

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

RAPPORT DE PROJET DE NEUF CRÉDITS PRÉSENTÉ À
L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

COMME EXIGENCE PARTIELLE
À L'OBTENTION D'UNE
MAÎTRISE EN GÉNIE LOGICIEL

PAR
Rodolfo RAMOS

ANALYSE DES BESOINS POUR UNE PLATEFORME
BIO-INFORMATIQUE : MODULE GÉNÉTIQUE

MONTRÉAL, LE AOÛT 2014



Rodolfo Ramos, 2014



Cette licence [Creative Commons](https://creativecommons.org/licenses/by-nc-nd/4.0/) signifie qu'il est permis de diffuser, d'imprimer ou de sauvegarder sur un autre support une partie ou la totalité de cette œuvre à condition de mentionner l'auteur, que ces utilisations soient faites à des fins non commerciales et que le contenu de l'œuvre n'ait pas été modifié.

PRÉSENTATION DU JURY

CE RAPPORT DE PROJET, DE 9 CRÉDITS A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE :

Prof. Alain April, directeur de projet
Département de génie logiciel et TI à l'École de technologie supérieure

IL A FAIT L'OBJET D'UNE ÉVALUATION

LE 29 AOÛT 2014

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

Je tiens à remercier, dans un premier temps, toute l'équipe d'enseignement de l'École de Technologie Supérieure de Montréal et ses intervenants professionnels responsables de la formation à la maîtrise en génie logiciel pour avoir assuré une bonne formation pendant mes études.

Je remercie tout particulièrement le professeur Alain April pour son aide et ses conseils qu'il m'a apportés lors des différents suivis du projet de maîtrise.

Je remercie également David Lauzon, Cédric Julien et toute l'équipe du groupe de recherche en pathologies musculosquelettiques de l'Hôpital Sainte-Justine pour sa continuelle collaboration pendant l'analyse des besoins de ce projet.

Finalement, je tiens exprimer toute ma gratitude à ma famille, à ma femme Giannina et mes filles Michelle et Heidi pour leur confiance et leur soutien précieux.

PLATAFORME BIO-INFORMATIQUE : MODULE GÉNÉTIQUE

Rodolfo RAMOS

RÉSUMÉ

Ce projet vise l'appui aux chercheurs du laboratoire de recherche en pathologies musculosquelettique du CHU Sainte-Justine. Le spécialiste en génétique, du laboratoire, traite, transforme et distribue l'information génétique aux différents chercheurs du laboratoire. Au moment de notre intervention, les données génétiques sont stockées à l'aide de fichiers Excel ce qui cause 1) de la duplication d'information et 2) de la difficulté d'accès à l'information. Le laboratoire profiterait grandement d'un outil de support centralisé accessible sur le web qui permettrait d'offrir l'accès aux données génétiques.

Le principal objectif de ce projet de recherche est de faire une analyse des besoins, rédiger des spécifications fonctionnelles et non-fonctionnelles de la plate-forme future. Le sous-système de gestion des données génomiques devrait avoir une interface web conviviale laquelle faciliterait les recherches des données génétique à partir d'une base de données centralisée.

PLATAFORME BIO-INFORMATIQUE: MODULE GÉNÉTIQUE

Rodolfo RAMOS

ABSTRACT

This project arises from the need to support the researchers of the musculoskeletal disorders research laboratory at the CHU Sainte-Justine. A genetician captures, transforms and distributes patient genetic information to the many researchers of the lab. At the moment of our involvement the genetic data was stored in Excel spreadsheets which causes 1) duplication of information and 2) difficulty in accessing the data. This research laboratory would profit greatly from a centralized web-based tool that allows to readily access genetic data.

The main objective of this research project is to analyze the needs and requirements, develop the functional and non-functional requirements of the future platform. The genetic data management sub-system should have a user-friendly web interface which will facilitate the query of genetic data in a centralized database.

2.8.2 Interface « Recherches des données génétiques» 23
2.8.3 Interface « Fichiers bruts » 26
2.9 Base de données 27
CHAPITRE 3 Discussion 29
3.1 Introduction..... 29
3.2 Apprentissage..... 29
3.3 Validation des interfaces..... 29
3.4 Travail future 29
3.4.1 Statistiques 30
3.4.2 Internationalisation. 30
3.4.3 Visualisation des graphiques des fichiers bruts. 30
3.4.4 Requêtes..... 30
CONCLUSION 31
LISTE DE RÉFÉRENCES BIBLIOGRAPHIQUES 51

Alain April 14-8-27 22:38
Deleted: 25
Alain April 14-8-27 22:38
Deleted: 28
Alain April 14-8-27 22:38
Deleted: 29
Alain April 14-8-27 22:38
Deleted: 31
Alain April 14-8-27 22:38
Deleted: 32
Alain April 14-8-27 22:38
Deleted: 33
Alain April 14-8-27 22:38
Deleted: 45

LISTE DES TABLEAUX

Page

Tableau 1. Résumé des intervenants.....	Error! Bookmark not defined,
Tableau 2. Résumé des utilisateurs.....	Error! Bookmark not defined,
Tableau 3. Niveaux de priorité.....	15,

Alain April 14-8-27 22:38
Deleted: 16
Alain April 14-8-27 22:38
Deleted: 16
Alain April 14-8-27 22:38
Deleted: 17

LISTE DES FIGURES

	Page
Figure 1. Processus d'obtention d'un fichier VCF.....	4
Figure 2. Plateforme Tute GENOMICS (Genomics 2014)	8
Figure 3.Statistiques de contrôle de qualités des données génétiques (Genomics 2014).....	9
Figure 4.Fenêtre de recherches des données génétiques(Genomics 2014).....	9
Figure 5.Plateforme Galaxy (Plume 2011)	11
Figure 6. Fichier Excel avec de données génétiques	14
Figure 7. Diagramme de cas d'utilisation du module génétique.	17
Figure 8. Interface de gestion des fichiers VCF.	23
Figure 9. Interface Recherche des données génétiques.	26
Figure 10. Interface pour gérer fichiers bruts	26
Figure 11. Modèle Entité-Relation de la base de données sous-système génétique.....	27
Figure 12. Fichier VCF(Marcketta 2013).	Error! Bookmark not defined.

Alain April 14-8-27 22:38

Deleted: 20

Alain April 14-8-27 22:38

Deleted: 25

Alain April 14-8-27 22:38

Deleted: 28

Alain April 14-8-27 22:38

Deleted: 28

Alain April 14-8-27 22:38

Deleted: 29

Alain April 14-8-27 22:38

Deleted: 45

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

CHU	Centre hospitalier universitaire
ADN	Acide désoxyribonucléique
VCF	« Variant Call Format »
BD	Base de données
CSV	« Comma-separated values »
CFR	Code des Règlements Fédéraux.

INTRODUCTION

La génomique permet de connaître l'ensemble du matériel génétique d'un être vivant, qu'il soit humain, une plante, un animal et même un virus. Avec le développement de la science de la génomique, en combinaison avec les avancements en bio-informatique, il est maintenant possible de faire des études sur les maladies génétiques afin d'avoir une meilleure compréhension des maladies et développement des traitements préventifs. Par exemple, « les chercheurs de l'université de Washington à Seattle (Pasquesoone 2014) sont parvenus à détecter, avec 98,1 % de certitude, 39 des 44 nouvelles mutations génétiques d'un fœtus. Ces variations génétiques ne sont partagées ni avec la mère ni avec le père. Étant donné que ces variations sont à la base d'une proportion importante de troubles génétiques, les identifier est crucial afin d'établir un diagnostic prénatal complet. Les chercheurs ont démontré que le séquençage total du génome d'un fœtus, la biologie informatique et les techniques statistiques actuelles permettent de localiser ces nouvelles mutations dans le génome du fœtus en développement ». Au Canada, le laboratoire de recherche en pathologies musculo-squelettique du docteur Alain Moreau localisé à l'hôpital Sainte-Justine effectue des recherches sur les maladies qui affectent les enfants. « La scoliose est définie comme une déformation de la colonne qui affecte quatre personnes sur cent. Elle affecte particulièrement les jeunes entre dix et dix-huit ans. En moyenne, 30% des enfants dont des parents souffrent déjà de la scoliose, ont une probabilité de développer cette maladie. Cependant, 85% des cas de scoliose se produisent sporadiquement chez les enfants sans antécédents familiaux » (Dallaire 2008).

Ce projet vise à analyser les logiciels et processus actuels de traitement des données génétiques dans ce laboratoire. Par la suite il proposera la conception et l'implémentation d'un prototype pour le sous-système de gestion des données génomiques pour le laboratoire musculo-squelettique du CHU Sainte-Justine. Ce rapport présente la synthèse des travaux effectués du projet de neuf crédits prescrit au programme de maîtrise en génie logiciel. Le but de ce rapport est de recueillir, analyser et documenter les besoins fonctionnels en génomique du laboratoire.

