UNIVERSIDAD POLITECNICA DE VALENCIA

PROS Centro de Investigación en Métodos de Producción de Software

Informe Técnico / Technical Report

Software for the Genetic Analysis Domain

Oscar Pastor, Francisco Valverde, and Maria Jose Villanueva

Ref. #:	ProS-TR-XXXX				
Title:	Software for the Genetic Analysis Domain				
Author (s):	Oscar Pastor, Francisco Valverde, and Maria Jose Villanueva				
Corresponding author (s):	opastor@dsic.upv.es fvalverde@pros.upv.es mvillanueva@pros.upv.es				
Document version number:		Final version:		Pages:	
Release date:					
Key words:					

Content

1 Introduction

The main purpose that promoted the research of the existent commercial alignment tools is to learn the different functionality that this kind of tools are offering. In order to accomplish this target will be necessary to obtain the differences between them and to find the strongest points and deficiencies of each one; but mostly which functionality may be missing in all of them.

Having a better comprehension about what has been already developed, what is actually being used and what are the needs of the users of that tools, could clarify if the objectives of the present project of the Genoma group could be a real contribution to the field.

The start point of this project had as a main objective to create a software that meets the requirements of the biologists when performing a DNA analysis from a patient sample searching for mutations that may cause some disease. As a consequence the software we intend to develop will try to cover all the activities involved on the process in order to provide a complete functionality that this experts claim all the commercial tools lack.

First of all it is essential to establish and delimit all the activities that the mentioned process comprises. However, inside this collection of activities we will find that some of them cannot be controlled by our software or cannot be automated by any software; activities like the sequencing the DNA sample and decision-making the correctness of the basecalling respectively.

The first step while analyzing a sample of a patient is to extract a very small fragment and perform the sequencing. This activity is done by the sequencer machine and it is not an activity that the software will take into account. However, the output of this activity, files in .ABI (Applied Biosystems Inc.) format, will be the input of the software. The sequences, the samples and the reference can be expressed in this format but also in the formats .SEQ, .GB (GeneBank) or FASTA. Before getting deeper in details about the analysis decomposition, it is important to emphasize that normally the analysis is restricted in only one gene at a time. Moreover, due biological difficulties on the sequencing process the DNA sample is sequenced in pieces about 800 bp (called contigs). For this reason, sometimes it is sequenced only the region of interest for the analysis, but at least it is sequenced twice (one reverse and other forward direction) in order to ensure the accuracy of the result. From both sequences it is obtained a consensus sequence.

DNA sequence analysis that the new software will have to manage consists on several phases:

1. Assembly each contig into its correct position: Independently of the number of contigs sequenced each one has to be located properly.

2. Clean the results from the sequencer machine: Using the tool and its biological knowledge biologists perform manually a cleaning on the basecalling of the contigs. Experts have to decide, aided with the reference sequence, if the basecalling of each sample and the consensus sequence reflects the veracity of the original sample, correcting thus all the errors that the sequence machine has introduced. The tool will have to perform as well a dropping of the beginnings and ends of sequenced contigs that normally are not useful for analysis because of the quality of the signal. This last cleaning is referred as "trimming".

3. Compare consensus with the reference sequence searching for variations: Each difference in the consensus sequence respect the reference sequence will be considered as a variation. The tool will have to search for insertions, deletions and indels. Heterozygosis (two different signals codifying two bases in the same position) is also important for the detection of variations.

4. Search what does mean each variation: One variation may be provoking some phenotype depending on the base changes and the position of them. For each possible variation-phenotype pair it would be crucial to know the first publication that supports the finding.

In this review there has been analyzed the following tools: Sequencher, SeqScape, Mutation Surveyor, CodonCodeAligner, Polyphred and InSNP. The following 6 sections will analyze the essentials of each tool concerning the installation and use of the tool, the variations detected in comparison with the conceptual model, the possible connection with bibliography and some other interesting features. The last two sections will contain the comparison between tools and the conclusion and recommendations of the author of this review. In the anex it is explained a firts contact about the behaviour of all tools under the same conditions.

2 Sequencher

Sequencher is a proprietary Software of Gene Codes Corporation [3] that offers the possibility to introduce sequences, perform assemblies, the exploration and edition of sequences in connection with its reference and finally the detection of the variations that differ from this reference.

Gene Codes Corporation offers a Demo version with restricted functionality and the possibility to purchase all the software with a license that once obtained it never expires.

It is available for MAC and Windows and it is possible to install it on a single computer or install the network-enabled version, with a server and its clients.

In order to help the users to familiarize with the tool there is a collection of easy and complete tutorials on its webpage (`http://www.genecodes.com`).

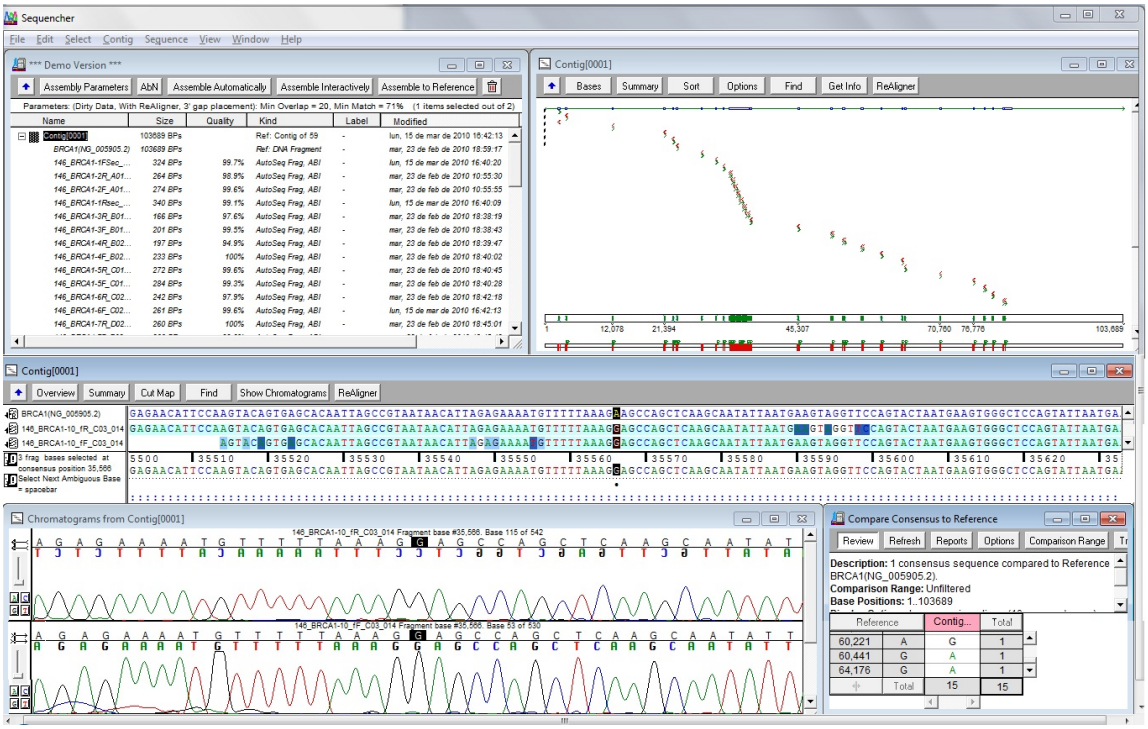


Figure 1: Different windows on Sequencher

Sequencher provides different views of the data (See Figure 1): files assembly (left and up window), providing the list of files containing the sequences assembled in the same contig; distribution of contigs around the gene (right and up window), displaying the reference sequence

(blue line) and all the contigs situated in its position (green lines for forward and red lines for reverse); discovered variations report (right and down window), describing the found changes in several fields inside a table; bases of each sequence (middle window), situating each sequence sequentially and the chromatograms of the samples (left and down window). When a variation is double clicked the bases involved are highlighted in the bases representation and in the chromatograms.

2.1 Types of detected mutations

Insertions, *deletions* and *indels* in homozygosis are detected and all of them are expressed base by base, as variations of length 1. If, for example, there is an insertion of 3 nucleotides, the variation is expressed as 3 insertions of 1 nucleotide.

However, *insertions* and *deletions* in heterozygosis cannot be detected. When they occur, the chromatograms of the contigs seem a mix of two signals that should be almost identical but now appear shifted several positions. This software detects *indels* in heterozygosis because this shift does not happen.

In order to identify a heterozygosis base (two different values in the same position) and differentiate it from a homozygosis one with noise, a sensibility value can be fixed. This sensibility is expressed in percentage and when the presence of two significant overlapping fluorescence peaks occurs in a concrete position, it represents the relation between the signal of lower intensity respect the other.

2.2 Return formats

All the sequences included the consensus sequence (once all contigs are assembled and cleaned) can be imported and exported in several formats: in plain text as ASCII plain format or unformatted (bases only); in specialist and legacy formats like AFDIL, Genentech, IG and Strider; in databases formats as GenBank, NBRF and EMBL; in commonly used formats like FASTA (normal and concatenated) and GCG; in phylogenic program formats like NEXUS/PAUP (interlieved and sequential), Phylip (normal, 3 and 4) and even in standard chromatogram formats like SCF (2.0 and 3.0). For some of those formats can be indicated as well the following options: export with upper/lower case, select the orientation of the sequence and leave or remove the gaps.

When there are variations in the sample they are reported in a table. The report gathers the data about position and value of the base on the reference, value of the base on the sample and number of different samples that have this variation. This variation table can only be exported in TXT and PDF.

2.3 Connection with bibliography

It is not possible to make any connection with data about variations neither in the reference sequence nor the variations table.

2.4 Other interesting features

During the import and assembly of contigs Sequencher offers the option of classify the samples automatically only using the name of the file. The samples can be divided in forward and reverse and at the same time in different specimen. It also allows the simultaneous analysis of several patients and all the differences among them would be reflected in the reports.

