Recommandations générales pour la gestion informatique des données et des analyses de séquençage à haut débit pour les laboratoires de diagnostic moléculaire de maladies génétiques.

Date de rédaction : 31/07/2014	Date de révision : 30/05/2016	**Version :** 1

Ce document a été préparé par les membres du GT1 « Cahier des charges informatique, bio-analyse/bio-informatique, bases de données mutations » dans le cadre du Réseau NGS Diagnostic.

Groupe de travail : Claire Bardel (Lyon), Christine Bellané-Chantelot (Paris), Christophe Béroud (Marseille), Céline Bonnet (Nancy), Audrey Briand (Paris), Laurent Castera (Caen), Yannis Duffourd (Dijon), Bénédicte Gérard (Strasbourg), Claude Houdayer (Paris), Eulalie Lasseaux (Bordeaux), Julie Leclerc (Lille), Cédric Le Maréchal (Brest), Jean Muller (Strasbourg), Julie Nocq (Dijon), Jean Baptiste Rivière (Dijon), David Salgado (Marseille).

Ce document a été relu en s'appuyant sur l'expertise et les compétences des équipes informatiques et/ou bioinformatiques de nos institutions, que nous remercions :

-Bioinformatique du « Centre Normand de Génomique et de Médecine Personnalisée » de Caen (Laurent Castéra, Antoine Rousselin, Baptiste Brault, Germain Paimparay) et de Rouen (Myriam Vezain, Sophie Coutant, Olivier Quenez, Raphaël Lanos).
-Bioinformatique de Lille (Emilie Ait Yahya, Fabrice Bonte, Christophe Demay et Martin Figeac)
-Hôpitaux Universitaires de Strasbourg (Amandine Velt, Anthony Le Béchec, Fabrice Stalter)
-Institut Curie-U900 (Nicolas Servant).

Sommaire

1 OBJET

Ce document contient les recommandations de l'ANPGM (Association Nationale des Praticiens de Génétique Moléculaire) pour la gestion des données informatiques dans le cadre de l'implémentation et du développement du séquençage à haut débit dans les laboratoires de diagnostic.

2 PORTEE & APPLICATIONS

Ce document cible les laboratoires qui désirent acquérir ou qui ont fait l'acquisition d'une plateforme de séquençage à haut débit. Il contient les recommandations pour la mise en œuvre de la partie informatique avec des points sur l'infrastructure informatique elle-même ainsi que des recommandations sur les formats à utiliser.

3 INTRODUCTION

Le séquençage à haut débit est une technologie de biologie moléculaire permettant d'accéder à l'intégralité de la séquence d'ADN d'un patient, ceci de manière plus simple et plus rapide que par la méthode classique de séquençage. Son application révolutionne la génétique humaine d'une part par l'amélioration des tests diagnostiques (augmentation du rendement diagnostique et implémentation de nouveaux tests) et d'autre part par l'accélération de la découverte de nouveaux gènes responsables de pathologies. L'implémentation de cette technologie dans les laboratoires de biologie moléculaire hospitaliers est une priorité qui nécessite des moyens et des compétences multiples qui ne sont pas nécessairement présentes au sein des laboratoires. Notamment, cette technologie nécessite un environnement informatique et bioinformatique qui n'est généralement pas négligeable. Cet environnement est l'un des garants de la bonne réalisation de ces analyses.

La Direction du Système Informatique (DSI), garante des infrastructures informatiques hospitalières devra faire pleinement partie des réflexions et des évolutions nécessaires à la mise en place du diagnostic en séquençage haut-débit. Il est ainsi judicieux d'inclure de nouveaux profils de poste (compétences en informatique et bioinformatique) au sein des équipes biologiques. Ces personnes devront développer une synergie avec la DSI. Une plateforme ou un automate de séquençage à haut débit produit des fichiers bruts différents selon la technologie employée, des images d'émission de fluorescence (pixel) ou encore des mesures de pH (Volt)… Ces données sont de taille très importante, pouvant aller jusqu'à plusieurs téraoctets. Ces mesures brutes sont ensuite traduites en nucléotides par un logiciel interne à l'automate de séquençage. L'information produit par ce logiciel contient alors différentes données : qualitative (base de l'ADN) et quantitative (intensité et valeur de qualité). Les séquences nucléotidiques sont par la suite alignées sur le génome humain de référence et puis comparées à ce génome de référence pour en extraire les variations. La pathogénicité des variations est par la suite évaluée par différents logiciels de prédiction (impact sur la fonction de la protéine, analyse du degré de conservation de l'acide aminé, analyse de l'effet sur l'épissage …)

Ces processus et traitements des différentes données qui sont stockées dans des fichiers de formats variés demandent des besoins informatiques plus ou moins importants incluant la capacité de paralléliser les analyses, une puissance de calcul importante, une mémoire vive et une capacité de stockage importante. Ce document vise à identifier ces besoins en fonction de la plateforme utilisée.

4 CONTEXTE TECHNIQUE

4.1 Unité de calcul

Le processeur, ou CPU (Central Processing Unit), est le composant de l'ordinateur qui exécute les programmes informatiques. Un processeur multi-cœur est un processeur possédant plusieurs cœurs physiques qui travaillent en parallèle. Un cœur physique est une sous-unité particulière du processeur capable d'exécuter des programmes de façon autonome. Ainsi un CPU possédant par exemple 8 cœurs est capable d'exécuter 8 programmes indépendants en parallèle, au contraire d'un CPU à un seul cœur qui ne pourra exécuter qu'un seul programme à la fois.

La puissance d'un processeur se mesure en Flops. Le Flop est un acronyme signifiant « opérations à virgule flottante par seconde » (Floating point Operations Per Second). Le nombre de Flops est une unité de mesure de la vitesse d'une machine ou d'un ensemble de machines. On parle généralement de téraflop, soit : 1Tflop = 10^{12}Flops.

4.2 Taille des données

Le bit (binary digit) et l'octet sont les unités de base de l'informatique. Un bit ne peut prendre que 1 ou 0 comme valeur et un octet équivaut à 8 bits. De manière générale, on peut considérer que le terme octet et le terme byte sont synonymes mais c'est considéré comme un abus de langage (1 byte est en général égal à 8 bit mais peut être égal à 7 ou 9 bits). On écrit plutôt 1b que 1bit, 1o que 1octet et 1B que 1byte. Lorsqu'on parle de ko, de Mo (ou Méga octet) ou de Go (ou Giga octet), on parle de 1000o, 1 000 000o et de 1 000 000 000o. I

MEDIA	TAILLE
1 CV en .doc	~500 ko
1 mp3	~3 Mo
1 photo HD	~15 Mo
1 DVD	~4,3 Go
1 Blu-ray	~50 Go

Table 1 Quelques exemples de fichiers avec leur taille.

4.3 Récapitulatif des caractéristiques des séquenceurs à haut débit

La technologie du séquençage à haut débit couvre en réalité un nombre important de technologies différentes qui se distinguent par leur fiabilité, leur débit et leur coût. En particulier, la capacité des séquenceurs à haut débit varie grandement selon les constructeurs et les gammes (Table 2). On distinguera les séquenceurs à très haut débit utilisés pour le séquençage de génome entier et d'exome entier (ex, Illumina HiSeq, Ion Proton…) des séquenceurs à moyen débit qui permettent de séquencer un nombre plus réduit de loci souvent inférieur à l'exome (Illumina MiSeq, Ion torrent PGM…).

