International Journal of Engineering & Technology, 7 (4) (2018) 2585-2593

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.16893
Research paper

Genomics big data hybrid depositories architecture to unlock precision medicine: a conceptual framework

Ummul H. Mohamad [1] *, Mohamad T. Ijab [2], Rabiah A. Kadir [3]

[1] *Institute of Visual Informatics, Universiti Kebangsaan Malaysia, 43600 UKM, Bangi, Selangor, Malaysia*
**Corresponding author E-mail: ummulhanan@ukm.edu.my*

Abstract

As the genome sequencing cost becomes more affordable, genomics studies are extensively carried out to empower the ultimate healthcare goal which is the precision medicine. By tailoring each individual medical treatment through precision medicine, it will potentially lead to nearly zero occurrence of the drugs side effects and treatment complications. Unfortunately, the complexity of the genomics data has been one of the bottlenecks that deter the advances of healthcare practices towards precision medicine. Therefore, based on the extensive literature review on the data driven genomics challenges towards precision medicine, this paper proposes two new contributions to the field; the conceptual framework for the genomics-based precision medicine and the architectural design for the development of hybrid depositories as the initial step to bridge the gap towards precision medicine. The genomics big data hybrid depositories architecture design is composed of few components; storage layer and service layer interconnected system such as visualization, data protection modeling, event processing engine and decision support, to carry out their purpose of merging the genomics data with the healthcare data.

Keywords: *Architecture Design of Hybrid Depositories; Data Driven Genomics; Personalized Medicine Framework.*

1. Introduction

Personalized medicine or now coined better as precision medicine, intends to ensure precise medical treatment can be delivered to each patient [1]. The accuracy on the customization of the medical treatment is believed to achieve a level in which it could eliminate any pre-incidences of any drugs' side effects, drugs complications, incorrect drugs dosage and to accurately determine the individual's disease predisposition [2]. The concept of precision medicine arises from the knowledge of human genetic variation. Three billion DNA base pairs that make up the whole human genome contain more than 20 000 protein coding genes [3]. Thus, genetic variation on these protein coding genes could be responsible for the different pharmacological responses despite a similar treatment regime. In addition, sequencing of the whole human genome yielded more than 100 gigabytes of data [4]. Thus, we can expect more than petabytes of genomics data to be generated as we are paving our way towards precision medicine [5], [6].

[7] claimed that the genomics study has now progress better with the advances in computational biology (also known as bioinformatics). Bioinformatics tools have allowed for easier genetic analysis, annotation, comparison, data interpretation and visualization of analyzed data. For example, Next Generation Sequencing major service provider (Illumina) has integrated cloud computing to aid sequencing projects [8]. Unfortunately, the outburst of genomics big data within a short span of time has surpass the capability of the currently available software and tools as many are developed to best suit the typical genomics data [9].

As the cost to sequence genome decreases, numerous extensive research are conducted to help understand the disease genomics leading to an exponential deluge in the genomics data. Genomics big data can be distinguished from the common data based on the 5V's parameter; volume, variety, velocity, veracity and value [10], [11]. The amount of genomics big data may reach petabytes and beyond in volume, surpassing the typical computational power for data processing. The second parameter, variety, defined that genomics big data accumulates from different data structures and sources; to which standardizing the data becomes a challenge.

Moreover, the speed of genomics big data accumulation (velocity) is also exponentially increasing in real time and continuous manner. Eventually, biological scientist is unable to keep up with the current and recent findings of genomics research without the help of big data tools. Veracity, another parameter of genomics big data, ensures data trustworthiness to eliminate bad data from the system as it may affect the downstream application of the genomics big data. Last but not least, genomics data comes with a value that must be statistically relevant before the data can be considered to guide the clinical decision making [12].

Upon the completion of Human Genome Project in 2003, new knowledge on human diseases and new challenges are unveiled concurrently [13]. For instance, patients were more exposed to their health information which will indirectly impact the lifestyle, health decision and treatments. Despite the extensive growth of genomics big data and advances of technology, we are still far from making precision medicine a reality. Thus, the objectives of this paper are to identify the challenges of the genomics big data towards precision medicine and discuss the framework to unify the concepts and ideas of genomics-dependent precision medicine system, focusing on the hybrid depositories as the first step to bridge the gap towards precision medicine.

2. Literature review

The literature review was conducted to collect and analyzed all the relevant papers in the field by the means of a structured search for literatures. The last search for papers was conducted in February 2018. In addition, the scope mainly was cross-fields studies, majorly in ICT, genomics, and big data but were not limited to these fields. Precision medicine, by definition, is a genetic-based approach that assess the individual health risk with the intention to design a precise, customized health plan to accurately manage a patient's medical treatment [14]. In line with this definition, several related keywords were used to search the online databases to facilitate the preparation of the groundwork for the subsequent literatures. These include data driven genomics, big data analytics, genomics data challenges and healthcare decision making. Main databases from major publishers were used to search for the related publications, such as Scopus (www.scopus.com), Science Direct (www.sciencedirect.com), Springer (www.springerlink.com), IEEE (http://ieeexplore.ieee.org/Xplore) and PubMed database (https://www.ncbi.nlm.nih.gov/pubmed). Initially, a total of 306 papers were identified. Through a quick check on the papers' contents, the journal articles were included or excluded from the analysis. To obtain a manageable number of papers for literature analysis, filtration through several criteria was performed. Some of the criteria include (i) removal of paper that were published before 2013 (however, several papers were included later to emphasis on the development of initial research), (ii) categorization of papers according to the issues such as challenges of genomics in the path of precision medicine, the technological advances of genomics and the healthcare data management and (iv) papers from cited references were also included as a secondary literature source.

Hence, as we are paving the steps towards precision medicine, there are many challenges that we need to overcome such as issues with the incompetent genomics big data storage, difficult transformation of the complex genomics big data into an understandable form, ineffective genomics big data management and high concern on the genomics big data privacy and protection.

2.1. Incompetent genomics big data storage

[6] Mentioned that data storage is one of the major restriction to having reliable collections of genomics big data. The cost to store the genomics data is higher when compared to the cost to sequence the genomics [15] due to the fact that genomics data storage require massive computing power, advanced software tools and feasible computational algorithms to support many downstream applications such as genomics data assembly, data compression and data analysis [16]. In addition, typical localized servers are no longer sufficient to handle the genomics big data as they lacked capacity, flexibility as well as mobility. This situation led to many genomics scientists resorting to the solution of storing the actual biological sample and performs re-sequencing of the genomes rather than storing the data itself. Therefore, there is a great likelihood that these practices are contributing to the endless gaps in understanding the genomics and diseases. In other words, unkempt genomics data will easily result into potential data loss [17].

Moreover, genomics data storage problem is also caused by the lack of depositories [18]. For instance, by 2018, at least 15 petabytes of genomics big data will be generated from 450 thousand individual genomes [19] solely from International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) research projects. This depicts the growth of the genomics big data in the near future and highlighted the strong need to fully equip the genomics big data storage with competent depositories, advance technology and compatible supporting tools before we can make sense of the genomics big data [20].

Although cloud computing has been the best storage solution to address the lack of depositories [21], we still encounter problems that contributed to the ineffective genomics big data storage [22]. Some of the problems were the accumulation of data in different data formats [23], imposed rules to restrict the access the depositories [24], data duplication and redundancy [25] as well as data standardization. In 2017, [25] suggested a deduplication framework to eliminate the data redundancies. This framework was based on three components; client, network and server, in which the client layer eliminate the duplication occurring from a client input while the network layer utilizes redundancy elimination devices to remove the redundancies from multiple clients. Next, the server layer exclude redundant data that comes from the different networks. However, the efficiency of eliminating data duplication and redundancy were depending on several factors such as the design of system capacity, characteristics of data sets, processing power and deduplication time [26]. Hence, there is yet to be any optimally best solution to address this issue.