Ce document est divisé en trois chapitres. Le chapitre 1 présente la problématique, les objectifs du projet et la revue bibliographique. Le chapitre 2 présente l'analyse des besoins, les spécifications des exigences ainsi que des prototypes d'interfaces graphiques pour la future l'application web. Finalement, le chapitre 3 présente les conclusions et travaux futurs.

CHAPITRE 1

PROBLÉMATIQUE ET REVUE LITTÉRAIRE

1.1 Identification de la problématique

Le laboratoire de recherche en pathologies musculosquelettique de l'hôpital Sainte-Justine effectue de la recherche sur la scoliose qui affecte les enfants (c.-à-d. les malformations congénitales, les tumeurs des os et du cartilage), les adolescents (c.-à-d. la scoliose idiopathique) et chez les adultes (c.-à-d. l'ostéoarthritis et la dégénérescence du cartilage articulaire).

Un problème important, pour les chercheurs du laboratoire, est la difficulté à gérer, diffuser et accéder à l'information génétique. Cette information provient des patients qui souffrent de la maladie de la scoliose et qui participent aux recherches du laboratoire. Ces données sont anonymisées et actuellement accessibles aux chercheurs dans différents chiffriers Excel. Actuellement, le généticien du laboratoire travaille à l'aide de fichiers Excel pour faire ses recherches de variations génétiques. Les données transformées par le généticien sont alors fournies aux autres chercheurs pour faire des analyses spécifiques. Pour obtenir des données, les chercheurs doivent faire une demande au généticien. Ce processus prend un certain temps et par la suite chaque chercheur sauvegarde ses fichiers localement sur son poste de travail et les utilise pour faire des recherches.

Ce projet de recherche en génie logiciel vise la conception, la création et l'implémentation d'une base de données centralisée qui permettra le stockage et la consultation des données génétiques d'une manière pratique, efficace et fiable. Actuellement, les données génétiques proviennent de fichiers VCF. Ces fichiers VCF (voir annexe I) contiennent l'information des variations génétiques des patients qui souffrent de scoliose ainsi que celle des patients qui n'ont pas la maladie. En plus, le projet vise à développer des interfaces WEB qui ont pour objectifs de faciliter l'accès à l'information stockée dans cette base de données. Les chercheurs pourront effectuer leurs requêtes, à l'aide de ce logiciel, plus efficacement. Les étapes pour l'obtention du fichier VCF sont les suivantes :

- 1) La première étape est l'obtention d'un échantillon prélevé sur un patient clinique par une infirmière (interne) ou d'un échantillon externe (ADN ou sang).
- 2) Ensuite, le séquençage est fait dans le laboratoire d'hôpital Sainte-Justine. Le temps de séquençage prend environ 2 semaines avec la technique Solid ou Illumina. Il génère les fichiers binaires « xsq » ou « fast » (la taille est environ 12.5 GB). Ces fichiers sont envoyés à Génome Québec.
- 3) À génome Québec, le bio-informaticien applique le processus ETL (« File Format Transformation ») aux fichiers. Le processus génère des fichiers « BAM » et finalement, on obtient le fichier « VCF ».

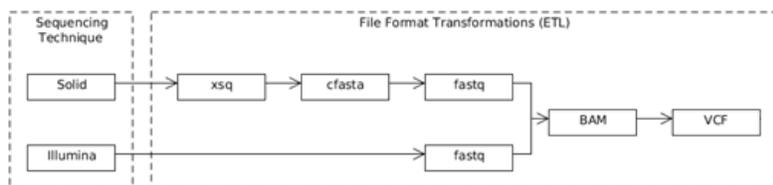


Figure 1. Processus d'obtention d'un fichier VCF

1.2 État de l'art

Cette partie du rapport présente une revue littéraire partielle. Dans un premier temps, on y décrit les concepts de haut niveau de développement du logiciel et ensuite on y présente une revue critique de différentes plateformes bio-informatiques existantes qui sont une source d'inspiration pour nos travaux.

1.2.1 La scoliose

La scoliose (scolios: qui dit tortueux en grec) « est une déformation tridimensionnelle (dans les trois plans de l'espace) de toute ou d'une partie de la colonne vertébrale (c.-à-d. cervicale, thoracique ou lombaire) entraînant une torsion d'une ou de plusieurs vertèbres sur elle(s)-

même(s) et provoquant une déformation du thorax, de l'abdomen et des zones para vertébrales (proches des vertèbres) » (Bruandet). « Contrairement à la scoliose « classique » dont les causes sont inconnues, la forme congénitale est liée à un développement anormal des vertèbres au stade embryonnaire. En France, un bébé sur 1 000 naît avec cette malformation caractérisée par une courbure pathologique de la colonne vertébrale. Pilotée par Sally Dunwoodie, de la faculté de médecine de Sydney, l'étude menée par cette équipe suggère que, pour se développer, cette pathologie nécessiterait la combinaison d'une mutation génétique, et d'une période de manque d'oxygène durant la grossesse » (Braly 2012).

1.2.2 La génomique

« La génomique vise l'étude de l'ensemble des gènes d'une personne (son génome), y compris les interactions entre les gènes et avec l'environnement. La génomique comprend l'étude scientifique des maladies complexes telles que les maladies cardiaques, l'asthme, le diabète et les maladies cancéreuses parce que ces maladies sont causées généralement par une combinaison de facteurs génétiques et environnementaux que par une seule mutation génétique. La génomique moderne offre de nouvelles possibilités pour les thérapies et les traitements de certaines maladies complexes, ainsi que de nouvelles méthodes de diagnostic (NCHGR 2014) ». Le génome est formé à partir des séquences de lettres « le gène » et « le chromosome », est ainsi l'ensemble de l'information héréditaire présente dans chaque cellule de chaque organisme vivant (et nécessaire au développement et au fonctionnement de l'organisme) y est décrit. « C'est un peu comme le mode d'emploi du fonctionnement des cellules » (GenomeQuébec 2014). Dû à la grande quantité de données qu'un génome génère . La taille d'un fichier de séquence est d'environ entre 12 à 100 giga-octets par personne dépendant des informations moléculaires (GenomeQuébec 2014), il est utile d'utiliser des outils informatiques pour le traitement des données et c'est ici que la Bio-informatique intervient.

1.2.3 La Bio-informatique

Selon le « National Institute of Mental Health », la bio-informatique est « la recherche, le développement et l'application d'outils et approches informatiques pour la manipulation des données biologiques, de la santé médicale, comportementale y compris celles nécessaires pour obtenir, stocker, analyser et afficher ces informations. » (Huerta, Haseltine et al. 2000)

La bio-informatique a pris un grand essor suite au projet de génome humain (Collins, Morgan et al. 2003) qui a nécessité le développement des premiers logiciels informatiques. Gilber définit la bio-informatique de la manière suivante, « c'est une « inter discipline » à la frontière de la biologie, de l'informatique et des mathématiques. Elle a pour but d'intégrer des données d'origines très diverses pour modéliser les systèmes vivants afin de comprendre et prédire leurs comportements. Ainsi, par exemple, le séquençage à très haut débit offre la possibilité de connaître de manière personnalisée le génome de chacun » (Deléage and Gouy 2013). « De cette manière, il est possible de connaître les différences entre individus, car dans deux génomes humains tirés au hasard, 99,9% de la séquence d'ADN est identique. Les 0,1% restants contiennent des variations de séquence dont le type le plus commun est le polymorphisme pour un nucléotide (SNP). Ces variations sont associées à de la diversité entre populations ou individus, une différence de sensibilité à des maladies et la réponse individuelle aux médicaments » (Masselot 2014).

1.2.4 Plateformes bio-informatiques

Les plateformes bio-informatiques offrent un grand appui aux chercheurs en génomique lesquelles facilitent la réalisation de leurs projets. Aujourd'hui ils existent différents plateformes bio-informatiques.

Cette section présente six plateformes : Gemini, Tute, Annovar, VAAST2, Galaxy et GeneSys.

