GeneRollup Documentation

Chris Gates and Jessica Bene
UM BRCF Bioinformatics Core
January 2016

Overview

The variant lines in the input file are grouped by gene. Each output line aggregates the mutation, SnpEff, and dbNSFP values, as described in the below sections.

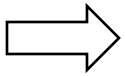
The number of output lines is equal to the number of distinct genes. The output lines are sorted such that the most impacted genes are listed first (based on *dbNSFP overall damaging rank* and *SnpEff overall impact rank*).

A sample is considered to be '**passed**' if it contains a value other than ".", "0", or blank. Otherwise, it is considered to be 'failed'.

A **sample-variant** is defined to be the intersection of a locus and a sample column.

Overview

Variant	SNPEFF_TOP_EFFECT_GENE_SYMBOL	JQ_SUMMARY_SOM_COUNT|Sample1	JQ_SUMMARY_SOM_COUNT|Sample2	JQ_SUMMARY_SOM_COUNT|Sample3	SNPEFF_TOP_EFFECT_IMPACT	dbNSFP_rollup_damaging
V1	ERBB2	1	1	0	HIGH	9
V2	ERBB2		1	1	HIGH	7
V3	ERBB2	1	1	1	LOW	3
V4	BRCA1	1	0	.	HIGH	9
V5	BRCA1	.	1	.	LOW	.
V6	BRCA1	1	1	.	LOW	2



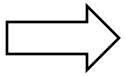
gene symbol	total impacted samples	distinct loci	total mutations	dbNSFP|overall damaging rank	dbNSFP|damaging votes|Sample1	dbNSFP|damaging votes|Sample2	dbNSFP|damaging votes|Sample3	SnpEff|overall impact rank	SnpEff|impact category|HIGH	SnpEff|impact category|MODERATE	SnpEff|impact category|LOW	SnpEff|impact category|MODIFIER	SnpEff|impact|Sample1	SnpEff|impact|Sample2	SnpEff|impact|Sample3
ERBB2	3	3	7	1	12	19	10	1	4	0	3	0	hl	hhl	hl
BRCA1	2	3	4	2	11	2		2	1	0	3	0	hl	ll	

gene symbol

Variant lines are merged by gene. In this example, the *SNPEFF_TOP_EFFECT_GENE_SYMBOL* column lists the genes.

gene symbol

Variant	SNPEFF_TOP_EFFECT_GENE_SYMBOL	JQ_SUMMARY_SOM_COUNT|Sample1	JQ_SUMMARY_SOM_COUNT|Sample2	JQ_SUMMARY_SOM_COUNT|Sample3	SNPEFF_TOP_EFFECT_IMPACT	dbNSFP_rollup_damaging
V1	ERBB2	1	1	0	HIGH	9
V2	ERBB2		1	1	HIGH	7
V3	ERBB2	1	1	1	LOW	3
V4	BRCA1	1	0	.	HIGH	9
V5	BRCA1	.	1	.	LOW	.
V6	BRCA1	1	1	.	LOW	2



gene symbol	total impacted samples	distinct loci	total mutations	dbNSFP|overall damaging rank	dbNSFP|damaging votes|Sample1	dbNSFP|damaging votes|Sample2	dbNSFP|damaging votes|Sample3	SnpEff|overall impact rank	SnpEff|impact category|HIGH	SnpEff|impact category|MODERATE	SnpEff|impact category|LOW	SnpEff|impact category|MODIFIER	SnpEff|impact|Sample1	SnpEff|impact|Sample2	SnpEff|impact|Sample3
ERBB2	3	3	7	1	12	19	10	1	4	0	3	0	hl	hhl	hl
BRCA1	2	3	4	2	11	2		2	1	0	3	0	hl	ll	

total impacted samples

For each gene, the count of passed samples is calculated. In this example, the *JQ_SUMMARY_SOM_COUNT* prefix depicts the columns for each sample.

total impacted samples

Variant	SNPEFF_TOP_EFFECT_GENE_SYMBOL	JQ_SUMMARY_SOM_COUNT|Sample1	JQ_SUMMARY_SOM_COUNT|Sample2	JQ_SUMMARY_SOM_COUNT|Sample3	SNPEFF_TOP_EFFECT_IMPACT	dbNSFP_rollup_damaging
V1	ERBB2	1	1	0	HIGH	9
V2	ERBB2		1	1	HIGH	7
V3	ERBB2	1	1	1	LOW	3
V4	BRCA1	1	0	.	HIGH	9
V5	BRCA1	.	1	.	LOW	.
V6	BRCA1	1	1	.	LOW	2

