Le Projet Zenome : Livre blanc
Écosystème Génomique basé sur la "Blockchain"

Nikolay Kulemin Sergey Popov Alexey Gorbachev

6 Octobre 2017

Traduction française par Eric Wursteisen, janvier 2018

Résumé

"Industrie 4.0" est la dénomination de la tendance actuelle d'automatisation, de mise à l'échelle et d'échange de données dans les technologies manufacturières. Elle inclut l'intelligence artificielle, la réalité virtuelle, l'internet des objets et l'analyse du "Big Data". La génomique est une représentation réelle de cette Industrie 4.0 qui requiert la résolution de problèmes urgents, comme le stockage et l'analyse des "Big Data" tout en gardant un accès public pour les chercheurs et gardant le respect de la confidentialité du particulier.

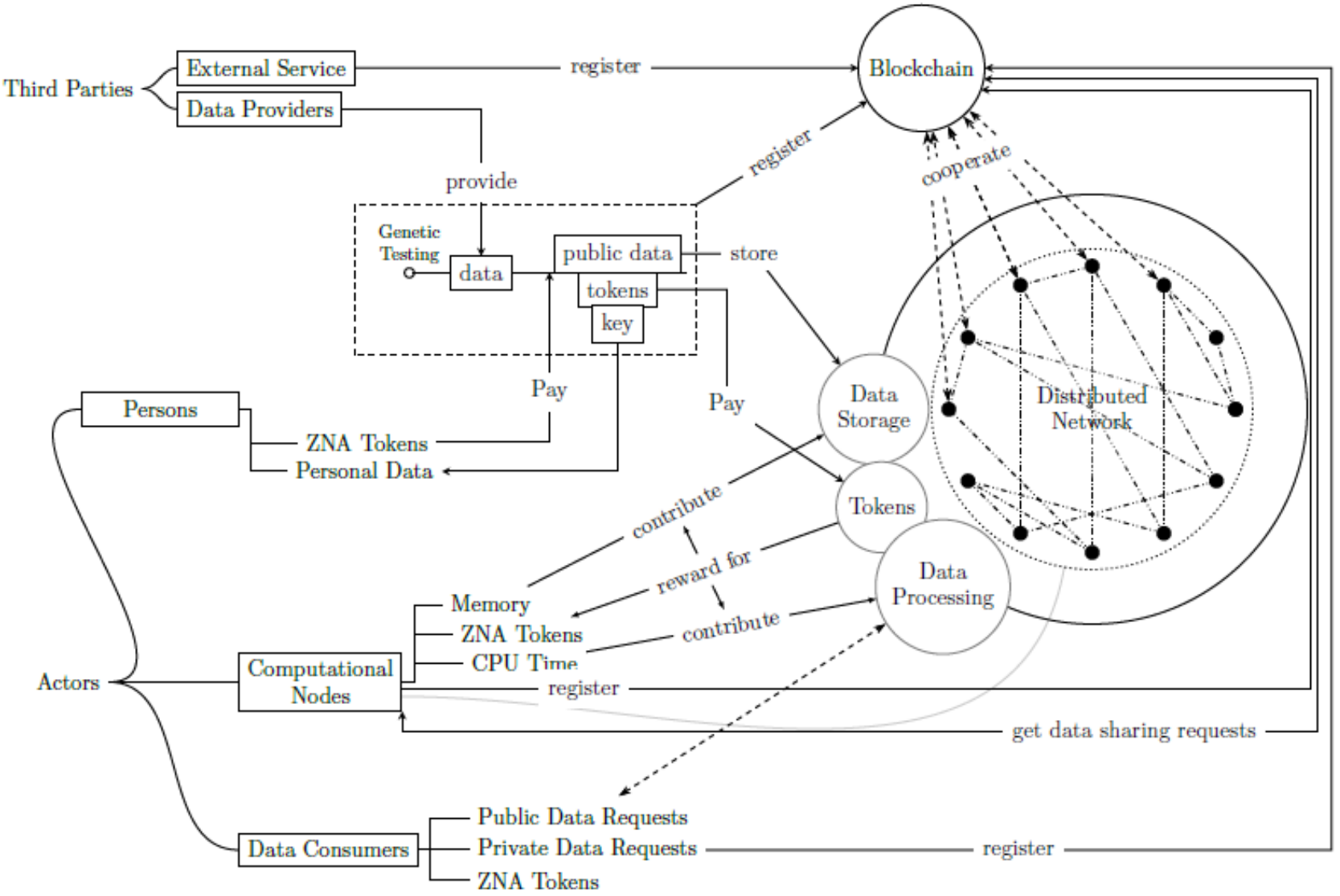
Pour l'instant il y a un problème d'inégalité dans l'industrie du génome. Les données génomiques personnelles sont concentrées dans des centres de données des organisations oeuvrant dans le génome, les gouvernements, les institutions scientifiques et médicales et les compagnies pharmaceutiques. Il y a également des limitations légales d'accès aux données génomiques personnelles en plus d'une absence de possibilité de partager et gérer ces données. Ce monopole de la donnée génomique empêche dramatiquement l'avancée de travaux pour de nombreux domaines scientifiques et médicaux.

Le développement de monnaies virtuelles et de technologies basées sur la "blockchain" ouvre la voie pour des transformations significatives dans de nombreux domaines économiques. Appliquer l'approche de la "blockchain" est un garde-fou permettant l'amélioration du développement de la génomique pour les particuliers. Cela rendra chaque personne propriétaire de ses propres données génomiques.

Le Projet Zenome est une base de données décentralisée d'information génomique utilisant la "blockchain". Cette plateforme supporte la possibilité de gérer vos données génomiques tout en maintenant le respect de vos données personnelles et vous offrant l'opportunité de capitaliser financièrement vos données en vendant un accès à différentes portions de votre génome. Cela va également établir des conditions égales pour le développement de nouveaux médicaments et le progrès de technologies scientifique et médicales.

Zenome est un nouvel environnement économique basé sur les données génomiques et la technologie "blockchain". L'implantation de notre modèle conceptuel résous les problèmes ci-dessous :

- Création d'une infrastructure pour stocker le "Big Data" génomique en utilisant une base de données distribuée
- Accès libre à des millions de génomes à travers le monde tout en ayant une protection de la confidentialité
- Opportunité pour chaque personne de participer à de la recherche clinique et scientifique et d'obtenir des revenus en échange
- Stimuler l'amélioration des sciences en génome dans les pays en développement et la démonopolisation des données génomiques dans les pays développés.


Third Parties
External Service
Data Providers
register
Blockchain
provide
register
cooperate
Genetic Testing
data
public data
tokens
key
store
Pay
Pay
Persons
ZNA Tokens
Personal Data
Data Storage
Distributed Network
Tokens
contribute
reward for
Data Processing
Memory
ZNA Tokens
CPU Time
contribute
Actors
Computational Nodes
register
get data sharing requests
Public Data Requests
Private Data Requests
ZNA Tokens
Data Consumers
register

L'ÉQUIPE

Alexey Gorbachev — Fondateur

Biologiste moléculaire et passionné de "Blockchain"
Ph.D en biologie moléculaire et biochimie

Université : Université d'État de Moscou, Département de Biologie Moléculaire

Alexey a un bagage scientifique significatif, avec expertise dans le domaines des affaires et en gestion de projet. Dans Zenome, Alexey est responsable de la vision globale ainsi que du développement des affaires.

Email : alex@zenome.io
Linkedin : https://www.linkedin.com/in/alexey-gorbachev-24b5305b/

Nikolay Kulemin — Fondateur

Ph.D en bio-informatique

Université : Institut de physique et Technologie de Moscou, Département de physique moléculaire et biologique

Spécialiste en bio-informatique et biologie mathématiques, Nikolay a une expérience académique en recherche et en application de travaux dans l'industrie du génome. Nikolay est également fondateur d'une compagnie développant de nouveaux algorithmes pour l'analyse du génome.

Email : nick@zenome.io
Linkedin : https://www.linkedin.com/in/nikolay-kulemin-50ab47a9/

Vladimir Naumov — Scientifique, Analyste du Génome

Scientifique de données dans le génome humain

Université : Université Nationale de Médecine de Recherche de Pirogov

Spécialiste en analyse de données et bio-informatique. Ancien CSO à iBinom, scientifique en bio-informatique à GERO, 8 années d'expérience dans l'industrie génomique. Travaille sur la création de pipeline et développant de nouvelles façons dans l'analyse et la visualisation de la donnée génomique.

Email : vov@zenome.io
Linkedin : https://www.linkedin.com/in/vladimir-naumov-8285a25b/

Sergey Popov — Développeur

Blockchain, P2P, systèmes distribués, mathématiques pures

Université : Institut de Physique et Technologie de Moscou, Département des Physiques Générales et Appliquées

Sergey a de l'expérience en physiques théoriques, l'informatique conceptuelle, les mathématiques pures, le développement de systèmes distribués et les "smart-contracts".

Email : sp@zenome.io

Dmitry Kwon — Conseiller

Ph.D en biologie moléculaire, gestionnaire du développement des affaires

Université : Université d'État de Moscou, département de biologie moléculaire

Spécialiste en génétique moléculaire. Dmitry a un bagage scientifique important, une expertise dan les technologies d'analyses génétiques, les marchés du diagnostic et du génome et une expérience réussie en biotechnologie et avec des compagnies mondialement reconnues dans le diagnostic.

Email : dk@zenome.io
Linkedin : https://www.linkedin.com/in/dmitry-kwon-2763b119/

SOMMAIRE

L'ÉQUIPE

Tokenomique du Génome

Introduction :
La “Tokenomique” du Génome

Dans ce chapitre, une description générale de la plateforme Zenome va être présentée. Le besoin de tokens (jetons) et les bénéfices pour de futurs investisseurs sont discutés.

_____

La plupart du temps, l’information génomique est stockée dans des bases de données, financées par des gouvernements ou de grandes compagnies. Isolées, chaque base de données ne contient pas assez de données pour faire le saut nécessaire dans une ère de médecine génomique et de précision. En même temps, chaque base contient tellement d’information qu’il est impossible pour une seule compagnie de passer à travers l’ensemble de ces données.

Il apparaît donc que l’échange d’information génétique est d’importance capitale. Le futur marché de la génétique doit s’assurer de se protéger contre tout mauvaise utilisation et la discrimination génétique en particulier. Il est notamment important de maintenir la transparence et un accès égal à ce marché.

Un échange mondialisé de l’information génétique devrait adresser les problématiques suivantes :

- La fragmentation de la donnée génétique
- L’accès limité des scientifiques, médecins et compagnies à la donnée génétique
- Les tests génétiques à un coût abordable
- Le manque de protection de la confidentialité de ceux qui acceptent de donner un libre accès à leur donnée génétique
- Ressources informatique insuffisantes

Zenome vise à créer l'infrastructure de la génomique du particulier, qui permettra à chaque participant de :

- Mettre en ligne son information génétique et en prendre contrôle
- Stocker de manière sécuritaire sa propre information génétique
- Retirer des revenus de la vente des accès à la donnée génétique ou à certaines portions de celle-ci
- Effectuer des tests génétiques en échange des droits à l’utilisation de l’information génétique
- Obtenir des recommandations diététiques personnalisées ou un programme d'entraînement basé sur la constitution génétique
- Utiliser d’autres services génétiques

Les principaux clients de l’information génétique sont des compagnies qui sont intéressées par le ciblage génétique, tel que Google, Facebook, Unilever et les compagnies pharmaceutiques.

Au sein de la plateforme Zenome, différents type d’information, notamment génétique, personnelle et financière sont étroitement liées. La nature spécifique de chaque type détermine la manière dont cette information va être stockée. Les données financières, qui inclut notamment les enregistrements des données, sont stockées sur la “blockchain”. Les données anonymes du génome, sont stockées sur le réseau distribué. Les données personnelles du participant sont conservées sur leur propre ordinateur. Traiter la donnée de différentes manières permet de la sécuriser ainsi que de s’assurer de l’expansion du système.

Puisque toutes les données de transaction, incluant l’achat et la vente de données, sont gouvernées par des “smart-contracts”, ce qui reflète bien la nature décentralisée de la plateforme, les interactions ne peuvent qu’inclure des soldes stockés sur des “blockchains”. L’utilisation de token déjà existant pour ce type d’interaction

provoquerait une dépendance déraisonnable à propos de l'évaluation de ce token ou monnaie externe. Ainsi, un token doit être émis, supportant l'économie des interactions sur la plateforme. Cela, en particulier, implique qu'il est impossible d'acheter de l'information génétique avec de la monnaie "normale", il est nécessaire d'obtenir des tokens auparavant.

