GOAT

Genetic Output Analysis Tool

Version : 1.0
Release date : May 3rd 2015
Revision date : Septembber 30th 2015

VISION DOCUMENT

Authors (software developers):

Dr. Alain APRIL – ÉTS Montréal, Canada
Beatriz Kanzki – ÉTS
David Lauzon – ÉTS
Cédric Urvoy – ÉTS
Supervisor :
Dr. Alain APRIL – ÉTS

Collaborators (users expressing their need)
Presented to : Pavel Hamet – CRCHUM
Michael Phillips - CRCHUM

approved on ___/___/___

approved on ___/___/___

Revision History

Date	Version	Description	Author
2015-05-02	0.1	First version	Beatriz Kanzki
2015-05-02	0.2	Review of version 0.1	Dr. Alain April
2015-05-02	0.3	Integration of David's comments	Beatriz Kanzki
2015-05-03	0.4	Integration of Alain's comments	Beatriz Kanzki
2015-05-03	0.5	Review of applicable sections and general document structure	David Lauzon
2015-05-05	0.6	Add responsibilities of users and put products available at competition	Beatriz Kanzki
2015-05-06	0.7	Added product overview diagram. Modified constraints and quality criteria (moved from B08). Moved B07 to FEA05. Moved MetaboAnalyst screenshots to Appendix B. Changed license.	David Lauzon
2015-05-06	0.8	Added Dre Tremblay's expectations and responsibilities	Beatriz Kanzki
2015-05-07	0.9	Added description and example of product expected behavior, workflow and user requirement for tool Appendix B,C, D	Beatriz Kanzki
2015-05-10	0.10	Updated product perspective, expanded user needs to 28 features.	David Lauzon
2015-05-14	0.11	Clarified sections 3.3, 4.1, 4.2, 5 + removed appendix B - D	David Lauzon
2015-05-15	0.12	Added feature definitions and URL's, and figure 1 of for scope of the project	Beatriz kanzki

2015-05-15	0.13	Reviewed format, licensing text	Alain April
2015-05-15	0.14	Added process diagrams	Beatriz Kanzki
2015-05-16	0.15	Reviewed english, scope text, requirements formulation, table number sequences and open comments	Alain April
2015-05-20	0.16	Reviewed Scope Figure, inserted potential UI examples, reviewed Gene Query process figure and simplify product features, updated feature attributes with state and effort values, assigned an open source creative commons license	Alain April Beatriz Kanzki David Lauzon
2015-05-25	0.16.1	Changed the scope figure and other minor display issues (table 7 and added appendix B)	David Lauzon
2015-05-27	0.16.2	Added SigmaPlot in competition	Beatriz Kanzki
2015-05-28	0.16.3	Workflow process image changed	Beatriz Kanzki
2015-09-20	1.0	final review	Alain April

Summary

Table list

Figures list

1. Introduction

1.1 Objective

The object of this document is to determine the project for the visualization and analysis tool for genomic datasets (coined GOAT) in order to have a common vision between the development team and the various project stakeholders. The figure 1 describes the whole research workflow. Notice that the scope of this document, GOAT, is restricted to the green section of the workflow (in the middle of the figure.

1.2 Scope

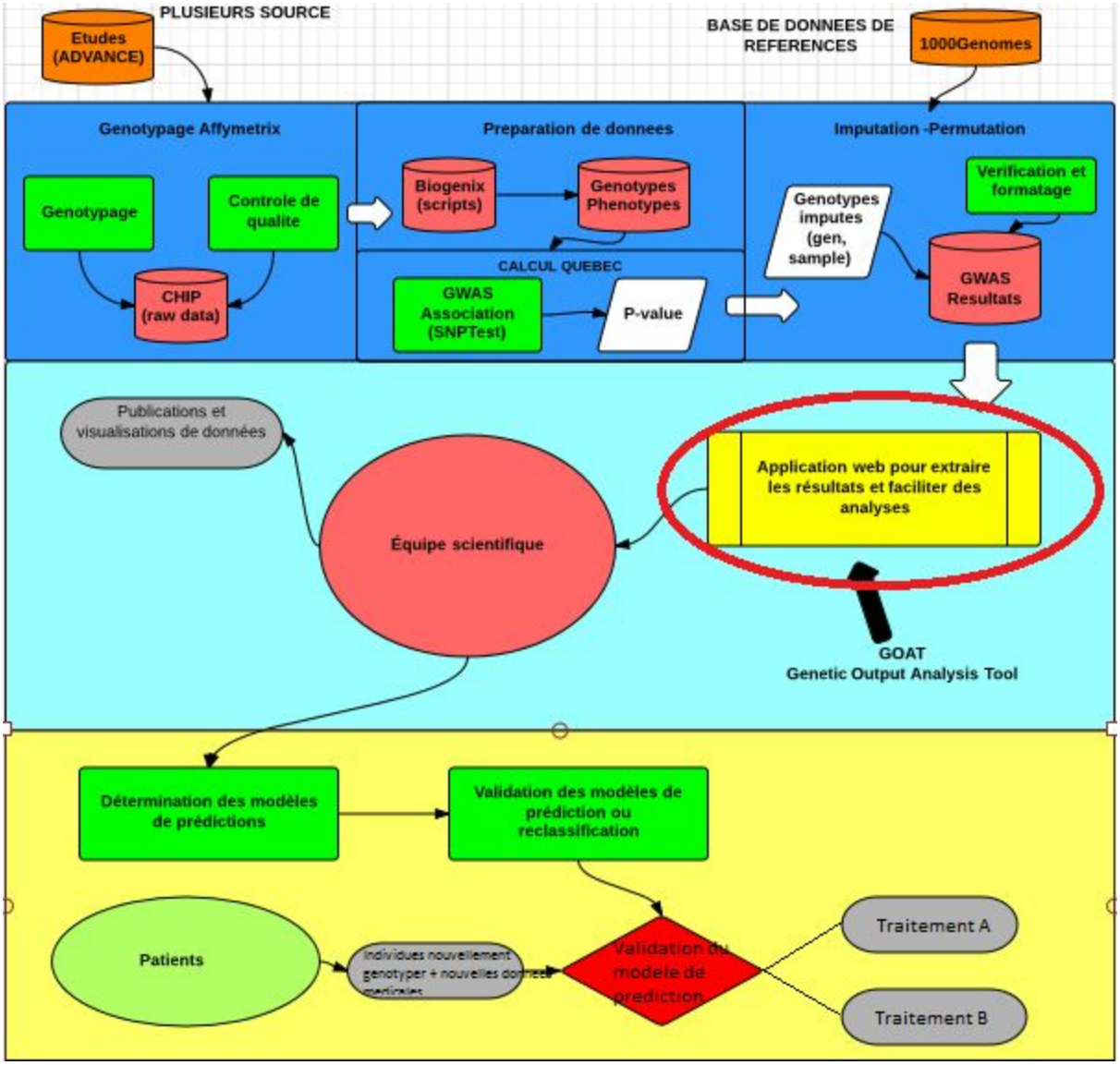


Figure 1 shows the whole process of the discovery workflow. **As indicated by the big black arrow and red circle, the scope of this Vision document is limited by the green zone**.

3 phases of software development are planned for GOAT:

1. Access and retrieval of data from the database
   - GWAS results
   - Intersection between GWAS and other resources.
   - Downloadable tables.
2. Visualization, analysis, phenotype queries.
   - Graph (interactive Manhattan)
   - BoxPlot (ratio case/control)
   - Gene Annotation
3. Post-processing analysis
   - Reanalyze processed data live
   - Create new preformatted
   - Boxplot of quantitative data compared to qualitative data
   - Statistics on the phenotype (comparison table)

Note that GOAT will be an independent open source software and can run on any genotyped dataset. The project does not require and is not dependent on the CRCHUM database content or its internal structure

1.3 Definitions, acronyms and abbreviations

Terms	Definition
ETL	Extract Transform Load. Process used Extract data from a source, Transform it, and Load into a destination database.
GWAS	Genome wide association study is an examination of many common genetic variants in different individuals to see if any variant is associated with a trait.
PI	Principal investigator
CAU	Caucasian population
AS	Asian population
AA	African American population
rs ID	Accession number for a SNP
SNP	Single nucleotide polymorphism is a DNA variation sequence
PubMed	Free full-text archive of biomedical and life sciences journal literature at the NIH
NIH	National Institutes of Health
1000 Genome	Catalog of human genetic variations
Genecards	Search, integrated database of human genes that gives concise genomic information, on all known predicted human genes.
ClinVar	Clinical variations is a freely accessible public archive of reports of the relationships among human variations and phenotypes hosted by NCBI
NCBI	National center for biotechnology information gives access to biomedical and genomic information.
SNP-TEST	Program for analysis of SNP association in genome-wide studies.
CFR21 part 11	Code of Federal Regulations Title 21: electronic records - electronic signatures

Table 1: Definitions, acronyms and abbreviations

1.4 Summary description of product

The final product is a web site based tool allowing the researcher to query its own database in order to extract info on phenotypes, genetic datasets, or biomarkers, visualize them in order to get a file that could be used for further statistical analysis, without having to go through the bioinformaticians office. This application would allow users to locate the information, interactively and visualize it, and get a table of significant results where each gene would be a hyperlink towards external sites for further information.