1.2.4.1 La plateforme GEMINI

« GEMINI, une plateforme qui permet de trouver les variations génétiques en utilisant comme source d'entrée les fichiers VCF(GEMINI). GEMINI (GENome MINIng) est conçu pour être un "framework" flexible pour explorer la variation génétique pour le génome humain. En mettant les variantes génétiques, les génotypes de l'échantillon, et les annotations du génome dedans la base de données. GEMINI fourni un système simple, flexible, mais très puissant pour explorer la variation génétique pour les maladies et pour la génétique de la population. Le « framework » GEMINI charge le fichier VCF dedans la base des données. GEMINI dispose d'un outil spécifique de type ligne de commande pour interroger la base de données. L'outil est appelé "gemini query". À continuation, on montre trois exemples des requêtes sur la base de données GEMINI pour donner une idée de la façon d'interagir avec la base de données gemini en utilisant l'outil "gemini query". »

- a. Extraire toutes les transitions avec un « call rate » > 95%

```
$ gemini query -q "select * from variants \
                    where sub_type = 'ts' \
                    and call_rate >= 0.95" my.db
```

- b. Extraire tous les « loss-of-function variants» avec un “ alternate allele frequency” < 1%:(GEMINI)

```
$ gemini query -q "select * from variants \
                    where is_lof = 1 \
                    and aaf >= 0.01" my.db
```

3. Extrait de la diversité nucléotidique pour chaque variante:

```
$ gemit query -q "select chrom, start, end, pi from variants" my.db
```

1.2.4.2 La plateforme TUTE

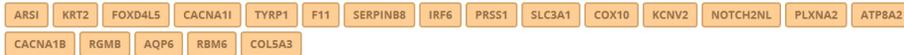
Une des contributions la plus importante dans ce domaine, c'est la plateforme Tute génomique (Genomics 2014). « Tute est un outil d'interprétation du génome qui permet aux chercheurs d'utiliser les données du génome humain pour la recherche scientifique et pour un traitement individualisé. La plateforme aide aux chercheurs à l'interprétation des variantes génétiques et à trouver les gènes liés à la maladie. La figure 1 montre la plateforme, chaque ligne est un échantillon des données génétiques chargé à partir d'un fichier VCF. »

Name	Version	Description	Family ID	Subject ID	Uploaded	Size	# Variants	
Ogden syndrome	v1.2	Demo data - Ogden syndrome	Ogden	ogden1	N/A	324.8 KB	1,479	View Annotation
ihafamily1	v1.2	Demo data - ihafamily1 affected child, hemolytic anemia	ihafamily	ihafamily1	N/A	8.98 MB	36,973	View Annotation
ihafamily2	v1.2	Demo data - ihafamily2 unaffected child	ihafamily	ihafamily2	N/A	9.25 MB	37,794	View Annotation
ihafamily3	v1.2	Demo data - ihafamily3 unaffected mother	ihafamily	ihafamily3	N/A	8.41 MB	34,577	View Annotation
ihafamily4	v1.2	Demo data - ihafamily4 unaffected father	ihafamily	ihafamily4	N/A	8.9 MB	36,761	View Annotation
ATCC HCC1187 TumorC	v1.2	Demo data - Control cancer data from illumina	0	ATCC_HCC1187_TumorC	N/A	1.24 GB	3,128,982	View Annotation
ATCC HCC1187 TumorBL	v1.2	Demo data - Case cancer data from illumina	0	ATCC_HCC1187_TumorBL	N/A	1.52 GB	3,838,392	View Annotation

Figure 2. Plateforme Tute GENOMICS (Genomics 2014)

Dans la figure 1, on choisit le fichier nommé «ihafamily4», le système de la plateforme va obtenir les résultats automatiquement les gènes qui contient variant pouvant causer la maladie. Voir la figure 2.

These genes were detected in this sample containing variants that may cause disease:



Quality Control Statistics

Variants Statistics

# Variants	Het/Hom Ratio	Ti/Tv Ratio	% dbSNP
36973	1.83	2.56	97.64%

Fraction	Het/Hom Ratio (Exonic)	Ti/Tv Ratio (Exonic)	%dbSNP (Exonic)
56.08	1.84	2.86	98.32%

Data Quality Flag

Flag	Concern
Ti/Tv ratio	Warning, Ti/Tv Ratio (2.56)
Large (>10Mb) LOH region	None (0)
Low Median Coverage	None (93.00X)
Excessive heterozygosity (sample mixup)	None (1.83)
%SNP with QUAL<=40	None (99.50%)

SNP Quality Statistics

QUAL value (from VCF)	Percentage
<20	0.00%
20-29	0.00%
30-39	0.50%
>=40	99.50%

Figure 3. Statistiques de contrôle de qualités des données génétiques (Genomics 2014)
La plateforme permet aux chercheurs d'appliquer des filtres sur les variations trouvées (voir figure 3).

Figure 4. Fenêtre de recherches des données génétiques (Genomics 2014)

À l'aide de la plateforme Tute, l'information est sauvegardée dans le nuage.

1.2.4.3 Le logiciel ANNOVAR

Le logiciel Annovar (Annovar) est « un logiciel utilisé afin d'interpréter et d'annoter les variantes génétiques détectées à partir de divers génomes (c.-à-d. humain, de la souris, du ver, de la mouche, de la levure et de bien d'autres organismes vivants). Il permet de dresser une liste de : variants, chromosome, position de départ, position de fin, nucléotidique de référence; et de nucléotides observés. L'utilisateur peut exécuter des scripts, en ligne de commande alors il doit maîtriser l'informatique » (Annovar).

1.2.4.4 Le logiciel VAAST2

« Le logiciel VAAST2 est un outil de recherche probabiliste qui permet d'identifier les gènes endommagés et leurs variants qui causent les maladies dans les séquences du génome d'un individu. Cet outil aide à identifier des variants rares qui pourraient être responsables de maladies génétiques rares, et il peut également utiliser des paires de variants rares et communes afin d'identifier les gènes responsables de maladies courantes. »(Yandell 2011)

1.2.4.5 La plateforme Galaxy

Galaxy est une plateforme WEB, « disponible en logiciel libre, qui permet d'effectuer la recherche intensive des données biomédicale (University 2014) (voir figure 4). Galaxy propose une "constellation" d'outils afin d'analyser, de manipuler et de visualiser des données génomiques, et ce sans avoir besoin de connaissance en programmation. Elle a été développée par le « Center for Comparative Genomics and Bioinformatics ». Galaxy est utilisée généralement pour la fouille de données. L'utilisateur peut réaliser quatre grands types d'opérations (Plume 2011): 1) la manipulation de fichiers; 2) des opération sur les données; l'analyse de séquences; et 4) la visualisation des données. »

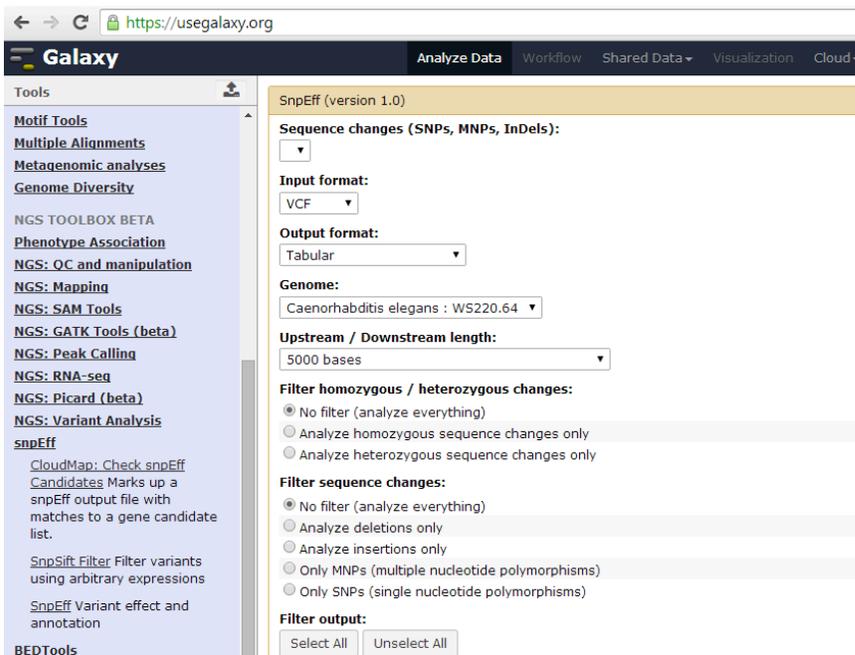


Figure 5. Plateforme Galaxy (Plume 2011)

1.2.4.6 La plateforme GeneSys

La plateforme bio-informatique GeneSys (Actasanitaria 2013) « permet une analyse intégrale des données qui provient des études de re-séquençage: à partir de l'évaluation des données brutes jusqu'à la génération d'une liste de variantes ou gènes priorités basés sur une maladie spécifique d'une façon rapide, automatisée et efficace. Ce système a été créé afin de permettre l'interprétation et l'exploitation efficace des données bio-informatiques pour l'appui au diagnostic médical et à la recherche biomédicale. Par rapport à la navigation, la plate-forme web spécialisée offre une interface conviviale pour accéder de façon simple et intuitive aux données génétiques. Quelques services inclus dans cette plateforme sont: 1) le filtrage avancé pour rechercher des résultats concluants; 2) la navigation graphique génomique pour une meilleure évaluation de l'implication d'une variation dans la maladie; et

3) la disponibilité des résultats dans le « nuage » qui peuvent être téléchargés à tout moment (offrant une grande accessibilité et disponibilité) ». Finalement, cette plateforme fournit un service de stockage et de conservation des données biologiques, ainsi que des conseils personnels lors de l'obtention et de l'interprétation des résultats (Genómicos 2013).