Instrument	Durée	Millions de Reads/run	Bases/read	Gb/Run
Applied Biosystems 3730	2h	0,000096	650	0,00006
454 GS Jr. Titanium	10h	0,1	400	0,1
454 FLX Titanium	10h	1	400	0,4
454 FLX+	20h	1	650	0,7
Illumina GA IIx v5 SE	2j	640	36	23
Illumina GA IIx v5 PE	14j	640	*288*	184,3
Illumina MiSeq v2 Nano	17h	1	*300*	0,3
Illumina MiSeq v2 Micro	19h	4	*300*	1,2
Illumina MiSeq v3	20h	22	*150*	3,3
Illumina MiSeq v3	55h	22	*600*	13,2
Illumina NextSeq 500 Mid	15h	130	*150*	19,5
Illumina NextSeq 500 High	18h	400	*150*	60
Illumina HiSeq 2500 Rapid run	27h	300	*200*	60
Illumina HiSeq 2500 v3	11j	1500	*200*	300
Illumina HiSeq X (2 flow cells)	3j	6000	*300*	1800
Ion Torrent – PGM 314 chip	2,3h	0,475	200	0,1
Ion Torrent – PGM 314 chip	3,7h	0,475	400	0,2
Ion Torrent – PGM 316 chip	3h	2,5	200	0,5
Ion Torrent – PGM 316 chip	4,9h	2,5	400	1
Ion Torrent – PGM 318 chip	4,4h	4,75	200	1
Ion Torrent – PGM 318 chip	7,3h	4,75	400	1,9
Ion Torrent - Proton I	4h	70	175	12,3
Ion Torrent - Proton II	5h	280	175	49
Ion Torrent - Proton III	6h	500	175	87,5
Life Technologies SOLiD 5500xl	8j	1410	110	155
Pacific Biosciences RS II	2h	0,03	3000	0,1
Oxford Nanopore MinION	≤6h	0,1	9000	0,9

Table 2. Tableau recensant les principales solutions de séquençage à haut débit en y incluant la durée d'un run, et la capacité de chaque solution en y incluant le nombre de read (lecture), la taille de ces reads et le débit final théorique.

4.4 Protocole et formats de fichiers

L'analyse des données issue d'une expérience de séquençage à haut débit requiert l'utilisation de nombreux algorithmes bioinformatiques. Le chainage de ces programmes constitue ainsi le protocole bioinformatique (ou *pipeline/workflow*). Ce protocole génère de nombreux types de fichiers différents d'un volume et d'une utilité pour l'utilisateur final très variable. De manière générale, chaque type de fichier correspond à une grande étape de ce protocole bioinformatique (Figure 1). En résumé l'étape du « *Raw Data Processing* » correspond aux traitements des données brutes issues des séquenceurs afin d'obtenir des séquences d'ADN, l'étape du « *Genome Mapping* » correspond à l'alignement (positionnement) des séquences du patient sur le génome de référence, l'étape du « *Variant Calling* » correspond à la détection et la quantification des différences (variants ou variations) par rapport au génome de référence, l'étape d'« *Variant Annotation* » permet d'attacher des éléments biologiques connus aux variants identifiés et l'étape optionnelle de « *Variant Ranking* » permet de hiérarchiser les variations afin d'identifier plus facilement la ou les variations responsables de la pathologie du patient.

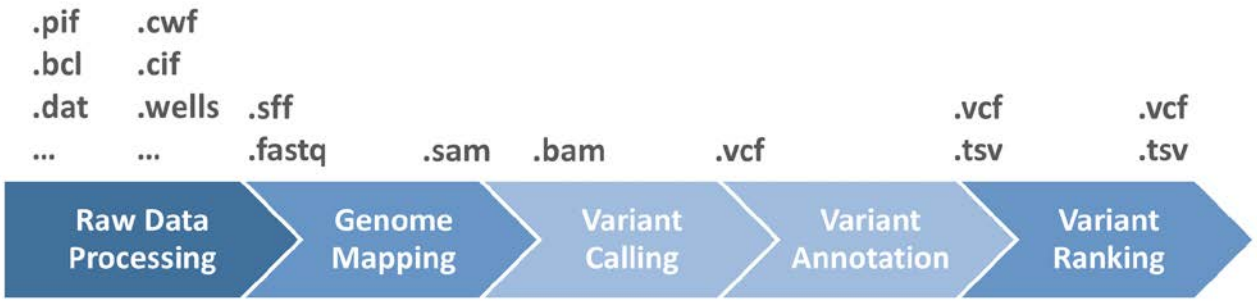

Figure 1. Protocole bioinformatique schématisant les 5 grandes étapes de l'analyse des données ainsi que les extensions des fichiers associées à ces étapes.

Dans ce document nous distinguerons 3 catégories de fichiers/formats, les **données brutes** (**données 0,** directement issues du séquenceur), les **données primaires** (**données I**, séquences et valeurs de qualité) et les **données secondaires** (**données II**, analyses additionnelles obtenues à partir des séquences).

Certains formats de fichiers sont propriétaires et dépendent des plateformes de séquençage utilisées. Ces formats spécifiques sont en général attachés aux données brutes manipulées par les logiciels propriétaires des fabricants de séquenceurs. Les autres formats de fichiers utilisés sont en général des formats libres, développés et validés par la communauté internationale et constituent des standards acceptés. Les types de fichiers décrits ci-dessous ne constituent pas une liste exhaustive des types de fichiers par plateforme mais un aperçu de ceux les plus communément utilisés :

4.4.1 Données brutes (données 0)

- Le fichier PIF (extension .pif) est un fichier image qui contient les images brutes de la plateforme Roche (GS Junior).
- Le fichier CWF (*Composite Wells Format*, extension .cwf) est un fichier qui contient les signaux non corrigés extraits des images de la plateforme Roche (séquenceurs 454).
- Le fichier DAT (extension .dat) est un fichier contenant la conversion des mesures brutes de pH en mesures numériques des voltages.
- Le fichier WELLS (extension .wells) est le fichier d'entrée pour la détection des bases. Il constitue le format brut pour la plateforme Ion Proton et Ion PGM.
- Le fichier BCL (*Base Call File*, extension .bcl) est un fichier binaire contenant les bases détectées, leur qualité pour chaque cycle et constitue le format brut pour la plateforme Illumina.

- Le fichier CIF (*Cluster Intensity Files*, extension .cif) est un fichier texte contenant les intensités extraites des images pour la plateforme Illumina.

4.4.2 Données primaires (données I)

- Le fichier FASTA (extension .fasta ou .fa) est un format de fichier texte utilisé pour stocker des séquences biologiques de nature nucléique ou protéique. Aucune information sur la qualité des séquences n'est donnée. Ce fichier texte doit être compressé pour le stockage.
- Le fichier FASTQ (extension .fastq ou .fq) est un fichier texte standard pour l'échange de données de séquence et de qualité utilisé pour tout type de séquenceur y compris Sanger. Il contient les noms des séquences, les séquences et la valeur de qualité des nucléotides. Ce fichier texte doit être compressé pour le stockage.
- Le fichier SFF (*Standard Flowgram File*, extension .sff) est un fichier texte utilisé pour stocker les séquences biologiques associées aux technologies Roche (séquenceurs 454) et Life Technologies (Ion torrent). Il contient les séquences, les données de qualité et les données de flux permettant un retraitement des données sans perte de signal. Ce fichier texte doit être compressé.

4.4.3 Données secondaires (données II)

- Le fichier SAM (*Sequence Alignment/Map*, extension .sam), est un fichier texte représentant des séquences alignées issus des programmes d'alignement.
- Le fichier BAM (*Binary Alignment/Map*, extension .bam), est la version binaire et compressée du fichier SAM. Il peut être accompagné d'un fichier d'index (« .bai »). On notera que ce format de fichier peut stocker également des séquences non alignées et il peut ainsi sous certaines conditions être considéré comme une donnée primaire.
- Le fichier VCF (*Variant Call Format*, extension .vcf), est un fichier texte répertoriant les variations observées dans un ou plusieurs patients. Celui-ci peut-être compressé (« vcf.gz ») et dans ce cas il peut être accompagné d'un fichier d'index (« vcf.gz.tbi »).
- Le fichier gVCF (Genomic Variant Call Format, extension .gvcf), est une variante du fichier VCF. Contrairement au fichier VCF il contient des informations pour toutes les régions génomiques même si une variation n'a pas été identifiée. Il permet ainsi de connaître les regions plus ou moins bien sequencées necessaire dans le cadre d'applications cliniques. Les fichiers gVCF sont généralement d'une taille inférieure à 1Go ou 1/100 de la taille du fichier BAM utilisé pour l'étape de variant calling. (dans la suite du document les gVCF sont assimilés aux fichiers VCF)
- Le fichier TSV (*Tab Separated Values*, extension .txt ou .tsv), est un fichier texte répertoriant les variations observées dans un patients ainsi que des annotations fonctionnelles variées utiles pour l'interprétation des données. Celui-ci est souvent lisible dans un tableur. Ce type de fichier peut-être compressé.