1.5 References

http://www.metaboanalyst.ca/faces/ModuleView.xhtml

Glossary : to come in a separate document

2. Positioning

2.1 Business opportunity

This software could be in demand by health researchers.

2.2 Problem definition

Problem	- Health researchers rely on bioinformatics specialists for data extraction from the database and for their visualization. - The current softwares are not efficient, or user friendly for their needs.
Affects	● Bioinformatics specialists job who cannot cope with all demands coming from stakeholders; ● And stakeholders turnaround time to analyze data
Impact	Generates a waiting list and a dependency on the bioinformatics specialist availability.
A good solution	A good solution would allow a researcher to locate the information that needs to be analyzed, from a database, conduct statistical analysis and visualize the resulting graphs, interactively (i.e. meaning a drill down facility) and allowing to obtain identification of gene using external best practice information interactively for further study. The researcher could save an ongoing study to continue the research another time.

Table 2: Problem Definition

2.3 Product positioning

For	Health research Labs that want to analyze genetic data
Who wants to	Improve information access from their internal databases, conduct statistical analysis and visualization seamlessly.
GOAT	is a web-based software as a service (SaaS) for bioinformatics genomic and genetic research.
Which allows	To display an interactive graph, varying depending on the information required, referring to significant genes that can be selected. To save the work session in order to access it later.
Unlike	-Having to call on a bioinformatics specialist to obtain the information. -Using existing open source tools that are currently available and do not offer the functionality required.
The product	Needs a professional and user-friendly interface that improves information retrieval efficiency and offers a modern visualization of genomic data. The product must also allow to save the user's session and work for furthering analysis.

Table 3: Product positioning

3. Description of stakeholders and users

3.1 Summary of stakeholders and users.

Name	Occupation
Dr Pavel Hamet	Presents his requirements : Principal investigator
Dre Tremblay	Presents her requirements
Michael Philips	User representative : Director
Francois Harvey	Bioinformatician Specialist
Francois Marois	Bioinformatician Specialist
Gilles Godefroid	Bioinformatician Specialist
Carole Long	Biologist
John Raelson	Geneticist
Mounsif Haloui	Biologist
Ramzan Tahir	Biostatistician
Paul Simon	Doctoral student, Observer
Alain April	Developers leader: Sftwr Project Supervisor (ÉTS)

Table 4: Stakeholders and users of GOAT

3.2 User and stakeholder profiles

3.2.1 Health Researcher Profile (Dr. Hamet)

Responsabilities	*Presents high level requirements and use tool to prepare conference and show data by doing analysis and extracting data. Add gene annotation on DNA region.*
Success Criteria	*Data extraction is successful and can download graphs* *Gene annotation recorded in database*

Deliverables	*Downloadable tables and graphs*
Comments / Issues	*Has to access database from inside the CHUM, Query in tool available now are slow, not user friendly and takes a lot of steps on different screens before getting the actual result.*

3.2.2 Health Researcher Profile (Dre. Tremblay)

Responsabilities	*Use tool to prepare conference and do search queries for genes, phenotype.* *Compare two phenotypes for a SNP, and then compare to GWAS catalog.*
Success Criteria	*Name of dataset where interesting results are is identified* *Whole SNPTest output is present* *Complete GWAS output with possibility to zoom* *Link to other interesting SNPs can be clicked easily after review of results in table.*
Deliverables	*Downloadable tables and graphs to use in powerpoint or reanalysis.*
Comments / Issues	*Has to access database from inside the CHUM, Has to access database from inside the CHUM, Query in tool available now are slow, not user friendly and takes a lot of steps on different screens before getting the actual result.*

3.2.3 Bioinformatician Profile (F. Harvey, F. Marois, G. Godefroid)

Responsabilities	*Enrich database with information that will be available for tool*
Success Criteria	*Data has been recorded*
Deliverables	*N/A*
Comments / Issues	*N/A*

3.2.4 ÉTS Stakeholder Profile

Responsabilities	● Development of the open source software ● Maintenance and long-term evolution of the software
Success Criteria	● Health researchers are more productive with the new product than with current solutions. ● New feature request can be implemented with minimal effort
Deliverables	● An open source software that can be implemented in the lab
Comments / Issues	● Frequent feedback from the CRCHUM's team, and the senior management board, is required to ensure success of the project.

3.2.5 Research Student

Responsabilities	*Extract data from database and analyse them* *Generate results, and make quality control* *Generate graphs*
Success Criteria	*Data extraction successful* *Analysis can be done by tool* *Graphs and tables are downloadable*
Deliverables	*Downloadable table and graphs*
Comments / Issues	*N/A*

3.2.6 Biostatician

Responsabilities	*Extract data from database and analyse them* *Quality control → (verify distribution of data)*
Success Criteria	*Data extraction successful* *Quality control of data successful with external tool* *Graphs and tables are downloadable*
Deliverables	*Downloadable table and graphs*
Comments / Issues	*N/A*

3.2.7 Geneticist

Responsabilities	*Extract data from database and analyse them* *Generate results, and make quality control* *Publish results and graph*
Success Criteria	*Data extraction successful* *External quality control tests passed by data* *Graphs and tables are downloadable*
Deliverables	*Downloadable table and graphs*
Comments / Issues	*N/A*

3.2.8 Biologists

Responsabilities	*Extract data from database and analyse them* *Generate results, and make quality control* *Generate graphs* *Locate publications according to genes*
Success Criteria	*Data extraction successful* *Analysis can be done by tool* *Graphs and tables are downloadable* *All available resources on internet appear on the tool.*
Deliverables	*Downloadable table and graphs*
Comments / Issues	*N/A*

3.3 Major needs/user requirements of all stakeholders and users

Needs	Concerns	Current Solution / Tool	Proposed Solution
B01 – Query by Gene	Researcher claim that some information is unavailable on the existing tool. When investigating with the bioinformatics specialists they show that it is available.	Information is difficult to find, non-existent, or lost with current software or it has display/query problems.	Modernize/review the query function UI for ease of use. The researcher should be able to filter the list of markers, from the database, using the most popular access keys just like popular open source catalogs.
B02 – Data visualization	- Today, resulting graphs have to be generated manually; - Users would like to delete or select significant data on visualization before saving session;	Current software only displays a table with results with links to external web sites.	Develop a new visualization function that can display interactive graphs: 1) of different types based on the type of data extracted; 2) that have data table(s) with integrated hyperlinks pointing to external websites and other sources (see B05).
B03 – Graph Export	- Users would like to use the graphs readily for publication.	Good tools exist to produce interactive graphs, but these are not readily acceptable for publications	The researcher will be able to export (i.e. download or copy/paste) a graph being visualized, as an image, in a publication ready format
B04 – Query recorder	Today, queries using the existing software cannot be saved and reused	This is not available today.	Develop a functionality to save current session work, using the user ID and description of the current analysis at any point. This session can be reopened and continued.
B05 – Contextual Gene Information	Obtain the up-to-date information for a gene from a reputed external source	Current tools available do not display all the updated publications available online or returns that they are unavailable	GOAT will provide information, from reputed online databases, in regards to the current visualization context. For example, if visualizing a gene, the researcher will be able to easily find more information on the engine.