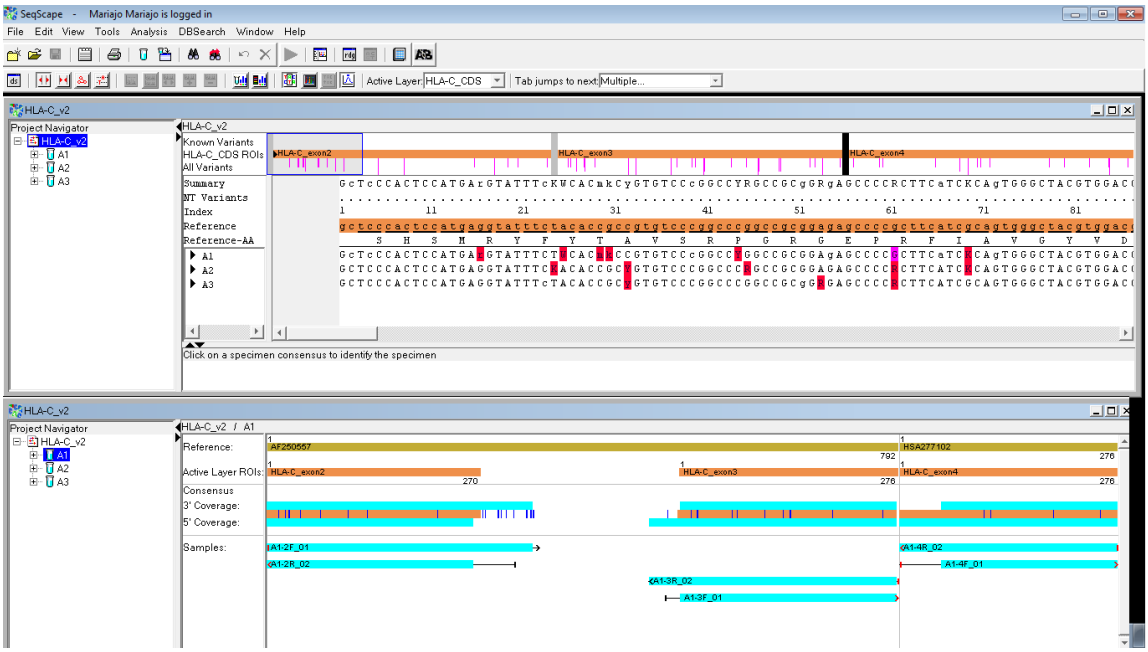


Figure 2: SeqScape window

3 SeqScape

SeqScape is a proprietary Software of Applied Biosystems [1] without demo available. Its functionality goes from importing, assembling, editing and analyzing samples searching for variations until comparison of the segments searching for patterns.

It is available only for Windows and can be configured a manage access security control with the use of logins and passwords allowing three different permission profiles. In addition to the security, it can prepare documentation for future audition, that is, it is possible to program some events to be recorded.

SeqScape changes the methodology of operation from actions-objective, where for each objective that the user want to achieve has to perform several concrete actions, to configuration-objectives, where it is necessary to configure all the needed options before performing any action and once all configure all of the objectives are executed together. Then it is necessary to learn how to configure the projects. It has several important configuration phases that finally will lead to one button that executes all the functionality performed in one step. Then the user will search for the data that wants to know in each moment (See Figure 2).

3.1 Types of detected mutations

Insertions, *deletions* and *indels* in homozygosis are detected and SeqScape claims that it also detects *insertions* and *deletions* in heterozygosis since the version 2.5. However the version provide by IMEGEN is the 2.0 and this kind of mutations can not be located. In version 2.0 when they occur these contigs are not included in the assembled ones. Still *indels* in heterozygosis are detected and it is possible to fix a sensibility value like was possible in Sequencher.

3.2 Return formats

The consensus sequence can be exported as FASTA, SEQ and QUAL format. It has the option of replacing unknown bases with a desired symbol, in case of gaps or bad quality bases.

The formats for exporting the reports about the variations are TXT, HTML, PDF or XML. The exported files will contain some analysis variables for each specimen (success of analysis, specimen score, mutations found) and information about variations (sample, position, size).

3.3 Connection with bibliography

It is possible to add known variations to the reference sequence in order to the SeqScape to be able to identify them as known or unknown. It is also possible to automate this introduction of data by creating an XLS file that will contain all the fields needed to describe each variation. The variations introduced can be classify as insertions, deletions, or basechange, indicating the ROI, the position, the reference base(s), the variant base(s) and its description.

3.4 Interesting features

In addition to all the security control around the access it is as well possible to export data signed electronically.

When the reference sequence has been imported from GeneBank, SeqScape creates regions of interest and layers that separates the different exons automatically.

In addition it exists the possibility of create libraries searching for patterns. A library is a collection of several segments (alleles, genotypes and haplotypes) with a fixed length of a region of interest (ROI). The consensus sequence is compared with the library searching for matches on any segment.

4 CodonCodeAligner

4.1 Types of detected variations

CodonCodeAligner is propietary software from CodonCode Corporation [2] used for sequence assembly, contig editing, and mutation detection.

CodonCode CorporationIt makes available a 30 day-demo that can be tryed under Windows and Mac.

Its interface offers a visualization (See Figure 3) of the assembled contigs and the reference sequence simultaneously with the codifying regions of the gene or (depending of the user preferences) the visualization of the differences between bases.

4.2 Return formats

CodonCodeAligner can export the samples with the formats FASTA and SCF but the consensus sequence only in FASTA (Normal bases or haplotypes). For both of them exists the options of including gaps, append the comments, replace problem characters in names and write FASTA quality files. If the target sequences are the disposition inside the whole assembly, it is possible to save it into an ACE project, a NEXUS/PAUD (interleaved and sequential) or a Phylip (interleaved and sequential) format.

The variations found are gathered in a report that can be exported in TXT and PDF. This report will contain the feature, the source, the type of source where the mutation has been founded, the parent contig, the start, the end, and the content.

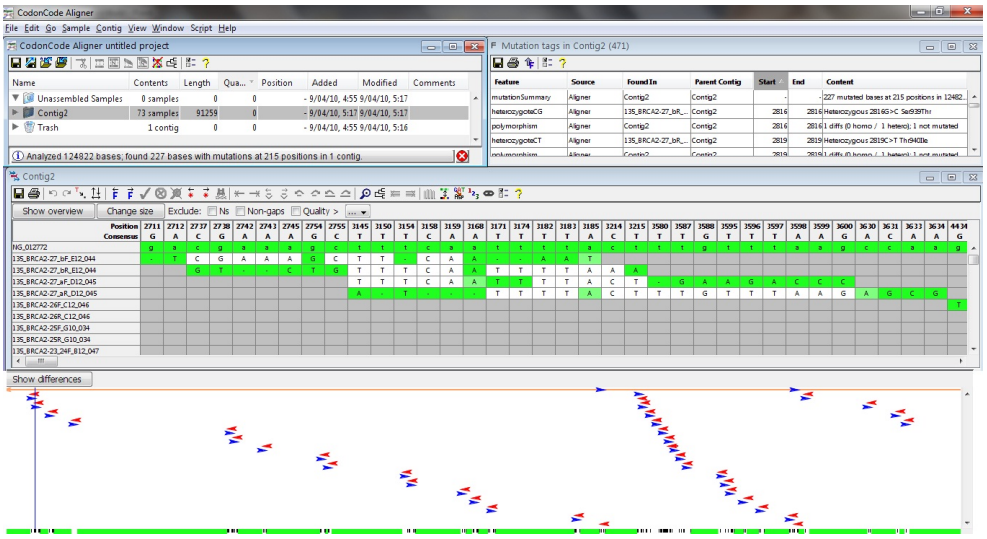


Figure 3: CodonCode Aligner

4.3 Connection with bibliography

There is no option to add any information about variations, bibliography or phenotypes.

4.4 Interesting Features

Graphical interface achieves the complete navigation along sequences, edition of sequences allowing all types of operations and visualization of all required data simultaneously.

Furthermore the reference sequence can be downloaded automatically from GeneBank by only indicating the accession number.

5 Mutation Surveyor

Mutation Surveyor is a proprietary Software of SoftGenetics [8] that compares several samples with a reference sequence searching for variations.

SoftGenetics offers a Demo version, the possibility of trying a fully functionality 30 days trail and a free training for genetic experts that will be the users of the tool.

The software is available for Windows (NT, 2000, XP, Vista and 7) and MAC (although it is necessary a file converter to PC files for those files that will be input of the software) and it is possible to install it on a single computer or install the network-enabled version (with a server and its clients).

The user-interface has been designed following Microsoft platform design guides to be a user-friendly software (See Figure 4). Different views compose the structure of the interface: the text view (right) and the graphical view (left), navigating easily along them.

There is an additional software called Mutation Surveyor Autorun that that permits the unattended analysis of multiple projects and a Log File Editor to configure the parameters of the hole unattended process.

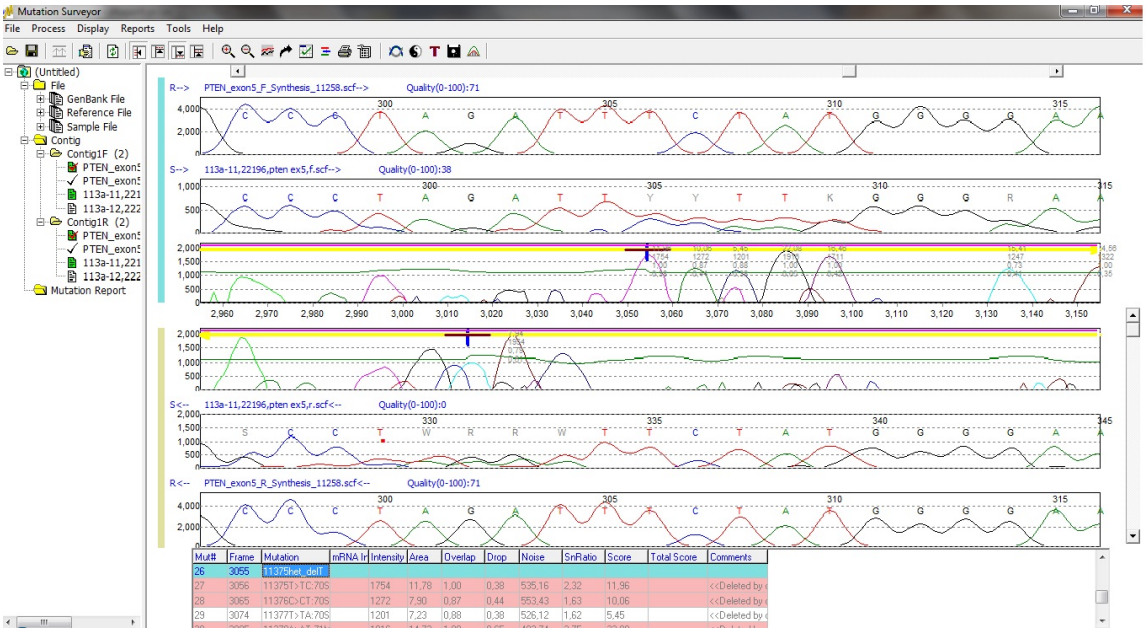


Figure 4: Mutation Surveyor views

5.1 Types of detected mutations

Insertions, *deletions* and *indels* homozygosis are detected and all of them are expressed with its correct length. Are also detected *indels* in heterozygosis and a sensibility value, now called dropping factor, can be fixed to find the heterozygosis bases.

Regarding *insertions* and *deletions* in heterozygosis, it has the ability to "identify heterozygous indels down to 5% of the primary peak". The sample is decomposed into two different samples representing both strands (one from the father, one for the mother). Then one of those is shifted and displayed below according the mutation detected in order to match the reference sequence. Sometimes due the nature of this kind of variations is difficult for the software to identify automatically the start point. For this reason, it allows the expert to verify if the obtained position is the correct one.