Remarques additionnelles :

Il est important de noter que la plupart des fichiers au format texte générés par les protocoles bioinformatiques peuvent être compressés et permettre ainsi un gain de place non négligeable. La taille de ces données peut être divisée par 4 environ. La plupart des logiciels d'analyse permettent de travailler directement sur ces fichiers compressés. La compression utilisée pour le séquençage haut débit est une compression sans perte utilisant généralement le format gzip correspondant à l'extension « .gz » (RFC 1952). Principalement utilisée sur des systèmes UNIX/Linux, il existe cependant certains programmes sous Windows capables de générer des fichiers compatibles avec gzip. **Nous recommandons fortement de compresser ces fichiers en utilisant si possible le plus haut taux de compression.**

4.5 Tableau synthétique des formats et volumes associés

Il est important de noter que les volumes indiqués sont établis à un moment donné pour une version de logiciels donnés et ne sont pas nécessairement valables indéfiniment.

	Roche		Illumina				Ion Torrent [a]						
Type et taille des fichiers générés/Run	**Type**	**Junior**	**Type**	**HiSeq[c]**	**NextSeq[d]**	**MiSeq**	**Type**	**Proton[b]**	**S5XL 520[e]**	**S5XL 530[e]**	**S5XL 540**	**PGM 318a**	**PGM 316a**	**PGM 314a**
Séquences (b=base)		100Mb		600Gb	40-120Gb	8Gb		10Gb	2Gb	4Gb	15Gb	1Gb	100Mb	40Mb
Données 0 (brutes)	PIF	10Go	BCL	9To	50Go	100Go	DAT	3To	356Go	869Go	~2To	393Go	39Go	29Go
Données I (fluorescence, volt)	CWF	750Mo	CIF	2,5To		20Go	WELLS	219Go	30Go	75Go	~180Go	23,3Go	2,3Go	1,9Go
Données I (séquences)	SFF	450Mo	FastQ compressé	600Go	50Go	8,5Go	SFF	112Go				11Go	1,1Go	0,5Go
Données II (alignements)			SAM	2To	50Go	50Go	Unaligned BAM		55Go	75Go	~85Go			
Données II (alignements)	BAM	15Mo		512Go	10Go	10Go		35Go	10Go	25Go	~55Go	3,6Go	0,36Go	0,2Go
Données II (variants[d])	Txt	3Mo	VCF	0,64Go	0,01Go	0,01Go		0,12Go				0,01Go	0,01Go	1Mo
TOTAL		**11,2Go**		**15To**	**150Go**	**190Go**		**3,4To**	**451Go**	**1,04To**	**2,32To**	**431Go**	**43Go**	**32Go**

Table 3. Récapitulatif des tailles approximatives des fichiers pour les principaux séquenceurs Illumina, Roche et Ion Torrent. [a]en utilisant la Torrent Suite Software 3.6 et le kit 200 bp. [b]en utilisant la Torrent Suite Software 3.6 et le kit 400 bp. [c]Les données sont proposées pour un run soit en moyenne 64 exomes (pour 1 exome le SAM représente 30Go et le BAM 8Go). [d]La taille des vcf est dépendante du nombre de variants et donc du type d'application utilisée et selon le cas des annotations ajoutées. Les données présentées sont des estimations basées sur des données et des performances à un moment donné. [d]Pour un run en mode « high output » avec 150 pb. [e]Pour un run à 400pb.

5 RECOMMANDATIONS : ARCHIVAGE DES DONNÉES INFORMATIQUES

5.1 Objectifs de l'archivage

Les objectifs d'une conservation à moyen ou long terme des données informatiques de séquençage haut débit sont doubles :

- Traçabilité technique (accréditation ISO15189) :
  - Traçabilité des fichiers de travail ayant permis d'aboutir au résultat final (données brutes, I et II). La durée de stockage de ces données doit couvrir au minimum la période entre deux visites du COFRAC. La durée de conservation des enregistrements techniques (données brutes, I ou II) est définie au sein de chaque laboratoire (Système de Management de la Qualité). Remarque, si le traitement des données brutes est captif de l'automate de séquençage, il n'est pas nécessaire de conserver ces données brutes.
  - Traçabilité des contrôles qualité ayant permis de valider l'analyse chez un patient donné : dans ce cas précis, valeur de qualité des bases analysées (différence entre le signal spécifique et le bruit de fond). Cette valeur de qualité est générée et est conservée dans les données I (voir chapitre suivant).
  - Traçabilité des versions des logiciels d'analyse (« *Raw Data Processing* », « *Genome Mapping* », « *Variant Calling* »,…) utilisés sur l'ensemble de la chaine d'analyse et des filtres automatiques
- Possibilité de réanalyser des données : il est donc nécessaire de conserver les données I d'entrée sur le pipeline. Par exemple, à la suite d'une amélioration du pipeline bioinformatique.

5.2 Contraintes liées aux données à conserver

Plusieurs facteurs techniques sont ainsi à considérer, incluant : la stabilité dans le temps des données générées, la capacité à régénérer les données sur un nouveau pipeline bioinformatique le cas échéant, le degré d'exhaustivité des données, l'utilité pour l'analyse biologique proprement dite, le volume de stockage, et les flux des données. Ces 2 derniers points influent directement sur le coût car ils visent la qualité de l'infrastructure informatique (stockage et réseau).

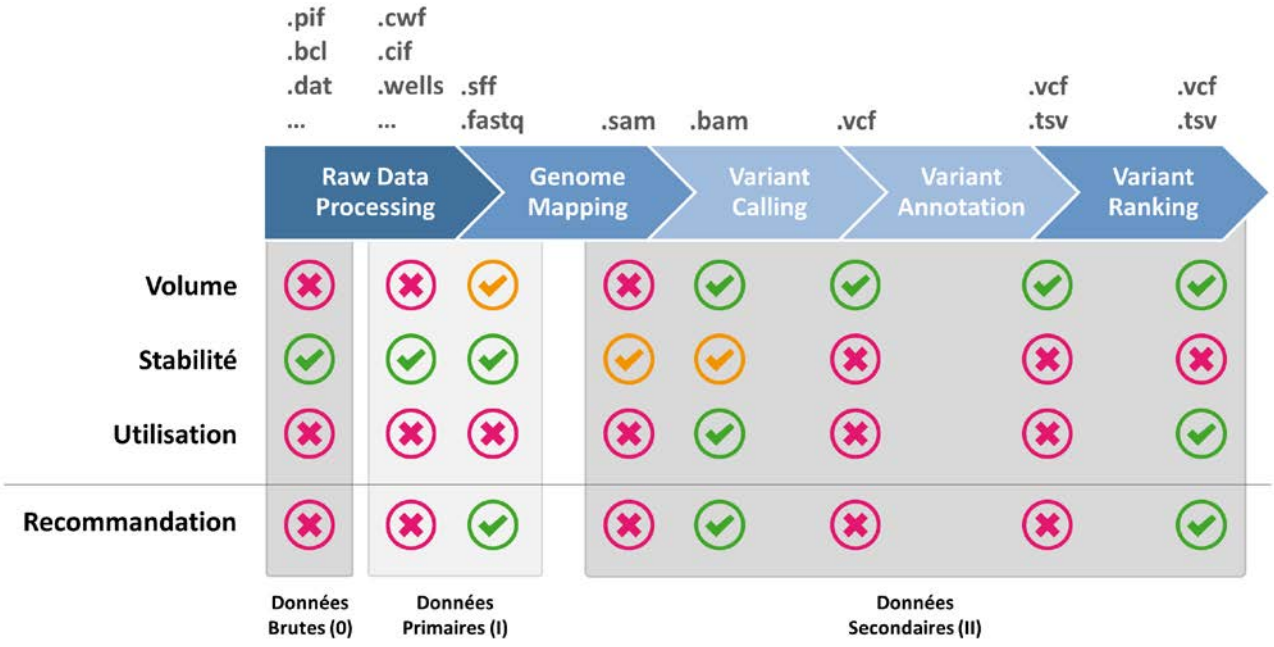

Figure 2. Format de fichiers associés aux grandes étapes du protocole bioinformatique.On notera une alternative possible pour les données primaires sous la forme de fichier bam (comportant toutes les séquences).