Variant	SNPEFF_TOP_EFFECT_GENE_SYMBOL	JQ_SUMMARY_SOM_COUNT|Sample1	JQ_SUMMARY_SOM_COUNT|Sample2	JQ_SUMMARY_SOM_COUNT|Sample3	SNPEFF_TOP_EFFECT_IMPACT	dbNSFP_rollup_damaging
V1	ERBB2	✓	✓		HIGH	9
V2	ERBB2		✓	✓	HIGH	7
V3	ERBB2	✓	✓	✓	LOW	3
V4	BRCA1	✓			HIGH	9
V5	BRCA1		✓		LOW	.
V6	BRCA1	✓	✓		LOW	2

✓ ✓ ✗

1 + 1 + 0 = 2

gene symbol	total impacted samples	distinct loci	total mutations	dbNSFP | overall damaging rank	dbNSFP | damaging votes | Sample1	dbNSFP | damaging votes | Sample2	dbNSFP | damaging votes | Sample3	SnpEff | overall impact rank	SnpEff | impact category | HIGH	SnpEff | impact category | MODERATE	SnpEff | impact category | LOW	SnpEff | impact category | MODIFIER	SnpEff | impact | Sample1	SnpEff | impact | Sample2	SnpEff | impact | Sample3
ERBB2	3	3	7	1	12	19	10	1	4	0	3	0	hl	hhl	hl
BRCA1	2	3	4	2	11	2		2	1	0	3	0	hl	ll	

distinct loci

Each line in the input file corresponds to a locus. For each gene, a locus is counted if there is at least one passed sample-variant at that locus. Duplicate loci are only counted once.

distinct loci

Variant	SNPEFF_TOP_EFFECT_GENE_SYMBOL	JQ_SUMMARY_SOM_COUNT|Sample1	JQ_SUMMARY_SOM_COUNT|Sample2	JQ_SUMMARY_SOM_COUNT|Sample3	SNPEFF_TOP_EFFECT_IMPACT	dbNSFP_rollup_damaging
V1	ERBB2	1	1	0	HIGH	9
V2	ERBB2		1	1	HIGH	7
V3	ERBB2	1	1	1	LOW	3
V4	BRCA1	1	0	.	HIGH	9
V5	BRCA1	.	1	.	LOW	.
V6	BRCA1	1	1	.	LOW	2

Variant	SNPEFF_TOP_EFFECT_GENE_SYMBOL	JQ_SUMMARY_SOM_COUNT|Sample1	JQ_SUMMARY_SOM_COUNT|Sample2	JQ_SUMMARY_SOM_COUNT|Sample3	SNPEFF_TOP_EFFECT_IMPACT	dbNSFP_rollup_damaging
V1	ERBB2	✓	✓		HIGH	9
V2	ERBB2		✓	✓	HIGH	7
V3	ERBB2	✓	✓	✓	LOW	3
V4	BRCA1	✓			HIGH	9
V5	BRCA1		✓		LOW	.
V6	BRCA1	✓	✓		LOW	2

1 + 1 + 1 = 3

gene symbol	total impacted samples	distinct loci	total mutations	dbNSFP | overall damaging rank	dbNSFP | damaging votes | Sample1	dbNSFP | damaging votes | Sample2	dbNSFP | damaging votes | Sample3	SnpEff | overall impact rank	SnpEff | impact category | HIGH	SnpEff | impact category | MODERATE	SnpEff | impact category | LOW	SnpEff | impact category | MODIFIER	SnpEff | impact | Sample1	SnpEff | impact | Sample2	SnpEff | impact | Sample3
ERBB2	3	3	7	1	12	19	10	1	4	0	3	0	hl	hhl	hl
BRCA1	2	3	4	2	11	2		2	1	0	3	0	hl	ll	

total mutations

For each gene, the total mutations is determined by the number of passed sample-variants.

total mutations

Variant	SNPEFF_TOP_EFFECT_GENE_SYMBOL	JQ_SUMMARY_SOM_COUNT|Sample1	JQ_SUMMARY_SOM_COUNT|Sample2	JQ_SUMMARY_SOM_COUNT|Sample3	SNPEFF_TOP_EFFECT_IMPACT	dbNSFP_rollup_damaging
V1	ERBB2	1	1	0	HIGH	9
V2	ERBB2		1	1	HIGH	7
V3	ERBB2	1	1	1	LOW	3
V4	BRCA1	1	0	.	HIGH	9
V5	BRCA1	.	1	.	LOW	.
V6	BRCA1	1	1	.	LOW	2