In addition, another essential issue that correlated to the incompetent genomics data storage was application of the suitable data compression method. For now, we may have general algorithms such as Lempel-Ziv [27] and GDC2 algorithm [28] to name a few. For instance, GDC2 claimed to be able to compress 1092 human diploid genomes almost 10,000 times with the speed of 200 MB/s. In other words, 1K human genome file size was only about 700MB after compression compared to 6.7 TB file size when uncompressed. Nevertheless, we may need to deploy customized algorithms to speed up the compression time and minimize the storage footprint of potential zettabytes worth of genomics big data in the future.

2.2. Difficult transformation of the complex genomics big data

Another challenge of data driven genomics is the complexity of genomics data [29]. Genomics data is not straightforward, harder to define and often involve numerous interacting variables [30]. A simple example can be explained in breast cancer genomics. We need to analyze and make sense of the (i) genomics data (for example, involving BRCA1 or BRCA2 genes mutations) which are responsible for increased risk of breast cancer [31], (ii) expression of HER2 protein [32] which are often associated with a greater risk of breast cancer recurrence, (iii) analysis of gene regulatory network to determine the expression of other related oncogenes and tumor suppressor genes, (iv) analysis of pharmacogenomics datasets to depict the suitability for the drug therapy and (v) detailed family histories, to name a few, before a conclusive decision can be made for precision medicine treatment of breast cancer.

To portray the genomics data complexity, several parameters are discussed. Among them are the scale of genomics data, the forms of genomics data and the information relatedness of the genomics data [33]. The parameters need to be embraced to ensure precise clinical prognostics can be made possible. The first parameter, which is the scale of genomics data, involves many dimensions such as genes expression, interacting proteins, gene copy numbers, and metabolic pathways. For instance, cancer disease has often linked to the abnormal expression of multiple oncogenes (cancerous genes) and repression of tumor suppressor genes, affecting the critical pathway of genes regulations that kept the cell in its normal state [34]. When combined together with other external factors such as environmental exposure and lifestyle, this may contribute to tumor aggressiveness and disease complexity. To add, [10] claimed that by 2025, up to 40 Exabyte is required just to store human genome data. Hence, it is possible that the genomics big data may reach beyond zettabytes in this 4IR era.

The second parameter which is the forms of genomics data are divided into four common types: sequencing data, annotations, quantitative data and read alignments [35]. In addition, these four types of genomics data come in different data formats. These many types and formats are responsible for the genomics data complexity, justifying why the conventional approaches tend to oversimplify and focus on individual data rather than the whole datasets [36]. Genomics sequencing data consists of the nucleotide sequence of the complete set of genes, contigs and transcript, and are usually in FASTA (unindexed) or 2bit format (indexed).

Meanwhile, annotation data encompasses the description to known features in the genome such as the conserved regions, coding genes, start codons, transcription activation factors and many more. Annotation data format is in BED, GTF2, GFFF3, PSL (unindexed) or BigBed (indexed). The quantitative data, on the other hand, is genomics data with numerical value in relation to the chromosomal locus (position). As example, there are some regions within the genome that are conserved across different organisms at certain degree. This kind of data is available in bedGraph or wiggle unindexed format and BigWig indexed format. Meanwhile, read alignment data consists of short sequences data that match identical sequence in the genome records using mapping. Read alignment data formats are bowtie, SAM, PSL (unindexed) and BAM (indexed).

In addition, although most of the genomics data is made available in the electronic health reports (EHR), the data is not readily accessible by the clinical decision system (CDS). Castaneda et al. (2015) also highlighted that the current system experienced difficulties such as the inability to store raw genomics data, the nonexistent link between the genetic abnormalities/mutations data with the pathological/clinical syndromes and the incomplete patients health data that deter the speed of clinical decision making.

Next parameter, the information relatedness of the genomics data, is often overlooked and viewed as an obstacle since it is difficult to explore, understand and describe the multiple, interacting and conflicting data among genetics, gene expression, clinical markers and variation patterns. The potentially relevant data must be examined in depth to uncover patterns and trends that can distinguish subtle phenotypes in the disease genomics. Describing genomics data using inconsistent terms is one of the factors that discourage the understanding of the relations between genetics and diseases phenotypes [38]. This can simply be seen from the different way biology scientists and clinicians classify diseases and describe symptoms. The situation of semantic irregularity tends to make it difficult and tedious to populate the required data [39]. For example, types of cancer such as carcinoma or sarcoma are not tagged as cancer without the standardization of data ontologies. Due to that, specific ontologies standards have been developed to support the growth of many genomics databases. Examples of the standardized ontologies include Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Medical Subject Headings thesaurus (MESH). In addition, the usefulness of biological ontologies is not only focused for better data classification [40], it also portrays biological data in a more understandable form, ensure effective data organization, supports statistical analysis and web simple search [41]. Therefore, it is essential that the data input followed the standardized ontology to minimize the gap in understanding genomics diseases.

2.3. Ineffective genomics big data management

Another major challenge towards precision medicine is genomics data management. This includes genomics data integration, data processing and data analysis. Difficulty in data integration mainly occurred due to the different data types [42]. Although the data may be in the form of digital data, the unavailability of tools to easily unify and incorporate these different data types becomes a restriction. To accomplish the objective of precision medicine, we need to be able to effectively integrate the genomics sequencing data with the electronic medical records [43], [44].

Moreover, genomics data integration is deterred when the data are kept isolated and enclosed from research purposes, data is managed poorly as well as data is not updated [45]. These practices discourage the application of precision medicine in healthcare practices and needs to be addressed in priority to ensure precision medicine vision is achieved [46]. An attempt using web-based ODMedit has successfully created more than a thousand data models with uniform semantic annotations from large public metadata repository [47]. In principal, ODMedit is practical tool to facilitate the data integration. However, there were some drawbacks of the tool such as the application was not fully automated, it worked only on structural metadata and not on raw genomics data and data completeness was not considered. In addition, ODMedit needed supportive effort by the scientific community to fully achieve its aim due to the semantic richness of patients' metadata. Unlike ODMedit, p-medicine, a medical informatics platform, was able to work with variable format of heterogenous patients data [48]. These data types which include imaging, genomic and clinical records data can also be assembled, linked and subjected to the analytics tools to better understand the data. [49].

Besides that, the genomics data processing and analysis are dependent on the compatibility and effectiveness of the available analytical tools and expertise [50]. Laboratory scientists mainly focused on getting the data from their research while bioinformatics software developers mostly lacked the experience in understanding the hands-on part [51]. This eventually led to the emergence of software, tools and computational infrastructures that did little to solve biological problems [52]. To make it worse, data analysis rarely becomes a priority as majority of the funds are spend to generate the data. This condition will eventually dampen collaborations, innovative development of software and also deters the potential to fully explore and unlock the personalized genomics data and use them to our advantage.

Subsequently after obtaining genomics data, the data will undergo more downstream processes such as genome assembly, annotation and alignment. These processes will generate different types of data in different formats. Genome assembly contains fragmented genome reads that requires very large memory capacity to assemble and arrange the genome sequences. The unavailability of sufficient computing power limits the progress of many genomics research [15]. This led to initiatives by few genomics sequencing service providers such as Pacific Biosciences (PacBio) and Oxford Nanopore to come up with long-read sequencing technologies [53] that allow for a maximum of 100-fold fragmented genome assembly without compromising the quality and quantity of data and at an affordable computational cost [54].

The assembled genome is then aligned with the reference genome to facilitate the genes annotation and comparison. Previously, the alignment requires comparison and alignment of the two sequences directly; however this takes longer time and more memory space. To overcome this problem, new tools [55] such Basic Local Alignment Search Tool (BLAST), Spliced Transcripts Alignment to a Reference (STAR), Burrows-Wheeler Aligner (BWA) and Bowtie are adopted to increase the speed to align genomes. The adoption of these new tools utilizes a two-step seed-and extends strategy [56]; generating indexes according to the query sequences or organize the database into compact binary files for quicker alignment time [57]. It is foreseen that there will be a greater need in the future to develop new algorithms as a game changer to execute the genomics data processing in a much practical way.