Zenome DNA (ZNA) est le token utilisé par la plateforme Zenome. L'évaluation de ce token est étroitement liée au succès de la plateforme.

Dans ce livre blanc, nous discuterons en plus grands détails les problématiques majeures dans le domaine de la génomique, autant que les solutions apportées par la plateforme Zenome.

Chapitre 1

La génomique

1.1 Contexte

Dans cette section, les définitions des termes "génome" et "génomique" sont données. L'histoire des premiers efforts dans le séquençage génomique et l'émergence de la technologie de séquençage à haut débit (NGS) sont considérées. Les principaux fabricants de réactifs et d'équipements utilisés pour obtenir de la donnée génomique sont décrits. Les problématiques en lien avec l'accumulation de donnée génomique et de la réduction des coûts des analyses sont considérées. Une revue des bases de données actuelles est fournie.

_____

Le Génome

Le génome est l'ensemble des instructions génétiques trouvées dans une cellule [1].
Le génome contient l'information biologique nécessaire pour le développement et le fonctionnement d'un organisme. Le génome humain est constitué de molécules d'ADN en double-hélice organisées en 22 paires de chromosomes et de deux autres chromosomes pour le sexe - X et Y. Toute l'information contenue dans le génome est cryptée en utilisant un code quaternaire à travers une séquence de 4 nucléotides désignés A, T, C et G. La terminologie "lire le génome" signifie "déterminer la séquence de nucléotides par un processus de séquençage" [2].

La séquence individuelle d'un génome défini une variété de fonctionnalités organiques, incluant l'apparence, la susceptibilité à certaines maladies, les habiletés athlétiques, le métabolisme, les préférences nutritionnelles, la compatibilité avec des partenaires sexuels (la possibilité de concevoir des enfants), et bien d'autres.

Le projet international du génome humain

Le Projet International du Génome Humain a été lancé sous la supervision du NIH (Institut National de la Santé) en 1990 pour déterminer le séquençage complet du génome haploïde humain. Le leader initial du projet était l'un des découvreurs de la structure de l'ADN, le prix Nobel James Watson.

Un premier séquençage du génome humain fut complété au milieu des années 2000 et publié au début de 2001 dans le journal Nature. Le coût de ce projet international, supporté par des fonds publiques est estimé approximativement à 3 milliards de dollars. En 1998, une compagnie privée, Celera Genomics, s'est joint à la course au séquençage humain. Le leader de ce projet privé, qui était développé en parallèle avec des institutions gouvernementales, était le célèbre scientifique et entrepreneur Craig Venter, qui a obtenu 300 millions de dollars en fonds privés pour le projet de Celera. En utilisant la nouvelle approche de séquençage "shotgun" et des méthodes de calcul plus productives, la séquence du génome de Craig Venter a été publiée pratiquement en même temps que le consortium international en 2001 [3] dans la journal Science. Le génome humain "entier" a été publié en 2007, et certaines régions du génome humain qui sont difficiles à séquencer, restent encore inconnues.

Développement de l'analyse du génome

Investissements importants, un nombre important de participants d'exception dans la communauté scientifique, et la compétition entre les organisations publiques et privées ont fourni une impulsion considérable au développement de technologies d'analyse du génome.
En conséquence, des technologies de séquençage modernes, comme la NGS (Séquençage Nouvelle Génération) ont émergé ensemble avec une nouvelle branche de la science appelée bio-informatique, un jeune domaine de recherche à l'intersection des mathématiques, de l'informatique et de la biologie, qui développe des techniques et algorithmes pour l'analyse de gros lots de données biologique par des voies plus productives et efficaces.

L'émergence de technologies de séquençage de deuxième et troisième génération (NGS) ont permis de réduire drastiquement les coûts d'analyse du génome. Alors que même en 2009 le coût d'une analyse complète du génome avoisinait les 100 000 USD, le prix moyen actuel pour la même analyse a chuté pour se trouver approximativement en dessous des 1 000 USD (voir figure 1 et tableau 3).

Tableau 1 : Fabricants d'équipement et fournisseurs de réactifs pour le séquençage du génome. Les capitalisations des marchés sont issues de Yahoo! Finances.

Company	Products	Capital-ization	Country
Illumina	Hardware, reagents, consumables, software	28.06 B	USA
Thermo Fisher Scientific	Hardware, reagents, consumables, software (a part of business)	68.98 B	USA
Oxford Nanopore Technologies	Hardware, reagents, consumables	534.41 M	UK
Pacific BioScience	Hardware, reagents, consumables	436,93 M	USA
Roche	Hardware, reagents, consumables, software (a part of business)	213.44 B	Switzerland
Agilent Technologies	Hardware, reagents, consumables, software (a part of business)	19.32 B	USA

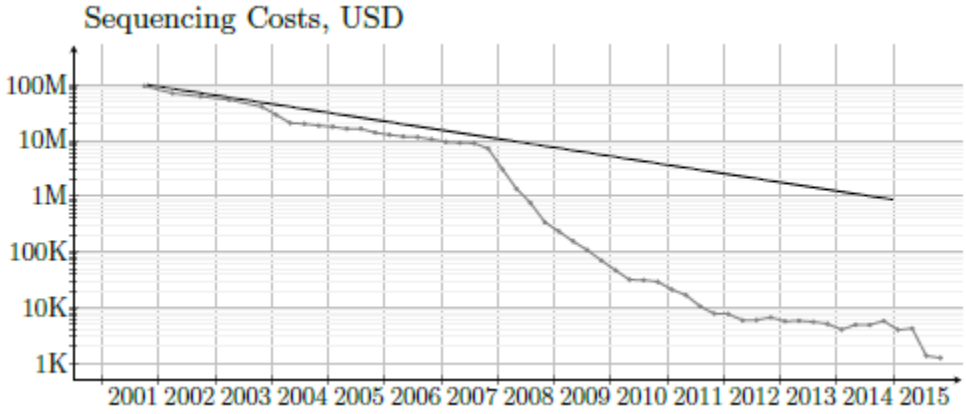


Figure 1 : La loi de Moore et la réduction du coût de l'analyse du génome. Une chute substantielle des prix en 2008 a été provoquée par l'arrivée des nouvelles générations de technologies de séquençage (NGS) [4].

Impact on other sciences

Le développement de la génomique (le domaine scientifique qui étudie différents génomes) a conduit à la transformation de nombreux autres domaines scientifiques, de la biologie à l'anthropologie en passant par la médecine et même les sciences sociales. Un grand nombre de compagnies commerciales reconnues comme Google, Apple, IBM, Amazon et Alibaba ont mis en place des objectifs d'utiliser la génomique pour ajuster leurs produits et services en fonction du profil génomique de leurs clients. De tels ajustements vont permettre à ces compagnies d'améliorer leur relation avec leurs clients et prédire leurs besoins et activités potentielles.

Bases de données génétiques

La réduction du coût de séquençage a mené à une augmentation exponentielle de la disponibilité de donnée génomique. Par exemple, un génome humain complet dans un format appelé "raw data" ("données brutes", ndtr) peut représenter de 50 Gb à 2 TB de données (dépendamment de la profondeur du séquençage requis). Pour stocker de si grands lots de données génomique, des bases de données spéciales ont été créées pour contenir différents types de données, comme la "raw data" obtenue de séquenceurs génomiques ("lire" ou "lecteur"), la séquence de gènes et protéines, des lots de régions d'un génome codées appelées exome et même des séquences de génomes complets ("scaffolds"); certaines de ces bases de données contiennent de l'information pertinente pour les essais cliniques ainsi que les liens entre des traits génétiques et des maladies. La majorité de ces bases de données sont gérées de manière centralisées et financées par des gouvernements ou de grandes corporations. Les scientifiques à travers le monde sont impliqués dans l'ajout de nouvelle donnée à ces bases, permettant une mise à jour et synchronisation rapide. Dans le tableau 2, nous présentons les bases de données les plus connues.

La majorité de telles bases de données sont hébergées dans des pays développées et sont gérées et contrôlées par des gouvernements. L'accès à certaines de ces bases est limité, même pour la communauté scientifique, ou limité par des abonnements commerciaux. Bien que les fondateurs de ces bases assurent qu'ils anonymisent et sécurisent les données génomiques, en réalité, la donnée n'est que stockée sous des pseudonymes, et dans certains cas, des individus ont pu être identifiés en se basant sur l'information génomique disponible [5].

1.2 Aperçu du marché du génome

Dans cette section, nous donnons une brève analyse du marché des technologies du génome. Des exemples des produits génomiques les plus populaires et des compagnies fournissant différents services autour du génome sont décrits.

_____

Dynamiques d'expansion du marché du génome

Le marché pour les technologies génomiques grandit rapidement et est hautement prometteur. Pour l'instant, le volume total du marché est approximativement de 25 milliards de $ avec une progression d'environ 10 fois, partant de 5,9 milliards de $ en 2010 jusqu'à 60 milliards de $ en 2020 (estimation).

Tableau 2. Bases de données génomiques

GenBank	**http://exac.broadinstitute.org**
Owner:	NCBI-NIH, USA
Product:	Genome sequences database
Stored:	More than 199,341,377 different genome sequences
ExaC	**www.ncbi.nlm.nih.gov/genbank**
Owner:	Broad Institute of MIT and Harvard, USA, ODC Open Database License (ODbL)
Product:	Exome Aggregation Consortium
Stored:	60,706 human exome samples/sequences
UniprotKB	**www.ebi.ac.uk/uniprot/**
Owner:	EMBL-EBI, SIB, PIR, UK, Switzerland, USA
Product:	Open Knowledge Base. Manual expert curation. Proteins and genes sequences.
Stored:	More than 555,100 manually reviewed and annotated record
ClinVar	**https://www.ncbi.nlm.nih.gov/clinvar/**
Owner:	NCBI-NIH, USA
Product:	freely available archive for interpretations of clinical significance of genomic variants for reported condition
Stored:	>158 000 submitted interpretations, representing >125 000 variants.
HGMD	**http://www.hgmd.cf.ac.uk/ac/index.php**
Owner:	QiaGen
Product:	Commercial database
Stored:	208,368 human mutation records with annotations
SNPedia	**https://www.snpedia.com/index.php/SNPedia**
Owner:	Open database
Product:	SNPedia is a wiki investigating human genetics
Stored:	107,073 SNPs and linked records
1000 Genomes Project	**www.1000genomes.org**
Owner:	EMBL-EBI, Wellcome Trust
Goal:	find most genetic variants with frequencies of at least 1%
Stored:	More than 2,504 Genome samples/sequences
100000 Genomes Project	**http://www.genomicsengland.co.uk/**
Owner:	NHS, Government of UK
Product:	UK government database containing a sequence of 100,000 genomes
Stored:	32,642 Whole genome sequences

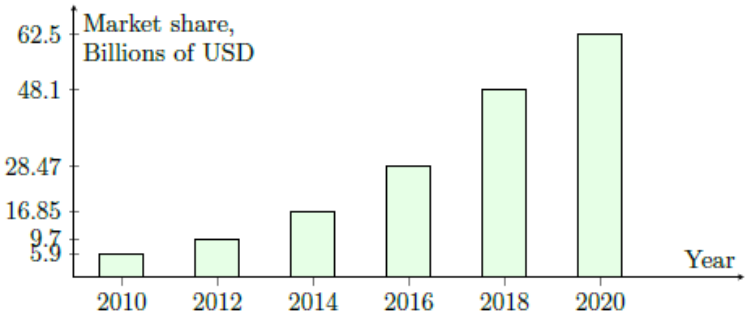


Figure 2. Dynamiques d'Expansion du marché du génome pour la période 2010-2024

Niches de produits

Composant clef de l'industrie 4.0, la génomique a généré une large spectre d'applications potentielles dans presque tous les domaines économiques. Les principaux produits du marché du génome actuels sont :

- **NIPT (Test pré-natal non invasif)**

Description :	basé sur l'ADN d'origine fœtal circulant dans le sang maternel. Le test peut potentiellement identifier l'aneuploïdie du fœtus et le sexe du fœtus après 6 semaines de grossesse seulement.
Parts de marché :	4 milliards USD
Leaders du marché	Illumina, Natera, Ariosa, Sequenom.