Le « ***Volume*** » indique le rapport bénéfice/coût du stockage informatique des fichiers générés. La « ***Stabilité*** » informe sur la stabilité relative des données générées dans le temps. Ainsi les données brutes pour lesquelles il y a peu de changements dans les programmes fournis par les fabricants de séquenceurs sont extrêmement stables. D'autres données comme l'annotation des variants sont dépendantes d'éléments extérieurs mis à jour régulièrement, par exemple, la version du génome utilisée (GRChXX) ou les références des séquences (identifiant RefSeq NM_XXX). L' « ***Utilisation*** » donne le degré d'utilité pour l'analyse biologique proprement dite.

Concernant l'« ***exhaustivité des données*** » certaines étapes du pipeline bioinformatique peuvent éliminer des données et donc conduire à une perte d'information en cas de réanalyse ultérieure. Par exemple, l'élimination dans les fichiers fastq des données de séquence ayant une valeur de qualité insuffisante ; l'élimination dans les fichiers .bam de données s'alignant de façon incorrecte (par exemple, lectures avec peu de bases alignées, localisations multiples sur le génome,…) ; l'élimination dans les fichiers vcf de variants de faible ratio allélique. **Ces filtrages dépendent généralement des paramétrages automatiques du pipeline bioinformatique et il est important de noter ces limitations.**

On notera ainsi que les fichiers bam peuvent constituer une alternative moins onéreuse au stockage des fichiers fastq ou sff. Dans ce cas, le fichier doit impérativement contenir les séquences alignées sans hard-clipping[1] ainsi que les séquences non alignées.

5.3 Recommandations pour l'archivage de données informatiques

De manière très générale et en reprenant les classifications énoncées dans le paragraphe précédent nous recommandons de conserver les formats de sortie des données primaires (données I), et les formats de sortie de certaines données secondaires (données II) qui traduisent un état final d'une analyse. Ainsi, dès que la qualité d'un run a pu être établie, les données brutes peuvent être supprimées.

Prenons l'exemple du séquenceur Illumina HiSeq 2500 qui permet d'obtenir 600 milliards de bases en 11 jours, ce qui correspond de manière théorique à un génome humain avec une profondeur de lecture de 200x (ou 4 à 5 génomes avec une profondeur de lecture de 40x) ou un exome humain avec une profondeur de lecture de 12000x (ou 120 exomes avec une profondeur de lecture de 100x). En pratique, on séquence plutôt 3 à 4 génomes ou 64 exomes humains.

Le volume total des données générées par run de ce séquenceur correspond à la taille des images brutes (9To), aux intensités extraites des images (2,5To), aux fichiers de séquences compressés (.fastq contenant les reads et leurs qualités pour 600Go). S'ajoutent encore à cela les fichiers .sam et .bam et les .vcf (3To). Ainsi pour un run de onze jours l'HiSeq produit environ 15To de données. En extrapolant l'activité à deux runs par mois pendant une année cela fait un total de 360To de données par an. Cette quantité importante de données ne doit pas nécessairement être conservée.

- Les **images** ont une taille de 9To par run, c'est la partie la plus volumineuse des données. Bien que les images permettent de réitérer l'analyse complète des données, leur transfert vers un serveur de stockage nécessitera une bande passante très importante ou alors un temps très long.
- Les fichiers d'**intensité** sont eux aussi volumineux, environ 2,5To de données par run.

En raison, de leur volume très important et la faible probabilité de leur réanalyse (du fait de la stabilité des algorithmes d'analyses fournis par les fabricants de séquenceurs), **nous ne recommandons pas de garder ces 2 types de fichiers (données brutes),** ceci dès que les données finales sont disponibles et validées.

Concernant les données primaires, plusieurs choix sont possibles selon la plateforme et la version du pipeline bioinforamtique utilisé. Hormis certaines incompatibilité notées ci-dessous,:

- Les fichiers **fastq**, d'une taille plus respectable (600Go par run), contiennent les séquences ainsi que l'information de qualité avant l'étape alignement. Du fait de sa taille acceptable et de son utilité, ce type de fichier peut être considéré comme **une donnée primaire à conserver,** une fois compressé, pour la **plateforme Illumina**. Pour les plateformes Life Technologies et Roche (454), le fichier fastq **ne peut pas être considéré comme une donnée primaire complète** en raison de l'impossibilité d'y

[1] Le « *hard clipping* » consiste à supprimer des fichiers d'alignement les bases non alignées des reads entrainant ainsi la perte définitive de l'information. Il est en opposition au « *soft clipping* » qui informe que le read n'est pas complètement aligné sans supprimer les bases pour autant.

stocker les données d'intensité. Ces données étant indispensables pour certaines étapes (p. ex « *Variant Calling* »). Les fichiers **SFF** quant à eux peuvent être considérés comme **une donnée primaire** pour ces plateformes. Nous recommandons donc de **conserver ces fichiers une fois compressés**

- Les fichiers **sam** sont des fichiers d'alignement qui ont une taille conséquente (2To par run). Il s'agit de fichiers intermédiaires, produits à partir des fichiers fastq, ils seront transformés en fichiers bam par la suite. Nous recommandons de **ne pas conserver ces fichiers**.
- Les fichiers **bam** ont une taille intermédiaire, 512Go par run (8Go/exome). Ces données secondaires sont utilisés à la fois pour visualiser les alignements par l'utilisateur final et produire des fichiers de variants plus globaux. Ce type de fichier contient des informations similaires au fichier sam dans un format plus optimal. Nous recommandons donc de **conserver ces fichiers**. Pour la plateforme Illumina, si le fichier fastq est conservé, le fichier bam peut être supprimé dès lors que l'analyse est finalisée.

  Comme indiqué précédemment, les fichiers bam peuvent également constituer une alternative efficace pour stocker les données I (on parlera souvent de « bam non aligné »). Dans ce cas, le fichier doit impérativement contenir les séquences alignées sans hard-clipping ainsi que les séquences non alignées.
- Les fichiers **VCF** (données secondaires) sont de petites tailles, 10 Mo par échantillon (<1Go pour les gVCF) (cas d'un exome). Il s'agit de fichiers contenant la liste des variants qui sont régulièrement utilisés pour comparer des échantillons. Conserver l'ensemble des variants détectés à un instant donné permet d'assurer une partie de la traçabilité des résultats et permettra de revenir sur les résultats en fonction de l'évolution des connaissances. Ces fichiers VCF peuvent également être complétés par des annotations fonctionnelles. Les fichiers gVCF permettent d'identifier les régions non séquencées et pouvant expliquer l'absence de variations dans une région donnée. Nous recommandons donc de **conserver ces fichiers**.
- Selon les protocoles bioinformatiques les fichiers VCF peuvent correspondre aux fichiers finaux (et contenir les annotations fonctionnelles nécessaires pour l'interprétation biologique des données). Néanmoins d'autres fichiers comme des fichiers **TSV** (ou tout autre extension) représentent les fichiers finaux (variants et annotations). Ce sont des fichiers de taille négligeable facilement compressibles. Nous recommandons donc de **conserver ces fichiers**.
- Les paramètres d'analyses et les versions des sources de données/programmes doivent être gardés pour assurer la traçabilité. Les fichiers bam et VCF permettent par exemple de stocker dans leur entête ces informations. Nous recommandons donc de **conserver ces fichiers** à ce titre.