Table 5: Major needs/user requirements for GOAT

3.4 Alternatives and competition

3.4.1 MetaboAnalyst, (*open source*)

Serves as a visualization and analysis tool but dedicated exclusively to metabolites. The client would like a user interface similar to MetaboAnalyst. Refer to **Appendix B** for MetaboAnalyst screenshots.
http://www.metaboanalyst.ca/

3.4.2 LocusZoom (*open source*)

Used to plot regional association results from genome-wide association scans or candidate gene studies but have to know locus region from data of the researcher prior to accessing it on this site in order to visualize it.
http://locuszoom.sph.umich.edu/locuszoom/

3.4.3 SNPTest (*open source*)

Program for the analysis of single SNP association in GWAS studies, but does not provide visualization tools.
https://mathgen.stats.ox.ac.uk/genetics_software/snptest/old/snptest.html

3.4.4 GWAS Diagram Browser (*open source*)

Pipeline is dedicated to GWAS studies but doesn't treat SNPTest file outputs.
http://www.ebi.ac.uk/fgpt/gwas/

3.4.5 HAPGEN (*open source*)

Program to simulate case control datasets at linked SNP markers conditional upon a set of known haplotypes.
https://mathgen.stats.ox.ac.uk/genetics_software/hapgen/hapgen2.html

3.4.6 Biopython packages (*open source*)

Demands programming knowledge in that specific language and for that package.
http://biopython.org/

3.4.7 UCSC genome browser (*proprietary closed source*)

Contains the reference sequence and working draft assemblies for a large collection of genomes. It is a very large software suite with many parameters; and only 1% of it is needed by the geneticist.
https://genome.ucsc.edu/cgi-bin/hgGateway

3.4.8 IGV

The **Integrative Genomics Viewer (IGV)** is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations. Great tool to visualize

region surrounding SNPs, but has to be access from external links with region of interest known prior to query.
https://www.broadinstitute.org/igv/

3.4.8 Sigmaplot (*proprietary closed source*)

SigmaPlot is **a software** product that helps researchers and engineers analyze their data, create precise plots and charts, develop publication-quality graphs and customize all analysis needs.

http://www.sigmaplot.com/products/sigmaplot/sigmaplot-details.php

4. Product overview

4.1 Product perspective

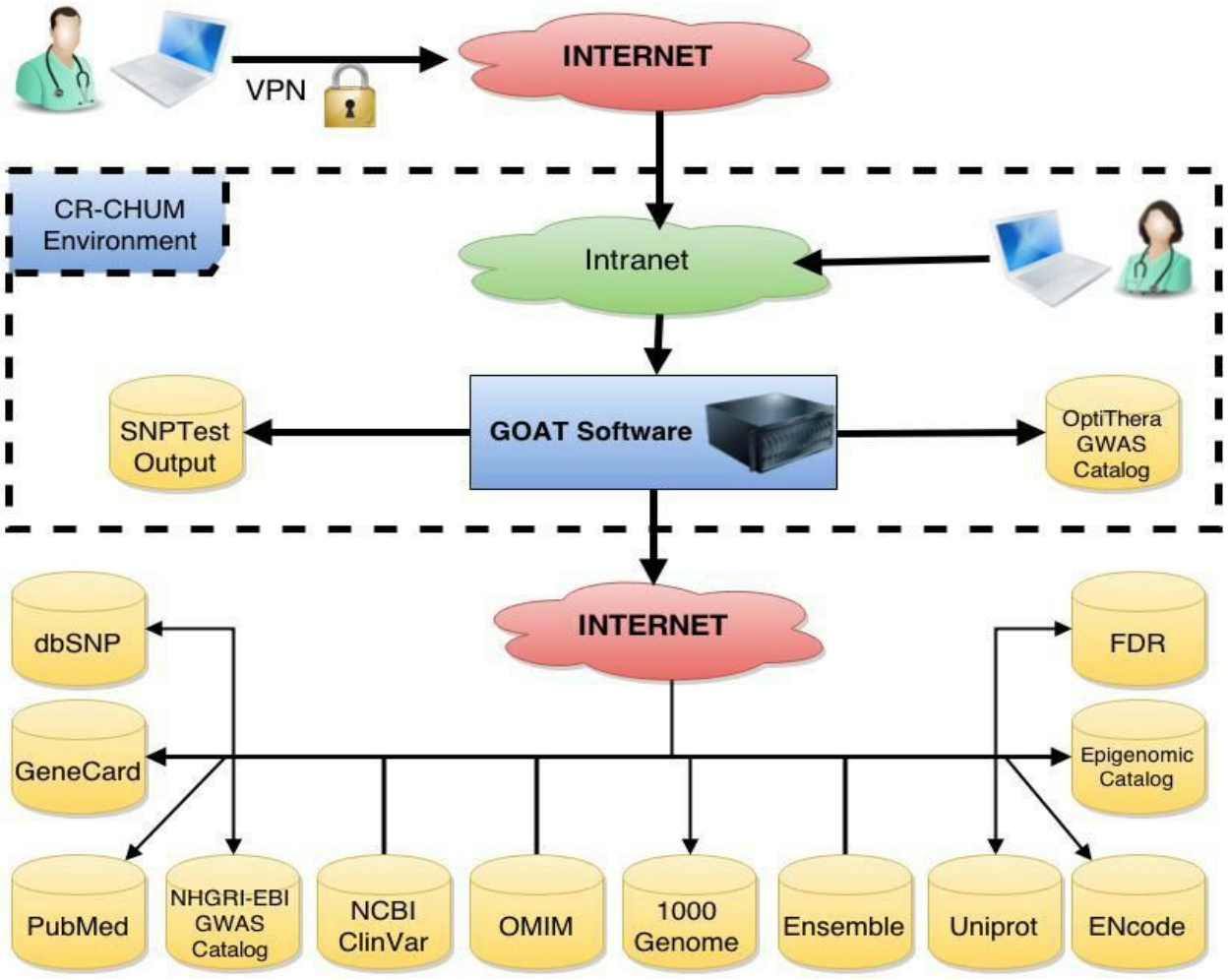
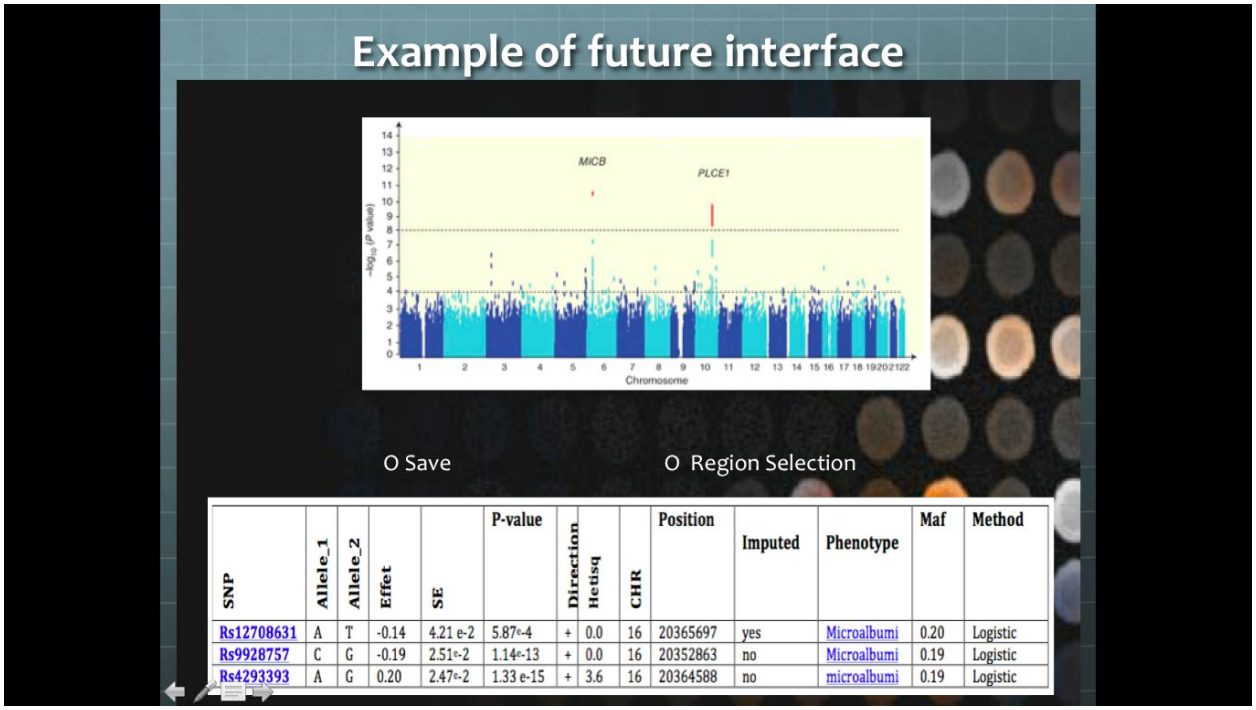


