Informe Técnico / Technical Report




DSL Development with Geneticists

Maria Jose Villanueva, Francisco Valverde, Oscar Pastor





Ref. #:	ProS-TR-XXXX				
Title:	DSL Development with Geneticists				
Author (s):	Maria Jose Villanueva, Francisco Valverde, Oscar Pastor				
Corresponding author (s):	mvillanueva@pros.upv.es fvalverde@pros.upv.es opastor@dsic.upv.es				
Document version number:		Final version:		Pages:	
Release date:					
Key words:					

DSL DEVELOPMENT WITH GENETICISTS

Authors (in alphabetical order):

Francisco Valverde and Maria José Villanueva

Universidad Politécnica de Valencia · Camino de Vera s/n · Edificio 1F · 46022 Valencia Spain ·
T. +34 96 387 70 07 Ext. 83530 · M. + 34 619 543 623 · F. +34 96 387 73 59 · info@pros.upv.es · www.pros.upv.es

Introduction

This document explains in detail the artefacts created in the different stages of DSL development. The general structure for each iteration is:

1. **Analysis:**
    a. Usage scenarios, User Stories, and Acceptance Tests
    b. Domain model: Feature model, Conceptual Model and Glossary of terms
2. **Design**:
    a. Abstract Syntax metamodel
    b. Concrete Syntax
    c. Semantic Restrictions
    d. Behavioural semantics
3. **Implementation**:
    a. Tests
    b. Generator rules
4. **Testing**:
    a. Questionnaire
    b. Demonstration screenshots

Further updates of this report can be asked to mvillanueva@pros.upv.es

UNIVERSIDAD POLITECNICA DE VALENCIA

PROS
Centro de Investigación en Métodos de Producción de Software

Iteration 1

Basics of Genetic Disease Diagnosis

In this iteration, we collaborated with IMEGEN to build the method and to apply it to get a preliminary version of the DSL

Universidad Politécnica de Valencia · Camino de Vera s/n · Edificio 1F · 46022 Valencia Spain ·
T. +34 96 387 70 07 Ext. 83530 · M. + 34 619 543 623 · F. +34 96 387 73 59 · info@pros.upv.es · www.pros.upv.es

Analysis

Usage Scenario Lactase Persistence

In order to diagnose the **Lactase Persistence disease**, I want to read the patient variations from a VCF file, annotate the variations with their Hgvs Notation, search the variations in HgvsDna NC_000002.11:g.136608646G>A and NC_000002.11:g.136616754C>A, and create a report with the variations found with the variations main properties and their hgvs notations.

Usage Scenario Alkaptonuria

In order to diagnose the **Alkaptonuria disease**, I want to read the patient variations from a VCF file, annotate the variations with their Hgvs Notation, search the variations in HgvsCoding NM_000187.3:c.688C>T (NC_000003.11:g.120363252G>A), NM_000187.3:c.899T>G (NC_000003.11:g.120357409A>C), NM_000187.3:c.174delA (NC_000003.11:g.120393750delT), NM_000187.3:c.16-1G>A (NC_000003.11:g.120394711C>T), NM_000187.3:c.342+1G>A (NC_000003.11:g.120371438C>T), NM_000187.3:c.140C>T (NC_000003.11:g.120393784G>A), and create a report with the variations found with the variations main properties and their hgvs notations.

Usage Scenario Achondroplasia Coding

In order to diagnose the **Achondroplasia disease**, I want to read the patient variations from a VCF file, annotate the variations with their Hgvs Notation, search the variations in HgvsCoding NM_000142.4:c.1123G>T (NC_000004.11:g.1806104G>T) NM_000142.4:c.1138G>A (NC_000004.11:g.1806119G>A) NM_000142.4:c.1138G>C (NC_000004.11:g.1806119G>C), and create a report with the variations found with the variations main properties and their hgvs notations.

Usage Scenario Achondroplasia Protein

In order to diagnose the **Achondroplasia disease**, I want to read the patient variations from a VCF file, annotate the variations with their Hgvs Notation, search the variations in HgvsProtein NP_000133.1:p.Gly375Cys NP_000133.1:p.Gly380Arg, and create a report with the variations found with the variations main properties and their hgvs notations.

User Story 1 Read patient's variations from a VCF File (PatientData)

	Role	Context	Action	Goal
US	**"As a geneticist, I want to read a patient's variations from a VCF file, so that I can analyse potential genetic diseases"**			
	Geneticist	File Path	Read Variations from VCF file	Variations data
AT1	**"When I choose the file 3Variations.vcf, I will see that the variations are: chr2:g.136438366A>G, chr11:g.111959693G>T and chr17:g.41245471C>T"**			
	Geneticist	File 3Variations.vcf	Read Variations from VCF file	chr2:g.136438366A>G chr11:g.111959693G>T chr17:g.41245471C>T
AT2	**"When I choose the file incomplete.vcf, I will see an error saying that the file has not all the required data columns"**			
	wrong.vcf	File wrong.vcf	Read Variations from VCF file	Error: "The file has not all the required data columns"
AT3	**"When I choose the file positionOutside.vcf, I will see an error saying that the file**			

	contains a variation whose position is outside the chromosome"			
	Tester	chr2:g.100000000000A>G	Read Variations from VCF file	Error: "The file contains a variation whose position is outside the chromosome"
AT4	**"When I choose the file wrong.sam I will see an error saying that the file has a wrong format"**			
	Tester	File variations.sam	Read Variations from VCF file	Error: "Wrong format"

DSL User role

	Role	Context	Action	Goal
US	**"As a DSL user, I want to specify the file path of a VCF file so that genetic variations can be read from it"**			
	DSL user	-	Specify file path	Variations data
AT1	**"When I write the path C:/Files/3Variations.vcf, variations will be read from the fiel 3Variations.vcf**			
	DSL user	-	Specify C:/Files/3Variations.vcf	Read from 3Variations.vcf

User Story 2: Annotate Variations with HGVS (Analysis)

	Role	Context	Action	Goal
US	**"As a geneticist, I want to annotate the patients' variations with the HGVS notation, so that I can see the change at the DNA, Coding and Protein level of each patient's variation expressed using a standard notation.**			
	Geneticist	A set of Variations	Annotate Variations with Hgvs	DNA level: Chromosome and position Coding level: Transcript and position Protein level: Identifier and aminoacid change
AT1	**"When I annotate the variation chr2:g.136438366A>G, chr11:g.111959693G>T with their Hgvs, I will see chr2:g.136438366A>G and a message saying that the variation is an intron"**			
	Geneticist	chr2:g.136438366A>G	Annotate Variations with Hgvs	chr2:g.136438366A>G Message: "The variation is an intron"
AT2	**"When I annotate the variation chr11:g.111959693G>T with their Hgvs, I will see chr11:g.111959693G>T, NM_003002.2:c.272G>T, NM_001276504.1: c.272G>T, NM_001276506.1:c.155G>T and NP_001263433.1:p.Arg52Met, NP_001263435.1:p.Arg91Met, NP_002993.1:p.Arg91Met"**			
	Geneticist	chr11:g.111959693G>T	Annotate Variations with Hgvs	chr11:g.111959693G>T NM_003002.3:c.272G>T, NM_001276504.1: c.155G>T

				NM_001276506.1: c.272G>T, NP_001263433.1:p.Arg52Met, NP_001263435.1:p.Arg91Met, NP_002993.1:p.Arg91Met
AT3	**"When I annotate the variation chr1:g.13211293delTC with the HGVS , I will see chr1:g.13211293delTC and a message saying that the variation is intergenic"**			
	Geneticist	chr1:g.13211293delTC	Annotate Variations with HgvsCoding	Message: "The variation is intergenic"

DSL User role

	Role	Context	Action	Goal
US	**"As a DSL user, I want to order the annotation of the patient's variations with the HGVS notation, so that the patient variations will be annotated with the hgvs notation."**			
	DSL user	-	Order annotate variations with hgvs	Variations will be annotated with hgvs
AT1	**"When I order the annotation of the patient variations, the patient variations will be annotated with the hgvs notation.**			
	DSL user	-	Order annotate variations with hgvs	Annotate variations with hgvs
AT2	**"When I order the annotation of the patient variations and variations have not been read, I will see an error saying that variations must be read before annotating"**			
	DSL user	Variations not read	Order annotate variations with hgvs	Error: "Variations must be read before annotating"

User Story 3: Search Variations in HGVSDna (Analysis)

	Role	Context	Action	Goal
US	**"As a geneticist, I want to search a set of variations in HGVSDna in the patient's variations, so that I can focus on the suitable variations for the diagnosis."**[1] [2]			
	Geneticist	A set of Variations	Search a set of variations in HGVSDna	Chromosome and genomic position of variations found
AT1	**"When I search in the patient's variations chr2:g.6438366A>G, chr11:g.111959693G>T, chr17:g.41245471C>T and chr17:g.41256103G>A the variations in HGVSDna NC_000017.10:g.41245471C>T and NC_000017.10:g.41256103G>A, I will see the variations chr17:g.41245471C>T and chr17:g.41256103G>A**			

[1] Assuming (at the moment) that variations and hgvs notation have the same reference and DSL users/geneticists control this fact.

[2] Assuming that the reference of the Hgvs variations is RefSeq (http://www.ncbi.nlm.nih.gov/refseq/)

	Geneticist	chr2:g.136438366A>G chr11:g.111959693G>T chr17:g.41245471C>T chr17:g.41256103G>A	Search Variations in hgvs dna: NC_000017.10:g.41245471C>T, NC_000017.10:g.41256103G>A	chr17:g.41245471C>T chr17:g.41256103G>A
AT2	**"When I search in the patient's variations chr2:g.6438366A>G, chr11:g.111959693G>T, chr17:g.41245471C>T and chr17:g.41256103G>A the variation in HGVSDna NC_000002.11:g.136608646G>A, I will see a message saying that none variation has been found"**			
	Geneticist	chr2:g.6438366A>G chr11:g.111959693G>T chr17:g.41245471C>T chr17:g.41256103G>A	Search Variations in hgvs dna NC_000002.11:g.136608646G>A	Message: "None variation has been found"

DSL User role

	Role	Context	Action	Goal
US	**"As a DSL user, I want to write the set of variations expressed in HGVS Dna, so that they can be search among the patient variations."**			
	DSL user	-	Write a set of variations in hgvs dna notation	Search a set of variations written in hgvs dna among the patient variations
AT1	**"When I write the hgvs dna variations set: "NC_000017.10:g.41245471C>T, NC_000017.10:g.41256103G>A", these will be searched among the patient variations.**			
	DSL user	-	Write NC_000017.10:g.41245471C>T, NC_000017.10:g.41256103G>A	Search Variations hgvs dna: NC_000017.10:g.41245471C>T, NC_000017.10:g.41256103G>A
AT2	**"When I write the variation in HGVSDna NC_000002.11: g.1234AC I will see an error saying that the HGVSDna notation is not correctly written"**			
	DSL user	-	Write NC_000002.11: g.1234AC	Error: "The HGVSDna notation is not correctly written"
AT3	**"When I write the variation in HGVS Dna NP_000235.3: p.Gly380Arg I will see an error saying that the HGVS notation is not expressed in DNA nomenclature"**			
	DSL user	-	Write: NC_000002.11: g.1234A>C	Error: "The HGVS notation is not expressed in DNA nomenclature"
AT4	**"When I write any variation in HGVS Dna and variations are not annotated with hgvs I will see an error saying that you should annotate the HGVS notation before searching by HGVS"**			
	DSL user	Variations not annotated with hgvs	Write any variation	Error: "You should annotate the HGVS notation before searching by HGVS"

User Story 4: Search Variations in HGVSCoding (Analysis)

	Role	Context	Action	Goal
US	**"I want to search a set of variations in HGVSCoding in the patient's variations, so that I can focus on the suitable variations only."**			
	Geneticist	A set of Variations	Search a set of variations in HGVSCoding	Transcript and coding position of variations found
AT1	**"When I search in the patient's variations chr2:g.6438366A>G, chr11:g.111959693G>T, chr17:g.41245471C>T and chr17:g.41256103G>A the variations in HGVSCoding NM_003002.2:c.272G>T, I will see the variation chr2:g.6438366A>G**			
	Geneticist	chr2:g.136438366A>G chr11:g.111959693G>T chr17:g.41245471C>T chr17:g.41256103G>A	Search Variations in hgvs coding NM_003002.3:c.272G>T	NM_003002.2:c.272G>T
AT2	**"When I search in the patient's variations chr2:g.6438366A>G, chr17:g.41245471C>T and chr17:g.41256103G>A the variation in HGVS Coding NM_015361.2:c.1003+20A>G, I will see a message saying that none variation has been found"**			
	Geneticist	chr2:g.6438366A>G chr17:g.41245471C>T chr17:g.41256103G>A	Search Variations in hgvs coding NM_003002.3:c.272G>T	Message: "None variation has been found"

DSL User role

	Role	Context	Action	Goal
US	**"As a DSL user, I want to write the set of variations expressed in HGVS Coding, so that they can be search among the patient variations."**			
	DSL user	-	Write a set of variations in hgvs coding notation	Search a set of variations written in hgvs coding among the patient variations
AT1	**"When I write the HGVS Coding variations set: "NM_003002.2:c.272G>T", these will be searched among the patient variations.**			
	DSL user	-	Write NM_003002.2:c.272G>T	Search Variations hgvs coding: NM_003002.2:c.272G>T
AT2	**"When I write the HGVS Coding variations set: NM_003002.2:c.272GT I will see an error saying that the HGVS Coding notation is not correctly written"**			
	DSL user	-	Search Variations in hgvs coding NM_003002.2:c.272GT	Error: "The HGVSCoding notation is not correctly written"
AT3	**"When I write the HGVS Coding variations set: NC_000002.11: g.1234A>C I will see an error saying that the HGVS notation is not expressed in Coding nomenclature"**			
	DSL user	-	Search: NC_000002.11: g.1234A>C	Error: "The HGVS notation is not expressed in Coding nomenclature"

AT4	**"When I write any variation in HGVS Coding and variations are not annotated I will see an error saying that you should annotate the HGVS notation before searching by HGVS"**			
	DSL user	Variations not annotated	Write any variation	Error: "You should annotate the HGVS notation before searching by HGVS"

User Story 5: Search Variations in HGVSProtein (Analysis)

	Role	Context	Action	Goal
US	**"I want to search a set of variations in HGVSProtein in the patient's variations, so that I can focus on the suitable variations only."**			
	Geneticist	A set of Variations	Search a set of variations in HGVSProtein	Focus on the suitable variations
AT1	**"When I search in the patient's variations chr2:g.6438366A>G, chr11:g.111959693G>T, chr17:g.41245471C>T and chr17:g.41244435T>C the variations in HGVSProtein NP_002993.1:p.Arg91Met, I will see the variation chr2:g.6438366A>G**			
	Geneticist	chr2:g.136438366A>G chr11:g.111959693G>T chr17:g.41245471C>T chr17:g.41256103G>A	Search variations in hgvs protein NP_002993.1:p.Arg91Met	chr11:g.111959693G>T
AT2	**"When I search in the patient's variations chr2:g.6438366A>G, chr11:g.111959693G>T, chr17:g.41245471C>T and chr17:g.41244435T>C the variation in HGVSProtein NP_002993.1:p.Arg30Lys, I will see a message saying that none variation has been found"**			
	Geneticist	chr2:g.136438366A>G chr11:g.111959693G>T chr17:g.41245471C>T chr17:g.41256103G>A	Search variations in hgvs protein NP_002993.1:p.Arg30Lys	Message: "None variation has been found"

DSL User role

	Role	Context	Action	Goal
US	**"As a DSL user, I want to write the set of variations expressed in HGVS Protein, so that they can be search among the patient variations."**			
	DSL user	-	Write a set of variations in hgvs protein notation	Search a set of variations written in hgvs protein among the patient variations
AT1	**"When I write the HGVS Protein variations set: "NM_003002.2:c.272G>T", these will be searched among the patient variations.**			
	DSL user	-	Write NP_002993.1:p.Arg91Met	Search Variations hgvs protein: NP_002993.1:p.Arg91Met
AT2	**"When I write the HGVS Protein variations set: NP_002993.1:c.Arg91Met I will see an error saying that the HGVS Protein notation is not correctly written"**			

	DSL user	-	Search Variations in hgvs coding NP_002993.1:c.Arg91Met	Error: "The HGVS Protein notation is not correctly written"
AT3	**"When I write the HGVS Protein variations set: NC_000002.11: g.1234A>C I will see an error saying that the HGVS notation is not expressed in Protein nomenclature"**			
	DSL user	-	Search: NC_000002.11: g.1234A>C	Error: "The HGVS notation is not expressed in Protein nomenclature"
AT4	**"When I write any variation in HGVS Protein and variations are not annotated I will see an error saying that you should annotate the HGVS notation before searching by HGVS"**			
	DSL user	Variations not annotated	Write any variation	Error: "You should annotate the HGVS notation before searching by HGVS"

User Story 6: Report Variations' properties (Report)

	Role	Context	Action	Goal
US	**"I want to create a report with a list of the variations and their main properties (chromosome, position, reference, alternative), so that I can see the main properties of the patient's variations"**			
	Geneticist	A set of Variations	Report chromosome, position, reference and alternative allele(s)	Table with rows for each variation and Chromosome, Position, Reference, Alternative Allele(s) as columns.
AT1	**"When I report the variations chr11:g.111959693G>T and chr17:g.41244435T>C I will see a report with one variation with the chr, position, ref, and alt of each variation"**			
	Geneticist	chr11:g.111959693 G>T	Report chromosome, position, reference and alternative allele(s)	**Chr** / **Pos** / **Ref** / **Alt** 11 / 111959693 / C / T

DSL User role

	Role	Context	Action	Goal
US	**"As a DSL user, I want to order the creation of a report with the patient's variations, so that a report with the diagnosis patient variations main properties chromosome, position, reference and alternative allele(s) is created"**			
	DSL user	-	Order the creation of a variation report	Report chromosome, position, reference and alternative allele(s)
AT1	**"When I order to create a variations report, a report with the diagnosis patient variations with their chromosome, position, reference and alternative allele(s) will be created.**			
	DSL user	-	Order create variation report	Report the diagnosis patient variations and their chromosome, position, reference and alternative allele(s)

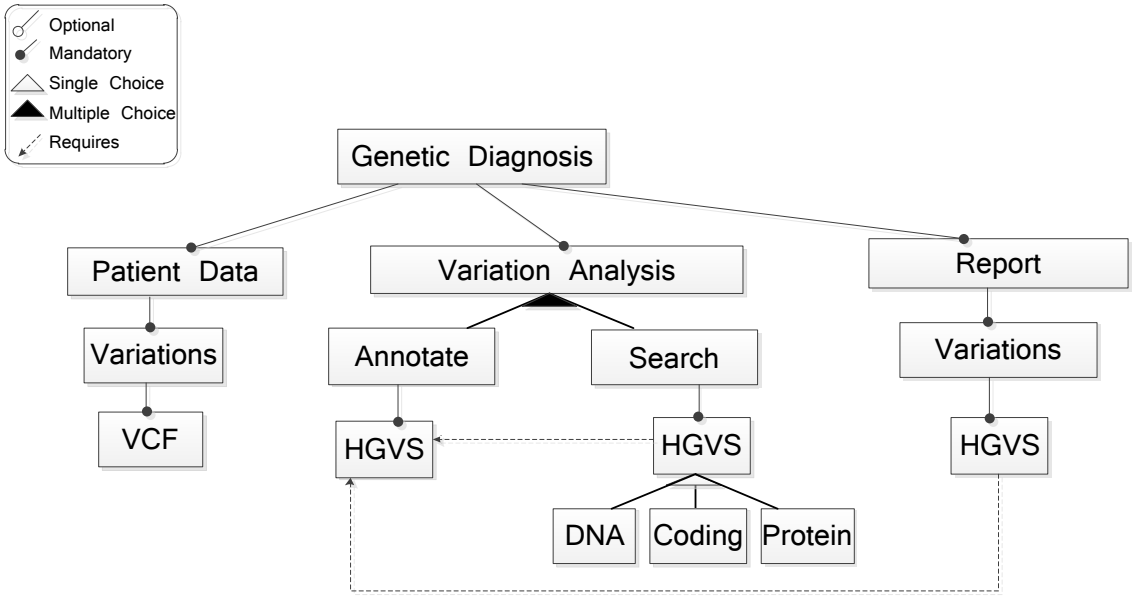
User Story 7: Report Variation's HGVS (Report)

<table>
<tr><th></th><th>Role</th><th>Context</th><th>Action</th><th>Goal</th></tr>
<tr><td>US</td><td colspan="4">"I want to add to a report with the variations their HGVS (Dna, Coding and Protein), so that I can see the patient variation's expressed in a standard notation"</td></tr>
<tr><td></td><td>Geneticist</td><td>Variations and Report</td><td>Report HGVS notation</td><td>New column Hgvs in the variations table.</td></tr>
<tr><td>AT1</td><td colspan="4">"When I report the variations chr11:g.111959693G>T I will see a table with one variation with the chr, position, ref, alt and hgvs"</td></tr>
<tr><td></td><td>Geneticist</td><td>chr11:g.111959693G>T
<table><tr><th>Chr</th><th>Pos</th><th>Ref</th><th>Alt</th></tr><tr><td>11</td><td>111959693</td><td>C</td><td>T</td></tr></table></td><td>Add HGVS notation to variation report</td><td><table><tr><th>Var</th><th>HGVS Dna</th><th>HGVS Coding</th><th>HGVS Protein</th></tr><tr><td>...</td><td>chr11:g.111959693G>T</td><td>NM_003002.2:c.272G>T
NM_001276504.1:c.272G>T,
NM_001276506.1:c.155G>T</td><td>NP_001263433.1:p.Arg52Met,
NP_001263435.1:p.Arg91Met,
NP_002993.1:p.Arg91Met</td></tr></table></td></tr>
</table>