As the demand to have more innovations and transformative tools to maximize and drive the integration of big data and data science into genomics, the National Institutes of Health (NIH) launched Big Data to Knowledge Initiative (BD2K) to ensure improvement in many genomics analysis tools is parallel to having trained biologist [58]. It is hoped that this initiative will encourage collaboration of many experts to address genomics big data issues and its solutions. Therefore, we must encourage more interactions, networking, knowledge and expertise sharing among the genomics scientists, bioinformaticians, data scientists, clinicians and IT experts, potentially through cloud collaborating platform [59] as the answer to understanding disease genomics sometimes are revealed from viewing the diseases from different angle.

2.4. High concern on the genomics big data privacy and protection

Addressing the concern on the genomics data security, certain data depositories such as the Biobank applied a different approach to prevent data loss and data corruption. Biobank contained data from nearly 200 000 donors which include data from personal health, genomics, proteomics and bio-specimens. Involvement of a

massive scale of data will definitely led to potential privacy threat which will interfere with the data security [60]. Therefore, Biobank utilizes Bio-PIN or 'Biological Personal Identification Number' is generated by the individual's unique non-genotypic single nucleotide polymorphisms (SNPs). The bio-specimen will be registered only with Bio-PIN and did not include the general donor identity data (to ensure confidentiality). Although the resulting PIN code cannot be linked back to the individual, the samples can still be distinguishable from each other [61].

In addition, [62] also highlighted that despite using either public or private cloud computing platforms, there will still be concerns on data privacy and security. This is because the advances in technology will always create opportunity for data manipulation [63] and integrative data requires broader security spectrum which will exert more pressure on data security components in the cloud [64]. Besides that, [65] mentioned that the genomics data protection is badly needed to encourage data sharing prior to the enactment of precision medicine. This is because concerns are raised over the fact that the genomics data can reveal more than just an individual's genetic information such as the health status, drug responses and disease predisposition, but also exposed the information on related family members [66]. Another extra measure such as having data protection policies must be derived to prevent any biocrime or data abuse that will endanger the safety of each individual. [66] described GeneCloud, a secure cloud computing as an alternative to public cloud computing with the added data security. Data is secured through the execution inside a secure sandbox that prevents any disclosure of sensitive data [67]. Similarly, an added data control and security can also be found in Bionimbus, a cloud computing platform [68] used to study the acute myeloid leukemia sequencing with data comparable to the 12 hours alignment-time using 8 CPUs.

In addition to that, data security may also be achieved through common approaches such as encryption [69], multiple-factors authentication [70], authorisation limits and blockchain monitoring [71], [72]. Moreover, there is also a need to implement policies to protect genomics data. The unavailability of these policies discourages voluntary genomics data capture, leading to insufficient data collection [73]. Without sufficient data, the findings will not be significant and relevant to be applied into a medical practise. These policies will also protect the individual rights on health related matter such as insurance to stop the insurance provider from using the genomics data to deny the insurance claims and entitlements [74].

3. Conceptual framework to bridge the gap towards precision medicine in the genomics context

[75] described the 4th industrial revolution (4IR) as an impactful fusion of the advanced technologies with the physical, digital and biological worlds. 4IR heavily relates to the genomics big data in terms of the utilization of the increasing volume, variable data that merge not only the genomics data, but include the clinical, environmental, and lifestyle information from individuals to larger populations. This had change the landscape of the precision medicine from hypothesis driven to data driven approach. As we are pacing our way in the 4IR era, the advances in technologies and bioinformatics have given rise to many important tools for data capture and storage, collaboration, analysis and decision purposes. Fig. 1 showed the conceptual framework to bridge the gaps towards precision medicine in the genomics context: (i) improve tools and pipelines, (ii) expand the cloud collaboration platform, (iii) create hybrid depositories and (iv) develop automated precision medicine system [76]. The following section will discuss more on the architectural design of the hybrid depositories as the main focus of this framework.

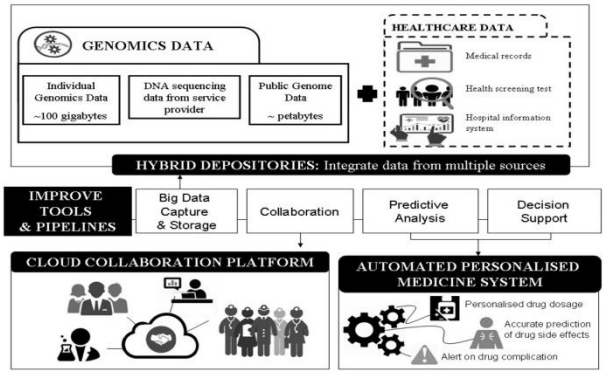

Fig. 1: Framework to Bridge the Gap Towards Precision Medicine in Genomic Context [76].

3.1. Improve tools and pipelines

Technology breakthrough has encouraged new, innovative developments and improvements to make data acquisition, management, sharing, analysis and application much easier and robust. It is best if we are able to improve the current tools and pipelines in these important areas; big data capture and storage, collaboration, predictive analysis and decision support. With respect to realizing precision medicine, the genomics big data capture and storage tools ensure efficient storage of genomics big data and related medical records to which the data can be shared openly among experts through collaboration tools, analyzed using fully-equipped predictive analysis tool before the clinical decisions can be guided using the decision support tools available in the healthcare institutions.

Precision medicine will just be a pipeline if we are still struggling with the basic requirements that involve small-scale data. Therefore, it is best to improve the currently available tools and pipelines rather than building new ones from scratch. In addition, we should avoid having so many tools that functions similarly as we should be focusing to develop tools that are missing and are crucially needed to manage the genomics big data [77]. For examples, there are yet to be any tools that are able to track the pattern or trends between the genomics big data and corresponding medical records or development of algorithm that can determine the efficacy of drug dosage based on the genomics data [78].

[79] mentioned that the health-related datasets were derived from three important sources which were health information systems (HIS), clinical decision support system (CDSS) and medical body area networks (MBANs). The raw data generated from these systems is a need to improve and optimize all the necessary tools that perform the query processing, data synchronization, real-time data accumulation and determination of automatic cloud storage capacity, despite the availability of these tools at this point of time.

In agreement to the need in improving the tools and having robust pipelines, National Cancer Institute in the United States initiate the NCI Cancer Research Data Commons (NCRDC) to improve the preventions, diagnostics, treatments on cancer diseases through open science efforts [80]. NDRDC is a cloud-based infrastructure consisting of multiple nodes that house processed data, raw data, metadata and analyzed data from cross-domains. It also supports data sharing and improve collaboration among researchers since NDRDC is accessible to all its users. Moreover, NCI also came out with more initiatives including the Genomics Data Commons (GDC), and three Cloud Resources [81]. GDC functioned as both the data repository and a system that applies bioinformatics pipelines to ensure data quality and allow utilization of user-dependent applications.

Other than that, the importance of precision medicine is distinctly shown by many vigorous efforts conducted including the recently initiated Bio-Nepresso Project [82]. This project was based on the success of the monoclonal antibody treatment strategy to treat cancer. However, the current strategy was costly and standardized to a point that variable responses resulted from the treatment. This condition gave rise to the Bio-Nepresso concept which focused on

producing a small-scale, lower cost and customized antibody. Nonetheless, since this machine eliminate the need for the pharmacist to have prerequisite knowledge on antibody production, the efficiency of Bio-Nepresso in aiding the personalized cancer treatment remains questionable.

3.2. Cloud collaboration platform

[83] stated that delivering the ultimate healthcare treatment to the patients requires the work and support from many parties as one could never have sufficient knowledge that can address the many variables arising from the clinical and genomics data. This justifies the need for the cloud collaboration platform, to ensure the objective of precision medicine can be met. The data obtained in this ever-expanding genomics big data research is increasing exponentially that it is impossible to keep up with the data using traditional approaches.

Implementation of cloud collaboration platform will allow for real-time knowledge sharing among many groups of experts worldwide. This platform will also be an interactive medium to solve the issues of genomics deadlock that block the road towards precision medicine. The traditional approach of knowledge and expertise sharing are often limited by several factors such as time, distance and availability. The progress has been slow with this typical approach, leading to more data losses and obsolete data.