5.2 Return formats

It is not possible to export the consensus sequence because this tool does not create any due the fact it searches for variations directly in each forward and reverse sample separately.

Per contra, there are a lot of possibilities of export and configure reports about the variations found. An standard report could contain the extra data: number, sample and reference file names, direction of the GeneBank reference sequence, reading frame, start and end of the sample, quality, a mutation code for each one found and several more information. All this information can be exported in TXT, XLS, HTML or XML.

The table with the mutations appeared on each sample can be exported only in a TXT file.

5.3 Connection with bibliography

It is not possible to link a mutation with a phenotype and neither with a bibliographic reference. In contrast, it is possible to add the positions and the values representing the possible changes of known variations to the reference sequence. The annotation of variations allows the introduction of the position, the gene name and the reference to the database. In each variation it can be added as well information about several alleles, indicating the population, the value of the base and the frequency.

5.4 Interesting features

Mutation surveyor downloads itself from GeneBank (using a saved URL) the reference sequence of the samples if no reference sequence is introduced. However the reference obtained is the mRNA sequence instead of Gene sequence.

When comparison between the samples and the reference sequence is being showed the mutation report table navigation is instanteneously coordinate with this comparison.

It has an additional representation, apart from the sample and the reference sequence, that allows a faster visualization of the variations. Together with the samples and its references appears a new frame in the middle of the window that calculates the difference of signals among the reference and the sample. For each kind of occurrence it has a color and a shape meaning. In this frame it also appears several scores that evaluate the potential variation found.

The reports are highly customizable adding and dropping all types of information: data about formats, about filters (indicating witch variations has to be included or grouped), about displaying, about variations, about colors and about nomenclature. Regarding the nomenclature, variations found can be expressed in different ways: Genomic, relative to CDS, relative to mRNA and HGVS nomenclature.

6 Polyphred

Polyphred is an academic and non-profit software created by researchers of the Department of Genome Science from the University of Washington [7].

It is free available under request or purchasing a commercial license for business use. It is available for the platforms Linux (x86 and x86-64) , Mac OS X (PowerPC and x86), Cygwin, Solaris (x86 and SPARC), SGI IRIX, Compaq Tru64 Alpha (formerly DEC OSF/1), HP-UX (PA-RISC) and AIX 5 (PowerPC). Polyphred is actually a integrator software of the existent tools Phred, Phrad and Consed. It uses Phred to perform the basecalling and the extraction of peak information, Phrad for sequence alignment in order to create a consensus sequence and Consed for graphical inspection of potential heterozygosis indels (See Figure 5). In order to compare the samples with a sequence of reference it is necessary an additional software, called Sudophred, that converts the formats to be the adequate one for Polyphred requirements.

6.1 Types of detected mutations

Polyphred is designed to identify *single nucleotide substitutions* which results on the identification of indels of length 1 in both homozygosis and heterozygosis. It is not possible to add a sensibility value because it shows all variations and scores it according its confidence, that is, depending on the fluorescence of the second signal.

It is not designed to detect insertions or deletions in any case. For this reason does not exist the problem of the sequences mixed due if it is produced a variation there is not any shift in the sequence.

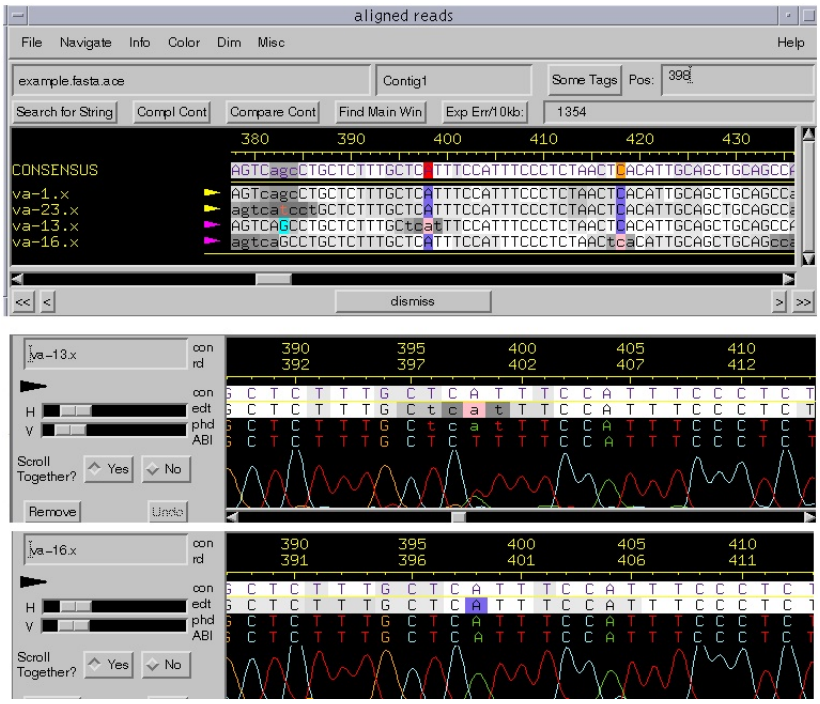


Figure 5: Polyphred windows

6.2 Return formats

Each of the programs which conforms Polyphred outputs its results in a file with different formats that lately the software will employ to do its activities. Phred outputs its data in PHD and POLY Files, Phrad uses this files and outputs its results in ACE files and finally Polyphred configures the output with these files using Consed.

Moreover the resultant report file is a TXT file divided in blocks demarcated by several tokens: POLY, GENOTYPE, COLUMNGENOTYPE, INDEL, POLYINDEL, COLUMNINDEL, MANUALGENOTYPE, VERIFIED, MICROSATELLITE, SAMPLE and COVERAGE. This structuring in blocks allows to export all the data also in XML.

6.3 Connection with bibliography

It is not possible to add any information to the sequences, so information about the located indel by the software and about its frequency in the population cannot be related.

6.4 Interesting features

Polyphred can be customized in order to fit the requirements of the user but it has to be done by configuring a Linux file with specific notation.

7 InSNP

InSNP is a free software, with free registration requirement, created by Mucosa Research Group [5, 6] which main objective was the improvement of the suite of programs Polyphred giving support for Windows users. As a consequence, at the moment it is only available for Windows.

This software is focused only on the detection of mutations SNPs and heterozygosis indels.

The main window of the InSNP can be observed in Figure 6. At the top is a line representing the entire reference sequence with arrows to indicate the forward and reverse primers. Below this line are the representations of amplicon frames, reverse and forward. The selected base is shown in the grouped traces frame (left and down) and the selected group in the middle can be visualized scrolling the frame. On the right, are situated all the buttons for all the available operations.

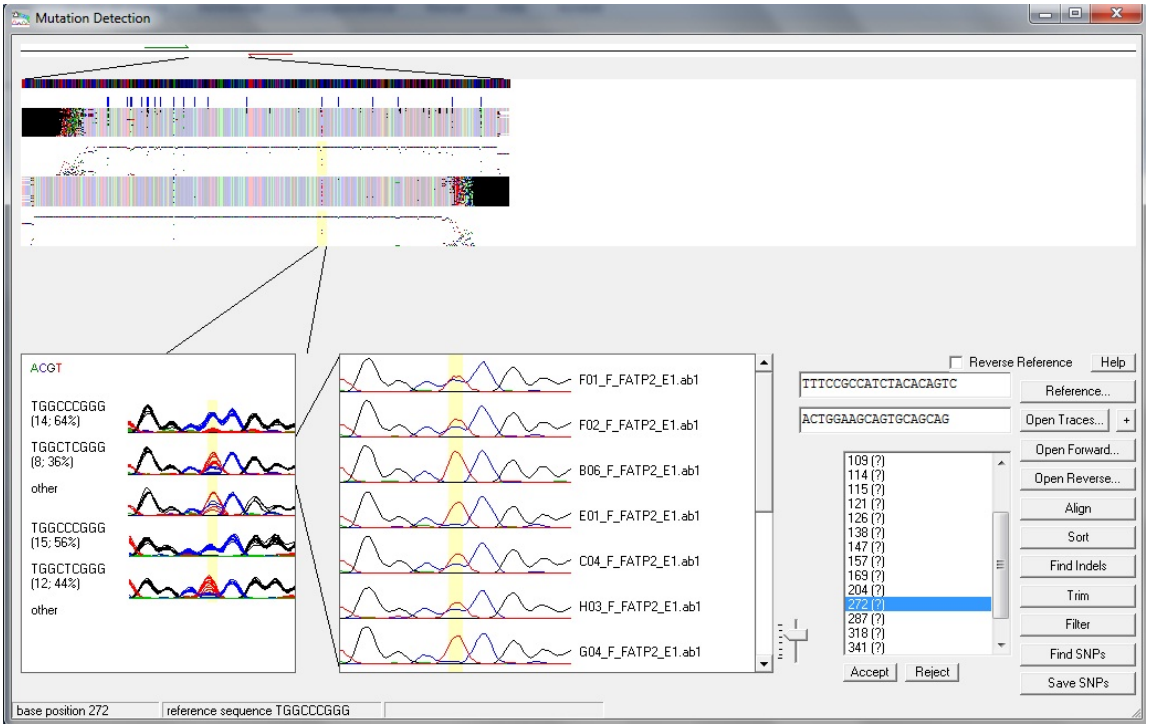


Figure 6: InSNP main window

7.1 Types of detected mutations

The location of variations is actually focused on variations in heterozygosis. For that reason it is able to locate in the first place those *single nucleotide polymorphisms* (*indels* of length 1) in heterozygosis. Applying the same operations for this type of variations location, it is possible to locate *indels* in homozygosis.

In addition InSNP locates the *insertions* and *deletions* in heterozygosis. This detection is based not in the reference sequence but in the own traces when there is a high ratio of number of peaks in several following bases. Then the initial position is located although the length has

to be obtained by the expert. InSNP helps with an additional window that decomposes the sequence in several signals showing the reference sequence at the same time.

7.2 Return formats

It is only possible to save the Indels and the SNP in TXT format including for each one its initial and final position, the replacement occurred and the calculated frequency where the variation arise respect all the samples analyzed.

7.3 Connection with bibliography

It is not possible to make any association of data to the variations found. The reference sequence cannot neither be modify in order to add information.

7.4 Interesting features

The most interesting features of InSNP reside on the user interface. The authors claim to present an innovator way of sequence representation offering two new display manners.

The first one is called sequence chart, in there the samples are represented using pixels, and depending on the color, it means one base or another. Using this representation it is possible to codify the sequences up to 128 times as dense as the character representation.

The second one is called frequency chart and it is used to locate mutations in a visual way. It consists in a representation of a pixel on a XY axes where X means the frequency of the samples that match the value of the reference and Y the position of the base. For that reason, when there is matching of several bases it appears a line and when there is a variation, the frequency is lower than the previous and the following bases giving then a visual effect of a broken line in that point.