Conclusion

Bien qu'il existe des plateformes bio-informatiques sur le marché. Il est préférable de développer un nouveau logiciel qui s'adapte aux besoins particuliers du laboratoire. Par exemple, c'est vrai que l'outil "gemini query" est très puissant pour faire des requêtes sur les données génétiques. Mais, il faut avoir une connaissance des sentences SQL. Un chercheur n'est pas familiarisé avec ce type de sentences. Donc, il faut avoir un système plus intuitif pour interroger une base de données avec information génétique. Avec l'outil Tute, l'information est sauvegardée dans le nuage, on préfère la sauvegarder sur les serveurs de l'hôpital Sainte-Justine pour des raisons de confidentialité.

Dans le chapitre suivant, on présente l'analyse des besoins et des spécifications du logiciel

CHAPITRE 2

ANALYSE DES BESOINS ET DES SPÉCIFICATIONS DU LOGICIEL

2.1 Introduction

Ce chapitre présente un sommaire du processus de développement logiciel, des besoins et des exigences fonctionnelles du futur système.

2.2 Processus de développement

Le processus de développement utilisé pour ce projet est une approche Agile. « Les méthodes agiles reposent sur une structure (cycle de développement) commune (itérative, incrémentale et adaptative) et elles visent la satisfaction réelle du client en priorité aux termes d'un contrat de développement » (Wikipedia 2014). La démarche d'identification des besoins pour les données de génomique a été réalisée en suivant les étapes suivantes :

- Analyse des besoins et des spécifications des exigences en génomique;
- Conception des interfaces utilisateur du système;
- Revue itérative du travail réalisé avec le spécialiste.

La première phase de cette démarche vise l'identification des besoins en génomique. Pour réaliser cette activité des réunions ont eu lieu avec le généticien afin de bien comprendre le processus actuel ainsi que ses besoins. Le généticien a décrit les chiffriers Excel contenant les données génomiques. La figure 5 présente une synthèse de ce fichier. Cette démarche a été faite lors de la rédaction du document de vision (voir annexe 2).

La deuxième phase visait la conception d'un premier prototype de base de données (voir section 2.9) et d'interfaces. Suite à l'inventaire, l'analyse, et la clarification des besoins et des exigences, les interfaces graphiques et la base de données a été conçue. Premièrement, les interfaces ont été dessinées en utilisant un outil « Mockup » (On a utilisé la version

« trial »), C’est un outil web nommé « mybalsamiq »(Guilizzoni 2014)et ensuite un prototype de la base de données et des interfaces ont été développées en utilisant la technologie de Microsoft ASP .NET et la base de données SQL Server.

Finalement, lors de la dernière phase, un bilan du premier prototype du sous-système génomique a été fait. Principalement, ce bilan consiste à lister les problèmes rencontrés et de lister les améliorations futures. Cette partie est présentée dans le chapitre 3 de ce rapport.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R			
1	chrom	start	end	ref	alt	gene	qual	Total_samples	%FG1	Sample_id	Total_sant	%FG2	Sample_id	Total_sant	%FG3	Sample_id	Total_sample	%C		
2	chr1	908928	908929	C	T	PLEKHN1	129	0	0	0	0	0	0	0	0	0	0	1	0.	
3	chr1	1325719	1325720	T	A	CCNL2	64.5	0	0	0	0	0	0	0	1	0.04347826	705	0	0	
4	chr1	1341205	1341206	T	C	MRPL20	133	1	0.0454545	400	0	0	0	0	0	0	0	0	0	
5	chr1	1500275	1500276	C	T	SSU172	116	0	0	0	0	0	0	0	0	0	0	0	1	0.
6	chr1	1721898	1721899	C	T	GNB1	999	0	0	0	0	0	0	0	0	0	0	0	1	0.
7	chr1	1848196	1848197	G	C	CALML6	999	0	0	0	0	0	0	0	0	0	0	0	1	0.
8	chr1	1900155	1900156	C	T	C1orf222	116	0	0	0	0	0	0	0	0	0	0	0	1	0.
9	chr1	1900210	1900211	C	T	C1orf222	999	0	0	0	0	0	0	0	0	0	0	0	1	0.
10	chr1	2129457	2129458	G	T	C1orf86	999	0	0	0	0	0	0	1	0.04347826	356	3	0	0	
11	chr1	2129497	2129498	G	A	C1orf86	999	0	0	0	0	0	0	0	0	0	0	0	1	0.
12	chr1	2234790	2234791	C	T	SKI	145	0	0	0	0	0	0	0	0	0	0	0	1	0.
13	chr1	2290142	2290143	G	A	MORN1	999	0	0	0	1	0.04545455	1600	0	0	0	0	0	1	0.
14	chr1	2422747	2422748	T	A	PLCH2	71.5	0	0	0	0	0	0	0	0	0	0	0	1	0.
15	chr1	2541236	2541237	G	A	MMEL1	154	0	0	0	0	0	0	0	0	0	0	0	1	0.
16	chr1	2938415	2938416	T	C	ACTRT2	999	0	0	0	0	0	0	0	0	0	0	0	1	0.
17	chr1	3424378	3424379	C	T	MEGF6	106	0	0	0	1	0.04545455	1606	0	0	0	0	0	1	0.
18	chr1	3548892	3548893	G	A	WRAP73	104	0	0	0	0	0	0	0	0	0	0	0	1	0.
19	chr1	3670711	3670712	G	A	CCDC27	136	0	0	0	0	0	0	0	0	0	0	0	1	0.
20	chr1	3680332	3680333	C	T	CCDC27	111	1	0.0454545	353	0	0	0	0	0	0	0	0	1	0.
21	chr1	3680344	3680345	C	T	CCDC27	999	1	0.0454545	1650	1	0.04545455	1676	0	0	0	0	0	1	0.

Figure 6. Fichier Excel avec de données génétiques

2.3 Identification des intervenants et des utilisateurs

On présente, à l’annexe 3 les rôles, et les personnes qui interviennent dans le projet, ainsi que la description des responsabilités des utilisateurs du sous-système.

2.4 Analyse des besoins

Cette section décrit les besoins en données génomiques des employés du laboratoire. Les besoins ont servis afin de concevoir le premier prototype d’application et pourront aussi servir pour faire les tests de vérification et de validation dans le futur. Chaque description d’un besoin possède une priorité. Trois niveaux de priorités ont été établis (voir tableau 3).

Critique	Ce besoin est indispensable pour l'implémentation de l'application.
Important	Ce besoin est nécessaire, mais l'absence de ce besoin ne généra pas des risques sur la mise en œuvre d'autres besoins.
Utile	L'absence de ce besoin ne mettra en risque le fonctionnement du système.

Tableau 1, Niveaux de priorité

Alain April 14-8-27 22:38

Deleted: 3

2.4.1 **B11 - Gérer les interfaces graphiques.**

Description du besoin

Il est nécessaire d'avoir une interface graphique pour gérer les données génétiques.

Priorité

Critique

Préoccupations

Accéder facilement aux données génétiques.

Situation actuelle

L'utilisateur travaille actuellement directement dans son fichier Excel ce qui n'est pas optimal.

Solution proposée

L'utilisateur devrait disposer d'interfaces graphiques faciles d'utilisation et qui fonctionnent rapidement pour accéder à la base des données centralisée.

2.4.2 **B12 - Rendre les données génétiques accessibles**

Description du besoin

Il est nécessaire que les chercheurs disposent d'un système de filtrage des données génétiques. Les données seront disponibles aux chercheurs sans qu'ils aient trop de manipulations à effectuer.

Priorité

Critique

Préoccupations

Les informations sur les mutations génétiques doivent être nettoyées et uniformisées.

Solution actuelle

Le généticien et le bio-informaticien manipulent et filtrent les données selon le besoin, et envoient ces données aux utilisateurs dans des feuilles Excel.

Solution proposée

Le format des données filtrées sera standardisé, et rendu directement accessible aux utilisateurs (sans nécessiter de manipulation supplémentaire au généticien et au bio-informaticien).

2.4.3 B13 – Normalisation / Re-transformation des données.**Description du besoin**

Il est important que l'utilisateur dispose d'une interface graphique qui permet de stocker les fichiers bruts de données génétiques dans le système.

Priorité

Important

Préoccupations

Pouvoir permettre un retraitement (ou normalisation) si nécessaire.

Solution actuelle

Il n'existe pas.

Solution proposée

Une interface utilisateur permettra l'accès aux fichiers bruts dans le système.

2.5 Cas d'utilisation

Le diagramme de cas d'utilisation décrit le fonctionnement du sous-système génétique de la nouvelle plateforme à développer. Le diagramme de synthèse de tous les cas d'utilisation est présenté à la figure 6 de la prochaine page. Il comporte onze cas d'utilisation. Ce diagramme a permis d'identifier quatre types d'utilisateur qui interagiront avec le nouveau sous-système: 1) un super utilisateur; 2) un généticien; 3) des chercheurs et; 4) un auditeur.

