# On distributed collaboration for biomedical analyses

Fatima-Zahra Boujdad, Alban Gaignard, Mario Südholt, Wilmer Garzón-Alfonso, Luis Daniel Benavides Navarro, Richard Redon

# On distributed collaboration for biomedical analyses

Fatima-zahra Boujdad*, Alban Gaignard†, Mario Südholt*, Wilmer Garzón‡, Luis Benavides‡, Richard Redon†

*STACK research team
IMT Atlantique, Inria, LS2N
Nantes, France
First.Last@imt-atlantique.fr

†Institut du Thorax
INSERM, CNRS, Nantes University
Nantes, France
First.Last@univ-nantes.fr

‡Dept. of Computer Engineering
Colombian School of Engineering
Bogotá, Colombia
First.Last@escuelaing.edu.co

***Abstract*—Cooperation of research groups is nowadays common for the development and execution of biomedical analyses. Multiple partners contribute data in this context, data that is often centralized for processing at some cluster-based or supercomputer-based infrastructure. In contrast, real distributed collaboration that involves processing of data from several partners at different sites is rare. However, such distributed analyses are often very interesting, in particular, for scalability, security and privacy reasons.**

**In this article, we motivate the need for real distributed biomedical analyses in the context of several ongoing projects, including the I-CAN project that involves 34 French hospitals and affiliated research groups. We present a set of distributed architectures for such analyses that we have derived from discussions with different medical research groups and a study of related work. These architectures allow for scalability, security/privacy and reproducibility issues to be taken into account. Finally, we illustrate that these architectures can serve as the basis of a development method for biomedical distributed analyses.**

***Index Terms*—bioinformatics, distributed systems, reproducibility, scalability, security, privacy, workflow**

## I. Introduction

The field of biomedical analyses, in particular genetic ones, abounds with scenarios involving the cooperation of multiple, often numerous, different stakeholders. The analyses are typically characterized by the application of new analysis algorithms to (very) large volumes of data. Distributed collaborations and the distributed execution of analyses should be a method of choice in this context, for instance, because some data may have to be remotely processed but other data — such as identifying clinical, imaging or genetic data — may have to be processed locally, for instance, due to legal or security reasons.

Genome Wide Association Studies (GWAS) are statistical analyses that require interrogating hundreds of thousands of genetic markers among large groups of human subjects, in order to identify genetic variants (or haplotypes) associated to particular phenotypes or diseases. Such large-scale approaches, which commonly requires international collaborations involving multiple research groups, have contributed to the elucidation of the genetic architecture of common diseases such as type-2 diabetes or Alzheimer's disease. [1]–[4]. These analyses usually consist in statistically comparing the minor allele frequencies of genetic variants across the whole genome, between two groups of affected versus unaffected individuals (case-control studies). The statistical significance of the discovered associations is often limited by the number of cases or control individuals included in the GWAS, especially in the context of less common diseases. It then becomes critical to join parallel efforts internationally and combine multiple case-control datasets into large-scale meta-analyses. However, joining genotype datasets in such meta-analyses raises several legal and ethical issues since individual genetic profiles are highly identifying : sharing case-control genotype datasets must be carefully controlled to conform with national data protection laws (implementations of the EU General Data Protection Regulation). In addition, once transferred, controlling how data is used by collaborating partner is particularly challenging : (1) there is no easy means to assess that genotype data are only used in the context of the original collaborative study, and (2) effective data removal is declarative and only based on mutual trust between partners. Modern distributed computing infrastructures should thus provide means to limit data transfer/centralization and enable data providers to keep full control on their own data, while allowing for large scale meta-analyses on individual genotype profiles.

Most medical research projects mainly harness distribution currently as a means to speed up computations through the use of cluster-like infrastructures. Computations are therefore mostly performed at a single site, so they do not involve or involve only little distributed collaboration. For example, often MapReduce is used to improve the performance of biomedical analyses, without distributed collaboration among different sites [5]–[8].

This is also the case for the current configuration of the I-CAN project[1], in which 34 French hospitals contribute their data to centralized servers for storage first. Data processing, which includes clinical, imaging and genetic data, is then performed on a different server but also in a centralized fashion, which implies large-scale data transmission at the moment of computation that requires a high-speed network to be available.

However, distributed collaborations become mandatory because of the following requirements:

- Flexible placement of computations and data in the context of locality constraints. Some data may have to be kept at a site where only a part of an analysis is performed.
- Security and privacy constraints on data usage and analysis execution may require that data from different sites cannot be pooled at one site.

[1]http://www.agence-nationale-recherche.fr/Project-ANR-15-CE17-0008

- Scalability for performance reasons may depend on parts of the analysis be performed at different sites.

These three requirements constitute, at the same time, difficult challenges in a distributed environment. In addition to these requirements, reproducibility is another important challenge in the context of distributed medical analyses. The I-CAN members, for example, envision three evolutions involving distributed collaboration because of these requirements.

Moreover, the structure of distributed collaborations for biomedical analyses strongly depends on architectural elements, such as the use of private, community or public clouds with high-speed network access, the availability of trusted parties and servers as well as whether input data must be hidden or not from the other stakeholders in the collaboration. As to our knowledge, no systematic analysis of collaboration architectures has been presented until now.

Based on our cooperation in multiple bioinformatics projects as well as a study of related work, we have performed a first systematic analysis of architectures for distributed biomedical analyses. The main goal of this paper is to fuel discussion on distributed biomedical collaborations based on three contributions:

- We *motivate the use of distributed biomedical cooperation* and discuss corresponding issues in the context of a real-world multi-center biomedical study, I-CAN that involves 34 French hospitals.
- We introduce a *set of architectures* for distributed biomedical analyses that are motivated by corresponding scenarios from real-world biomedical projects.
- We present a preliminary version of a method for the design and implementation of distributed biomedical analyses that meets the requirements introduced above. We apply the resulting approach to the I-CAN project and show how it could benefit from the proposed distributed cloud-based architectures.

This paper is structured as follows. We motivate distributed collaborations for biomedical analyses based on the I-CAN project in Sec. II. Sec. III presents distributed collaboration architectures. Our approach to the definition of distributed biomedical analyses and its application to the I-CAN project is presented in Sec. IV. We then discuss related work in Sec. V and conclude.

## II. The I-CAN biomedical study

We now present a real-world biomedical study, the I-CAN project, which involves collaboration among 34 French hospitals and research groups. I-CAN is a multi-center study aimed at bridging clinical observations, genomic markers as well as quantitative imaging biomarkers to better understand the formation, the development and the possible rupture of intracranial aneurysms. Fig. 1 shows the currently deployed architecture that leverages multiple infrastructures.