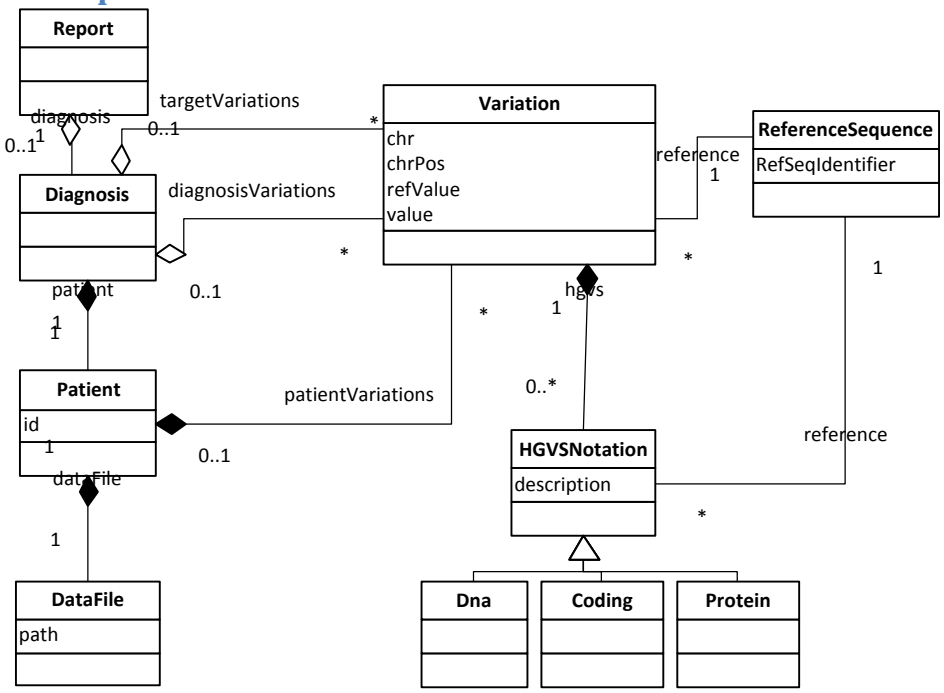
DSL User role

<table>
<tr><th></th><th>Role</th><th>Context</th><th>Action</th><th>Goal</th></tr>
<tr><td>US</td><td colspan="4">"As a DSL user, I want to order the report of the HGVS notation, so that the HGVS notation is added to the patient's variation report.</td></tr>
<tr><td></td><td>DSL user</td><td>-</td><td>Order the creation of a variation report with the hgvs notation</td><td>Report hgvs notation</td></tr>
<tr><td>AT1</td><td colspan="4">"When I order to create a variations report with the hgvs notation, the hgvs notation will be added to the variation report</td></tr>
<tr><td></td><td>DSL user</td><td>-</td><td>Order report hgvs notation</td><td>Add hgvs notation to variation report</td></tr>
<tr><td>AT2</td><td colspan="4">"When I order to create a variations report with the hgvs notation and variations are not annotated with the hgvs notation I will see an error saying that you should annotate the hgvs notation before reporting hgvs notation"</td></tr>
<tr><td></td><td>DSL user</td><td>Variations not annotated with hgvs notation</td><td>Order report hgvs notation</td><td>Error: "You should annotate the hgvs notation before reporting hgvs notation"</td></tr>
</table>

Feature Model



Conceptual Model



Glossary

Report: Gathering of relevant information to show to geneticists/clinicians the result of a genetic disease diagnosis.

Diagnosis: Analysis that is performed to a patient to find out if the genotype indicates a potential risk to have a genetic disease.

Patient: Object of study to perform the diagnosis.

Datafile: Set of data saved in a file system.

Variation: Each of the nucleotides that a patient has different in regards to a reference sequence.

Reference Sequence: A representative sequence of nucleotides that theoretically represents the sequence of a “disease free” human.

HGVS Notation: Standard nomenclature the describe variations

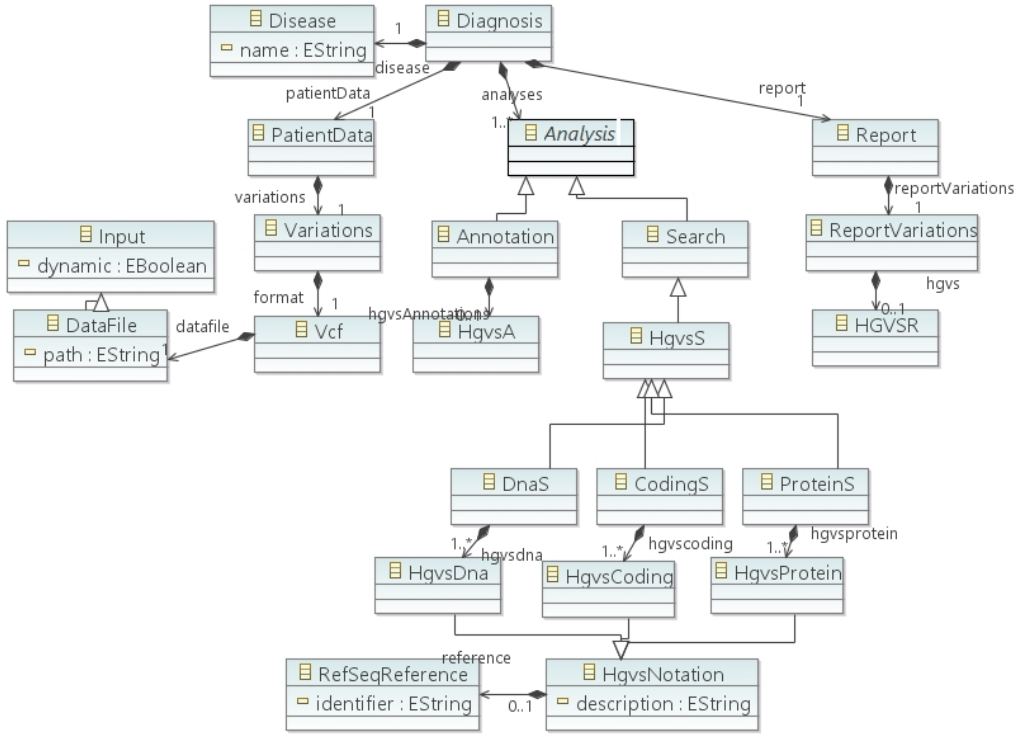
(HGVS Notation) DNA: HGVS Nomenclature that represents the value of the variation at nucleotide level.

(HGVS Notation) Coding: HGVS Nomenclature that represents the value of the variation at the coding level.

(HGVS Notation) Protein: HGVS Nomenclature that represents the value of the variation at the amino acid level.

Design

Abstract Syntax



Syntax Questionnaire

https://drive.google.com/open?id=1O4xaBuPfijHOK-LJ6Ivg4TGDjLRcI_JMFOZ57fZ8SZg

Syntax Examples

Take a look to the next 4 possible syntaxes to describe the usage scenario "Diagnose of Diabetes Mellitus type II" previously described

*Obligatorio

Syntax 1

Diagnosis: Achondroplasia
Variations VCF file: Patient1.vcf
Variations Annotations: hgvs
Analysis Search: by coding {NM_000142.4:c.1123G>T, NM_000142.4:c.1138G>A, NM_000142.4:c.1138G>C}
Variation report fields: hgvs

Give your opinion from 1 to 5 (being 1 the lowest rate and 5 the highest) about this syntax *

	1 (I don't like it at all)	2 (I don't like it)	3 (Neutral)	4 (It's ok)	5 (I like it a lot)
My opinion about Syntax 1	○	○	○	○	○

Syntax 2

```
Diagnose Achondroplasia
Read Variation's from VCF file Patient1.vcf
Annotate Variations with hgvs
Search Variations by coding {NM_000142.4:c.1123G>T, NM_000142.4:c.1138G>A,
NM_000142.4:c.1138G>C}
Report Variations with hgvs
```

Give your opinion from 1 to 5 (being 1 the lowest rate and 5 the highest) about this syntax *

	1 (I don't like it at all)	2 (I don't like it)	3 (Neutral)	4 (It's ok)	5 (I like it a lot)
My opinion about Syntax 2	○	○	○	○	○

Syntax 3

```
Diagnosis.Disease("Achondroplasia")
Diagnosis.Patient.Variations. ("Patient1.vcf", VCF)
Diagnosis.Patients.Variations.Annotations (hgvs)
Diagnosis.Patients.Variations.Analysis.Search.ByCoding(NM_000142.4:c.1123G>T,
NM_000142.4:c.1138G>A, NM_000142.4:c.1138G>C)
Diagnosis.Patients.Variations.Report.Fields(hgvs)
```

Give your opinion from 1 to 5 (being 1 the lowest rate and 5 the highest) about this syntax *

	1 (I don't like it at all)	2 (I don't like it)	3 (Neutral)	4 (It's ok)	5 (I like it a lot)
My opinion about Syntax 3	○	○	○	○	○

Syntax 4

```
<Diagnose>
    <Disease>Achondroplasia</Disease>
    <PatientsData>
        <Variations>
            <VCF>Patient1.vcf</VCF>
        </Variations>
    </PatientsData>
    <Analyses>
        <Annotate><hgvs/></Annotate>
        <Search><Coding>
            <variation>NM_000142.4:c.1123G>T</variation>
            <variation>NM_000142.4:c.1138G>A</variation>
            <variation>NM_000142.4:c.1138G>C</variation>
        </Coding></Search>
    </Analyses>
    <Report>
        <Variations>
            <hgvs/>
        </Variations>
    </Report>
</Diagnose>
```

Give your opinion from 1 to 5 (being 1 the lowest rate and 5 the highest) about this syntax *

	1 (I don't like it at all)	2 (I don't like it)	3 (Neutral)	4 (It's ok)	5 (I like it a lot)
My opinion about Syntax 4	○	○	○	○	○

Prefered Syntax

My prefered syntax is... *

- Syntax 1
- Syntax 2
- Syntax 3
- Syntax 4
- I'd like to propose a new one

Concrete Syntax

Complete Grammar

```
// automatically generated by Xtext
grammar diagnosis.it1.mydsl.MyDiag with org.eclipse.xtext.common.Terminals

import "diagnosis"
import "http://www.eclipse.org/emf/2002/Ecore" as ecore

diagnosis returns Diagnosis:
       'Diagnose' disease=disease
        patientData=patientData
     analyses+=analysis+
       report=report;

/*PATIENT DATA */
patientData returns PatientData:
       'Read'
       variations=variations;

variations returns Variations:
       'variations'
       format=vcf;

vcf returns Vcf:
       'from' 'a VCF file'
       datafile=dataFile;

/*ANALYSES */
analysis returns Analysis:
       annotation |
       search;

//Variation Annotation
annotation returns Annotation:
        'Annotate variations'
        'with' hgvs=hgvsA;
//HGVS Annotation
hgvsA returns HgvsA:
       'hgvs'{HgvsA};


//Variations Search
search returns Search:
       'Search' 'variations'
```

```
	hgvsS;
hgvsS returns HgvsS:
	dnaS|
	codingS|
	proteinS;
dnaS returns DnaS:
	hgvsdna+=hgvsdna+;
codingS returns CodingS:
	hgvscoding+=hgvscoding+;
proteinS returns ProteinS:
	hgvsprotein+=hgvsprotein+;


/*REPORT */
report returns Report:
	'Report'
	 reportVariations=reportVariations;
reportVariations returns ReportVariations:
	'variations' {ReportVariations}
	('with' hgvs=hgvsAR)?;
hgvsAR returns HgvsR:
	'hgvs' {HgvsR};


/*DataModel Types */
disease returns Disease:
	name=EString;

dataFile returns DataFile:
	'from'{DataFile}
	(dynamic?=INPUT|path=EString);


hgvsdna returns HgvsDna:
	reference=refSeqReference ':''g.'description=HGVSEXPR;
hgvscoding returns HgvsCoding:
	reference=refSeqReference ':''c.'description=HGVSEXPR;
hgvsprotein returns HgvsProtein:
	reference=refSeqReference ':''p.'description=HGVSEXPR;

refSeqReference returns RefSeqReference:
	identifier=(REFSEQ|ASSEMBLY);



/* Data Types ecore */
EBoolean returns ecore::EBoolean:
	'true' | 'false';
EString returns ecore::EString:
	STRING | ID;
EInt returns ecore::EInt:
	'-'? INT;
```

```
/*Terminals */
terminal INPUT:
      'input';
terminal REFSEQ:
      'N'('C'|'G'|'M'|'P')'_'   INT'.'INT;
terminal ASSEMBLY:
      ('Hg'INT) |
      ('NCBI'INT);
terminal HGVSEXPR:
      (INT(('+'|'-')INT)?('ins'|'del')('A'|'T'|'G'|'C')+)|//ins/del
      (INT(('+'|'-')INT)?('A'|'T'|'G'|'C')+'>'('A'|'T'|'G'|'C')+)|//indel
      (('A'..'Z'|'a'..'z')+INT('A'..'Z'|'a'..'z')+);//Protein
```

Scenarios

```
Diagnose LactoseIntolerance
Read variations from a VCF file from input
Annotate variations with hgvs
Search variations
NC_00002.11:g.136608646G>A
NC_000002.11:g.136616754C>A
Report variations

Diagnose Alkaptonuria
Read variations from a VCF file from input
Annotate variations with hgvs
Search variations
NM_000187.3:c.688C>T
NM_000187.3:c.899T>G
NM_000187.3:c.174delA
NM_000187.3:c.16-1G>A
NM_000187.3:c.342+1G>A
NM_000187.3:c.140C>T
Report variations with hgvs

Diagnose AchondroplasiaCoding
Read variations from a VCF file from input
Annotate variations with hgvs
Search variations
NM_000142.4:c.1123G>T
NM_000142.4:c.1138G>A
NM_000142.4:c.1138G>C
Report Variations with hgvs

Diagnose AchondroplasiaProtein
Read variations from a VCF file from input
Annotate variations with hgvs
Search variations
NP_000133.1:p.Gly375Cys
NP_000133.1:p.Gly380Arg
Report variations with hgvs
```

Semantic restrictions

Syntax Validation

HGVSNotation errors:

1. An EBNF rule that describes the structure of the hgvsnotation

```
terminal HGVSEXPR:
        (INT(('+'|'-')INT)?('ins'|'del')('A'|'T'|'G'|'C')+)|//ins/del
        (INT(('+'|'-')INT)?('A'|'T'|'G'|'C')+'>'('A'|'T'|'G'|'C')+)|//indel
        (('A'..'Z'|'a'..'z')+INT('A'..'Z'|'a'..'z')+);//Protein
```

Behaviour Validation

1. Patient Data must be read before any analysis. Error message provided by geneticists:
    - "Variations must be read before annotating"
2. It is mandatory to always annotate hgvs before search by hgvs and before reporting. Error messages provided by geneticists
    - "You should annotate the HGVS notation before searching by HGVS"
    - "You should annotate the hgvs notation before reporting hgvs notation"

Behavioral semantics

Preliminary Templates

Read the patient's variations from a VCF File

Inputs/Output	Description	CM Entities
Input1	A data file with the file path.	DataFile
Output1	A list of Variations with their main properties and their reference sequence	Patient-> Variation patientVariations ->ReferenceSequence

Tool Name	UploadFile		
Source	Galaxy		
Inputs	**Name**	**Type**	**Mapping/Predefined Value**
Input1	Format	Enumeration	VCF
Input2	File	String	Input1
Input3	Genome	Enumeration	Human GRCH37/hg19 (Output1)
Outputs	**Name**	**Type**	**Mapping/Predefined Value**
Output1	output	VCF File	Output1

Output2	format	Enumeration	VCF
Output3	database	Enumeration	hg19 (Output1)

Annotate Variations with HGVS notation (Dna, Coding, Protein)

Inputs/Output	Description	CM Entities
Input1	A list of variations with their main properties and the reference sequence	Patient->Variation-> ReferenceSequence
Output1	A list of Variations with the HGVSDna, HGVSCoding and HGVSProtein notation (with their associated reference sequence).	Patient->Variation-> Hgvs{DNA,Coding,Protein} ->ReferenceSequence,

Tool Name	SNPEff (**Dna/Coding**/Protein) not use for protein because it does not provide NP identifier		
Source	Galaxy		
Inputs	**Name**	**Type**	**Mapping/Predefined Value**
Input1	input	File	Input1
Input2	inputformat	Enumeration	VCF
Input3	Genome	Enumeration	Hg19
Input4	hgvs	Boolean	True
Input5	outputformat	Enumeration	VCF
Inputs	**Name**	**Type**	**Mapping/Predefined Value**
Output1	output	File (VCF)	Output1

Tool Name	Variant Effect Predictor (Dna/Coding/**Protein**) only used for protein due to integration issues		
Source	Galaxy		
Inputs	**Name**	**Type**	**Mapping**
Input1	input	File	Input1
Input2	species	String	homo_sapiens
Input3	Refseq	Boolean	true
Input4	hgvs	Boolean	true
Input6	outputformat	Enumeration	VCF
Inputs	**Name**	**Type**	**Mapping**
Output1	output	File (VCF)	Output1

Search Variations by HGVS Dna/Coding nomenclature

Inputs/Output	Description	CM Entities
Input1	A list of Variations with the HGVSDna/Coding notation	Patient->Variation->HGVSDna/Coding ->ReferenceSequence
Input2	A list of HGVSDna/Coding notations	Diagnosis->Variation TargetVariations ->HGVSDna/Coding->ReferenceSequence
Output1	A list of Variations with the HGVSDna/Coding notation	Patient->Variation->HGVSDna/Coding ->ReferenceSequence

Tool Name	SNPSift Filter		
Source	Galaxy		
Inputs	**Name**	**Type**	**Mapping**
Input1	input	VCF File	Input1
Input2	expression	String	Input2
Inputs	**Name**	**Type**	**Mapping**
Output1	output	File (VCF)	Output1

Search Variations by HGVS Protein

Inputs/Output	Description	CM Entities
Input1	A list of Variations with the HgvsProtein notation	Patient->Variation->HgvsProtein ->ReferenceSequence
Input2	A list of HgvsProtein notations	Diagnosis->Variation TargetVariations ->HGVSProtein->ReferenceSequence
Output1	A list of Variations with the HgvsProtein notation	Patient->Variation->HgvsProtein ->ReferenceSequence

Tool Name	VEP Filter		
Source	Galaxy (it has been altered to solve errors to filter by hgvsprotein as it did not support the use of ":" in the filter)		
Inputs	**Name**	**Type**	**Mapping**
Input1	input	VCF File	Input1
Input2	filters	String	Input2
Inputs	**Name**	**Type**	**Mapping**
Output1	output	File (VCF)	Output1

Report Variations (main properties)

Inputs/Output	Description	CM Entities
Input1	A list of Variations with their main properties	Diagnosis-> Variation DiagnosisVariations
Output1	A report showing the main properties of each variation	Report->Diagnosis-> Variation DiagnosisVariations

Tool Name	ShowVariations		
Source	Galaxy (New tool)		
Inputs	**Name**	**Type**	**Mapping/Predefined Value**
Input1	File	String	Input1
Outputs	**Name**	**Type**	**Mapping/Predefined Value**
Output1	output	HTML File	Output1

Report Variations with HGVSNotation

Inputs/Output	Description	CM Entities
Input1	A list of Variations with their main properties and their hgvs notation	Diagnosis-> Variation DiagnosisVariations-> HGVS (Dna, Coding, Protein)
Output1	A report showing the main properties of each variation and their hgvs notation	Report->Diagnosis-> Variation DiagnosisVariations -> HGVS (Dna, Coding, Protein)->Reference

Tool Name	ShowVariations		
Source	Galaxy (New tool)		
Inputs	**Name**	**Type**	**Mapping/Predefined Value**
Input1	File	String	Input1
Input2	Dna	Boolean	True
Input3	Coding	Boolean	True
Input4	Protein	Boolean	True
Outputs	**Name**	**Type**	**Mapping/Predefined Value**
Output1	output	HTML File	Output1

Implementation

Semantic Tests

- **The patient data must be read before any analysis**. The following test is used to check this restriction. This restriction is implemented as a EBNF rule that establishes the order of diagnosis steps

Input:
Diagnose LactoseIntoleranceHGVS
//Read Variations genotypes from VCF file Patient1.vcf
Annotate Variations with hgvs
Search Variations NC_000002.11:g.136608646G>A NC_000002.11:g.136616754C>A
Report variations with hgvs
Output:
"Variations must be read before annotating"

- **It is mandatory to always annotate hgvs before search by hgvs and before reporting**. The following test is used to check this restriction. This restriction is implemented in the xtend class of the validator.