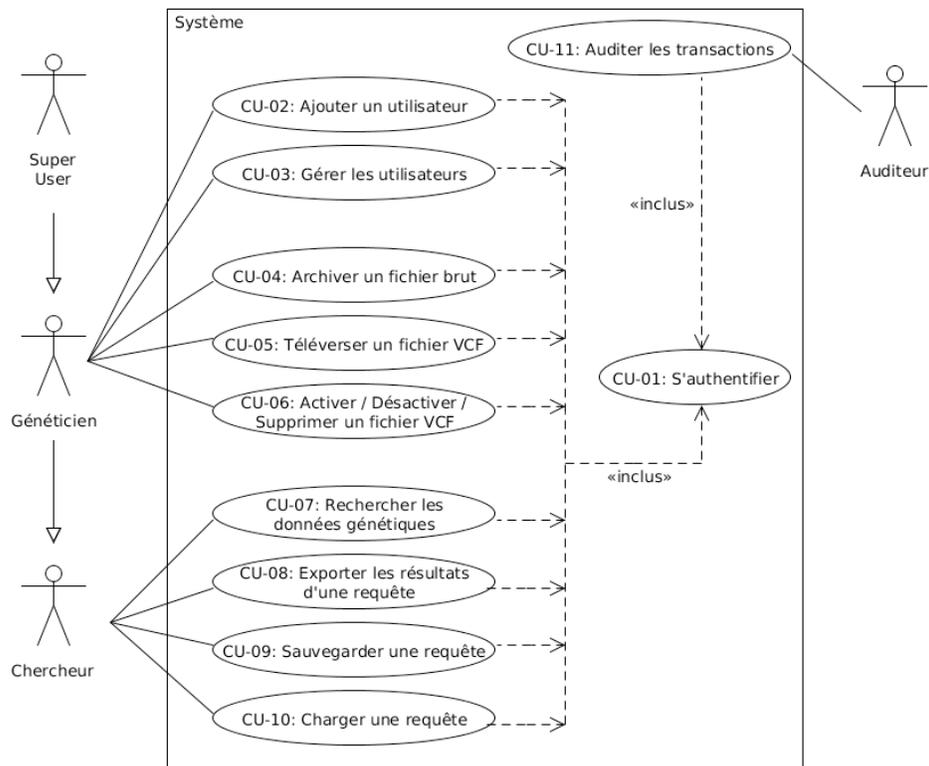


Figure 7. Diagramme de cas d'utilisation du module génétique.

2.5.1 CU-01: S'authentifier

Permet à un utilisateur d'accéder aux autres cas d'utilisation du système. L'utilisateur doit avoir un compte et un mot de passe pour accéder au système

2.5.2 CU-02: Ajouter un utilisateur

Permet d'ajouter un nouvel utilisateur ayant accès au sous-système et de lui attribuer un rôle (généticien, chercheur, auditeur, etc..).

2.5.3 CU-03: Gérer les utilisateurs

Permet de modifier le rôle et les permissions d'un utilisateur existant. Seul le super utilisateur joue le rôle d'administrateur du sous-système.

2.5.4 CU-04: Archiver un fichier brut

Permet au généticien de conserver les fichiers bruts de séquençage, et ainsi conserver une traçabilité.

2.5.5 CU-05: Téléverser un fichier VCF

Permet au généticien de téléverser un nouveau fichier VCF dans la base de données. L'état du nouveau fichier sera inactif par défaut pour lui permettre de valider les données contenues dans le VCF avant de rendre les données accessibles aux chercheurs.

2.5.6 **CU-06: Activer / Désactiver / Supprimer un fichier VCF**

Permet de changer l'état d'un fichier VCF dans le sous-système. Avant de rendre un fichier VCF disponible aux chercheurs, il doit être activé. Lorsqu'un fichier VCF est désactivé, il demeure disponible aux chercheurs l'ayant déjà accédé préalablement; mais il n'est plus accessible pour les autres chercheurs. Lorsqu'un fichier VCF est supprimé, il est éliminé de la base de données.

2.5.7 **CU-07: Rechercher les données génétiques**

Permet aux chercheurs d'effectuer des requêtes sur la base de données génétique.

2.5.8 **CU-08: Exporter les résultats d'une requête**

Permet aux chercheurs d'exporter les résultats d'une recherche vers Microsoft Excel (format CSV).

2.5.9 **CU-09: Sauvegarder une requête**

Permet aux chercheurs de sauvegarder, dans le système, une requête existante effectuée sur un fichier de format VCF.

2.5.10 **CU-10: Charger une requête**

Permet aux chercheurs de retrouver une requête préalablement sauvegardée et de la charger dans le système pour la consulter, ou l'exécuter à nouveau.

2.5.11 **CU-11: Auditer les transactions**

Permet à l'auditeur de consulter toutes les transactions effectuées sur le sous-système.

2.6 Les acteurs

Cette section décrit les acteurs interagissant avec le sous-système de gestion des données génétiques.

2.6.1 Chercheur

Utilise le sous-système pour chercher efficacement et extraire des informations sur la base de données génétiques. C'est le principal utilisateur du système.

2.6.2 Généticien

Responsable de la mise à jour des données génétiques. Il peut réaliser tous les cas d'utilisation des chercheurs, en plus des siens.

2.6.3 Super utilisateur

Le super utilisateur est l'administrateur du sous-système. Il peut réaliser tous les cas d'utilisation du généticien et des chercheurs, en plus des siens. Par contre, il ne peut pas modifier les transactions déjà enregistrées du système, ni effectuer les audits.

2.6.4 Auditeur

Responsable de la vérification des transactions du sous-système et de s'assurer que la norme CFR 21 partie 11 est respectée. Son rôle est limité au CU-11. L'auditeur vérifie seulement les fichiers « logs ». Il n'est pas accès aux autres fonctionnalités du système.

2.7 Analyse des exigences

2.7.1 Exigences non fonctionnelles

Cette prochaine section présente les exigences non-fonctionnelles qui doivent être pris en compte pour développer le l'application web. Une exigence non-fonctionnelle est une qualité que le logiciel devrait atteindre telles que le temps d'exécution, la fiabilité, la sécurité(ISO/IEC 2000)

2.7.1.1 Exigence non-fonctionnelle : " facilité d'utilisation "

L'interface de l'application devra être conviviale, facilement compréhensible et qui fonctionnera correctement afin de permettre aux utilisateurs d'accéder facilement et rapidement aux données génétiques de manière intuitive.

2.7.1.2 Exigence non-fonctionnelle : " temps de réponse pour une recherche des données "

Le nouveau sous-système devra être capable de répondre aux requêtes des chercheurs dans un temps raisonnable (en moins d'un seconde dans la plupart des cas).

2.7.1.3 Exigence non-fonctionnelle : « Sécurité »

Le système devra comporter des mesures de protection des données génétiques.

2.8 Description des interfaces graphiques

Cette section décrit quelques maquettes des interfaces utilisateur à titre de recommandation, pour le premier prototype. La forme et l'aspect de l'information montrée dans chaque interface n'est pas irrévocable et sera raffinée suite aux commentaires du généticien. Cependant, le contenu de chaque interface doit être présent dans la version finale du logiciel.

L'interface graphique, voir figure 7, contiendra cinq onglets : 1) l'accueil pour afficher le message d'introduction lors du lancement du sous-système; 2) utilisateur afin de gérer les accès au sous-système; 3) fichiers VCF afin de gérer les fichiers VCFs; 4) fichiers bruts afin de gérer les fichiers bruts génomiques; et finalement 5) Recherches pour effectuer des recherches sur les données génomiques.

2.8.1 Interface « gestion de fichiers VCF »

Cas d'utilisation associé(s) : CU-05, CU-06

La figure 7 présente l'interface principale. On y présente trois boutons: 1) le bouton « Téléverse » qui permet de charger un fichier VCF dans le sous-système; 2) le bouton « Afficher » qui permet de visualiser les données contenues dans le fichier VCF; et 3) le bouton « Supprimer » qui permet d'éliminer les données d'un fichier VCF. Ainsi qu'un tableau composé de trois colonnes : 1) « activer/désactiver » permet de gérer l'accès aux données du fichier VCF (quand la case est sélectionnée cela signifie que le fichier est disponible pour consultation); 2) le nom du fichier VCF; et 3) l'information concernant la taille du fichier VCF.



Figure 8. Interface de gestion des fichiers VCF.

2.8.2 Interface « Recherches des données génétiques »

Cas d'utilisation associé(s) : CU-07, CU-08, CU-09, CU-10

Cette interface, voir figure 8, permet aux chercheurs d'effectuer des requêtes sur les données génétiques. L'interface contient quatre boutons: 1) « Charger requête » qui permet charger un fichier contenant une requête pour faire des recherches; 2) « Sauvegarder req » qui permet la sauvegarde d'une requête pour l'utiliser ultérieurement; 3) « Afficher VCF » qui affiche le contenu du fichier VCF à l'écran. L'interface de la figure 8 présente aussi un tableau qui liste les tables du sous-système. La première colonne est un « checkBox » qui permet de sélectionner la table. La deuxième colonne affiche le nom de la table de la base de données. La recherche de données génétiques fonctionne de la manière suivante: L'utilisateur choisit les tables dans le tableau, puis il choisit le bouton « afficher tables sélectionnées » lequel affiche les tables sélectionnées avec leurs attributs. Dans cet exemple, il y a deux champs sélectionnés: 1) variation_id et 2) sample_id dans la table sample_variation.