- **PGS (Dépistage génétique pré-implantation)**

Description :	Basé sur l'ADN-microarray ou le profilage génétique par séquençage NGS des embryons avant leur implantation par la procédure de fertilisation in-vitro.
Parts de marché :	336.4 millions USD
Leaders du marché	Illumina, Agilent Technologies

- **Tests génétiques DTC (Direct-To-Customer), SNP-génotypage**

Description :	Test génétique basé sur l'ADN-microarray et le SNP-génotypage pour obtenir certaines recommandations telles que : généalogique, besoin nutritionnels, statut du porteur, exercices optimaux.
Parts de marché :	2 milliards USD
Leaders du marché	AncestryDNA, 23andMe, DNAfit, deCode genetics

- **Compagnies de diagnostics (incluant l'oncogénétique)**

Description :	Différents types de procédures de diagnostic, incluant la génétique. Séquençage de lots de gènes, des exomes, de génomes complets et de biopsie liquide.
Parts de marché :	16 milliards USD
Leaders du marché	Pathway Genomics, Human Longevity, Inc., Laboratory Corporation of America, Quest Diagnostics.

1.3 Défis pour la génomique

Dans cette section, les principaux problèmes avec la génomique moderne sont considérés. Ces problèmes doivent être résolus pour permettre le concept de la génomique 2.0 : l'expansion générale et l'application des technologies génomiques au particulier.

_____

Faible disponibilité des analyses du génome

Le coût des différentes analyses du génome sont répertoriées dans le tableau 3 et, en général, commencent à 100$. Cette échelle de prix est très basse par rapport au coût de lecture du génome il y a 5 ou 10 ans [7].
Cependant, le prix du séquençage et de l'interprétation bio-informatique reste élevé, restreignant l'accès public à ces analyses du génome, notamment dans les pays en développement.

Tableau 3 : Types d'analyses du génome.

DNA microarray	
Description:	analysis of 1-5 million pre-selected SNPs
Price range:	$100-500
Information amount	0.033% of full genome
Service providers	23andMe, AncestryDNA, DNAfit.
Genes panels	
Price range:	$100-2000
Information amount	0.001-1% of full genome
Service providers	Pathway Genomics, CeGaT.
Exome sequencing	
Description:	Sequencing for the coding part of a genome (exome)
Price range:	$250-3000
Information amount	2% of full genome
Service providers	BGI, CeGaT.
Whole Genome Sequencing	
Price range:	$600-10000
Information amount	80-98% of full genome
Service providers	BGI, FullGenomes, Human Longevity.

Note : en fait, une portion du génome ne va pas être séquencée. Sa taille dépendra de la profondeur du séquençage et les détails de la procédure de préparation de l'échantillon. Ainsi, le terme "séquençage complet du génome" signifie en fait, l'obtention de la séquence d'un peu plus de 80% du génome.

Obtenir un grand montant d'information génomique (séquences de génomes incluant les caractéristiques phénotypiques) pour les résidents de pays en voie de développement est extrêmement important dans la perspective d'avoir la plus large diversité possible dans l'information génétique qui, au final, va stimuler le développement du marché du "big data" du génome. Même dans les pays développés, seulement 2% de la population a réalisé une analyse du génome (microarray, séquençage d'exomes, séquençage complet du génome).

Compromis sur la confidentialité comme revers de médaille lors de la participation des enquêtes biomédicales

Un consentement informé pour la collecte et l'utilisation des données personnelles est une problématique clef pour toutes les études de recherches biomédicales. Tout projet incluant des études du génome humain commence avec la collecte de signatures des individus attestant le fait qu'ils comprennent les conséquences et acceptent les termes de l'étude. Cette forme consentements varie grandement selon le projet et peut inclure de donner la

permission d'utiliser les données pour des projets futurs, dont les conséquences ne peuvent être connues d'avance.

Dans le Projet du Génome Personnel, supervisé par la Harvard Medical School, les participants ont volontairement accepté que leurs données et échantillons de matériel génétique puissent être utilisés plusieurs fois et peuvent être mis à disposition d'autres laboratoires. Les participants de ces projets sont spécifiquement informés que leur identité peut être désanonymisée et que leur donnée privée peut être rendue publique. Ce projet vise à s'assurer que les données génomiques du plus grand nombre de personnes soient rendu disponibles pour stimuler des nouvelles recherches et le développement de l'industrie génomique.

Les auteurs de ce projet sont persuadés que si nous ne donnons pas un accès libre à la donnée génomique et à son échange, il y a un risque de finir avec des collections de données génomiques isolées, stockées par des entités privées (notamment les compagnies pharmaceutiques, les corporations génomiques et les centres scientifiques), et que chacune de ces bases de données indépendantes ne contiennent pas suffisamment de données pour faire des découvertes majeures.

Conduire des études cliniques et scientifiques internationales centralisées

Pour réaliser de la recherche et développement dans le domaine de la génomique, il est nécessaire de faire des études cliniques et scientifiques sur de grands échantillons et de groupes de population variés. Actuellement, il est difficile d'assurer la collecte d'échantillons d'individus de groupes ethniques variés, car des projets spéciaux doivent être créés, des expéditions doivent être menées et des permissions des autorités locales obtenues.

Aujourd'hui, seule une start-up cherche à adresser cette problématique https://www.dnasimple.org/, pour un petit frais et la promesse d'anonymat.

Création de bio-banques stockant les matériels de divers individus

Les bio-banques servent en tant que médiateur entre les donneurs de matériel biologique (échantillons de sang, extraits d'os, etc.) et les chercheurs en traitant les matériaux reçus et les stockant pour une utilisation future. En général, les bio-banques sont un outil précieux pour la progression de la médecine personnalisée (précision) et le développement de médicaments. Une des plus importantes fonctions des bio-banques est la collecte de matériel de la part de donneurs pour une utilisation future, incluant du sang, les os et même des cellules germinales.

En ce moment, les bio-banques sont de plus en plus développées dans de nombreux pays. Cependant, il y a la possibilité que les échantillons les plus importants appartiennent aux bio-banques les plus solides financièrement, par exemple, les bio-banques appartenant à des compagnies pharmaceutiques, ce qui mènera à un accès inégal au matériel de ces bio-banques pour de nombreuses catégories de chercheurs. Ainsi, il est nécessaire de créer des bio-banques dans le maximum de villes et pays possibles jusqu'à atteindre une quantité suffisante de bio réservoirs personnels pour chaque personne.

Processus : la rapidité d'analyse du génome est limitée par son processus bio-informatique. Pour l'ajustement des logiciels analytiques pour une application généralisée

Actuellement, la vitesse d'obtention de donnée génomique en utilisant les technologies d'analyse du génome (séquençage) est grande et surpasse la vitesse d'analyse de ces données. Si nous considérons toute étude scientifique de grande envergure portant sur un grand nombre de génomes, 20% de la durée de cette étude est dédiée aux étapes expérimentales pour obtenir l'information génomique, alors qu'une majorité du temps restant

est utilisé pour analyser ces données. Ici, l'analyse de données comprend les étapes nécessaires pour obtenir les données de séquençage brutes, l'interprétation des résultats et la recherche d'associations diverses.

Un autre problème avec le logiciel moderne est que ces logiciels d'analyse du génome ont été créés par des scientifiques, pour des scientifiques et donc nécessitent des ajustements pour être utilisés de manière générale par les médecins et le grand public. Comme des analyseurs de génome de la taille d'une clef usb existent déjà, l'utilisation d'un séquenceur personnel de la même manière n'est pas de la science-fiction, mais au contraire, réaliste dans un futur proche.

Interprétation de la donnée génomique : Modèles mathématiques des risques de développement d'une maladie

Plusieurs modèles et algorithmes sont utilisés pour classer les risques de développement d'une maladie en se basant sur la donnée génétique. Les types primaires de ces modèles sont basés sur le type d'héritage considéré : monogénique, polygénique ou multifactoriel. Déterminer le risque de développer des maladies complexes ou multifactorielles implique de compter l'influence et l'interférence de plusieurs gènes ainsi que de facteurs environnementaux. Pour une description plus grande des méthodes disponibles pour déterminer les risques de maladies complexes, voir [8,9].

Pour développer une nouveau modèle déterminant le risque de développer une maladie, il est nécessaire d'inclure un grand nombre de scientifiques. Il est également nécessaire d'étudier de nombreuses publications en regard de la maladie analysée, d'identifier les types d'héritages, déterminer les polymorphismes et mutations qui contribuent au développement de la maladie, et développer "l'algèbre génomique", soit le lot de règles pour l'estimation du risque. Quand un modèle est établi et validé in silico, des études cliniques doivent être conduites pour valider son application. Cette approche est actuellement celle qui est la plus précise, mais également la plus chère que ce soit pour les efforts ou le temps requis.

Interprétation de la donnée génomique : Application de l'apprentissage automatique

L'utilisation d'algorithme d'apprentissage automatique pour déterminer le risque de maladies multifactorielles est grandement étudiée, mais jusqu'à maintenant, du fait du manque d'échantillons d'entrainement suffisants, les modèles mathématiques existant développés par des biologistes surclassent les approches d'auto-apprentissage.

Cependant, l'auto-apprentissage est déjà utilisé pour prédire certaines caractéristiques complexes du corps humain. Un exemple est l'apparition de la prédiction dans le travail de Craig Venter et ses collègues. L'essence de leur travail implique l'analyse de génomes et d'environ 30 000 points faciaux de plusieurs milliers de volontaires. Basés sur les données obtenues, des échantillons d'entrainement pour les algorithmes d'auto-apprentissage sont construits et des dépendances entre des traits génomiques et des apparences individuelles ont été déterminées. Grâce à ce travail, les ordinateurs ont appris à restaurer de manière précise l'apparence d'une personne basés sur ses données génomiques [10,11].

Les résultats de ce projet ont permis la prédiction de l'apparence d'un criminel ou d'un enfant pas encore né lors des premiers mois de grossesse. En obtenant un échantillon de sang d'une femme enceinte et en extrayant l'ADN foetal du sang, on peut déterminer avec précision l'apparence à 18 ans d'un enfant pas encore né.

Pour permettre ce projet, Craig Venter a recruté l'un des meilleurs spécialiste de l'auto-apprentissage de Google, Franz Och, une célébrité informatique, connu pour être l'architecte principal de Google Translate [12].

En ce moment, l'auto-apprentissage n'est pas encore utilisé de manière répandue pour les maladies, car de nombreux échantillons d'entraînement et correctement structurés sont requis. La création de bases de données de

génomes humains, tout autant que la disponibilité de questionnaires détaillés reflétant l'état de santé des individus, peut améliorer le développement de l'entrainement informatique dans la génomique et permettra une grande précision avec peu de marge d'erreur dans la détermination du risque de développement d'une maladie. En même temps, ces données seront publiques et disponible à tous les utilisateurs du système, excluant la possibilité d'un monopole. Cette disponibilité est très importante, car la concentration de grandes quantités de données dans les bases de données corporatives provoquera un monopole dans le domaine de l'auto-apprentissage du génome.

Stockage sécurisé de grandes quantités de données

La sécurité des données personnelles est très importante; nous essayons tous de nous prémunir du vol des données de carte de crédits, des informations d'assurances, des numéros des comptes bancaires, et de l'information médicale de tout type. Le vol d'information génomique peut paraître bénin pour de nombreuses personnes aujourd'hui. Cependant, cela peut mener à de terribles conséquences qui sont difficiles à prédire, comme par exemple, la possibilité qu'une portion de la séquence d'un génome d'un individu soit synthétisée et déposée sur le lieu d'un crime ou d'un acte terroriste.

Les solutions actuelles pour ce type de problèmes impliquent l'encryptage du stockage sur un serveur central comme [13], https://www.pathway.com/, https://www.23andme.com/, ou http://www.humanlongevity.com/. Ce type de stockage de données "fermé" est relativement sécurisant, mais il empêche la possibilité de partager la donnée et son accès aux scientifiques du monde entier, ce qui devient une limitation majeure au développement de la science génomique moderne.

Un autre problème associé avec le stockage est la grandeur du génome lui-même, et l'expansion exponentielle de la disponibilité de données génomiques, puisque de plus en plus de personnes font du séquençage génomique. Une étude [14] estime qu'en 2025, le volume total de données génomiques stockées (en considérant qu'un génome complet fait 100Gb) atteindra 40 exabytes par année et que le stockage de génomes deviendra le plus grand consommateur des capacités de stockage et d'analyses informatiques.

