SVI Results

I. Sample Usecase

In this paper, we aimed to investigate with two different usecases: Simple Variant Interpretation (SVI) and Freebayes Variant Calling. Both usecases are tested for ”Effect Analysis” and ”Cause Analysis” which explains how different the results are and what is the cause for the difference. The investigations of SVI workflows are representated in tables whereas the investigations of Freebayes are visulaised in graph format.

A. Usecase 1 - Simple Variant Interpretation

The Simple Variant Interpretation (SVI), process was implemented as part of Re-Comp project (recomp.org.uk). SVI is a workflow which takes advantage of two external data sources OMIM GeneMap and NCBI ClinVar to provide interpretation of human variants to facilitate clinical diagnosis of genetic diseases say for example, Alzheimer’s disease. We have observed above stated Why-Diff scenario in SVI Workflow. The overall structure of the SVI Workflow is depicted in the figure 3:

As per our terminologies, $\{X, D, W\}$ are the set of artefacts in the execution of Workflow which resulted in Y, where X are the set of inputs [x_1, x_2, x_3,...], D are the set of dependencies [d_1,d_2,...], W are set of activities [w_1, w_2,..], Y are set of outputs [y_1,y_2,..] . To give a common picture, the two external databases OMIM GeneMap and NCBI ClinVar are considered as $d1$ and $d2$ of the SVI workflow W which has set of activity blocks [w_1, w_2, w_3,..] which has either generated an entity say x_1 or used an entity say x_2 where x_1 and x_2 belongs to X. There are totally 14 “entity”, 11 “activity” Neo4j nodes and 11 “generated” and 13 “used” relationships are used to visualise a single SVI invocation. The Table III best describes our analogy for the SVI Workflow.

Terminologies	SVI Analogies
X= $\{x_1\}$ (Input)	x_1 = Patient Input
D = $\{d_1, d_2\}$ (Dependencies)	d_1 = Genemap, d_2 = Clinvar
W = $\{w_1, w_2, \dots\}$ (Workflow)	Number of Activity W = $\{w_1, w_2, \dots, w_{11}\}$
Y = $\{y_1\}$ (Output)	y_1 = svi_classification

TABLE I: Mapping terminologies for SVI Workflow

<table>
<tr><td colspan="5"><W,X,D’> - “Varying Dependency keeping Workflow, input unchanged”</td></tr>
<tr><td colspan="5">d_1 = “genemap2-160607-esc.txt” and d_2 = “variant_summary-1605.txt” and $d_2{}'$ = “variant_summary-1604.txt”</td></tr>
<tr><td>Invocation Id</td><td>Patient Input (X)</td><td>GeneMap Version</td><td>ClinVar Version</td><td>Total nodes added to Δ Graph</td></tr>
<tr><td>132076</td><td>MUN0785</td><td>d_1</td><td>d_2</td><td></td></tr>
<tr><td>132084</td><td>MUN0785</td><td>d_1</td><td>$d_2{}'$</td><td>3</td></tr>
<tr><td colspan="5">where d_1 = “genemap2-svi-161130-161102.csv” and $d_1{}'$ = “genemap2-161031-esc.txt” and d_2 = “variant_summary-1604.txt”</td></tr>
<tr><td>130409</td><td>B_0198</td><td>d_1</td><td>d_2</td><td></td></tr>
<tr><td>13903</td><td>B_0198</td><td>$d_1{}'$</td><td>d_2</td><td>9</td></tr>
</table>

TABLE II: Varying Dependency

Table IV, V and VI describes the evaluation result of the 3 scenarios $\{W,\ X,\ D'\}$, $\{W',\ X,\ D\}$, $\{W,\ X',\ D\}$ respectively.