Input:
Diagnose LactoseIntoleranceHGVS
Read Variations genotypes from VCF file Patient1.vcf
//Annotate Variations with hgvs
Search Variations NC_000002.11:g.136608646G>A NC_000002.11:g.136616754C>A
Report variations with hgvs
Output:
"You should annotate the HGVS notation before searching by HGVS"
"You should annotate the hgvs notation before reporting hgvs notation"

Target platform tests:

User Story	Test Files	Galaxy Workflow	Tools
Read Variations from VCF File	3Variants.VCF	US1	InputFile and CheckCustomErrors
Annotate HGVS	1IntronVariant.vcf, 1ExonVariant.vcf, 1IntergenicVariant.vcf	US2	Snpeff (dna/coding) and VEP (protein)
Search variations by hgvsDna	4VariantsAnnotated.vcf	US3	SnpSift
Search variations by hgvsCoding	4VariantsAnnotated.vcf 3VariantsAnnotated.vcf	US4	SnpSift
Search variations by	4VariantsAnnotated.vcf	US5	Vep filter

hgvsProtein			
Show the main properties of a variation	1ExonVariant.vcf	US6_US7	ShowVariations
Add HGVS notation to report	4Variants.vcf	US6_US7	ShowVariations

Generator Test Example

Test Annotate HGVS

```
def void testSetupOnce() {
   DiagnosisPackage.eINSTANCE.eClass();
   diagnosis = parser.parse ('''Diagnose LactoseIntoleranceHGVSIt1
   Read variations from a VCF file from input
   Annotate variations with hgvs
   Search variations NC_000002.11:g.136608646G>A
   NC_000002.11:g.136616754C>A
   Report variations with hgvs''')
   fsa= new InMemoryFileSystemAccess()
   generator.doGenerate(diagnosis.eResource, fsa)
}
@Test def testAnnotateHgvs(){
   Assert.assertFalse("The workflow fragment of AnnotateHgvsDNA is
   different to the generated one",
   checkGeneratorGalaxy(fsa.getTextFiles().values(),"AnnotateHgvsD
   na.txt", "vep"))
}
```

Testing

Testing Questionnaire

Did you find any error while using the DSL editor?

Usage scenario: -

Error description: -

Did you detect any error or warning that the DSL editor did not inform? Did you receive any message that you did not understand?

Usage scenario: LactasePersistance

Error description: Name of the disease does not allow to write spaces to separate words. I wanted to write “Lactase Persistance”

Did you find any step missing or that you could not write to specify a usage scenario (diagnosis)? Would you like to add one for the next iteration that you consider important?

Usage scenario (and User Story if applies): -

Step description: -

Had the language all and the suitable elements to write all the information to complete the usage scenario (diagnosis)?

Comments: No, I wanted to specify the genome version of the analysis in order to ensure that the analyses performed (annotation and search) query the suitable databases and report the suitable information. At the moment the genome version is by defect “GRCh37/hg19”. I want to specify this value using the DSL.

Had the language the suitable structure (words order, symbols, indentation, labeling, etc…)?Would you change, add, remove or reorder any word of the language?

Usage scenario: -
Changes proposed: -

Do you know new software that better accomplishes any of the steps in any of the usage scenarios (diagnosis)?

User Story: -
Software information: -

Did you find any error when executing the code generated by the DSL editor (Galaxy Workflow)?

Usage scenario: Alkaptonuria
Error description: Some variations were not detected by the workflow. There were errors in the tests of annotation of hgvs because of the tool SNPEff does not annotate the hgvs notation of the variations:
NC_000003.11:g.120394711C>T
NC_000003.11:g.120371438C>T

UNIVERSIDAD POLITECNICA DE VALENCIA

PROS Centro de Investigación en Métodos de Producción de Software

Iteration 2

Annotations from external data sources

In this iteration, we collaborated with IMEGEN to refine the method and to apply it to get a refined but still preliminary version of the DSL

Universidad Politécnica de Valencia · Camino de Vera s/n · Edificio 1F · 46022 Valencia Spain ·
T. +34 96 387 70 07 Ext. 83530 · M. + 34 619 543 623 · F. +34 96 387 73 59 · info@pros.upv.es · www.pros.upv.es

Analysis

Usage Scenario Mammalian Cancer

In order to diagnose the **Mammalian Cancer disease (Analysis 1)**, I want to read the patient variations from a VCF file, annotate the variations with their Hgvs Notation, annotate the variations with their genes, annotate their variations with their rsId from DbSNP, filter the variations by the genes BRCA1 and BRCA2, and create a report with the variations found with their main properties, their hgvs notations, their genes and their rsIds.

Usage Scenario Mamalian Cancer

In order to diagnose the **Mammalian Cancer disease (Analysis 2)**, I want to read the patient variations from a VCF file, annotate the variations with their Hgvs Notation, annotate the variations with their genes, annotate their variations with their rsId from DbSNP, filter the variations by the gene RAD51C, and create a report with the variations found with their main properties, their hgvs notations, their genes and their rsIds.

User Story 1 Annotate Variations with Gene (Analysis)

	Role	Context	Action	Goal
US	**"I want to annotate the patient variations with the gene names provided by HGNC, so that I can see all the genes involved each variation"**			
	Geneticist	A set of variations	Say "annotate with gene"	Genes involved in each variation
AT1	**"When I annotate the gene in the variations chr2:g.136438366A>G and ch17:g.41245471C>T, I will see that the genes are R3HDM1, { SDHD, C11orf57 and TIMM8B } and BRCA1, respectively"**			
	Geneticist	chr2:g.136438366A>G chr11:g.111959693G>T chr17:g.41245471C>T chr17:g.41256103G>	Annotate Gene	R3HDM1 SDHD, C11orf57 and TIMM8B BRCA1 BRCA1
AT2	**"When I annotate the gene in the variation chr1:g.13211292Del2 I will see an message saying that the variation is not located in a gene"**			
	Geneticist	chr1:g.13211292Del2	Annotate Gene	Message: "Variation is intergenic"

DSL User role

	Role	Context	Action	Goal
US	**"As a DSL user, I want to order the annotation of the patient's variations with the gene, so that the patient variations will be annotated with the gene."**			
	DSL user	-	Order annotate variations with gene	Variations will be annotated with gene
AT1	**"When I order the annotation of the patient variations with the gene, the patient variations will be annotated with the gene notation"**			
	DSL user	-	Order annotate variations with gene	Annotate variations with gene

User Story 2 Annotate Variations with rsId (Analysis)

	Role	Context	Action	Goal
US	**"I want to annotate the patient variations with the rsId from dbSNP, so that I can see if a variation has been identified with an rsId from dbSNP and get additional information about it afterwards"**			
	Geneticist	A set of variations	Say "annotate with rsId from dbSNP"	Know if a variation has been identified as rsId from dbSNP
AT1	**"When I search for the rsId from dbSNP of the variation chr17:g.41245471C>T, I will get that the rsId is rs386597001"**			
	Geneticist	chr17:g.41245471C>T	Annotate rsId from dbSNP	rs386597001
AT2	**"When I search for the rsId of the variation chr11:g.111959693G>T I will get a message saying that the variation does not have rsId"**			
	Geneticist	chr11:g.111959693G>T	Annotate rsId from dbSNP	Message "the variation does not have rsId"

DSL User role

	Role	Context	Action	Goal
US	**"As a DSL user, I want to order the annotation of the patient's variations with the rsId from dbSNP, so that the patient variations will be annotated with the rsId of this source."**			
	DSL user	-	Order annotate variations with rsId from dbSNP	Variations will be annotated with the rsId from this data source
AT1	**"When I order the annotation of the patient variations with the rsId from dbsnp, the patient variations will be annotated with the rsId from dbsnp"**			
	DSL user	-	Order annotate variations with rsId from dbSNP	Annotate variations with rsId from dbSNP

User Story 3: Filter Variations by Gene (Analysis)

	Role	Context	Action	Goal
US	**"I want filter the patient's variations by a set of genes (by HGNC Gene Name), so that I can focus on the suitable variations only"**			
	Geneticist	Variations read and annotated with gene	Filter variations of a list of genes	Focus on the suitable variations
AT1	**"When I filter the variations chr2:g.6438366A>G R3HDM1, chr11:g.111959693G>T {SDHD, c11orf75, TIMM8B}, chr17:g.41245471C>T BRCA1 and chr17:g. 41256103G>A BRCA1, by the genes BRCA1 and BRCA2, I will see only the variations chr17:g.41245471C>T and chr17:g. 41256103G>A"**			
	Geneticist	chr2:g.136438366A>G R3HDM1 chr11:g.111959693G>T {SDHD, c11orf75, TIMM8B} chr17:g.41245471C>T BRCA1	Filter by BRCA1 and BRCA2	Chr17:g.41245471C>T, BRCA1 Chr17:g.41256103G>A, BRCA1

		chr17:g.41256103G>A BRCA1		
AT2	**"When I filter the variations chr2:g.6438366A>G R3HDM1, chr11:g.111959693G>T {SDHD, c11orf75, TIMM8B} by the genes BRCA1 and BRCA2, I will see a message saying that none variation belongs to the indicated genes"**			
	Geneticist	chr2:g.136438366A>G R3HDM1 chr11:g.111959693G>T {SDHD, c11orf75, TIMM8B}	Filter by BRCA1 and BRCA2	Message: "None variation belongs to the indicated genes"
AT3	**"When I filter the variations chr17:g.41245471C>T BRCA1 and chr17:g.41244435T>C BRCA1 by the gene BRCA I will see an error saying that the gene is not correct or is not in HGNC nomenclature"**			
	DSL user	-	Write BRCA	Error: "The gene does not exist"

DSL User role

	Role	Context	Action	Goal
US	**"As a DSL user, I want to write the set of genes expressed in hgnc nomenclature, so that patient variations can be filtered by these genes"**			
	DSL user	-	Write a set of genes in hgnc notation	Filter the patient variations by genes written in hgnc notation
AT1	**"When I write the genes set: BRCA1, BRCA2, patient variations will be filtered by these genes "**			
	DSL user	-	Write BRCA1, BRCA2	Filter Variations by BRCA1 BRCA2
AT2	**"When I write the gene ENS_00000012048, I will see an error saying that the gene is not expressed in hgnc nomenclature"**			
	DSL user	-	Write ENS_0000012048	Error: "The gene is not expressed in hgnc nomenclature"
AT3	**"When I write any gene and variations are not annotated I will see an error saying that you should annotate the gene before searching by HGVS"**			
	DSL user	Variations not annotated with gene	Write any gene	Error: "You should annotate the gene notation before filter by gene"

User Story 4: Report Variation's Gene (Report)

	Role	Context	Action	Goal
US	**"I want to add to a report with the variations their gene, so that I can locate easily each variation of the report."**			
	Geneticist	Variations	Report gene	See variation gene of each variation
AT1	**"When I report the variations chr11:g.111959693G>T with the gene I will see a table with one variation with the chr, position, ref, alt and gene"**			
	Geneticist	chr11:g.111959693G>T **Chr** **Pos** **Ref** **Alt**	Report gene	**Var** **Gene** ... SDHD, C11orf57 TIMM8B

		11 111959693 C T			

DSL User role

	Role	Context	Action	Goal
US	**"As a DSL user, I want to order the report of the gene, so that the gene is added to the patient's variation report.**			
	DSL user	-	Order the creation of a variation report with the gene	Add gene to the variations report
AT1	**"When I order to create a variations report with the gene, the gene will be added to the variation report**			
	DSL user	-	Order report gene	Add gene to variation report
AT2	**"When I order to create a variations report with the gene and variations are not annotated with the gene I will see an error saying that you should annotate the gene before reporting the gene"**			
	DSL user	Variations not annotated with gene	Order report gene	Error: "You should annotate the gene before reporting gene"

User Story 5: Report Variation's rsId (Report)

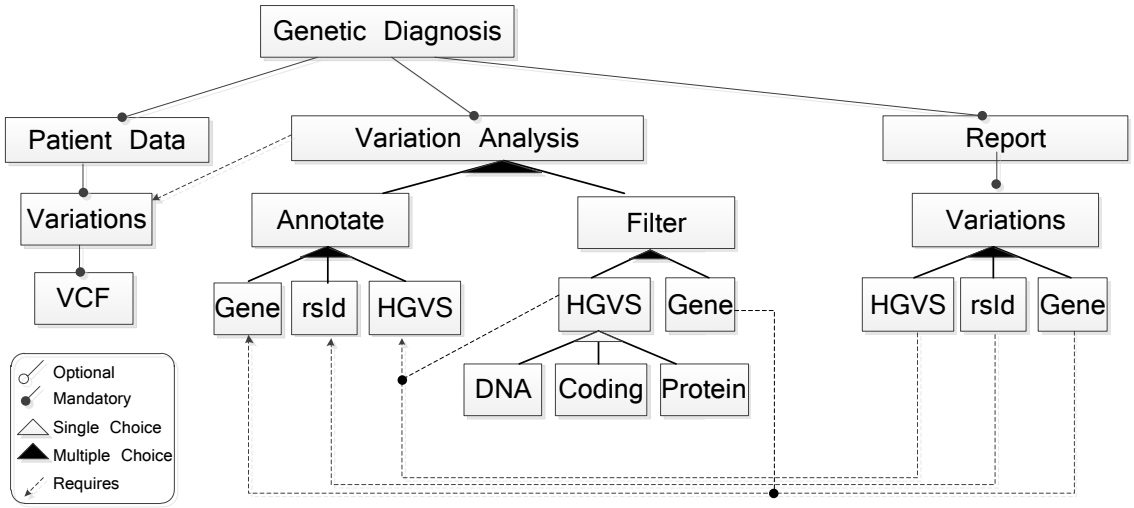
	Role	Context	Action	Goal
US	**"I want to add to a report with the variations their rsId, so that I can easily see which variations of the report are known SNPs"**			
	Geneticist	Variations	Report rsId	See variation rsId of each variation
AT1	**"When I report the variations chr17:g.41245471C>T with the rsId I will see a table with one variation with the chr, position, ref, alt and rsId"**			
	Geneticist	chr17:g.41245471C>T **Chr** **Pos** **Ref** **Alt** 17 41245471 C T	Report rsId	**Var** **rsId** ... rs4986850

DSL User role

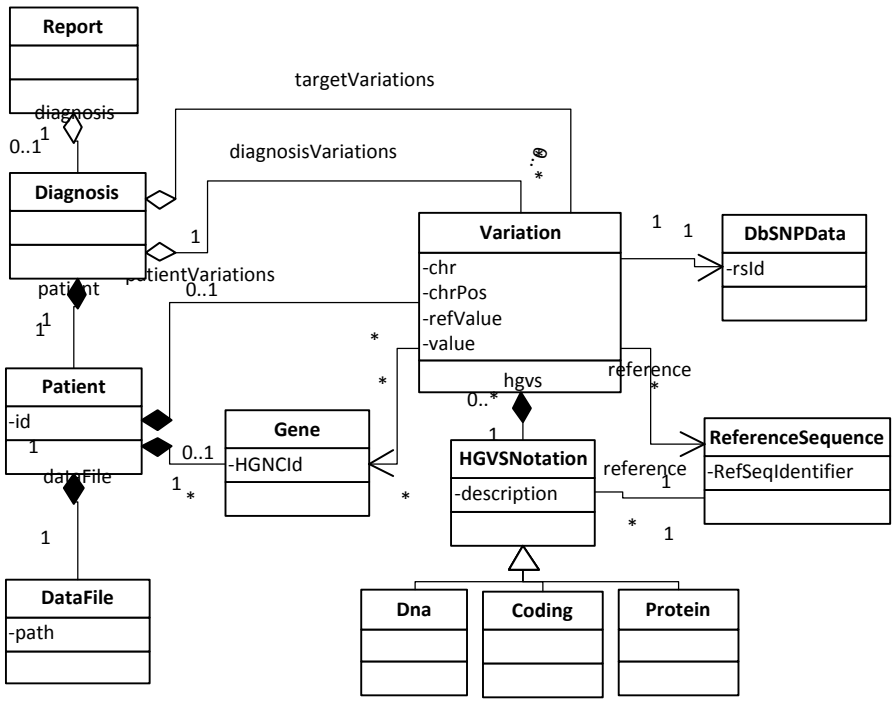
	Role	Context	Action	Goal
US	**"As a DSL user, I want to order the report of the rsId, so that the rsId is added to the patient's variation report.**			
	DSL user	-	Order the creation of a variation report with the rsId	Add rsId to the variations report
AT1	**"When I order to create a variations report with the rsId, the rsId will be added to the variation report**			
	DSL user	-	Order report rsId	Add rsId to variation report
AT2	**"When I order to create a variations report with the rsId and variations are not annotated with the rsId I will see an error saying that you should annotate the rsId**			

	before reporting the rsId"			
	DSL user	Variations not annotated with rsId	Order report rsId	Error: "You should annotate the rsId before reporting rsId"

Feature Model



Conceptual Model



Glossary

Report: Gathering of relevant information to show to geneticists/clinicians the result of a genetic disease diagnosis.

Diagnosis: Analysis that is performed to a patient to find out if the genotype indicates a potential risk to have a genetic disease.

Patient: Object of study to perform the diagnosis.

Datafile: Set of data saved in a file system.

Variation: Each of the nucleotides that a patient has different in regards to a reference sequence.

Reference Sequence: A representative sequence of nucleotides that theoretically represents the sequence of a "disease free" human.

HGVS Notation: Standard nomenclature the describe variations

(HGVS Notation) DNA: HGVS Nomenclature that represents the value of the variation at nucleotide level.

(HGVS Notation) Coding: HGVS Nomenclature that represents the value of the variation at the coding level.

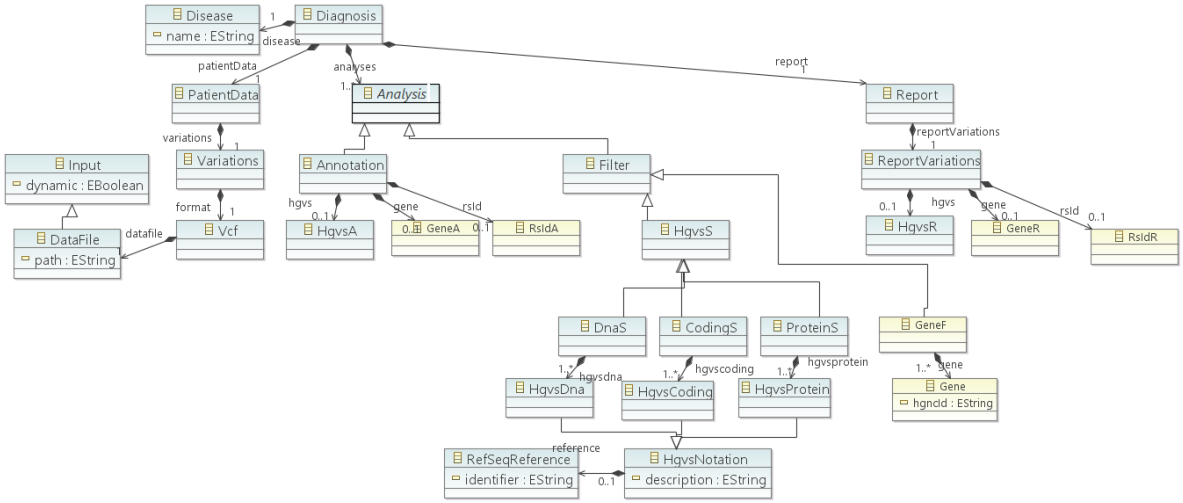
(HGVS Notation) Protein: HGVS Nomenclature that represents the value of the variation at the amino acid level.

Gene: Functional unit that delimiters a subset of nucletides from the DNA sequence that is responsible to regulate a function of the body.

DbSNPData: Information from the database of SNPs dbSNP, a reference database in the field.