8 Future Research

The analysis of the different tools will require a second sight of the functionality concerning the bibliography, studying what exactly can be done with the tools where it is possible to annotate sequences (SeqScape and Mutation Surveyor).

Finally the study about a 7th tool, Staden, has been obviated for lack of time.

9 Comparison

In first place the Figure 7 offers a comparison between all the studied tools.

As can be observed on the comparison of tools table, regarding the functionality some of the tools, Mutation Surveyor, Polyphred and InSNP, have only the functionality of searching and reporting variations. On the contrary, the remaining ones are worried about the quality of the samples, allowing the biologists to take part in the process, and also about the location and reporting of the variations found on the samples.

A performance analysis between Mutation Surveyor, Polyphred and InSNP is shown in Figure [6]: Falta

	Secuencher	SeqScape	Mutation Surveyor	CodonCode Aligner	Polyphred	InSNP
Review of Samples						
Trimming	√	√	×	√	×	√
Editing of samples	√		×	√	×	×
Several sequences Support						
Identification fw/rev	√	√	√	√	×	√
Comparative display (samples & reference)	√	√	× (not in the same window)	√	√	√ (with dots not chromatograms/bases)
Assembly	√	√	√	√	√	√ (primers needed)
Among samples	√	√	×	√	×	×
In relation reference sequence	√	√	√	√	√	√
Search of variations						
Homozygosis	ALL	ALL	ALL	ALL	NOT ALL	NOT ALL
Insertions	√	√	√	√	×	×
Deletions	√	√	√	√	×	×
Indels	√	√	√	√	√ (only SNP)	√ (only SNP)
Heterozygosis	NOT ALL	ALL	ALL	ALL	NOT ALL	ALL
Insertions	×	√ (Since v2.5?)	√	√	×	√
Deletions	×	√ (Since v2.5?)	√	√	×	√
Indels	√	√	√	√	√ (only SNP)	√ (only the start position)
Connection Bibliography						
Annotation sequences	×	√	√	×	×	×
Report formats						
PDF	√	√	×	×	×	×
TEXT	√	√	√	√	√	√
XLS	√ (Copying rows and columns manually)	√	×	×	×	×
XML	×	√	×	×	√	×
HTML	√	×	×	×	×	×
OTHERS	×	×	×	×	√ (PHD, POLY and ACE)	×

Figure 7: Tools Comparison

10 Conclusion

Once analyzed some of the available tools for DNA analysis it is possible to determine which essentials need to be solved in this domain.

By one hand biologysts require one tool that integrates all the process or at least several interconnected tools performing all the analysis. In any case, all the activities included in the analysis process have to be covered and controlled by them at any moment. By the other hand, there is a lack in the coverage of the last activity of the process: to relate the variations found with a phenotype and the bibliography that supports this relation. Among the studied tools it is possible to find some approaches to this idea but none of them accomplish all the activity with the necessary detail. Some of the tools, SeqScape and MutationSurveyor allow the annotation of the reference sequence with variations but none of them allow the inclusion of data associated to these variations (phenotype and reference) or even the automatic insertion of new variations. In all the cases this activity does not seem to be highly developed or it is not explained or emphasized in the manuals and tutorials about the tools.

In summary there are three tools, Sequencher, SeqScape and CodonCodeAligner, that cover the three first phases of the analysis and only two that does an attempt to cover the last phase, SeqScape and MutationSurveyor.

There is a tool, CodonCodeAligner, that beats the other ones in relation to its interaction. CodonCodeAligner not only offers a very friendly interface but the possibilities of operations: perform changes, moves, visualizations, etc, and in a very intuitive way, helped by multiple icons on the toolbar.

However, due it it very tempting to choose an intuitive and flexible tool, in these case the functionality is the first requirement. That is why, althoug SeqScape has the most difficult methodology of analyisis, it is the most option configurable and the tool that covers more requirements and allows exportation into multiple formats. Additionally, considering that it is not sure 100% that SeqScape detects insertions and deletions of heterozygosis and neither the procedure of detection is unknown, it is highly recomended Mutation Surveyor for this concrete functionality, because the location and visual decomposition it is coincident with the requirements of biologysts.

References

[1] SeqScape Applied Biosystems. http://www3.appliedbiosystems.com/abhome/index.htm.

[2] CodonCodeAligner CodonCode Corporation. http://www.codoncode.com/aligner/.

[3] Sequencher GeneCodesCorporation. http://www.genecodes.com/.

[4] IMEGEN (Instituto MEdicina GEnómica). http://www.imegen.es/.

[5] InSNP Mucosa Research Group. http://www.mucosa.de/insnp/.

[6] Carl Manaster, Weiyue Zheng, Markus Teuber, Stefan Wächter, Frank Döring, Stefan Schreiber, and Jochen Hampe. Insnp: a tool for automated detection and visualization of snps and indels. *Human mutation*, 26(1):11–19, July 2005.

[7] Polyphred Department of Genomic Sciences (University of Washington). http://droog.gs.washington.edu/polyphred/.

[8] Mutation Surveyor Softgenetics. http://www.softgenetics.com/.

Genomic Analysis

Introduction

The genomic analysis addresses three steps: 1) first takes as a base a complete VCF (containing one individual or triplets); 2) then adds all relevant information to the VCF file (annotation process) and finally 3) depending on the analysis the VCF annotated is filtered according some criteria:

The relevant information they want to annotate is: 1) Structural Ids (normally from ENSEMBL) (and hgvs notation); 2) Snp ids (rs from dbSNP); 3) Population Allele frequency (from 1000G data or Exome Sequencing Project); 4) Coverage; 5) Pair Effect prediction-Transcript (Algorithm: SIFT, POLYPHEN or combined, and transcripts: all transcripts that are affected, transcript of the tissue where it expresses the most damaging consequence, most important transcript for disease); and 6) combined analysis with triplets (Phasing and Combined heterocigosis of several variations that affect at gene level).

And the different filtering criteria are: 1) Structural: Chr, Gene, Type of variation (ins, del, indel, MNPs, SNPs and CNVs) type of polymorphism (homozygosis, heterozygosis, combined heterozygosis, de-novo); 2) Position range; 3) Allele frequency; 4) Coverage; 5) Transcript; 6) Loss of function.

Annotations can be classified in three categories: 1) structural information of the variation, such as gene, exon, transcripts; 2) database information, such as the rs from dbSNP; and 3) effects in different transcripts, including scores of SIFT and POLYPHEN and the hgvs notation.

In order to retrieve these data and annotate the variation file, several options are available:

- Download database file manually –in GVF, GFF, BED formats- or using biomart (in TXT file) and annotate the file with this data afterwards.
- Run the suitable commands of the suites Annovar, SnpEff and/or VEP.

Retrieving and annotating data manually

A1. The GFF Format

The GFF format (Generic Feature Format Version) specifies genomic/genetic features (genes, exons, CDS, etc.) and their properties (name, sequence, dbxref, etc.) using some predefined fields and rules and also ontology terms. It is widely used by the community to annotate variations with structural data.

Example (current version 3):

```
0  ##gff-version 3
 1  ##sequence-region   ctg123 1 1497228
 2  ctg123 . gene       1000  9000  .  +  .  ID=gene00001;Name=EDEN
 3  ctg123 . TF_binding_site 1000  1012  .  +  . ID=tfbs00001;Parent=gene00001
 4  ctg123 . mRNA       1050  9000  .  +  . ID=mRNA00001;Parent=gene00001;
 5  ctg123 . mRNA       1050  9000  .  +  . ID=mRNA00002;Parent=gene00001;
 6  ctg123 . mRNA       1300  9000  .  +  . ID=mRNA00003;Parent=gene00001;
 7  ctg123 . exon       1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
 8  ctg123 . exon       1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
 9  ctg123 . exon       3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00003
10  ctg123 . exon       5000  5500  .  +  . ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
11  ctg123 . exon       7000  9000  .  +  .  ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
```

It can be downloaded from the NCBI ftp: ref_GRCh37.p13_top_level.gff3.gz (Note: I choose this file observing the name and assuming that it is the one that contains all the information)

A2. The BED format

The BED format is defined by the USCS also to describe annotations.

```
browser position chr7:127471196-127495720
browser hide all
track name="ItemRGBDemo" description="Item RGB demonstration" visibility=2
itemRgb="On"
chr7    127471196  127472363  Pos1  0  +  127471196  127472363  255,0,0
chr7    127472363  127473530  Pos2  0  +  127472363  127473530  255,0,0
chr7    127473530  127474697  Pos3  0  +  127473530  127474697  255,0,0
chr7    127474697  127475864  Pos4  0  +  127474697  127475864  255,0,0
chr7    127475864  127477031  Neg1  0  -  127475864  127477031  0,0,255
chr7    127477031  127478198  Neg2  0  -  127477031  127478198  0,0,255
chr7    127478198  127479365  Neg3  0  -  127478198  127479365  0,0,255
chr7    127479365  127480532  Pos5  0  +  127479365  127480532  255,0,0
chr7    127480532  127481699  Neg4  0  -  127480532  127481699  0,0,255
```

A3. The GVF Format

The format GVF (Genome Variation Format) is a specialization of the GFF format to describe variations relative to a reference genome. It is widely used by the community to annotate variations with database data.

GVFLinkGCF

```
##gvf-version 1.06
##genome-build NCBI B36.3
##sequence-region chr16 1 88827254

chr16   samtools SNV 49291141 49291141 .  +   . ID=ID_1;Variant_seq=A,G;Reference_seq=G;
chr16   samtools SNV 49291360 49291360 .  +   . ID=ID_2;Variant_seq=G;Reference_seq=C;
chr16   samtools SNV 49302125 49302125 .  +   . ID=ID_3;Variant_seq=T,C;Reference_seq=C;
chr16   samtools SNV 49302365 49302365 .  +   . ID=ID_4;Variant_seq=G,C;Reference_seq=C;
chr16   samtools SNV 49302700 49302700 .  +   . ID=ID_5;Variant_seq=T;Reference_seq=C;
chr16   samtools SNV 49303084 49303084 .  +   . ID=ID_6;Variant_seq=G,T;Reference_seq=T;
chr16   samtools SNV 49303156 49303156 .  +   . ID=ID_7;Variant_seq=T,C;Reference_seq=C;
chr16   samtools SNV 49303427 49303427 .  +   . ID=ID_8;Variant_seq=T,C;Reference_seq=C;
chr16   samtools SNV 49303596 49303596 .  +   . ID=ID_9;Variant_seq=T,C;Reference_seq=C;
```

Different datasets containing this data can be downloaded from ENSEMBL(ftp) and NCBI(ftp). According to the README in the ENSEMBL ftp, we should choose the file "*homo_sapiens_incl_consequences.gvf.gz*".

Biomart

Biomart provides a user interface to query and retrieve tabular files with the data of their databases.