Ensuite, l'utilisateur clique sur le bouton « obtenir colonnes » afin d'afficher les colonnes sélectionnées dans un deuxième tableau. Ce tableau est formé pour un "checkbox", le nom de la table de la colonne et une colonne "Filtrer" pour saisir une valeur. En plus, le sous-système affiche l'énoncé SQL généré pour la requête, laquelle sera exécutée avec l'action du bouton "Executer la requête". Le résultat de la requête sera affiché à l'écran.

localhost:2410/WebSite1/ x
localhost:2410/WebSite1/Default2.aspx

MODULE GÉNÉTIQUE

Accueil Utilisateur Fichiers VCF Fichiers Bruts Recherches

Recherche données génétiques

Charger requête Sauvegarder req Afficher VCF

<input type="checkbox"/>	Name
<input checked="" type="checkbox"/>	rs_numbers
<input checked="" type="checkbox"/>	sample variations
<input checked="" type="checkbox"/>	samples
<input type="checkbox"/>	biochemical_groupe
<input checked="" type="checkbox"/>	GenTable
<input type="checkbox"/>	variations

Obtenir Tables sélectionnées

rs_numbers

Nom
<input type="checkbox"/> rs_id

sample_variations

Nom
<input checked="" type="checkbox"/> variation_id
<input checked="" type="checkbox"/> sample_id
<input type="checkbox"/> genotype
<input type="checkbox"/> genotype_cual

samples

Nom
<input type="checkbox"/> sample_id

GenTable

Nom
<input type="checkbox"/> Codego
<input type="checkbox"/> Nom
<input type="checkbox"/> Description

Obtenir colonnes

	Colonne	Table	Filtrer
<input type="checkbox"/>	variation_id	sample_variations	<input type="text"/>
<input type="checkbox"/>	sample_id	sample_variations	<input type="text"/>

Select sample_variations.variation_id,sample_variations.sample_id From sample_variations

Executer la requete

Figure 9. Interface Recherche des données génétiques.

2.8.3 Interface « Fichiers bruts »

Cas d'utilisation associé(s) : CU-04

Cette interface, voir figure 9, présente les fonctionnalités suivantes : le bouton « Téléverser » charge un fichier brut et le bouton « Supprimer » élimine les données d'un fichier. Il comporte aussi un tableau de deux colonnes : 1) la première décrit le nom du fichier VCF; et 2) la deuxième présente la taille du fichier.



MODULE GÉNÉTIQUE - CHU SAINTE-JUSTINE

Accueil Utilisateur Fichiers VCF Fichiers Bruts Recherches

FICHIERS BRUTS

Téléverser Supprimer

<input type="checkbox"/>	Nom	Taille
<input type="checkbox"/>	Fichier1	14 gb
<input type="checkbox"/>	Fichier2	15 gb
<input type="checkbox"/>	Fichier3	33 gb

Figure 10. Interface pour gérer fichiers bruts

2.9 Base de données

La base de données a été conçue en suivant un modèle de base de données relationnelle comme il est illustré à la figure 10. La table Variations contient les champs de base pour une variation génétique telles que le chrom (chromosome), start (la position de départ) end (la position de fin), ref (référence allèle), alt (Allèle suppléant pour la variante), qual (le score de qualité l'ALT). La table sample_variations contient l'information le type de genotype (Par exemple :0,1,2) et la qualité de genotype de chaque échantillon. L'information sur l'échantillon est dans la table « sample ». L'information relative au type de groupe des patients est dans la table biochemical_groups. La table rs_numbers contient le code anonymisé de chaque patient. Finalement, la table GenTable contient la liste des gènes.

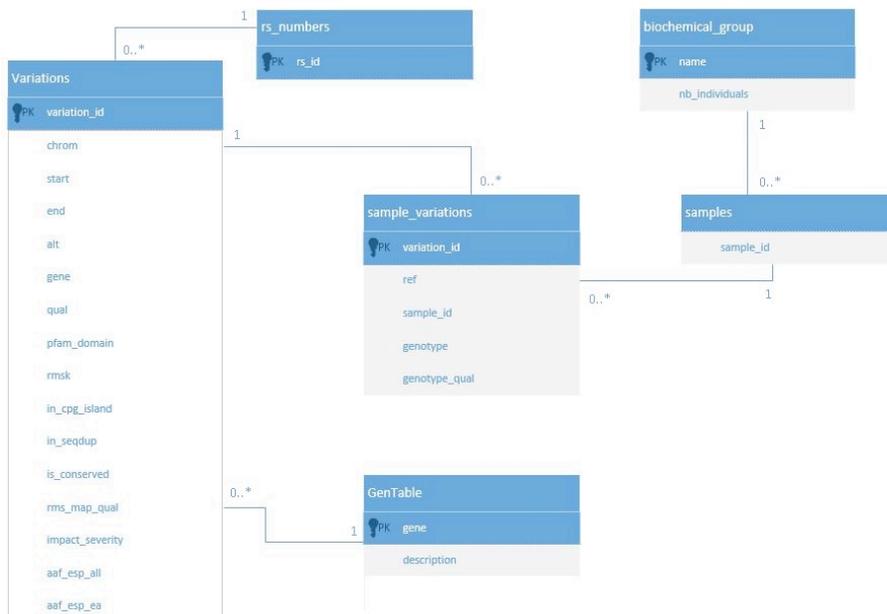
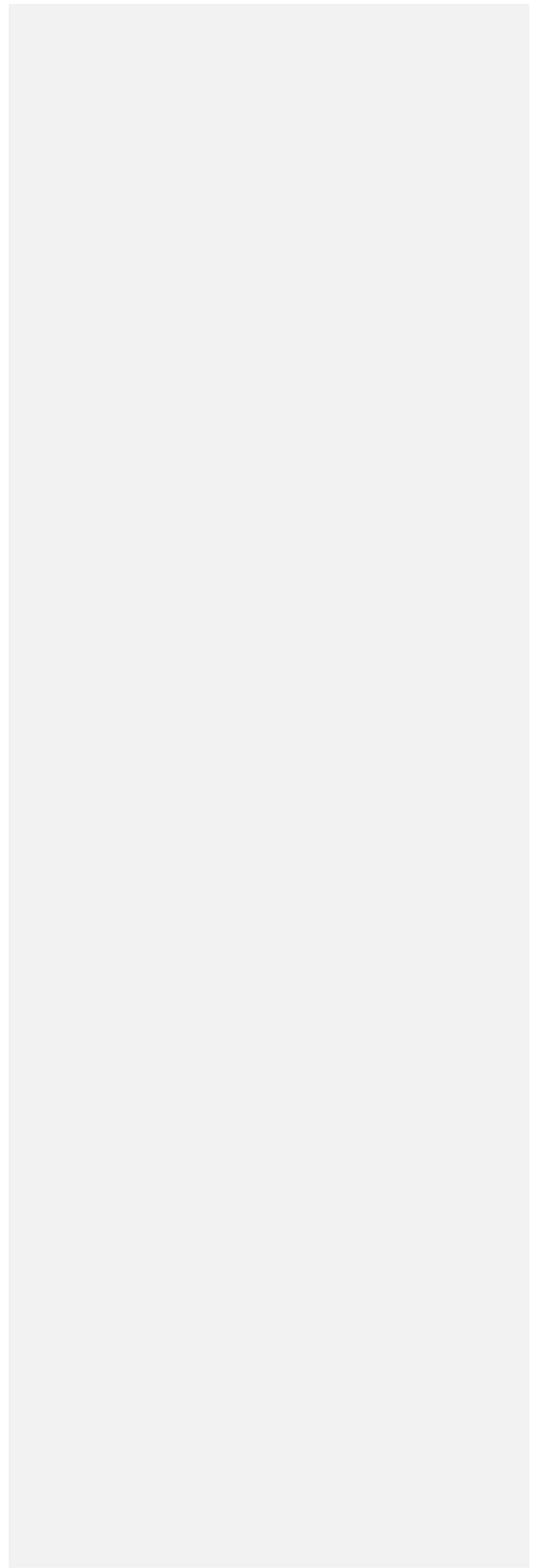


Figure 11. Modèle Entité-Relation de la base de données sous-système génétique

Conclusion

Les cas d'utilisation 4 à 10 ont été spécifiés mais il reste les cas 1,2,3 et 11 à effectuer dans une autre étude.



CHAPITRE 3

Discussion

3.1 Introduction

Cette partie présente les observations sur le projet et ce qui reste à faire pour améliorer ce prototype lors d'une prochaine itération.