1) Varying dependency: In the Table IV, Comparisons are made between invocations 132076 and 132084 as well as 130409 and 13903. In the first case, both the invocations 132076 and 132084 have used the same Patient input (i.e. MUN0785) as well as same Genemap dependency. However, the invocation 132076 has used dependency “variant_summary-1605.txt” whereas 132084 has used dependency “variant_summary-1604.txt”, which leads to different output Y'. The total number of divergent nodes added to the Δ Graph is 3 sets (3 from the two invocations) including the transient data passed inside workflow pipeline. In the second case, both the invocations 130409 and 13903 have used the same Patient input (i.e. B_0198) as well as same ClinVar dependency. However, the invocation 130409 has used dependency “genemap2-svi-161130-161102.csv” where as 13903 has used dependency “genemap2-161031-esc.txt”, which leads to different output Y'. The total number of divergent nodes added to the Δ Graph is 9 sets (9 from the two invocations) including the transient data passed inside workflow pipeline. Out of the 2 comparisons we made, first one records 3 sets of divergent nodes and second one records 9 sets of divergent nodes. The reason is that changing the ClinVar dependency from “variant_summary-

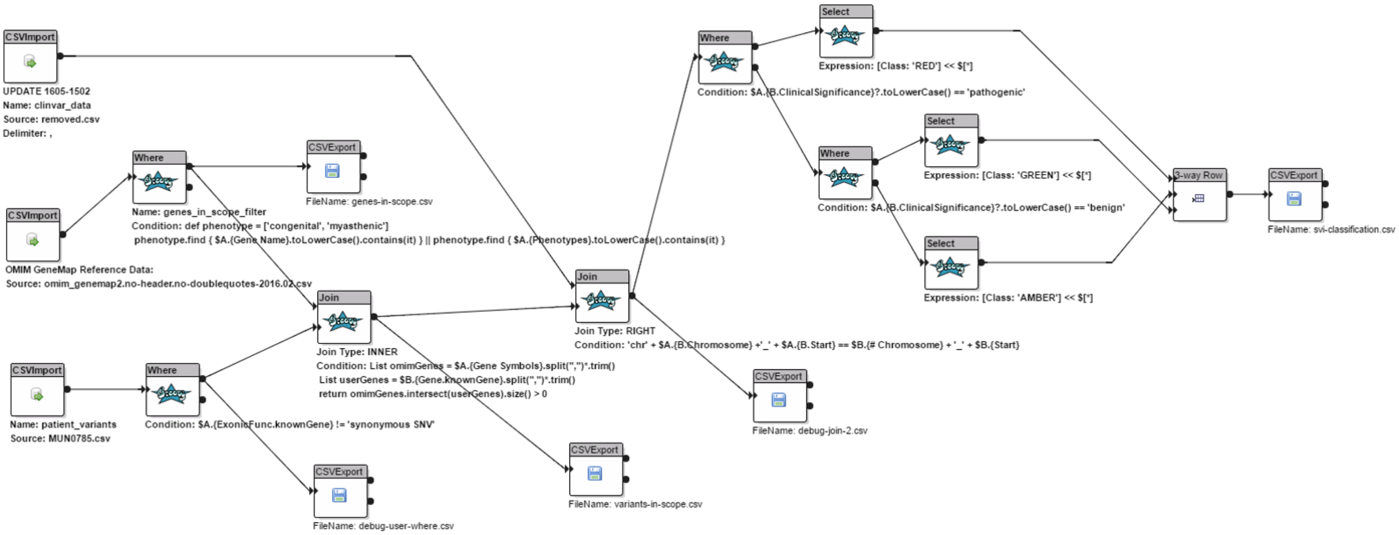
CSVImport
UPDATE 1605-1502
Name: clinvar_data
Source: removed.csv
Delimiter: ,
Where
Name: genes_in_scope_filter
Condition: def phenotype = ['congenital', 'myasthenic']
phenotype.find { $A.{Gene Name}.toLowerCase().contains(it) } || phenotype.find { $A.{Phenotypes}.toLowerCase().contains(it) }
CSVExport
FileName: genes-in-scope.csv
CSVImport
OMIM GeneMap Reference Data:
Source: omim_genemap2.no-header.no-doublequotes-2016.02.csv
Join
Join Type: INNER
Condition: List omimGenes = $A.{Gene Symbols}.split(",")*.trim()
List userGenes = $B.{Gene.knownGene}.split(",")*.trim()
return omimGenes.intersect(userGenes).size() > 0
CSVImport
Name: patient_variants
Source: MUN0785.csv
Where
Condition: $A.{ExonicFunc.knownGene} != 'synonymous SNV'
CSVExport
FileName: debug-user-where.csv
CSVExport
FileName: variants-in-scope.csv
Join
Join Type: RIGHT
Condition: 'chr' + $A.{B.Chromosome} +'_' + $A.{B.Start} == $B.{# Chromosome} + '_' + $B.{Start}
CSVExport
FileName: debug-join-2.csv
Where
Condition: $A.{B.ClinicalSignificance}?.toLowerCase() == 'pathogenic'
Select
Expression: [Class: 'RED'] << $[*]
Where
Condition: $A.{B.ClinicalSignificance}?.toLowerCase() == 'benign'
Select
Expression: [Class: 'GREEN'] << $[*]
Select
Expression: [Class: 'AMBER'] << $[*]
3-way Row
CSVExport
FileName: svi-classification.csv