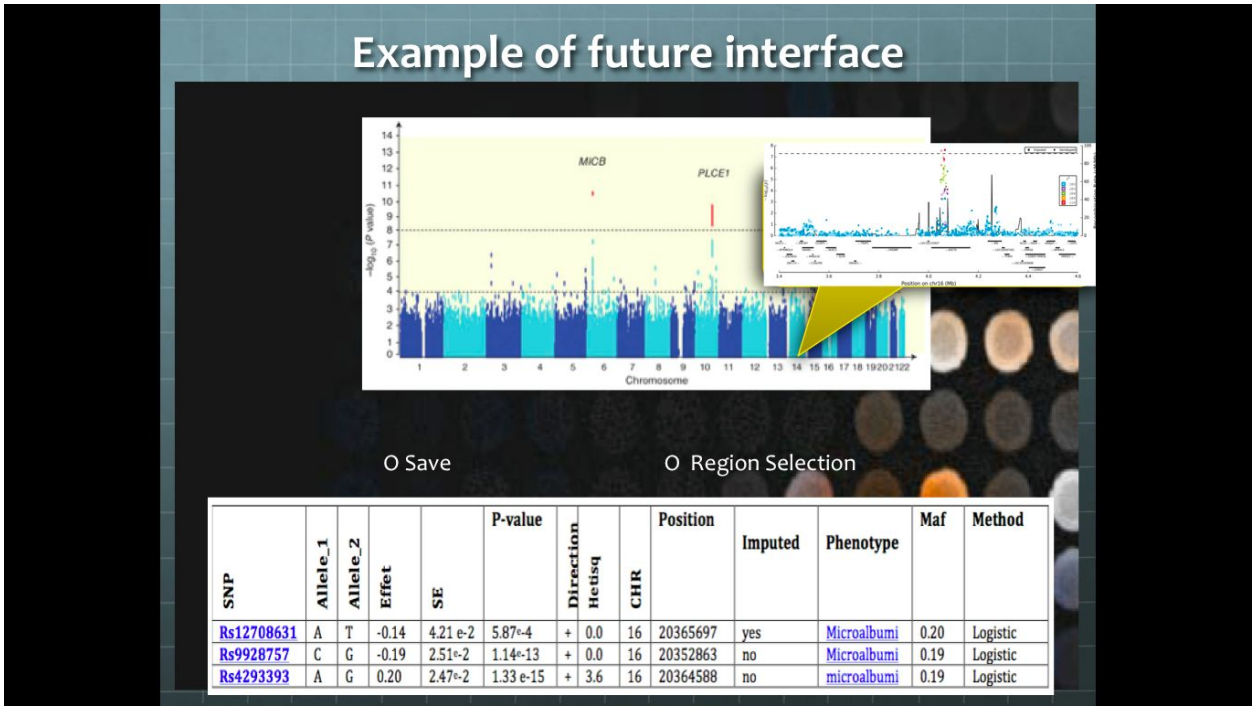
Figure 2. GOAT user environment and related open source projects involved

Ideas of User Interfaces to be developed



SNP	Allele_1	Allele_2	Effet	SE	P-value	Direction	Hetisq	CHR	Position	Imputed	Phenotype	Maf	Method
Rs12708631	A	T	-0.14	4.21 e-2	5.87e-4	+	0.0	16	20365697	yes	Microalbumi	0.20	Logistic
Rs9928757	C	G	-0.19	2.51e-2	1.14e-13	+	0.0	16	20352863	no	Microalbumi	0.19	Logistic
Rs4293393	A	G	0.20	2.47e-2	1.33 e-15	+	3.6	16	20364588	no	microalbumi	0.19	Logistic

Figure 2. First page to appear after gene Query



SNP	Allele_1	Allele_2	Effet	SE	P-value	Direction	Hetisq	CHR	Position	Imputed	Phenotype	Maf	Method
Rs12708631	A	T	-0.14	4.21 e-2	5.87e-4	+	0.0	16	20365697	yes	Microalbumi	0.20	Logistic
Rs9928757	C	G	-0.19	2.51e-2	1.14e-13	+	0.0	16	20352863	no	Microalbumi	0.19	Logistic
Rs4293393	A	G	0.20	2.47e-2	1.33 e-15	+	3.6	16	20364588	no	microalbumi	0.19	Logistic

Figure 3. Region selection generates a popup.

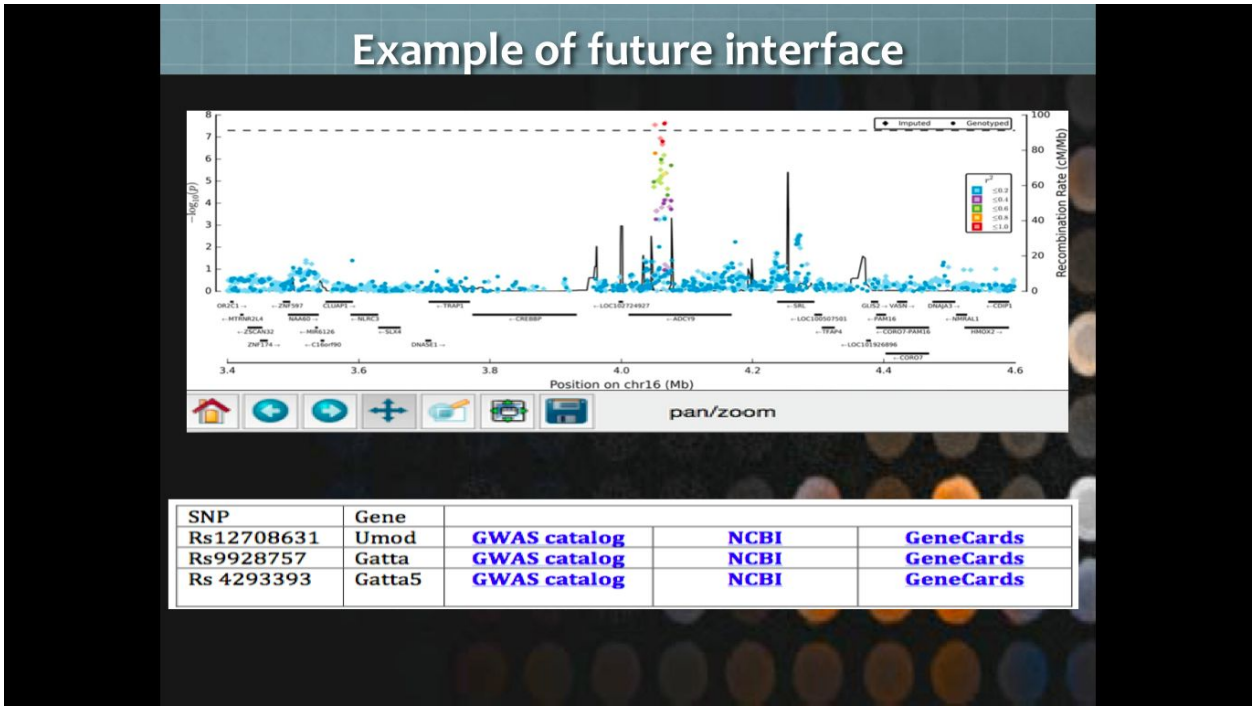


Figure 4. Information of region selection, and gene information from external links