Variant	SNPEFF_TOP_EFFECT_GENE_SYMBOL	JQ_SUMMARY_SOM_COUNT|Sample1	JQ_SUMMARY_SOM_COUNT|Sample2	JQ_SUMMARY_SOM_COUNT|Sample3	SNPEFF_TOP_EFFECT_IMPACT	dbNSFP_rollup_damaging
V1	ERBB2	✓	✓		HIGH	9
V2	ERBB2		✓	✓	HIGH	7
V3	ERBB2	✓	✓	✓	LOW	3
V4	BRCA1	✓			HIGH	9
V5	BRCA1		✓		LOW	.
V6	BRCA1	✓	✓		LOW	2

1 + 1 + 1 + 1 = 4

gene symbol	total impacted samples	distinct loci	total mutations	dbNSFP | overall damaging rank	dbNSFP | damaging votes | Sample1	dbNSFP | damaging votes | Sample2	dbNSFP | damaging votes | Sample3	SnpEff | overall impact rank	SnpEff | impact category | HIGH	SnpEff | impact category | MODERATE	SnpEff | impact category | LOW	SnpEff | impact category | MODIFIER	SnpEff | impact | Sample1	SnpEff | impact | Sample2	SnpEff | impact | Sample3
ERBB2	3	3	7	1	12	19	10	1	4	0	3	0	hl	hhl	hl
BRCA1	2	3	4	2	11	2		2	1	0	3	0	hl	ll	

dbNSFP | damaging votes | SampleX

For each gene, the *dbNSFP_rollup_damaging* values are summed within each sample if their corresponding sample-variant is passed.

In other words, for a particular passed sample-variant within a gene, the *dbNSFP_rollup_damaging* value which corresponds to that locus is added to the total damaging votes for that sample.

dbNSFP | damaging votes | SampleX

Variant	SNPEFF_TOP_EFFECT_GENE_SYMBOL	JQ_SUMMARY_SOM_COUNT|Sample1	JQ_SUMMARY_SOM_COUNT|Sample2	JQ_SUMMARY_SOM_COUNT|Sample3	SNPEFF_TOP_EFFECT_IMPACT	dbNSFP_rollup_damaging
V1	ERBB2	1	1	0	HIGH	9
V2	ERBB2		1	1	HIGH	7
V3	ERBB2	1	1	1	LOW	3
V4	BRCA1	1	0	.	HIGH	9
V5	BRCA1	.	1	.	LOW	.
V6	BRCA1	1	1	.	LOW	2

Variant	SNPEFF_TOP_EFFECT_GENE_SYMBOL	JQ_SUMMARY_SOM_COUNT|Sample1	JQ_SUMMARY_SOM_COUNT|Sample2	JQ_SUMMARY_SOM_COUNT|Sample3	SNPEFF_TOP_EFFECT_IMPACT	dbNSFP_rollup_damaging
V1	ERBB2	9	9		HIGH	9
V2	ERBB2		7	7	HIGH	7
V3	ERBB2	3	3	3	LOW	3
V4	BRCA1	9	0	2	HIGH	9
V5	BRCA1		.		LOW	.
V6	BRCA1	2	2		LOW	2

9 + 0 + 2 = 11

0 + 0 + 0 = 0

0 + 0 + 2 = 2

gene symbol	total impacted samples	distinct loci	total mutations	dbNSFP | overall damaging rank	dbNSFP | damaging votes | Sample1	dbNSFP | damaging votes | Sample2	dbNSFP | damaging votes | Sample3	SnpEff | overall impact rank	SnpEff | impact category | HIGH	SnpEff | impact category | MODERATE	SnpEff | impact category | LOW	SnpEff | impact category | MODIFIER	SnpEff | impact | Sample1	SnpEff | impact | Sample2	SnpEff | impact | Sample3
ERBB2	3	3	7	1	12	19	10	1	4	0	3	0	hl	hhl	hl
BRCA1	2	3	4	2	11	2		2	1	0	3	0	hl	ll	

dbNSFP | overall damaging rank

For each gene, the sum of the dbNSFP votes across all samples is calculated. These sums are then ranked such that the highest sum corresponds to a rank of 1.