Few of the well-known genomics based cloud collaboration platforms were CAVATICA, Cancer Genomics Cloud Pilot and OpenGeneMed. Each of this platform served the main purpose to engage data sharing and encourage new discoveries. For instance, CAVATICA, a cloud-based analytic platform provided an open access environment, focusing on large pediatric brain tumor big data (inclusive of raw genomic data, whole genome, RNA sequencing data, annotations and specimen data) to allow collaboration among its users [84] This platform also allow access to public datasets such as The Cancer Genome Atlas (TCGA) and other National Cancer Institute (NCI) datasets, provide additional supports such as pipelines, computation storage power and visualizations.

Besides that, Cancer Genomics Cloud Pilot, a project by NCI, aimed to improve cancer genomics data sharing by enabling researchers to incorporate their own datasets and tools in addition to the available datasets and in-house analytic tools. This feature was favored by the researchers as it solved most of the common data sharing issues such as the need to download, store and secure the large-scale datasets locally. In other words, fusion of technological advances in big data analytics will support meaningful collaboration among team science to reveal new knowledge that benefited the cancer research, thus realizing precision medicine [85].

In addition, [86] described OpenGeneMed as an informatics hub with automated flexibility to manage next-generation sequencing datasets to support the precision medicine clinical trials. One of the key feature of this system is that it allowed different research team involved in daily management of a clinical trial (sequencing lab, treatment review team, clinical team, statisticians to name a few) to communicate in an open access cloud environment. This feature was advantageous as it minimized the data transferring error between groups and support clear documentations to ensure data integrity. Another important feature of OpenGeneMed was automation. Incorporated tools and pipelines assisted the generation of summarized reports that present the mutation findings from the sequencing result in a timely manner. This information is then used by the team to assign patients to different arms based on their detected mutations.

Therefore, it could be clearly seen that genomics big data requires big infrastructure to support the steps towards precision medicine. Since the abovementioned cloud collaboration platforms have proven to work in encouraging the data sharing and expertise transfer through cloud computing within the genomics field, the proposed framework suggesting expansion of cloud collaboration platform is deemed reasonable and achievable. The aim is similar, only to cover a broader scale of data types inclusive of not only the genomics big data, but healthcare big data as well.

3.3. Automated precision medicine system

On the road to 4IR epoch, the healthcare facilities will not be effective to tackle precision medicine without any reliable system. The precision medicine system needs to incorporate many informatics technologies [87] such as artificial intelligence, augmented reality, machine learning, simulations and visualization to expedite the individualized treatment decisions based on the genomics big data. Moreover, flexibility of the systems is required to ensure continuity of the real-time data accumulation. Another criterion to be fulfilled by this system is automation. This helps to minimize the human error in determining unbiased and accurate individualized treatment to patients. Nonetheless, precision medicine system should be heavily protected with cyber security approaches to prevent typical digital abuse such as data corruption, data loss and data hack [88]. A successful implementation of automated precision medicine system will be able to improve diagnosis, classification and treatment at a larger scale beyond the capability of the existing practices.

[89] highlighted two key areas that will enhanced the precision medicine system which are to build a precision medicine knowledge base (KB) and enhance electronic health records (EHR). A decent precision medicine KB need to fulfill these criteria; contains fruitful information on the diseases (subtypes, prevalence, diagnostics, prognosis and treatment) gained from the data analytics, flexible in terms of supporting different data types that came from genomics research and clinical datasets, scalable as the data will continue to grow and may reach beyond petabytes, extensible to allow modification and addition of modules to the current KB system and readability that encompasses human and machine. The reason behind these criteria is that the currently available KBs were mostly focused on a field such as genomics, remained inaccessible, not connected to each other and fail to execute merged querying. Therefore, new architectural design of precision medicine KBs must be developed to enable clinical decision-making support based on the most recent genomic discoveries and clinical evidences. In addition, electronic health records will be one of the main data sources incorporated into the automated precision medicine system to assist the clinical decision. This is because EHR has been isolated from the genomics data so far. Improvement to the currently available EHR include better structure, improved collection and data display to allow users to obtain meaningful patient information.

To address the issue with regard to the isolation of EHR from the genomics data, an ontology-based system named ONTOFUSION was developed to facilitate the biomedical database integration. It carried out two important processes of mapping and unification. The versatility of ONTOFUSION that differs from p-medicine would be the methodologies used for ontological unification which were top down (existing ontology), bottom-up (build new domain ontology) and hybrid combination [90]. Nevertheless, despite the promising use of ONTOFUSION, the system was only semi-automated since there is a need for human input for to link the database schema with the virtual schema.

Other than that, the importance of integrating the genomics data with EHRs was also supported by IGNITE (Implementing GeNomics In pracTicE). These projects strategized multiple data extraction using data warehousing to integrate the data into a central repository. However, at the moment, IGNITE experience several drawbacks similar to ONTOFUSION, in which certain extent of the data extraction was subjected to manual curation [91].

3.4. Hybrid depositories

Until now, the genomics data still remained isolated from the medical records, personalized genetics screening data and many more [92]. In the long run, this will be a detrimental factor that discourages the progress towards precision medicine. Hybrid de-

positories can simply be defined as data storage places to ensure safekeeping of multiple forms [93] of health related data, comprising of structured data (e.g. genomics data), semi-structure and unstructured data (e.g. electronic health records, lifestyle tracks, environmental exposure health records and medical imaging) [94]. Therefore, the hybrid depositories is visualized with the capability to manage different data types ranging from the genomics data from sequencing, personalized genomics data from genetic screening, laboratory test result information, health medical records, and family health history data. These depositories will expedite the tracking of the required information, from individual to community to whole population.

The reason behind the nonexistent integration between the genomics data and the health records is due to the complexity of the genomics data [95]. Thus, we are continuously dealing with gaps between these related data such as limited data scope, underutilized data and missing information. Without reliable data comparison and integration, the information to understand the genetic mechanism or molecular interaction of the disease remained latent and unexplored. Genomics data mainly contains data from personal genetics, DNA sequencing from service providers and public genome data while healthcare data comprises of medical records, health screening test and data from hospitals' laboratory information system (LIMS).

In addition, a simple model to portray the benefit of data integration as intended by these hybrid depositories is proven through Framingham Heart Study. This Heart Study intended to determine the risk factors for coronary artery diseases [96]. The predictive model [97] combined variable data such as demographics (age, gender), lifestyle factor (tobacco use), clinical data (diabetes, hypertension, body mass index, Low Density Lipoprotein/High Density Lipoprotein - LDL/HDL ratio, cholesterol) and family history data. In other words, this demonstrated that to achieve maximum power to understand disease phenotypes, all relevant data, genomic or otherwise, should be considered. Other similar data integrative benefit is also mentioned by [98]. An approach under Initiative in Precision Medicine was hoped to progress the individualized concept of medical treatment by harnessing various types of data such as the clinical data, genomic evaluations, environmental factors, lifestyle activities to name a few to depict a wide perspective of the patient's health state and its future path.

Hybrid depositories come with a potential to elevate the outcome of getting reliable and accurate disease risk predictors as all data will be taken into consideration. This is not the case with the traditional clinical approach. A breast cancer study by [99] that utilizes 70-genes predictor as the prognostic markers in breast cancer was found to be not unique, probably due the similarity in the dataset samples. Therefore, integrating multiple data from greater scale of data sets will diminish this problem [33] and other consequences such as having misleading strategy for disease treatment.

In other words, it is impractical to utilize genomics big data without the support of big data technology. In the following section, the proposed genomics big data hybrid depositories system will be discussed based on the work by [100]–[102].

4. Genomics big data hybrid depositories system architecture

Since many of the challenges mentioned earlier (genomics big data storage, data complexity, management and security) are majorly reflecting on the lack of depositories that can house different types of related data, this is where the hybrid depositories are of importance. The architectural design of hybrid depositories must address these issues such as linkage of data from multiple sources, elimination of data duplication, improved data compression through customized algorithm, versatility to deposit many types and data formats as well as automated integration of genomics and EHR data.