Tableau 4. 4 domaines du “Big Data” en 2025

Astronomy	
Acquisition	25 zetta-bytes/year
Storage	1 EB/year
Analysis	In situ data reduction; Real-time processing; Massive volumes;
Distribution	Dedicated lines from antennae to server (600 TB/s)
Twitter	
Acquisition	0.5-15 billion tweets/year
Storage	1-17 PB/year
Analysis	Topic and sentiment mining; Metadata analysis
Distribution	Small units of distribution
YouTube	
Acquisition	500-900 million hours/year
Storage	1-2 EB/year
Analysis	Limited requirements
Distribution	Major component of modern user’s bandwidth (10 MB/s)
Genomics	
Acquisition	1 zetta-bases/year
Storage	2-40 EB/year
Analysis	Heterogeneous data and analysis; Variant calling, $\sim$ 2 trillion central processing unit (CPU) hours; All-pairs genome alignments, $\sim$ 10,000 trillion CPU hours.
Distribution	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

Pour des bases de données communes avec des questionnaires continuellement mis à jour

En même temps, le manque d’une base de donnée publique basée sur le concept du stockage distribué et utilisant du logiciel libre peut mener à une domination totale de ce marché par des compagnies comme Google ou Amazon du fait de leur serveurs puissants. Si des corporations ou des compagnies pharmaceutiques possède le monopole dans le domaine de l’information génomique, nous observerons un développement lent et très couteux de la médecine, avec les approches de traitements actuels qui perdureront au lieu d’un futur dans lequel une maladie peut être prévenue avant même son développement ou dans ces premiers signes de manifestation.

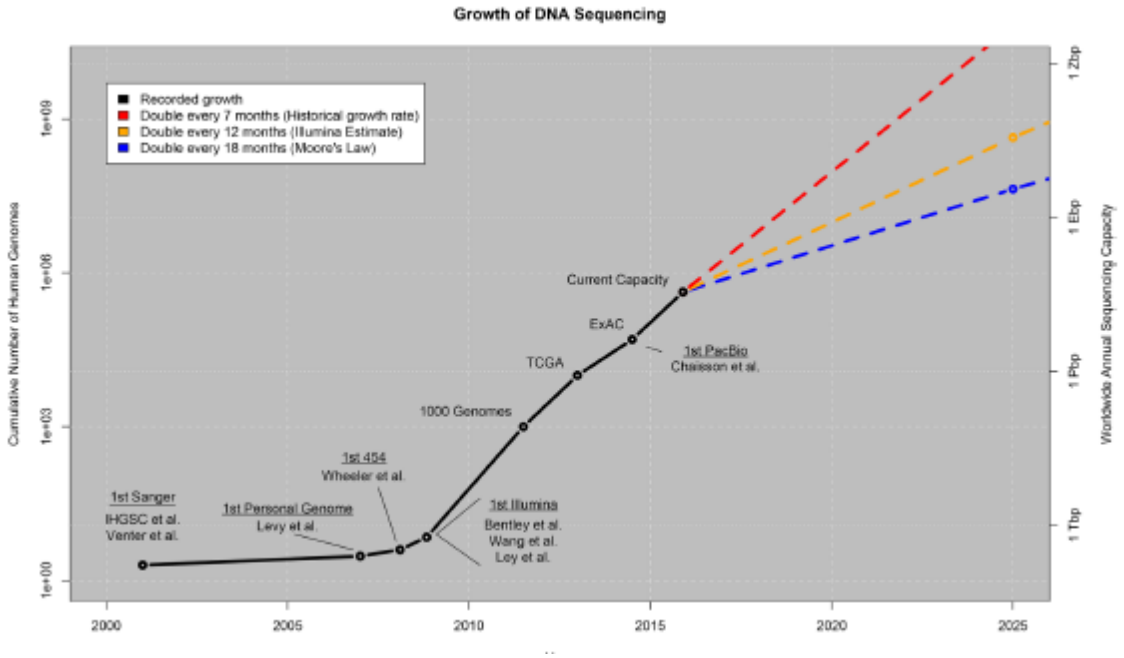


Figure 3 : Expansion des données de séquençage de l'ADN. Cette courbe illustre l'expansion du séquençage de l'ADN en considérant le total de génomes humains séquencés (axe de gauche) et la capacité maximale mondiale de séquençage (axe de droite : Tera-basepairs (Tbp) ("basepairs" pour le nombre de paires identifiées (ndtr)), Peta-basepairs (Pbp), Exa-basepairs (Ebp), et Zetta basepairs (Zbp)). Les données de 2015 sont basées sur des données historiques publiées et liées avec des étapes majeures dans le séquençage (premier séquençage Sanger avec la première publication du génome humain PacBio), ainsi que trois projets référence utilisant du séquençage à grande échelle : le 1000 Genome Project qui en 2012 a accumulé des centaines de génomes humains, The Cancer Genome Atlas (TCGA) qui accumula plusieurs milliers de paires de génome tumeur/normal, et Exome Aggregation Consortium (ExAC), qui accumule plus de 60 000 exomes humains. La majorité des génomes séquencés jusqu'à aujourd'hui sont en fait des exomes, mais nous nous attendons à ce que dans le futur, ce ratio s'inverse au profit des génomes complets. Les valeurs au-delà de 2015 sont des projections en fonction de trois scénarios possibles, décrits dans le texte principal. Source [14].

Le futur proche de la génomique

En conclusion de ce chapitre, il est important de décrire quelques perspectives hypothétiques mais réalisables techniquement, ainsi que les dangers auxquels feront face l'industrie génomique et la société en général :

- Une réduction du coût des analyses du génome et la miniaturisation des appareils de séquençage du génome au point d'être au niveau de plug-ins pour téléphones;
- L'expansion explosive des données génomiques, leur stockage et l'émergence des pirates du génome et les problématiques de confidentialité (protection de la donnée)
- La distribution générale de la médecine génomique et la télésanté
- Changement dans l'industrie alimentaire avec la nutrition personnalisée basée sur le génome
- Le développement de traitements à base de médicaments personnalisés
- Les rencontres basées sur la compatibilité génomique
- Identification des individus en utilisant l'information génomique, incluant la possibilité de faire des paiements et d'obtenir des services
- Une augmentation de la longévité dans tous les pays, l'extension de la vie active, et les changements de préjugés envers les populations plus âgées, les grossesses tardives et les taux de naissance décroissants
- Le développement de technologies d'édition du génome

- Design de l'apparence de futurs enfants et autres développements qui sont difficiles à imaginer. Par exemple, la possibilité technique de sélectionner les enfants en santé et de déterminer leur apparence

future autant dès l'étape de l'embryon, soit en obtenant l'ADN du fœtus du sang d'une femme enceinte [10].

1.4 Problématiques éthiques de la génomique personnelle

Dans cette section, les problèmes éthiques associés au développement et la distribution mondiale des technologies génomiques sont considérés, tel que la confidentialité, les bases de données publiques, le libre accès pour les chercheurs, les mauvaises utilisations et menaces possibles aux libertés individuelles du fait de l'expansion de la génomique.

NOTE : Dans ce livre blanc, nous fournissons une brève revue des principaux problèmes et perspectives de l'industrie du génome. Pour de plus amples détails et discussions sur la confidentialité et les problèmes de sécurité, voir https://www.smeal.psu.edu/fcfe/documents/innovations-in-medical-genomics-pdf.

La confidentialité

L'information génomique personnelle est très sensible pour beaucoup de gens. Cependant, beaucoup de personnes ne comprennent pas que, basé sur leur information génomique, il est possible de déterminer leur longévité, leur propension à prendre des décisions émotionnelles (manipulation du processus de décision), possibilité de développer différentes maladies mentales et le risque de mort subite due, par exemple, à l'arythmie du cœur.

De telles informations pourraient être désavantageuses lors de la recherche d'emploi, la participation à une élection ou l'établissement du prix d'une assurance médicale. Il y a également la possibilité qu'une personne malintentionnée qui connait la séquence d'un génome puisse laisser des fragments d'ADN identique pour, par exemple, faussement impliquer ou accuser une personne d'un acte terroriste. Une autre personne pourrait se voir refuser des soins médicaux (ou devrait avoir à payer plus) ou encore se voir refuser un travail.

Les corporations et les gouvernements pourraient délibérément influencer les décisions d'une personne et leurs actes connaissant leurs "faiblesses" identifiée par le génome. Ainsi, la protection des données génomiques est nécessaire pour protéger l'égalité des droits de différentes catégories de gens.

En même temps, certaines études ont déjà permis d'identifier des individus en se basant sur leurs génomes anonymisés [5,15].

De plus, certaines compagnies (http://www.humanlongevity.com/media/) possèdent des algorithmes d'auto-apprentissages qui peuvent reconstruire avec précision l'apparence d'une personne en utilisant seulement leurs données génomiques. [16].

Publicité

Une telle approche va permettre le développement de la médecine préventive à travers laquelle l'analyse de grands lots de données vont permettre la prédiction du développement des maladies (avant leur apparition), permettant des actions augmentant la durée de vie et améliorant la qualité de vie [17] tout autant que l'identification de donneurs mondiaux pour répondre à certains besoins médicaux. Malheureusement, il n'existe pas actuellement de solution valide pour le public pour l'utilisation de l'information génomique tout en protégeant la confidentialité. Toutefois, des start-ups de ce domaine travaillant sur la technologie "blockchain" devraient être mentionnées.

Encrypgen, une start-up qui récemment réalisé son ICO, décrit l'existence de problèmes et la relation entre confidentialité et publicité en utilisant des "blockchains" [18]. Cependant, le livre blanc associé à ce projet n'offre pas la description d'une implémentation technique qui apporterait une solution au problème de confidentialité et de disponibilité.

Un autre projet dans ce domaine est la start-up DNAbits [19], dont le fondateur, Dror Samuel Brama, a déposé un brevet pour une approche générale dans le stockage de données et le transfert en utilisant les technologies de "blockchain" [20]. Toutefois, cette compagnie n'a pas mis en place, techniquement parlant, son concept lors des trois dernières années.

Le droit de posséder les données génomiques ?

Aujourd'hui, il n'y a pas de définition légale du droit de posséder sa propre information génétique. Dans certains pays développés, incluant les USA, l'Allemagne, et l'Autriche, les citoyens n'ont pas le droit d'accéder ni de posséder leur données génétiques dans le contexte de son interprétation [21]. Un intermédiaire, soit un médecin ou un centre médical qui possède de tels droits, est nécessaire. Cette voie est utilisée par les compagnies Pathway Genomics aux USA et CeGat en Allemagne (http://www.cegat.de/en/).

Pour entreprendre une analyse génétique, les recommandations d'un médecin qui peut être un fournisseur de tests génétiques est nécessaire, et seulement un médecin a les droits d'interpréter l'information issue de l'analyse génétique.

Aux USA, il y a des fournisseurs de services dans le domaine de la "génétique pour le plaisir", tels que les compagnies 23andMe et Ancestry.com, qui vendent des tests génétiques directement à l'utilisateur, mais ces compagnies ne peuvent seulement que fournir des informations sur les origines ethniques et certaines caractéristiques reliées à la santé (par exemple, les habiletés sportives) et n'ont pas la permission de fournir les informations médicales importantes. Ces limitations, imposées par des régulateurs comme la FDA, n'empêchent pas 23andMe de vendre l'accès aux données génétiques aux grandes compagnies pharmaceutiques. Ce genre de contrats sont connus : un accord a été passé avec Genentech (une sous-division du géant pharmaceutique Roche) pour 60 millions USD [22] pour une étude de la maladie de Parkinson, et un accord avec un autre géant pharmaceutique, Pfizer, pour une étude sur les maladies inflammatoires des intestins (notamment, la maladie de Crohn) [23]. Certains rapports mentionnent également des négociations entre 23andMe et Novartis pour une étude sur la maladie d'Alzheimer [24].

Ainsi, nous donnons actuellement le droit aux grands groupes de gérer notre information génomique, de la stocker et d'en retirer des profits. Les corporations, derrière le voile des bonnes intentions, monopolisent le "big data" du génome, et nous ne pouvons prédire comment ce monopole influencera le prix de futurs médicaments et les découvertes en médecine.