In this project, each participating university hospital is populating data repositories provided by domain-specific infrastructures. Time-Of-Flight MRIs and arteriographies are centralized

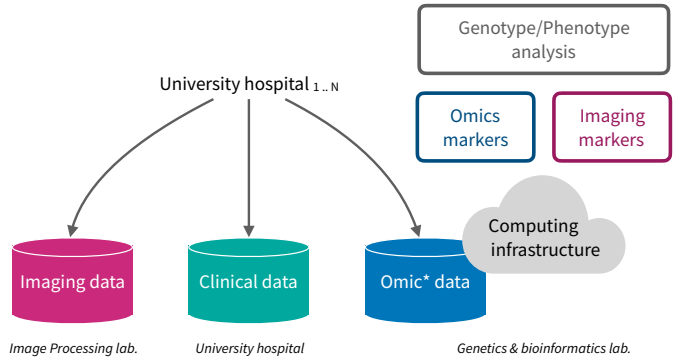



Fig. 1. Leveraging heterogeneous data repositories when studying intracranial aneurysms development and rupture cases

in a medical imaging repository, clinical observations are centrally hosted by the university hospital coordinating the clinical study through its clinical research database, and omics data is centrally hosted elsewhere by the genetics research lab responsible for the sequencing of biological samples and the required bioinformatics analyses.

In this architecture, compute-intensive analyses are performed at the genetics and bioinformatics lab, which operates a cluster originally dedicated to bioinformatics applications. This approach is pragmatic since no computing infrastructure is available close to the partners that principally generate medical imaging data. However it presents several drawbacks. First, medical imaging data must be transferred between two sites, which often represent a waste of time, energy and raises security risks. Moreover, the bioinformatics cluster must be tooled with the required image processing software components so that the MRI quantification workflow can be run and parallelized. However, the image processing software is rarely of use at that cluster.

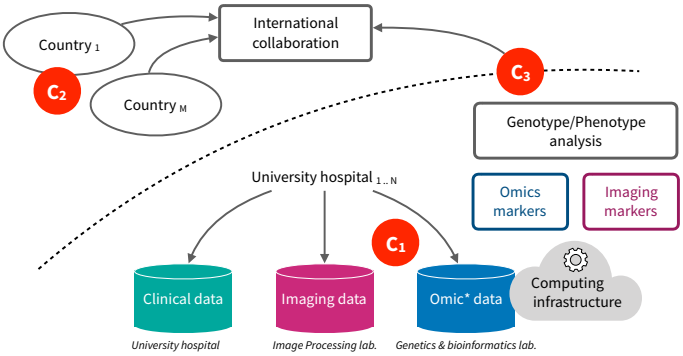



Fig. 2. Distributed computing strategies can provide means to tackle security, scalability as well as reproducibility challenges in large-scale biomedical studies.

The I-CAN partners currently discuss different extension scenarios. For instance, we show in Fig. 2 that genetic data sequenced at the Genetics & bioinformatics lab. can be contributed to an international cohort in the context of an international collaborative analysis. In this paper, we pro-

pose corresponding distributed architectures that address the distributed analysis challenges mentioned in the introduction. Concretely, three major challenges important to the project are illustrated in Fig. 2:

- $C_1$: **Scalability**. For a single individual, the combination of novel imaging techniques as well as the availability of genome-wide sequencing technologies require an unprecedented amount of computation time. Centralizing computations hardly cope with the size of the studied cohorts: 3000 considered cases in the context of the ICAN project.
- $C_2$: **Reproducibility**. When joining international efforts in the context of genome-wide association studies for instance, data coming from each partner are pooled to increase findings statistical power. This requires to be sure that the very same data analysis workflow is reproduced on each participating dataset.
- $C_3$: **Security & Privacy** The privacy of individuals who contribute their genetic markers for research should be protected not only by hiding their identity, but also — and especially — by not divulging any information about their genomes. In the context of GWAS, even aggregate data from each country cannot be shared because of possible identifying attacks [9]. Therefore, GWA studies require ad-hoc security protocols and architectures to be applied at this level *i.e.,* privacy-preserving GWAS as in [10].

## III. Distributed collaboration for biomedical analyses

Collaboration between multiple clinical and research groups is common and there is a growing need for the coordination of distributed analyses that are executed at different sites. The notion of architecture is central to the design and implementation of such distributed collaborative analyses that have to meet the challenges introduced before. However, currently no systematic study of architectures for distributed biomedical analyses has been performed. In this section we first review common architectural elements that have been used in this context. We then present the main contribution of this paper, a series of architectures that we have extracted from our review of related work and discussions with different research groups and projects from the medical domain.

### A. Current architectural elements for biomedical analyses

Three architectural elements are principally used in current biomedical analyses: cloud infrastructures, trusted parties and hybrid architectures.

*1) Cloud infrastructures:* Cloud infrastructures, notably private clouds hosted, for example, by a research organization, are already commonly used in biomedical research. Public clouds are often suited for analyses involving massive volumes of data but are more rarely used, principally because of security and privacy issues. Public clouds use is widely advocated, however, by computer science research community where such constraints are already resolved, such as GWAS analyses that can operate on encrypted genetic data. In such cases, the security and privacy properties of the corresponding data sets are well known and the analyses can be conceived in a such way that keeps data protected from other participants and other intruders.

*2) Trusted parties:* Trusted parties that protect data and computations all the while providing efficient infrastructures for collaborative analyses is a common deployment model. Data in such architectures has to be protected only during its transfer to and from the processing trusted party using secure communication protocols.

*3) Hybrid infrastructures:* Because biomedical data are heterogeneous and its sensitivity to privacy issues differs from one data set to another, one can consider a hybrid infrastructure that mix public and private infrastructures for analyses in order to process data in the most fitting environment. Non-identifying data like de-identified neuroimages can be analyzed on public clouds while genomes should be analyzed in an isolated secure platform, *e.g.,* locally. Another scenario motivating the use of hybrid architectures is when encrypted data has to be decrypted for processing at a secure infrastructure and/or trusted party.