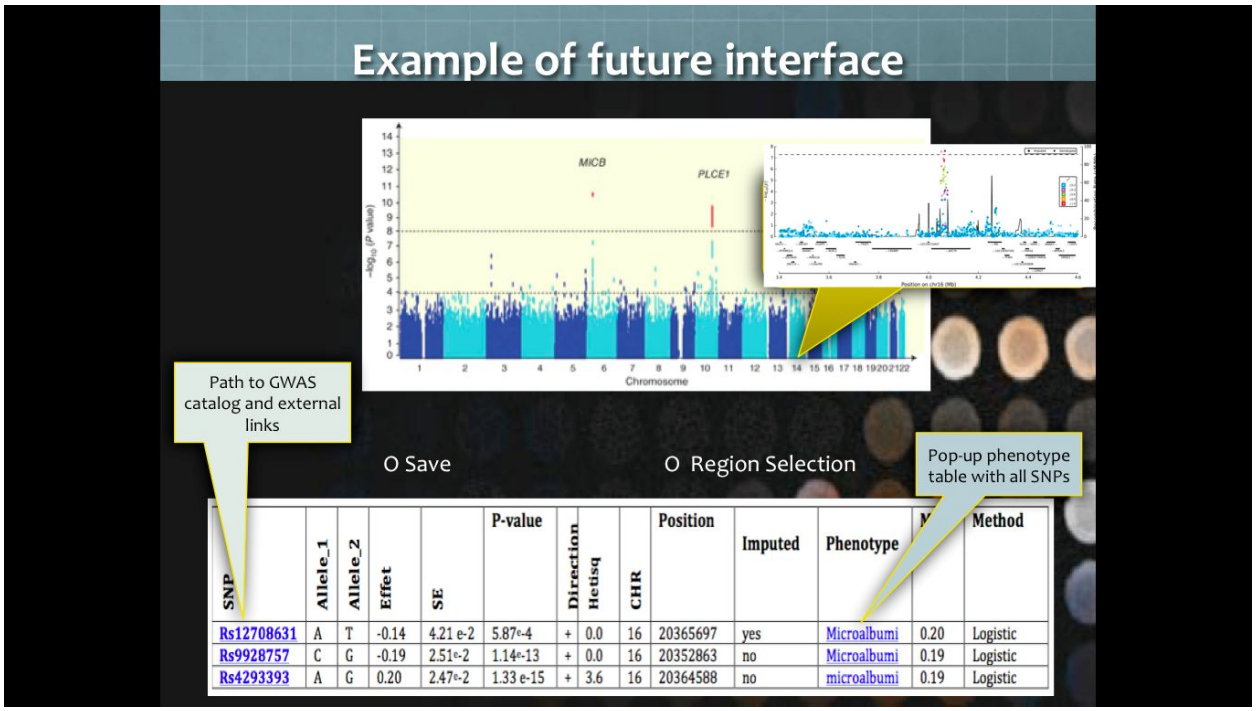


Figure 5. Overall view of GOAT gene query interface.

4.2 Assumptions and dependencies

- GOAT will collect external database information for which the rights are free or can be acquired.

- CRCHUM will provide test data to the development team in order to facilitate development.

4.3 Cost and Price

This software will be developed free of charge by the members of the project team: Dr A. April (professor at ÉTS), D. Lauzon (a PhD student in software Engineering at ÉTS). B.Kanzki is the principal developer and C.Urvoy will improve the UI of the prototype developed by B.Kanzki.

4.4 Licensing

In this document the ÉTS/UdeM team have translated/enriched user requirements (expressed by users) into more detail in order to capture the overall long term product requirement of GOAT. The document refers to public information about the following open-source projects: MetaboAnalyst, LocusZoom, SNPTest, GWAS Diagram Browser, HapGen, BioPython, IGV and many other open source projects. Before GOAT uses any of the information of these sources, their individual licences will have to be investigated. Only open-source and publicly available material will be integrated to GOAT.

GOAT Software License: The resulting GOAT software will be licensed: **General Public License: Gpl v3 or any later versions**
Installation: The open source software will be available publicly. Installation of the software can easily be made by CRCHUM bioinformatics specialists at Dr. Hamet CRCHUM Lab.

GOAT Vision Document License: The resulting GOAT vision document is licensed: Creative commons - **attribution-shareAlike 4.0 International license**

5. Product features

This section highlights the key features of this new software product. Figure 3 presents the GOAT overall process.

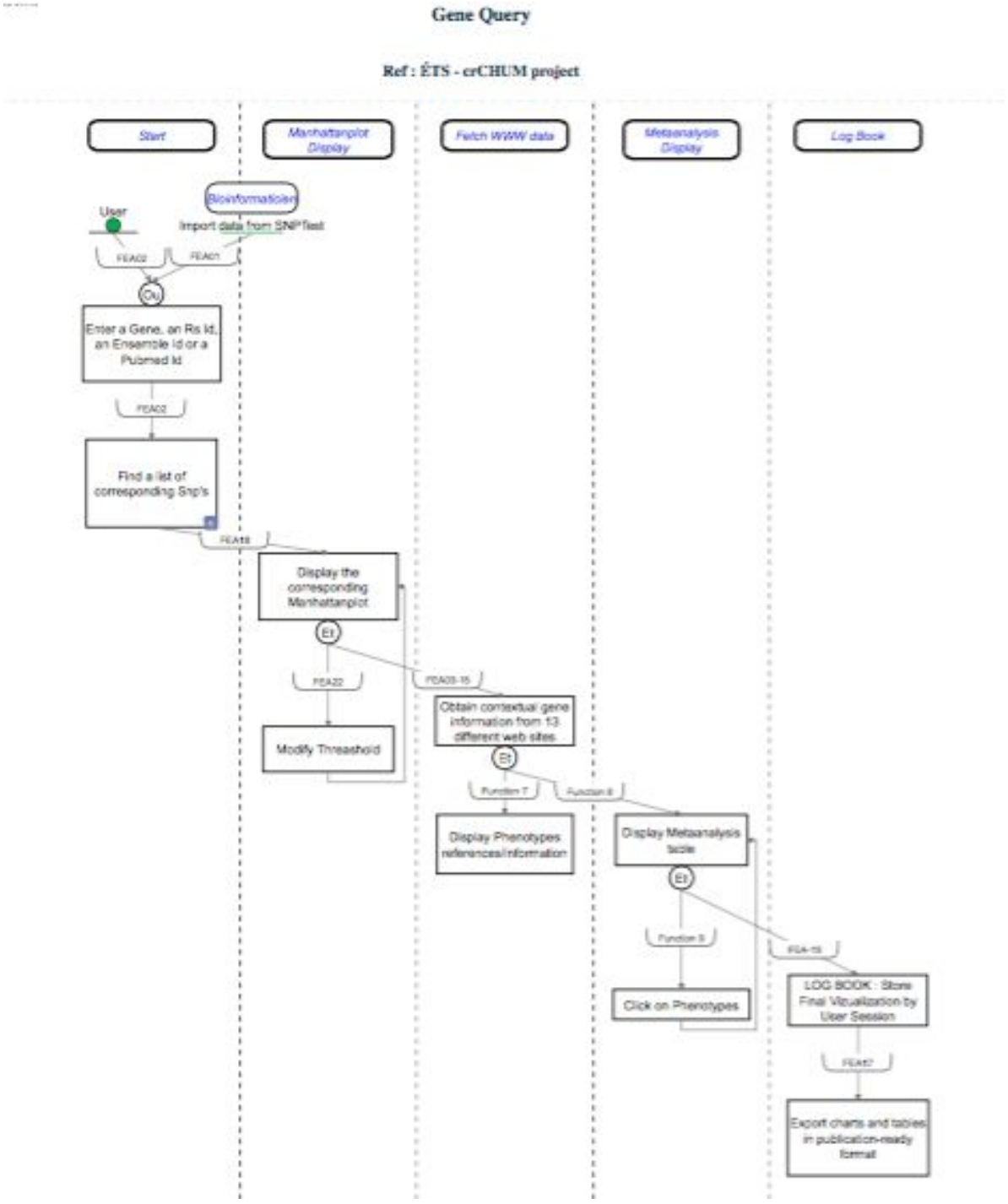


Figure 6: GOAT process diagram

FEA01 – Import data from SNPTest

The bio-informaticians would like a way to import data into the GOAT's database. The file format is an output from SNPTest that is currently stored in a MySQL database. Therefore a functionality is required to load this data from the current sql tables of the existing database. GOAT also needs to be able to support custom columns from SNPtest.

FEA02 – Filter SNPs by access key

Users would like to be able to find the list of SNPs associated with a gene using one of the following four access keys:

- Gene name
- RS ID
- Ensemble ID
- Pubmed ID

Each of these access keys can select one or more SNPs. A minimum of 2 SNPs are required to be able to visualize the data.

FEA03 – Contextual gene information from: GeneCards

When visualizing data, the users would like to have an hyperlink to the *GeneCards* website directly to gene page information. This means that gene information would be a parameter automatically passed to the corresponding website (when this functionality is allowed) to contextualize the link precisely on the information required.