Le droit d'accès à l'information génomique

Une autre problématique éthique qui doit être discutée; auparavant, nous avons souligné que les violations de confidentialité et de l'accès à l'information génomique pourraient être utilisées illégalement par un employeur, par exemple. Une personne pourrait donc être remerciée ou se voir refuser une promotion en se basant sur son information génétique. Pour ceux qui travaillent pour la sécurité de la population et des systèmes, comme des conducteurs de camion, de bus, des pilotes, des opérateurs de centrales nucléaires ou d'autres avec des occupations similaires, l'état de santé est un élément critique, et l'information génomique pourrait prévenir un accident ou un désastre. Dans certaines professions, un danger potentiel peut menacer un travailleur plutôt que son entourage, comme un mineur avec des problèmes de poumons. Pour ces cas, une discussion impliquant professionnels et experts ainsi que le grand public est nécessaire, notamment pour développer des cadres législatifs pour l'utilisation de l'information génomique par les employeurs.

Chapitre 2

Concept

2.1 Le projet Zenome

Vision philosophique

L'intérêt du grand public envers la médecine génomique reste encore assez bas dans les pays développé et encore moins dans les pays en développement. Cela signifie que les gens, en général, ont peu de compréhension des bénéfices possibles de la génomique et de ses dangers. Dans de nombreux pays, cela a conduit au développement de procédures institutionnelles inutilement complexes pour protéger l'information génétique d'une éventuelle mauvaise utilisation, qui plus est a ralenti le progrès scientifique.

La plateforme Zenome cherche à augmenter l'intérêt envers la médecine génomique, pour permettre aux utilisateurs de prendre des décisions éclairées quant à leurs données. Pour s'assurer d'y parvenir, la plateforme Zenome se base sur les principes fondamentaux suivants :

Propriété de l'information génomique personnelle Chaque participant à tous les droits sur ses données génomiques personnelles

Liberté de choix Chaque participant décide comment l'information génétique doit être utilisée. Une personne peut choisir de participer, ou non à une étude clinique ou scientifique.

Le droit de partager Le participant peut autoriser l'accès à l'information génétique à une tierce partie tout en interdisant sa copie

Confidentialité L'encryptage des données les rendant confidentielles rend impossible l'accès à l'information génétique de l'individu sans sa permission

Stockage de données distribué Une architecture de bases de données distribuées permet une haute disponibilité et une tolérance aux pannes grâce à la réplication et la possibilité d'expansion

Analyse des données distribuée Les données sont analysées sur le maximum de nœuds du réseau possible en même temps. Chaque utilisateur peut devenir un nœud en fournissant de l'Espace disque et du temps de son CPU au réseau.

Expansion L'architecture de la plateforme permet une expansion et une flexibilité du système.

L'écosystème de la plateforme Zenome

Sur la plateforme Zenome, l'utilisateur a la possibilité de réaliser de multiples interactions à travers le système. Ces interactions prennent place à différents niveaux des systèmes, et n'interfèrent pas entre eux tout en impliquant

différents type d'interactions. Ainsi, ils devraient être représentés en tant qu'entités distinctes ayant des rôles différents:

Nœud (Calcul/stockage) qui fournit de l'espace disque et de la puissance processeur pour une récompense.

Personne qui a importé ces données génétiques personnelles sur la plateforme et qui va potentiellement utiliser des services génétiques.

Analyste qui est intéressé dans l'analyse de l'information génétique sur la plateforme. Peut représenter : un scientifique de données, une organisation scientifique et ainsi de suite.

Un fournisseur de service qui implante un espace pour l'utilisateur afin d'obtenir un service génétique (probablement payant) sur la plateforme. En fait, c'est une organisation qui utilise les données génétiques comme raison d'affaires.

En bref : Chaque utilisateur se voit offrir de multiples interactions de types différents, utilisant différents rôles. Certains de ces rôles, tel que fournisseur de service et analyste, nécessitent des connaissances spécifiques, contrairement à Nœud et Person.

Chaque rôle sera décrit en détails plus tard.

Aperçu de l'architecture du système

La plateforme Zenome est une application distribuée qui est constituée de 3 couches principales.

Couche Réseau (et d'accès aux données) : fournit un niveau d'abstraction qui encapsule les interactions réseau et fournit une interface à un environnement distribué pour les couches supérieures.

	Blockchain	DHT Kademlia
Data Storage Cost	High	Low
Data Immutability	True	False
Performance	Low	High
Deterministic Result	True	False

Cette couche est constituée de deux systèmes distribués de nature complètement différente :

Portefeuille distribué (basé sur la "blockchain") qui enregistre les transactions entre les participants d'une manière vérifiable et permanente. Pour accéder au nœud de la "blockchain", le logiciel utilise le client Ethereum inclus.
Réseau de Table de hachage distribuée (DHT)(basé sur le protocole Kademlia) qui combine les nœuds physiques dans un réseau superposé et qui permet aux messages de passer entre les nœuds et le stockage distribué des données.

Rôle : (Calculs/Storage) **Nœud** opère au niveau de la couche réseau.

Logiciel intermédiaire (Middleware) contient la gestion des comptes, est l'interface cohérente avec les fonctions de sécurité de la couche sous-jacente et des APIs haut niveau pour le logiciel qui fonctionne au niveau applicatif.

Rôle : Le **fournisseur de service** fonctionne avec le Middleware. Les API de services externes de la plateforme permettent aux tierces parties de démarrer les services génétiques sur la plateforme.

Couche applicative L'application Zenome a pour fonction une interface avancée pour l'utilisateur final qui traduit les actions de l'utilisateur au Middleware de manière cohérente. L'interface est extensible par design ainsi les services génétiques peuvent rouler nativement sur la pile de logiciel.

Marché des services génétiques

Le marché des services génétiques est actuellement en développement dans les domaines suivants :

1. **Recherche et adoption de technologies** dans le marché
2. Mise à disposition de **services de diagnostic génomique**
3. **Certification gouvernementale** de technologies génétiques
4. **Développement d'un cadre légal.** En particulier, des mesures législatives pour préserver l'information génétique.

Note : La structure du marché est assez complexe. Certains joueurs sont en fait en train de se développer dans plusieurs directions pour trouver leur place dans le marché pour répondre aux besoins grandissant des consommateurs.

Les joueurs suivants sont représentés sur le marché des services génétiques :

Les corporations scientifiques qui travaillent sur la découverte et l'adaptation de nouvelles technologies sur le marché. Celles-ci incluent :

- Les compagnies pharmaceutiques, biotechnologiques et les compagnies de diagnostic, comme Pfizer et Myriad
- Les compagnies qui développent et vendent tous les réactifs chimiques nécessaires (comme Life Technologies)

Les compagnies bio-informatiques qui sont engagées dans l'invention et le développement de méthodes d'analyse des données par ordinateur. Les joueurs dans ce secteur sont encore en difficulté avec les types et les volumes des données obtenus.

Les centres scientifiques et médicaux qui jouent un rôle de leader dans la mise à disposition et le développement de services de diagnostics génétiques.

Les laboratoires commerciaux qui fournissent des services de diagnostics génétiques rapides, efficaces et habituellement à un prix relativement abordable. Ils possèdent de nombreuses ressources compétentes et financières.

Les compagnies qui vendent directement au consommateur les diagnostics génétiques qui augmentent l'intérêt du grand public envers les diagnostics génétiques. Pour l'instant ce segment de marché est relativement petit, mais dans le futur, il peut devenir la part de marché principale et qui peut être adaptée aux pratiques cliniques.

Tableau 5. Comparaison sur le marché de produits similaires

	We	GeneCoin	Encrypgen	23andMe	Pathway Genomics	Snpedia (Promethease)	Human longevity
Decentralized	✓	✓	✓	–	–	–	–
Suitable for non-human organisms	✓	✓	✓	–	–	–	–
Customer is the owner of his data	✓	✓	✓	–	–	✓	–
Possibility to load your own data	✓	✓	✓	–	–	✓	–
Opened nonprivate data	✓	✓	–	–	–	–	✓
Performs its own data analysis	✓	✓	–	✓	✓	✓	✓
Provides a report for customers	✓	–	–	✓	✓	✓	–
Uses AI and Machine Learning	✓	–	–	–	–	–	✓
Sharing without transmitting huge data	✓	–	✓	–	–	–	–
Earn using your data	✓	–	–	–	–	–	–
Opened for scientists	✓	–	–	–	–	✓	–
Is a platform for other tools	✓	–	✓	–	–	–	–

2.2 Rôles sur la plateforme Zenome

Perspectives des nœuds des ressources

Le système du point de vue du nœud informatique
Nœud - un participant qui fournit les ressources de son ordinateur (espace disque et temps de processeur) pour le besoin de stockage distribué et l'analyse de données génétiques pour obtenir une récompense en Tokens ZNA.

Pour devenir un nœud de calcul, les utilisateurs lancent le logiciel Zenome sur leur ordinateur et activent le rôle de nœud à partir de l'interface utilisateur graphique. Le logiciel doit être utilisé continuellement en arrière-plan.

Note : une version du logiciel Zenome en ligne de commande sera également disponible pour spécifiquement lancer un nœud de calcul.

La politique d'allocation des ressources de son système et la gestion des tâches peut être configurée par le propriétaire du nœud :

- Espace disque maximal que le logiciel peut utiliser
- Règle d'utilisation des ressources de calculs :
    - Utilisation fixe du CPU/GPU, nombre de cœurs et chargement maximal (en pourcentage) de chaque cœur
    - Utilisation dynamique du CPU/GPU, allocation des ressources pour ne pas interférer avec les autres applications de l'utilisateur
- Choix des processus de calculs (par id) qui auront la plus haute priorité d'exécution
- Arrêt temporaire du nœud : effectue une requête au réseau pour transférer les données non disponible ailleurs, aux autres nœuds. Attente jusqu'à ce que les données soient transférées puis déconnexion.
- Arrêt complet du nœud : effectue une requête pour transférer toutes les données aux autres nœuds, attente jusqu'à ce que le transfert soit terminé et effacement des données.

Personne/Utilisateur

Personne - un individu fournissant volontairement son information génétique au système Zenome dans l'objectif d'en tirer profit en vendant ses données personnelles ou en utilisant des services génétiques disponibles sur la plateforme.

Un utilisateur installe le logiciel Zenome pour pouvoir travailler avec l'information génétique personnelle.

En utilisant l'interface graphique, l'utilisateur sera en mesure de :

- Créer un compte personnel
- Gérer ses tokens : réception, transfert, utilisation, paiement pour du stockage de données, dépense pour des services génétiques payants
- Importation de données génétiques (le format du fichier est dans la majorité des cas détecté automatiquement)
- Gestion des données personnelles : mise à disposition des données personnelles, remplissage de questionnaire, consultation d'une liste des questionnaires les plus populaires
- Travailler avec des offres ciblées en utilisant un sous-système de recommandations sécurisé
- Utilisation de services génétiques et configuration du niveau de confidentialité pour chacun des services. Individuellement.

En chargeant les données génétiques, le niveau de confidentialité peut être sélectionné parmi:

Confidentialité Totale Dans ce cas, les données sont stockées sous une forme encryptée, et le prix complet est facturé pour stocker ce type de données

Confidentialité Normale les données génétiques sont stockées de manière fragmentée et il est impossible d'identifier l'utilisateur. Chaque fragment est stocké ouvertement sur le système. L'information pour lier les fragments aux ID des utilisateurs est privée. Dans ce cas, le stockage est moins cher puisque subventionné.

Accès public les données sont stockées ouvertement. Le stockage est gratuit puisque subventionné.

Note : Attention ! Bien qu'il n'y ait pas de restrictions techniques pour augmenter le niveau de confidentialité, seulement les données futures seront affectées. Les données rendues publiques une première fois ne pourront redevenir privées. Considérez ceci lorsque vous importez vos données la première fois.

Consommateur de données

Consommateur de données - un scientifique, une société commerciale, une organisation scientifique ou tout autre participant à la plate-forme, qui s'intéresse à l'analyse de données génétiques en utilisant les capacités de la plate-forme. Le consommateur de données est en mesure de faire des demandes aux utilisateurs et de fixer un remboursement qui sera payé aux utilisateurs qui répondent.
Note: il existe certaines restrictions pour les requêtes conçues pour empêcher la désanonymisation de l'utilisateur du système et autre mauvaise utilisation de la plate-forme. Ces restrictions dépendent également du classement des consommateurs dans le système.

Fournisseur de services

Fournisseur de service - une organisation qui utilise les données génétiques pour réaliser ses affaires et implémente ainsi un service à l'utilisateur sur la plateforme.