Hybrid depositories' basic function is to provide an integrative view that considers all forms of data, emphasizing on making use of full spectrum of clinical and demographic data in conjunction to the genomics data. This integration is hoped to provide a sense of data completeness that may help to unveil the relation between the disease genomics and phenotypes and identify not only the link but the potentially conflicting disease risk predictors that defined the disease. The importance of integrating data from healthcare and genomics was also greatly highlighted by the Electronic Medical Records and Genomics (eMERGE) network [103]. This network has successfully demonstrated the usability of linking genomics and medical records data through the identifying the genomics association on age-related cataract diseases from the linked biobanks-EHRs data [104].

Fig. 2 depicted the general architecture of the potential genomics big data hybrid depositories system. The system consists of cross-layered integration of different components. The data model is based on the ontological-relational approach; in which it offers flexibility and logical inference for accurate and high quality data steering while integrating the use of relational databases and stores of triples. However, this paper will not elaborate more on the system development of the proposed system (via the depicted architectural design) since that will come in the next stage of the study.

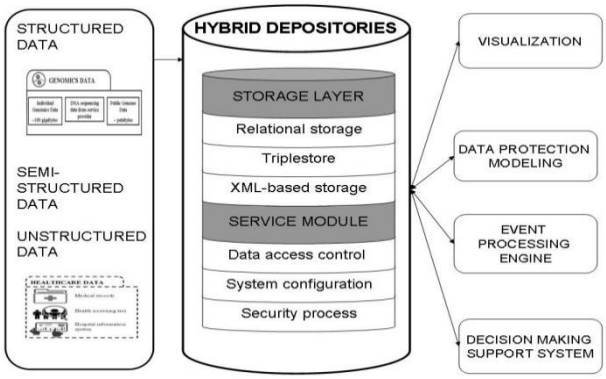

Fig. 2: General Architecture of Genomics Big Data Hybrid Depositories System.

The genomics big data hybrid depositories incorporate several components; which include storage layer and service module. Heterogeneous forms of data; structured, semi-structured and non-structured, will be stored in the hybrid depositories's storage layer. As there is variety of data structures involved, this depositories combines the facility of relational storage, triplestore and XML–based storage. Since many of the genomics data came from public databases, sequencing projects databases and personalized genomics service providers' databases, this suit the purpose of the relational storage. Relational storage works best with the structured data through a relational database management system (RDBMS). In addition, triplestore will function as a database built for the storage and retrieval of triples through semantic queries. Another storage layer of the hybrid depositories is XML-based storage. It allows data to be stored in XML format, queried, transformed, exported and returned to a calling system. This type of storage also minimizes the need for extraction or entry of metadata to support searching and navigation.

Moreover, the second part of the hybrid depositories will be the services module. Hybrid depositories service layer permit communication between the components of the system and extraction of relevant data through three modules which are data access control, system configuration, security process. The data access control module interpret enquiries for data retrieval and grant access right according to the defined permissions in the depositories while the system configuration module manages information for systems, networks, applications and services. Next, the security processes module aims to detect, log and resolve system problems. The hybrid depositories will also be of support for other interrelated system such as visualization, data protection modeling, event processing engine and decision support. Visualization system depicts the relevant data to aid the understanding of data in the context of disease genomics. Meanwhile, data protection mod-

eling system focuses on preventing any vicious threat to ensure safekeeping of data in the hybrid depositories. Event processing engine executes end-to-end processes from filtering, capturing, enrichment, formatting, and importing data from multiple, heterogeneous sources. Finally, data clustering through application of machine learning approach will guide the decision support system.
The need to have a reliable hybrid depositories was strengthen when [105] mentioned that the drawback towards accurate diagnosis in healthcare was due to the lack of available system that can provide decision making through big data analytics approach. However, his proposed framework was more focused towards improving the current diagnosis approach and was based hugely on the health data rather than integrating the genomics data with the health data to define individualized treatment for precision medicine. At the moment, there were no examples of the implemented and used hybrid depositories solely on genomics-healthcare big data with respect to personalized medicine. Therefore, for the purpose of this study, we adapt the concept for the architecture design of the genomics big data hybrid depositories from the work done by [100].
The issues that led to the designing of the hybrid ontological-relational data repository to enhance computer network security [106] were almost similar to the challenges that led to the need for genomics big data hybrid depositories. As such, the availability of security data in variable types (including vulnerabilities databases, attacks databases, platforms databases, weaknesses databases) and format as well as no integration between the different security data. Their study also highlighted that in order to determine the relationship between various, related data, we need to consider as many data sources as possible. Other than that, since [100] work addressed security issue, this is also in par with the concern on the designing of the genomics hybrid depositories.

5. Conclusion

Precision medicine can be potentially materialized with the understanding of knowledge derived from the integrated genomics data and other related health data such as the clinical data, personalized genetic screening, laboratory DNA test, gene expression analysis and health reports. Without this, we will be left clueless especially in transforming the future of healthcare treatment from the current approach of 'one size fits all' to the personalized medicine plan.
From this study, we had identified the major gaps that deter the progress towards precision medicine. Despite an exponential increase in the genomics data, we were still facing challenges such as inefficient genomics big data storage, the difficult integration between the genomics data with electronic health records and more as well as the lack of versatile depositories that can house the complex genomics-healthcare data. As the data remained isolated from each other, we were unable to make full use of data that may hold the key to unveil the knowledge towards understanding disease better or the link between genetic-dependent individual response on drug or treatment. Although this paper discussed part of findings from larger studies which aims to build the actual big data hybrid depository, this paper is the first of its series (the initial work of the larger research).
Based on the conceptual framework to bridge the gap towards precision medicine in the genomics context, we put emphasis on the hybrid depositories. This is due to the potential ability of the hybrid depositories to address the major problems listed earlier. Hybrid depositories will be of importance as it can capture, link, organize and perform analysis from different types of data that will feed the automated precision medicine system. The design of this hybrid depositories will take into account on the ability to correlate events in cross-domain manner with high scalability. The function is supported by two depository's layers; which are storage layer and service module, to carry out these basic requirements of data storage, metadata storage, different level of data management, simultaneous data access, data integrity support, data privacy and support of multi-version management.
To begin, development of the hybrid depositories could be done in a smaller scale, using mock data before we attempt on large-scale data in the hybrid depositories. From here, we will be able to identify the interactions, complementarities and conflicts that occurred among the genes expressions, genetics, clinical markers and other risk factors. This outcome can again be validated using the available medical records. Other than that, the hybrid depositories need to be able to assist the data migration and allow efficient data accessibility. Once this is proven to work, we can apply this platform using real genomics big datasets and health informatics data. In addition, the materialized hybrid depositories could be extended in the future into the adaptation of hybrid depositories into a mobile-size application for more flexibility and mobility.