Design

Abstract Syntax



Syntax Questionnaire

https://drive.google.com/open?id=1sayNzyjXyb6Jiennxa7kcKVQqMNOdO5DLT9o-xEmmo0

UNIVERSIDAD POLITECNICA DE VALENCIA

PROS Centro de Investigación en Métodos de Producción de Software

Syntax Examples

Take a look to the next 4 possible syntaxes to describe the usage scenario "Diagnose of Diabetes Mellitus type II" previously described

*Obligatorio

Syntax 1

recortesDiagnosis: Breast Cancer
Variations VCF file: Patient1.vcf
Variations Annotations: gene, rsId (dbSNP)
Analysis Filters: by genes{BRCA1, BRCA2}
Variation report fields: gene, rsId (dbSNP)

Give your opinion from 1 to 5 (being 1 the lowest rate and 5 the highest) about this syntax *

	1 (I don't like it at all)	2 (I don't like it)	3 (Neutral)	4 (It's ok)	5 (I like it a lot)
My opinion about Syntax 1	○	○	○	○	○

Syntax 2

Diagnose Breast Cancer
Read Variation's from VCF file Patient1.vcf
Annotate Variations with gene and rsId (dbSNP)
Filter Variations by genes {BRCA1, BRCA2}
Report Variations with gene and rsId (dbSNP)

Give your opinion from 1 to 5 (being 1 the lowest rate and 5 the highest) about this syntax *

	1 (I don't like it at all)	2 (I don't like it)	3 (Neutral)	4 (It's ok)	5 (I like it a lot)
My opinion about Syntax 2	○	○	○	○	○

Syntax 3

Diagnosis.Disease("Breast Cancer**")**
Diagnosis.Patient.Variations. ("Patient1.vcf**", VCF)**
Diagnosis.Patients.Variations.Annotations (gene, rsId (dbSNP)**)**
Diagnosis.Patients.Variations.Analysis.Filter.ByGene(BRCA1, BRCA2**)**
Diagnosis.Patients.Variations.Report.Fields(gene, rsId(dbSNP)**)**

Give your opinion from 1 to 5 (being 1 the lowest rate and 5 the highest) about this syntax *

	1 (I don't like it at all)	2 (I don't like it)	3 (Neutral)	4 (It's ok)	5 (I like it a lot)
My opinion about Syntax 3	○	○	○	○	○

Syntax 4

```
<Diagnose>
    <Disease>Breast Cancer</Disease>
    <PatientsData>
        <Variations>
            <VCF>Patient1.vcf</VCF>
        </Variations>
    </PatientsData>
    <Analyses>
        <Annotate><gene/><rsId source=dbSNP/></Annotate>
        <Filter><genes>
            <gene>BRCA1</gene>
            <gene>BRCA2</gene>
        </genes></Filter>
    </Analyses>
    <Report>
        <Variations>
            <gene/><rsId source=dbSNP/>
        </Variations>
    </Report>
</Diagnose>
```

Give your opinion from 1 to 5 (being 1 the lowest rate and 5 the highest) about this syntax *

	1 (I don't like it at all)	2 (I don't like it)	3 (Neutral)	4 (It's ok)	5 (I like it a lot)
My opinion about Syntax 4	○	○	○	○	○

Prefered Syntax

My prefered syntax is... *

- ○ Syntax 1
- ○ Syntax 2
- ○ Syntax 3
- ○ Syntax 4
- ○ I'd like to propose a new one

Concrete Syntax

Examples of Syntaxes proposed to geneticists

Syntax 1:Textual declarative

Diagnosis: **Breast Cancer (Analysis1)**
Variations VCF file: Patient1.vcf
Variations Annotations: hgvs, gene, rsId (dbSNP)
Analysis Filters: by genes{BRCA1, BRCA2}
Variation report fields: hgvs, gene, rsId (dbSNP)

Diagnosis: **Breast Cancer (Analysis2)**
Variations VCF file: Patient1.vcf
Variations Annotations: hgvs, gene, rsId (dbSNP)
Analysis Filters: by genes{RAD51C}
Variation report fields: hgvs, gene, rsId (dbSNP)

Syntax 2:Textual imperative

Diagnose **Breast Cancer (Analysis1)**
Read Variation's from VCF file Patient1.vcf
Annotate Variations with hgvs, gene and rsId (dbSNP)
Filter Variations by genes {BRCA1, BRCA2}
Report Variations with hgvs, gene and rsId (dbSNP)

```
Diagnose Breast Cancer (Analysis2)
Read Variation's from VCF file Patient1.vcf
Annotate Variations with hgvs, gene and rsId (dbSNP)
Filter Variations by genes {RAD51C}
Report Variations with hgvs, gene and rsId (dbSNP)
```

Syntax 3: Object oriented

```
Diagnosis.Disease("Breast Cancer (Analysis1)")
Diagnosis.Patient.Variations. ("Patient1.vcf", VCF)
Diagnosis.Patients.Variations.Annotations(hgvs, gene, rsId (dbSNP))
Diagnosis.Patients.Variations.Analysis.Filter.ByGene(BRCA1, BRCA2)
Diagnosis.Patients.Variations.Report.Fields(hgvs, gene, rsId (dbSNP))

Diagnosis.Disease("Breast Cancer (Analysis2)")
Diagnosis.Patient.Variations. ("Patient1.vcf", VCF)
Diagnosis.Patients.Variations.Annotations(hgvs, gene, rsId (dbSNP))
Diagnosis.Patients.Variations.Analysis.Filter.ByGene(RAD51C)
Diagnosis.Patients.Variations.Report.Fields(hgvs, gene, rsId (dbSNP))
```

Syntax 4: XML_like

```
<Diagnose>
  <Disease>Breast Cancer (Analysis 1)</Disease>
  <PatientsData>
    <Variations>
      <VCF>Patient1.vcf</VCF>
    </Variations>
  </PatientsData>
  <Analyses>
    <Annotate><hgvs/><gene/><rsId source=dbSNP/></Annotate>
    <Filter><genes>
      <gene>BRCA1</gene>
      <gene>BRCA2</gene>
    </genes></Filter>
  </Analyses>
  <Report>
    <Variations>
      <hgvs/><gene/><rsId source=dbSNP/>
    </Variations>
  </Report>
</Diagnose>


<Diagnose>
  <Disease>Breast Cancer (Analysis 1)</Disease>
  <PatientsData>
    <Variations>
      <VCF>Patient1.vcf</VCF>
    </Variations>
  </PatientsData>
  <Analyses>
    <Annotate><hgvs/><gene/><rsId source=dbSNP/></Annotate>
    <Filter><genes>
```

```
      <gene>BRCA1</gene>
      <gene>BRCA2</gene>
    </genes></Filter>
  </Analyses>
  <Report>
    <Variations>
      <hgvs/><gene/><rsId source=dbSNP/>
    </Variations>
  </Report>
</Diagnose>
```

Syntax Questionnaire

https://docs.google.com/forms/d/1sayNzyjXyb6Jiennxa7kcKVQqMNOdO5DLT9o-xEmmo0/viewform

Concrete Syntax Grammar

```
// automatically generated by Xtext
grammar diagnosis.it2.mydsl.MyDiag with org.eclipse.xtext.common.Terminals

import "diagnosis"
import "http://www.eclipse.org/emf/2002/Ecore" as ecore

diagnosis returns Diagnosis:
       'Diagnose' disease=disease
        patientData=patientData
     analyses+=analysis+
        report=report;

/*PATIENT DATA */
patientData returns PatientData:
       'Read'
       variations=variations;

variations returns Variations:
       'variations'
       format=vcf;

vcf returns Vcf:
       'from' 'a VCF file'
       datafile=dataFile;

/*ANALYSES */
analysis returns Analysis:
       annotation |
       search;

//Variation Annotation
annotation returns Annotation:
        'Annotate variations'
        'with'
        {Annotation}
        (hgvs=hgvsA)? (gene=geneA)? (rsId=rsIdA)?;
```

```
//Annotation Fields
hgvsA returns HgvsA:
       'hgvs'{HgvsA};
geneA returns GeneA:
       'gene'{GeneA};
rsIdA returns RsIdA:
       'rsId'{RsIdA};


//Variations Filter
search returns Filter:
       hgvsS|
       ('Filter' 'variations' geneF);
hgvsS returns HgvsS:
       'Search' 'variations'
       (dnaS|codingS|proteinS);
dnaS returns DnaS:
       hgvsdna+=hgvsdna+;
codingS returns CodingS:
       hgvscoding+=hgvscoding+;
proteinS returns ProteinS:
       hgvsprotein+=hgvsprotein+;
geneF returns GeneF:
       'by gene' gene+=gene+;



/*REPORT */
report returns Report:
       'Report'
        reportVariations=reportVariations;
reportVariations returns ReportVariations:
       'variations'
       {ReportVariations}
       ('with' (hgvs=hgvsR)? (gene=geneR)? (rsId=rsIdR)?)?;
hgvsR returns HgvsR:
       'hgvs' {HgvsR};
geneR returns GeneR:
       'gene' {GeneR};
rsIdR returns RsIdR:
       'rsId' {RsIdR};

/*DataModel Types */
disease returns Disease:
       name=EString;

dataFile returns DataFile:
       'from'{DataFile}
       (dynamic?=INPUT|path=EString);
```

```
hgvsdna returns HgvsDna:
	reference=refSeqReference ':''g.'description=HGVSEXPR;
hgvscoding returns HgvsCoding:
	reference=refSeqReference ':''c.'description=HGVSEXPR;
hgvsprotein returns HgvsProtein:
	reference=refSeqReference ':''p.'description=HGVSEXPR;
gene returns Gene:
	{Gene}
	hgncId= HGNCGENE;

refSeqReference returns RefSeqReference:
	identifier=(REFSEQ|ASSEMBLY);



/* Data Types ecore */
EBoolean returns ecore::EBoolean:
	'true' | 'false';
EString returns ecore::EString:
	STRING | ID;
EInt returns ecore::EInt:
	'-'? INT;

/*Terminals */
terminal HGNCGENE:
	(('A'..'Z')+ ((('0'..'9')+('A'..'Z')+)* |('0'..'9')+) );

terminal INPUT:
	'input';
terminal REFSEQ:
	'N'('C'|'G'|'M'|'P')'_'   INT'.'INT
;
terminal ASSEMBLY:
	('Hg'INT) |
	('NCBI'INT);
terminal HGVSEXPR:
	(INT(('+'|'-')INT)?('ins'|'del')('A'|'T'|'G'|'C')+)|//ins/del
	(INT(('+'|'-')INT)?('A'|'T'|'G'|'C')+'>'('A'|'T'|'G'|'C')+)|//indel
	(('A'..'Z'|'a'..'z')+INT('A'..'Z'|'a'..'z')+);//Protein
```

Semantic restrictions

Syntax Validation

Gene errors:

1. An EBNF rule that describes the structure of the gene according to the hgcn nomenclature

```
terminal HGNCGENE:
	(('A'..'Z')+ ((('0'..'9')+('A'..'Z')+)* |('0'..'9')+) );
```

Behaviour Validation

1. Gene existence. Error message provided by geneticists
    - "The gene does not exist
2. It is mandatory to always annotate (gene/rsId) before filtering (by gene) and reporting (gene/rsId). Error messages provided by geneticists
    - "You should annotate the gene notation before filter by gene"
    - "You should annotate the gene before reporting gene"
    - You should annotate the rsId before reporting rsId

Behavioral semantics

Annotate Variations with Gene

User Story	Annotate Variations with Gene			
Service Identifier	**SNPEff**			
Source description	**Galaxy**			
Inputs	Description	Type	Constant	Value
input	File that gathers the variations	File	No	-
inputformat	Format of the file containing the variations	Enumeration	Yes	VCF
Genome	Identifier of the genome	Enumeration	Yes	Hg19
symbol	Flag that indicates if the gene must be annotated	Boolean	Yes	True
outputformat	Format of the file containing the annotated variations	Enumeration	Yes	VCF
Outputs	Description	Type	Visibility	
Output	File that gathers the annotated variations	File (VCF)	Yes	

Annotate Variations with rsId from dbSNP

User Story	Annotate Variations with rsId from dbSNP			
Service Identifier	SNPSift Annotate			
Source description	Galaxy			
Inputs	Description	Type	Constant	Value
input	File that gathers the variations	File	No	-
dbSNP	File that gathers the annotations from dbSNP	File	Yes	ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606_b141_GRCh37p13/VCF/All.vcf
Outputs	Description	Type	Visibility	
Output	File that gathers the annotated	File (VCF)	Yes	

	variations		

Filter Variations by Gene

User Story	Filter Variations by Gene			
Service Identifier	SNPSift Filter			
Source description	Galaxy			
Inputs	Description	Type	Constant	Value
input	File that gathers the variations	File (VCF)	No	-
FilterCriteria	Expression that evaluates the genes to use as filter	String	No	Example: BRCA1 and BRCA2
Outputs	Description	Type	Visibility	
Output	File that gathers the annotated variations	File (VCF)	Yes	

Report Variations with Gene

User Story	Filter Variations by Gene			
Service Identifier	ShowVariations			
Source description	Galaxy			
Inputs	Description	Type	Constant	Value
input	File that gathers the variations	File (VCF)	No	-
gene	Flag that indicates if the gene must be showed	Boolean	Yes	True
Outputs	Description	Type	Visibility	
Output	Html that shows the variations and their annotations	File (HTML)	Yes	

Report Variations with rsId from dbSNP

User Story	Report Variations with rsId from dbSNP			
Service Identifier	ShowVariations			
Source description	Galaxy			
Inputs	Description	Type	Consta	Value

			nt	
input	File that gathers the variations	File (VCF)	No	-
rsId	Flag that indicates if the rsId must be showed	Boolean	Yes	True
Outputs	Description	Type	Visibility	
Output	Html that shows the variations and their annotations	File (HTML)	Yes	

Implementation

Semantic Tests

1. **Gene errors:** The gene must be described using the HGCN nomenclature**:** This restriction is described in the validator xtend class.

> **Input:**
> **Diagnose** LactoseIntoleranceHGVS
> **Read Variations genotypes from VCF file** Patient1.vcf
> **Annotate Variations with** hgvs
> **Filter Variations by gene BRCA**
> **Report variations** with hgvs
> **Output**:
> *"The gene does not exist"*

2. **It is mandatory to annotate the gene before filtering by gene and reporting by gene**. This restriction is described in the validator xtend class.

> **Input:**
> **Diagnose** LactoseIntoleranceHGVS
> **Read Variations genotypes from VCF file** Patient1.vcf
> **//Annotate Variations with** gene
> **Filter Variations** by gene BRCA1
> **Report variations** with gene
> **Output**:
> *"You should annotate the gene before filter by gene"*
> *"You should annotate the gene before reporting gene"*

3. **It is mandatory to annotate the rsId before filtering by rsId and reporting by rsId**. This restriction is described in the validator xtend class.

Input:
Diagnose LactoseIntoleranceHGVS
Read Variations genotypes from VCF file Patient1.vcf
//Annotate Variations with gene
Filter Variations by rsId rs1230009
Report variations with gene
Output:
"When the entity rsIdR is created, the entity rsIdA should be present"
"You should annotate the rsId before reporting"

Target platform Tests

User Story	Test Files	Galaxy Workflow	Tools
Annotate Gene	4Variants.vcf, 1IntergenicVariant.vcf	US1	Snpeff
Annotate rsId	RsIdVariant.vcf, NoRsVariant.vcf	US2	SNPSift Annotate
Filter variations by gene	4VariantsWithGene.vcf, 2VariantsWithGene.vcf	US3	SnpSift
Add Gene to report	4VariantsWithGene.vcf	US4_US5	ShowVariations
Add rsId to repor	VariantWithRsId.vcf	US4_US5	ShowVariations

Generator Test Example

Test Annotate Gene

```
@Before
def void testSetupOnce() {
   DiagnosisPackage.eINSTANCE.eClass();
   diagnosis = parser.parse ('''Diagnose BreastCancerAll
   Read variations from a VCF file from input
   Annotate variations with hgvs gene rsId
   Filter variations by gene BRCA1 BRCA2
   Report variations with hgvs gene rsId''')
   fsa= new InMemoryFileSystemAccess()
   generator.doGenerate(diagnosis.eResource, fsa)
}
@Test def testAnnotateGene(){
   Assert.assertFalse("The workflow fragment of AnnotateGene is
   different to the generated one",
   checkGeneratorGalaxy(fsa.getTextFiles().values(),"AnnotateGene
   .txt", "snp_eff"))
}
```

Testing

Testing Questionnaire

Did you find any error while using the DSL editor?

Usage scenario: -
Error description: -

Did you detect any error or warning that the DSL editor did not inform? Did you receive any message that you did not understand?

Usage scenario: Breast Cancer
Error description: Name of the disease does not allow writing spaces to separate words. I wanted to write “Breast Cancer”

Did you find any step missing or that you could not write to specify a usage scenario (diagnosis)? Would you like to add one for the next iteration that you consider important?

Usage scenario (and User Story if applies): -
Step description: -

Had the language all and the suitable elements to write all the information to complete the usage scenario (diagnosis)?

Comments: No, I wanted to specify the genome version of the analysis in order to ensure that the analyses performed (annotation and filter) query the suitable databases and report the suitable information. At the moment the genome version is by defect “GRCh37/hg19”. I want to use the DSL to specify this value.

Had the language the suitable structure (words order, symbols, indentation, labeling, etc…)?Would you change, add, remove or reorder any word of the language?

Usage scenario: -
Changes proposed: -

Do you know new software that better accomplishes any of the steps in any of the usage scenarios (diagnosis)?

User Story: Annotate Gene
Software information: Ensemble VEP also annotates the gene using the hgcn nomenclature.

Did you find any error when executing the code generated by the DSL editor (Galaxy Workflow)?

Usage scenario: Breast Cancer
Error description: The gene correction is not checked. I wrote BRCA1A1 and no error was retrieved while writing the DSL or in the execution. *Response: As it is a dynamic semantic error it cannot be checked in compilation time. At the moment this error is taken into account but it is not completely implemented (stub)

Universidad Politécnica de Valencia · Camino de Vera s/n · Edificio 1F · 46022 Valencia Spain ·
T. +34 96 387 70 07 Ext. 83530 · M. + 34 619 543 623 · F. +34 96 387 73 59 · info@pros.upv.es · www.pros.upv.es

Iteration 3

Additional Annotations

In this iteration, we collaborated with INCLIVA to validate the method and to apply it to get a first version of the DSL

Universidad Politécnica de Valencia · Camino de Vera s/n · Edificio 1F · 46022 Valencia Spain ·
T. +34 96 387 70 07 Ext. 83530 · M. + 34 619 543 623 · F. +34 96 387 73 59 · info@pros.upv.es · www.pros.upv.es

Analysis

Usage Scenarios, User Stories and Acceptance tests

Usage Scenario	**Usage Scenario Diabetes Mellitus Type 2 (Analysis 1)**
Description	In order to research the **diabetes mellitus type II disease**: • I want to read the genotypes of several samples from a VCF file. • I want to annotate the variations o With their genes o with all the names of the transcripts that they hit. o and the score and effect of SIFT and POLYPHEN. • I want to filter the variations o by the diabetes genes “ABCC8, CAPN10,KCNJ11, GCGR, SLC2A2, HNF4A, INS, INSR, PPARG, TCFl2, ADIPOQ, AKT2, PAX4, MAPK81p1, GPD2, MNTR1B”, o by the “deleterious” variations according to SIFT o and “possibly damaging” or “probably damaging” variations according to POLYPHEN • I want to create a report with o the variations main properties, o their genes o their transcript names o their Sift and Polyphen predictions.

Usage Scenario	**Usage Scenario Diabetes Mellitus Type 2 (Analysis 2)**
Description	In order to research the **diabetes mellitus type II disease**: • I want to read the genotypes of several samples from a VCF file. • I want to annotate the variations o with all the names of the transcripts that they hit. o and the sample MAF. • I want to filter the variations o by the diabetes genes “ABCC8, CAPN10,KCNJ11, GCGR, SLC2A2, HNF4A, INS, INSR, PPARG, TCFl2, ADIPOQ, AKT2, PAX4, MAPK81p1, GPD2, MNTR1B”, • I want to prioritize o by the sample Sift [0,0.5] “ascendant” • I want to create a report with o the variations main properties, o the genes o the Sift prediction

Usage Scenario	Usage Scenario Diabetes Mellitus Type 2 (Analysis 3)
Description	In order to research the **diabetes mellitus type II disease**: • I want to read the genotypes of several samples from a VCF file. • I want to annotate the variations o with all the names of the transcripts that they hit. o and the sample MAF. • I want to filter the variations o by the diabetes genes "ABCC8, CAPN10,KCNJ11, GCGR, SLC2A2, HNF4A, INS, INSR, PPARG, TCFl2, ADIPOQ, AKT2, PAX4, MAPK81p1, GPD2, MNTR1B", • I want to prioritize o by the sample MAF [0.1,0.5] "ascendant" • I want to create a report with o the variations main properties, o the sample MAF

User Story 1: Read genotypes of several samples from a VCF File

User Story	Description in Natural Language	Role	Mandatory	Action	Goal
Read genotypes of several samples from a VCF File	As a geneticist, I want to read several samples' genotypes from a VCF file, so that I can perform several analysis over those samples	Geneticist	No	Read patient's variations genotypes from a VCF file	See genotypes
Acceptance Tests	**Description in Natural Language**	**Role**	**Input**	**Action**	**Response**
AT1	When I choose the file 3Pat_3Var.vcf, I will see that the variations and genotypes are: chr2:g.136438366A>G [0/0,1/1,1/0], chr11:g.111959693G>T [1/1,1/1,1/1], chr17:g.41245471C>T [0/0,0/1,0/1]	Geneticist	File 3Pat_3Var.vcf	Read genotypes from VCF	chr2:g.136438366A>G [0/0,1/1,1/0], chr11:g.111959693G>T [1/1,1/1,1/1], chr17:g.41245471C>T [0/0,0/1,0/1]
AT2	When I choose the file WrongGenotype.vcf, I will see an error saying that The VCF file contains a variation whose allele is missing according to sample genotype values	Geneticist	File WrongGenotype.vcf chr2:g.136438366A>G [0/2,1/1,1/0]	Read genotypes from VCF	Error: "The variation chr2:g.136438366A>G contains 1 alternative allele, but the genotype index points to the alternative allele 2"

AT3	When I choose the file IncompleteWithoutINFO.vcf I will see an error saying that the VCF file does not have the mandatory column INFO according to the VCF format	Geneticist	IncompleteWithoutINFO.vcf	Read genotypes from VCF	Error: "VCF File does not have the mandatory column INFO according to the VCF format"
AT4	When I choose the file 1Pat_3Var.vcf I will see an error saying that the VCF file is not multisample	Geneticist	1Pat_3Var.vcf	Read genotypes from VCF	Error: "The VCF File not is multisample"
AT5	When I choose the file Alignment.sam I will see an error saying that the file has a wrong format	Geneticist	Alignment.sam	Read genotypes from VCF	Error: "Wrong format. You provided a SAM file and it was expected a VCF"