### B. Architectures for distributed biomedical analyses

We now motivate and present a series of architectures for collaborations involving distributed analyses. They are characterized in terms of the requirements introduced before: placement of computations and data, security/privacy considerations, as well as scalability/performance. Additionally, our analysis of related work and discussions with medical researchers have resulted in a set of major architectural elements (the first of which coincide with those commonly used and discussed in the previous subsection):

- The use of public/community clouds and the involvement of trusted parties.
- The availability of local computation and storage resources.
- Whether data can be moved to a remote site or has to be processed locally (*e.g.,* because of legal/security reasons, the volume of data involved or clinical data that has to be generated locally).
- Whether the researcher performing the analysis is part of the stakeholders owning the data and computation facilities or whether he is external.
- Whether the involved parties can be able to have access to (part of) the data of the other stakeholders or whether data should be hidden from all stakeholders and only results be available.

Fig. 3 gives an overview of real world and state-of-the-art architectures that cover different kinds of existing biomedical analyses as well as some future analyses our partners are planning. We classify architectures according to the locality properties of data and computations and explain their characteristics with respect to the scalability and privacy criteria. We symbolize community clouds and public clouds with red and blue colored clouds respectively. Storage is indicated with a database icon while computations are represented with a

gear. Most of the architectures that include public clouds do not have current scenarios in the biomedical field, namely Fig. 3(e), Fig. 3(f) and Fig. 3(h). These architectures, however, can be very interesting in handling biomedical workflows with resource demanding computations on highly heterogeneous data.

*1) Local storage and computations:* Using this architecture hospitals can provide limited storage and data processing resources that are used for local processing at each site. Therefore, the architecture in Fig. 3(a) requires an intermediate hospital between researchers and data contributors that orchestrates queries to the system and guarantees transparency for researchers. From the researcher's point of view, this scenario supposes that data can be fully processed remotely and that he only can get the final results. Queries can be sent to hospitals via the mediating server that is also part of the study. clearly, this type of architecture does not afford good scalability and correct computational reproducibility is not guaranteed.

*2) Local storage and processing on community clouds:* Fig. 3(b) depicts an architecture where a community cloud is harnessed to speed up computations. This architecture is used for example by numerous projects leveraging the cluster infrastructure of the French Institute of Bioinformatics[2]. This model offers better performance than on-site processing (Fig. 3(a)), but data need to be transferred to processing clusters each time as they are stored locally. In addition, community and private clouds often support scalability only in a limited fashion. No issues related to reproducibility should emerge with this model because executing the same workflow on input data is straightforward given that data is pooled to a single location for processing, that is, no workflow replication is needed.

*3) Local storage and public cloud based processing:* Most proposals of secure GWAS advocate the use of public cloud computing following the architecture of Fig. 3(c) [11]–[14]; data is first encrypted locally by each data owner, then sent to be processed in a public cloud. With the availability of special (aka. homomorphic) encryption algorithms, the public cloud can process the data in its encrypted form and yield encrypted results to the researcher. This way of processing and data transfer is very important when dealing with highly privacy-sensitive data, like genetic one. Finally, the researcher can decrypt the results. In this protocol, encryption of input data should be locally performed on the fly which supposes that hospitals (or other entities) are endowed with sufficient computation resources.

*4) Community cloud storage and processing:* The architectures in Fig. 3(d) and Fig. 3(i) illustrate the case where hospitals are either deprived of sufficient resources and expertise in which case they need to outsource a big amount of data along with computations or willing to grant easy and permanent data access to researchers. For researchers, it is easier to work with already available data and perform computations on the corresponding site. Otherwise, the architecture can hardly be scalable given that an important number of hospitals pool their data and computations into the same community cloud almost whenever new data is gathered. The current I-CAN project architecture runs in the model of Fig. 3(d) regarding genetic data and of Fig. 3(i) for neuroimaging data.

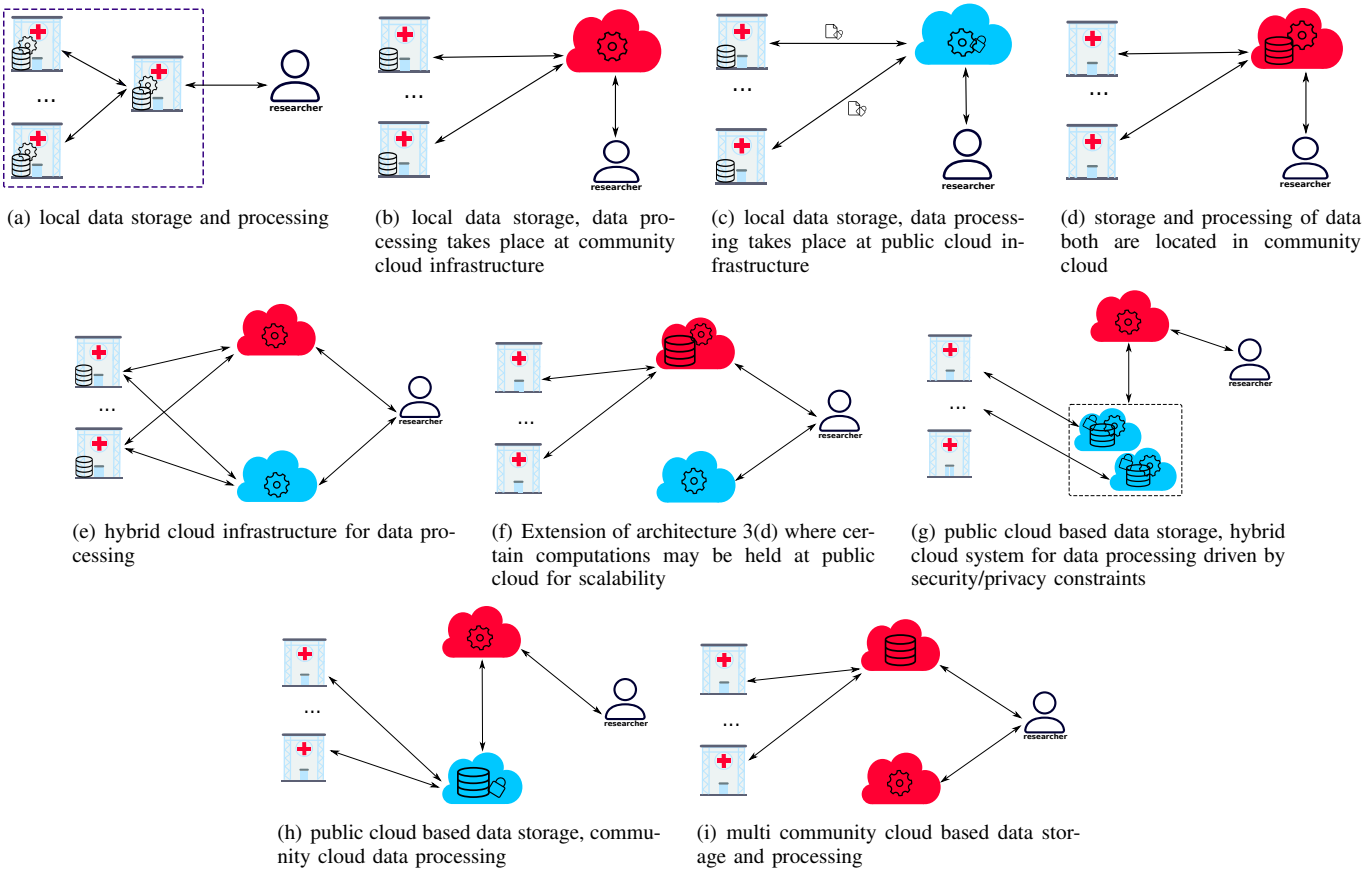
*5) Hybrid cloud system for data processing:* There are cases where data owners, hospitals for instance, would prefer local storage for, among others, security reasons. However, in order to benefit from good performance they would rather consider using the cloud computing for data processing, in which case a hybrid cloud infrastructure as shown in 3(e) can be of use, This is useful to satisfy different policies on different data types: in the I-CAN project, for instance, genetic data is processed on a private/community cloud and non-identifying imaging data can be analyzed on a public cloud.