References

[1] J. Jameson and D. Longo, "Precision medicine—personalized, problematic, and promising," *Obstet. Gynecol. Surv.*, vol. 70, no. 10, pp. 612–614, 2015. https://doi.org/10.1097/01.ogx.0000472121.21647.38.
[2] E. A. Ashley, "The precision medicine initiative: a new national effort," *Jama*, vol. 313, no. 21, pp. 2119–2120, 2015. https://doi.org/10.1001/jama.2015.3595.
[3] I. Ezkurdia *et al.*, "Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes," *Hum. Mol. Genet.*, vol. 23, no. 22, pp. 5866–5878, 2014. https://doi.org/10.1093/hmg/ddu309.
[4] M. Grossglauser and H. Saner, "Data-driven healthcare: from patterns to actions," *Eur. J. Prev. Cardiol.*, vol. 21, no. 2_suppl, pp. 14–17, Nov. 2014.
[5] G. Mendel, "Mendel's Journey from Peas to Petabytes," *Biol. Imagin. Innov. Biosci.*, p. 121, 2014.
[6] A. O'Driscoll, J. Daugelaite, and R. Sleator, "'Big data', Hadoop and cloud computing in genomics," *J. Biomed. Inform.*, vol. 46, no. 5, pp. 774–781, 2013. https://doi.org/10.1016/j.jbi.2013.07.001.
[7] T. A. Peterson, E. Doughty, and M. G. Kann, "Towards precision medicine: advances in computational approaches for the analysis of human variants," *J. Mol. Biol.*, vol. 425, no. 21, pp. 4047–4063, 2013. https://doi.org/10.1016/j.jmb.2013.08.008.
[8] S. Zhao *et al.*, "Rainbow: a tool for large-scale whole-genome sequencing data analysis using cloud computing," *BMC Genomics*, vol. 14, no. 1, p. 425, 2013. https://doi.org/10.1186/1471-2164-14-425.
[9] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mob. Networks Appl.*, vol. 19, no. 2, pp. 171–209, 2014. https://doi.org/10.1007/s11036-013-0489-0.
[10] Z. D. Stephens *et al.*, "Big Data: Astronomical or Genomical?," *PLoS Biol.*, vol. 13, no. 7, p. e1002195, Jul. 2015. https://doi.org/10.1371/journal.pbio.1002195.
[11] M. Viceconti, P. Hunter, and R. Hose, "Big data, big knowledge: big data for personalized healthcare," *IEEE J. Biomed. Heal. Informatics*, vol. 19, no. 4, pp. 1209–1215, 2015. https://doi.org/10.1109/JBHI.2015.2406883.
[12] J. Andreu-Perez, C. C. Y. Poon, R. D. Merrifield, S. T. C. Wong, and G.-Z. Yang, "Big data for health," *IEEE J. Biomed. Heal. informatics*, vol. 19, no. 4, pp. 1193–1208, 2015.
[13] F. S. Collins and V. A. McKusick, "Implications of the Human Genome Project for medical science," *Jama*, vol. 285, no. 5, pp. 540–544, 2001. https://doi.org/10.1001/jama.285.5.540.
[14] K. Offit, "Personalized medicine: new genomics, old lessons," *Hum. Genet.*, vol. 130, no. 1, pp. 3–14, 2011. https://doi.org/10.1007/s00439-011-1028-3.
[15] P. Muir, S. Li, S. Lou, and D. Wang, "The real cost of sequencing: scaling computation to keep pace with data generation," *Genome*, vol. 17, no. 1, p. 53, 2016.
[16] M. H.-Y. Fritz, R. Leinonen, G. Cochrane, and E. Birney, "Efficient storage of high throughput DNA sequencing data using reference-based compression," *Genome Res.*, vol. 21, no. 5, pp. 734–740, 2011. https://doi.org/10.1101/gr.114819.110.
[17] N. Khan *et al.*, "Big data: survey, technologies, opportunities, and challenges," *Sci. World J.*, vol. 2014, 2014.
[18] N. S. Mauthner and O. Parry, "Open Access Digital Data Sharing: Principles, Policies and Practices☆," *Soc. Epistemol.*, vol. 27, no. 1, pp. 47–67, 2013. https://doi.org/10.1080/02691728.2012.760663.
[19] J. L. Jennings and T. J. Hudson, "Abstract 130: International Cancer Genome Consortium (ICGC)," *Cancer Res.*, vol. 76, p. 130, 2016. https://doi.org/10.1158/1538-7445.AM2016-130.

[20] V. Marx, "Biology: The big challenges of big data," *Nature*, p. 255, 2013. https://doi.org/10.1038/498255a.

[21] E. S. Dove, Y. Joly, and A. Tassé, "Genomic cloud computing: legal and ethical points to consider," *Eur. J. Hum. Genet.*, vol. 23, no. 10, pp. 1271–1278, 2015. https://doi.org/10.1038/ejhg.2014.196.

[22] S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, "Big data: Issues and challenges moving forward," in *System Sciences (HICSS), 2013 46th Hawaii International Conference on*, 2013, pp. 995–1004.

[23] N. Levin, R. M. Salek, and C. Steinbeck, "From Databases to Big Data," *Metab. Phenotyping Pers. Public Healthc.*, p. 317, 2016.

[24] S. Choudhury, J. R. Fishman, M. L. McGowan, and E. T. Juengst, "Big data, open science and the brain: lessons learned from genomics," *Front. Hum. Neurosci.*, vol. 8, 2014.

[25] D. Kim, S. Song, and B.-Y. Choi, "Introduction," in *Data Deduplication for Data Optimization for Storage and Network Systems*, Springer, 2017, pp. 3–21. https://doi.org/10.1007/978-3-319-42280-0_1.

[26] D. Kim, S. Song, and B.-Y. Choi, "Existing Deduplication Techniques," in *Data Deduplication for Data Optimization for Storage and Network Systems*, Springer, 2017, pp. 23–76. https://doi.org/10.1007/978-3-319-42280-0_2.

[27] H. H. Do, J. Jansson, K. Sadakane, and W.-K. Sung, "Fast relative Lempel–Ziv self-index for similar sequences," *Theor. Comput. Sci.*, vol. 532, pp. 14–30, 2014. https://doi.org/10.1016/j.tcs.2013.07.024.

[28] S. Deorowicz, A. Danek, and M. Niemiec, "GDC 2: Compression of large collections of genomes.," *Sci. Rep.*, vol. 5, p. 11565, Jun. 2015. https://doi.org/10.1038/srep11565.

[29] W. Christopher and M. Simon, "Review on Genomics APIs," *Comput. Struct. Biotechnol. J.*, 2016.

[30] E. Wang, N. Zaman, S. Mcgee, J.-S. Milanese, A. Masoudi-Nejad, and M. O'Connor-McCourt, "Predictive genomics: a cancer hallmark network framework for predicting tumor clinical phenotypes using genome sequencing data," in *Seminars in cancer biology*, 2015, vol. 30, pp. 4–12. https://doi.org/10.1016/j.semcancer.2014.04.002.

[31] N. Tung, C. Battelli, B. Allen, R. Kaldate, and S. Bhatnagar, "Frequency of mutations in individuals with breast cancer referred for BRCA1 and BRCA2 testing using next-generation sequencing with a 25-gene panel," *Cancer*, vol. 121, no. 1, pp. 25–33, 2015. https://doi.org/10.1002/cncr.29010.

[32] T. Cooke, J. Reeves, A. Lanigan, and P. Stanton, "HER2 as a prognostic and predictive marker for breast cancer," *Ann. Oncol.*, pp. 23–28, 2001. https://doi.org/10.1093/annonc/12.suppl_1.S23.

[33] M. West, G. S. Ginsburg, A. T. Huang, and J. R. Nevins, "Embracing the complexity of genomic data for personalized medicine," *Genome Res.*, vol. 16, no. 5, pp. 559–566, 2006. https://doi.org/10.1101/gr.3851306.

[34] L. Chin, W. C. Hahn, G. Getz, and M. Meyerson, "Making sense of cancer genomic data," *Genes Dev.*, vol. 25, no. 6, pp. 534–555, 2011. https://doi.org/10.1101/gad.2017311.

[35] J. G. Dunn and J. S. Weissman, "Plastid: nucleotide-resolution analysis of next-generation sequencing and genomics data," *BMC Genomics*, vol. 17, no. 1, p. 958, 2016. https://doi.org/10.1186/s12864-016-3278-x.

[36] D. C. Koboldt, K. M. Steinberg, D. E. Larson, R. K. Wilson, and E. R. Mardis, "The next-generation sequencing revolution and its impact on genomics," *Cell*, vol. 155, no. 1, pp. 27–38, 2013. https://doi.org/10.1016/j.cell.2013.09.006.

[37] C. Castaneda *et al.*, "Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine," *J. Clin. Bioinforma.*, vol. 5, no. 1, p. 4, 2015. https://doi.org/10.1186/s13336-015-0019-3.

[38] L. Schriml, C. Arze, S. Nadendla, and Y. Chang, "Disease Ontology: a backbone for disease semantic integration," *academia.edu*, vol. 40, no. D1, pp. 910–946, 2011.

[39] D. Gomez-Cabrero *et al.*, "Data integration in the era of omics: current and future challenges," *BMC Syst. Biol.*, vol. 8, no. 2, p. I1, 2014. https://doi.org/10.1186/1752-0509-8-S2-I1.

[40] G. O. Consortium, "Expansion of the Gene Ontology knowledgebase and resources," *Nucleic Acids Res.*, vol. 45, no. D1, pp. 331–338, 2017. https://doi.org/10.1093/nar/gkw1108.