User Story	**Description in Natural Language**	**Role**	**Mandatory**	**Action**	**Goal**
Read genotypes of several samples from a VCF File	As a DSL user, I want to specify the file path of a VCF file so that genetic variations can be read from it	DSL user	No	Specify file path	Variations data
Acceptance Tests	**Description in Natural Language**	**Role**	**Input**	**Action**	**Response**
AT1	When I write the path C:/Files/3Variations.vcf, variations will be read from the fiel 3Variations.vcf	DSL user	-	Specify C:/Files/3Pat_3Var.vc	Read patient's variations genotypes from 3Pat_3Var.vcf

User Story 2: Annotate Variations with Transcripts Names

User Story	**Description in Natural Language**	**Role**	**Mandatory**	**Action**	**Goal**
Annotate Variations with Transcripts Names	As a geneticist, I want to annotate the patients' variations with the transcripts names (provided by RefSeq) that each variation hits (exons), so that I can see the different transcription patterns that the variation hits	Geneticist	No	Annotate variation with transcripts names	Variations annotated with the list of Transcripts Names that each variation hits (exon)
Acceptance Tests	Description in Natural Language	Role	Input	Action	Response

AT1	When I annotate the variation chr11:g. 76255523A>G with the transcripts names that it hits , I will see that the transcripts names are XM_005274109.1 XM_005274112.1 XM_005274108.1 XM_005274110.1 XM_005274106.1 XM_005274107.1 NM_020193.3 XM_005274113.1 XM_005274111.1	Geneticist	chr11: g. 76255523G>T	Annotate variation with transcripts names	XM_005274109.1 XM_005274112.1 XM_005274108.1 XM_005274110.1 XM_005274106.1 XM_005274107.1 NM_020193.3 XM_005274113.1 XM_005274111.1
AT2	When I annotate the variation Chr17:g.41256103G>A with the transcripts names that it hits, I will see a message saying that the variation does not hit any transcript (intron)	Geneticist	Chr17: g.41256103G>A	Annotate variation with transcripts names	Message: "The variation does not hit any transcript (intron)"

User Story	**Description in Natural Language**	**Role**	**Mandatory**	**Action**	**Goal**
Annotate Variations with Transcripts Names	As a DSL user, I want to order the annotation of the patient's variations with the transcript, so that the variations will be annotated with the transcript	DSL user	No	Order "annotate variations with transcript"	Variations will be annotated with transcript
Acceptance Tests	**Description in Natural Language**	**Role**	**Input**	**Action**	**Response**
AT1	When I order the annotation of the patient variations with the transcript, I will obtain the source code that annotates variations with transcripts	DSL user	-	Order "annotate variations with transcript"	Source code that annotates variations with transcripts

User Story 3: Annotate Variations with SIFT prediction

User Story	**Description in Natural Language**	**Role**	**Mandatory**	**Action**	**Goal**

Annotate Variations with SIFT prediction	I want to annotate the patients' variations with the prediction of the SIFT algorithm score and effect for each variation transcript, so that I can preliminary assess the predicted effect of each variation taking into account to this algorithm	Geneticist	No	Annotate variation with SIFT prediction Score and Effect	Variations annotated with the prediction score and effect of each variation transcript according to the SIFT algorithm
Acceptance Tests	**Description in Natural Language**	**Role**	**Input**	**Action**	**Response**
AT1	When I annotate the variation chr11:g. 76255523G>T with the SIFT prediction score and effect, I will get that for the following transcripts the predictions are XM_005274109.1: 0.36 tolerated, XM_005274108.1:0.26 tolerated, XM_005274106.1: 0.26 tolerated, NM_020193.3:0.35 tolerated, XM_005274113.1:0.54 tolerated	Geneticist	chr11:g. 76255523G>T	Annotate variation with SIFT prediction Score and Effect	XM_005274109.1 0.36 tolerated XM_005274108.1 0.26 tolerated XM_005274106.1 0.26 tolerated NM_020193.3 0.35 tolerated XM_005274113.1 :0.54 tolerated
AT2	When I annotate the variation Chr17:g.41256103G>A with the SIFT prediction score and effect, I will see a message saying that SIFT does not provide any prediction for this variation	Geneticist	Chr17:g.41256103G>A	Annotate variation with SIFT prediction Score and Effect	Message: "SIFT does not provide any prediction for this variation"

User Story	**Description in Natural Language**	**Role**	**Mandatory**	**Action**	**Goal**
Annotate Variations with SIFT prediction	As a DSL user, I want to order the annotation of the variations with the SIFT prediction, so that variations will be annotated with the SIFT score and effect	DSL user	No	Order annotate variations with SIFT	Variations will be annotated with SIFT score and effect
Acceptance Tests	**Description in Natural Language**	**Role**	**Input**	**Action**	**Response**

AT1	When I order the annotation of the patient variations with the SIFT prediction, the patient variations will be annotated with the SIFT score and effect	DSL user	-	Order annotate variations with SIFT	ource code that annotates variations with SIFT score and effect

User Story 4: Annotate Variations with POLYPHEN prediction

User Story	**Description in Natural Language**	**Role**	**Mandatory**	**Action**	**Goal**
Annotate Variations with POLYPHEN prediction	I want to annotate the patient's variations with the prediction of the POLYPHEN algorithm score and the effect, for each transcript, so that I can preliminary assess the predicted effect of each variation taking into account to this algorithm	Geneticist	No	Annotate variation with POLYPHEN prediction Score and effect	Variations will be annotated with prediction score of each variation transcript according to the POLYPHEN algorithm
Acceptance Tests	Description in Natural Language	Role	Input	Action	Response
AT1	When I annotate the variation chr11:g. 76255523G>T with the POLYPHEN prediction score and effect, I will get that for the following transcripts the predictions are XM_005274109.1: 0.991 probably damaging, XM_005274108.1: 0.899 possibly damaging, XM_005274106.1: 0.899 possibly damaging, NM_020193.3: 0.899 possibly damaging, XM_005274113.1:0 unknown	Geneticist	chr11:g. 76255523G>T	Annotate variation with POLYPHEN prediction Score and effect	XM_005274109.1: 0.991 Probably damaging XM_005274108.1: 0.899 possibly damaging XM_005274106.1: 0.899 possibly damaging NM_020193.3: 0.899 possibly damaging XM_005274113.1: 0 unknown
AT2	When I annotate the variation chr17:g.41256103G>A with the POLYPHEN prediction score, I will see a message saying that POLYPHEN does not provide any prediction for this variation	Geneticist	chr17: g.41256103G>A	Annotate variation with POLYPHEN prediction Score and effect	Message: "POLYPHEN does not provide any prediction for this variation

User Story	**Description in Natural Language**	**Role**	**Mandatory**	**Action**	**Goal**

Annotate Variations with POLYPHEN prediction	As a DSL user, I want to order the annotation of the patient's variations with the POLYPHEN prediction, so that variations will be annotated with the POLYPHEN score and effect.	DSL user	No	Order annotate variations with POLYPHEN	Variations will be annotated with POLYPHEN score and effect
Acceptance Tests	**Description in Natural Language**	**Role**	**Input**	**Action**	**Response**
AT1	When I order the annotation of the patient variations with the POLYPHEN prediction, the patient variations will be annotated with the POLYPHEN score and effect	DSL user	-	Order annotate variations with POLYPHEN	Source code that annotates variations with POLYPHEN score and effect

User Story 5: Annotate Variations with the sample Minor Allele frequency

User Story	Description in Natural Language	Role	Mandatory	Action	Goal
Annotate Variations with the sample Minor Allele frequency	I want to annotate the patients' variations with the sample Minor Allele Frequency, so that I can see the frequency of the allele that has minor occurrence in the analysed samples	Geneticist	No	Annotate with Minor Allele Frequency	Know the frequency of allele with minor occurrence
Acceptance Tests	**Description in Natural Language**	**Role**	**Input**	**Action**	**Response**
AT1	When I annotate the MAF of the variations chr2:g.136438366A>G [0/0,1/1,1/0], chr11:g. 76255523G>T [1/1,0/0,1/1], chr11:g.111959693G>T [1/1,1/1,1/1], chr17:g.41245471C>T [0/0,0/1,0/1] the MAFS are G(0.5), G(0.33), G(0.0), T(0.33) respectively	Geneticist	chr2:g.136438366A>G [0/0,1/1,1/0] chr11:g. 76255523G>T [1/1,0/0,1/1] chr11:g.111959693G>T [1/1,1/1,1/1] chr17:g.41245471C>T [0/0,0/1,0/1]	Annotate with Minor Allele Frequency	G(0.5) G(0.33) G(0.0) T(0.33) Warning: "The MAF can not be ensured because the number of sample is below 20".

User Story	Description in Natural Language	Role	Mandatory	Action	Goal

Annotate Variations with the sample Minor Allele frequency	As a DSL user, I want to order the annotation of the patient's variations with the MAF, so that patient variations will be annotated with the MAF allele and frequency	DSL user	No	Order annotate variations with MAF	Variations will be annotated with Minor Allele Frequency
Acceptance Tests	Description in Natural Language	Role	Input	Action	Response
AT1	When I order the annotation of the patient variations with the MAF, the patient variations will be annotated with the MAF allele and frequency	DSL user	-	Order annotate variations with MAF	Source code that annotates variations with MAF allele and frequency

User Story 6: Filter Variations by SIFT effect

User Story	Description in Natural Language	Role	Mandatory	Action	Goal
Filter Variations by SIFT effect	As a geneticist, I want to filter the patient variations by the effect predicted by SIFT (tolerated/deleterious), so that I can see only the variations that pass the filter	Geneticist	No	Filter by a qualitative value of SIFT prediction (tolerated/deleterious	See only the variations that pass the filter
Acceptance Tests	Description in Natural Language	Role	Input	Action	Response
AT1	When I filter the variations chr2:g.136438366A>G {}, chr11:g. 76255523 G>T {deleterious}, chr11:g.111959693G>T {}, chr17:g.41245471C>T {tolerated, deleterious} by the SIFT effect deleterious I will see the variations chr11:g.76255523 G>T and chr17:g.41245471C>T	Geneticist	chr2: g.136438366A>G {} chr11: g.76255523 G>T {deleterious} chr11: g.111959693G>T {} chr17: g.41245471C>T {tolerated, deleterious}	Filter by a qualitative value of SIFT prediction deleterious	chr11: g.76255523 G>T chr17: g.41245471 C>T
AT2	When I filter the variations chr2:g.136438366A>G {}, chr11:g. 76255523 G>T {deleterious}, chr11:g.111959693G>T {}, by the SIFT effect tolerated I will see a message saying that none variation has been annotated with the effect tolerated	Geneticist	chr2: g.6438366A>G {} chr11: g. 76255523 G>T {deleterious} chr11: g.111959693G>T {}	Filter by a qualitative value of SIFT prediction harmful	Message: "None variation has been annotated by SIFT with the effect harmful"

User Story	Description in Natural Language	Role	Mandatory	Action	Goal
Filter Variations by SIFT effect	As a DSL user, I want to write an effect (according to the SIFT algorithm category), so that patient variations can be filtered by this effect	DSL user	No	Write a SIFT effect	Filter the patient variations by SIFT effect
Acceptance Tests	Description in Natural Language	Role	Input	Action	Response
AT1	When I write the effect damaging, patient variations will be filtered by this effect	DSL user	-	Write SIFT damaging	Source code that filters variations by SIFT damaging
AT2	When I write the effect harmful, I will see an error saying that the effect harmful is not a SIFT effect	DSL user	-	Write SIFT harmful	Error: "The effect harmful is not a SIFT effect. The effect must be tolerated or damaging"
Dependencies	Description in Natural Language	Precondition		Action	Error Message
DP1	When I write any SIFT effect and variations are not annotated with SIFT prediction I will see an error saying that you should annotate the SIFT prediction before filtering by effect	Annotate with SIFT prediction		Write any SIFT effect	Error: "You should annotate the SIFT prediction before filtering by SIFT effect"

User Story 7: Filter Variations by POLYPHEN effect

User Story	Description in Natural Language	Role	Mandatory	Action	Goal
Filter Variations by POLYPHEN effect	s a geneticist, I want to filter the patient variations by a set of effects predicted by POLYPHEN (benign, probably damaging, possibly damaging, unknown), so that I can see only the variations that pass the filter	Geneticist	No	Filter by a set of POLYPHEN effect (benign/ probably damaging/ possibly damaging)	See only the variations that pass the filter
Acceptance Tests	Description in Natural Language	Role	Input	Action	Response

AT1	When I filter the variations chr2:g.136438366A>G {},chr11:g. 76255523 G>T {probably damaging}, chr11:g.111959693G>T{}, chr17:g.41245471C>T {benign}, by the POLYPHEN effect possibly damaging I will see the variation chr11:g.111959693G>T	Geneticist	chr2: g.136438366A>G {} chr11: g. 76255523 G>T {probably damaging} chr11: g.111959693G>T{} chr17: g.41245471C>T {benign}	Filter by POLYPHEN probably damaging	chr11: g.111959693G>T
AT2	When I filter the variations chr2:g.136438366A>G {},chr11:g. 76255523 G>T {probably damaging}, chr11:g.111959693G>T{}, chr17:g.41245471C>T {benign}, by the effect probably damaging I will see a message saying that none variation has the desired effect	Geneticist	chr2: g.136438366A>G {} chr11: g. 76255523 G>T {probably damaging} chr11: g.111959693G>T{} chr17: g.41245471C>T {benign}	Filter by POLYPHEN possibly damaging	Message: "None variation has been annotated by POLYPHEN with the desired effect

User Story	**Description in Natural Language**	**Role**	**Mandatory**	**Action**	**Goal**
Filter Variations by POLYPHEN effect	As a DSL user, I want to write a set of effects (according to POLYPHEN algorithm category), so that patient variations can be filtered by these effects	DSL user	No	Write a POLYPHEN effect	Filter the patient variations by effect
Acceptance Tests	Description in Natural Language	Role	Input	Action	Response
AT1	When I write the effect tolerated, patient variations will be filtered by this effect	DSL user	-	Write POLYPHEN damaging	Source code that filter s Variations by POLYPHEN damaging
At2	When I write the effect harmful, I will see an error saying that the effect harmful is not a SIFT effect	DSL user	Write POLYPHEN harmful		Error: "The effect harmful is not a POLYPHEN effect. The effects must be tolerated, possibly damaging or probably damaging"

Dependencies	Description in Natural Language	Precondition	Action	Error Message
DP1	When I write any POLYPHEN effect and variations are not annotated with POLYPHEN prediction I will see an error saying that you should annotate the POLYPHEN prediction before filtering by effect	annotated with POLYPHEN prediction	Write any POLYPHEN effect	Error: "You should annotate the POLYPHEN prediction before filtering by POLYPHEN effect"

User Story 8: Prioritize Variations by sample Minor Allele Frequency

User Story	Description in Natural Language	Role	Mandatory	Action	Goal
Prioritize Variations by sample Minor Allele Frequency	I want to prioritize the patient's variations by an interval (between 0 and 1) of the sample minor allele frequency and an order,Min2Max or Max2Min, so that I can focus and on the most important variations according to this frequency.	Geneticist	No	Prioritize by a MAF interval and an order	Focus on the most important variations
Acceptance Tests	Description in Natural Language	Role	Input	Action	Response
AT1	When I prioritize the variations chr2:g.136438366A>G 0.5, chr11:g. 76255523 G>T 0.33 chr11:g.111959693G>T 0, chr17:g.41245471C>T 0.33, by the interval [0.15, 0.40] Min2Max , I will see only the variations chr11:g. 76255523 G>T and chr17:g.41245471C>T	Geneticist	chr2:g.136438366A>G 0.5 chr11:g. 76255523 G>T 0.33 chr11:g.111959693G>T 0 chr17:g.41245471C>T 0.33	Prioritize by MAF [0.15, 0.40] Min2Max	chr11:g. 76255523 G>T chr17:g.41245471C>T

AT2	When I prioritize the variations chr2:g.136438366A>G 0.5, chr11:g. 76255523 G>T 0.33 chr11:g.111959693G> T 0, chr17:g.41245471C>T 0.33, by the interval [0, 0.2] and by Min2Max, I will see a message saying that none variation has a MAF inside the interval	Geneticist	chr2:g.136438366A>G 0.5 chr11:g. 76255523 G>T 0.33 chr11:g.111959693G> T 0 chr17:g.41245471C>T 0.33	Prioritize by MAF [0, 0.2] Min2Max	Message "None variation has a MAF inside the interval.

User Story	**Description in Natural Language**	**Role**	**Mandatory**	**Action**	**Goal**
Prioritize Variations by sample Minor Allele Frequency	As a DSL user, I want to write a numeric interval and an order (Min2Max or Max2Min), so that patient variations can be filtered by this MAF range and ordered	DSL user	No	Order a prioritization and Write a MAF range and an order	Filter and order the patient variations by MAF frequency
Acceptance Tests	Description in Natural Language	Role	Input	Action	Response
AT1	When I write the range [0.15, 0.30] and the order ascendant, patient variations will be filtered and ordered by this MAF range and criteria	DSL user	-	Order prioritize and write MAF [0.15, 0.30] Min2Max	Source code that filters variations by MAF [0.15, 0.30] and orders ascendant
AT2	When I write the range [0.15, 2], I will see an error saying that the MAF range values must be between 0 and 1	DSL user	-	Order a prioritization and Write a MAF [0.15,2] and an order Min2Max	Error: "The MAF range values must be between 0 and 1"
AT3	When I write the MAF order down I will see an error saying that the MAF order must be ascendant or descendant	DSL user	-	Order a prioritization and write a MAF range and an order down	Error: "The MAF order must be ascendant or descendant"
Dependencies	Description in Natural Language	Precondition		Action	Error Message
DP1	When I write any MAF range and order and variations are not annotated with MAF I will see an error saying that you should annotate the MAF before filtering by MAF	annotate with MAF		Order a prioritization and Write a MAF range and an order	Error: "You should annotate the MAF prediction before filtering by MAF"

User Story 9: Prioritize Variations by SIFT score

User Story	Description in Natural Language	Role	Mandatory	Action	Goal
Prioritize Variations by SIFT score	As a geneticist, I want to prioritize (filter and order) the patient's variations by a range (between 0 and 1) and an order of the SIFT prediction (min2Max, Max2Min), so that I can focus and on the most important variations for the analysis based on this algorithm	Geneticist	No	Prioritize by a SIFT interval and an order	Focus on the most important variations for the analysis
Acceptance Tests	Description in Natural Language	Role	Input	Action	Response
AT1	When I prioritize the variations chr2:g.136438366A>G {}, chr11:g. 76255523 G>T {0.3,0.4,0.5}, chr11:g.111959693G>T {}, chr17:g.41245471C>T {0.01,0.16}, by the SIFT interval [0.01, 0.40] min2max, I will see the variations chr17:g.41245471C>T andchr11:g.111959693G>T following this order	Geneticist	chr2: g.136438366A>G {} chr11: g. 76255523 G>T {0.3,0.4,0.5} chr11: g.111959693G>T {} chr17: g.41245471C>T {0.01,0.16}	Prioritize by SIFT [0.01, 0.40] min2Max	chr17: g.41245471C>T chr11: g.111959693G>T
AT2	When I prioritize the variations chr2:g.6438366A>G SIFT={}, chr11:g.111959693G>T SIFT={0, 0.01}, chr17:g.41245471C>T SIFT={0.01, 0.16}, chr17:g.41244435T>C SIFT={}, by the SIFT interval [0.30, 1] min2max , I will see a message saying that none variation has a SIFT prediction inside the interval	Geneticist	chr2:g.136438366A>G {} chr11:g. 76255523 G>T {0.03,0.04,0.05} chr11:g.111959693G>T {} chr17:g.41245471C>T {0.01,0.16}	Prioritize by SIFT [0.2, 1] min2Max	Message "None variation has a SIFT prediction inside the interval"

User Story	Description in Natural Language	Role	Mandatory	Action	Goal

Prioritize Variations by SIFT score	As a DSL user, I want to write a numeric interval (between 0 and 1) and an order (Min2Max), so that patient variations can be filtered by this SIFT range and ordered	DSL user	No	Write a SIFT range and an order	Filter and order the patient variations by SIFT effect. Help: "lower values are deleterious and higher values tolerated"
Acceptance Tests	**Description in Natural Language**	**Role**	**Input**	**Action**	**Response**
AT1	When I write the range [0.15, 0.30] and the order ascendant, patient variations will be filtered and ordered by this SIFT range and criteria	DSL user	-	Write SIFT [0.15, 0.30] Min2Max	Source code that filters by SIFT [0.15, 0.30] and order ascendant.
AT2	When I write the range [0.15, 2], I will see an error saying that the SIFT range values must be between 0 and 1	DSL user	-	Write SIFT [0.15,2]	Error: "The SIFT range values must be between 0 and 1"
AT3	When I write the SIFT order down I will see an error saying that the SIFT order must be ascendant or descendant"	DSL user	-	Write SIFT down	Error: "The SIFT order must be min2Max or Max2Min"
Dependencies	Description in Natural Language	Precondition		Action	Error Message
DP1	When I write any SIFT range and order and variations are not annotated with SIFT I will see an error saying that you should annotate the SIFT before filtering by SIFT	annotated with SIFT		Write an SIFT rang and order	Error: "You should annotate the SIFT prediction before filtering by SIFT"