*6) Hybrid cloud system for data storage and processing:* Fig 3(f) is an integration of architectures $d$ and $e$ into one system where data can be stored at a community cloud while data processing can be held in both private and public clouds. The novelty of this combined system is that it i) grants straightforward data access to researchers, ii) enhances scalability/performance of computations and iii) helps to enforce security/privacy constraints by processing private parts of the data in a relatively more secure server *i.e.,* community cloud.

*7) Public cloud storage, hybrid cloud system for data processing:* The architecture of Fig. 3(g) is that of the solution proposed in [10] intended to secure collaborative analyses on medical and genetic data. In this architecture, medical and genetic data is secured and stored in a public cloud that can additionally handle computations that do not threaten data privacy; otherwise, there is a trusted party *e.g.,* community cloud which handles computations where *e.g.,* data decryption would be needed. Encryption is not the only technique used to preserve security/privacy constraints, data splitting can also be used in order to avoid sensitive associations and intensive encryption, hence splitting permits to enhance the overall system performance. Scalability and security are the criteria that define the architecture, thus, they are assured by construction. However, considering a trusted party for *e.g.,* processing genomes in plaintext is unlikely to allow for an international collaboration, for instance, in the context of GWAS because of legal and political considerations. A similar architecture is depicted in Fig. 3(h), the difference with Fig. 3(g), is that public cloud here is used only for storage, which means that data cannot be analyzed in the protected state, for instance, when they are symmetrically encrypted. This architecture can be interesting for scenarios where storage is very resource demanding while computations on the stored data do not occur very often.

We evaluate architectures of Fig. 3 with respect to i) scalability/performance of storage and computations, ii) security/privacy of medical and genetic data and iii) reproducibility. The evaluation is reported in table I. We choose to evaluate scalability/performance along with security/privacy with three levels: low (-), medium (+) and high (++) which is an attempt to covering all depicted architectures. Reproducibility is evaluated with straightforward ($+$), not straightforward ($-$).

[2]https://www.france-bioinformatique.fr/




(a) local data storage and processing

(b) local data storage, data processing takes place at community cloud infrastructure

(c) local data storage, data processing takes place at public cloud infrastructure

(d) storage and processing of data both are located in community cloud

(e) hybrid cloud infrastructure for data processing

(f) Extension of architecture 3(d) where certain computations may be held at public cloud for scalability

(g) public cloud based data storage, hybrid cloud system for data processing driven by security/privacy constraints

(h) public cloud based data storage, community cloud data processing

(i) multi community cloud based data storage and processing

Fig. 3. Some forms of distributed collaboration architectures in Bioinformatics

TABLE I
DISTRIBUTED ARCHITECTURES EVALUATION

| | **scalability/performance** | | **security/privacy of data** | **computational reproducibility** |
|---|---|---|---|---|
| **Architecture** | storage | computations | | |
| (a) | - | - | ++ | – |
| (b) | - | + | + | + |
| (c) | - | + | ++ | + |
| (d) | + | + | + | + |
| (e) | - | ++ | + | + |
| (f) | + | ++ | + | + |
| (g) | ++ | + | + | + |
| (h) | ++ | + | + | + |
| (i) | + | + | + | + |

We drop the following main conclusions from the table:

- Reproducibility is straightforward when data is pooled in a single server for processing and when virtualization can be integrated in the system, which is an asset of cloud computing.
- Data privacy is relatively high when data and computations do not migrate to beyond the control scope of its owner or when it is fully processed while encrypted
- Data privacy decreases when parts of the data are temporarily outsourced to be processed in the clear domain or when it is outsourced in plaintext for storage
- When the architecture is hybrid, computations performance increases due to easy scalability which alleviates eventual network exchanges slowdown.

## IV. CONSTRUCTING DISTRIBUTED BIOMEDICAL ANALYSES

We now consider the problem of how to design and implement distributed biomedical analyses that meet the requirements and challenges motivated before (placement, scalability, security/privacy and reproducibility). The discussion before, notably Table I, suggests that three of these challenges are essentially linked to the choice of architecture from Fig. 3(a). Most prominently, the data and computation placement strategy as well as the approach to scalability directly depend on the architecture. Reproducibility can be facilitated much by the

right choice of architecture. This leaves security and privacy issues as the challenge that have to be tackled using means less directly coupled to the choice of the architecture.

We therefore advocate the following *development method*:

1) Choose an architecture for the deployment and execution of the analysis that determines the placement strategy of data and computations depending on the requirements of the collaboration between research groups and available infrastructure (hospitals, clusters etc.)
2) Define strategies for scalability and reproducibility based on the elements of the architecture.
3) Define a security and privacy strategy in terms of local and remote data placement and computations as well as privacy-enhancing techniques, such as encryption and database fragmentation.
4) Implement the distributed analysis.

In the previous section we have mainly discussed the issues related to the first two steps. We now concentrate on how to handle the security/privacy issues. We then discuss the challenges the outlined development method faces in practice. Finally, we show how such a method can be harnessed for different evolutions of the collaboration model of the I-CAN study.