In order to retrieve the required data we can:

1) Choose the database of interest (**FROM**),

2) Restrict our query choosing among the different filtering criteria provided (regions, gene ontology terms, etc…) (**WHERE**)

3) Restrict the fields of the result choosing among the different fields (ENSEMBL Ids, HGCN Ids, etc….) (**SELECT**)

```
Gene Start (bp) Gene End (bp) Ensembl Gene ID
```

```
16573334 16678949 ENSG00000037637
78028101 78149104 ENSG00000036549
29814705 29823405 ENSG00000225011
30117392 30117525 ENSG00000221126
30181698 30182394 ENSG00000228176
```

VCFTools

Description	Perl scripts that perform operation over vcf files
Troubleshooting	Requires installing other linux packages (tambix, gbzip) and perl modules (Test::Most), and setting several environment variables (PATH and PERL5LIB). Requires preprocessing of vcf files: compressing and index creation. Few documentation about internal details of commands (Ex, vcf-annotate has a –d option not documented, but required when annotating a vcf file)
Version installed	vcf_tools_0.1.11
Commands reviewed	vcf-to-tab, vcf-query, vcf-annotate
General opinion from SwEngineering perspective	Very interesting commands over vcf files, however, those operations (merge, filtering, intersection, annotation), should be performed over variations instead of its corresponding text representation.

***vcf-query*:** Converts VCF files into format defined by the user.

```
Usage: vcf-query [OPTIONS] file.vcf.gz
Options:
   -c, --columns <file|list> List of comma-separated column names or
one column name per line in a file.
   -f, --format <string> The default is '%CHROM:%POS\t%REF[\t%SAMPLE=%GT]\n'
   -l, --list-columns  List columns.
   -r, --region chr:from-to Retrieve the region. (Runs tabix.)
   --use-old-method Use old version of API, slower but more robust.
Expressions:
   %CHROM    The CHROM column (similarly also other columns)
   %GT    Translated genotype (e.g. C/A)
   %GTR   Raw genotype (e.g. 0/1)
   %INFO/TAG    Any tag in the INFO column
   %LINE     Prints the whole line
   %SAMPLE   Sample name
   []     The brackets loop over all samples
   %*<A><B>  All format fields printed as KEY<A>VALUE<B>
Examples:
   vcf-query file.vcf.gz 1:1000-2000 -c NA001,NA002,NA003
   vcf-query       file.vcf.gz       -r       1:1000-2000       -f
'%CHROM:%POS\t%REF\t%ALT[\t%SAMPLE:%*=,]\n'
   vcf-query file.vcf.gz -f '[%GT\t]%LINE\n'
   vcf-query file.vcf.gz -f '[%GT\ ]%LINE\n'
   vcf-query file.vcf.gz -f '%CHROM\_%POS\t%INFO/DP\t%FILTER\n'
```

vcf-to-tab Converts the VCF file into a tab-delimited text file listing the actual variants instead of ALT indexes

```
Usage: vcf-to-tab [OPTIONS] < in.vcf > out.tab
Options: -i, --iupac Use one-letter IUPAC codes
```

vcf-annotation Adds custom annotations to VCF files.

```
Usage: vcf-annotate [OPTIONS] > out.vcf
Options:
-a, annotation.gz
-d key=INFO,ID=ANN,Number=1,Type=Integer,Description='MyAnnotation'
-c CHROM,FROM,TO,INFO/ANN > out.vcf
```

Vcf-filter: Add filtering criteria to filter VCF Files: Cromosome, gene, rsId (one or a file), Allele frequency

```
Usage: vcftools –vcf file.vcf –-bed BEDfilewithVariations>
vcf.filtered
```

Annovar

Description	Perl scripts that annotate vcf files with structural data and database information.
Troubleshooting	Requires a lot of space in disk. 80M package, 12M RefGene and 8,5G dbsnp.
Version installed	Dowloaded latest version in October2013
Commands reviewed	annotate_variation, convert2annovar
General opinion from SwEngineering perspective	Commands only work with their custom format, so users should manage themselves the transformation of their files to this format (even for the standard format vcf). It requires downloading all databases locally, which for some users could be problematic. Still, very interesting functionality about annotations, structural data and database data and its separation in different files. However, I find textual annotations in general problematic because textual data is tedious to read, misses the well-formed structured rules to control errors, and leads to redundancy-which could affect to access latency and space.

Convert2annovar: Converts VCF files into format used by annovar (.avinput)

```
Command: convert2annovar.pl -format vcf4 file.vcf > file.avinput
```

Annotate_variation:

- **To download a database**: Concretely refSeq, dbsnp and 1000G project (allele frequency data)

```
Command: annotate_variation.pl -downdb -buildver hg19 -webfrom annovar refGene
humandb

Command:annotate_variation.pl -downdb -buildver hg19 -webfrom annovar dbsnp135
humandb

Command:annotate_variation.pl -downdb 1000g2012apr humandb –buildver hg19
```

- **To search variation properties:** Gene, region.

```
Command: annotate_variation.pl -buildver hg19 –geneanno file.avinput humandb/
```

This command creates the file `*.variation_function`, which tells whether the variant hit a structural region and the name of the gene (or neighbouring genes).

- **To search variation in databases:** Concretely in dbsnp retrieving the rsId

```
Command: annotate_variation.pl -buildver hg19 –filter –dbtype dbsnp135
file.avinput humandb/
```

This command creates two files: `*.filtered` that contains the SNPs not in dbSNP and `*.dropped` that contains variants that are annotated in dbSNP together with their rs identifiers

- **To search variation properties: S**tructural effect and hgvs notation.

```
Command: annotate_variation.pl -buildver hg19 -hgvs file.avinput humandb/
```

This command creates the file `*.exonic_variant_function` that contains the amino acid changes as a result of the exonic variant(several hgvsNotation with its refSeqIdentifiers)

VEP

Description	Perl Scripts that annotate variation files (VCF, TXT and BED) with the effects of a variation.
Troubleshooting	Requires to download 5,5 G. Refseq didn't download correctly and I had to solve the problem manually with linux commands (download and extract in the suitable directory: home/.vep/human).
Version installed	Dowloaded latest version in October2013
Commands reviewed	Variant_effect_predictor, filter vep
General opinion from SwEngineering perspective	Very interesting functionality about effects. However, the textual format has the same problem mentioned in annovar and Snpeff, the difference is that VEP expresses in the field "extra" a set of properties expressed using a pair "Key=value" and separated with ";". Still, to improve visualization the analysis creates a report containing statistics a visual diagrams.

Variant_effect_predictor: Annotates a variation file (VCF, Pileup, HGVSIdentifiers) with structural data: Gene (HGNC identifier, symbol), CDS positions, Intron/Exon number, and database data: rsId from dbSNP.

```
Command: perl variant_effect_predictor --cache -i file.vcf --symbol -o file.o.vcf
```

Variant_effect_predictor: Annotates a variation file (VCF, Pileup, HGVSIdentifiers) with aminoacid and codon change and allele frequency.

```
Command: perl variant_effect_predictor --cache -i file.vcf -o file.o.vcf
```

It can also annotate variations with the effect of SIFT and POLYPHEN, which analyse if the variant changes the protein function; it calculates the HGVS notation (genomic, coding and sometimes protein).

```
Command: perl variant_effect_predictor --cache -i file.vcf --hgvs --refseq --sift b --polyphen b -o file.o.vcf
```

Filter_vep: Filters the output file of VEP using a textual expression

```
Command: perl filter_vep --cache -i file.vcf --hgvs --sift b --polyphen b -o file.o.vcf
```

SamTools

SNPEff

Description	Java program that annotate variation files (VCF, TXT and BED) with the effects of a variation.
Troubleshooting	Requires to download 1,7G. And have a disk Space of 10G.
Version installed	Dowloaded latest version in October2013 (v3.3)
Commands reviewed	SnpSif, snpEff
General opinion from SwEngineering perspective	It requires downloading all databases locally, which for some users could be problematic. Very interesting functionality about annotations of effects and database data. However, the textual format has the same problem mentioned in annovar. Still, to improve visualization the analysis creates a report containing statistics a visual diagrams. SnpEff provides a very powerful option to filter variations, as it support expressions involving a big set of operations and any field of a VCF file. However, this operation should not be performed at the textual level.

SnpSif: Annotates the ID field of a variation file (VCF, TXT or BED). This functionality can be used to annotate the rs of dbSNP. The dbsnp.vcf can be downloaded from ncbi.

```
Command: java -jar SnpSif.jar annotate -v dbSnp.vcf file.vcf > file.dbSnp.vcf
```

SnpEff:

- **To download the database**: From the reference Genome.

```
Command: java -jar SnpEff.jar download -v GRCh37.69
Command: java -jar SnpEff.jar download -v hg19
```

- **To calculate structural data: G**ene, reference ids. #TODO test this command

```
Command: java -Xmx4g -jar SnpEff.jar -v GRCh37.69 file.dbSnp.vcf> file.str.vcf
```

- **To calculate variation effects:** List of effects, their impact, codon/aminoacid changes, gene, reference ids and the loss of function. Also the hgvs notation.

```
Command: java -Xmx4g -jar SnpEff.jar eff -hgvs -v GRCh37.69 file.dbSnp.vcf>
file.eff.vcf
```

*More options can be used in the analysis: Apply filters, choose regions of application, provide a list of transcripts, use custom annotations, etc.

- **To download databases with variation effects:** The database dbNSFP contains information about SIFT and POLPHEN as well as MAFs from 1000G among others.

```
Command:wget http://dbnsfp.houstonbioinformatics.org/dbNSFPzip/dbNSFPv2.3.zip
Command: unzip dbNSFP2.3.zip
```

And annotate variations with data from this database

```
Command: java -jar SnpSift.jar dbnsfp -v dbNSFP2.3.txt –f SIFT_score,
Polyphen2_HVAR_pred file.annotated.vcf > file.annotated2.vcf
```

Pipeline Development Environments

Biopython, BioPerl, Biojava, Bio*

Bio* is a family of libraries for the development of personalized genetic analysis tools on top of well-established programming languages. BioPerl (STAJICH2002) in Perl, BioPython (COCK2009) in Python, or BioJava (HOLLAND2008) in Java.

Their common aim is to provide common functionality regarding DNA sequences manipulation and functionality regarding integration among software components. These libraries provide a set of modules, classes and methods that implement algorithms, data structures, advanced string manipulation operations and so on. Additionally, as the domain still lacks of a standard nomenclature to express the output results, they provide several format conversion operations to transform these results among different tools.

For example, BioPerl provides support for: a) Indexation, transformation and annotation of sequences; b) Sequence alignment; c) Sequence search; d) Format transformations; e) Pattern matching algorithms for sequence analysis; f) Wrappers for database retrieval or online services execution; and g) 3D representation of proteins.

```
use Bio::DB::EMBL;
use Bio::SeqIO;
my $db= new Bio::DB::EMBL();
my $seq=$db->get_Seq_by_acc("U14680);
my $seqout=new Bio::SeqIO(-format=> "genbank");
if(defined $seq){
    $seqout->write_seq($seq);
}
```

Figure 1 BioPerl example

Figure 6 shows an example written with BioPerl to access to the EMBL database in order to retrieve a sequence whose identifier is “U14680”. After retrieval, this sequence is transformed to the Genebank format.