3.2 Apprentissage

Au début du projet, j'avais une méconnaissance du domaine de la bioinformatique. Pour y remédier, j'ai fait une revue bibliographique de ce qui se fait dans ce domaine. Pendant ce projet, le généticien a été très collaboratif en donnant l'information nécessaire. Il y a eu plusieurs réunions, avec les parties prenantes, afin de bien identifier les besoins du groupe génomique. C'est à ce moment que j'ai appliqué mes connaissances du génie logiciel et conçu des interfaces utilisateur. Dans le projet, j'ai appris que la partie technologique est importante, mais qu'il y a aussi d'autres aspects à maîtriser telles que la satisfaction du client et la bonne relation de cordialité et de communication avec lui.

3.3 Validation des interfaces

Depuis le début du projet, on a fait des rencontres avec les parties prenantes et générés des documents. Les documents générés (c.-à-d. plan du projet, document vision, etc.) ont été validés par le généticien. Aussi, il a démontré sa satisfaction et son accord concernant la conception et le contenu des maquettes et des interfaces utilisateur de l'application.

3.4 Travail futur

Il y a toujours possibilité de faire des améliorations à ce premier prototype:

3.4.1 Statistiques

Le système devrait fournir des statistiques sur les données génétiques de façon automatique. En plus, l'utilisateur doit pouvoir définir des statistiques personnalisées.

3.4.2 Internationalisation

Afin que le système soit adapté aux préférences d'utilisateur. L'application devrait pouvoir afficher les interfaces en anglais.

3.4.3 Visualisation des graphiques des fichiers bruts

Étant donné que l'application permet de sauvegarder ce type de fichier. Le système devrait avoir l'option d'afficher des graphiques sur l'écran des données contenu dans les fichiers bruts.

3.4.4 Requêtes

Actuellement, l'application permet qu'un utilisateur exécute des requêtes complexes sur les données génétiques en générant des résultats que l'utilisateur puisse sauvegarder. Dans le futur, le système devrait permettre de faire des requêtes sur ces résultats.

CONCLUSION

La scoliose est une maladie qui affecte de jeunes enfants. Pour cette raison, il est souhaitable de développer une solution informatique qui appuie les chercheurs afin d'accélérer la découverte de traitement et médicaments. Ce projet, en génie logiciel, a étudié les processus et outils logiciels existants et les besoins futurs du laboratoire. Ainsi on a spécifié les besoins fonctionnels et non fonctionnels d'une application proposée (c.-à-d. un module génomique). Cette application devait avoir des interfaces conviviales qui accèdent à une base de données centralisée. De cette manière, les chercheurs auront un accès plus efficace aux données génétiques des patients. En plus, ils n'auront plus besoin de se préoccuper de la sauvegarde et de la gestion de ces données sur Excel.

ANNEXE I

Fichier VCF ("Variant Call Format")

« Le fichier VCF est un fichier de format texte. Il contient des lignes de méta-information, une ligne d'en-tête, et les lignes des données qui contiennent l'information sur une position dans le génome. »(Marcketta 2013). La figure 12 présente une partie de ce fichier.

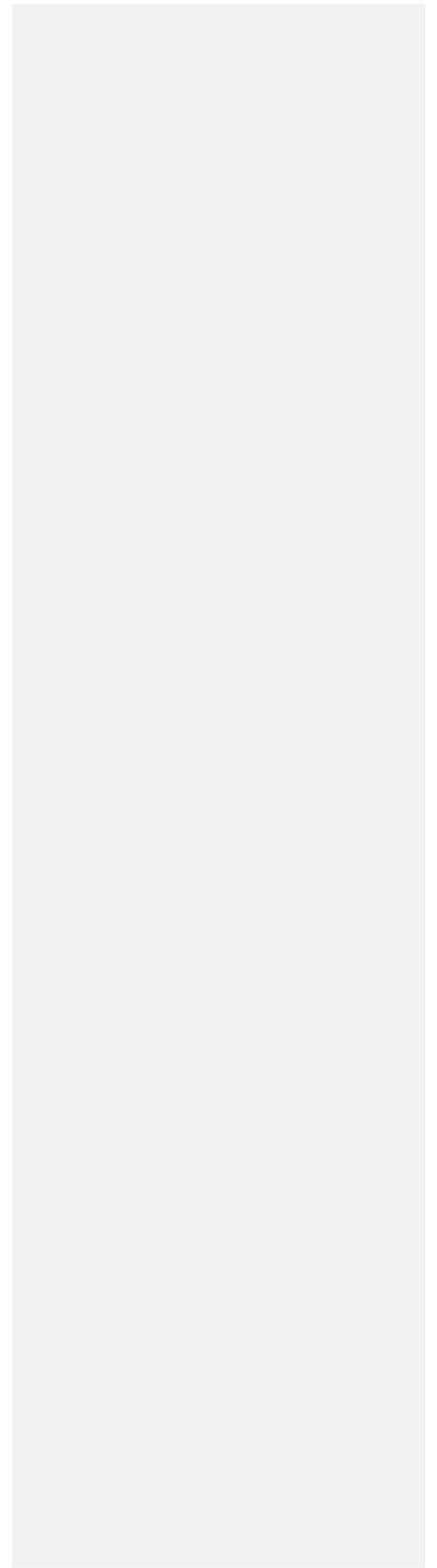
```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",t
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT MA00001
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4
```

Figure 12. Fichier VCF(Marcketta 2013)

ANNEXE II

Document Vision

<confidentiel>



ANNEXE III

Identification des intervenants et des utilisateurs

Nom	Description	Responsabilités
Gestionnaire de projet (Alain April professeure)	Responsable du déroulement et du succès du projet.	Définit les exigences de l'application conjointement avec le responsable du prototype. Fait un suivi détaillé de chacune des étapes. S'assure de l'avancement du projet. Définit, vérifie et valide les livrables. Aide l'équipe de projet à trouver des solutions en cas de difficulté.
Responsable du prototype du module génétique. (Rodolfo Ramos, étudiant en maîtrise de génie logiciel)	Personne responsable de la collection, de l'analyse, de la définition de besoins et spécification et de la conception prototype.	Assiste le gestionnaire de projet pour définir les exigences de l'application. S'occupe de la conception de l'application. S'occupe de colliger, de analyser les besoins du client. Élabore la cartographie des processus du client. Définit les fonctions du système

		à développer pour adresser ces besoins.
Alain Moreau	Directeur de la recherche au CHU Sainte-Justine.	Valide les spécifications du projet d'informatisation du CHU Sainte-Justine, et prend les décisions finales.
Cédric Julien	Généticien	Procède aux analyses génétiques sur les données du laboratoire, et extrait les informations génétiques demandées par les étudiants.
Robert Eveleigh	Bio-informaticien	Effectue le traitement, filtrage et transformations des données sur le « cluster » de Génome Québec, sur instructions du généticien.

Tableau 2. Résumé des intervenants

Alain April 14-8-27 22:38

Deleted: 1

Nom	Description	Responsabilités	Intervenant
Les étudiants	Utilisateurs du nouveau système	Faire des recherches sur les données génétiques dans le cadre de leurs travaux.	Anita Franco
Généticien	Utilisateur privilégié	Mets à jour les données génétiques de la BD Éliisa accessible aux étudiants.	Cédric Julien

Tableau 3. Résumé des utilisateurs

Alain April 14-8-27 22:38

Deleted: 2

ANNEXE IV

SPÉCIFICATIONS DES CAS D'UTILISATION

CU-01: S'authentifier

Description générale: Permet à un utilisateur d'accéder aux autres cas d'utilisation du système.

Acteurs : généticien, super utilisateur, chercheur, auditeur.

Règle d'initiation : L'utilisateur doit avoir un compte et un mot de passe pour accéder au système.

Scénario nominal :

1. Le système affiche la fenêtre d'authentification.
2. L'utilisateur saisit le nom et le mot de passe.
3. Le système vérifie le nom d'utilisateur et le mot de passe dans le serveur.
4. La connexion est réussie et le système affiche l'application sur le navigateur web.

Règle de terminaison : L'utilisateur est authentifié dans le système.

Scénario alternatif

Le nom d'utilisateur ou le mot de passe sont invalides.

- 1- Le système affiche le message d'erreur que les données saisies par l'utilisateur sont invalides.
- 2- Le système propose à l'utilisateur de saisir de nouveau le nom et le mot de passe.

CU-02: Ajouter un utilisateur

Description générale: Permet d'ajouter un nouvel utilisateur ayant accès au système et de lui attribuer un rôle (généticien, chercheur).

Acteurs : super utilisateur

Règle d'initiation : L'utilisateur doit être authentifié en tant que super utilisateur et il est dans l'interface utilisateur.

Scénario nominal :

5. Le super utilisateur choisit l'option « ajouter utilisateur »
6. Le super utilisateur saisit le nom d'utilisateur, son rôle et le mot de passe.
7. Le système envoie les données au serveur.
8. Le système affiche le message que l'utilisateur a été ajouté.

Règle de terminaison : L'utilisateur a été ajouté dans le système.

