GeneRollup Overview

Chris Gates and Jessica Bene
UM BRCF Bioinformatics Core
January 2016

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	

1

- This document explains how the GeneRollup command-line tool translates input variant file into the output gene file.
- After reading this document you should understand how each cell of the output is derived well enough to explain it to a computationally-interested PI.
- Each of the following pages focuses on how a single element of the output is derived.

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.

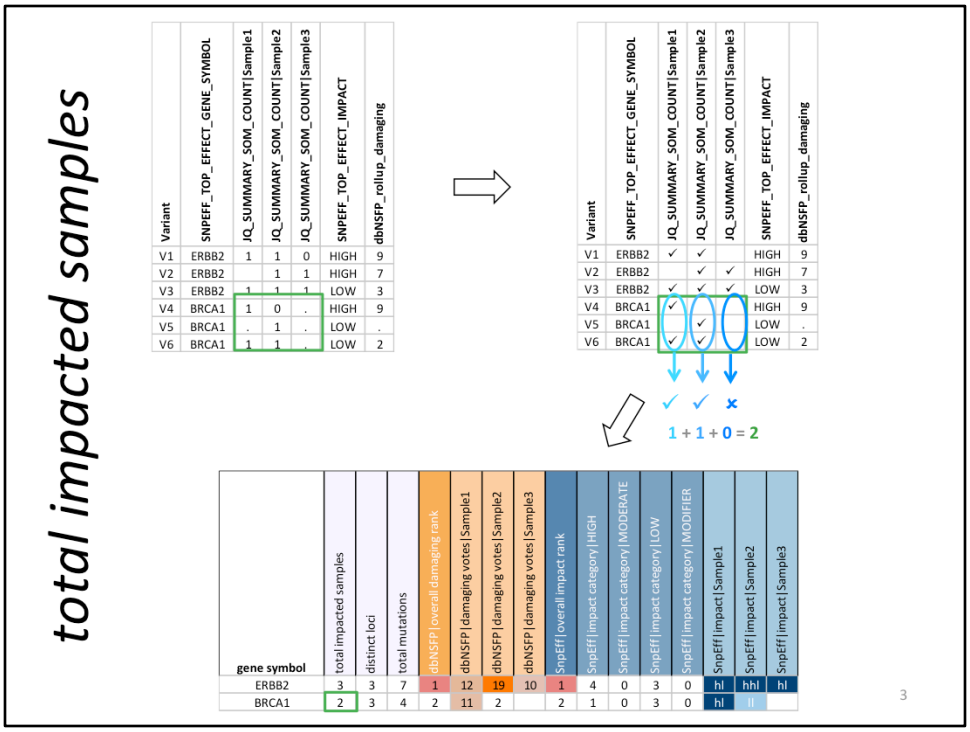
gene symbol