Taverna

Taverna provides an environment to design, edit and execute workflows using graph representations: nodes that represent tasks, and arrows that represent links for communications among tasks. Geneticists can choose from a list of components (that provide genetic functionality), the tasks they want to accomplish, drag and drop them on a graphical worksheet and eventually they compose a workflow that executes the desired genetic analysis. Taverna integrates functionality through myExperiment (Deroure et al. 2008), a social network to share scientific workflows, and the Biocatalogue (Bhagat et al.), a curated catalogue of web services for the life sciences.

Taverna offers a user interface (Figure 2) to facilitate the creation of workflows. The right side, offers the user a set of tabs to work with the environment; for example, a tab to discover services or to overview the details of a workflow. The left side of the interface shows the graphical representation of all the tasks and dataflow among tasks gathered in the workflow.

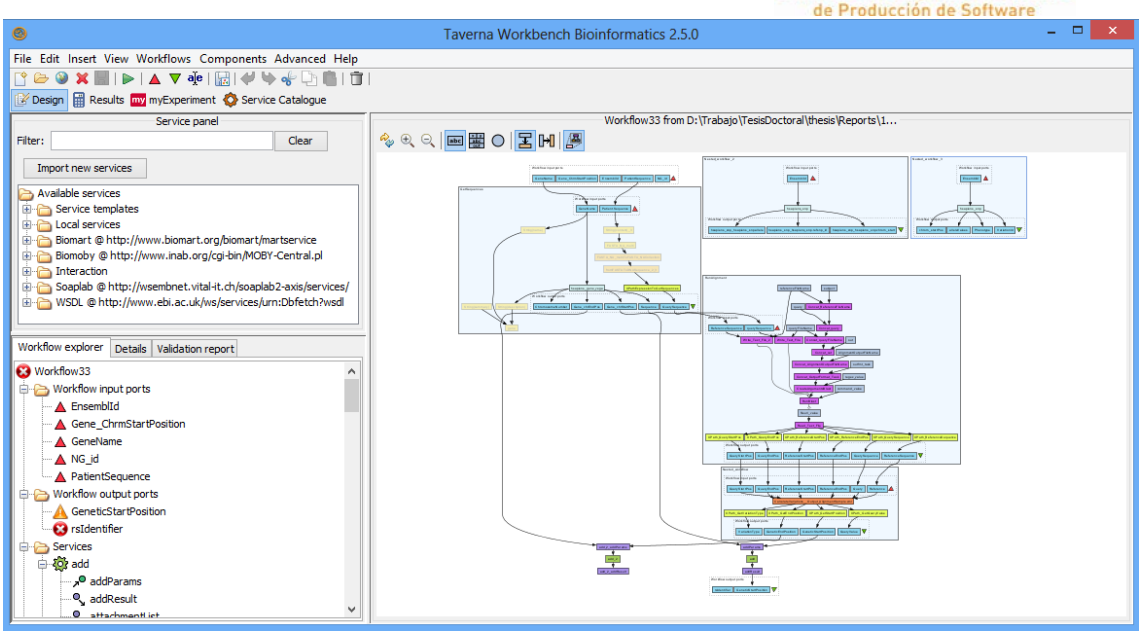


Figure 2 Workflow Example on Taverna

Benefits

After a workflow execution, it provides intermediary results from all services that allow the traceability of the results.

It provides a high abstraction to integrate command line tools, rest services, biomart services, and others. This abstraction makes easier the integration, but technological details are still required.

Disadvantages:

The Taverna interface has a tedious workflow representation: Both simple and complex examples contain overloading elements (a big amount of services).

It is not possible to represent the objects of a data model.

It does not provide clear description of functionality and parameters from the integrated services.

General Issues:

- Service discovery: Services not ordered semantically.
    - Same family services are scattered: The available taxonomy shows the technology and the provider in the top. As a second level some semantics are used, however, they contain similar and overlapping categories such as {conversion, converting} and {Alignment, Bioinformatics}. In the third level, semantics is lost again, and services are ordered by package name.
    - Filter not enough helpful: Service search is facilitated by a key word filter. However, although service list is reduced discovery is still difficult because of the catalog taxonomy previously described.

- Data management: Using non-formatted fields or a provided ontology (Biomoby). The use of this ontology presents some issues:
    - Non-editable data model: New data entities cannot be defined and existent ones cannot be modified for the use in a workflow.
    - Biological, bioinformatic and informatics concepts mixed: Example: {Allele, gene}, {BlastJob, AlignedSequences} {Array, base64_encoded}. All of them are in the same level/category.
    - Only the relationship "is-a" is represented: Other relationships cannot be distinguished until the object is used in the workflow.
    - Object content inaccessible: Once an object is created as a result of a service their properties can only be accessed using xpath expressions over its xml representation.
    - **How to manage high level data**: Conceptual model representation will be used in the DSL. In order to instantiate them in Taverna there are several options: 1) Try to map some entities whose meaning is the same that a Biomoby object (ontology mapping); or 2) Create a biomoby object "textXML" or a "readTextFile" service that reads the xml object representation (less intuitive in Taverna). In both cases information from an object is retrieved using XPath expressions.
- Graphical notation: Issues when creating and testing workflows
    - Workflow design performance is compromised when workflow starts to grow.
    - Graphical representation of objects is user overloading: each property is represented with a box and situated in a different level from bottom to up. Example: A class with 3 properties and a relationship with another class with 5 properties will be represented as 10 boxes distributed in 3 levels.
    - Graphical representation of objects is confusing: Although different colors are used for different type of services, data objects are also represented and treated (executed) as services.
- Database access: Low level configuration jdbc
    - A service is available to query a database, however, it is required to download the driver into the suitable taverna workspace, it should be linked to the corresponding service, and the syntax of the service must be known.

Report

Taverna has been used to create the workflows previously defined.

Each workflow has been developed using the available services as much as possible (trying to avoid the integration of local scripts, programs or tools) and using the most intuitive services available (taking into account service description and input/output data descriptions). The main purpose of this procedure is to identify the most agile way that taverna provides to geneticists in order to develop their workflows.

The main objective of this analysis is to detect those issues that could be improved to be more usable for geneticists and also to identify what actions cannot be performed by them as a consequence of lacking of programming knowledge.

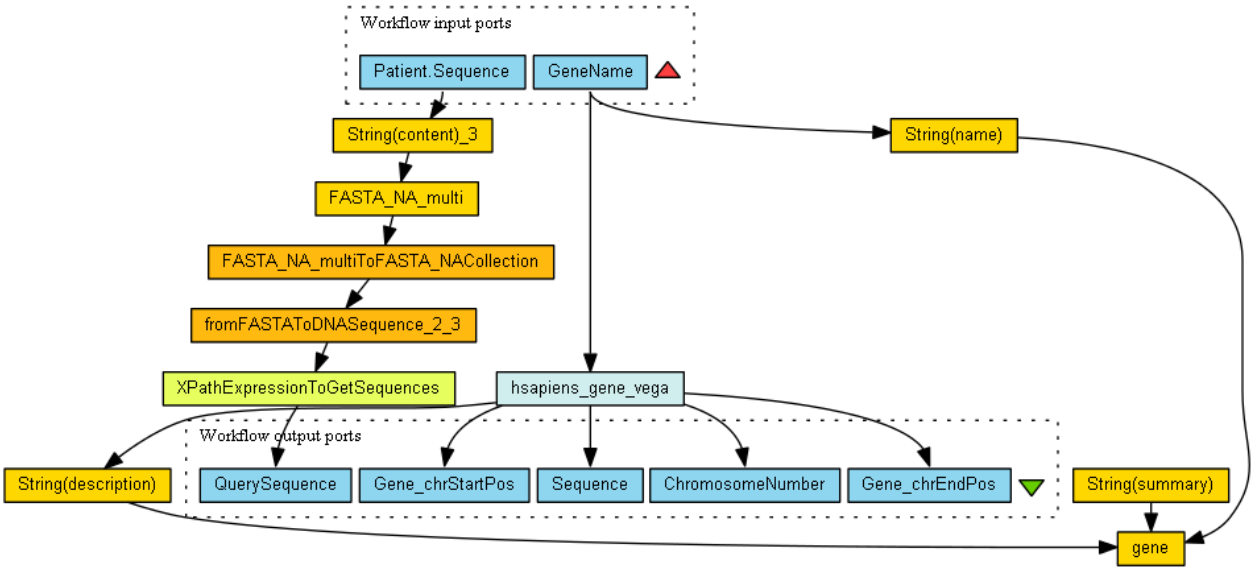
Description: Taverna is a framework for the design, edition and execution of workflows based on the integration of web services. Taverna is specially focused on the biological domain, providing the interoperability with biological resources such as the BioCatalogue or myExperiment.

PROS: It has an intuitive graphical user interface that allows drag and drop elements of the workflow design. The integration with the BioCatalogue and myExperiment allow geneticists to search biological services or other geneticists' workflows.

CONS: Services provided by the BioCatalogue are not completely described: it offers the number, the name and the type of parameters but it does not provide the explanation of the parameters (concept) either the format or the allowed values.

*Workflow 1: Diagen (*Disease diagnosis BREAST Cancer from variation detection)

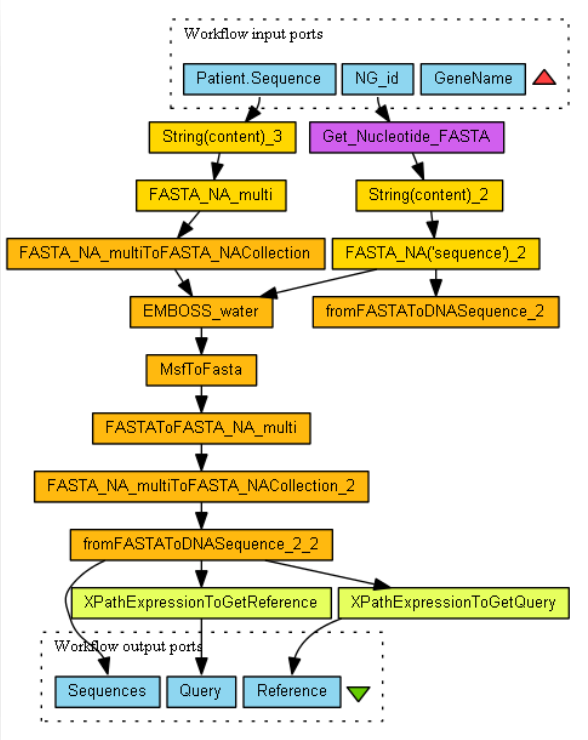
1) Describe Gene, retrieve Patient.Sequence, retrieve Gene{BRCA1}.sequence from NCBI. The three tasks can be performed in Taverna.
   Problems detected:
   - **The required data entity is not available:** In order to describe a Gene a MOBY object Gene{name,summary,description} has been used. However, several properties cannot be expressed (chromosome number, chromosome start position and chromosome end position). Moreover, it is not possible to create a personalized entity.
   - **A simple task must be performed using different services**: Five services (plus one XPath service) are required in order to retrieve a Patiente.Sequence expressed in fasta/multifasta format.
   - **NCBI data retrieval is up to the user**: Entrez utilities (8 server-side programs to access to NCBI) are provided, but user has to learn how those utilities and then construct the query. As a solution, genetic data can be easily retrieved using a Biomart service, which provides a user interface when the user is guided to create a query against a concrete data source. This user interface provides the available fields to filter the data source and the entities that can be retrieved and their attributes. Using the biomart service, NCBI is not available, but we used another database called VEGA (Sanger UK), where all the required properties for this workflow were available.
   - **Database identifiers must be taken into account**: The VEGA database accessed allows the retrieval of the sequence using the HGCN identifier provided. However, other databases require other external identifiers as Entrez gene, RefSeq, GO, UniProt, etc.