Note : Un utilisateur peut choisir lui-même les données qu'il désire partager avec le fournisseur de service. L'utilisateur sera notifié si les données requises peuvent être utilisées pour l'identifier.

Note : Par exemple, un fournisseur de service ne peut voir plus de 70% de la liste de mutation dans un format lisible en utilisant des requêtes directes ou en obtenant l'information correspondante aux données brutes (comme des fichiers fastq) avec le questionnaire de l'utilisateur.

2.3 Données génomiques

Types de données d'omiques (génomiques)

Dans le cadre du concept, on distingue 3 types de données d'omiques (génomiques) dépendamment du paiement et de la valeur de l'information :

- Données ouvertes, qui n'a pas ou peu de valeur pour ses propriétaires, mais qui est importante pour les scientifiques
  Exemple : le génome de la souche de la bactérie Helicobacter pylori.

- Données ouvertes, qui ont de la valeur autant pour le propriétaire et les consortiums.
  Exemple : les génomes de la majorité des participants du réseau

- Données restreintes qui sont simplement stockées sur le réseau
  Pour des projets à accès limité de diverses institutions publiques ou commerciales

Pré-analyse des données génétiques

La gestion des données génomiques issues de NGS (et de la majorité des autres lots de données omiques) consiste habituellement en deux étapes indépendantes :

1. Analyse préliminaire des données "brutes"
2. Analyse dédié des séquences génétique pour le développement de recommandations personnelles ou dans le cadre une recherche

Note : un génome de référence - est un lot de données d'ADN digital assemblé pour être un exemple représentatif d'un ensemble de gènes d'une espèce.

La pré-analyse des données génomiques issues des NGS inclus les étapes suivantes :

1. Alignement des lectures des NGS au génome de référence
2. Recherche de mutations et autres différences par rapport au génome de référence, en enregistrant leur liste dans le format gVCF

Note : le protocole est le même pour les données d'organismes non humains. Bien sûr, dans ce cas, un génome de référence approprié est requis.

Si les lots de données personnelles représentent le résultat de la technologie de génotypage par microarray (type de fichier de 23andMe), ils peuvent être importés dans la plateforme en tant que gVCF puisque les formats de données de ces types de fichier sont similaires.

Le problème des fausses données

Si les données génétiques d'autres organismes (non humains) sont importés à la place d'un bon génome (de manière accidentelle ou volontaire), cela sera détecté pendant la pré analyse des données brutes et l'utilisateur sera notifié. Si l'utilisateur importe volontairement des données génétiques erronées (fausses) dans le système, cela peut être détecté en utilisant des méthodes connues de vérification avant de les stocker.

Note : l'incitatif économique pour éviter l'importation de fausses données est que le paiement pour le stockage de données devrait être fait en avance, pour l'ensemble de l'année à venir.

Le problème de l'identification de l'utilisateur en se basant sur les données génomiques

L'accès libre à l'information génétique apporte le problème de l'identification des utilisateurs par leurs données génomiques et autres. Si un utilisateur décide de ne pas rendre disponible complètement ses données génétiques, des mesures appropriées doivent être prises à chaque étape pour analyser et stocker les données. Pour résoudre ce problème de l'identification de l'utilisateur, l'interaction doit être pensée de telle manière qu'à chaque étape aucun nœud ne peut déterminer la propriété du matériel génétique pour un individu spécifique, ou même la ville dans laquelle l'individu réside.

Note : Les différences entre des résidents d'une même ville sont d'environ 0,01% de la séquence

A chaque étape, l'objectif est atteint de différentes manières :

1. Dans la phase d'analyse préliminaire - en divisant le fichier source en morceaux de sorte que la couverture moyenne soit inférieure au seuil de confiance (6 copies)
2. À l'étape du stockage - le compte de stockage est fragmenté par longueur.

Stockage des données génomiques

Les données génomiques sont localisées dans un réseau distribué basé sur le protocole DHT Kademlia. Les participants qui fournissent les ressources pour opérer ce réseau (voir les détails au rôle *Nœud*) reçoivent des paiements sous la forme de tokens ZNA. Pour recevoir ce paiement, ils doivent prouver au réseau qu'il stocke bien les données. La procédure pour cette vérification est basée sur l'utilisation de la "blockchain". L'encryptage est utilisée si nécessaire.

Note : Une intégration avec les noeuds Storj et FileCoin sera également réalisée.

Tel que mentionné précédemment, les données peuvent être "brutes" et traitées.

Type of data	Raw	Processed
Features		
Format	fastq / bam	(lists of mutations) gvcf / vcf + bed / 23me(txt)
Size, Gb	50	2
Value	To improve the technology of sequencing and processing (for the equipment development market)	To conduct research, as well as to make a report.
Storage conditions		
Number of copies	At least 3 in the independent nodes	At least 5 in the independent nodes

Les données génomiques sont divisées en fragments lorsque stockées, de telle manière que la longueur des fragments ne permettent pas d'identifier sans ambiguïté leur appartenance à un individu spécifique.

Note : L'information qui détermine quels fragments du génome constitue le génome de l'utilisateur est également confidentiel et ne peut être obtenue qu'avec l'autorisation de l'utilisateur.

2.4 Profils personnels

Le remplissage de questionnaires augmente significativement la possibilité d'utilisation des données génomiques. Les utilisateurs remplissent les questionnaires en utilisant l'interface graphique.

Note : Si certains questionnaires deviennent populaires, l'application demande à l'utilisateur de le remplir. Chaque analyste peut créer son questionnaire et le déposer sur la plateforme.

Caractéristiques d'un questionnaire

Un certain nombre de questionnaires peuvent être énormes; il est donc nécessaire d'introduire la notion de caractéristique du questionnaire.

La caractéristique d'une question - est la description complète de tous les champs du questionnaire et des valeurs autorisées dans ses champs à remplir.

Note : Officiellement, la caractéristique du questionnaire contient une référence à l'auteur, une description et des lots rassemblés enregistrés; chacun correspondant à un champ du questionnaire.

Les champs peuvent être sous différents types :

Champs numériques la valeur du champ est un nombre entier.

Choix multiples la valeur du champ est un nombre ou une réponse.

Note : Les réponses à ce type de questions seront en accès libre (sans aucune référence à l'utilisateur), puisqu'elles ne menacent pas la confidentialité.

Champs d'une chaîne de caractères La valeur de champ est une chaîne. C'est un champ privé car il peut potentiellement compromettre l'identité de l'utilisateur en se basant sur une réponse spécifique.

Bloc privé Permet de rendre n'importe quel champ ou lot de champs privés, peu importe leur type.

Requêtes au système

Les données statistiques concernant certains états génomiques seront ouvertes au public si le propriétaire n'a pas décidé de les encrypter. De plus, l'information concernant les réponses disponibles au questionnaire le sera également. Ainsi, tout le monde saura, par exemple, quel est le nombre d'usager du réseau ayant 25 ans ou ayant une mutation rs6025 (facteur V de coagulation).

L'architecture du système empêche l'Extraction complète de toute la base de données :

- Lors de la création de requêtes conjointes, le client n'a pas accès aux données "brutes".
- Un frais de base inclus seulement un nombre limité de requêtes quotidiennes. Les frais pour des requêtes supplémentaires durant un même jour augmentent exponentiellement.
- Si le résultat d'une requête associative contient moins de 100 utilisateurs, le résultat ne sera pas fourni au client.

Si certains utilisateur encryptent complètement leurs données, alors seulement eux pourront décider à qui leurs données personnelles, y compris les fragments constituant leur génome, pourront être transférées. Aucun analyste ne sera en mesure de savoir quels types de données sont encryptées.

Transfert sécurisé des données personnelles

Le processus de transfert des données personnelles entre des participants sur le système devraient avoir les propriétés suivantes :

1. Les données complètes à être transmises ne doivent être accessibles qu'à l'acheteur et au vendeur
2. La transmission de token ne devrait avoir lieu que si le transfert des données est réalisé avec succès
3. La tentative de vendre des données erronées doit être détectée et bloquée
4. La tentative d'accuser faussement un vendeur de fournir des données erronées doit être révélée
5. La transmission des données ne doit pas être confiée à une tierce partie.

Note : la technologie "blockchain" sera utilisée pour sécuriser le transfert des données. Toutefois, il doit être considéré que stocker (et transférer) de grandes quantités d'information à la "blockchain" utilise beaucoup de ressources. C'est pourquoi:

6. Il est seulement autorisé de transférer de petites quantités d'information par la "Blockchain". Les données restantes peuvent être transmises par un simple canal de communication encrypté.

2.5 Système de notation

La plateforme va créer un classement pour :

- Séparément pour chaque fragment génétique ou bloc de données personnelles
- Organisations
- Fournisseurs de services
- Fournisseurs de données (tels que les laboratoires de séquençage d'ADN)

Note : Il n'y aura pas de classement individuel des utilisateurs, puisque cela représente la somme des notes de ses fragments génétiques, qui ne sont pas disponibles au public. Si les données génétiques ont été importées par un compte Organisation, alors la note initiale sera automatiquement augmentée par la note de l'organisation.

Les facteurs qui affectent les notes des fragments génétiques sont :

La confirmation du laboratoire augmente la note des données importées en proportion de la note du laboratoire à partir duquel elles ont été obtenues.

Vérification de plausibilité permet de vérifier l'information génétique d'un utilisateur en utilisant des modèles statistiques prédéfinis sur les fréquences polymorphiques et de lien génétique. Ce module est en développement.

Participation aux recherches la note augmente avec le nombre de recherches réussies impliquant le fragment. Si le résultat des recherches sur l'information a été considérée peu probable, la note redescend.

2.6 Cas d'utilisation

Utilisation individuel

Pour l'utilisateur individuel il y a une opportunité d'obtenir leur information génétique et de la transformer en source de revenu. La combinaison du génome et des interactions avec l'environnement est une source d'information importante. Notre plateforme permet à l'utilisateur de gérer de manière sécurisée cette ressource.

La plateforme fournit l'option de stocker et partager en sécurité l'information génétique, permettant aux utilisateurs de recevoir une grande variété de services génétiques. Voici quelques exemples :

- Rapports et recommandations sur la nutrition, risques de maladies, cosmétique, régime, mise en forme
- Recherche de liens familiaux et clarification ancestrale
- Services de rencontres
- Sélection individuelle de vêtements, chaussures, paramétrage de climat dans la maison, destinations voyages et zones de résidences
- Différentes variations des rapports génétiques pour une groupe d'individus, par exemple des équipes sportives ou des groupes de travail
- Presque tous les aspects de la vie humaine sont influencés par la génétique, donc regardons ce que de nouvelles compagnies peuvent apporter en utilisant notre plateforme.

Plus que la multiplicité des services, un utilisateur se voit donner l'opportunité de faire un profit avec l'unicité de sa génétique en fournissant des réponses aux questionnaires aux compagnies pour leurs besoins de recherche. Ainsi, les données individuelles combinées à l'information génétique deviennent l'équivalent de produits de base ou des ressources minérales.

Santé

La santé moderne et la médecine personnalisée ne peuvent être imaginées sans l'utilisation de technologies génomiques. La plateforme va permettre à des patients de partager en toute sécurité l'information génétique utilisée en clinique avec le personnel médical :

- Les dosages individuels et l'intolérance aux médicaments (par exemple, le dosage individuel de l'anticoagulant warfarin basé sur des caractéristiques génétiques)

- Portée acceptable personnelle des paramètres biochimiques du corps (par exemple, les marqueurs PSA)

- La prédisposition génétique à certaines maladies (par exemple, un fort risque de dégénérescence maculaire et le besoin de recherche et prévention supplémentaire)

- La transplantation ou le don d'organes. Les utilisateurs peuvent, en sécurité, partager l'information concernant le type de leurs antigènes HLA qui détermine la compatibilité entre deux individus pendant une transplantation. Ainsi, il sera possible de créer une base de données sécurisée de donneurs et volontaires pour sauver des vies grâce à la transplantation.

Compagnie

Il y a deux types de compagnies, tout d'abord celles qui fournissent aux utilisateurs des services basés sur les données génomiques, et celles qui sont intéressées par l'obtention des données génétiques des utilisateurs pour conduire leurs recherches.