[41] M. Subhani, A. Anjum, and A. Koop, "Clinical and genomics data integration using meta-dimensional approach," *Proc. 9th*, pp. 416–421, 2016.

[42] B. Louie, P. Mork, F. Martin-Sanchez, and A. Halevy, "Data integration and genomic medicine," *J. Biomed. Inform.*, vol. 40, no. 1, pp. 5–16, 2007. https://doi.org/10.1016/j.jbi.2006.02.007.

[43] P. Appleby, "Linking Genomic Data with Phenotypes Derived from Electronic Health Records.," *Int. J. Popul. Data Sci.*, vol. 1, no. 1, 2017.

[44] M. D. Ritchie, M. De Andrade, and H. Kuivaniemi, "The foundation of precision medicine: integration of electronic health records with genomics through basic, clinical, and translational research," *Front. Genet.*, vol. 6, 2015.

[45] P. Khatri and S. Drăghici, "Ontological analysis of gene expression data: current tools, limitations, and open problems," *Bioinformatics*, vol. 21, no. 18, pp. 3587–3595, 2005. https://doi.org/10.1093/bioinformatics/bti565.

[46] S. Palaniappan and N. Y. Huey, "A tool for healthcare information integration," *J. ICT*, vol. 5, pp. 29–44, 2006.

[47] M. Dugas, A. Meidt, P. Neuhaus, M. Storck, and J. Varghese, "ODMedit: uniform semantic annotation for data integration in medicine based on a public metadata repository.," *BMC Med. Res. Methodol.*, vol. 16, p. 65, 2016. https://doi.org/10.1186/s12874-016-0164-9.

[48] J. Marés *et al.*, "p-medicine: A medical informatics platform for integrated large scale heterogeneous patient data," in *AMIA Annual Symposium Proceedings*, 2014, vol. 2014, p. 872.

[49] F. Schera, G. Weiler, E. Neri, S. Kiefer, and N. Graf, "The p-medicine portal—a collaboration platform for research in personalised medicine," *Ecancermedicalscience*, vol. 8, 2014.

[50] A. Alyass, M. Turcotte, and D. Meyre, "From big data analysis to personalized medicine for all: challenges and opportunities," *BMC Med. Genomics*, vol. 8, no. 1, p. 33, 2015. https://doi.org/10.1186/s12920-015-0108-y.

[51] F. Cheng, J. Zhao, and Z. Zhao, "Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes," *Brief. Bioinform.*, vol. 17, no. 4, pp. 642–656, Jul. 2016. https://doi.org/10.1093/bib/bbv068.

[52] J. Howison and J. Bullard, "Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature," *J. Assoc. Inf. Sci. Technol.*, vol. 67, no. 9, pp. 2137–2155, 2016. https://doi.org/10.1002/asi.23538.

[53] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nat. Rev. Genet.*, vol. 17, no. 6, pp. 333–351, 2016. https://doi.org/10.1038/nrg.2016.49.

[54] M.-A. Madoui *et al.*, "Genome assembly using Nanopore-guided long and error-free DNA reads," *BMC Genomics*, vol. 16, no. 1, p. 327, 2015. https://doi.org/10.1186/s12864-015-1519-z.

[55] T. Madden, "The BLAST sequence analysis tool," 2013.

[56] R. Wilton, T. Budavari, B. Langmead, S. J. Wheelan, S. L. Salzberg, and A. S. Szalay, "Arioc: high-throughput read alignment with GPU-accelerated exploration of the seed-and-extend search space," *PeerJ*, vol. 3, p. e808, 2015. https://doi.org/10.7717/peerj.808.

[57] F. E. Faisal, L. Meng, J. Crawford, and T. Milenković, "The post-genomic era of biological network alignment," *EURASIP J. Bioinforma. Syst. Biol.*, vol. 2015, no. 1, p. 3, 2015. https://doi.org/10.1186/s13637-015-0022-9.

[58] R. Margolis, L. Derr, M. Dunn, and M. Huerta, "The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data," *J. Am. Med. Informatics Assoc.*, vol. 21, no. 6, pp. 957–958, 2014. https://doi.org/10.1136/amiajnl-2014-002974.

[59] T. Barreto, A. Mand, M. Spielberg, D. MacKenzie, and S. Ghods, "Managing updates at clients used by a user to access a cloud-based collaboration service." Google Patents, 21-Apr-2015.

[60] T. Takai-Igarashi *et al.*, "Security controls in an integrated Biobank to protect privacy in data sharing: rationale and study design," *BMC Med. Inform. Decis. Mak.*, vol. 17, no. 1, p. 100, 2017. https://doi.org/10.1186/s12911-017-0494-5.

[61] E. S. Dove, "Biobanks, Data Sharing, and the Drive for a Global Privacy Governance Framework," *J. Law, Med. Ethics*, vol. 43, no. 4, 2015.

[62] F. Carrasco-Ramiro, R. Peiró-Pastor, and B. Aguado, "Human genomics projects and precision medicine," *Gene Ther.*, vol. 24, no. 9, p. 551, 2017. https://doi.org/10.1038/gt.2017.77.

[63] T. Schultz, "Turning healthcare challenges into big data opportunities: A use-case review across the pharmaceutical development lifecycle," *Bull. Assoc. Inf. Sci. Technol.*, vol. 39, no. 5, pp. 34–40, 2013. https://doi.org/10.1002/bult.2013.1720390508.

[64] J. Luo, M. Wu, D. Gopukumar, and Y. Zhao, "Big data application in biomedical research and health care: A literature review," *Biomed. Inform. Insights*, vol. 8, p. 1, 2016. https://doi.org/10.4137/BII.S31559.

[65] A. Alzu'bi, L. Zhou, and V. Watzlaf, "Personal genomic information management and personalized medicine: challenges,