User Story 10: Prioritize Variations by POLYPHEN score

User Story	Description in Natural Language	Role	Mandatory	Action	Goal

Prioritize Variations by POLYPHEN score	As a geneticist, I want to prioritize (filter and order) the patient's variations by a range(between 0 and 1) and order of the POLYPHEN prediction, so that I can focus and on the most important variations for the analysis based on to this algorithm	Geneticist	No	Prioritize by a POLPYHEN interval and an order (min2Max, max2Min)	Focus on the most important variations for the analysis
Acceptance Tests	Description in Natural Language	Role	Input	Action	Response
AT1	When I prioritize the variations chr2:g.136438366A>G {}, chr11:g. 76255523 G>T {0,0.984,0988}, chr11:g.111959693G>T {}, chr17:g.41245471C>T {0.01,0.006}, by the POLYPHEN interval [0.001, 1] ascendant , I will see the variations chr17:g.41245471C>T and chr11:g.111959693G>T"	Geneticist	chr2: g.136438366A>G {} chr11: g. 76255523 G>T {0,0.984,0988} chr11: g.111959693G>T {} chr17: g.41245471C>T {0.01,0.006}	Prioritize by POLYPHEN [0.001, 1] min2Max	chr17: g.41245471C>T chr11: g.111959693G>T
AT2	When I prioritize the variations chr2:g.136438366A>G {}, chr11:g. 76255523 G>T {0,0.984,0988}, chr11:g.111959693G>T {}, chr17:g.41245471C>T {0.01,0.006}by the POLYPHEN interval [0.15, 0.85] ascendant , I will see a message saying that none variation has a POLYPHEN prediction inside the interval	Geneticist	chr2: g.136438366A>G {} chr11: g. 76255523 G>T {0,0.984,0988} chr11: g.111959693G>T {} chr17: g.41245471C>T {0.01,0.006}	Prioritize by POLYPHEN [0.15,0.85] min2Max	Message "None variation has a POLYPHEN prediction inside the interval"

User Story	**Description in Natural Language**	**Role**	**Mandatory**	**Action**	**Goal**
Prioritize Variations by POLYPHEN score	As a DSL user, I want to write a numeric interval and an order, so that patient variations can be filtered by this POLYPHEN range and ordered	DSL user	No	Write a POLYPHEN range and an order (min2Max, Max2Min)	Filter and order the patient variations by POLYPHEN effect. Help: "lower values are benign and higher values damaging"

Acceptance Tests	Description in Natural Language	Role	Input	Action	Response
AT1	When I write the range [0.15, 0.30] and the order ascendant, patient variations will be filtered and ordered by this POLYPHEN range and criteria	DSL user	-	Write POLYPHEN [0.15, 0.30] ascendant	Source code that filters by POLYPHEN [0.15, 0.30] and order ascendant
AT2	When I write the range [0.15, 2], I will see an error saying that the POLYPHEN range values must be between 0 and 1	DSL use	-	Write POLYPHEN [0.15,2]	Error: "The POLYPHEN range values must be between 0 and 1"
AT3	When I write the POLYPHEN order down I will see an error saying that the POLYPHEN order must be ascendant or descendant	DSL use	-	Write POLYPHEN down	Error: "The POLYPHEN order must be ascendant or descendant"
Dependencies	Description in Natural Language	Precondition		Action	Error Message
DP1	When I write any POLYPHEN range and order and variations are not annotated with POLYPHEN I will see an error saying that you should annotate the POLYPHEN before filtering by POLYPHEN	annotated with POLYPHEN		Write an POLYPHEN range an order	Error: "You should annotate the POLYPHEN prediction before filtering by POLYPHEN"

User Story 11: Report Variation's MAF

User Story	Description in Natural Language	Role	Mandatory	Action	Goal
Report Variation's MAF	I want to add the MAF to the variations' report, so that I can see which allele has the minimum frequency and the value of this frequency	Geneticist	No	Report variations' MAF	See MAF of each variation
Acceptance Tests	**Description in Natural Language**	**Role**	**Input**	**Action**	**Response**

AT1	When I report the MAF of the variation chr11:g.111959693 C>T I will see a table with the one variation row and the columns: “11” “111959693” “C” “T” and	Geneticist	chr11:g.111959693C>T	Report MAF prediction	Report with 5 columns and one row with the fields: “11”,”1111959693”, ”C”,”T”, ”G(0.0)”

User Story	**Description in Natural Language**	**Role**	**Mandatory**	**Action**	**Goal**
Report Variation’s MAF	As a DSL user, I want to order the report of the MAF, so that the MAF is added to the patient’s variation report.	DSL user	No	Order the creation of a variation report with the MAF	Add MAF to the variations report
Acceptance Tests	Description in Natural Language	Role	Input	Action	Response
AT1	When I order to create a variations report with the MAF, the MAF will be added to the variation report	DSL user	-	Order report MAF	Source code to add MAF to a variation report
Dependencies	Description in Natural Language	Precondition		Action	Error Message
DP1	When I order to create a variations report with the MAF and variations are not annotated with the MAF I will see an error saying that you should annotate the MAF before reporting the MAF	annotated with MAF		Order report MAF	Error: “You should annotate the MAF before reporting MAF”

User Story 12: Report Variation’s SIFT Prediction

User Story	**Description in Natural Language**	**Role**	**Mandatory**	**Action**	**Goal**
Report Variation’s SIFT Prediction	I want to add to a variation report their SIFT predictions, so that I can see which variations have an effect in codification	Geneticist	No	Report variations’ SIFT prediction	See transcripts of each variation
Acceptance Tests	Description in Natural Language	Role	Input	Action	Response

AT1	When I report the SIFT prediction of the variation chr11:g.111959693 C>T I will see a table with the one variation row and the columns: "11" "111959693" "C" "T" and "NM_001276504.1 0.01 damaging, NM_001276506.1 0 damaging, NM_003003.3 0 damaging	Geneticist	chr11:g. 111959693C>T	Report SIFT prediction	Report with 5 columns and one row with the fields: "11","1111959693", "C","T", "G(0.0)". and "NM_001276504.1", NM_001276506.1, NM_003003.3"

User Story	**Description in Natural Language**	**Role**	**Mandatory**	**Action**	**Goal**
Report Variation's SIFT Prediction	As a DSL user, I want to order the report of the SIFT, so that the SIFT prediction score and effect is added to the patient's variation report	DSL user	No	Order the creation of a variation report with the SIFT	Add SIFT prediction score and effect to the variations report
Acceptance Tests	**Description in Natural Language**	**Role**	**Input**	**Action**	**Response**
AT1	When I order to create a variations report with the SIFT, the SIFT will be added to the variation report	DSL user	-	Order report SIFT	Source code to add SIFT prediction score and effect to a variation report
Dependencies	Description in Natural Language	Precondition		Action	Error Message
DP1	When I order to create a variations report with the SIFT and variations are not annotated with the SIFT I will see an error saying that you should annotate the SIFT before reporting the SIFT	annotate with SIFT		Order report SIFT	Error: "You should annotate the SIFT prediction before reporting SIFT prediction"

User Story 13: Report Variation's POLYPHEN Prediction

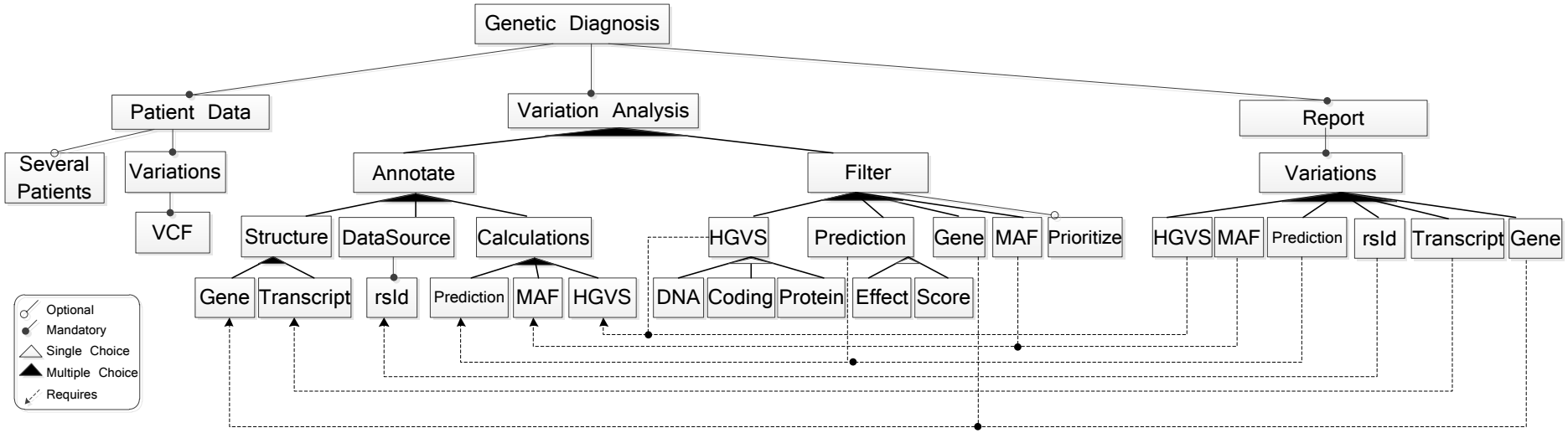
User Story	**Description in Natural Language**	**Role**	**Mandatory**	**Action**	**Goal**

Report Variation's POLYPHEN Prediction	I want to add to a variation report their POLYPHEN predictions (score and effect), so that I can see which variations have an effect in codification	Geneticist	No	Report variations' POLYPHEN prediction	See transcripts of each variation
Acceptance Tests	**Description in Natural Language**	**Role**	**Input**	**Action**	**Response**
AT1	When I report the POLYPHEN prediction of the variation chr11:g.111959693 C>T I will see a table with the one variation row and the columns: "11" "111959693" "C" "T" and "NM_001276504.1 0.01 benign, NM_001276506.1 0 possibly damaging, NM_003003.3 0 possibly damaging	Geneticist	chr11:g. 111959693C>T	Report POLYPHEN prediction	Report with 5 columns and one row with the fields: "11","1111959693", "C","T", "G(0.0)". and "NM_001276504.1: 0.01 damaging", NM_001276506.1: 0 damaging, NM_003003.3: 0 damaging

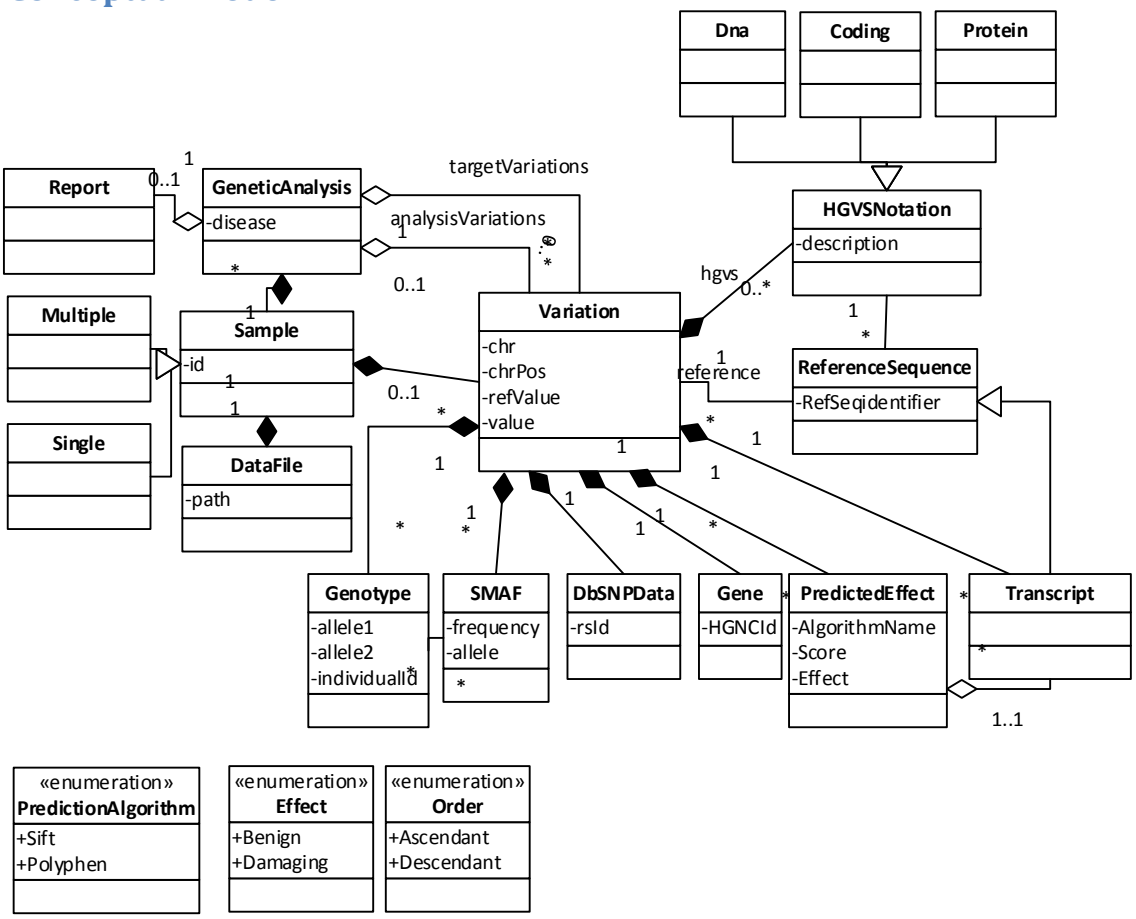
User Story	**Description in Natural Language**	**Role**	**Mandatory**	**Action**	**Goal**
Report Variation's POLYPHEN Prediction	As a DSL user, I want to order the report of the POLYPHEN, so that the POLYPHEN prediction score and effect is added to the patient's variation report	DSL user	No	Order the creation of a variation report with the POLYPHEN	Add POLYPHEN prediction score and effect to the variations report
Acceptance Tests	**Description in Natural Language**	**Role**	**Input**	**Action**	**Response**
AT1	When I order to create a variations report with the POLYPHEN, the POLYPHEN will be added to the variation report	DSL user	-	Order report POLYPHEN	Source code add POLYPHEN prediction score and effect to variation report
Dependencies	Description in Natural Language	Precondition		Action	Error Message

DP1	When I order to create a variations report with the POLYPHEN and variations are not annotated with the POLYPHEN I will see an error saying that you should annotate the POLYPHEN before reporting the POLYPHEN	Annotate with POLYPHEN	Order report POLYPHEN	Error: "You should annotate the POLYPHEN prediction before reporting POLYPHEN prediction"

Feature Model



Conceptual Model



Glossary

Genetic Analysis: Analysis that is performed to a sample observing genetic data.
Report: Relevant information gathered as a result of a genetic analysis.
Sample: Object of study to perform a genetic analysis (one or several individuals).
Single (Sample): When the object of study is a single individual
Multiple (Sample): When the object of study are several individuals.
Datafile: Genetic data of the sample saved in a textual file.
Variation: Each of the nucleotides that each individual of the sample has different in regards to a reference sequence.
Reference Sequence: A representative sequence of nucleotides that theoretically represents the sequence of a "disease free" human.
HGVS Notation: Standard nomenclature the describe variations
(HGVS Notation) DNA: HGVS Nomenclature that represents the value of the variation at nucleotide level.
(HGVS Notation) Coding: HGVS Nomenclature that represents the value of the variation at the coding level.
(HGVS Notation) Protein: HGVS Nomenclature that represents the value of the variation at the amino acid level.

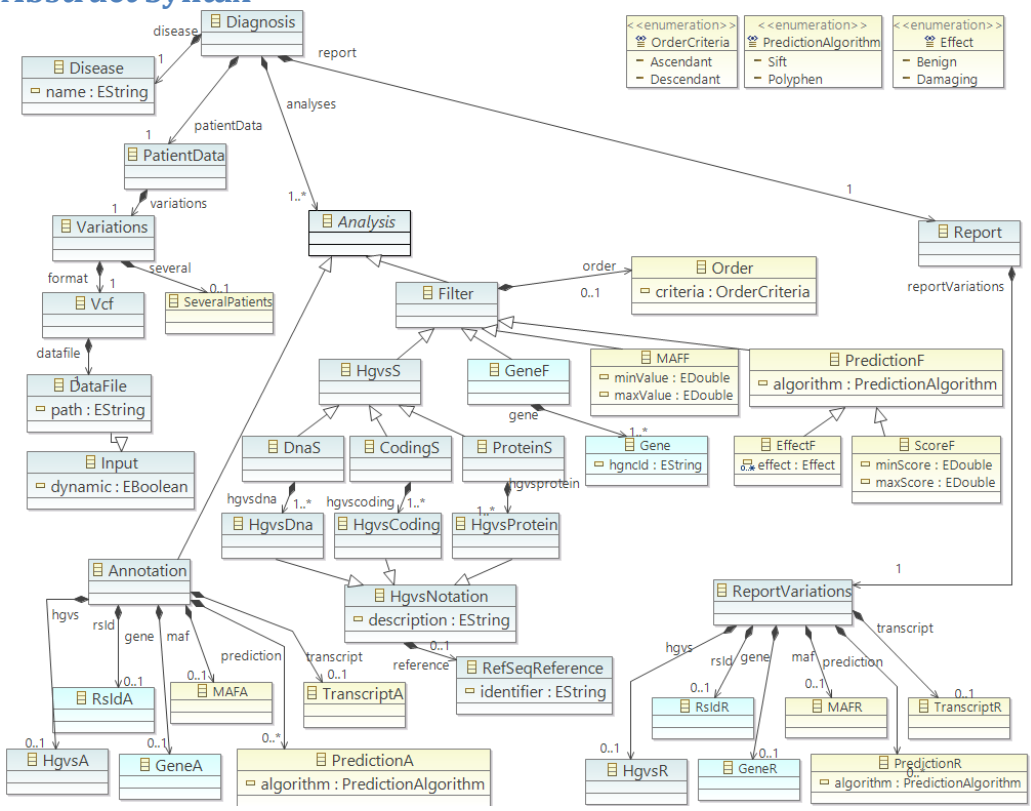
Gene: Functional unit that delimiters a subset of nucletides from the DNA sequence that is responsible to regulate a function of the body.
DbSNPData: Information from the database of SNPs dbSNP, a reference database in the field.
Transcript: Functional structure of the gene that represents the parts that play a role in the transcription of the nucleotides of the genes to proteins.
Predicted Effect: Result of the execution of a prediction algorithm that assesses the effect of the variation in an individual.
Genotype: Two alleles of an individual in a position in the chromosome.
sMAF: Abbreviation of sample Minimum Allele Frequency. Calculation that represents the allele has the minimum frequency among the individuals of the sample.

Relationships between Feature model and Conceptual model

- Feature *VCF*-> Entity *DataFile*
- Feature *Annotate.Calculations.EffectPrediction*-> Entity *PredictedEffect*
- Feature *Filter.Gene*->Entity *Gene*
- Feature *Filter.EffectPrediction*->Entity *PredictedEffect*
- Feature *Filter.sMAF*->Entity *sMAF*
- Feature *Priotitize*->Entity *Interval*
- Feature *Priotitize*->Entity *Order*
- Feature *Hgvs.HgvsDna*->Entity *DNA*
- Feature *Hgvs.HgvsCoding*->Entity *Coding*
- Feature *Hgvs.HgvsProtein*->Entity *Protein*

Design

Abstract Syntax



Examples of concrete Syntaxes

Syntax 1:Textual declarative

Diagnosis: **Diabetes Mellitus Type 2 (Analysis 1)**
Variations Genotypes VCF file: Patient1.vcf
Variations Annotations: gene, transcript, SIFT, POLYPHEN
Analysis Filters: by genes {ABCC8, CAPN10,KCNJ11, GCGR, SLC2A2, HNF4A, INS, INSR, PPARG, TCFl2, ADIPOQ, AKT2, PAX4, MAPK81p1, GPD2, MNTR1B}
Analysis Priorizations: by SIFT [0, 0.1]
Variation report fields: gene, transcript, SIFT, POLYPHEN

Syntax 2:Textual imperative

Diagnose **Diabetes Mellitus Type 2 (Analysis 1)**
Read Variations genotypes from VCF file Patient1.vcf
Annotate Variations with gene, transcript, SIFT and POLYPHEN
Filter Variations by genes {ABCC8, CAPN10,KCNJ11, GCGR, SLC2A2, HNF4A, INS, INSR, PPARG, TCFl2, ADIPOQ, AKT2, PAX4, MAPK81p1, GPD2, MNTR1B }
Report Variations with gene, transcript, SIFT, POLYPHEN

Syntax 3: Object oriented

Diagnosis.Disease("Diabetes Mellitus Type 2 (Analysis 1)")

Diagnosis.Patient.Variations.Genotypes ("Patient1.vcf", VCF)
Diagnosis.Patients.Variations.Annotations (gene, transcript, SIFT, POLYPHEN)
Diagnosis.Patients.Variations.Analysis.Filter.ByGene(ABCC8, CAPN10,KCNJ11, GCGR, SLC2A2, HNF4A, INS, INSR, PPARG, TCFl2, ADIPOQ, AKT2, PAX4, MAPK81p1, GPD2, MNTR1B)
Diagnosis.Patients.Variations.Analysis.Priorization.Prediction.Score(SIFT, 0, 0.1)
Diagnosis.Patients.Variations.Report.Fields(gene, transcript, SIFT, POLYPHEN)

Syntax 4: XML_like

```
<Diagnose>
  <Disease>Diabetes Type2</Disease>
  <PatientsData>
    <Genotypes>
        <VCF>Patient1.vcf</VCF>
    </Genotypes>
  </PatientsData>
  <Analyses>
    <Annotate>
      <gene/>
      <transcript/>
      <SIFT/>
      <POLYPHEN/>
    </Annotate>
    <Filter>
      <genes>
        <gene>ABCC8</gene>
        <gene>CAPN10</gene>
        <gene>KCNJ11</gene>
        <gene>GCGR</gene>
        <gene>SLC2A2</gene>
        <gene>HNF4A</gene>
        <gene>INS</gene>
        <gene>INSR</gene>
        <gene>PPARG</gene>
        <gene>TCFl2</gene>
        <gene>ADIPOQ</gene>
        <gene>AKT2</gene>
        <gene>PAX4</gene>
        <gene>MAPK81p1</gene>
        <gene>GPD2</gene>
        <gene>MNTR1B</gene>
      </genes>
    </Filter>
    <Prioritize order="descendant">
      <SIFT>
        <min>0</min>
        <max>0.1<max>
      </SIFT>
    </Prioritize>
  </Analyses>
  <Report>
    <Variations>
```

```
    <gene/>
    <transcript/>
    <SIFT/>
    <POLYPHEN/>
  </Variations>
</Report>
```

Syntax Questionnaire

https://docs.google.com/forms/d/1AiNWZqcewhCyeQpL0xO8977c7iViY5tmGSpsBF5euZQ/viewform?usp=send_form

Syntax Questionnaire

This questionnarie proposes four different syntaxes to describe the usage scenario "Diagnose of Diabetes Mellitus type 2 (Analysis 2)" and asks for your opinion about them.