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

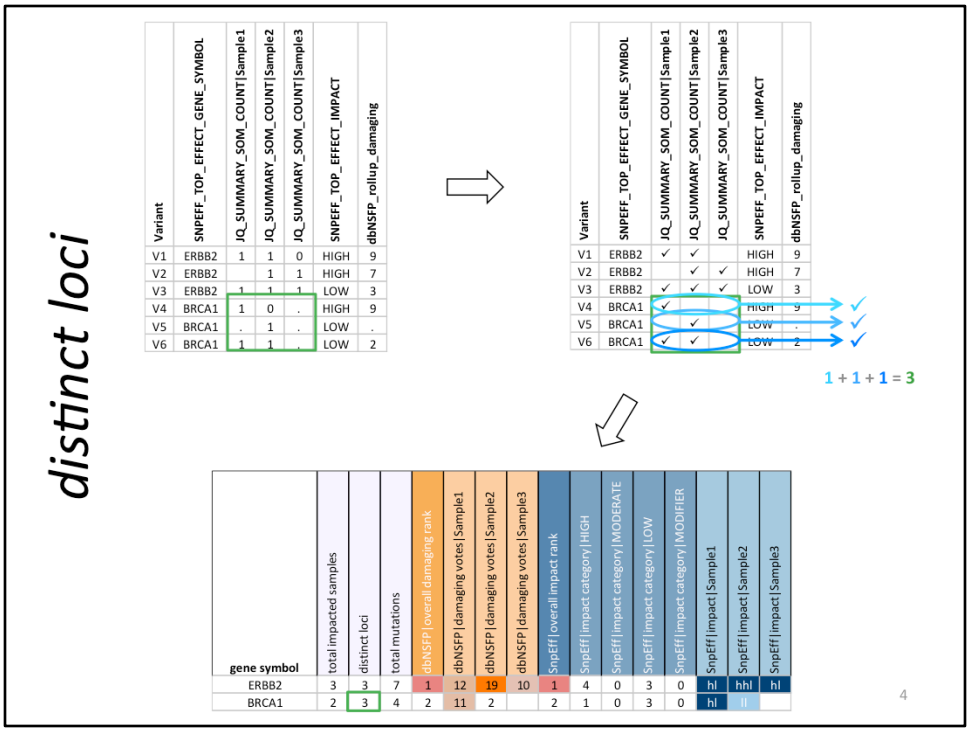
gene symbol	total impacted samples	distinct loci	total mutations	dbNSFP| overall damaging rank	dbNSFP| damaging votes|Sample1	dbNSFP| damaging votes|Sample2	dbNSFP| damaging votes|Sample3	SnpEff| overall impact rank	SnpEff| impact category| HIGH	SnpEff| impact category| MODERATE	SnpEff| impact category| 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	

2

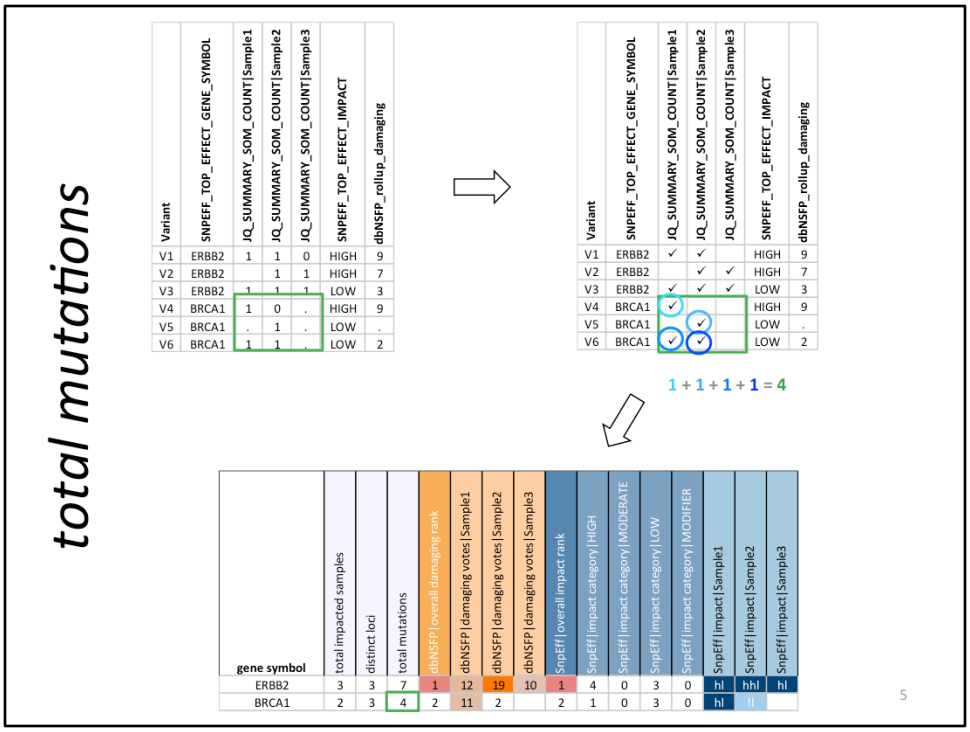
Variant lines are merged by gene. In this example, the *SNPEFF_TOP_EFFECT_GENE_SYMBOL* column lists the genes.