Le premier type est déjà décrit dans la section des cas d'utilisation des utilisateurs. Le second type de compagnies peut être décrit comme étant des utilisateurs acheteurs de données génomiques pour conduire leurs propres recherche et améliorer les propriétés consommatrices de produits, effectuer du ciblage génétique de produits et de publicités, comme les exemples suivants :

- Par exemple, une compagnie pharmaceutique prévoit de lancer un nouveau médicament qui agit contre la protéine de mutation du cancer. La compagnie peut trouver des utilisateurs dans le système, qui ont survécu à la maladie, pour les payer pour leurs données génétiques pour obtenir la fréquence des mutations dans un gène encodant la protéine qui est ciblée pour le principe actif.

- Une compagnie de produits de consommation planifie de pénétrer un nouveau marché et à besoin de tester comment les utilisateurs perçoivent le goût du produit. Il est connu que certaines saveurs provoquent des réactions parmi les porteurs d'une variante génétique spécifique du récepteur du gène gustatif. La compagnie peut envoyer une offre à travers le réseau pour étudier spécifiquement ces porteurs de la variante génétique et choisir une autre saveur, ou bien déterminer la fréquence d'apparition de cette variante dans différents marchés et adapter le produit en fonction des marchés.

La communauté scientifique

Pour la communauté scientifique, le système ouvre la voie au stockage, le partage et la réalisation de recherches avec des données génomiques variées. Parce que la plateforme n'est pas restreinte aux génomes humains, elle peut être utilisée pour stocker sécuritairement et analyser les données génomiques, par exemple, en lien avec l'Agriculture (plantes, animaux, micro-organismes).

En général, la présence d'écosystèmes mène à l'enrichissement de la communauté scientifique grâce à l'accès aux données d'une population générale, même sans référence à des questionnaires individuels. De plus, avec le consentement des utilisateurs, ils peuvent faire partie des recherches scientifiques.

La plateforme fournit également de la puissance de calcul distribuée, accessibilité qui va permettre d'analyser de gros volumes de données génétiques (similaire a un AWS pour travailler avec les données génétiques).

Chapitre 3

Partie Technique

3.1 Objets distribués

Concept de sous-systèmes distribués

Sous-système (Distribué) - ensemble de fractions des composants et processus du système (plateforme) qui peuvent être représentés en orienté-objet comme une entité qui possède une identité distincte et démontre un comportement visible à l'externe très bien défini.

Pour donner des détails spécifiques d'un sous-système, les aspects suivants doivent être décrits :

1. **Structure :** les éléments et processus composant le système
2. **Comportement externe** : interaction d'un sous-système dans son ensemble avec d'autres participants. En particulier :
    a. **Interface** un ensemble de requêtes possible au sous-système dans son ensemble
    b. **Actions** : actions entreprises par un sous-système concernant l'autre système de participants
3. **État interne :** l'état interne agrégé d'un sous-système

Les sous-systèmes peuvent être représentés comme des **quasis objet**, avec lesquels les autres participants peuvent interagir.

Note : le préfixe "quasi" indique des interactions avec des composants du sous-système existent, qui, a leur tour, ont des interactions complexes entre eux autres au titre de cette interaction réelle, pour que tout cela soit considéré comme une interaction avec un objet agrégé.

Note : Ci-dessous nous allons effectuer la distinction entre un sous-système et sa représentation en quasi objet.

Les interactions avec un sous-système peuvent être représentées comme la suite d'un petit nombre d'opérations de base :

- L'interface du sous-système, qui est, les actions des autres participants en regard du sous-système
- Les actions concernant les autres participants de la plateforme
- Les processus internes qui changent l'état d'un système

Processus internes

Les processus internes au sens large - représentent l'ensemble de tous les processus internes de chaque composant de sous-système et les interactions entre ces éléments.

Les processus internes (dans un sens plus restreint) - représentent les processus à l'intérieur d'un sous-système qui change son état interne. Une description complète des processus internes contient tous les comportements du sous-système excluant les problématiques spécifiques de son implémentation.

3.2 Les principaux sous-systèmes distribués de la plateforme

La plateforme a les niveaux d'organisation suivants :

- Couche du système de base (infrastructure critique)
- Niveau d'analyse et de stockage des données
- Interactions à haut niveau

Les sous-systèmes suivants sont inclus dans la plateforme :

1. Couche du système de base

   Interopérabilité de bas niveau sous-système de base d'échange de messages entre les nœuds du réseau. Il permet également la création des tables de hachage distribuées.

   Autorisations L'infrastructure des comptes et de la gestion des accès aux informations confidentielles.

2. Niveau d'analyse et de stockage des données

   Stockages Couche abstraite pour les accès au système de fichier distribué.

   Analyse Infrastructure pour les calculs distribués.

3. Interactions haut-niveau

   Requêtes sécurisées Un sous-système qui permet la création des offres pour l'achat des données génomiques qui peuvent être montré aux utilisateurs pertinents seulement.

   Opérations sur les données libres Fournit les outils pour utiliser les données libres (pas confidentielles).

   Plateforme de services externes Fournit l'API pour connecter les services gérés et centralisés externes à la plateforme. Organise le transfert sécurisé des données en utilisant les protocoles web conventionnels.

3.3 Couche d'interopérabilité de bas niveau

Réseau P2P distribué

Une des fondations de la plateforme est son réseau P2P basé sur le protocole Kademlia. Le réseau implique les ordinateurs des utilisateurs qui ont installé le logiciel Zenome.

Nœud - est un nœud du réseau P2P distribué qui représente un ordinateur d'un utilisateur sur lequel est installé le logiciel Zenome.

Note : En conséquence, un réseau global est créé entre les appareils participant au réseau. Cela représente un réseau virtuel dans lequel le "nodeId" de chaque participant est assigné, et qui n'a pas de lien avec les adresses IP réelles de l'appareil. Chaque nœud stocke une liste des nœuds les plus centraux, ou la distance entre les nœuds est calculée en fonction des "NodeId" et non de la notion de distance classique.

Les nœuds stockent les données en utilisant des tables de hachage distribuées.

Caractéristiques de la mise en œuvre d'un réseau distribué

Une spécification modifiée du protocole est utilisée. Les différences majeures sont les suivantes :

- Les nœuds peuvent échanger des messages aléatoirement entre eux. Un nœud est capable de transférer un message à un autre seulement en connaissant son "NodeId".

- Plusieurs tables de hachage peuvent exister à travers le réseau. Ces tables sont identifiées par une clef en chaînes de caractères.

Des règles différentes, de stockage et suppression des valeurs, peuvent être appliquées à des tables différentes.

- Les données transférées d'un nœud à un autre sont encryptées (voir ci-dessous).

Échange de messages dans un environnement distribué

Ainsi, les participants du réseau P2P sont capables d'échanger les messages suivants :

PING	Vérifie qu'un nœud est toujours existant
STORE(T,K,V)	Stocke la valeur 'V' de la clef 'K' dans la table 'T' dans le nœud recevant le message
FIND_NODE(N)	Un nœud recevant ce message enverra les données concernant les nœuds les plus proches du nœud 'N' parmi les nœuds qu'il connait
FIND_VALUE(T,K)	Si la paire '(K,V)' est stockée dans le nœud en réception du message, il enverra la valeur 'V', sinon les données concernant les nœuds connus les "plus proches" du fichier.
SEND(M,N,D?)	Envoie le message 'M' qui peut également contenir la donnée 'D' au nœud 'N'. 'FIND_NODE' est utilisé pour localiser le nœud.

Note : Ce niveau d'interaction est au niveau de transport de la plateforme.

3.4 La “Blockchain”

Information de base

La plateforme utilise la “blockchain” Ethereum qui représente une machine virtuelle décentralisée unique (EVM). Le système logique désiré peut être mis en place en utilisant les “smart contracts”.

Citation : “Un contrat est une collection de code (ses fonctionnalités) et de données (ses états) qui résident à une adresse spécifique sur la “blockchain” Ethereum. “
--- Introduction aux “Smart Contracts” (manuel Solidity)

Ainsi, les “smart contracts” sont capables de stocker des données. Pour la donnée qui est dans la “blockchain”, le type de système abstrait décrit ci-dessus est encore valide (mais avec des limitations).

Un sous-système pour travailler avec la “blockchain”

Note : Attention : les détails de mise en place peuvent être différents dépendamment de la plateforme utilisée. Les descriptions ci-dessous est pertinente pour la plateforme PC.

Pour fournir un accès à la “blockchain”, le logiciel du nœud inclus la mise en place complète d’un nœud Ethereum donnant accès par ‘JSON-RPC 2.0’.
Les clés secrètes sont stockées dans le stockage crypté. Un utilisateur est invité à définir un mot de passe au premier lancement de l'application

Note : Bien qu’il soit techniquement faisable d’utiliser un compte déjà existant, il est recommandé d’en créer un nouveau.*

L’interface de l’application permet de :

1. Créer un nouveau compte
2. Importer un compte existant
3. Établir la sauvegarde du stockage confidentiel

Note : Attention : Il est recommandé de mettre en place une sauvegarde dans le “cloud” pour permettre de garder un accès au compte, dans le cas où l’ordinateur serait physiquement indisponible.

Note : Le logiciel permet d’Avoir accès au nœud Ethereum par ligne de commandes. Cette fonction est prévue initialement pour le débogage. Il est n’est pas recommandé d’utiliser les lignes de commandes si vous ne comprenez pas leur raison d’être.

Tokens internes

Les interactions économiques à l'intérieur du système sont fournies en utilisant les tokens ZNA internes. Ils sont des tokens Ethereum valides et peuvent être achetés et vendus sur des échanges.

Concept de compte

Un compte dans la plateforme permet à l’utilisateur d’interagir avec le système en jouant plusieurs rôles en même temps. Chaque rôle dans un compte correspond à un “smart contract” distinct dans la “blockchain”.

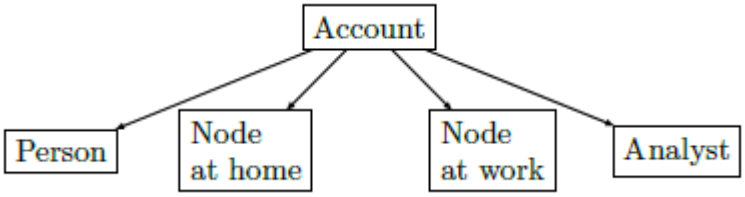


Figure 4. Le compte d'un utilisateur qui a fourni son génome (rôle de Personne) à la plateforme, qui travaille comme bio-informaticien et utilise deux noeuds de calcul, un à la maison, l'autre au travail.

Rôle - est une caractéristique du participant au système au regard des types spécifiques d'interactions dans lesquelles il prend part. Chaque participant peut jouer plusieurs rôles simultanément.

Note : Un sous-système séparé est responsable de la distribution des rôles.

3.5 Réseau distribué de stockage des données

Les principes des opérations d'un réseau distribué

Le réseau distribué basé sur le protocole DHT Kademlia est utilisé pour stocker les données de type arbitraire. Les nœuds de ce réseau sont représentés en enregistrant les fournisseurs de ressource, qui obtiennent un bonus en stockant des données et en réalisant les calculs requis. Le participant qui a exporté les données paie une récompense pour le stockage des données, mais, dans de nombreux cas pratiques importants, ces récompenses sont totalement ou en partie subventionnées par le système. Par exemple, il existe un mécanisme de subvention du stockage des données à grande valeur scientifique.

Les unités de stockage sont représentées par un bloc arbitraire de données et une clef unique à partir de laquelle il peut être accessible.

Note : la taille d'un bloc de données est restreinte pour limiter les mauvaises utilisations.

Support des services externes. Une possibilité d'interaction avec les nœuds Storj et FileCoin, tout autant que les centres de stockage gérés, vont être mis en place.

Assurance de la fiabilité du stockage des données

Pour assurer la fiabilité du stockage des données, en considérant que les noeuds peuvent librement se connecter et se déconnecter du réseau, les données sont stockées indépendamment sur plusieurs noeuds en même temps. Le type de données stockées détermine le nombre de noeud sur lesquels l'information est dupliquée.

Tableau 6. Exemples de paramètres de donnée génétique

	Raw	Processed
Format	fastq/bam	gvcf/vcf + bed/23me(txt)
Size	50Gb	2Gb
Nodes	≥ 3	≥ 5

Lorsque le nombre de nœuds change, les données stockées sont redistribuées pour satisfaire les requis pour le nombre minimum de nœuds.