dbNSFP | overall damaging rank

Variant	SNPEFF_TOP_EFFECT_GENE_SYMBOL	JQ_SUMMARY_SOM_COUNT|Sample1	JQ_SUMMARY_SOM_COUNT|Sample2	JQ_SUMMARY_SOM_COUNT|Sample3	SNPEFF_TOP_EFFECT_IMPACT	dbNSFP_rollup_damaging
V1	ERBB2	1	1	0	HIGH	9
V2	ERBB2		1	1	HIGH	7
V3	ERBB2	1	1	1	LOW	3
V4	BRCA1	1	0	.	HIGH	9
V5	BRCA1	.	1	.	LOW	.
V6	BRCA1	1	1	.	LOW	2

gene symbol	total impacted samples	distinct loci	total mutations	dbNSFP|overall damaging rank	dbNSFP|damaging votes|Sample1	dbNSFP|damaging votes|Sample2	dbNSFP|damaging votes|Sample3	SnpEff|overall impact rank	SnpEff|impact category|HIGH	SnpEff|impact category|MODERATE	SnpEff|impact category|LOW	SnpEff|impact category|MODIFIER	SnpEff|impact|Sample1	SnpEff|impact|Sample2	SnpEff|impact|Sample3
ERBB2	3	3	7	1	12	19	10	1	4	0	3	0	hl	hhl	hl
BRCA1	2	3	4	2	11	2		2	1	0	3	0	hl	ll	

12 + 19 + 10 = 41

11 + 2 + 0 = 13

41 > 13

1. 41
2. 13

SnpEff | impact category | X

In this case, the *SNPEFF_TOP_EFFECT_IMPACT* column contains impact categories. For each impact category within each gene, the total number of mutations is stored in its corresponding *SnpEff | impact category* column.

In other words, for each gene, the data is pivoted on *SNPEFF_TOP_EFFECT_IMPACT*, and the count of passed samples-variants is listed for each impact category.

SnpEff | impact category | X

Variant	SNPEFF_TOP_EFFECT_GENE_SYMBOL	JQ_SUMMARY_SOM_COUNT|Sample1	JQ_SUMMARY_SOM_COUNT|Sample2	JQ_SUMMARY_SOM_COUNT|Sample3	SNPEFF_TOP_EFFECT_IMPACT	dbNSFP_rollup_damaging
V1	ERBB2	1	1	0	HIGH	9
V2	ERBB2		1	1	HIGH	7
V3	ERBB2	1	1	1	LOW	3
V4	BRCA1	1	0	.	HIGH	9
V5	BRCA1	.	1	.	LOW	.
V6	BRCA1	1	1	.	LOW	2

Variant	SNPEFF_TOP_EFFECT_GENE_SYMBOL	JQ_SUMMARY_SOM_COUNT|Sample1	JQ_SUMMARY_SOM_COUNT|Sample2	JQ_SUMMARY_SOM_COUNT|Sample3	SNPEFF_TOP_EFFECT_IMPACT	dbNSFP_rollup_damaging
V1	ERBB2	h	h		HIGH	9
V2	ERBB2		h	h	HIGH	7
V3	ERBB2	l	l	l	LOW	3
V4	BRCA1	h			HIGH	9
V5	BRCA1		l		LOW	.
V6	BRCA1	l	l		LOW	2

h → 1 l → 1 + 1 + 1 = 3

m → 0 x → 0

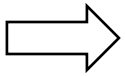
gene symbol	total impacted samples	distinct loci	total mutations	dbNSFP | overall damaging rank	dbNSFP | damaging votes | Sample1	dbNSFP | damaging votes | Sample2	dbNSFP | damaging votes | Sample3	SnpEff | overall impact rank	SnpEff | impact category | HIGH	SnpEff | impact category | MODERATE	SnpEff | impact category | LOW	SnpEff | impact category | MODIFIER	SnpEff | impact | Sample1	SnpEff | impact | Sample2	SnpEff | impact | Sample3
ERBB2	3	3	7	1	12	19	10	1	4	0	3	0	hl	hhl	hl
BRCA1	2	3	4	2	11	2		2	1	0	3	0	hl	ll	

SnpEff | overall impact rank

Impact categories are weighted such that HIGH has the most significance and MODIFIER has the least significance. The *SnpEff | impact category* values are multiplied by their corresponding weight to obtain an impact score. These scores then ranked such that the highest score corresponds to a rank of 1.