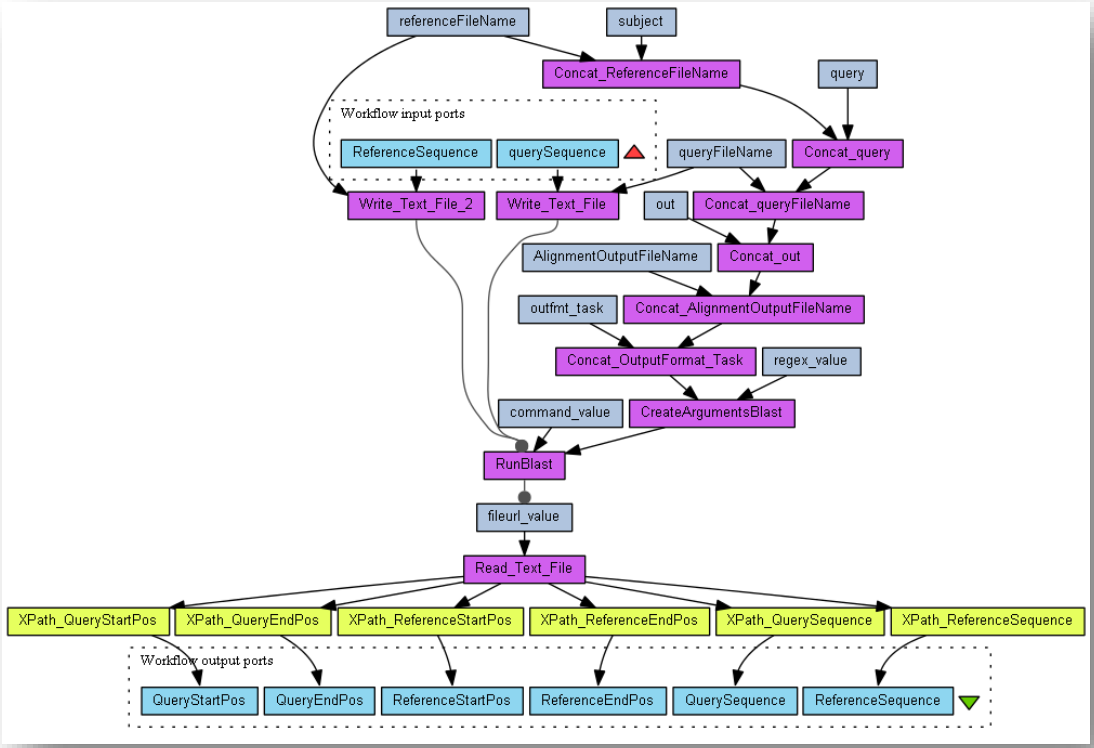


2) Single Alignment. This task can be performed in Taverna, but in order to obtain the list of Differences a proprietary algorithm has to be implemented.
   Problems detected:
   - **Tedious Alignment service use:** Several ways to use a service
     - Search by keyword "alignment": A huge set of different services appear spread around several categories. It is provided service name and sometimes a brief description in natural language. User should try one by one until finding the suitable service. We try different ones until the algorithm "EmbossWatter" worked, however it does not support long sequences. The problem with this service is that it does not provide the position of the sequence where the query aligns.
     - Create a "command line app" and invoke a local application using a command. In this case, the user has to manage the corresponding paths and permissions to execute the command, the construction of the command by concatenating different Strings, and the retrieval of results.

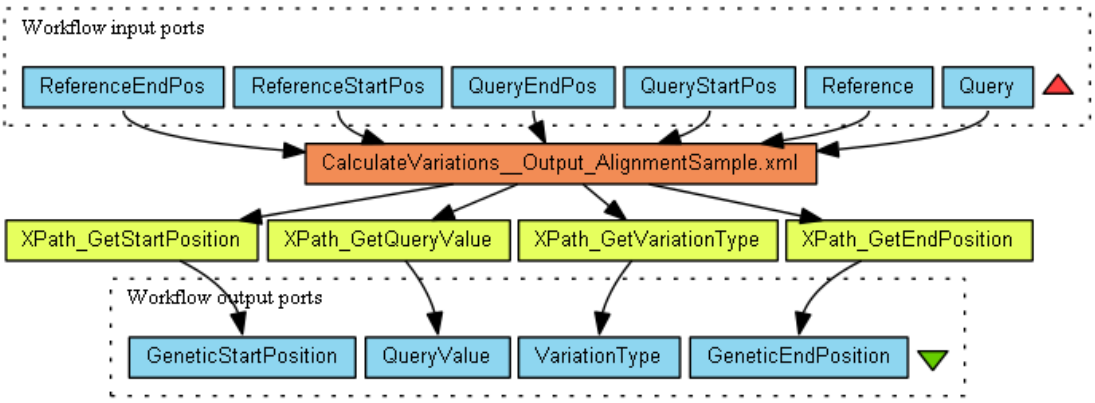
   - **Data types of service inputs/outputs should be taking into account**: As each service uses a different input/output data entity user has two possible solutions:
     - Manage format conversion by concatenating convertor services if available.
     - Manage format conversion extracting data using xPath and use it as input to the attribute of the object required.
       - If inputs are Moby objects, Taverna provides an utility to search compatible services. However, the retrieved list only provides the name of the service and

user has to try, like explained before, until one of them matches their preferences. We searched a service that consumes a DNASequence MOBY Object and we tried four services and they did not work (runNCBIBlastn, runBlat, runNCBIBlastn2seq, runWUBlastn2seq)


Workflow input ports
Patient.Sequence
NG_id
GeneName
String(content)_3
Get_Nucleotide_FASTA
FASTA_NA_multi
String(content)_2
FASTA_NA_multiToFASTA_NACollection
FASTA_NA('sequence')_2
EMBOSS_water
fromFASTAToDNASequence_2
MsfToFasta
FASTAToFASTA_NA_multi
FASTA_NA_multiToFASTA_NACollection_2
fromFASTAToDNASequence_2_2
XPathExpressionToGetReference
XPathExpressionToGetQuery
Workflow output ports
Sequences
Query
Reference

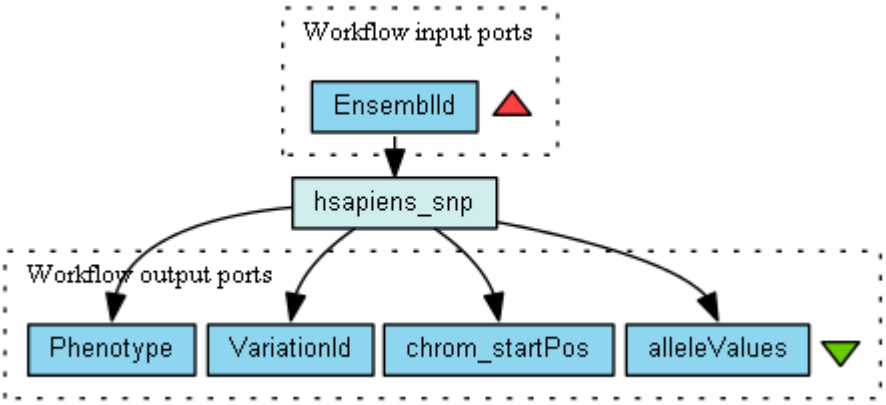


3) Obtain List Differences. This task can be performed in Taverna, but a new algorithm must be defined and integrated.
   Problems detected:



- **The required data entities are not available**: Alignment cannot be expressed (the most similar MOBY object is “Sequence Alignment Report”, which contains an attribute with the content in raw text). Equally, variations cannot be expressed (the most similar MOBY object is “so_sequence_alteration”, which is decomposed in, “insertion”, “deletion”, “indel” and “inversion”. However they cannot be filled by the user and there is not any compatible service to do it.

4) Retrieve List SNPs of GeneBRCA1 from dbSNP, Retrieve List mutations of Gene from HGMD, LOVD, BIC.
   All this tasks can be performed in Taverna, but information is retrieved from "ENSEMBL VARIATION 67 (SANGER UK) "
   Problems detected
   - **Non-consistent data**: Fields retrieved from Biomart services have non-restricted content. Some fields are not fulfilled, others do not correspond with the field meaning and others are expressed in natural language. In our case, it is not possible to retrieve the type of variation: insertion, deletion, indel.
   - **The required entities for saving data retrieved are not available**: Fields retrieved from any data source should be saved in one html/csv/xls file or linked to a port.



5) Compare List differences vs List SNPs (To be done)

Galaxy

Galaxy (Giardine et al. 2005) is an open-source web-based environment for the execution of biological services. Its main purpose is to help geneticists with their data intensive biological research through the definition of web interfaces for biological data retrieval and services execution. With this purpose, it provides different interfaces that access to some popular genetic databases and toolkits.

Galaxy offers a user interface (Figure 3) to facilitate the creation of workflows made of three panels. The right panel, shows a list of tools that can be executed and used for the workflow. The central panel shows the graphical representation of all the tasks and dataflow among tasks gathered in the workflow. And finally, the left panel shows the details of the tool selected in the central panel so that it can be configured by the user.