Puisque tout le monde qui veut, peut devenir un nœud du réseau distribué, ce n'est pas sécurisé de croire que le nœud stocke réellement le fichier lorsqu'il l'affirme. Ainsi, une vérification de disponibilité des données est réalisée périodiquement en respectant le protocole de la couche de sécurité. Les données sur les résultats de vérification sont introduites sur la "blockchain" et peuvent être la raison pour donner une récompense ou revoir la notation d'un nœud.

Assurance de la confidentialité des données stockées

La confidentialité des données stockées dans un tel réseau distribué devient possible grâce à l'application d'encryptage asymétrique. La méthode actuelle d'encryptage est déterminée par le protocole de sécurité correspondant.

Note : Il doit être souligné qu'un mécanisme de subvention ne peut être appliqué pour le stockage de données encryptées.

Fonctionnalités spécifiques au stockage d'information génomique

L'appartenance à un individu spécifique de certains fragments de génome peut être identifiée sans ambiguïté si le fragment est suffisamment grand. C'est pourquoi les génomes, dans le réseau distribué, sont stockés par petits fragments.

La fragmentation du génome est toujours réalisée en fonction du génome de référence. Pour chaque génome de référence (par exemple, une autre version du génome humain de référence ou un génome d'un organisme d'une espèce différente) la fragmentation est choisie seulement une fois, et chaque fragment est assigné avec un identifiant qui est unique pour ce génome de référence.

3.6 Les données personnelles de l'utilisateur

Enregistrement - est une unité minimale d'échange des données. Il n'est pas possible de transférer (par exemple, de vendre pour une récompense) une partie de l'information d'un enregistrement. Il est possible de créer un nouvel enregistrement qui contiendra seulement une portion des données, mais dans un tel cas, ce nouvel enregistrement cela significativement moins intéressant pour des acheteurs potentiels du fait de sa note basse.

Le schéma des données détermine quelle information et sous quel format un enregistrement doit les contenir. L'usage de schémas est un outil flexible pour l'unification du format d'échange des données entre les participants du réseau. Un identificateur de schémas de données peut être n'importe quelle chaîne qui permet aux individus de trouver sa description sur le réseau, par exemple, l'adresse d'une URL ou d'un "smart contract" avec une description. Il doit être compris que l'objectif premier d'un schéma de données est d'unifier la création de ce qui est inclus dans les données entre les acheteurs et les vendeurs de ces données.

Caractéristiques des questionnaires

Remplir des questionnaires personnels augmente substantiellement la valeur des données génomiques. Un grand nombre de questionnaires peuvent exister, il est donc nécessaire d'introduire le concept des caractéristiques d'un questionnaire.

Caractéristique d'un questionnaire - représente une description complète de tous les champs du questionnaire et les valeurs permises de ses champs.

Officiellement, les caractéristiques d'un questionnaire contiennent une référence à l'auteur, une description et une liste classée des enregistrements dont chacun correspond à un champ du questionnaire.

Champ numérique une réponse d'un utilisateur doit être comprise dans la gamme de valeurs permises *[a,b]*. La réponse est codée par l'entier non signé, où le décompte commence par la limite gauche de l'intervalle.

Note : Une réponse à ce type de question est disponible publiquement puisque cela ne menace pas la confidentialité de l'information.

Choix multiple La réponse de l'utilisateur est codée comme un entier non signé correspondant au numéro de série de la réponse dans une liste. Si la valeur est égale à zéro, alors l'utilisateur a préféré ne pas répondre à la question.

Note : Attention : une réponse à ce type de question est disponible publiquement puisque cela ne menace pas la confidentialité de l'information.

Réponse à chaine de caractère La réponse de l'utilisateur est stockée dans une chaîne.

Note : Une réponse à ce type de question est de l'information confidentielle.

Questionnaire rempli - est une structure de données contenant les réponses de l'utilisateur aux questions du questionnaire.

Note : *cette structure est complètement encryptée au besoin. Elle est au moins stockée sur l'ordinateur de l'utilisateur. Un utilisateur peut exporter une sauvegarde encryptée sur la "blockchain" s'il le désire.*

Les requêtes de données personnelles de la part des utilisateurs ont des fonctions distinctives :

- Pas tous les utilisateurs ne satisfassent aux critères spécifiques de la recherche. Pour vérifier si les données sont appropriées, un accès à l'information personnelle est requis.

- L'information personnelle ne devrait pas être transférée en dehors de l'ordinateur de l'utilisateur que ce soit explicitement (directement) ou implicitement ("warrant canary").

- Dans le cas où l'utilisateur satisfait le critère, il recevra une offre pour partager ses données personnelles pour une récompense.

Conclusion 1 : Puisque le transfert d'information n'est pas permis, la vérification devrait être réalisée dans un environnement isolé. Le code exécutable dans cet environnement doit avoir un accès complet à l'information privée mais ne peut pas interagir avec d'autres parties du système.

Conclusion 2 : Le résultat de l'exécution de cette fonction isolée ne peut être envoyé comme une réponse à un client potentiel puisqu'il menace la confidentialité.

Conclusion 3 : Il est raisonnable de choisir le lot de données personnelles restreintes, puisqu'il est nécessaire pour rendre une conclusion, lors d'une étape de vérification distincte. L'utilisateur verra sur quelles données la décision a été rendue sur l'écran d'offre d'échange.

Conclusion 4 : Les données qui seront transférées explicitement sont listées sur l'écran d'offre d'échange. Seulement les données qui ont été requises pour faire une vérification pourront être transférées directement, et la liste de ces données sera également disponible à l'utilisateur.

Requêtes de données personnelles :

- À la première étape un accès est requis pour les données personnelles d'un utilisateur Seulement les identifiants de données qui ont été listés dans cette requêtes seront disponibles pour la fonction de vérification.

- À la seconde étape, le code pour la fonction de vérification est exécuté dans un environnement isolé, et son résultat constitue l'offre formulée en regard de l'échange de données ou de l'annulation de l'offre. Dans le dernier cas, aucune donnée n'est transférée.

- Dans le cas précédent, l'utilisateur est notifié lorsque l'offre d'échange est formulée, alors il peut examiner l'offre, la liste des données utilisées lors de la vérification et la liste des données qui seront transférées si l'utilisateur accepte l'offre.

- Si l'utilisateur rejette l'offre, aucune donnée ne sera transférée.

- Si l'utilisateur accepte l'offre, seulement les données explicitement montrées à l'utilisateur seront transférées.

Bibliographie

[1] NHGRI. Talking Glossary of Genetic Terms. Word «Genome».
URL: https://www.genome.gov/glossary/index.cfm?id=90.

[2] Venter J.C., Smith H.O., Adams M.D. “The Sequence of the Human Genome”.
In: Clinical Chemistry 61.9 (2015), pp. 1207–1208.
URL: http://clinchem.aaccjnls.org/content/61/9/1207.long.

[3] Adams M.D. Venter J.C. Smith H.O. “The Sequence of the Human Genome”.
In: Science 291.5507 (2001), pp. 1304–1351. ISSN: 0036-8075.
DOI: 10.1126/science.1058040.
URL: http://science.sciencemag.org/content/291/5507/1304.

[4] Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).
URL: www.genome.gov/sequencingcostsdata.

[5] Melissa Gymrek et al. “Identifying Personal Genomes by Surname Inference”.
In: Science 339.6117 (2013), pp. 321–324. ISSN: 0036-8075.
DOI: 10.1126/science.1229566.
URL: http://science.sciencemag.org/content/339/6117/321.

[6] H. Christina Fan et al. “Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood”. In: Proceedings of the National Academy of Sciences 105.42 (2008), pp. 16266–16271.
DOI: 10.1073/pnas.0808319105.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2562413/.

[7] Sboner, Andrea and Mu, Xinmeng Jasmine and Greenbaum, Dov and Auerbach, Raymond K. and Gerstein, Mark B. “The real cost of sequencing: higher than you think!” In: Genome Biology 12.8 (Aug. 2011), p. 125. ISSN: 1474-760X.
DOI: 10.1186/gb-2011-12-8-125.
URL: https://doi.org/10.1186/gb-2011-12-8-125.

[8] Rachel R. J. Kalf et al. “Variations in predicted risks in personal genome testing for common complex diseases”. In: Genet Med 16.1 (Jan. 2014), pp. 85–91. ISSN: 1098-3600.
DOI: 10.1038/gim.2013.80.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3883880/.

[9] Karen Norrgard. “Calculation of Complex Disease Risk”. In: Nature (2008).
URL: https://www.nature.com/scitable/topicpage/calculation-ofcomplex-disease-risk-756.

[10] Kathryn Nave. “How Craig Venter is fighting ageing with genome sequencing”. In: WIRED UK (2016).
URL: http://www.wired.co.uk/article/craig-venter-human-longevity-genome-diseases-ageing.

[11] Amalio Telenti et al. “Deep sequencing of 10,000 human genomes”. In: Proceedings of the National Academy of Sciences 113.42 (2016), pp. 11901–11906.
DOI: 10.1073/pnas.1613365113.
eprint: http://www.pnas.org/content/113/42/11901.full.pdf.
URL: http://www.pnas.org/content/113/42/11901.abstract.

[12] Luke Timmerman. “Google Translate Star Leaves Venter’s Human Longevity For Illumina-Backed Grail”. In: Forbes (2016).
URL: https://www.forbes.com/sites/luketimmerman/2016/09/27/google-translate-star-leaves-venters-human-longevity-for-illumina-backed-grail.

[13] João Sá Sousa et al. “Efficient and secure outsourcing of genomic data storage”. In: BMC Medical Genomics 10.2 (July 2017), p. 46. ISSN: 1755-8794.
DOI: 10.1186/s12920-017-0275-0.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5547444/.

[14] Zachary D. Stephens et al. “Big Data: Astronomical or Genomical?” In: PLOS Biology 13.7 (July 2015), pp. 1–11.
DOI: 10.1371/journal.pbio.1002195.
URL: https://doi.org/10.1371/journal.pbio.1002195.

[15] Erika Check Hayden. “Privacy protections: The genome hacker”. In: Nature (2013).
URL: http://www.nature.com/news/privacy-protections-the-genomehacker-1.12940.

[16] Laure-Anne Pessina. “Reconstructing a face from DNA: an EPFL alumnus takes the stage at the 2016 TED Conference”. In: School of Engineering (Federal Institute of Technology Lausanne) (2016).
URL: http://sti.epfl.ch/page-129921-en.html.

[17] Melanie Swan. “Health 2050: The Realization of Personalized Medicine through Crowdsourcing, the Quantified Self, and the Participatory Biocitizen”. In: Journal of Personalized Medicine 2.3 (2012), pp. 93–118.
ISSN: 2075-4426.
DOI: 10.3390/jpm2030093.
URL: http://www.mdpi.com/2075-4426/2/3/93.

[18] Patrick Lin. “Blockchain: The Missing Link Between Genomics and Privacy?” In: Forbes (2017).
URL: https://www.forbes.com/sites/patricklin/2017/05/08/blockchain-the-missing-link-between-genomics-and-privacy.

[19] Justin Zimmerman. “DNA Block Chain Project Boosts Research, Preserves Patient Anonymity”. In: CoinDesk (2014).
URL: https://www.coindesk.com/israels-dna-bits-moves-beyondcurrency-with-genes-blockchain/.

[20] D.S. Brama. “Method, System and Program Product for Transferring Genetic and Health Data”. US Patent App. 14/218,865. July 2015.
URL: https://www.google.com/patents/US20150205929.

[21] Melanie Swan. Blockchain: Blueprint for a New Economy.
URL: https://www.goodreads.com/book/show/24714901-blockchain.

[22] Matthew Herper. “Surprise! With $60 Million Genentech Deal, 23andMe Has A Business Plan”. In: Forbes (2015).
URL:
https://www.forbes.com/sites/matthewherper/2015/01/06/surprise-with-60-million-genentech-deal-23andme-has-a-business-plan.

[23] “23andMe, Pfizer to Launch Inflammatory Bowel Disease Genetics Study”. In: GenomeWeb (2014).
URL:
https://www.genomeweb.com/clinical-genomics/23andme-pfizer-launch-inflammatory-bowel-disease-genetics-study.

[24] “Genetic Wild West: 23andMe Raw Data Contains 75 Alzheimer’s Mutations”. In: Alzforum (2017).
URL:
http://www.alzforum.org/news/community-news/genetic-wildwest-23andme-raw-data-contains-75-alzheimers-mutations.