SnpEff | overall impact rank

Variant	SNPEFF_TOP_EFFECT_GENE_SYMBOL	JQ_SUMMARY_SOM_COUNT|Sample1	JQ_SUMMARY_SOM_COUNT|Sample2	JQ_SUMMARY_SOM_COUNT|Sample3	SNPEFF_TOP_EFFECT_IMPACT	dbNSFP_rollup_damaging
V1	ERBB2	1	1	0	HIGH	9
V2	ERBB2		1	1	HIGH	7
V3	ERBB2	1	1	1	LOW	3
V4	BRCA1	1	0	.	HIGH	9
V5	BRCA1	.	1	.	LOW	.
V6	BRCA1	1	1	.	LOW	2



gene symbol	total impacted samples	distinct loci	total mutations	dbNSFP|overall damaging rank	dbNSFP|damaging votes|Sample1	dbNSFP|damaging votes|Sample2	dbNSFP|damaging votes|Sample3	SnpEff|overall impact rank	SnpEff|impact category|HIGH	SnpEff|impact category|MODERATE	SnpEff|impact category|LOW	SnpEff|impact category|MODIFIER	SnpEff|impact|Sample1	SnpEff|impact|Sample2	SnpEff|impact|Sample3
ERBB2	3	3	7	1	12	19	10	1	4	0	3	0	hl	hhl	hl
BRCA1	2	3	4	2	11	2		2	1	0	3	0	hl	ll	

$4w_1 + 0w_2 + 3w_3 + 0w_4 = 4000003000$

$1w_1 + 0w_2 + 3w_3 + 0w_4 = 1000003000$

$w_1 = 10^9$
$w_2 = 10^6$
$w_3 = 10^3$
$w_4 = 10^0$

$4000003000 > 1000003000$

1. 4000003000
2. 1000003000

SnpEff | impact | SampleX

The *SNPEFF_TOP_EFFECT_IMPACT* values are mapped to a single letter such that: HIGH = h, MODERATE = m, LOW = l, MODIFIER = x. For each gene, its single letters are concatenated within each sample if their corresponding sample-variant is passed.

In other words, for a particular passed sample-variant within a gene, the single letter for the *SNPEFF_TOP_EFFECT_IMPACT* value which corresponds to that locus is appended to the SnpEff impact.

SnpEff | impact | SampleX

Variant	V1	V2	V3	V4	V5	V6
SNPEFF_TOP_EFFECT_GENE_SYMBOL	ERBB2	ERBB2	ERBB2	BRCA1	BRCA1	BRCA1
JQ_SUMMARY_SOM_COUNT|Sample1	1		1	1	.	1
JQ_SUMMARY_SOM_COUNT|Sample2	1	1	1	0	1	1
JQ_SUMMARY_SOM_COUNT|Sample3	0	1	1	.	.	.
SNPEFF_TOP_EFFECT_IMPACT	HIGH	HIGH	LOW	HIGH	LOW	LOW
dbNSFP_rollup_damaging	9	7	3	9	.	2

Variant	V1	V2	V3	V4	V5	V6
SNPEFF_TOP_EFFECT_GENE_SYMBOL	ERBB2	ERBB2	ERBB2	BRCA1	BRCA1	BRCA1
JQ_SUMMARY_SOM_COUNT|Sample1	h		l	h		l
JQ_SUMMARY_SOM_COUNT|Sample2	h	h	l		l	l
JQ_SUMMARY_SOM_COUNT|Sample3		h	l			
SNPEFF_TOP_EFFECT_IMPACT	HIGH	HIGH	LOW	HIGH	LOW	LOW
dbNSFP_rollup_damaging	9	7	3	9	.	2

gene symbol	ERBB2	BRCA1
total impacted samples	3	2
distinct loci	3	3
total mutations	7	4
dbNSFP|overall damaging rank	1	2
dbNSFP|damaging votes|Sample1	12	11
dbNSFP|damaging votes|Sample2	19	2
dbNSFP|damaging votes|Sample3	10	
SnpEff|overall impact rank	1	2
SnpEff|impact category|HIGH	4	1
SnpEff|impact category|MODERATE	0	0
SnpEff|impact category|LOW	3	3
SnpEff|impact category|MODIFIER	0	0
SnpEff|impact|Sample1	hl	hl
SnpEff|impact|Sample2	hhl	ll
SnpEff|impact|Sample3	hl	