Scénario alternatif:

Le nom d'utilisateur existe dans le système.

1. Le système affiche le message que le nom d'utilisateur existe dans le système.
2. Le système suggère de saisir de nouveau nom d'utilisateur.

CU-03: Gérer les utilisateurs

Description générale: Permet de modifier le rôle et les permissions d'un utilisateur.

Acteurs : super utilisateur

Règle d'initiation : L'utilisateur doit être authentifié en tant que super utilisateur et il est dans l'interface utilisateur.

Scénario nominal :

1. L'administrateur choisit l'option « gérer les utilisateurs »
2. Le système affiche la liste des utilisateurs.
3. L'administrateur choisit un utilisateur.
4. Le système affiche les données d'un utilisateur.
5. L'administrateur modifie le rôle ou la permission d'utilisateur et soumet l'opération.
6. Le système affiche le message confirmant les modifications des données d'un utilisateur.

Règle de terminaison : Le rôle et/ou les permissions de l'utilisateur ont été mise à jour dans le système.

CU-04: Archiver un fichier brut

Description générale: Permet au généticien de conserver les fichiers bruts de séquençage, et ainsi conserver une traçabilité.

Acteurs : généticien.

Règle d'initiation : L'utilisateur doit être authentifié et il est dans l'interface des fichiers bruts.

Scénario nominal :

1. Le généticien choisit l'option « téléverser un fichier »
2. Le généticien choisit un fichier brut de l'ordinateur et soumet l'opération.
3. Le système affiche le message confirmant que le fichier a été ajouté.

Règle de terminaison : Le fichier brut a été ajouté dans le système.

CU-05: Téléverser un fichier VCF

Description générale: Permet au généticien de téléverser un nouveau fichier VCF dans la base de données. L'état du nouveau fichier sera inactif par défaut pour lui permettre de valider les données contenues dans le VCF avant de rendre les données accessibles aux chercheurs.

Acteurs : généticien.

Règle d'initiation : L'utilisateur doit être authentifié et il est dans l'interface des fichiers vcf.

Scénario nominal :

4. Le généticien choisit l'option « téléverse un fichier »
5. Le généticien choisit un fichier brut de l'ordinateur et soumet l'opération.
6. Le système affiche le message confirmant que le fichier a été ajouté.

Règle de terminaison : « Le fichier vcf » a été ajouté dans le système.

CU-06: Activer / Désactiver / Supprimer un fichier VCF

Description générale: Permet de changer l'état d'un fichier VCF dans le système. Pour rendre un fichier VCF disponible aux chercheurs, il doit être activé. Lorsqu'un fichier VCF est désactivé, il demeure disponible aux chercheurs l'ayant déjà accédé préalablement; mais

il n'est plus accessible pour les autres chercheurs. Lorsqu'un fichier VCF est supprimé, il est éliminé de la base de données.

Acteurs : super utilisateur.

Règle d'initiation : L'administrateur est dans l'interface des « fichiers vcf ».

Scénario nominal :

7. L'administrateur choisit un fichier de la liste de fichiers VCFs.
8. L'administrateur active ou désactive le fichier choisi.
9. Le système affiche le message confirmant que le fichier a été activé ou désactivé.

Règle de terminaison : L'état du « fichier vcf » est modifié dans le système.

CU-07: Rechercher les données génétiques

Description générale: Permet au chercheur d'effectuer une requête sur la base de données génétique.

Acteurs : généticien, chercheur.

Règle d'initiation : L'utilisateur est authentifié et il est dans l'interface « Recherches des données génétiques »

Scénario nominal :

1. L'utilisateur choisit les tables sur la liste de tables.
2. L'utilisateur choisit l'option « afficher tables sélectionnées ».
3. Le système affiche les tables sélectionnées avec leurs attributs.
4. L'utilisateur choisit l'option « obtenir colonnes »
5. Le système affiche les colonnes sélectionnées dans un tableau et génère la requête SQL.
6. L'utilisateur choisit l'option « exécuter » pour soumettre la requête.
7. Le système affiche les résultats de la requête sur l'écran.

Règle de terminaison : Les résultats de la requête sont affichés sur l'écran.

CU-08: Exporter les résultats d'une requête

Description générale: Permet au chercheur d'exporter les résultats d'une recherche au format Microsoft Excel (CSV).

Acteurs : généticien, chercheur.

Règle d'initiation : Les résultats de la requête sont affichés sur l'écran.

Scénario nominal :

1. L'utilisateur choisit l'option « sauvegarder résultats »
2. Le système affiche une fenêtre pour que l'utilisateur saisi un nom de fichier et la localisation où le fichier va être sauvegardé.
3. Le système affiche un message que le fichier a été sauvegardé.

Règle de terminaison : Un fichier au format Microsoft Excel (CSV) a été créé.

CU-09: Sauvegarder une requête

Description générale: Permet au chercheur de sauvegarder dans le système une requête sur un fichier VCF.

Acteurs : généticien, chercheur.

Règle d'initiation : Les résultats de la requête sont affichés sur l'écran.

Scénario nominal :

1. L'utilisateur choisit l'option « sauvegarder requête »
2. Le système affiche une fenêtre pour que l'utilisateur saisi un nom de fichier et la localisation où le fichier va être sauvegardé.
3. Le système affiche un message que le fichier a été sauvegardé.

Règle de terminaison : Un fichier avec la requête a été créé.

CU-10: Charger une requête

Description générale: Permet au chercheur de retrouver une requête ayant été sauvegardée et de la charger dans le système pour la consulter, ou l'exécuter de nouveau.

Acteurs : généticien, chercheur.

Règle d'initiation : L'utilisateur est dans l'interface « Recherches des données génétiques »

Scénario nominal :

1. L'utilisateur choisit l'option « charger une requête »
2. L'utilisateur charge la requête dans le système
3. Le système affiche un message dont la requête a été chargée.

Règle de terminaison : la requête a été chargée dans l'application.

CU-11: Auditer les transactions

Description générale: Permet à l'auditeur de consulter toutes les transactions effectuées sur le système.

Acteurs : auditeur.

Règle d'initiation : L'utilisateur est authentifié dans le système.

Scénario nominal :

1. Le système affiche toutes les transactions effectuées dans le système.
2. L'utilisateur recherche les transactions par date ou par utilisateur.
3. L'utilisateur peut imprimer les résultats des transactions.

Règle de terminaison : Les résultats des transactions sont imprimés

LISTE DE RÉFÉRENCES BIBLIOGRAPHIQUES

Actasanitaria (2013). "Sistemas Genómicos lanza GeneSys, una innovadora herramienta de análisis e interpretación de datos masivos." from <http://www.actasanitaria.com>.

Annovar. "ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data." from <http://www.openbioinformatics.org/annovar/>.

Braly, J.-P. (2012). "SCOLIOSE CONGÉNITALE Une interaction entre gènes et environnement." from <http://www.larecherche.fr/actualite/sante/scoliose-congenitale-interaction-entre-genes-environnement-01-07-2012-91314>.

Bruandet, J. "LA SCOLIOSE." from http://www.med.univ-rennes1.fr/sisrai/art/scoliose_p_291-294.html.

Collins, F. S., et al. (2003). The Human Genome Project : Lessons from Large Scale Biology.

Dallaire, M. (2008). "CHU Saint-Justine Research Center - Inauguration of World-Leading Musculoskeletal Disorders Laboratory." CNW Telbec.

Deléage, G. and M. Gouy (2013). Bioinformatique.

GEMINI. "GEMINI a flexible framework for exploring genome variation." from <http://gemini.readthedocs.org/en/latest/index.html>.

GenomeQuébec (2014). "GÉNOMIQUE." from <http://www.genomequebec.com/genomique-101.html>.

Genómicos, S. (2013). GeneSystems Platform.

Genomics, T. (2014). from <http://www.tutegenomics.com/>.

Guilizzoni, G. (2014). "Balsamiq." Balsamiq Studios.

Huerta, M., et al. (2000) NIH working definition of bioinformatics and computational biology.

ISO/IEC (2000). "Information technology — Software product quality."

Marcketta, A. (2013). "A Deep Catalog of Human Genetic Variation." 1000 Genomes(Wednesday 9th October, 2013 21:07).

Masselot, M. (2014). "Génétique." from http://www.edu.upmc.fr/sdv/masselot_05001/polymorphisme/snp.html.

NCHGR (2014). "Questions About Genetic and Genomic Science." from <http://www.genome.gov/19016904>.

Pasquesoone, V. (2014). "Une méthode de dépistage des maladies génétiques plus performante et sans risques."

Plume, P. (2011). "GALAXY." from <https://www.projet-plume.org/fiche/galaxy>.

University, J. H. (2014). "GALAXY." from <https://usegalaxy.org/>.

Wikipedia (2014). "Méthode agile." [Wikipedia](#).

Yandell, M. (2011). "Variant Annotation, Analysis and Search Tool - VAAST 2." from <http://www.yandell-lab.org/software/vaast.html>.