GeneCards is a searchable, integrated database of human genes that provides comprehensive, updated, and user-friendly information on all known and predicted human genes. It includes Ensemble, Uniprot, OMIM, NCBI, ClinVar, Pubmed, dbSNP. GeneCards extracts and integrates gene-related data, including genomic, transcriptomic, proteomic, genetic, clinical, and functional information. This is automatically mined from >100 carefully selected web sources, thereby allowing one-stop access to a very broad information base. GeneCards overcomes barriers of data format and heterogeneity, and uses standard nomenclature and approved gene symbols. It presents a rich subset of data for each gene, and provides deep links to the original sources for further scrutiny. GeneCards is widely used, and assists in the understanding of gene-related aspects of biology and medicine. http://www.genecards.org/

dbSNP is a free public archive for genetic variation within and across different species developed and hosted by NCBI. http://www.ncbi.nlm.nih.gov/projects/SNP/

PubMed comprises more than 24 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites. http://www.pubmed.org/

ClinVaraggregates information about genomic variation and its relationship to human health. http://www.ncbi.nlm.nih.gov/clinvar/

OMIM (*Online Mendelian Inheritance in Man*) is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily. OMIM is authored and edited at the McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, under the direction of Dr. Ada Hamosh. http://www.ncbi.nlm.nih.gov/omim or http://www.omim.org/

1000 Genomes Project finds most genetic variants that have frequencies of at least 1% in the populations studied. The team wants to see allele frequencies per population.
http://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.
http://useast.ensembl.org/index.html?redirect=no

The mission of **UniProt** is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.
http://www.uniprot.org/

FEA04 – Contextual gene information from: NHGRI-EBI GWAS Catalog

When visualizing data, the users would like to have an hyperlink to the *NHGRI GWAS Catalog* website directly to gene page information.

GWAS Catalog is a quality controlled, manually curated, literature-driven collection of all published genome-wide association studies, produced by a collaboration between EMBL-EBI and NHGRI. These GWAS studies assays at least 100,000 SNPs and all SNP-trait associations have p-values < 1.0×10^{-5}.
https://www.ebi.ac.uk/gwas/

FEA05 – Contextual gene information from: ENcode

When visualizing data, the users would like to have an hyperlink to the *ENcode* website directly to gene page information.

The Encyclopedia of DNA Elements (**ENCODE**) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.
http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg18&position=chrX%3A151073054-151383976&hgsid=426788787_i7aYCjc4DgZSO4MS6dmiTzZlAmma

FEA06 – Contextual gene information from: Epigenomic catalog

When this project is operation, the users would like to have an hyperlink to the *Epigenomic* website directly to gene page information (*this project is not yet operational at this time*).

Human epigenomic catalog (HEP) aims to identify, catalogue and interpret genome wide DNA methylation patterns of all human genes in all major tissues.
http://www.epigenome.org/?page=project

FEA07 – Contextual calculations of gene information from: FDR

Users would like to have to see, in the presented table of information, the FDR calculations done for the information visualized.

False discovery rate (FDR) is is the expected proportion of Type I errors among the rejected hypotheses FDR = E(V/R | R>0)P(R>0). In many cases (particularly in genomics) we can live with a certain number of false positives. In these cases, the more relevant quantity to control is the false discovery rate (FDR). False discovery rate (FDR) is designed to control the proportion of false positives among the set of rejected hypotheses.

FEA08– Log Book (provenance/traceability)

Users would like to have provenance/traceability of the discovery actions.

Every user actions will be recorded in a "Log Book" database (refer to CFR21 part 11 requirements). This will allow the user to easily find, select and replay any previously saved exploration sessions including visualizations. This functionality would be accessible by user sessions.

FEA09 – Export charts and tables.

Users would like to allow visualized Charts to be exported in pdf format and visualized tables to be exportable to .csv format. This action should be simple, resulting figures should be of high quality.

FEA10 – Visualization: GWAS Manhattan Plot

User would like to see this kind of plot which will give them an idea of which SNPs are the most interesting, by chromosome.

A **Manhattan plot** is a type of scatter plot, usually used to display data with a large number of data-points - many of non-zero amplitude, and with a distribution of higher-magnitude values. In GWAS Manhattan plots, genomic coordinates are displayed along the X-axis, with the negative logarithm of the association *P*-value for each single nucleotide polymorphism (SNP) displayed on the Y-axis, meaning that each dot on the Manhattan plot signifies a SNP. Because the strongest associations have the smallest *P*-values (e.g., 10^{-15}), their negative logarithms will be the greatest (e.g., 15).

FEA11 – Visualization: Box Plot

In some cases, user would like to see a Box-plot representation of some data in order to see the distribution of the data.

Box-Plot is n descriptive statistics, a **box plot** or **boxplot** is a convenient way of graphically depicting groups of numerical data through their quartiles. Box plots may also have lines extending vertically from the boxes (*whiskers*) indicating variability outside the upper and lower quartiles, hence the terms

box-and-whisker plot and **box-and-whisker diagram**. Outliers may be plotted as individual points. This is also called a "box and whisker plot".

FEA12 – Interactive visualization: Genome Viewer

Users would like to be able to interact with a graph currently visualised. The interaction would be similar to a biopython based graph with selected region feature, gene name, and other information to be assessed in the SRS stage.

FEA13 – Interactive visualization: Region Selection

User would like to be able to determine gene region visualization, by selecting an interval underlying before and after the SNP of interest. Could be 1000 bases before and 1000 after.

FEA14 – Interactive visualization: Modify threshold

Users would like to have a system set threshold default default that can be changed by the user interactively. The user would be able to input the desired threshold that he wishes to apply on the dataset and changes would appear on the manhattan plot interactively.

FEA15 – LD Regression Score

Users would like GOAT to Computes LD (Linkage Disequilibrium) regression score for a region of interest in the graph.

LD Score regression, that quantifies the contribution of each by examining the relationship between test statistics and linkage disequilibrium (LD). The LD Score regression intercept can be used to estimate a more powerful and accurate correction factor than genomic control.

FEA16 – MAF per population (from 1000 Genome Project)

Users would like GOAT to display the MAF per population for region of interest.

Allele Frequency. Population: AA, AS, CAU.
This information would be obtained directly from the 1000 Genome Project.

6. Constraints

The first version/prototype of GOAT needs should be delivered within 40 days of the acceptance of this vision document (a usable proof of concept).

7. Quality criteria (non-functional requirements)

CN01 – Data must be secured: During SRS security requirements must be explicitly defined
CN02 – Maintainability: GOAT should use Software Engineering best practices to allow a design pattern approach for easier maintainability.
CN03 – Usability. The user interface must be easy to use

8. Features attributes

This section summarizes the features of the system according to the benefits they bring to the customer, the effort required to implement the risk associated with their implementation and their stability (probability of change). Each value of these attributes are detailed in Appendix A.

Features	State	Benefits	Effort	Risk
FEA01 – Generic interface from external data sources (import data from SnpTest)	Partial		Low	just import a .csv file for now
FEA02 – Filter SNPs by access key	Approved		Low	
FEA03 – Contextual gene information from GeneCards	Approved		high	licences, availability of ap
FEA04 – Contextual gene information from: NHGRI-EBI GWAS Catalog	Approved		low	licences, availability of APii
FEA05 – Contextual gene information from: ENcode	Approved		low	licences, availability of APi
FEA13 – Contextual gene information from: Epigenomic catalog	Delayed		low	project not available
FEA07 – Contextual calculations of gene information from FDR	Approved		low	formula details
FEA08 – Log Book (provenance/traceability)	Delayed		high	other project and vision document will be done
FEA09 – Export charts and tables	Approved		low	
FEA10 – Visualization: GWAS Manhattan Plot	Approved		medium	
FEA11 – Interactive visualization: Box Plot	Delayed		medium	
FEA12 – Interactive visualization: GWAS Manhattan Plot	Approved		medium	
FEA13 – Interactive visualization: Region Selection	Approved		low	

FEA14 – Interactive visualization: Modify threshold	Approved		low	
FEA15 – LD Regression Score calculations	Approved		low	
FEA16 – MAF per population (from 1000 Genome Project)	Approved		low	

Table 7: GOAT features state, benefit, effort, risk and stability

9. Other Product Requirements

9.1 Applicable standards

9.1.1 Internal

crCHUM representatives have not presented any specific internal standards requirements to be imposed on this software.