### A. Security and privacy-preserving technologies

Some security and privacy-preserving technologies for secure biomedical analyses are well known: symmetric and asymmetric encryption as well as homomorphic encryption, for example. However, other advanced security techniques, for instance, secure execution environments (Intel's SGX) and constraint-based data fragmentation are not yet largely known. Moreover, few mechanisms apart from encryption and access control mechanisms have been integrated into current biomedical systems. Furthermore, not all biomedical studies in need of multiparty collaboration got sufficient attention from security and cryptography research communities, as does the widely-used GWAS infrastructure for instance.

In this part, we briefly introduce advanced security and privacy mechanisms that have, for most, not yet been integrated into state-of-the-art secure distributed biomedical analyses.

*1) Homomorphic encryption:* With the ever increasing emergence of cloud based services, there was a need in developing algorithms that allow data processing on cloud-like infrastructures while preserving data privacy. Homomorphic encryption schemes have emerged as a solution to this challenge. Actually, Homomorphically encrypted data can be processed without the need to decrypt it first, that is, no need for a key. In addition, the resulting computations yield encrypted results. Homomorphic encryption has already been applied to biomedical analyses and to the GWAS in particular [12]–[14] employing architectures equivalent to the one depicted in 3(c). In fact, by using homomorphic encryption schemes, data owners can contribute to medical research without divulging anything about their data sets to researchers but the final results.

Homomorphic encryption generally has a significant impact on the efficiency of calculations. It is currently only reasonably usable for simple algorithms that perform additions or multiplications only. *Fully homomorphic encryption* that allows arbitrary computations to be performed over encrypted data is still very inefficient [15].

*2) Secure processing environments (SPE):* Secure environments for clear data processing can be constructed based on software support for trusted parties [10] or specialized hardware support, such as Intel's Software Guard Extension (SGX), which provides a notion of secure *enclaves*. SPEs are of particular interest if homomorphic encryption is not efficient enough or, more generally, if processing encrypted data is not applicable.

Recently, there is an increasing tendency in adopting secure hardware component based computations for biomedical analyses, in particular for the GWAS. The first was suggested by Canim *et al.* [16]. Interestingly, the secure component can be part of an untrusted server, for instance in a public cloud [17]. Actually, its characteristics of isolating critical computations on sensitive data from privileged software, for instance the OS, makes it a special case of trusted party based analyses.

Combinations of homomorphic encryption and SGX technology can be used, for example, when the encryption scheme is partially homomorphic but the analysis requires operations on data that cannot be performed in the encrypted domain. For example, Fisher's exact statistical test can be securely processed inside an SGX enclave after contingency tables of a single nucleotide polymorphism (SNP) from several contributors were added to each other using Paillier's additive homomorphic cryptosystem at the untrusted server level [18].

*3) Data fragmentation:* Data fragmentation is quite common in real-life deployments of biomedical infrastructures. For example, in I-CAN, data is fragmented and the fragments are distributed to different poles of expertise. Doing so has other virtues, indeed, fragmentation and distribution of (bio)medical data to separate servers alleviates some privacy-related risks in the sense that linking *e.g.,* clinical health records, images and genomes of patients for computations is then much more difficult for intruders. There are also formal models for fragmentation such as constraint-based fragmentation [19]: constraints specify which attributes should be physically separated, and the corresponding heuristics return a minimal number of fragments that satisfy the constraints; in practice, this determines the number of non-communicating servers needed to securely deploy the fragmented database. This type of fragmentation has also been integrated into a framework for secure distributed biomedical analyses [10].

### B. Developing distributed medical analyses

Secure distributed biomedical analyses is a relatively recent research domain that is centered around the execution of analyses on cluster or high-performance calculator infrastructures. Existing frameworks therefore do not involve advanced features for distributed analyses that we advocate in this paper, such as architecture-driven placement and scalability as well

as advanced security and privacy mechanisms. Using current tools that typically support features like sequential workflows and the parallel execution of analyses on cluster infrastructures using, for example, Apache Spark.

Secure biomedical analysis engineering today calls for specific workflow management tools aware of the locations of data and computation, privacy levels, scalability properties of different infrastructures in order to conduct distributed collaboration studies. For illustration purposes, consider an analysis workflow running data from Fig. 2 architecture, suppose that the steps are the following: 1) neuroimaging data is analyzed in order to select interesting patients, 2) genomes of the selected patients are sequenced and analyzed, 3) the relevant genomes will be contributed to an international association study cohort for deeper discovery. Unfortunately, to this date, there is no real-life workflow management tool that will manage the corresponding distributed deployment and execution properties, notably security and privacy issues, for instance, when data is transferred to the international realm at step 3.

### C. I-CAN revisited

Figure 4 shows two extensions of the distributed computing infrastructure supporting the I-CAN multi-centric study. These extensions are aimed at better supporting the scalability of the massive and heterogeneous data analyses ($C_1$), enhancing the computational reproducibility ($C_2$), as well as better preserving security and privacy ($C_3$) in the context of sensitive human data.

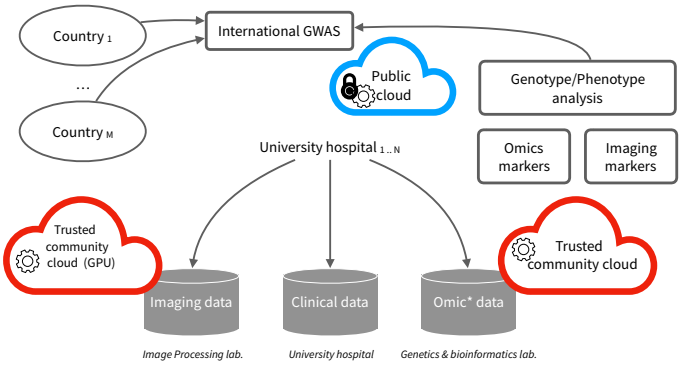



Fig. 4. Safe public clouds and trusted community clouds provide means to address reproducibility, privacy and scalability in the context of the intracranial aneurysm multi-center studies.

Among the state-of-the-art architectures summarized in Fig. 3 we argue that a combination of multiple community cloud infrastructures (Fig. 3 (d)) and secured public cloud infrastructures (Fig. 3 (c)) are the cornerstone to develop large-scale multi-center research studies.

For instance, the medical image processing could be scaled-up thanks to a dedicated computing infrastructure equipped with GPUs (red cloud entitled "Trusted community cloud (GPU)"). The scalability of the whole data processing could also be enhanced if multiple cloud infrastructures can be mobilized to the nearest data sources. This would prevent data relocation in a single cloud infrastructure which can cause network bandwidth, storage or compute bottlenecks and thus provide means to address challenge $C_1$.