- Résumé des fichiers à conserver à moyen ou long terme

	Roche		Illumina			Ion Torrent[a]						
Type et taille des fichiers générés par run	**Type**	**Junior**	**Type**	**HiSeq[c]**	**NextSeq[d]**	**MiSeq**	**Type**	**Proton[b]**	**S5XL 520[e]**	**S5XL 530[e]**	**S5XL 540**	**PGM 318a**	**PGM 316a**	**PGM 314a**
Séquences (b=base)		100Mb		600Gb	40-120Gb	8Gb		10Gb	2Gb	4Gb	15Gb	1Gb	100Mb	40Mb
Données I (séquences)	SFF	450Mo	FastQ compressé	600Go	50Go	8,5Go	SFF	112Go				11Go	1,1Go	0,5Go
Données II (alignements)									55Go	75Go	~85Go			
Données II (alignements)	BAM	15Mo		512Go	10Go	10Go		35Go	10Go	25Go	~55Go	3,6Go	0,36Go	0,2Go
Données II (variants[d])	Txt	3Mo	VCF	0,64Go	0,01Go	0,01Go		0,12Go				0,01Go	0,01Go	1Mo
TOTAL		**~500Mo**		**1,2To**	**60Go**	**19Go**		**150Go**	**65Go**	**100Go**	**~145Go**	**~15Go**	**~1,5Go**	**~0,7Go**

Table 4 Récapitulatif des tailles approximatives des fichiers pour les principaux séquenceurs Illumina, Roche et Ion Torrent. [a]en utilisant la Torrent Suite Software 3.6 et le kit 200 bp. [b]en utilisant la Torrent Suite Software 3.6 et le kit 400 bp. [c]Les données sont proposées pour un run soit en moyenne 64 exomes (pour 1 exome le SAM représente 30Go et le BAM 8Go). [d]La taille des vcf est dépendante du nombre de variants et donc du type d'application utilisée et selon le cas des annotations ajoutées. Les données présentées sont des estimations basées sur des données et des performances à un moment donné.

Ainsi selon les générations et les technologies de séquenceurs, le gain en terme de stockage informatique peut aller d'un facteur 2,5x à >40x.

5.4 Flux de données et réseau

Dans la majorité des cas, il n'est pas possible, voire non recommandé de réaliser les analyses secondaires des données primaires issues du séquençage directement sur le PC pilotant le séquenceur.

L'archivage à long terme des données de séquence ne peut être réalisé sur le séquenceur ou sur le PC fourni par le constructeur du séquenceur. Les données de séquence doivent donc être transférées et stockées de préférence sur un serveur disposant d'un stockage sécurisé, permettant par la suite d'effectuer les opérations d'analyse et/ou d'archivage à long terme des données. Plusieurs moyens sont envisageables pour ce transport incluant l'utilisation de périphériques USB (type disque dur transportable), néanmoins l'utilisation d'un réseau filaire est l'option la plus courante.

Au niveau du transfert, au regard du volume à transférer et du délai acceptable pour ce transfert (Table 5), il est donc nécessaire de vérifier la structure du réseau local pour s'assurer que les flux de données seront maitrisés. A titre d'exemple, pour un HiSeq 2500, les 600Go de fichiers fastq vont être transférés à la fin du run d'une seule traite à travers le réseau. Si le réseau n'est pas correctement dimensionné, le transfert risque d'être très lent ou parfois peut même induire un arrêt du fonctionnement du réseau (et donc d'éventuelles pertes d'information).

Données primaires à transférer pour 1 run	Temps de transfert sur un réseau 100Mb	Temps de transfert sur un réseau 1Gb	Temps de transfert sur un réseau 10Gb
HiSeq2000 (600Go)	13h20min	1h20min	8min
MiSeq (8,5Go)	~11min	1min	6,8sec
Proton (112Go)	2h29min	~14min	~1min30sec
PGM (1,1Go)	1min28sec	8,8sec	~1sec

Table 5 Durées théoriques des temps de transfert des données primaires à travers les réseaux basé sur le fait qu'1 octet = 8 bits donc un taux de transfert de 100Mb/s = 12.5 Mo/s, 1Gb =125Mo/s, 10Gb = 1,25Go/s. Les temps de transfert indiqués sont théoriques, et ne prennent pas en compte les pertes inhérentes à la technologie et le trafic réseau « autre » généré par le réseau de votre institution (accès internet pour les usagers, etc. ...)

Au niveau du stockage à long terme, le volume de stockage, la vitesse d'accès aux données, la sécurisation des données doivent être dimensionnées aux besoins du laboratoire.

5.4 Vérification de l'intégrité des données transférées

L'intégrité de la copie des fichiers d'un support à un autre (via un réseau ou via un disque dur externe) doit être contrôlée à chaque transfert. Cela se fait à l'aide de sommes de contrôles (*checksum*). Il s'agit de comparer les empreintes des fichiers avant et après copie pour s'assurer qu'aucune différence n'est apparue pendant la copie[2]. Certains outils ou protocoles garantissent ce point.

6 LES INFRASTRUCTURES HARDWARE

6.1 Stockage

Dès lors que l'estimation des volumes et des flux de données est réalisée, il est important de distinguer le stockage de l'archivage. Le stockage et l'archivage ont des niveaux de sécurité, de disponibilité et de performance différents.

La sécurité des données correspond à la tolérance admise aux problèmes matériels pouvant impliquer des pertes de données. La disponibilité correspond aux délais pendant lesquels les données peuvent être

[2] Ces méthodes de vérification calculent l'empreinte d'un fichier en créant une clé ou *hash* unique (*checksum*) calculée avant et après la copie. Seules ces clés sont comparées à l'issue de la copie et permettent de garantir l'intégrité des données.

indisponibles suite à une panne. Une disponibilité faible implique un délai d'attente important afin de récupérer les données perdues. La performance correspond à la réactivité du matériel concernant les lectures/écritures des données sur le support.

Ces trois points influencent directement les coûts et ne sont pas tous compatibles entre eux.

- le stockage propose généralement un niveau de sécurité moyen, avec un niveau de disponibilité élevé et une performance moyenne à élevée.
- L'archivage propose un niveau de sécurité fort, une disponibilité faible et des performances faibles.

Ainsi, pour la même quantité de données, le stockage est généralement plus cher que l'archivage.

Les disques de stockage permettent d'accéder aux données afin d'effectuer les différentes étapes d'analyse des séquences. Ils sont rapides pour permettre un accès simultané à une plus grande quantité de données et redondants pour assurer la sécurité de celles-ci.

La technologie utilisée pour assurer la sécurité et la disponibilité est le RAID (*Redundant Array of Independent Disks*). Cette technologie accepte la perte d'1 ou plusieurs disques durs selon le mode séléctionné. Néanmoins cette solution a un cout plus élevé en raison, d'une part, de la qualité supérieure des disques utilisés (résistance aux pannes) et d'autre part, pour le nombre de disques plus important à stockage égal pour la mise en place de la solution (ex : un RAID5 avec 4 disques de 1To permettront de stocker 3To d'espace utilisable et non 4). Il est important de considérer que le stockage, même en RAID5[3] ou RAID6[4], n'est en aucun cas une solution d'archivage. Malgré la redondance, les données ne sont pas sécurisées et la perte de 2 (RAID5) ou 3 (RAID6) disques durs peut entraîner la perte de l'intégralité des données.

L'archivage ou sauvegarde (backup) qui a pour but de répliquer les données avec une possibilité de restauration en cas de perte est souvent réalisées sur des bandes magnétiques. Par opposition aux médias de stockage, les médias d'archivage sont moins performants en termes de vitesse, et ne permettent donc pas d'effectuer des opérations d'analyse. Ils sont cependant plus fiables, redondants car peu couteux et souvent délocalisés. Concernant les sauvegardes ou archivages, il est important de définir leur fréquence (ex : 1x par semaine et le nombre de sauvegardes conservées). La fréquence des sauvegardes correspond à ce que l'utilisateur accepte ou non de perdre en cas d'incident. Dans l'idéal, il faut sauvegarder toutes les données que vous avez décidé de conserver : fastq, bam et vcf. La donnée primaire doit être sauvegardée, car elle permet de régénérer tout le reste des données. L'espace d'archivage doit être évolutif à moindre coût car la capacité d'archivage doit continuer à augmenter au fur et à mesure de la génération des données.