9.1.2 External

The CFR21 Part 11 requirements may be imposed on GOAT. This will be assessed in another vision document addressing provenance and another separate project leaded by D.Lauzon.

9.2 System requirements

9.2.1 Security

9.2.2 Accessibility: to be handled by Dr.Hamet informatics staff:

- GOAT should be available internally, at the crCHUM, by Dr Pavel Hamet's staff, but Dr Hamet will ask the CRCHUM to give him external accesses in order to access GOAT from outside the crCHUM.

9.2.3 Portability

- The client software will be accessible from any desktop workstation with the with the following web browser installed: Firefox version 35
- The server will be compatible with Ubuntu Linux 14.04 or later.

9.3 Performance requirements

GOAT should be faster the the existing tool used. SRS should assess design approaches that allow the following functions to be responsive. The following items requires good response time:

- FEA01-Extraction of data from the internal database.
- FEA10-Statistical analysis execution
- FEA11-Exploring a graph
- FEA3,4,5,6 and 16 - obtaining information from external sources

- FEA7, 16-computing FDR and LD
- FEA09-Exporting a graph or a table
- FEA8-LogBook capture

9.4 Environmental requirements.

A web server, in the CRCHUM, accessible to Dr Hamet's staff and having a VPN access for secure external access is planned.

10. Documentation requirements.

10.1 User's Manual

No user manual is planned for GOAT

10.2 Online help

No online help is planned for GOAT

10.3 Installation Guides, configuration, and README file

At the end of the project, the software will be installed, by the bioinformatics specialist, and will be operated a crCHUM web server for internal use only. Configuration and readme file will be developed for GOAT and made available publicly.

APPENDIX

Appendix A - Feature attributes.

This section defines the attributes we associate with different characteristics of the system.

State

Proposed	This status indicates that the feature is available and must be the subject of discussion within the project team for its acceptance or rejection.
Partial	Will be partially implemented in the current system
Approved	This status indicates that the feature has been selected for an upcoming development.
Incorporated	This status indicates that this feature has been incorporated into the system during development .
Delayed	Feature will be implemented in a future version

Benefits

Critical	This level indicates that the feature is essential for the software. This means that the customer will not want to have a system without this feature.
Important	This level indicates that this feature is important. However, if it is not implemented, the software can be used. It is up to the customer to decide whether to have the software without this feature or if the project stops.
Useful	The system will integrate all the important features to achieve the anticipated profit. However if time permits and if the project team is available, some features can be added to increase the customer's benefit.

Effort

High	This level indicates that the amount of effort required is at least two weeks of work.
Average	This level indicates that the amount of effort required is between one and two weeks of work.

Low	This level indicates that the amount of effort required is below a workweek.

Risk

High	This level indicates that there is a high level of uncertainty about the duration or cost of implementation, or even a risk of cancellation .
Average	This level indicates that there is some uncertainty about the length or cost of the implementation.
Low	This level indicates that the duration and costs are well defined and are unlikely to change.

Stability

High	This level indicates that this feature will not undergo change. This also shows that it was well understood.
Average	This level indicates that the feature may be subject to change.
Low	This level of stability shows the probability that this feature can be changed over time.

Appendix B - License

This document has been assigned a Creative commons - **attribution-shareAlike 4.0 International license**

You are free to:

Share — copy and redistribute the material in any medium or format
Adapt — remix, transform, and build upon the material for any purpose, even commercially.
The licensor cannot revoke these freedoms as long as you follow the license terms.

Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

Creative Commons Corporation (“Creative Commons”) is not a law firm and does not provide legal services or legal advice. Distribution of Creative Commons public licenses does not create a lawyer-client or other relationship. Creative Commons makes its licenses and related information available on an “as-is” basis. Creative Commons gives no warranties regarding its licenses, any material licensed under their terms and conditions, or any related information. Creative Commons disclaims all liability for damages resulting from their use to the fullest extent possible.

Using Creative Commons Public Licenses

Creative Commons public licenses provide a standard set of terms and conditions that creators and other rights holders may use to share original works of authorship and other material subject to copyright and certain other rights specified in the public license below. The following considerations are for informational purposes only, are not exhaustive, and do not form part of our licenses.

Considerations for licensors: Our public licenses are intended for use by those authorized to give the public permission to use material in ways otherwise restricted by copyright and certain other rights. Our licenses are irrevocable. Licensors should read and understand the terms and conditions of the license they choose before applying it. Licensors should also secure all rights necessary before applying our licenses so that the public can reuse the material as expected. Licensors should clearly mark any material not subject to the license. This includes other CC-licensed material, or material used under an exception or limitation to copyright. More considerations for licensors.
Considerations for the public: By using one of our public licenses, a licensor grants the public permission to use the licensed material under specified terms and conditions. If the licensor's permission is not necessary for any reason–for example, because of any applicable exception or limitation to copyright–then that use is not regulated by the license. Our licenses grant only permissions under copyright and certain other rights that a licensor has authority to grant. Use of the licensed material may still be restricted for other reasons, including because others have copyright or other rights in the material. A licensor may make special requests, such as asking that all changes be marked or described. Although not required by our licenses, you are encouraged to respect those requests where reasonable. More considerations for the public.
Creative Commons Attribution-ShareAlike 4.0 International Public License

By exercising the Licensed Rights (defined below), You accept and agree to be bound by the terms and conditions of this Creative Commons Attribution-ShareAlike 4.0 International Public License ("Public License"). To the extent this Public License may be interpreted as a contract, You are granted the Licensed Rights in consideration of Your acceptance of these terms and conditions, and the Licensor grants You such rights in consideration of benefits the Licensor receives from making the Licensed Material available under these terms and conditions.

Section 1 – Definitions.

Adapted Material means material subject to Copyright and Similar Rights that is derived from or based upon the Licensed Material and in which the Licensed Material is translated, altered, arranged, transformed, or otherwise modified in a manner requiring permission under the Copyright and Similar Rights held by the Licensor. For purposes of this Public License, where the Licensed Material is a musical work, performance, or sound recording, Adapted Material is always produced where the Licensed Material is synched in timed relation with a moving image.
Adapter's License means the license You apply to Your Copyright and Similar Rights in Your contributions to Adapted Material in accordance with the terms and conditions of this Public License.
BY-SA Compatible License means a license listed at creativecommons.org/compatiblelicenses, approved by Creative Commons as essentially the equivalent of this Public License.
Copyright and Similar Rights means copyright and/or similar rights closely related to copyright including, without limitation, performance, broadcast, sound recording, and Sui Generis Database Rights, without regard to how the rights are labeled or categorized. For purposes of this Public License, the rights specified in Section 2(b)(1)-(2) are not Copyright and Similar Rights.
Effective Technological Measures means those measures that, in the absence of proper authority, may not be circumvented under laws fulfilling obligations under Article 11 of the WIPO Copyright Treaty adopted on December 20, 1996, and/or similar international agreements.

Exceptions and Limitations means fair use, fair dealing, and/or any other exception or limitation to Copyright and Similar Rights that applies to Your use of the Licensed Material.
License Elements means the license attributes listed in the name of a Creative Commons Public License. The License Elements of this Public License are Attribution and ShareAlike.
Licensed Material means the artistic or literary work, database, or other material to which the Licensor applied this Public License.
Licensed Rights means the rights granted to You subject to the terms and conditions of this Public License, which are limited to all Copyright and Similar Rights that apply to Your use of the Licensed Material and that the Licensor has authority to license.
Licensor means the individual(s) or entity(ies) granting rights under this Public License.
Share means to provide material to the public by any means or process that requires permission under the Licensed Rights, such as reproduction, public display, public performance, distribution, dissemination, communication, or importation, and to make material available to the public including in ways that members of the public may access the material from a place and at a time individually chosen by them.
Sui Generis Database Rights means rights other than copyright resulting from Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, as amended and/or succeeded, as well as other essentially equivalent rights anywhere in the world.
You means the individual or entity exercising the Licensed Rights under this Public License. Your has a corresponding meaning.
Section 2 – Scope.