In addition, relocating sensitive human data outside data production sites can raise legal or ethical issues. However, it is generally required when participating in an international collaboration. In this context, we advocate the use of public cloud infrastructures (blue cloud in Fig. 4) for the GWAS part, provided that data is protected beforehand using fully homomorphic encryption or combining partially homomorphic encryption with secure hardware component enabled at the public cloud server, both strategies allow full (privacy-preserving) processing in public cloud infrastructures (Fig. 3(c)). More precisely, an international GWAS can be used in a public cloud infrastructures in the following steps:

1) Each data owner computes locally the contingency tables for each SNP
2) Each data owner encrypts the tables with a homomorphic encryption scheme using the same encryption key
3) Each data owner sends the encrypted tables to the public cloud infrastructure
4) The public cloud adds the tables values together to get the final contingency table (of each SNP) corresponding to data from all participants
5) at this level, there are two possibilities:
   - Whether the statistical test is performed on the encrypted final contingency table, but this requires use of a fully homomorphic scheme at step 2. Otherwise,
   - The homomorphic scheme can be additive only, in which case, we require the public cloud to enable secure hardware based computations *e.g.,* sgx-enabled cloud, that will decrypt the table to calculate the statistical test and encrypts the obtained results
6) Encrypted results are communicated to researchers

Finally, by nature, both public and community cloud infrastructures leverage virtualization technologies and thus provide a key building block for more reproducible scientific computations (challenge $C_2$). Here, the collaborating partners agree on a precisely-defined software environment. These environments are then stored into a virtual machine responsible for the execution of the whole data analysis. Thanks to this virtual machine, the whole data analysis can be re-executed on a local computer or on any community cloud infrastructure thus enhancing the reproducibility of the analysis.

Summarizing this section, we think distributed biomedical analyses have to be supported by specialized tooling that supports architecture-based placement and scalability as well as largely automatic handling of security and privacy issues, in particular, at the level of biomedical workflows.

## V. Related work

The biomedical scientific community creates knowledge through collaborative and distributed workflows. We call a

scientific workflow *distributed and collaborative* when it links together computational tasks and datasets located in several geographically separated sites. In this section, we explore approaches that address the problems found when processing these distributed and collaborative workflows. Concretely, we present related work comparing it with the categories discussed earlier in this paper: data and computation placement, data security and privacy, scalability and performance. At the end of the section, we also consider how these studies compare with our proposal.

### A. Placement of computations and data

As seen before, when realizing collaborative and distributed workflows several constraints arise (e.g., legal or technical) restricting data movement and computation placement. In this section, we present studies addressing the problem of data and computation placement in these workflows. The problem of data placement is a prominent problem to solve in scientific workflow systems [20]. It is a big challenge in biomedical analysis, making it necessary to minimize data transfer among distributed data centers. In many cases, to perform such optimization, it is required to know the data dependency, the bandwidth limit, and the storage capacity on each data center. However, finding the optimal placement of a dataset in a distributed system is an NP-hard problem [21], and then, several strategies have been proposed to find approximate solutions. Yuan et al. [22] propose an approach based on k-means clustering, guaranteeing a balanced distribution of data, even for the amount of data, using a data placement matrix. This matrix is created using the information of the data centers (size and storage capacity). Similarly, Zhao et al. [21] implement a data placement based using genetic algorithms using a similar matrix.

In contrast, several proposals have addressed the problem of optimal computation placement under a given set of constrains [23]–[26]. As expected, these studies have shown that moving the computation near the data is much "cheaper" than moving the data to the computation.

### B. Data security and privacy

Security and privacy must be granted during extraction, transfer, and processing of data in collaborative and distributed workflows. Addressing these requirements is challenging, especially considering the amount of biomedical data involved, the inherent insecurity and heterogeneity of the network, the lack of strong information security policies on the organizations, and the technical complexity of information security solutions (see for example the discussion on [27]).

There are different approaches to minimize these risks. Ken et al. [28] propose an analysis technique capable of processing ciphered data located on third-party servers, without accessing the actual data, reducing then, the risk of unauthorized access or a data leak [29]. Similarly, Vinod et al. [30] propose a homomorphic method of probabilistic encryption, supported on the additive and multiplicative properties of homomorphisms and the Euler's theorem, to process data without looking at it.

Other approaches have investigated security strategies for storage, transfer, and sharing of biomedical data. These approaches concentrate on preserving the integrity, preserving confidentiality, and respecting data ownership. For example, GIFT-Cloud [31] is a collaborative platform sharing biomedical data and images, that is supported on an advanced technical architecture. In the platform, all data transferred to the server is encrypted and passed through a chain of gateways and firewalls guaranteeing the protection against intruders. Similarly, XNAT [3] adapts to the special needs of highly controlled data exchange and provides different levels of protection for private, protected and public information. These collaborative platforms have been subject of several studies addressing compliance with biomedical information security policies, for example, iDASH [32] is a collaborative platform compatible with the Health Insurance Portability and Accountability Act (HIPAA) [4].

Finally, other studies propose an integrated solution where security is considered at the workflow definition. In [33], [34], authors propose techniques to enforce security constraint on workflow definitions. Similarly, Chebotko et al. [35] propose three levels of security specification within workflows: at task level, at port level, and at data channel level.

### C. Scalability and performance

Several approaches addressing scalability and performance have been proposed. In this section, we discuss three categories of studies: studies based on the optimal distribution of computation on a cluster located in one site; approaches combining workflow support with distributed computation; and standalone approaches.

Hadoop[5] is a framework for distributed and parallel data processing and analysis of large amounts of data. Hadoop, was originally based on the MapReduce heuristic but currently, it can be used to implement any kind of distributed algorithm. GATK (Genome Analysis Toolkit) [36] was the first bioinformatics applications implemented using MapReduce and is widely used in genomic data analysis. Crossbow [37] supports different processing models including Hadoop cluster, a single computer, or Amazon EMR (Elastic Map Reduce). Similarly, MetaSpark, Halvade, CloudBurst, DistMap, SeqWare, and Hydra process the data using Hadoop or Spark[6] [6]. MapReduce and Hadoop have been widely used in Next Generation Sequencing (NGS) problems due to the way they parallelize and distribute the data [8]. However, many bioinformatic problems cannot be solved using those tools because they require advanced computer skills to design efficient solutions.