Les infrastructures de stockage et d'archivage sécurisées, de hautes disponibilités et de hautes performances, demandent un investissement important tant au niveau **des équipements que du personnel**.

Remarques additionnelles :

Il est important de noter qu'au cours de l'analyse, un nombre considérable de fichiers temporaires sont générés (ex : les fichiers sam). Ces fichiers seront effacés par la suite, cependant il est nécessaire de prendre en compte un espace dit « **tampon** » dans lequel seront faites ces analyses. Cet espace sera vidé régulièrement après la fin des analyses bioinformatiques et la validation des données de qualité. Cet espace est d'autant plus important que le volume à considérer est non négligeable du fait qu'une analyse est souvent réalisée en parallèle pour plusieurs patients/échantillons. La taille de ce tampon correspond à la taille totale de vos données (fastq + sam + bam + vcf) multipliée par le nombre d'échantillon maximal que vous analyserez en simultané.

[3] Le RAID5 permet au système de stockage de conserver l'intégralité des données malgré la perte d'un disque.
[4] Le RAID6 permet au système de stockage de conserver l'intégralité des données malgré la perte de deux disques

6.2 Calcul

Le protocole bioinformatique d'analyse des données issues des expériences de séquençage à haut débit fait largement appel à des programmes développés par la communauté scientifique qui sont pour la grande majorité disponibles pour des systèmes d'exploitation de type UNIX/Linux. La nature des analyses (plusieurs patients séquencés en parallèle) et la capacité des logiciels permettent dans la plupart des cas une parallélisation des analyses bioinformatiques. Ainsi la ou les machines de calcul doivent être multi CPU et multi cœur.

Prenons l'exemple de l'analyse d'un exome sur une machine type (1 CPU 8 cœurs à 2,5 GHz/ 32Go de RAM sous Unix/Linux), dont l'analyse représente 98h de temps de calcul sur un seul CPU ou bien 22h en utilisant la capacité de la machine à pouvoir exécuter des processus en parallèle (Table 6).

Etape du protocole	Parallélisation	Durée
QC totaux		2h
Alignement local	2x4CPU	6h
Alignement global	1xCPU	4h
Clean / sort / dedup	1xCPU	4h
Indel realignement	8xCPU	2h
Recalibration	1xCPU	1h
Recherche de variants	8xCPU	2h
Annotation (en ligne)		1h
Total		**22h/8CPU**
		98h/CPU

Table 6. Tableau recensant les principales étapes et leur durée sur un serveur type pour l'analyse d'un exome avec le protocole d'analyse standard GATK.

Dans ce cas précis nous avons rapporté le résultat pour un seul exome, néanmoins en pratique l'analyse concerne plusieurs exomes en simultané issus du séquençage de plusieurs échantillons. Dans le cas de la machine décrite précédemment, l'analyse de 8 exomes nécessiterait une semaine d'analyse (22 heures x 8 = 176h, soit 7,3 jours). Il peut donc être intéressant en fonction du volume de données à analyser de multiplier les machines d'analyse afin d'optimiser au mieux le temps de calcul. Ainsi, si le nombre de machines était de 8 au lieu d'une seule, les 8 exomes pourraient être traités en 22 heures au lieu d'une semaine.

L'intérêt de la parallélisation est donc démontré, mais il est nécessaire de mesurer le coût des machines par rapport au débit et à l'efficacité d'analyse voulue.

A titre d'exemple une configuration type pourrait comporter les caractéristiques suivantes :

- un serveur de type lame ou rack, qui permet d'intégrer une infrastructure déjà existante,
- un ou plusieurs CPU multi-cœurs, ayant une fréquence importante (1,5 à 3GHz) et de 2 à 24 cœurs par CPU. ; c'est cela qui va permettre la parallélisation de certains algorithmes,
- un volume important de mémoire vive ou RAM. La capacité doit être proportionnelle au nombre de cœur, entre 3 et 6Go/cœur minimum, et avoir une fréquence élevé (>1066 MHz),
- des disques durs en RAID5, pour la stabilité du système et la redondance des données,
- un contrôleur SAS, pour l'attachement des disques de stockage des données,

- un système d'exploitation basé sur UNIX, pour plusieurs raisons : fiabilité, stabilité, souplesse, performance et compatibilité avec les logiciels disponibles.
- La capacité de gestion à distance de l'administration des serveurs de calcul (carte IMPI).

L'hébergement des serveurs de calcul doit se faire dans une salle dans laquelle la puissance électrique et la climatisation sont adaptées. Les serveurs doivent être branchés sur des prises ondulées.

7 SOLUTIONS EXISTANTES

Il existe plusieurs solutions adaptées aux différentes situations de chaque laboratoire, en fonction de la taille de votre laboratoire, au débit du séquenceur, et de l'infrastructure de l'institution. Ces solutions sont les suivantes :

7.1 Solutions locales

7.1.1 Infrastructures fournies par les fabricants de séquenceurs et NAS

Une première solution consiste à se contenter des équipements informatiques fournis par le constructeur du séquenceur et à y adjoindre une solution de stockage pour pallier le manque d'espace sur ces serveurs et/ou la gestion des sauvegardes.

Certaines compagnies produisant des séquenceurs fournissent des infrastructures informatiques avec leur séquenceur, généralement sous forme de serveur lié au séquenceur. Certaines de ces infrastructures sont comprises dans l'achat des séquenceurs tandis que d'autres sont en plus. Le prix et les solutions proposées varient suivant la compagnie et le séquenceur. Voici un tableau récapitulatif de ce que proposent les compagnies :

Instrument	Type de ressources informatique	Coût en k$
Illumina MiSeq	cloud	inclus
Illumina NextSeq 500	Cloud/cluster	Inclus/50
Illumina HiSeq 2500	cluster[a]	222
Ion Torrent – PGM	desktop[b]	16,5
Ion Torrent – Proton	cluster[c]	75
PacBio RS	cluster	65
Roche 454 GS Jr.	desktop	5
Roche 454 FLX+	desktop	5
SOLiD – 5500xl	cluster	35

Table 7. Solutions proposées par les constructeurs de séquenceur à haut-débit. [a]Ordinateur Illumina, système au niveau 1, 3xCPU 8 cœurs, 144Go de RAM partagés et ~24To de stockage. [b]Dekstop signifie des ordinateurs de bureau haut de gammes équipés de plusieurs processeurs, au moins 8Go de RAM, au moins 1To de stockage. [c]Cluster fourni par Life Technologies, 2xCPU 8 cœurs, 128Go de RAM, 2xGPU et ~ 27To de stockage.

Cette solution reste viable pour un débit faible, type Ion Torrent PGM et Illumina MiSeq. Dans certains cas ce type de machine n'est pas fourni, alors l'achat d'un serveur (ou « gros » PC) au format tour est possible à des prix raisonnables (2000~3000 €) avec des spécifications permettant d'analyser ce type de données.

Il faudra cependant obligatoirement investir dans une solution de stockage réseau pour archiver les données car l'espace de stockage sur ces machines reste faible et ne peut assurer à la fois le calcul, le stockage à court/moyen terme et l'archivage des données.

Il peut sembler simple et peu coûteux (~5000 €) de s'équiper d'un NAS (*Network Attached Storage*) pour pallier à ces problèmes de stockage. Le NAS est un boitier de stockage autonome, qui peut être branché directement

sur un réseau ou attaché à une machine, pouvant stocker jusqu'à plusieurs To de données et supportant le RAID pour une sécurité moyenne des données. Il est possible de passer à une sécurité et une disponibilité forte en doublant ce système et en délocalisant géographiquement le doublon. La solution obtenue peut servir d'archivage mais à un coût élevé. Les évolutions de ce type de matériel sont limitées et ne permettent pas de suivre l'augmentation de la demande en séquençage haut débit.

Avantages	Inconvénients
Temps d'analyse adapté à la technologie	Convient pour un débit limité seulement
Facile et rapide à mettre en place	Temps d'analyse long
Pas besoin d'infrastructure informatique particulière (salle machine, réseau 10Gb, etc ....)	Stockage limité en taille maximale
Pas besoin de personnel IT qualifié	Solution couteuse au final

Table 8. Avantages et inconvénients des solutions fournies + NAS.