License grant.
Subject to the terms and conditions of this Public License, the Licensor hereby grants You a worldwide, royalty-free, non-sublicensable, non-exclusive, irrevocable license to exercise the Licensed Rights in the Licensed Material to:
reproduce and Share the Licensed Material, in whole or in part; and
produce, reproduce, and Share Adapted Material.
Exceptions and Limitations. For the avoidance of doubt, where Exceptions and Limitations apply to Your use, this Public License does not apply, and You do not need to comply with its terms and conditions.
Term. The term of this Public License is specified in Section 6(a).
Media and formats; technical modifications allowed. The Licensor authorizes You to exercise the Licensed Rights in all media and formats whether now known or hereafter created, and to make technical modifications necessary to do so. The Licensor waives and/or agrees not to assert any right or authority to forbid You from making technical modifications necessary to exercise the Licensed Rights, including technical modifications necessary to circumvent Effective Technological Measures. For purposes of this Public License, simply making modifications authorized by this Section 2(a)(4) never produces Adapted Material.
Downstream recipients.
Offer from the Licensor – Licensed Material. Every recipient of the Licensed Material automatically receives an offer from the Licensor to exercise the Licensed Rights under the terms and conditions of this Public License.
Additional offer from the Licensor – Adapted Material. Every recipient of Adapted Material from You automatically receives an offer from the Licensor to exercise the Licensed Rights in the Adapted

Material under the conditions of the Adapter's License You apply.
No downstream restrictions. You may not offer or impose any additional or different terms or conditions on, or apply any Effective Technological Measures to, the Licensed Material if doing so restricts exercise of the Licensed Rights by any recipient of the Licensed Material.
No endorsement. Nothing in this Public License constitutes or may be construed as permission to assert or imply that You are, or that Your use of the Licensed Material is, connected with, or sponsored, endorsed, or granted official status by, the Licensor or others designated to receive attribution as provided in Section 3(a)(1)(A)(i).
Other rights.

Moral rights, such as the right of integrity, are not licensed under this Public License, nor are publicity, privacy, and/or other similar personality rights; however, to the extent possible, the Licensor waives and/or agrees not to assert any such rights held by the Licensor to the limited extent necessary to allow You to exercise the Licensed Rights, but not otherwise.
Patent and trademark rights are not licensed under this Public License.
To the extent possible, the Licensor waives any right to collect royalties from You for the exercise of the Licensed Rights, whether directly or through a collecting society under any voluntary or waivable statutory or compulsory licensing scheme. In all other cases the Licensor expressly reserves any right to collect such royalties.
Section 3 – License Conditions.

Your exercise of the Licensed Rights is expressly made subject to the following conditions.

Attribution.

If You Share the Licensed Material (including in modified form), You must:

retain the following if it is supplied by the Licensor with the Licensed Material:
identification of the creator(s) of the Licensed Material and any others designated to receive attribution, in any reasonable manner requested by the Licensor (including by pseudonym if designated);
a copyright notice;
a notice that refers to this Public License;
a notice that refers to the disclaimer of warranties;
a URI or hyperlink to the Licensed Material to the extent reasonably practicable;
indicate if You modified the Licensed Material and retain an indication of any previous modifications; and
indicate the Licensed Material is licensed under this Public License, and include the text of, or the URI or hyperlink to, this Public License.
You may satisfy the conditions in Section 3(a)(1) in any reasonable manner based on the medium, means, and context in which You Share the Licensed Material. For example, it may be reasonable to satisfy the conditions by providing a URI or hyperlink to a resource that includes the required information.
If requested by the Licensor, You must remove any of the information required by Section 3(a)(1)(A) to the extent reasonably practicable.

ShareAlike.
In addition to the conditions in Section 3(a), if You Share Adapted Material You produce, the following conditions also apply.

The Adapter's License You apply must be a Creative Commons license with the same License Elements, this version or later, or a BY-SA Compatible License.
You must include the text of, or the URI or hyperlink to, the Adapter's License You apply. You may satisfy this condition in any reasonable manner based on the medium, means, and context in which You Share Adapted Material.
You may not offer or impose any additional or different terms or conditions on, or apply any Effective Technological Measures to, Adapted Material that restrict exercise of the rights granted under the Adapter's License You apply.
Section 4 – Sui Generis Database Rights.

Where the Licensed Rights include Sui Generis Database Rights that apply to Your use of the Licensed Material:

for the avoidance of doubt, Section 2(a)(1) grants You the right to extract, reuse, reproduce, and Share all or a substantial portion of the contents of the database;
if You include all or a substantial portion of the database contents in a database in which You have Sui Generis Database Rights, then the database in which You have Sui Generis Database Rights (but not its individual contents) is Adapted Material, including for purposes of Section 3(b); and
You must comply with the conditions in Section 3(a) if You Share all or a substantial portion of the contents of the database.
For the avoidance of doubt, this Section 4 supplements and does not replace Your obligations under this Public License where the Licensed Rights include other Copyright and Similar Rights.
Section 5 – Disclaimer of Warranties and Limitation of Liability.

Unless otherwise separately undertaken by the Licensor, to the extent possible, the Licensor offers the Licensed Material as-is and as-available, and makes no representations or warranties of any kind concerning the Licensed Material, whether express, implied, statutory, or other. This includes, without limitation, warranties of title, merchantability, fitness for a particular purpose, non-infringement, absence of latent or other defects, accuracy, or the presence or absence of errors, whether or not known or discoverable. Where disclaimers of warranties are not allowed in full or in part, this disclaimer may not apply to You.
To the extent possible, in no event will the Licensor be liable to You on any legal theory (including, without limitation, negligence) or otherwise for any direct, special, indirect, incidental, consequential, punitive, exemplary, or other losses, costs, expenses, or damages arising out of this Public License or use of the Licensed Material, even if the Licensor has been advised of the possibility of such losses, costs, expenses, or damages. Where a limitation of liability is not allowed in full or in part, this limitation may not apply to You.
The disclaimer of warranties and limitation of liability provided above shall be interpreted in a manner that, to the extent possible, most closely approximates an absolute disclaimer and waiver of all liability.
Section 6 – Term and Termination.

This Public License applies for the term of the Copyright and Similar Rights licensed here. However, if You fail to comply with this Public License, then Your rights under this Public License terminate automatically.
Where Your right to use the Licensed Material has terminated under Section 6(a), it reinstates:

automatically as of the date the violation is cured, provided it is cured within 30 days of Your discovery of the violation; or
upon express reinstatement by the Licensor.
For the avoidance of doubt, this Section 6(b) does not affect any right the Licensor may have to seek remedies for Your violations of this Public License.
For the avoidance of doubt, the Licensor may also offer the Licensed Material under separate terms or conditions or stop distributing the Licensed Material at any time; however, doing so will not terminate this Public License.
Sections 1, 5, 6, 7, and 8 survive termination of this Public License.
Section 7 – Other Terms and Conditions.

The Licensor shall not be bound by any additional or different terms or conditions communicated by You unless expressly agreed.
Any arrangements, understandings, or agreements regarding the Licensed Material not stated herein are separate from and independent of the terms and conditions of this Public License.
Section 8 – Interpretation.

For the avoidance of doubt, this Public License does not, and shall not be interpreted to, reduce, limit, restrict, or impose conditions on any use of the Licensed Material that could lawfully be made without permission under this Public License.
To the extent possible, if any provision of this Public License is deemed unenforceable, it shall be automatically reformed to the minimum extent necessary to make it enforceable. If the provision cannot be reformed, it shall be severed from this Public License without affecting the enforceability of the remaining terms and conditions.
No term or condition of this Public License will be waived and no failure to comply consented to unless expressly agreed to by the Licensor.
Nothing in this Public License constitutes or may be interpreted as a limitation upon, or waiver of, any privileges and immunities that apply to the Licensor or You, including from the legal processes of any jurisdiction or authority.
Creative Commons is not a party to its public licenses. Notwithstanding, Creative Commons may elect to apply one of its public licenses to material it publishes and in those instances will be considered the “Licensor.” The text of the Creative Commons public licenses is dedicated to the public domain under the CC0 Public Domain Dedication. Except for the limited purpose of indicating that material is shared under a Creative Commons public license or as otherwise permitted by the Creative Commons policies published at creativecommons.org/policies, Creative Commons does not authorize the use of the trademark “Creative Commons” or any other trademark or logo of Creative Commons without its prior written consent including, without limitation, in connection with any unauthorized modifications to any of its public licenses or any other arrangements, understandings, or agreements concerning use of