To help researchers to design efficient experiments on the tools described above, several studies investigate the use of distributed workflow languages or environments to define distributed scientific workflows. Galaxy, Taverna, and Wings [38], [39] are Data Workflow Systems (DWFSs). Galaxy is

[3]https://www.xnat.org
[4]https://www.ncbi.nlm.nih.gov/books/NBK500019/
[5]https://hadoop.apache.org/
[6]https://spark.apache.org/

very popular in the bioinformatics community and allows large scale data analysis from different formats [40], and is designed under the model PaaS. In contrast, Taverna was conceived under the model SaaS. It has been widely used in several fields such as bioinformatics, biodiversity, social sciences, and astronomy [38], [40]. Finally, Wings has been used in different domains, such as sciences, text analytics, geosciences, and multi-omics analysis (see Malcolm et al. [40] for a more detailed discussion on workflows system).

Finally, there are still many standalone applications that perform data analysis in single computers. VaRank has been used locally in the identification of genes in some diseases. Similarly, KD3 and S-MART allow extracting knowledge from biological data locally, and CAFE allows alignment-free genome and metagenome comparisons [41]–[44].

Based on the above and according to [38], most of the current DWFSs are not efficient, nor are they lightweight, nor contribute to large-scale data analysis in an efficient way. For this reason, great challenges are identified in the design of workflow systems that cover the three categories discussed in this section.

### D. Discussion

Table II compares the tools discussed above against the architectural features proposed for a distributed and collaborative framework for biomedical analyses (see Sec. IV), which are: Data and computation placement; Security and confidentiality; and scalability and performance. Each feature, is further divided in sub-categories, this allows us to have more detailed analysis. The 'X' indicates that the property is satisfied in some proportion, and the blanks indicate otherwise.

In the table, applications designed for local computations and local data placement, were considered more likely to have stronger security policies and easier implementation of those policies. So they were marked as having very secure features in storage, data transfer, and processing. In contrast, tools with distributed data or computation placement, were marked as secure only when it was stated explicitly in the literature.

Many applications grant fragmentation and distribution of data in a cloud infrastructure (e.g., Map Reduce services on Amazon Web Services), but they do not necessarily ensure the placement of data in different locations. In these systems, the security and privacy risk with the data increases, since there is an intervention of third parties, while the scalability and performance of the data improve significantly because such systems are designed to scale dynamically.

It is important to note, that very few tools provide a mechanism for describing distributed workflows, thus the challenge of reproducibility is a very hard constrain for biomedical researchers. Reproducibility, in the current state of the art depends on the skills and expeerience of reasearchers and not on the availability of strong and standard workflow languages.

In summary, this classification allows us to verify that there are not collaborative frameworks, supporting scientific workflow description, and satisfying all the architectural features discussed in this work. This is a great challenge and an opportunity for research and implementation of optimal collaborative frameworks.

## VI. Conclusion

In this paper we have motivated that multi-party studies involving biomedical analyses often benefit from distributed collaborations. They are, however, subject to difficult challenges concerning the computation and placement of data and computations, scalability and reproducibility of computations, as well as the security and privacy, notably of sensitive data.

We have introduced a set of nine architectures for such distributed analyses and sketched a development method that starts by selecting an architecture in order to determine the placement and scalability strategies, develop an appropriate security and privacy enforcement strategy before implementing the analysis itself. We have also illustrated the challenges and solutions we propose in the context of a real-world multi-stakeholders biomedical study, the I-CAN project.

As future work, we plan to extend an existing framework [10] in order to provide full-fledged support for the security and privacy challenge of distributed biomedical analysis. Furthermore, we plan to extend tools that are used by biomedical researchers to implement a comprehensive development method for distributed analyses.

## References


[1] T. Manolio, "Collaborative genome-wide association studies of diverse diseases: programs of the nhgris office of population genomics," *Pharmacogenomics*, vol. 10(2), 2009.

[2] Y. Zhang, M. Blanton, and G. Almashaqbeh, "Secure distributed genome analysis for gwas and sequence comparison computation," in *BMC medical informatics and decision making*, vol. 15, no. 5. BioMed Central, 2015.

[3] Q. Li, T. Yang, and otherse, "Large-scale collaborative imaging genetics studies of risk genetic factors for alzheimers disease across multiple institutions," in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016.

[4] C. Fuchsberger, J. Flannick *et al.*, "The genetic architecture of type 2 diabetes," *Nature*, 2016.

[5] J. Luo, M. Wu *et al.*, "Big data application in biomedical research and health care: a literature review," *Biomedical informatics insights*, vol. 8, 2016.

[6] G. Cattaneo, R. Giancarlo *et al.*, "Mapreduce in computational biology-a synopsis," in *Italian WS on Artificial Life and Evolutionary Computation*. Springer, 2016.

[7] L. Dai, X. Gao *et al.*, "Bioinformatics clouds for big data manipulation," *Biology direct*, vol. 7, no. 1, 2012.

[8] R. Taylor, "An overview of the hadoop/mapreduce/hbase framework and its current applications in bioinformatics," in *BMC bioinformatics*, vol. 11, no. 12. BioMed Central, 2010.

[9] N. Homer, S. Szelinger *et al.*, "Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays," *PLOS Genetics*, vol. 4, no. 8, 08 2008.

[10] F.-z. Boujdad and M. Südholt, "Constructive Privacy for Shared Genetic Data," in *CLOSER 2018 - 8th Int. Conf. on Cloud Computing and Services Science*, ser. Proceedings of CLOSER 2018, Mar. 2018.

[11] J. S. Sousa, C. Lefebvre *et al.*, "Efficient and secure outsourcing of genomic data storage," *BMC medical genomics*, vol. 10(Suppl 2), 46, 2017.

[12] Y. Zhang, W. Dai *et al.*, "Foresee: Fully outsourced secure genome study based on homomorphic encryption," *"BMC Medical Informatics and Decision Making*, vol. 15, no. 5, Dec. 2015.

[13] W. Lu, Y. Yamada, and J. Sakuma, "Privacy-preserving genome-wide association studies on cloud environment using fully homomorphic encryption." *BMC medical informatics and decision making*, vol. 15 Suppl 5, 2015.

TABLE II
CLASSIFICATION OF SOME BIOMEDICAL APPLICATIONS

| | Categories and Features | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Placement of computations and data (Sec. V-A) | | | Security and privacy (Sec. V-B) | | | Scalability and performance (Sec. V-C) | | |
| **Name** | **Locally** | **Distributed data** | **Distributed processes** | **Storage** | **Transfer** | **Processing** | **Distributed** | **Local** | **DWFS** |
| Alfree | X | | | X | X | X | | X | |
| CAFE | X | | | X | X | X | | X | X |
| CloudBurst | | X | X | X | | | X | | |
| Crossbow | | X | X | X | | | X | | |
| CloudMan | | X | X | X | | | X | | |
| DistMap | X | | | X | X | X | X | | X |
| Galaxy | | X | X | X | | X | X | X | X |
| GIFT-Cloud | | X | | X | X | X | X | | X |
| GATK | X | X | X | X | | | X | | |
| Halvade | | X | | X | | | X | | |
| iDASH | | X | X | X | | X | X | | |
| S-MART | X | | | X | X | X | X | | |
| SeqWare | X | | | X | X | | X | | |
| Sparkhit | | X | X | X | | | X | | |
| Taverna | X | X | | X | | | X | | X |
| VaRank | X | | | X | X | X | X | | |
| VisTrails | X | X | X | X | | | | | X |
| Wings | | X | | X | | | X | | X |


[14] S. Wang *et al.*, "Healer: homomorphic computation of exact logistic regression for secure rare disease variants analysis in gwas," *Bioinformatics*, vol. 32, no. 2, 2016.

[15] A. Acar *et al.*, "A survey on homomorphic encryption schemes: Theory and implementation," *ACM Comput. Surv.*, vol. 51, no. 4, Jul. 2018.

[16] M. Canim, M. Kantarcioglu, and B. Malin, "Secure management of biomedical data with cryptographic hardware," *Trans. Info. Tech. Biomed.*, vol. 16, no. 1, Jan. 2012.

[17] F. Chen, C. Wang *et al.*, "Presage: Privacy-preserving genetic testing via software guard extension," *BMC medical genomics*, vol. 10(Suppl 2), 48, 2017.

[18] N. Sadat, M. A. Aziz *et al.*, "SAFETY: secure gwas in federated environment through a hybrid solution with intel SGX and homomorphic encryption," *CoRR*, vol. abs/1703.02577, 2017.

[19] V. Ciriani, S. Vimercati *et al.*, "Combining fragmentation and encryption to protect privacy in data storage," *ACM Trans. Inf. Syst. Secur.*, vol. 13, no. 3, Jul. 2010.

[20] X. Li, L. Zhang *et al.*, "A novel workflow-level data placement strategy for data-sharing scientific cloud workflows," *IEEE Trans. on Services Computing*, 2016.

[21] Z. Er-Dun, Q. Yong-Qiang *et al.*, "A data placement strategy based on genetic algorithm for scientific workflows," in *2012 Eighth Int. Conf. on Computational Intelligence and Security*. IEEE, 2012.

[22] D. Yuan, Y. Yang *et al.*, "A data placement strategy in scientific cloud workflows," *Future Generation Computer Systems*, vol. 26, no. 8, 2010.

[23] R. Stewart, P. Trinder *et al.*, "Comparing high level mapreduce query languages," in *Int. WS on Adv. Parallel Proc. Techn.* Springer, 2011.

[24] M. Ebrahimi, "Data placement and task mapping optimization for big data workflows in the cloud," Ph.D. dissertation, Wayne State University Dissertations, 2017.

[25] D. Agrawal, A. El Abbadi *et al.*, "Data management challenges in cloud computing infrastructures," in *Int. WS on Databases in Networked Information Systems*. Springer, 2010.

[26] M. Ebrahimi, A. Mohan *et al.*, "Bdap: a big data placement strategy for cloud-based scientific workflows," in *2015 IEEE First Int. Conf. on Big Data Computing Service and Applications*. IEEE, 2015.

[27] C. Tan, L. Sun, and K. Liu, "Big data architecture for pervasive healthcare: A literature review," in *ECIS*, 2015.

[28] K. Naganuma *et al.*, "Privacy preserving analysis technique for secure, cloud based big data analytics," *Hitachi Rev*, vol. 63, no. 9, 2014.

[29] C. Hasti and A. Hasti, "Data security in cloud-based analytics," in *Big Data Analytics*. Springer, 2018.

[30] V. Kumar, R. Kumar *et al.*, "Fully homomorphic encryption scheme with probabilistic encryption based on eulers theorem and application in cloud computing," in *Big Data Analytics*. Springer, 2018.

[31] T. Doel and D. o. Shakir, "Gift-cloud: A data sharing and collaboration platform for medical imaging research," *computer methods and programs in biomedicine*, vol. 139, 2017.

[32] L. Ohno-Machado, V. Bafna *et al.*, "idash: integrating data for analysis, anonymization, and sharing," *J. of the American Medical Informatics Association*, vol. 19, no. 2, 2011.

[33] Y. Gil, W. Cheung *et al.*, "Privacy enforcement in data analysis workflows," in *Proceedings of the 2007 Int. Conf. on Privacy Enforcement and Accountability with Semantics-Volume 320*. Citeseer, 2007.

[34] S. Davidson, S. Khanna *et al.*, "Privacy issues in scientific workflow provenance," in *Proceedings of the 1st Int. WS on Workflow Approaches to New Data-centric Science*. ACM, 2010.

[35] A. Chebotko, S. Chang *et al.*, "Scientific workflow provenance querying with security views," in *2008 The Ninth Int. Conf. on Web-Age Information Management*. IEEE, 2008.

[36] A. McKenna, M. Hanna *et al.*, "The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data," *Genome research*, vol. 20, no. 9, 2010.

[37] J. Gurtowski, M. Schatz, and B. Langmead, "Genotyping in the cloud with crossbow," *Current protocols in bioinformatics*, vol. 39, no. 1, 2012.

[38] R. Karim, A. Michel *et al.*, "Improving data workflow systems with cloud services and use of open data for bioinformatics research," *Briefings in bioinformatics*, vol. 19, no. 5, 2017.

[39] S. Cohen-Boulakia, K. Belhajjame *et al.*, "Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities," *Future Generation Computer Systems*, vol. 75, 2017.

[40] M. Atkinson, S. Gesing *et al.*, "Scientific workflows: Past, present and future," 2017.

[41] V. Geoffroy, C. Pizot *et al.*, "Varank: a simple and powerful tool for ranking genetic variants," *PeerJ*, vol. 3, 2015.

[42] A. Dander, M. Handler *et al.*, "[kd 3] a workflow-based application for exploration of biomedical data sets," in *Trans. on large-scale data-and knowledge-centered systems IV*. Springer, 2011.

[43] Y. Lu, K. Tang *et al.*, "Cafe: accelerated alignment-free sequence analysis: Supplementary material," *The University of Southern California*, 2017.

[44] M. Zytnicki and H. Quesneville, "S-mart, a software toolbox to aid rna-seq data analysis," *PloS one*, vol. 6, no. 10, 2011.