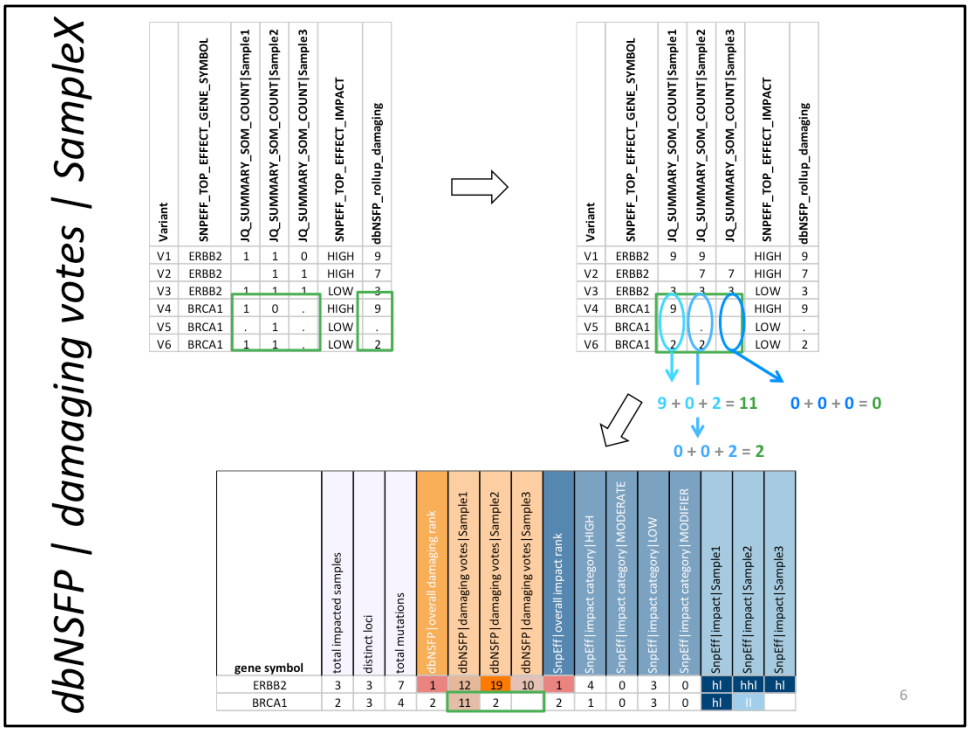
For each gene, the count of passed samples is calculated. In this example, the *JQ_SUMMARY_SOM_COUNT* prefix depicts the columns for each sample.



- 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.



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



- 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 | 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

