

Column	Field	Description
1	GENE	Gene symbol
2	TRINFO	Transcript information
3	ENST	Ensembl transcript ID
4	CaseTotalRarePTV	Sum of CountCase
5	CaseDifferentRarePTV	Number of CountCase rows
6	CaseSingletonPTV	Number of CountCase=1
7	ICR1000TotalRarePTV	Corresponding column from GeneSummaryPTV_ICR1000.txt
8	ICR1000DifferentRarePTV	Corresponding column from GeneSummaryPTV_ICR1000.txt
9	ICR1000SingletonPTV	Corresponding column from GeneSummaryPTV_ICR1000.txt
10	InHouse419TotalRarePTV	Corresponding column from GeneSummaryPTV_InHouse419.txt
11	InHouse419DifferentRarePTV	Corresponding column from GeneSummaryPTV_InHouse419.txt
12	InHouse419SingletonPTV	Corresponding column from GeneSummaryPTV_InHouse419.txt