7.1.2 Solution de calcul et stockage intégrées au système informatique de l'institution.

Une solution couramment utilisée est la mise en place d'une infrastructure complète incluant : le calcul, le stockage et l'archivage des données au sein de l'infrastructure informatique existante de l'institution. Ainsi l'investissement se composera de machines de calculs dédiées avec des spécifications suggérées précédemment dans le document qui seront intégrées dans la salle serveur de l'institution les accueillant, d'une baie de stockage pouvant accueillir les données à des fins de calcul et enfin une solution d'archivage pour la sauvegarde des données.

Cette solution est coûteuse à mettre en place dans sa globalité (pour l'institution) mais permet de grandement mutualiser les coûts pour les bénéficiaires (les laboratoires par ex). Il est ainsi possible de dimensionner au mieux les besoins et de les faire évoluer dans le temps. On notera cependant la nécessité de disposer d'un personnel technique hautement qualifié pour maintenir cette structure. Ce personnel devra être compétent dans la mise en place d'un logiciel de type « gestionnaire de ressources pour le calcul parallèle » (TORQUE, PBS, SGE, slurm, OAR, Maui). Ce type de logiciel permet de gérer l'ordonnancement des calculs en fonction des ressources de calcul disponibles (mémoire et CPU).

Prenons en exemple une structure informatique avec 4 CPUs à 16 cœurs à 2,5 GHz, avec 256Go RAM, 3 disques dur en RAID5, 1 carte contrôleur SAN, un système d'exploitation Linux et une maintenance de 3 ans. Le coût d'une telle structure est estimé autour de 23k€, auquel il est nécessaire d'ajouter le salaire d'un administrateur Système/Réseau (~50k€) qui peut être mutualisé.

Remarques additionnelles :

Dans cet environnement, il est possible de déployer des machines virtuelles, c'est à dire l'illusion d'un appareil informatique créée par un logiciel d'émulation. Le logiciel d'émulation simule la présence de ressources matérielles et logicielles telles que la mémoire, le processeur, le disque dur, voire le système d'exploitation et les pilotes, permettant d'exécuter des programmes dans les mêmes conditions que celles de la machine simulée. Un des intérêts des machines virtuelles est de pouvoir s'abstraire des caractéristiques de la machine physique utilisée. Cette solution permet notamment au sein d'une grande institution de mutualiser les coûts et de répartir au mieux les ressources. Les machines virtuelles peuvent également être sauvegardée dans différents états (initial, mois après mois…) ce qui permet également de les restaurer en cas de problème majeur.

Avantages	Inconvénients
Solution peu couteuse si la structure est existante	Solution couteuse dans sa globalité
Temps d'analyse court	Mise en place difficile
Stockage et calcul théoriquement illimités	Personnel IT qualifié indispensable
Gère facilement un débit important	Nécessité d'infrastructure informatique préexistante
Adaptable si les besoins évoluent	Nécessite une infrastructure réseau dédiée entre les séquenceurs et la solution informatique

Table 9. Avantages et inconvénients des solutions de calcul achetées.

Cette solution est préconisée en cas de débits importants pour des séquenceurs du type HiSeq2500, Solid etc...

7.2 Solutions externalisées

Le journal officiel du 6 juin 2010 défini le « *Cloud Computing* » comme « un mode de traitement des données d'un client, dont l'exploitation s'effectue par l'internet, sous la forme de services fournis par un prestataire. » C'est « une forme particulière de gérance de l'informatique, dans laquelle l'emplacement et le fonctionnement du nuage ne sont pas portés à la connaissance des clients. » Il convient de distinguer nuage public, nuage privé (dédié à une organisation) et nuage communautaire (partagé par plusieurs organisations identifiées).

L'Agence nationale de la sécurité des systèmes d'information (ANSSI) sous l'action du Centre opérationnel de la sécurité des systèmes d'information (COSSI) a émis un certain nombre de recommandations pour le bon déroulement de l'utilisation d'une solution d'externalisation de solution informatique (http://www.ssi.gouv.fr/externalisation). Le recours à l'externalisation est une pratique courante qui présente un certain nombre d'avantages, mais également des risques qu'il convient d'évaluer avant de choisir une prestation de ce type. Comme le rappelle le COSSI, « il convient à cet égard de ne pas opposer sécurité et externalisation ». L'utilisation d'une solution externe peut être une excellente solution pour pallier à l'absence ou l'insuffisance de solutions par l'institution, à condition que l'évaluation des risques (via le Plan d'Assurance Sécurité) ait été faite et que le prestataire s'y engage. Le risque est évaluable à la fois pour le système d'information et pour les données (intégrité, disponibilité, confidentialité). Le plan d'assurance sécurité décrit l'ensemble des dispositions spécifiques à prendre pour garantir le respect des exigences de sécurité du donneur d'ordre (voir références).

Parmi les différents points importants on notera par exemple pour :

- le transfert des données : l'intégrité et la confidentialité des données. Le temps de transfert des données.
- l'exploitation : la localisation du service et des données (loi du 6 janvier 1978 modifiée, notamment la localisation des données au sein de la communauté européenne). Sécurisation des moyens informatiques. Mutualisation des ressources informatiques par d'autres applications (ex un serveur est utilisé pour l'analyse de données de santé ainsi que des applications bancaires).
- la réversibilité : restitution des données (absence de copie résiduelle dans le cloud).

Parmi les solutions externalisées on distinguera les centres de calcul que l'on peut qualifier de *cloud* privé, des fournisseurs de solutions commerciales que l'on qualifie tout simplement de *cloud public* ou *cloud*.

7.2.1 Centre de calcul ou cloud privé

Un certain nombre de centres de calculs sont disponibles dans toute la France et dans certains DOM-TOM (Figure 3). Ils sont accessibles de partout, et proposent à des tarifs très avantageux de disposer de temps de calcul et d'espace de stockage en quantité et qualité importante, permettant aussi la mutualisation des moyens de calcul et donc une utilisation plus optimisée.

Prenons en exemple le centre de calcul de Dijon, il s'agit du 9ème centre de calcul de France, avec plus de 40 Tflops. Il possède deux salles de machines, 2.5 ETP + 2.5 ETP de messagerie, 100 utilisateurs de recherche, avec plus de 1000 heures de CPU/an, 600 comptes enseignements, plus de 3000 cœurs de calcul et 10To de mémoire RAM. Pour 20To d'espace d'archivage (*Network File Ssytem*), 5To d'espace work et 2To d'espace permanent (*NFS*) et 2 000 000 d'heures de calcul, comptez environ 4000€ par an.

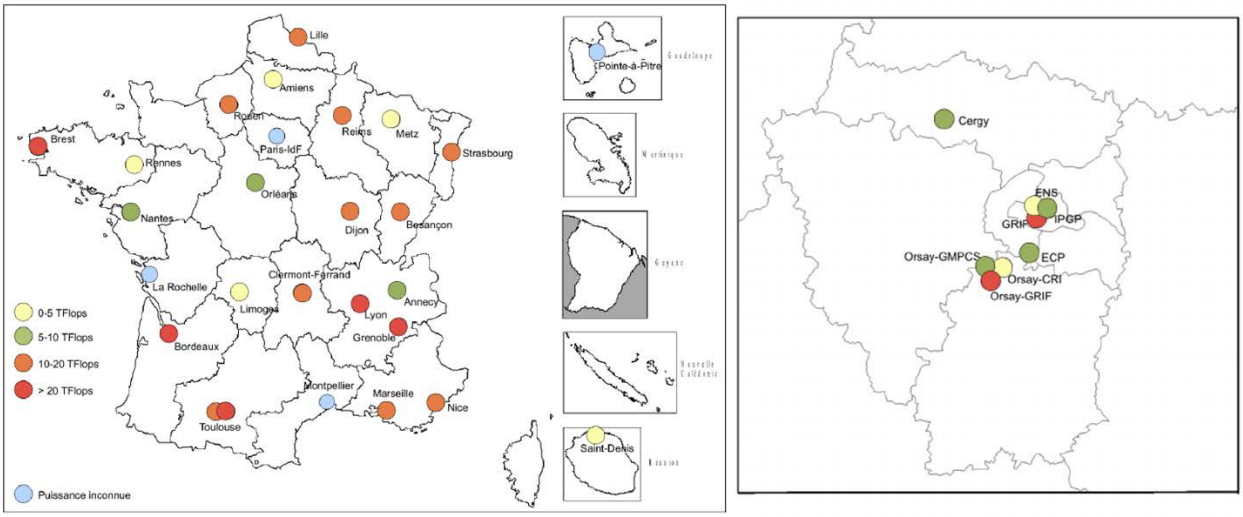

Figure 3. Localisation des centres de calculs en France.