Instructions:
1) Set the timestamp when you start the questionnarie.
2) For each syntax, you must rate their suitability to describe the scenario. By suitability, we mean if the syntax is CONCISE (goes to the point) and EASY to understand.

*Obligatorio

Start Time *

h : min : s

Syntax 1

Diagnosis: Diabetes Mellitus Type 2 (Analysis 2)
Genotypes VCF file: Patients.vcf
Variations Annotations: gene, transcript, sample MAF, SIFT and POLYPHEN
Analysis Filters: by gene {ABCC8, CAPN10,KCNJ11, GCGR, SLC2A2, HNF4A, INS, INSR, PPARG, TCFl2, ADIPOQ, AKT2, PAX4, MAPK81p1, GPD2, MNTR1B}
by SIFT effect deleterious
Analysis Priorizations by sample MAF [0.1,0.5] ascendant
Variation report fields: gene, transcript, sample MAF, SIFT and POLYPHEN

Rate from 1 to 5 (being 1 the lowest rate and 5 the highest) the suitability of syntax 1 *

	1 (lowest)	2	3	4	5 (highest)
Suitability of Syntax 1	○	○	○	○	○

Syntax 2

Diagnose Diabetes Mellitus Type 2 (Analysis 2)
Read Variation's genotypes from VCF file Patients.vcf
Annotate Variations with gene, transcript, sample MAF, SIFT and POLYPHEN
Filter by genes {ABCC8, CAPN10,KCNJ11, GCGR, SLC2A2,
HNF4A, INS, INSR, PPARG, TCFl2, ADIPOQ, AKT2, PAX4,
MAPK81p1, GPD2, MNTR1B}
Filter by SIFT effect deleterious
Prioritize by sample MAF [0.1,0.5] ascendant
Report Variations with gene, transcript, sample MAF, SIFT and POLYPHEN

Rate from 1 to 5 (being 1 the lowest rate and 5 the highest) the suitability of syntax 2 *

	1 (lowest)	2	3	4	5 (highest)
Suitability of Syntax 2	○	○	○	○	○

Syntax 3

Diagnosis. Disease("Diabetes Mellitus Type 2 (Analysis 2)")
Diagnosis.Patients.Variations.Genotypes ("Patients.vcf", VCF)
Diagnosis.Patients.Variations.Annotations (gene, transcript, sample MAF, SIFT, POLYPHEN)
Diagnosis.Patients.Variations.Analysis.Filters.ByGene(ABCC8, CAPN10,KCNJ11,
GCGR, SLC2A2, HNF4A, INS, INSR, PPARG, TCFl2, ADIPOQ, AKT2, PAX4,
MAPK81p1, GPD2, MNTR1B)
Diagnosis.Patients.Variations.Analysis.Filters.bySIFTEffect(deleterious)
Diagnosis.Patients.Variations.Analysis.Priorizations.bySampleMAF(0.1,0.5,ascendant)
Diagnosis.Patients.Variations.Report.Fields(gene, transcript,sample MAF, SIFT, POLYPHEN)

Rate from 1 to 5 (being 1 the lowest rate and 5 the highest) the suitability of syntax 3 *

	1 (lowest)	2	3	4	5 (highest)
Suitability of Syntax 3	○	○	○	○	○

Syntax 4

```
<Diagnose>
    <Disease>Diabetes Mellitus Type 2 (Analysis 2)</Disease>
    <PatientsData>
        <Genotypes>
            <VCF>Patient.vcf</VCF>
        </Genotypes>
    </PatientsData>
    <Analyses>
        <Annotation>
            <Gene/><Transcript/><sMAF/><SIFT/><POLYPHEN/>
        </Annotation>
        <Filter><Gene>ABCC8, CAPN10,KCNJ11, GCGR, SLC2A2,
            HNF4A, INS, INSR, PPARG, TCF12, ADIPOQ, AKT2,
            PAX4, MAPK81p1, GPD2, MNTR1B</Gene></Filter>
        <Filter><SIFT><effect>deleterious</effect></Sift>
        <Prioritize order="ascendant"><sMAF>
            <min>0.1</min>
            <max>0.5</max>
        </sMAF></Prioritize>
    </Analyses>
    <Report>
        <Variations>
            <Gene/><Transcript/><sMAF/><SIFT/><POLYPHEN/>
        </Variations>
    </Report>
</Diagnose>
```

Rate from 1 to 5 (being 1 the lowest rate and 5 the highest) the suitability of syntax 4 *

	1 (lowest)	2	3	4	5 (highest)
Suitability of Syntax 4	○	○	○	○	○

Taking into account the previous 4 examples, which syntax would you use for describing a diagnosis workflow? *

- ○ I'd like to propose a new one
- ○ Syntax 1
- ○ Syntax 2
- ○ Syntax 3
- ○ Syntax 4

Syntax 1

For the following statements expressed with Syntax 1, would you change anything?

Describe diagnosis information: "Diagnosis: Diabetes Mellitus Type 2 (Analysis 2)"

Suggestions or changes for improvement

Read genotypes: "Genotypes VCF file: Patients.vcf"

Suggestions or changes for improvement

Annotate variations: "Variations Annotations: gene, transcript, sample MAF, SIFT and POLYPHEN"

Suggestions or changes for improvement

Apply filters to variations: "Analysis Filters: by gene {ABCC8, CAPN10,KCNJ11, GCGR, SLC2A2, HNF4A, INS, INSR, PPARG, TCFI2, ADIPOQ, AKT2, PAX4, MAPK81p1, GPD2, MNTR1B}"

Suggestions or changes for improvement

Apply priorizations to variations "Analysis Priorizations: by sample MAF [0.1, 0.5] ascendent"

Suggestions or changes for improvement

Report variations: "Variation report fields: gene, transcript, sample MAF, SIFT and POLYPHEN"

Suggestions or changes for improvement

Concrete Syntax Grammar (Syntax 2)

Complete Concrete Syntax

```
grammar diagnosis.it3.mydsl.MyDiag with org.eclipse.xtext.common.Terminals

import "diagnosis"
import "http://www.eclipse.org/emf/2002/Ecore" as ecore

diagnosis returns Diagnosis:
       'Diagnose' disease=disease
        patientData=patientData
     analyses+=analysis+
        report=report;
```

```
/*PATIENT DATA */
patientData returns PatientData:
	'Read'
	variations=variations;


variations returns Variations:
	'variations'
	several=severalPatients
	format=vcf;

severalPatients returns SeveralPatients:
	'genotypes'{SeveralPatients};

vcf returns Vcf:
	'from' 'a VCF file'
	datafile=dataFile;

/*ANALYSES */
analysis returns Analysis:
	annotation |
	search;

//Variation Annotation
annotation returns Annotation:
	 'Annotate variations'
	 'with'
	 {Annotation}
	 (hgvs=hgvsA)?                    (gene=geneA)?(transcript=transcriptA)?
prediction+=predictionA* (rsId=rsIdA)?;

//Annotation Fields
hgvsA returns HgvsA:
	'hgvs'{HgvsA};
geneA returns GeneA:
	'gene'{GeneA};
transcriptA returns TranscriptA:
	'transcript'{TranscriptA};
predictionA returns PredictionA:
	algorithm=predictionAlgorithm;
rsIdA returns RsIdA:
	'rsId'{RsIdA};



//Variations Filter
search returns Filter:
	hgvsS
	|('Filter' 'variations' 'by' (geneF | predictionF))
	|('Prioritize' 'variations' 'by' (geneF | predictionF) order=order);

hgvsS returns HgvsS:
	'Search' 'variations'
```

```
	(dnaS|codingS|proteinS);
dnaS returns DnaS:
	hgvsdna+=hgvsdna+;
codingS returns CodingS:
	hgvscoding+=hgvscoding+;
proteinS returns ProteinS:
	hgvsprotein+=hgvsprotein+;
geneF returns GeneF:
	'gene' gene+=gene+;
predictionF returns PredictionF:
	effectF|scoreF;

effectF returns EffectF:
	algorithm=predictionAlgorithm 'effect'
	effect+=effectEnum+;

scoreF returns ScoreF:
	algorithm=predictionAlgorithm 'score'
	'['minScore=EDouble','maxScore=EDouble']';
order returns Order:
	criteria=orderCriteria;



/*REPORT */
report returns Report:
	'Report'
	 reportVariations=reportVariations;
reportVariations returns ReportVariations:
	'variations'
	{ReportVariations}
	('with' (hgvs=hgvsR)? (gene=geneR)? (rsId=rsIdR)?)?;
hgvsR returns HgvsR:
	'hgvs' {HgvsR};
geneR returns GeneR:
	'gene' {GeneR};
rsIdR returns RsIdR:
	'rsId' {RsIdR};
transcriptR returns TranscriptR:
	'transcript'{TranscriptR};
predictionR returns PredictionR:
	algorithm=predictionAlgorithm;

/*DataModel Types */
disease returns Disease:
	name=EString;

dataFile returns DataFile:
	'from'{DataFile}
	(dynamic?=INPUT|path=EString);


hgvsdna returns HgvsDna:
```

```
	reference=refSeqReference ':''g.'description=HGVSEXPR;
hgvscoding returns HgvsCoding:
	reference=refSeqReference ':''c.'description=HGVSEXPR;
hgvsprotein returns HgvsProtein:
	reference=refSeqReference ':''p.'description=HGVSEXPR;
gene returns Gene:
	hgncId=(EString|HGNCGENE);

refSeqReference returns RefSeqReference:
	identifier=(REFSEQ|ASSEMBLY);



/* Data Types ecore */
EBoolean returns ecore::EBoolean:
	'true' | 'false';
EString returns ecore::EString:
	STRING | ID;
EInt returns ecore::EInt:
	'-'? INT;
EDouble returns ecore::EDouble:
	'-'? INT '.' INT;


/*Terminals and Enumerations */
terminal HGNCGENE:
	(('A'..'Z')+ ((('0'..'9')+('A'..'Z')+)* |('0'..'9')+) );
terminal INPUT:
	'input';
terminal REFSEQ:
	'N'('C'|'G'|'M'|'P')'_'   INT'.'INT
;
terminal ASSEMBLY:
	('Hg'INT) |
	('NCBI'INT);
terminal HGVSEXPR:
	(INT(('+'|'-')INT)?('ins'|'del')('A'|'T'|'G'|'C')+)|//ins/del
	(INT(('+'|'-')INT)?('A'|'T'|'G'|'C')+'>'('A'|'T'|'G'|'C')+)|//indel
	(('A'..'Z'|'a'..'z')+INT('A'..'Z'|'a'..'z')+);//Protein

enum predictionAlgorithm returns PredictionAlgorithm: Sift='Sift' |
Polyphen='Polyphen';
enum orderCriteria returns OrderCriteria: AlphAsc='AlphAsc'|
AlphDes='AlphDes'| Max2Min='Max2Min' |Min2Max='Min2Max';
enum effectEnum returns Effect: Tolerated='tolerated'| Damaging='damaging'|
ProbablyD='probably damaging' |PossiblyD='possibly damaging';
```

Semantic Restrictions

- **Effect errors**: The prediction effect has two possible values: *tolerated* and *damaging*. This restriction is described using an EBNF rule.

```
enum effectEnum returns Effect:
Tolerated='tolerated'| Damaging='damaging';
```

- **Order errors:** There are four possible ways to order variations: *alphabetically descendant, alphabetically ascendant, from max to min, from min to max*. This restriction is described as an EBNF rule that describes with an enumeration the possible ways to order variations.

```
enum orderCriteria returns OrderCriteria:
AlphAsc='AlphAsc'|AlphDes='AlphDes'|Max2Min='Max2Min'
Min2Max='Min2Max';
```

- I**t is mandatory to annotate POLYPHEN/SIFT before filtering by POLYPHEN/SIFT and reporting by POLYPHEN**. The restriction in natural language is "*When the entity PredictionF or PredictionR are created, the entity PredictionA should be present and the attribute algorithm of both entities must be equal*". If this restriction is not fulfilled, the custom error messages are: "*You should annotate the prediction from POLYPHEN notation before filtering by a POLYPHEN/SIFT effect*" and "*You should annotate the prediction from POLYPHEN/SIFT before reporting the POLYPHEN/SIFT effect*".
- **It is mandatory to always annotate (Transcript/SIFT/POLYPHEN/MAF) before filtering (SIFT/POLYPHEN/MAF), prioritizing (SIFT/POLYPHEN/MAF) and reporting (Transcript /SIFT/POLYPHEN/MAF)**
- **Filter/Prioritize range by SIFT/POLYPHEN/MAF must be between 0 and 1**. "The attributes minValue and maxValue of MAFF/ScoreF must be >=0 and <=1". If this restriction is not fulfilled, the custom error messages is: "*The range to filter variations is outside limits, it should be amgon 0 and 1*".

Behavioral semantics (semantic stories)

- Read Genotypes from VCF
- Annotate Transcript
- Annotate SIFT
-Annotate POLYPHEN
-Annotate MAF
- Filter variations by SIFT effect
- Filter variations by POLYPHEN effect
- Prioritize variations by MAF
- Prioritize variations by SIFT score
- Prioritize variations by POLYPHEN score
- **Report of variations:**
    1) Add Transcript
    2) Add MAF

3) Add SIFT
4) Add POLYPHEN

The DSL will be realized in the Galaxy Environment (Giardine, B: A platform for interactive large-scale genome analysis. Genome Research).

Read Genotypes from VCF

User Story	Read Genotypes from VCF			
Service Identifier	UploadFile			
Source description	Galaxy			
Inputs	**Description**	**Type**	**Constant**	**Value**
Format	Format of the file containing the genotypes data	Enumeration	Yes	VCF
File	A data file with the file path.	String	No	-
Genome	Identifier of the genome	Enumeration	Yes	Human GRCH37/hg19
Outputs	**Description**	**Type**	**Visibility**	
output	A list of Variations with their main properties including the genotypes and their reference sequence	VCF File	Yes	
format		Enumeration	No	
database		Enumeration	No	

Annotate Variations Transcript

User Story	Annotate Variations Transcript			
Service Identifier	Variant Effect Predictor *retrieves variation's gene transcripts instead variation transcripts (exonic regions)			
Source description	Galaxy			
Inputs	**Description**	**Type**	**Constant**	**Value**
Input	File that gathers the variations	DataFile (VCF)	False	-
Species	Name of the species	String	True	homosapiens
Refseq	If annotations are expressed according to refSeq references	Boolean	True	True
Hgvs	If hgvs must be annotated	Boolean	True	True
Outputs	**Description**	**Type**	**Visibility**	
annotated_vcf	File that gathers the annotated variations	DataFile (VCF)	True	

Annotate Variations SIFT /POLYPHEN

User Story	Annotate Variations SIFT /POLYPHEN

Service Identifier	Variant Effect Predictor			
Source description	Galaxy			
Inputs	Description	Type	Constant	Value
Input	File that gathers the variations	DataFile (VCF)	False	-
Species	Name of the species	String	True	homosapiens
Refseq	If annotations are expressed according to refSeq references	Boolean	True	True
Sift/Poly	If Sift/Polyphen must be annotated	Boolean	True	True
Outputs	Description	Type	Visibility	
annotated_vcf	File that gathers the annotated variations	DataFile (VCF)	True	

Annotate Variations MAF

User Story	Annotate Variations with MAF			
Service Identifier	Allele Frequencies			
Source description	Galaxy			
Inputs	Description	Type	Constant	Value
Input	File that gathers the variations	DataFile (VCF)	False	-
AnnotateVCF	If the input file must be annotated	Boolean	True	True
Outputs	Description	Type	Visibility	
annotated_vcf	File that gathers the annotated variations	DataFile (VCF)	True	

Filter by SIFT/POLYPHEN EFF

User Story	Filter Variations by SIFT /POLYPHEN			
Service Identifier	Variant Effect Predictor filter			
Source description	Galaxy			
Inputs	Description	Type	Constant	Value
Input	File that gathers the variations	DataFile (VCF)	False	-
FilterCriteria	Evaluation expression that indicates the sift/polyphen criteria to filter	String	False	Examples: “Sift is deleterious” “Polyphen is benign”
Outputs	Description	Type	Visibility	
annotated_vcf	File that gathers the annotated variations	DataFile (VCF)	True	

Prioritize by MAF

User Story	Prioritize Variations by MAF			
Service Identifier	Sort MAF			
Source description	Galaxy			
Inputs	Description	Type	Constant	Value
Input	File that gathers the variations	DataFile (VCF)	False	-
FilterCriteria	Expression that evaluates the MAF criteria to filter	String	False	Example: "MAF <0.5"
Order	If variations must be ordered	Boolean	True	True
orderCriteria	Order to organize the filtered variations	Enumeration	False	"Ascendant/Descendant"
Outputs	Description	Type	Visibility	
annotated_vcf	File that gathers the annotated variations	DataFile (VCF)	True	

Prioritize by SIFT/POLYPHEN Score

User Story	Prioritize Variations by SIF/POLYPHEN Score			
Service Identifier	Sort SIFT and POLYPHEN			
Source description	Galaxy			
Inputs	Description	Type	Constant	Value
Input	File that gathers the variations	DataFile (VCF)	False	-
FilterCriteria	Expression that evaluates the SIFT/POLY criteria to filter	String	False	Example: "SIFT <0.5" "POLY >0.1"
Order	If variations must be ordered	Boolean	True	True
orderCriteria	Order to organize the filtered variations	Enumeration	False	"Ascendant/Descentant"
Outputs	Description	Type	Visibility	
annotated_vcf	File that gathers the annotated variations	DataFile (VCF)	True	

Report Variations with Transcript/MAF/ SIFT/ POLYPHEN

User Story	Report Variations with Transcript/MAF/ SIFT/ POLYPHEN

Service Identifier	ShowVariations			
Source description	Galaxy			
Inputs	Description	Type	Constant	Value
Input	File that gathers the variations	DataFile (VCF)	False	-
transcript	If transcript must be reported	Boolean	False	Yes/No
MAF	If MAF must be reported	Boolean	False	Yes/No
SIFT/POLY	If SIF/POLY must be reported	Boolean	False	Yes/No
Outputs	Description	Type	Visibility	
report_html	File that gathers the variations and annotations organized in a table	DataFile (HTML)	True	

Implementation

Semantic Tests

- **Effect errors**: The prediction effect has two possible values. This restriction is described using an EBNF rule.

> **Input:**
> **Diagnose** Diabetes Mellitus Type 2 (Analysis 1)
> **Read Variations genotypes from VCF file** Patient1.vcf
> **Annotate Variations with** polyphen
> **Filter Variations** by Polyphen effect harmful
> **Report variations** with hgvs
> **Output**:
> *"The effect harmful is not a POLYPHEN effect. The effects must be tolerated or damaging"*

- **Order errors:** An EBNF rule that describes with an enumeration the possible ways to order variations. This restriction is described in the Validator xtend class.