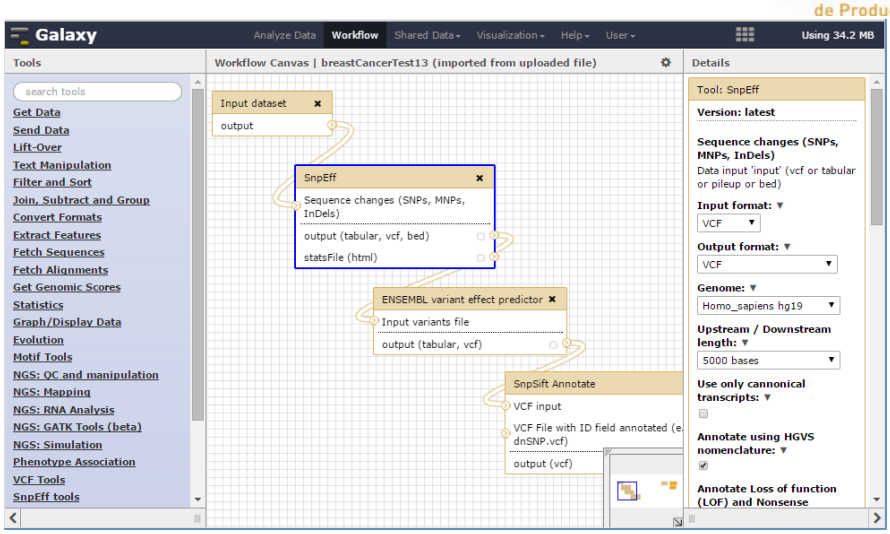


Figure 3 Workflow Example on Galaxy

Report

Summary: Galaxy is an environment that provides the user the possibility to run different biological services and create workflows combining those services. The environment can be executed locally, using a web interface or in the cloud. Galaxy allows the users to retrieve local and USCS database' data sets to be used in their experiments. It allows to combine data from independent queries, to perform calculations over these data sets (such as filtering a data set, combining several data sets and transforming data using a biological service) and, finally, to visualize the results.

PROS: The most common biological services used by geneticists are integrated in Galaxy (such as database retrieval, biological algorithms and visual display units). The users are provided with a user interface for each service, concretely a form with all the parameters to be filled. Galaxy allows the user to record all the steps or services that are executed, and afterwards, those selected by the user can be composed in a workflow. All these tasks are available to execute as many times as required.

CONS: Galaxy operates using low level data descriptions: its functionality is based in the use of raw data sets saved in files. All services receive data from a file and, as a result, they obtain another file. Data is organized in rows and columns where each field implies a specific meaning. As a consequence, users configure the service's parameters taking into account the rows and the columns instead of the underlying concepts.

Additionally, when composing a workflow, the user has to worry about data flow. The services' interfaces are configured to accept files expressed in a specific format, so it can differ within different services. If the format is different a transformation is required. Hence, the user has to find the way to provide inside the Galaxy environment a mechanism that executes this transformation. The authors claim that new functionality can be added, but deep knowledge of Galaxy and some programming skills are required.

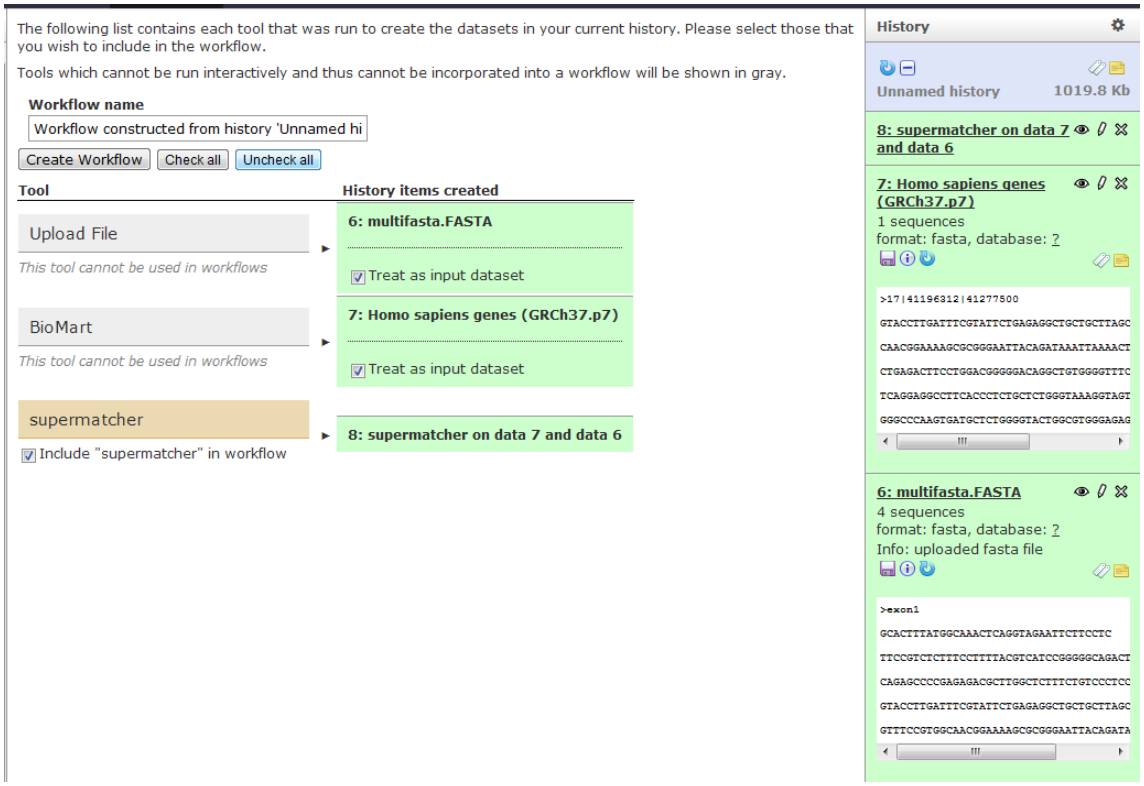
*Workflow 1: Diagen (*Disease diagnosis BREAST Cancer from variation detection)

1) Describe Gene, retrieve Patient.Sequence, retrieve Gene{BRCA1}.sequence from NCBI.

   The three tasks cannot be performed in Taverna.

Problems detected

- **Entities cannot be described:** Data is managed using a "dataset entity" with some metadata associated (such a name, format, etc). Gene and Patient are represented as a dataset that contains one or several sequences respectively.
- **NCBI data retrieval cannot be performed**: USCS Genome browser and ensemble database can be queried. Additionally, genetic data can be easily retrieved using a Biomart service, which provides a user interface when the user is guided to create a query against a concrete data source. This user interface provides the available fields to filter the data source and the entities that can be retrieved and their attributes. Using the biomart service, NCBI is not available, but we used another database called VEGA (Sanger UK), where all the required properties for this workflow were available.
-



Description: Proteus is a problem solving environment based on the Grid technology for composing, compiling and running bioinformatic applications. Proteus is based on the use of ontologies to aid the users to define their applications. The bioinformatic ontology used represents the bioinformatics domain by means of the definition of biological data sources, software components and bioinformatic processes or tasks. The user designs their personalized tool browsing and querying the ontology to find for the components that will be used. Additionally, each component is provided with metadata to allow the user to configure it according their requirements.

PROS: The ontology is suitable to find software components that fit biologists' requirements. Additionally, the browser is easy to use because the ontology is categorized in different taxonomies and displayed using showing labeled relationships with other items of the ontology.

CONS: The environment does not explain the interaction between a domain ontology and the bioinformatics ontology, either how the workflow management system transforms the graphical design into grid scripts. As a consequence, it is not clear how data flow among tasks is designed or executed.

This work was developed in 2004-2005, and the environment has not been made available. Their authors focused their efforts in applying these ideas in mass-spectrometry proteomics and created a specific platform (MS-Analyzer) that integrates algorithms and tools to design tools regarding this domain.

EBioFlow

eBioFlow (WASSINK2010) is an open-source workflow management system to design and execute biological workflows developed in the academic environment as a proof of concept of a series of Phd dissertations. Its main purpose is to improve other workflow development environments by providing a better usability for workflow design (multiple perspectives to model data and control flow), a better workflow enactment with support for late binding of services, and finally, the support of data provenance for improving workflow sharing and reuse.

Figure 14 shows the interface of eBioflow. On the left panel, a set of tabs are available to navigate between workflows, manage data provenance, search for services, and see previous workflow runs. On the right panel, several tabs provide the different perspectives available to manage all the different concerns while designing, executing and sharing workflows.

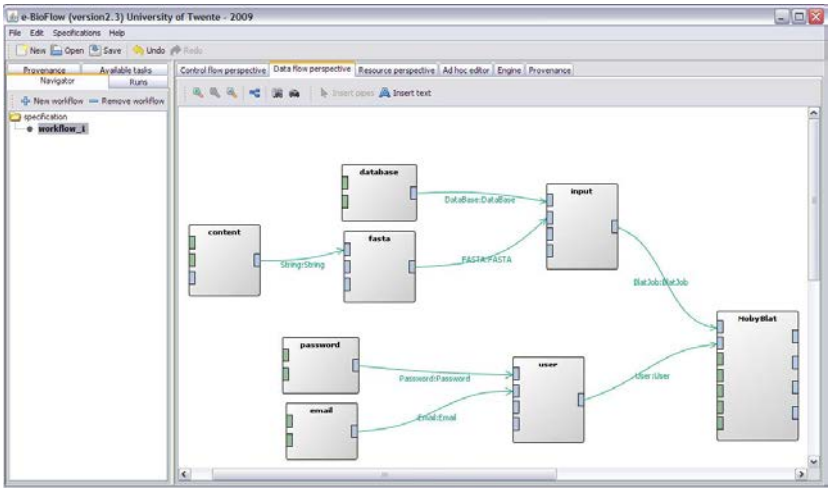


Figure 4 Workflow Example on eBioFlow

Report

Description: BiosFlow is a workflow platform that allows the user to discover web services using ontology constructs. It defines the entity "node" as a basic unit with a set of features that represents a service. This platform provides a user environment to discover, compose and execute predefined service nodes integrated in the platform.

PROS: This work explains the need to simplify the creation of applications by biologists. With this purpose, they add semantics to workflow composition, using ontologies, and develop an easy user interface, based on icons, to drag and drop services and allow their subsequent execution.

CONS: There is only a paper that describes this work (2009) and the platform is not available. They claim that ontologies are used to search for web services, but it does not explain how services are related with ontologies to allow the service discovery, neither which ontologies they use, and neither an example of use.

BSIS

Report

Description: BioService Integration System (BSIS) is a framework for the development of biological workflows. BSIS supports semantic discovery and composition of services because provides a mechanism to annotate webservices with ontologies. Concretely the ontologies used are: 1) Service ontologies, which describe processes and transformations implemented by bioinformatic services; and 2) Data ontologies, which describe biological data.

BSIS provides a graphical workflow language (formally described) that allows the user to create concrete or abstract workflows that execute web services. The concrete workflows are specified by the users while the abstract ones are instantiated by the framework attending semantic constrains added by the user.

PROS: The framework provides a graphical user interface easy to use for the user. Moreover, this proposal takes into account the use of semantics for the definition of bioinformatic workflows. As a consequence, the user is provided with a high level of abstraction of implementation details. Concretely, the user can browse the biological ontology (service or data) and use it directly in the workflow design environment by drag and drop.

CONS: The phd was developed in 2007 and its main publication was on 2010. The contact information is not valid and the framework is not available. As it is not available and the project seems to be over, recent biological ontologies coverage cannot be ensured. The "service" basic unit is only applicable to web service.