Fig. 1: SVI Workflow

1605.txt” to “variant_summary-1604.txt” does not make much difference compared to the GeneMap dependency change (ie. from “genemap2-svi-161130-161102.csv” to “genemap2-161031-esc.txt”). Hence in this particular case, the impact of ’GeneMap’ variation is more than the ClinVar variation.

<W’,X,D> - “Varying Workflow keeping Dependency, input unchanged”			
w_7 = “Source:variant_summary-1605.txt” and $w_7{'}$ = “Source:variant_summary-1604.txt”			
Invocation Id	Patient Input (X)	Activity Version	Total nodes added to Δ Graph
132076	MUN0785	w_7	
132084	MUN0785	$w_7{'}$	1
where w_3 = “Source:genemap2-svi-161130-161102.csv” and $w_3{'}$ = “Source:genemap2-161031-esc.txt”			
130409	B_0198	w_3	
13903	B_0198	$w_3{'}$	1

TABLE III: Varying Workflow

2) Varying Workflow: In the Table V, Comparisons are made between invocations 132076 and 132084 as well as 130409 and 13903. In the first case, both the invocations 132076 and 132084 have used the same Patient input (i.e. MUN0785). However, the invocation 132076 has an activity w$_7$ (seventh Workflow block) with the configuration “Source:variant_summary-1605.txt” whereas 132084 has the activity $w_7{'}$ (variation of w$_7$) with the configuration “Source:variant_summary-1604.txt”. The total number of divergent nodes added to the Δ Graph is 1 set of activity node (1 from each invocation). In the second case, both the invocations 130409 and 13903 have used the same Patient input (i.e. B_0198.entire). However, the invocation 130409 has an activity w$_3$ (third Workflow block) with the configuration “Source:genemap2-svi-161130-161102.csv” whereas 13903 has the activity $w_3{'}$ (variation of w$_3$) with the configuration “Source:genemap2-161031-esc.txt” which leads to different output Y'. The total number of divergent nodes added to the Δ Graph is 1 set of activity node (1 from each invocation). The reason behind the different output from the two comparison is because of the dependency variations hence it has recorded 1 divergent activity set. The activity configuration is associated with many properties out of which the ’Source’ label indicates which paticular input has been used by that activity. Apart from the “Source” label, the activity variation could be detected from the sets of libraries used,activity version as these details are collected by the eScience WfMS.

<W,X’,D> - “Varying Input keeping Dependency, Workflow unchanged”		
X = “MUN0785.csv” and X' = “MUN1000.csv”		
Invocation Id	Patient Input (X)	Total nodes added to Δ Graph
720	MUN0785	
867	MUN1000	11

TABLE IV: Varying Input

3) Varying Input: In the Table VI, Comparisons are made between invocations 720 and 867. Both invocations 720 and 867 have used different inputs “MUN0785” and “MUN1000” respectively. But both have used the Workflow blocks and same GeneMap and ClinVar dependencies. This comparison has recorded 11 sets of divergent “entity” nodes out of total 14 “entity” nodes. The other 3 entity nodes which are excluded form the divergent set are GeneMap, ClinVar and ClinVar imported Transient data. As there is change in the input X, we can expect large set of divergent entity nodes.

So far, we have oberserved differences due to a change in only one artefact (which could be input, dependency or workflow). However, there could be multiple artefacts responsible for different output. The multiple reasons for the output difference can be found by observing the Δ Graph as it records all the diverging nodes. We have conducted experiments on varied datasets. The entites comparison are at file label level. We intend to find the difference between two entites at file-content level when the corresponding entities are different.