current solutions, and roles of HIM professionals.," *Perspect. Heal. Inf. Manag.*, vol. 11, no. Spring, p. 1c, 2014.
[66] M. Beck, V. Haupt, J. Roy, J. Moennich, and R. Jäkel, *Genecloud: Secure cloud computing for biomedical research*. Springer, Cham., 2014.
[67] M. D. Assunção, R. N. Calheiros, S. Bianchi, M. A. S. Netto, and R. Buyya, "Big Data computing and clouds: Trends and future directions," *J. Parallel Distrib. Comput.*, vol. 79, pp. 3–15, 2015. https://doi.org/10.1016/j.jpdc.2014.08.003.
[68] A. P. Heath *et al.*, "Bionimbus: a cloud for managing, analyzing and sharing large genomics datasets," *J. Am. Med. Informatics Assoc.*, vol. 21, no. 6, pp. 969–975, Nov. 2014. https://doi.org/10.1136/amiajnl-2013-002155.
[69] S. Datta, K. Bettinger, and M. Snyder, "Practical Guidelines for Secure Cloud Computing using Genomic Data," *bioRxiv*, p. 34876, 2015.
[70] Q. Jiang, M. K. Khan, X. Lu, J. Ma, and D. He, "A privacy preserving three-factor authentication protocol for e-Health clouds," *J. Supercomput.*, vol. 72, no. 10, pp. 3826–3849, 2016. https://doi.org/10.1007/s11227-015-1610-x.
[71] A. Park *et al.*, "The Blockchain for Personalized Medicine," 2017.
[72] Z. Shae and J. J. P. Tsai, "On the Design of a Blockchain Platform for Clinical Trial and Precision Medicine," in *Distributed Computing Systems (ICDCS), 2017 IEEE 37th International Conference on*, 2017, pp. 1972–1980.
[73] D. Milius *et al.*, "The International Cancer Genome Consortium's evolving data-protection policies," *Nat. Biotechnol.*, vol. 32, no. 6, pp. 519–523, 2014. https://doi.org/10.1038/nbt.2926.
[74] R. C. Green, D. Lautenbach, and A. L. McGuire, "GINA, genetic discrimination, and genomic medicine," *N. Engl. J. Med.*, vol. 372, no. 5, pp. 397–399, 2015. https://doi.org/10.1056/NEJMp1404776.
[75] C. Auffray *et al.*, "Making sense of big data in health research: towards an EU action plan," *Genome Med.*, vol. 8, no. 1, p. 71, 2016. https://doi.org/10.1186/s13073-016-0323-y.
[76] U. H. Mohamad, M. T. Ijab, and R. A. Kadir, "Bridging the Gap in Personalised Medicine Through Data Driven Genomics," in *International Visual Informatics Conference*, 2017, pp. 88–99. https://doi.org/10.1007/978-3-319-70010-6_9.
[77] A. Shachak, K. Shuval, and S. Fine, "Barriers and enablers to the acceptance of bioinformatics tools: a qualitative study," *J. Med. Libr. Assoc. JMLA*, vol. 95, no. 4, p. 454, 2007. https://doi.org/10.3163/1536-5050.95.4.454.
[78] L. Samuel, "Drug dosing goes digital with new algorithm," *Stat*, 2016. [Online]. Available: https://www.statnews.com/2016/04/06/tailoring-dosages-patients/. [Accessed: 19-Jan-2018].
[79] L. Wang, R. Ranjan, J. Kolodziej, A. Y. Zomaya, and L. Alem, "Software Tools and Techniques for Big Data Computing in Healthcare Clouds.," *Futur. Gener. Comp. Syst.*, vol. 43, pp. 38–39, 2015. https://doi.org/10.1016/j.future.2014.11.001.
[80] S. Wilson *et al.*, "Developing Cancer Informatics Applications and Tools Using the NCI Genomic Data Commons API," *Cancer Res.*, vol. 77, no. 21, pp. e15–e18, 2017. https://doi.org/10.1158/0008-5472.CAN-17-0598.
[81] I. V Hinkson, T. M. Davidsen, J. D. Klemm, I. Chandramouliswaran, A. R. Kerlavage, and W. A. Kibbe, "A Comprehensive Infrastructure for Big Data in Cancer Research: Accelerating Cancer Research and Precision Medicine," *Frontiers in Cell and Developmental Biology*, vol. 5. p. 83, 2017. https://doi.org/10.3389/fcell.2017.00083.
[82] A. Bisnajak, "The Bio-Nespresso Project: The design of a small-scale manufacturing unit for personalized medicine production," 2018.
[83] A. B. of Directors, "Laboratory and clinical genomic data sharing is crucial to improving genetic health care: a position statement of the American College of Medical Genetics," *Genet. Med.*, 2017.
[84] A. C. Resnick *et al.*, "Abstract LB-008: The Pediatric Brain Tumor Atlas: building an integrated, multi-platform data-rich ecosystem for collaborative discovery in the cloud." AACR, 2017.
[85] E. R. Hsu, J. D. Klemm, A. R. Kerlavage, D. Kusnezov, and W. A. Kibbe, "Cancer Moonshot Data and Technology Team: Enabling a National Learning Healthcare System for Cancer to Unleash the Power of Data," *Clin. Pharmacol. Ther.*, vol. 101, no. 5, pp. 613–615, 2017. https://doi.org/10.1002/cpt.636.
[86] A. Palmisano, Y. Zhao, M.-C. Li, E. C. Polley, and R. M. Simon, "OpenGeneMed: a portable, flexible and customizable informatics hub for the coordination of next-generation sequencing studies in support of precision medicine trials," *Brief. Bioinform.*, vol. 18, no. 5, pp. 723–734, 2016. https://doi.org/10.1093/bib/bbw059.
[87] D. R. Leff and G.-Z. Yang, "Big data for precision medicine," *Engineering*, vol. 1, no. 3, pp. 277–279, 2015. https://doi.org/10.15302/J-ENG-2015075.
[88] K. Lauter, A. López-Alt, and M. Naehrig, "Private Computation on Encrypted Genomic Data.," in *International Conference on Cryptology and Information Security in Latin America*, 2014, pp. 3–27.
[89] J. D. Tenenbaum *et al.*, "An informatics research agenda to support precision medicine: seven key areas," *J. Am. Med. Informatics Assoc.*, vol. 23, no. 4, pp. 791–795, 2016. https://doi.org/10.1093/jamia/ocv213.
[90] D. Pérez-Rey *et al.*, "ONTOFUSION: Ontology-based integration of genomic and clinical databases," *Comput. Biol. Med.*, vol. 36, no. 7–8, pp. 712–730, 2006. https://doi.org/10.1016/j.compbiomed.2005.02.004.
[91] N. R. Sperber *et al.*, "Challenges and strategies for implementing genomic services in diverse settings: experiences from the Implementing GeNomics In pracTicE (IGNITE) network.," *BMC Med. Genomics*, vol. 10, no. 1, p. 35, May 2017. https://doi.org/10.1186/s12920-017-0273-2.
[92] B. M. Welch, K. Eilbeck, G. Del Fiol, L. J. Meyer, and K. Kawamoto, "Technical desiderata for the integration of genomic data with clinical decision support," *J. biomed. info*, vol. 51, pp. 3–7, 2014.
[93] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 97–107, 2014. https://doi.org/10.1109/TKDE.2013.109.
[94] H. Chang and M. Choi, "Big data and healthcare: building an augmented world," *Healthc. Inform. Res.*, vol. 22, no. 3, pp. 153–155, 2016. https://doi.org/10.4258/hir.2016.22.3.153.
[95] N. V Chawla and D. A. Davis, "Bringing big data to personalized healthcare: a patient-centered framework," *J. Gen. Intern. Med.*, vol. 28, no. 3, pp. 660–665, 2013. https://doi.org/10.1007/s11606-013-2455-8.
[96] C. W. Tsao and R. S. Vasan, "Cohort Profile: The Framingham Heart Study (FHS): overview of milestones in cardiovascular epidemiology," *Int. J. Epidemiol.*, vol. 44, no. 6, pp. 1800–1813, 2015. https://doi.org/10.1093/ije/dyv337.
[97] T. Gordon, W. P. Castelli, M. C. Hjortland, W. B. Kannel, and T. R. Dawber, "High density lipoprotein as a protective factor against coronary heart disease: the Framingham Study," *Am. J. Med.*, vol. 62, no. 5, pp. 707–714, 1977. https://doi.org/10.1016/0002-9343(77)90874-9.
[98] I. S. Kohane, "Ten things we have to do to achieve precision medicine," *Science (80-.).*, vol. 349, no. 6243, pp. 37–38, 2015.
[99] M. J. Van De Vijver *et al.*, "No TitleA gene-expression signature as a predictor of survival in breast cance," *N. Engl. J. Med.*, vol. 347, no. 25, pp. 1999–2009, 2002. https://doi.org/10.1056/NEJMoa021967.
[100] I. Kotenko, O. Polubelova, A. Chechulin, and I. Saenko, "Design and implementation of a hybrid ontological-relational data repository for siem systems," *Futur. internet*, vol. 5, no. 3, pp. 355–375, 2013.
[101] H. Garcia-Molina, *Database systems: the complete book*. Pearson Education India, 2008.
[102] D. Marco, "Building and managing the meta data repository," *A full lifecycle Guid.*, 2000.
[103] J. W. Smoller *et al.*, "An eMERGE clinical center at partners personalized medicine," *J. Pers. Med.*, vol. 6, no. 1, p. 5, 2016. https://doi.org/10.3390/jpm6010005.
[104] M. D. Ritchie *et al.*, "Electronic medical records and genomics (eMERGE) network exploration in cataract: several new potential susceptibility loci," *Mol. Vis.*, vol. 20, p. 1281, 2014.
[105] M. I. Babar, M. Jehanzeb, M. Ghazali, D. N. A. Jawawi, F. Sher, and S. A. K. Ghayyur, "Big data survey in healthcare and a proposal for intelligent data diagnosis framework," in *2nd IEEE International Conference on Computer and Communications (ICCC)*, 2016, pp. 7–12.
[106] A. V Fedorchenko, I. V Kotenko, E. V Doynikova, and A. A. Chechulin, "The ontological approach application for construction of the hybrid security repository," in *Soft Computing and Measurements (SCM), 2017 XX IEEE International Conference on*, 2017, pp. 525–528.