> **Input:**
> **Diagnose** Diabetes Mellitus Type 2 (Analysis 1)
> **Read Variations genotypes from VCF file** Patient1.vcf
> **Annotate Variations with** polyphen
> **Prioritize variations Variations** by Polyphen effect tolerated AlphAsc
> **Report variations** with hgvs
> **Output**:
> *"The POLYPHEN order must be AlphAsc, AlphDesc, Min2Max or Max2Min"*

- **It is mandatory to annotate POLYPHEN/SIFT before filtering by POLYPHEN/SIFT and reporting by POLYPHEN**. This restriction is described in the Validator xtend class

> **Input:**
> **Diagnose** Diabetes Mellitus Type 2 (Analysis 1)
> **Read Variations genotypes from VCF file** Patient1.vcf
> **//Annotate Variations with** polyphen
> **Filter Variations** by predicted effect polyphen damaging
> **Output**:
> *"You should annotate the prediction from POLYPHEN notation before filtering by a POLYPHEN effect"*
> *"You should annotate the prediction from POLYPHEN/SIFT before reporting the POLYPHEN/SIFT effect"*

- **It is mandatory to always annotate (Transcript/SIFT/POLYPHEN/MAF) before filtering (SIFT/POLYPHEN/MAF), prioritizing (SIFT/POLYPHEN/MAF) and reporting (Transcript /SIFT/POLYPHEN/MAF)** This restriction is described in the Validator xtend class

> **Input:**
> **Diagnose** Diabetes Mellitus Type 2 (Analysis 1)
> **Read Variations genotypes from VCF file** Patient1.vcf
> **//Annotate Variations with** polyphen
> **Filter Variations** by predicted effect polyphen damaging
> **Output**:
> *"You should annotate the prediction from POLYPHEN notation before filtering by a POLYPHEN effect"*
> *"You should annotate the prediction from POLYPHEN/SIFT before reporting the POLYPHEN/SIFT effect"*

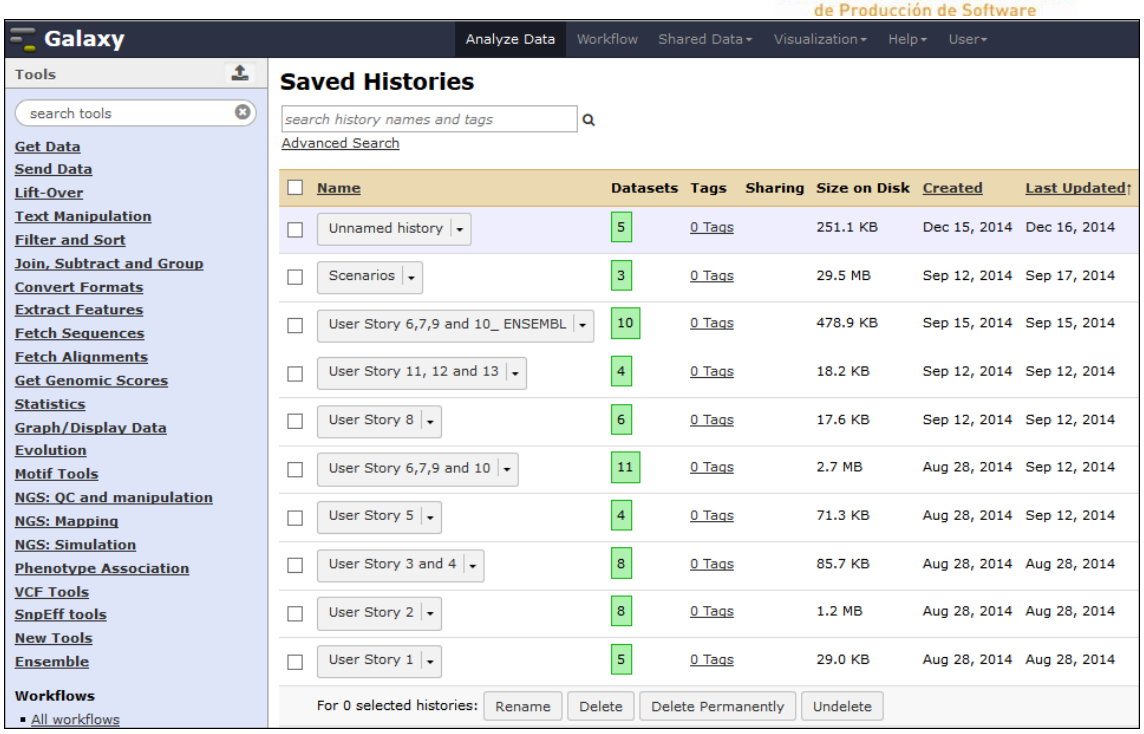
- **Filter/Prioritize range by SIFT/POLYPHEN/MAF must be between 0 and 1**. "The attributes minValue and maxValue of MAFF/ScoreF must be >=0 and <=1". If this restriction is not fulfilled, the custom error messages is: "*The range to filter variations is outside limits, it should be amgon 0 and 1*". This restriction is described in the Validator xtend class.

> **Input:**
> **Diagnose** Diabetes Mellitus Type 2 (Analysis 1)
> **Read Variations genotypes from VCF file** Patient1.vcf
> **Annotate Variations with** polyphen
> **Prioritize variations Variations** by Polyphen effect tolerated [0.15,2]
> **Report variations** with hgvs
> **Output**:
> *"The range to filter variations is outside limits, it should be between 0 and 1"*

Target Platform Tests:

These tests are created in a local Galaxy Server

User Story	Test Files	Galaxy Workflow	Tools
Read Genotypes	10Variants.vcf	US1	InputFile, CheckCustErrors
Annotate Transcript	1Exon.vcf, 1Intron.vcf, 10Variants.vcf, 4Variants.vcf	US2	VEP
Annotate SIFT	1Missense.vcf, 1Intron.vcf, 10Variants.vcf, 4Variants.vcf	US3_US4	VEP
Annotate POLYPHEN	1Missense.vcf, 1Intron.vcf, 10Variants.vcf, 4Variants.vcf	US3_US4	VEP
Annotate MAF	10Variants.vcf, 4Variants.vcf	US5	AlleleFrequencies
Filter SIFT Effect	4VariantsAnnotatedSIFT.vcf, 3VariantsAnnotatedSIFT.vcf	US6_7_9_10	SiftAndPolyphen
Filter POLYPHEN	4VariantsAnnotatedSIFT.vcf	US6_7_9_10	SiftAndPolyphen
Prioritize by MAF	4VariantsAnnotatedMAF.vcf, 10UnrealVariantsAnnotatedMAF.vcf	US8	MAFSort
Prioritize by SIFT	4VariantsAnnotatedSIFT.vcf	US6_7_9_10	SiftAndPolyphen
Prioritize by POLYPHEN	4VariantsAnnotatedSIFT.vcf	US6_7_9_10	SiftAndPolyphen
Report MAF	4VariantsAnnotatedMAF.vcf	US11_12_13	ShowVariations
Report SIFT	4VariantsAnnotatedSIFTPOLY.vcf	US11_12_13	ShowVariations
Report POLYPHEN	4VariantsAnnotatedSIFTPOLY.vcf	US11_12_13	ShowVariations



Generator Test Example

Example: TestFilterPolyphenEffect

```
@Before
def void testSetupOnce() {
DiagnosisPackage.eINSTANCE.eClass();
diagnosis = parser.parse ('''Diagnose Diabetes Mellitus Type 2
(Analysis 1)
Read Variations genotypes from VCF file Patient1.vcf
Annotate Variations with gene, transcripts, polyphen
Filter Variations by genes {ABCC8, CAPN10, KCNJ11, GCGR, SLC2A2,
HNF4A, INS, INSR, PPARG, TCFl2, ADIPOQ, AKT2, PAX4, MAPK81p1,
GPD2, MNTR1B}
Filter Variations by predicted effect polyphen damaging
Report Variations with gene, predicted_effect''')
fsa= new InMemoryFileSystemAccess()
generator.doGenerate(diagnosis.eResource, fsa)
}
@Test def testFilterPolyphenEffect(){
Assert.assertFalse("The workflow fragment of
filterByPolyphenEffect is different to the generated one",
checkGeneratorGalaxy(fsa.getTextFiles().values(),
"Galaxy_Fragment_PolyphenEffect.txt", "ensembl_id"))
}
```

Compiler example:

```
def steps(Resource resource)'''
            «/*PatientData */»
```

```
            «var patient=new PatientDataGenerator()»

      «patient.readPatientData(resource.allContents.toIterable.filter(Patien
tData).get(0))»«//Only one patient at the moment

            /*Analyses */»
            «FOR Analysis a:resource.allContents.toIterable.filter(Analysis)
SEPARATOR ','»«
                  var analysis=new AnalysisGenerator()»«
                  analysis.runAnalysis(a)»«
            ENDFOR»
            ,«/*Report */»«
            var report=new ReportGenerator()»«

      report.generateReport(resource.allContents.toIterable.filter(ReportVar
iations).get(0))»'''//Only one reportVariations at the moment

}
def annotateVaritionsWithVCFTools(boolean maf)'''«
      var step=galaxy.getLastStep+1»
      "«step»": {
            "annotation": "Annotate MAF",
            "id": «step»,
            "input_connections": {
                "input": {
                "id": «galaxy.getLastWorkflowStep»,
                   "output_name": "output"
                }
             },
                "inputs": [],
          "name": "Allele Frequencies",
          "outputs": [
              {
                  "name": "output1",
                  "type": "tabular"
              },
              {
                  "name": "output",
                  "type": "vcf"
              }
          ],
          "position": {
                        "left": «(step)*200»,
                        "top": «(step)*200»
                    },
          "post_job_actions": {
              "HideDatasetActionoutput": {
                  "action_arguments": {},
                  "action_type": "HideDatasetAction",
                  "output_name": "output"
              }
          },
          "tool_errors": null,
          "tool_id": "allele_frequencies",
```

```
            "tool_state": "{«
             »\"__page__\": 0, \"input\": \"null\",
\"__rerun_remap_job_id__\": null, «
             »\"mafOption\": \"{«mafTranslator(maf)»}\"«
             »}",
            "tool_version": "latest",
            "type": "tool",
            "user_outputs": []
        }«
»'''
def mafTranslator(Boolean maf)'''«
             »\\\"mafFieldname\\\": \\\"«maf»\\\",«
             »\\\"mafCheckbox\\\": \\\"«IF maf»True«ELSE»False«ENDIF»\\\", «
             » \\\"__current_case__\\\": «IF maf»0«ELSE»1«ENDIF»«
      »'''
```

Complete implementation:

http://personales.upv.es/mavilde2/PhD/TechnicalReports/iteration3.zip

Testing

Testing Questionnaire:

Analysis

Assess if the DSL complies with all the requirements described in user stories.

1. **Did you find in the language any erroneous step/instruction? (Coverage)**
2. **Did you find in the language any step that contains some erroneous aspect? (Coverage)**
3. **Did you miss any essential step/instruction? How important is it for the usage scenario? (Coverage)**

Syntax:

Assess if the DSL syntax fulfills the preferences of end-users.

4. **Would you add, change, remove or reorder any word of the language? (Expressivity)**
5. **Is the language easy to understand (Expressivity)?**
6. **Is the language intuitive to use (Expressivity)?**
7. **Did you find a combination of words that were incorrect but they could be written with the DSL (Coverage)?**

Semantics:

Assess if the DSL semantics is well specified.

8. **Did you find a combination of steps that were incorrect but they could be written with the DSL (coverage)?**

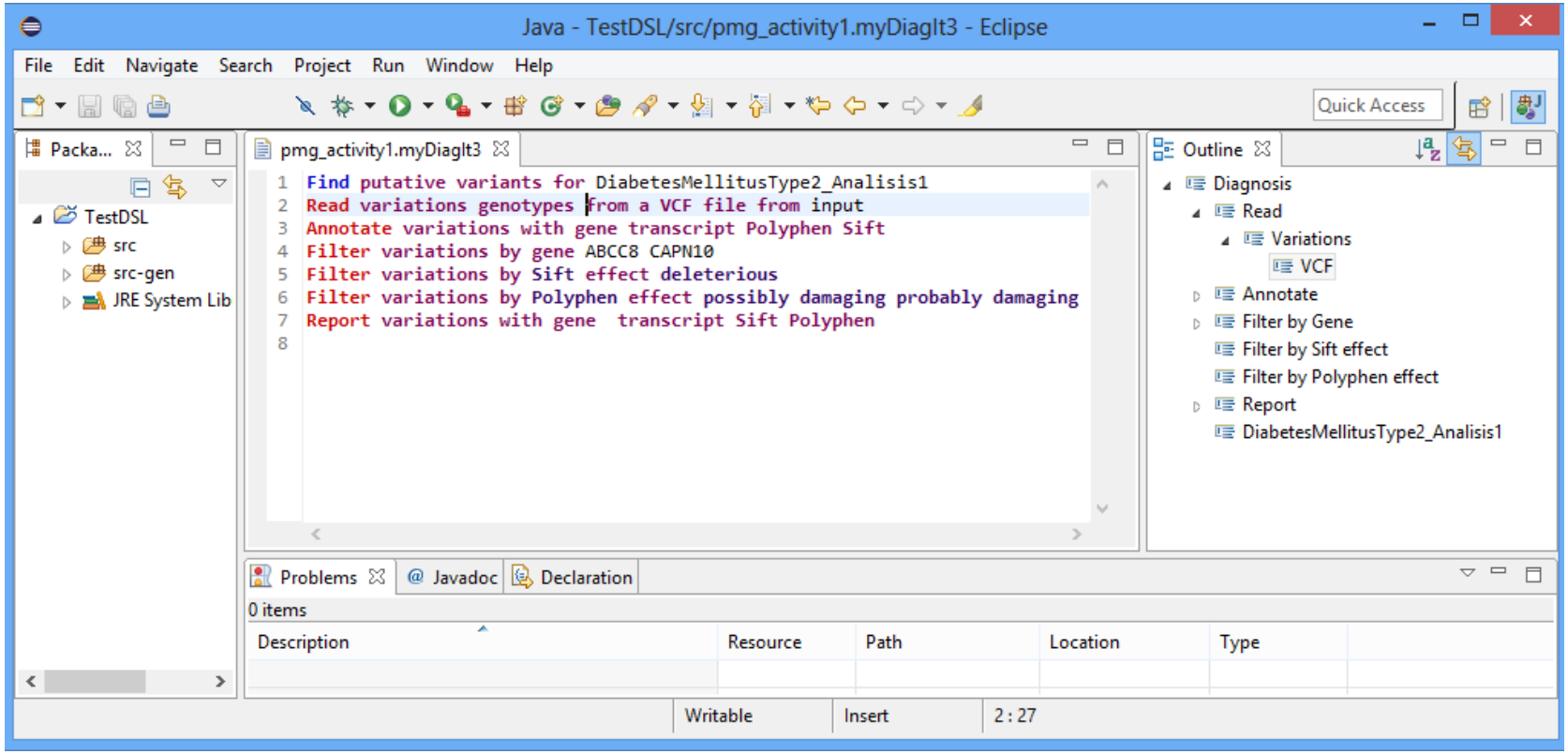
9. **Did you find any step that was dependent of another one but it could be written with the DSL alone (coverage)?**
10. **Did you detect any other error that could be written with the DSL editor?**
11. **Do you know new software that suits better to implement any user story of the DSL (Completeness)?**
12. **Is the Galaxy workflow generated equivalent to the specification written with the DSL?**

Implementation:

Assess if the DSL editor represents the syntax and semantics design.

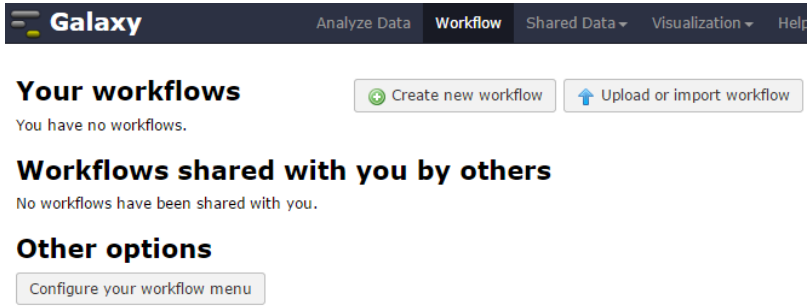
13. **Did you find any problem/error while using the DSL editor (Completeness)?**
14. **Were you informed with any message that you did not understand (expressivity)?**
15. **Was there any problem with the creation of the .ga file?**
16. **Was there any problem uploading the .ga file into Galaxy?**
17. **Did you find any error after executing the Galaxy workflow (completeness)?**
18. **Did you miss help by the DSL editor in a specific manner?**

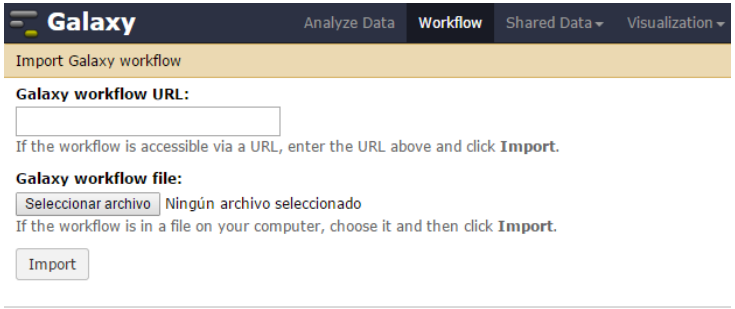
Demonstration Screenshots

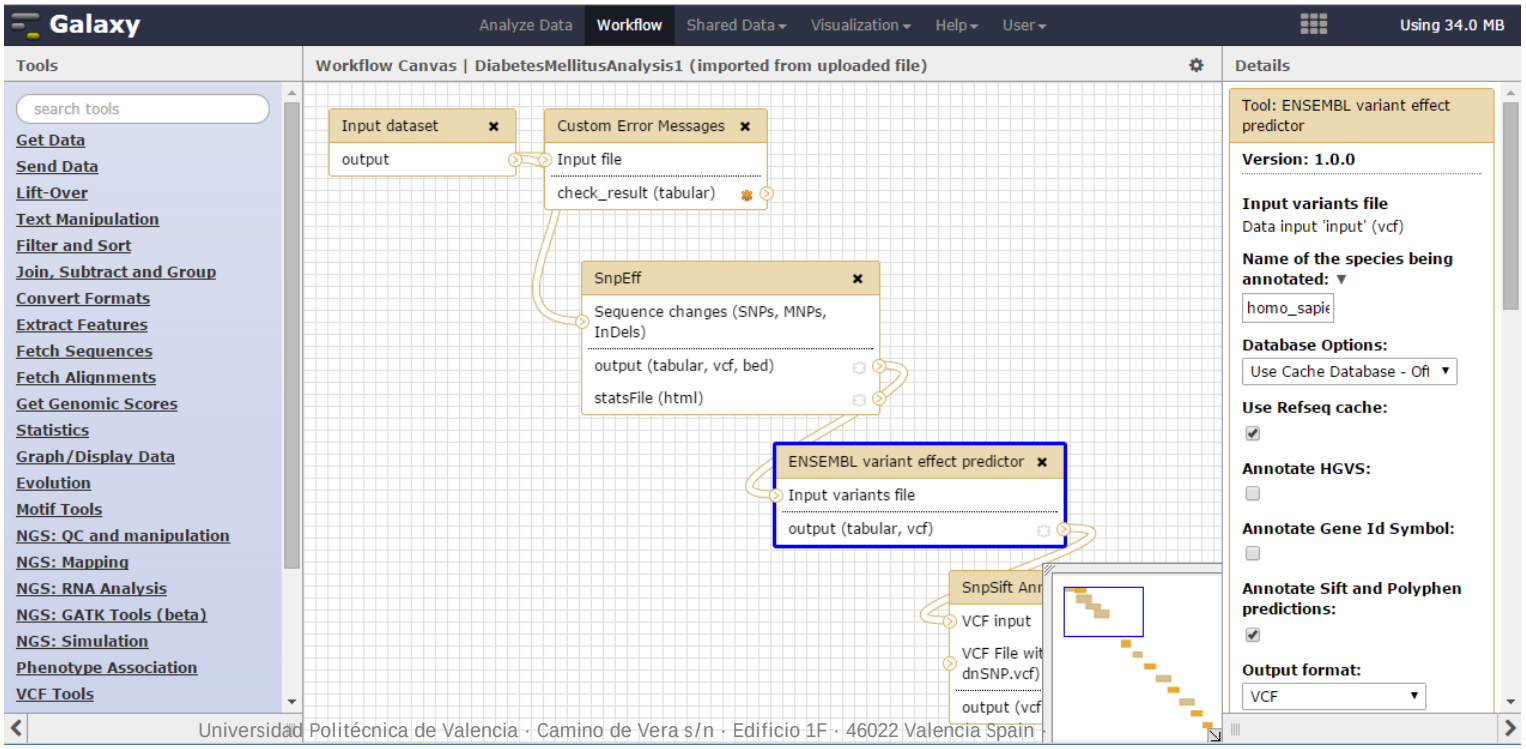


UNIVERSIDAD POLITECNICA DE VALENCIA

PROS
Centro de Investigación en Métodos de Producción de Software







Universidad Politécnica de Valencia · Camino de Vera s/n · Edificio 1F · 46022 Valencia Spain ·
T. +34 96 387 70 07 Ext. 83530 · M. + 34 619 543 623 · F. +34 96 387 73 59 · info@pros.upv.es · www.pros.upv.es

Universidad Politécnica de Valencia · Camino de Vera s/n · Edificio 1F · 46022 Valencia Spain ·
T. +34 96 387 70 07 Ext. 83530 · M. + 34 619 543 623 · F. +34 96 387 73 59 · info@pros.upv.es · www.pros.upv.es