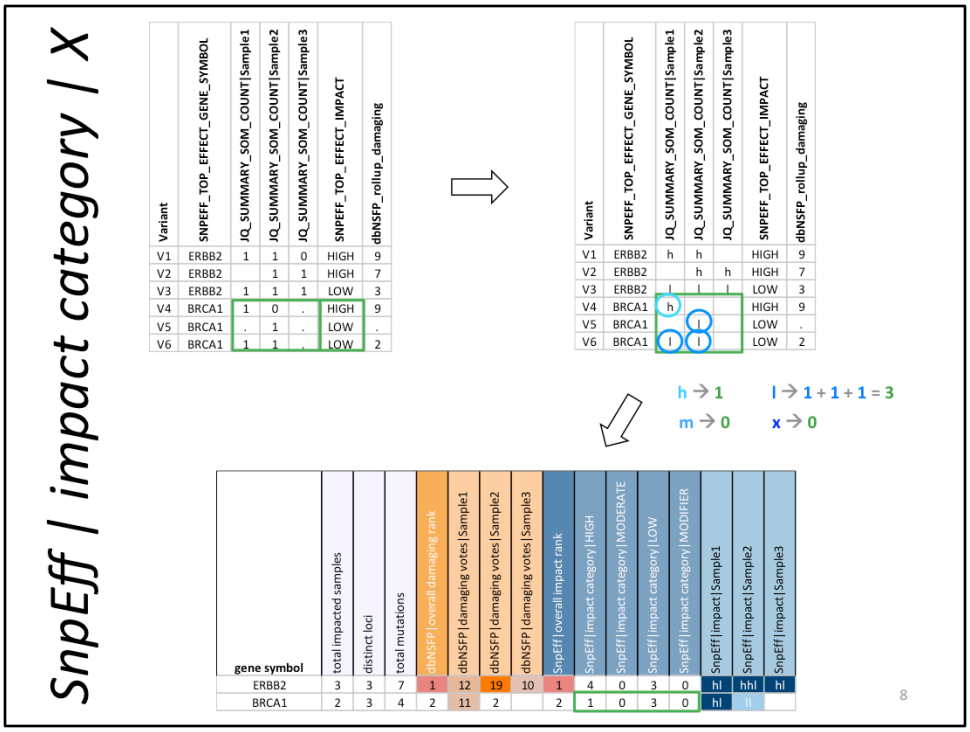
11 + 2 + 0 = 13

41 > 13

1. 41
2. 13

7

- 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.



- 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 | 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	

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

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

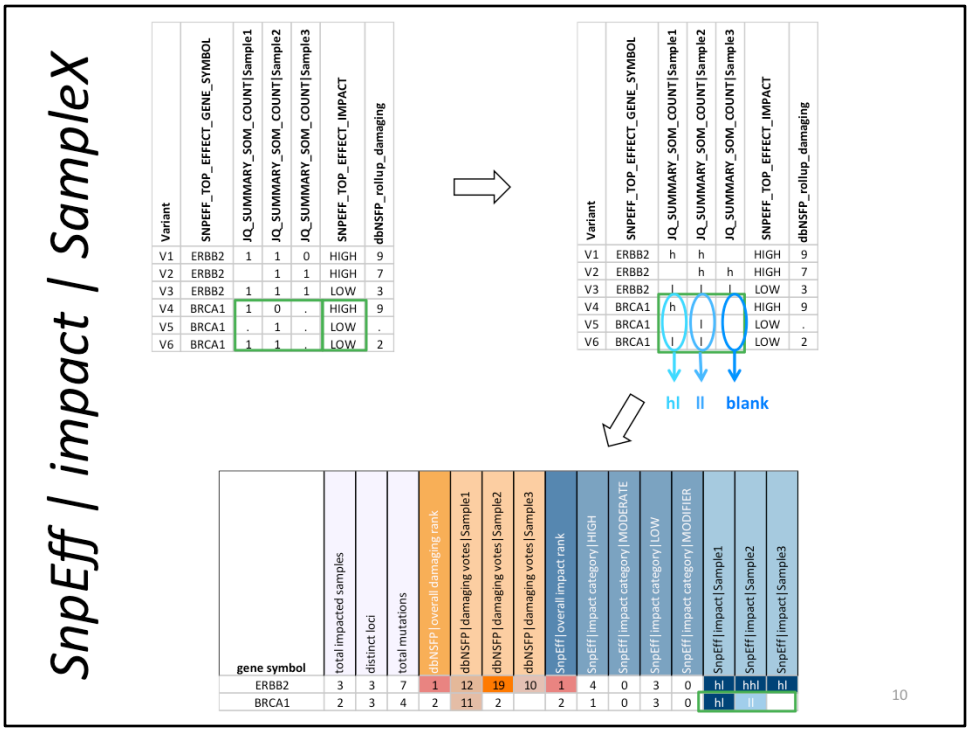
$w_1 = 10^9$
$w_2 = 10^6$
$w_3 = 10^3$
$w_4 = 10^0$

4000003000 > 1000003000

1. 4000003000
2. 1000003000

9

- 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.



- 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.