Avantages	Inconvénients
Solution très peu chère	Structure académique de recherche n'intégrant pas toujours la problématique de la disponibilité des ressources de calcul à visée diagnostique
Pas besoin de personnel IT qualifié	Mise à disposition ingale des ressources et dépendante des utilisateurs.
Mutualisation des ressources	Nécessite l'anonymisation complète des données ou bien la structure doit être hébergeur de santé
Mise en place rapide après configuration de la solution	N'inclus ni le stockage, ni l'archivage

Table 10. Avantages et inconvénients de la solution du centre de calcul.

7.2.2 Le cloud public

Les principaux fournisseurs de *cloud* sont Amazon, ITS Integra, ASPSERVEUR, Google, IBM, Intrinsex, Orange Business Services, OVH et SFR Business Team. Ces fournisseurs peuvent vous fournir une puissance de calcul très importante à la demande, et il en est de même pour le stockage. La mise à disposition des machines est quasiment immédiate après la demande. Les tarifs sont variables et si la solution est viable pour le calcul, elle reste très onéreuse pour le stockage dû à un système de paiement par mois par volume cumulatif, rendant la solution rapidement très chère pour le stockage à long terme.

Par exemple les tarifs chez Amazon IC2 pour une machine équivalente à celle citée dans la section précédente est de 0,26 €/heure pour le calcul, et de 150 €/To/mois pour le stockage.

Le prix du stockage à long terme devient donc très important à partir du moment où un volume de données archivées occupe vos disques.

Remarques additionnelles :

Le problème du rapatriement de l'intégralité des données au bout de quelques années en cas de rupture de contrat doit être posé. Le transfert de données volumineuses nécessite de prévoir une bande passante adaptée. Il faudra également s'assurer que la limite de quantité de données transférées soit en adéquation avec la volumétrie. Les solutions reposent généralement sur des lignes télécom dédiées.

Avantages	Inconvénients
Temps d'analyse court	Solution couteuse pour le stockage
Stockage et calcul illimités	Sécurité et confidentialité des données à vérifier
Gère facilement un débit important	Statuts juridiques encore flous
Adaptable très rapidement	Dépendance d'Internet pour accès et utilisation
	Besoin d'une bande passante adaptée
	Nécessite un hébergement en Europe
Pas besoin de personnel IT qualifié	Les données doivent être complètement anonymisées ou le service de cloud doit être accrédité hébergeur de données de santé
Accessibilité de partout	Rapatriement des données à la rupture du contrat

Table 11. Avantages et inconvénients des solutions de type cloud.

8 Perspectives du Big Data

Le « *Big Data* » est une expression générique qui fait référence à la fois aux outils et aux procédures permettant la création, la manipulation, la visualisation et la gestion de grosses quantités de données dans un environnement hautement distribué (ressources informatiques partagées). L'utilisation du « *Big Data* » dans des domaines variés (gestion des risques, météorologie, physique, biologie, épidémiologie, médecine, etc.) est considérée aujourd'hui comme un enjeu majeur permettant de répondre à des questions complexes difficilement résolubles avec des outils non adaptés.

8.1 Pourquoi pour la génomique ?

L'évolution des technologies de séquençage ces dernières années a permis d'augmenter d'un facteur 10 000 la quantité de données générées. Cette évolution technique s'est également accompagnée d'une baisse importante des coûts du séquençage mais également d'une augmentation des coûts et des problématiques informatiques.

Afin d'être utilisées dans un cadre diagnostique, les données de séquençage doivent être analysées le plus efficacement possible et souvent le plus rapidement possible pour en extraire les variations observées (variant calling) et ainsi générer un fichier VCF. Ces contraintes fortes sont difficiles à satisfaire avec les formats de données (fichiers SAM/BAM) et les pipelines actuels, orientés vers des processus non hautement distribués. Pour pallier à ces problèmes, l'Université de Berckley, par exemple, propose depuis 2013 *ADAM* qui regroupe des formats de fichier, un framework de programmation et des outils adaptés aux environnements fortement distribués. Cette solution permet à l'étape de « *Variant Calling* » d'être 50x plus rapide sur un fichier de

séquençage de génome entier de 250GB pour une couverture de 60x (utilisation d'un cluster de 100 nœuds de calcul). Il s'agit d'un cas concret d'utilisation du « big data » pour la génomique.

8.2 Perspectives

Considérant l'arrivée plus ou moins rapidement du séquençage complet du génome de patients, il est difficilement envisageable de concevoir du séquençage génomique généralisé sans disposer d'outils orientés « *Big Data* ». L'utilisation efficace du « *Big Data* » doit se faire dans un environnement hautement distribué (calcul et stockage) nécessitant une infrastructure informatique dimensionnée. Les solutions externalisées et locales décrites précédemment, avec les avantages et inconvénients correspondants, peuvent répondre à ce type de contraintes. Il faudra cependant s'assurer que l'infrastructure informatique est compatible avec ces approches distribuées en disposant de compétences adéquates au sein des services informatiques et bioinformatiques tout en s'assurant que la réglementation en vigueur pour les données de santé et en particulier la génétique soit respectée..

9 REFERENCES

- Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. **The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants.** Nucleic Acids Res. 2010 Apr;38(6):1767-71
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R; 1000 Genomes Project Analysis Group. **The variant call format and VCFtools.** Bioinformatics. 2011 Aug 1;27(15):2156-8.
- Glenn TC. **Field guide to next-generation DNA sequencers.** Mol Ecol Resour. 2011 Sep;11(5):759-69.
- Greene, Casey S. and Tan, Jie and Ung, Matthew and Moore, Jason H. and Cheng, Chao. **Big Data Bioinformatics** J. Cell. Physiol. 229: 1896–1900.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. **The Sequence Alignment/Map format and SAMtools.** Bioinformatics. 2009 Aug 15;25(16):2078-9
- Lipman DJ, Pearson WR. **Rapid and sensitive protein similarity searches.** Science. 1985 Mar 22;227(4693):1435-41
- Massie M, Nothaft F, Hartl C, Kozanitis C, Schumacher A, Joseph AD and Patterson DA. **ADAM: Genomics Formats and Processing Patterns for Cloud Scale Computing** UCB/EECS-2013-207.
- Sboner A, Jasmine Mu X, Greenbaum D, Auerbach RK and Gerstein MB. **The real cost of sequencing: higher than you think!** Genome Biology 2011, 12:125
- Cours de Y. Duffourd. DU « Séquençage haut débit et maladies génétiques »
- Site web du cloud Amazon : http://aws.amazon.com/fr/ec2/
- Centre de calcul de l'université de Bourgogne : https://haydn2005.u-bourgogne.fr/dsi-ccub/
- Life Technologies, Torrent Suite™ Software 3.6.2 USER GUIDE AND ADMINISTRATION GUIDE Revision Date July 2013
- Illumina, CASAVA 1.8.2 Quick Reference Guide
- Wikipedia
- Hébergement (territoire) : http://esante.gouv.fr/services/referentiels/securite/hebergement-faq#13
- Hébergement (hébergeur données de santé) : http://esante.gouv.fr/services/referentiels/securite/hebergement-faq#22
- http://www.ssi.gouv.fr/externalisation,
- http://www.ssi.gouv.fr/IMG/pdf/2010-12-03_Guide_externalisation.pdf
- https://www.ietf.org/rfc/rfc1952.txt
- Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés