EGI-Engage

Analysis of requirements on biobank and study workflows

D6.8

Date	02 November 2016
Activity	SA2
Lead Partner	BBMRI-ERIC
Document Status	FINAL
Document Link	https://documents.egi.eu/document/2931

Abstract

Human biobanks, which are the core of BBMRI-ERIC medical research infrastructure, are repositories of biological material and data associated with the research participants (donors or patients willing to participate in the research). The associated data covers a broad range of data types: from data collected directly from the research participants and medical processes related to them, to data generated from the biological material. This document focuses on describing biobank data processing workflows that were selected for piloting in EGI-Engage the biobanks by the BBMRI.nl and BBMRI.cz (national nodes of BBMRI-ERIC) together with their associated biobanks. The main focus is on proteomics and genomics workflows, which cover both extremes of privacy-sensitive data processing spectrum: from relatively non-sensitive applications to very sensitive ones.

The EGI-Engage project is co-funded by the European Union (EU) Horizon 2020 program under Grant number 654142 http://go.egi.eu/eng

COPYRIGHT NOTICE

DELIVERY SLIP

	Name	***Partner/Activity***	***Date***
From:	Morris Swertz, Pieter Neerincx, Ondřej Vojtíšek, Petr Holub	BBMRI Competence Centre	2016-10-19
Moderated by:	Małgorzata Krakowian	EGI.eu/NA1	2016-10-31
Reviewed by	Luca Pirredu Gergely Sipos	CRS4 EGI.eu	2016-10-26 2016-10-22
Approved by:	AMB and PMB		

DOCUMENT LOG

Issue	***Date***	***Comment***	***Author/Partner***
v.1	2016-10-19	First version of the document for review.	Morris Swertz, Pieter Neerincx, Ondřej Vojtíšek, Petr Holub
FINAL	2016-10-31	Final version	Petr Holub

Contents

1 Introduction

Human biobanks, which are the core of BBMRI-ERIC medical research infrastructure, are repositories of biological material and data associated with the research participants (i.e., donors or patients willing to participate in the research). The associated data spans a broad range of data types: from data collected directly from the research participants and medical processes pertaining to them, to data generated from their biological material via, e.g., laboratory experiments or imaging devices, to additional data such as environmental exposure or lifestyle data (which will become particularly abundant due to availability of wearables devices) – as shown in Figure 1. Biobanks also provide services based on their capacities (e.g., hosting of samples and data) and their expertise (e.g., analysis of data from molecular experiments, data integration). All of these resources are made available for research purposes. This requires access to the data in a suitable way (SFTP, permanent storage, HPC storage), but also access to the necessary processing resources (HPC/compute, pipelines & software). Because they are dealing with human biological material and data, BBMRI-ERIC biobanks need to pay particular attention to aspects of data protection, since privacy protection of donors (population biobanks) or patient subjects (from disease biobanks) are one of the keys to trustworthiness of the research infrastructure.

Following the unprecedented growth of the size of research data in medicine and biology, as witnessed in genomics, proteomics, metabolomics and other types of so called “omics” data, as well as large scale imagery and lifestyle data, BBMRI-ERIC has become part of the EGI-Engage project in BBMRI Competence Centre (BBMRI CC), in order to explore how cloud-based scalable data processing and storage can be used to improve the research process. BBMRI-ERIC is participating in the BBMRI CC together with several national nodes, BBMRI.cz, BBMRI.nl and BBMRI.se, which agreed to contribute their selected workflows and explore their applicability to the cloud scenarios.

This deliverable provides an overview of the workflows, that were identified by the national nodes as a good candidates for their piloting within BBMRI CC. We have tried to cover a broad spectrum of workflows: from privacy-sensitive workflows that are specific to medical research dealing with individual research participants data, to less sensitive or non-sensitive workflows that are representatives of more generic biological and chemical analytical processes. This should span a significant breadth of needs of biobanks: from applications, that are typically restricted to very protected storage and processing environments, to applications that can unleash the full potential of scalability of the cloud computing (and that are expected to be shared with many other domains). The workflows are, however, not covering the whole breadth of workflows that are run by the biobanks – we know this from the participating national nodes (we are aware that other national nodes are also running other types of processing such as large-scale image analysis in BBMRI.it and BBMRI.nl), but this selections had to be made due to constrained work capacity available in the BBMRI CC.

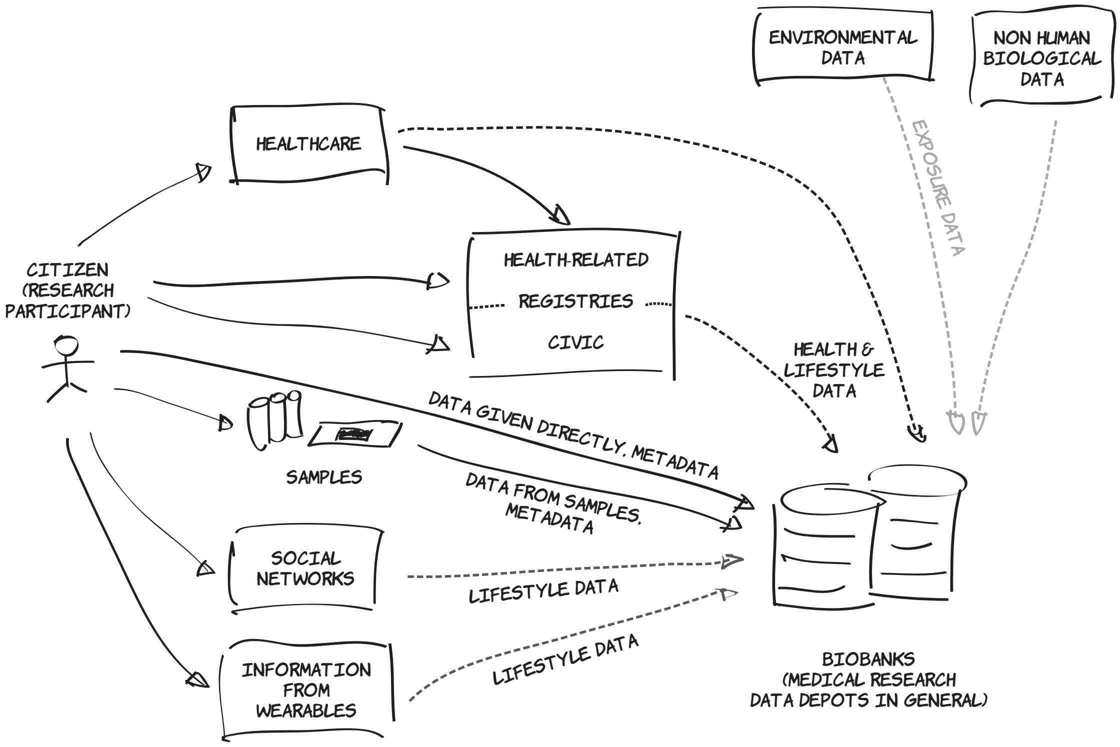
ENVIRONMENTAL DATA
NON HUMAN BIOLOGICAL DATA
HEALTHCARE
EXPOSURE DATA
HEALTH-RELATED
REGISTRIES
CIVIC
CITIZEN (RESEARCH PARTICIPANT)
HEALTH & LIFESTYLE DATA
DATA GIVEN DIRECTLY, METADATA
SAMPLES
DATA FROM SAMPLES, METADATA
SOCIAL NETWORKS
LIFESTYLE DATA
LIFESTYLE DATA
INFORMATION FROM WEARABLES
BIOBANKS (MEDICAL RESEARCH DATA DEPOTS IN GENERAL)

Figure 1: Flow of data into the biobanks.

2 Description of Selected Representative Use Cases

Below we describe representative use cases from the biobank domain. These use cases have in common the following aspects that are to some extent specific to the expectations that biobanks and their users have from EGI and other cloud services:

1. ability to provide access to isolated areas of storage and compute using federated authentication,
2. ability for user groups to deploy standard software and/or pipelines,
3. (optionally for specific workflows) a proof that the 'digital research environments' are up to hospital standards.

Below we first summarize example use cases from proteomics and genomics and subsequently discuss additional security constraints that are required by some biobanks (such as larger biobanks with extensive regulations and/or patient biobanks).

2.1 Proteomics

Proteomics is the analysis of proteins in samples from the human body (this is specific to human biobanks) typically using mass spectrometry. It relies on identifying different proteins in the samples by matching their known "mass spectrum footprints" based on available databases. Proteomics is not used for diagnostic purposes as of now, since it is expensive (cost of equipment purchase and maintenance) and still not standardized and precise enough for clinical use. Proteomics plays an important role in medical research because it helps find biomarkers used in diagnostic and health procedures.

Cloud infrastructure can be effectively utilized for the proteomics analysis for those users that either do not have the sufficient hardware, or who perform the analyses at high volumes and need elastic compute capacity that can scale up based on the requirements.[1] If third party cloud infrastructures are used, they can analyse the data extremely fast and the network transmissions of data may become the bottleneck and the improvement would not be significant; hence in order to utilize an infrastructure provided by a third party, network links with appropriate capacity are required.

Data are not sensitive from perspective of patient data privacy: the output from mass spectrometry or from downstream analysis itself does not contain any information that could be reasonably used to identify the original patient.

[1] Hardware for mass spectrometry is substantially more costly than the computing hardware, hence for small-scale analyses with stable compute capacity requirements, this is less attractive.

2.2 Processing of Genomics Data

New 'next-generation sequencing' technology (NGS) are enabling increasingly large DNA and RNA experiments. DNA can be considered as a "book" that consists of 23 chapters (chromosomes) containing a total of ∼30,000 words (regions of the DNA we call genes). As a whole, the book containing DNA is transcribed (copied) to RNA molecules that, in turn, result in the proteins that constitute much of the human body. The human genome consists of about three billion building blocks (base pairs, or letters in our book analogy) from an alphabet of four letters (ACTG). Of these 3 billion only <1% makes up known genes; the other 99% of letters between the genes we think have a 'regulatory' function that influences which genes are switched 'on' or 'off', thus controlling how many copies of RNA are made from DNA during over the cells' lifetime. There is still much to be studied and understood in this area.

At the DNA level, NGS enables the measurement of multiple or even all genes at once (also called 'exome') or even the measurement of the whole of the DNA (whole genome). These observations enable fine grained characterisation of all DNA variation points between individuals, i.e., differences in 'genotype'. This information forms the basis for analyses to identify the genetic determinant of heritable traits and to diagnose diseases caused by DNA mutations. At the RNA level one can measure the quantity of RNA products per gene (transcripts) and/or even genotype each of the RNA products to investigate differences at the RNA level. Much of the focus of this research is now on 'personalized medicine and health'—for instance how can we predict what (expensive) medicine is most appropriate given a patient's DNA and RNA profiles.

To extract information from these NGS data a significant amount of computational analysis is required. The analysis consists of a series of computational steps, which are often collectively referred to as a 'pipeline'. First the data need to be made ready for analysis (what is called 'processing') because the NGS process does not actually produce complete sequences of whole chromosomes; instead it measures small fragments of 100–200 base pairs. Then, the analysis starts with large files (GBs) that contains thousands of fragments ('reads'), i.e., short text sequences of CATG. The first processing step is to reconstruct these fragments into complete chromosomes, which is typically done by aligning the fragments to the known human 'reference' genome. Subsequently, the differences between the experimentally derived sequence and the reference are computed by variant calling, resulting in a listing of the differences. On average these variations amount to about 1 letter per stretch of 1000 base pairs although also larger differences are common (deletions, insertion, translocations). During processing many intermediate steps are needed. In case of RNA, the number of reads per gene is quantified as a proxy for quantifying 'gene expression'—i.e., the amount of RNA per gene. Finally, large statistical analysis can be done comparing DNA variation of many individuals to phenotypes (such as disease or height) and/or comparing RNA expression. Also, in case of diagnostics, individual genotypes/RNA expression can be compared to known reference data as basis for diagnosis.

A key aspect of the procedures is that analysis environments must be portable such that data and pipelines can be rapidly deployed on a new facility. For example, in UMC Groningen thre are two research clusters and two diagnostic clusters next to the BBMRI.nl central clusters, and it is essential that data and pipelines can be moved or replicated across these clusters in a reproducible way.

2.3 Sensitivity considerations for human subject research clouds

For routine analysis with constant workloads many biobanks and research institutes have acquired adequate HPC facilities. However, in many cases data from multiple biobanks needs to be integrated in order to reach sufficient statistical power. Typically, such large analyses are implemented as a large multi-center (and often even multi-national) consortiums where researchers from many institutes need to collaborate around the data, requiring centralized access to data and analysis procedures. Such large-scale facilities are beyond what individual institutes can provision, so there is a demand for 'cloud' solutions (IaaS, SaaS) that enable research consortia to have a 'digital research environment for human subject research data' on which to conduct their analyses with suitable facilities. Naturally, this demand for scale-out facilities is very much present among BBMRI-ERIC Members.

New research methods such as NGS and 'personalized medicine' are also rapidly uptaken in the context of healthcare. In addition the speed at which new analysis methods are translated from research to health care is increasing. As diagnostics facility often don't have access to HPC facilities, BBMRI members are often asked to also make analysis pipelines available to diagnostics labs, where they have to meet or exceed high operating standards (Stand Operating Procedures = SOPs) and privacy constraints. However, next to the requirement of pipeline portability that was previously mentioned, hospitals often also have additional requirements before the hospital, such as high standards for validation/verification and SOPs (which the BBMRI-ERIC Members can provide). Moreover, hospitals also need access to suitable storage and compute facilities that implement high availability measures while considering ethical, legal and societal implications, such as ensuring privacy, which introduce significant technical complexity.

Some biobanks and all hospitals will therefore require proof that the research environment adheres to sufficient measures for information security. This is in particular relevant because DNA is very identifiable data of which a fraction is sufficient to re-identify a person. Also, DNA is often compared to all kinds of phenotypes that also might be sensitive, such as (predisposition for) disease. So any implementation of cloud solutions will need to be evaluated against these considerations.

3 Proposed Pilot Workflows and Evaluation

This section provides an overview of the particular workflows that will be implemented in the BBMRI Competence Centre of EGI-Engage. They fall into two categories

- workflows that do not handle privacy-sensitive data and are therefore good candidates to be processed on the wider EGI FedCloud infrastructure for improved scalability;
- privacy-sensitive workflows that will use the private cloud infrastructures built in the biobanks using EGI FedCloud (or possibly by BBMRI-ERIC National Nodes and provided to the biobanks as virtual private cloud infrastructure).

The first category covers several different proteomics worflows, which are the main focus of the BBMRI.cz National Node and RECAMO biobank involved in the EGI-Engage, related to identification and quantification of proteins in the biological samples and analysis of their properties. These workflows are part of the production processes embedded in the medical research of the RECAMO biobank. In the future, these can be extended with computationally more demanding workflows, such as OpenSWATH-based identification of candidate proteins for biomarkers based on heatmap differences of proteins between healthy tissues and tumors for patient groups with identical diagnoses. The data formats used in the proteomics applications and further referenced in this report are described in [1].

On behalf of the BBMRI.nl, we plan to try and implement the processing pipeline for NGS DNA alignment and genotyping in the cloud, typically the first and most computationally intensive analysis step before the biobank data analysis can commence. This pipeline is an example for many more pipelines and once this pilot is successful we expect to also deploy RNA genotyping and and GWAS genotype array imputation pipelines. Finally, we evaluate the pilots against the requirements with particular attention to portability (which is a key requirement for these pipelines) and information security (which is essential before sensitive data from biobanks and hospitals can be allowed to be sent to cloud providers).

3.1 Workflow: Identification of proteins

The objective of protein identification is to identify an unknown protein. In this identification procedure, the unknown protein is typically broken up into its component peptides; the peptides are then ionized and introduced into the mass spectrometer which will determine their mass. The measured mass of the peptides (raw outcome from mass spectrometry in .raw format) is used as the input to an analytic software (Proteome discoverer) which searches database of known proteins and matches the hit based on similarity of mass spectra. The result is list of proteins and probability of hit.

- GOAL: identify unknown protein
- INPUT:
  - .raw file (approx. 1 GB), containing mass of peptides

- PROCESS:
  - Find the best probable hit in database of known proteins (free downloadable,[2] approx. 300 MB, text file: name and sequence, .fasta).
  - Approx. 1–6 hours on a workstation PC (depends on protein database size and on parameters).
  - Used software: Proteome Discoverer.[3]
- OUTPUT:
  - .msf [2]
  - List of proteins from the database that best match the measured peptide profile, along with the probability of hit.
- PARALLELISM
  - at least on the level of individual analyses

3.2 Workflow: Protein quantitation

The goal is to measure the mass of protein in a sample. The obtained information can be further used to understand cancer behaviour by comparing the level of protein in healthy tissue with the tumor, since the change of protein level might be an important disease biomarker.

There are two basic types of quantitation: label-free or label. In the label-free method all samples are measured separately and then analyzed; this results is a lot of files and the method is more time consuming. On the other hand, the label method uses the isotopes to mark sample. Two or more samples are mixed together and put into mass spectrometer. Thanks to the isotopes the device can distinguish between the peptides from different samples and it can measure all of them in one single run.

In summary, the label-free: more data generated and needed for analysis, reproducible, more precise, more time and computational resources needed; label: generates less data, faster, less computational resources needed for analysis, more expensive (labels), less reproducible. Quantitation can be relative or absolute. Relative is faster and for many purposes fully sufficient (output is the ratio among measured samples).

- GOAL: Measure amount of proteins in sample. It can be used to compare protein levels in tumor and in healthy tissue.
- INPUT:

[2] http://www.uniprot.org/downloads

[3] https://www.thermofisher.com/order/catalog/product/IQLAAEGABSFAKJMAUH

  - .wiff (label of label-free method) or .raw (label method)

- PROCESS:

  - Find the best probable hit in database of known proteins (free downloadable,[4] approx. 300 MB, text file: name and sequence, .fasta).
  - First you identify relative amount of protein types in sample and then compare tuples. For more precise analysis the identification is done with X outputs from mass spectrometer (i.e., x .wiff files of tumor and x .will files of healthy tissue). The analysis is more computationally intensive than identification and it needs more RAM (i.e., $2 * x * 2$–5 GB).
  - Approx. 24 hours on a workstation PC.
  - Used SW:
    * Label free method: Protein Pilot[5] (identification), PeakView[6] (quantitation), MarkerView[7] (statistical analysis, visualization of results)
    * Label method: Proteome discoverer (both identification and quantitation)

- OUTPUT:

  - Relative amount of protein types in sample, difference between/among given samples.

- PARALLELISM

  - at least on the level of individual analyses

3.3 Workflow: HDX analysis

Information about change of protein structure after interaction with ligand (based on the change of deuteration) - comparison of protein deuteration without ligand with protein deuteration with ligand. This is used to explore protein folding and drug interactions (ligands) with proteins.

- INPUT:

  - .raw data (approx. 0.5 GB)

- PROCESS:

  - Computationally intensive
  - Used SW: HDExaminer[8]

[4] http://www.uniprot.org/downloads
[5] http://sciex.com/products/software/proteinpilot-software
[6] http://sciex.com/products/software/peakview-software
[7] http://sciex.com/products/software/markerview-software
[8] http://www.massspec.com/HDExaminer.html

- OUTPUT:
  - 5–15 GB of output data per experiment (size depends on experiment extent)
- PARALLELISM
  - at least on the level of individual analyses

3.4 Workflow: DNA alignment and genotyping

As pilot we propose to use the BBMRI.nl DNA pipeline. This pipeline converts raw sequence fragments (reads) for each sample into a genotype file, which contains the DNA differences of each sample compared to a common reference. This is an example for more of these pipelines and consist of a series of command-line tools that need to run in specific order as jobs. To enable portability of these pipelines, the environment is configured using an Ansible playbook, which installs EasyBuild for the reproducible deployment of the pipeline and the binaries it uses and the Lmod module system for loading versioned software during pipeline execution. We expect also a job scheduler such as SLURM to manage the jobs as well as shared storage when spreading the load across multiple computer nodes. It is essential that the pipelines can be deployed without root/admin permissions.

- The most recent version of the pipeline including installation instructions is documented here: `http://molgenis.github.io/pipelines/`.
- For software dependency management this pipeline relies on EasyBuild[9] @ compile / installation time and Lmod[10] @ run time.

  Lmod is installed as root by sys admins and from the repos of the Linux distro used. EasyBuild on the other hand is installed by deploy admins (bioinformaticians) without root privileges.

The DNA pipelines contains 20+ steps, which can be divided in 4 parts:

1. Demultiplexing: conversion of raw sequence machine output for multiplexed samples into sequence reads per sample. For example for Illumina machines the raw data consists of base calls per cycle in BCL file format. The raw data formats are sequencing platform specific; hence there is no standard. For sequence reads the de facto standard file format is gzip compressed FastQ (*.fq.gz), which contains strings of nucleotides A, T, C or G and associated quality scores.
2. Alignment of sequence reads to a reference genome. De facto standard file format for aligned reads: Binary SAM (*.bam)

[9] `https://github.com/hpcugent/easybuild`
[10] `https://www.tacc.utexas.edu/research-development/tacc-projects/lmod`

3. Variant calling to determine where samples differ compared to the reference genome. De facto standard file format for variant calls: bgzip compressed VCF (*.vcf.gz)

4. Quality control. No de facto standards exist, but QC reports are usually a PDF with tables and plots of summary statistics.

Due to the nature of the raw data, the first 'demultiplexing' step is performed either directly on the sequence machine itself or on a dedicated server close to the sequence equipment. (The raw sequence data usually contains many and small files, which makes them inefficient to transfer over networks or store on large parallel storage systems). Once the raw data is converted into FastQ files these can be processed efficiently in parallel on clusters. Processing does not require the use of MPI as the data can be split in chunks (per sample or per chromosome or per per chromosome arm or even smaller regions when necessary), which can be processed independently of each other.

The size of the data will depend on the size of the genome of the species investigated and the genomic region analysed. Typical for human data:

1. Gene panels:

   a) 50–150 genes.

   b) ≪2% of the complete genome

2. Whole Exome Sequencing (WES)

   a) 23,000 genes.

   b) 2–5% of the complete genome depending on amount of included flanking sequence data

3. Whole Genome Sequencing (WGS)

   a) 100%

   b) 60–100 GB per sample for reads in FastQ format

   c) Another 60–100 GB per sample for aligned reads in BAM format

   d) <1 GB for variants in VCF format, QC report, logs, etc.

Resource requirements differ per type of job:

- Walltime ranges from 1 minute to 64 hours

- Nodes is always 1 with cores per node ranging from 1 to 21

- Memory ranges from 1 GB to 32 GB

The smallest amount of resources is consumed by the jobs that produce the final QC reports: 1 minute walltime, 1 core and 1 GB RAM. (These jobs may even complete with less resources, but these are the smallest numbers we can specify in our current job scheduling system.)

Implementation of this pipeline consists of many applications developed by various research institutes / projects and written in various languages. The heterogeneity of the software stack results in heterogeneous resource requirements for different steps of the analysis. In addition the heterogenity can make it challenging to get all dependencies installed, which makes automated deploy management a must. The core apps include:

- FastQC[11] for quality control of FastQ files (written in Java)
- BWA[12] for alignment (written in C)
- Sambamba[13] for BAM file processing (written in D)
- GATK[14] "best practices" for variant calling (written in Java)
- R[15] for statistics (written in R, C, C++ and Fortran)
- Molgenis Compute[16] for job orchestration (written in Java)
- Scripts and "glue" written in various scripting languages like Perl, Python, Bash, Ruby, Lua, etc.

Full description of all the steps can be found at `http://molgenis.github.io/pipelines/ngs-protocols`. A summary overview is visible in Figure 2.

The BBMRI.cz has similar workflows to the ones described above for the BBMRI.nl, with the following computational and storage capacities used: input 60–100 GB for a single run, .fq.gz (gzipped compressed FastQ standard text files that contain reads); output 60–100 GB of BAM files (binary alignment format standard) and <1 GB VCF genotype file (variant calling format) and additionally small script and log files. Depending on the size of input data, the BBMRI.cz has various capacity needs: 1–20 CPU cores, RAM memory ranges from 1 GB to 32 GB, walltime ranges from 1 minute to 64 hours.

[11] `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`

[12] `http://bio-bwa.sourceforge.net/`

[13] http://lomereiter.github.io/sambamba/

[14] `https://software.broadinstitute.org/gatk/`

[15] `https://www.r-project.org/`

[16] `http://molgenis.github.io/software/compute`

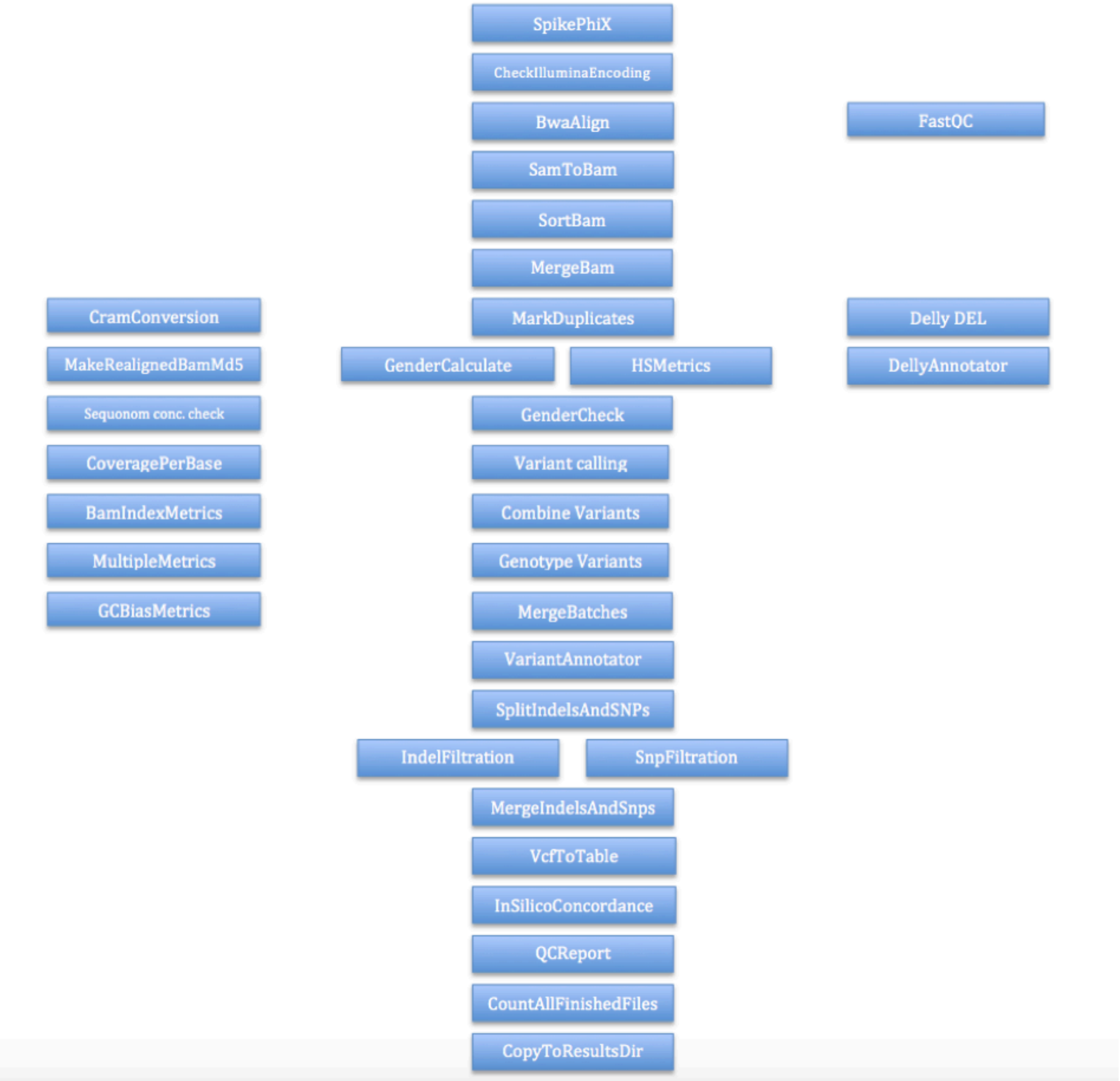

Figure 2: Summary of the NGS_DNA pipeline. Dataflow from top to bottom; parallel tasks are shown next to each other. At various steps the analysis is run split in smaller parallel jobs (not shown), e.g., alignment is parallelised per NGS machine run, variant calling is run for each of the 23 chromosomes separately. An analysis can consist of up to hundreds of samples resulting in a sizeable analysis task.

4 Evaluation of information security considerations for diagnostics

The cloud provider and the intermediary who implements the pipelines must demonstrate that the facility appropriately addresses information integrity and security considerations. This is because many biobank data, such as NGS results, are classified as ‘high risk’ with regard to data confidentiality due to privacy concerns. When NGS data is used in a diagnostics setting the risks with regard to data integrity as well as data availability are also high; When data is required for treatment, it cannot arrive late or be wrong, because a system failed. In case of research a lower availability may be acceptable.

Many of these requirements are defined in international standard such as ISO/27002:2013 [3] (H5-15), while national regulation might apply as described by BBMRI-ERIC in the report “Security and Privacy Architecture of BBMRI-ERIC IT” [4]. Below we summarize the most important constraints for which we will evaluate the cloud pilot implementations. For this pilot, we assume that permanent data storage with backups is addressed by the biobank/health care provider and that we only need address information security for the temporary storage used during the analysis in the cloud. Some of these requirements will need to be demonstrated by the cloud provider and some by the pipeline provider/operator. Below we summarize examples of the issues on which we will need to report and evaluate measures.

4.1 General

A Risk Assessment (RA) has been performed and regularly updated. This RA consists of the following four phases:

- Phase 1: Business Impact Analysis (BIA) to determine risk with regard to data integrity, data availability and data confidentiality (none, low, medium, high)
- Phase 2: Decide which management measures from (ISO) standards are relevant based on the risk as determined in the BIA.
- Phase 3: A self assessment is made where the infrastructure describes how the relevant management measures have been implemented. There will always be a “Risidual Risk” left. This must be acceptable to the client/customer that wants to use the infrastructure. If the Residual Risk is deemed unacceptable it’s either back to the drawing board to redesign part of the infrastructure and update the self assessment or the infrastructure cannot be used.
- Phase 4: Regular audits to check whether the management measures as described in the self assessment are implemented “as advertised”

4.2 Demonstrate measures to ensure data integrity

When NGS data is used in a professional research or diagnostic setting there is high risks with regard to data integrity, e.g., in light of reproducibility of the analyses or errors that impact patient lives. Before a cloud can be used for these use cases the security officer of the user institutes may require measures such as:

- Check for corruption or manipulation (e.g., compute and verify md5 checksums) after data is moved/copied.
- Ensure correct input of the pipelines, for instance by using input validation of files and parameters.
- Separation of development, acceptance and production environments.
- Validation of analysis pipeline after development, automated testing and verification when a previously validated pipeline is deployed in the acceptance or production environment.
- Version management of the software (e.g. Git, Mercurial) and of the deployment configurations (e.g. Ansible, Puppet) to ensure that exactly the same setup as previously validated on another infra site can be deployed on new site. In addition this allows for rollback in case a setup fails validation tests after updates/upgrades.
- SOPs and checklists up to diagnostics standards.
- Minimize the contact interfaces to other systems (in particular those having lower information security levels).

4.3 Demonstrate measures to ensure data confidentiality

Many biobank data, such as NGS results, and also diagnostic data are classified as ‘high risk’ with regard to data confidentiality due to privacy concerns. Before a public cloud can be used for these use cases the security officer of the user institutes may require measures such as:

- Isolation of users and groups so that data cannot be shared accidentally.
- Chain of trust, federated identity and authorization (e.g., OpenConext + COmanage[17]) to ensure that when employees change function or leave the institute, their access is revoked within a reasonable time frame.
- Secure upload/download data transfer channel.

[17] http://www.internet2.edu/products-services/trust-identity/comanage/

- Access to a patient's data limited to employees involved in the treatment of that patient. When part of the work is outsourced to a third party (e.g., maintenance of hardware) this third party must have an adequate 'data processing agreement'.

- Ensure that no identifiable data is left behind after a certain amount of time to comply with data destruction laws/guidelines and also when the contract is terminated.

- Ensure that physical access to hardware / data centres is logged and sufficiently protected by means of keys. Access must be restricted to a known set of individuals for which legal IDs (e.g., passports) were verified.

- Data is stored within the EU, cannot leave the EU and any company involved operates exclusively under EU jurisdiction (e.g., company stock not listed on New York Stock Exchange.)

- Procedure for reporting data 'leakage' incidents.

- Firewalls, stealth check, logging of attempts of unauthorized access.

- Minimizing the impact of unauthorized access, e.g., by minimizing the volume and ease of abuse of the data by only keeping fractions that in itself are not enough to derive identification or abuse whenever possible.

- Installation of new software via change management (possibly including virus scanners).

- Ensure inactive users are disconnected to prevent unauthorized access from client machines.

- Procedure to safely dispose of hardware at the end-of-life (e.g., use self-encrypting hard drives or put hard disks into a shredder).

4.4 Demonstrate measures to ensure reliable operations/runtimes

In case of diagnostics, also the operational risks are classified as 'high'—it would be unacceptable to excessively delay or lose a result critical for a diagnosis because of a system failure. Before a cloud can be used for these use cases the security officer of the user institutes may require measures such as:

- Redundancy by means of fall-back sites. Automatic failover is not necessary. There should be no single points of failure; Hence infra running at different sites should be completely independent. This includes not only the hardware but also the staff managing the sites. Too little staff or staff managing multiple sites from a single location may result in a single point of failure.

- Redundancy for components at a certain site to reduce the chance of a site going down. E.g. redundant power supplies.

- Ensure quality of server hosting. For instance, climate control, emergency power to survive outages, etc.

- Quota for resources like storage, cores and RAM to prevent exhausting them.

- Change management ⟹ validation/verification or support contracts for the software used. Configuration management to guarantee reproducible analysis environments so that these do not lead to accidental changes that may impact diagnostic outcome.

- Inventory management to ensure services are not running on the wrong or deprecated machines.

- Monitoring of system availability and auditability. This includes online status (e.g., Nagios,[18] node health check), logging, proving immutability of these logs and enabling filtering of these logs to derive operational insight as well as being able to investigate the audit trail after an incident has occurred.

- Backup of pipeline software, configuration and data.

- Data loss to a maximum of 24 hours.

- Predictable maintenance windows.

- Contract possible within the cloud provider and the cloud consumer covering all of above.

[18] https://www.nagios.org/

5 Conclusions

In this Deliverable we have tried to select both workflows that are typical for the biobanks aiming for the private clouds (using FedCloud as a technology to deploy inside private clouds), as well as workflows we will attempt at using the EGI infrastructure (using FedCloud as an infrastructure to get access to public cloud from the biobank perspective). Particularly for the private clouds, we need to look into supporting the Section 4 and look also to the extent we can apply them to the public infrastructure.

References

[1] E. W. Deutsch. “File formats commonly used in mass spectrometry proteomics”. In: *Molecular & Cellular Proteomics* 11.12 (2012), pp. 1612–1621.

[2] P. Aiyetan, B. Zhang, L. Chen, Z. Zhang, and H. Zhang. “M2Lite: An open-source, light-weight, pluggable and fast proteome discoverer MSF to mzIdentML tool”. In: *Journal of bioinformatics* 1.2 (2014), p. 40.

[3] *Information technology – Security techniques – Code of practice for information security controls*. ISO 27002:2013. Oct. 2013.

[4] P. Holub and C. S. IT. *Security and Privacy Architecture*. Sept. 2016. DOI: 10.5281/zenodo.159444. URL: https://doi.org/10.5281/zenodo.159444.

Acknowledgments

Many members of BBMRI have contributed to the knowledge accumulated in this report, including Jeroen Laros, Peter van het Hof, Leon Mei, Roan Kanninga, Gerben van der Vries, Wim Nap, Fokke Dijkstra, Bob Dröge, Dalibor Valík, Petra Dvořáková, Jakub Faktor. We apologize to those we failed to name.