EGI-Engage

Analysis of requirements on biobank and study workflows

D6.8

Date	19 October 2016
Activity	SA2
Lead Partner	BBMRI-ERIC
Document Status	FINAL
Document Link	https://documents.egi.eu/document/XXXX

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	2016-02-26
Reviewed by			
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		Final version	

Contents

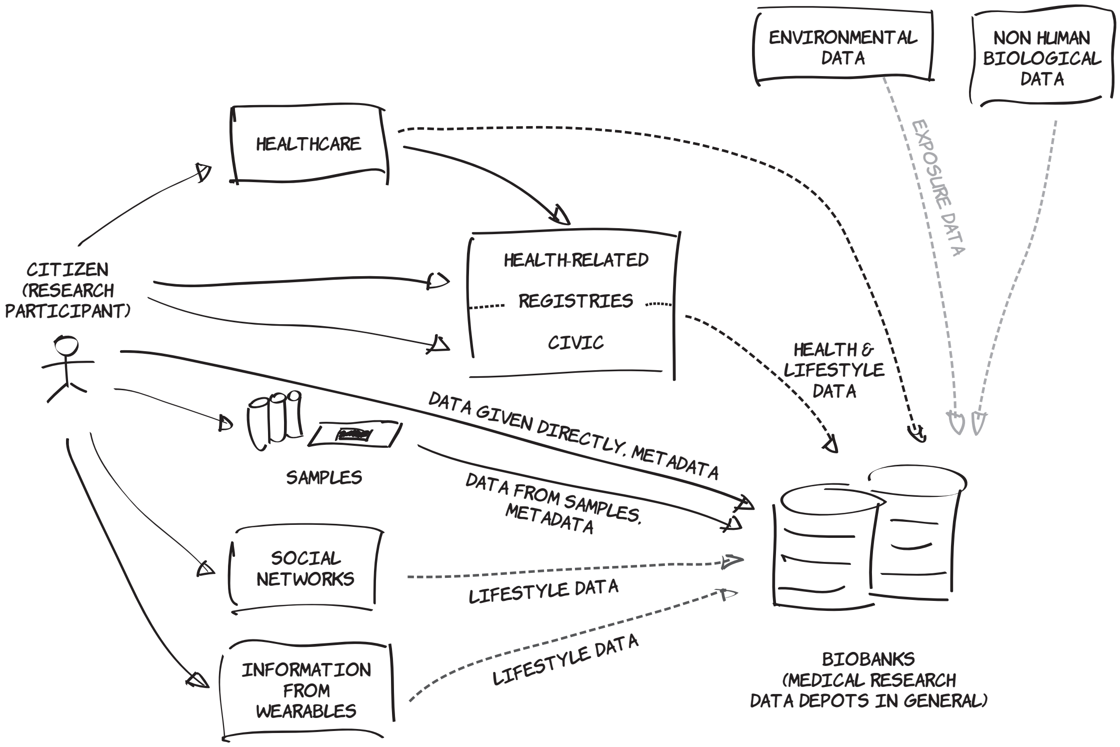
ENVIRONMENTAL DATA
NON HUMAN BIOLOGICAL DATA
EXPOSURE DATA
HEALTHCARE
HEALTH-RELATED
REGISTRIES
CIVIC
CITIZEN (RESEARCH PARTICIPANT)
HEALTH & LIFESTYLE DATA
DATA GIVEN DIRECTLY, METADATA
SAMPLES
DATA FROM SAMPLES, METADATA
SOCIAL NETWORKS
LIFESTYLE DATA
INFORMATION FROM WEARABLES
LIFESTYLE DATA
BIOBANKS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

Figure 1: Flow of data into the biobanks.

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 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 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 of dealing with human biological material and data, BBMRI-ERIC biobanks need to pay particular attention to aspects of data protection, as privacy protection of research participants (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 by 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 for improvement of 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 worflows and explore their applicability to the cloud scenarios.

This deliverable provides an overview of the workflows, that we 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 strike good representation 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 - this is given both by the participating national nodes (we are aware that other national nodes are running also other types of processing, e.g., large scale image analysis in BBMRI.it and BBMRI.nl) and constrained work capacity available in the BBMRI CC.

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 analysis of the proteins in the 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 the elastic compute capacity that can scale up based on the volume of analyses performed.[1] If third party cloud infrastructures are used, they can calculate the data extremely fast and the network transmissions of data may become the bottleneck and the improvement would not be significant; hence either on-site clouds are needed or very fast network links to the cloud infrastructures provided by third parties.

Data are not sensitive from perspective of patient data privacy: the output from mass spectrometry of from further analysis itself does not contain any information that could point to 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 considered as a “book” that consists of 23 chapters (chromosomes) containing total of ∼30.000 words (regions of the DNA we call genes). As a whole, the book contains DNA is transcribed (copied) to RNA molecules that in the end result in the proteins that make-up much of the human body. 3 billion building blocks (base pairs) from an alphabet that consists of four letters (ACTG). Of this 3 billion only <1% matches to known genes; the 99% of letters between the genes we think have a ‘regulatory’ function that influences which genes are switched ‘on’ or ‘off’ but there is much to learn there. With regulation we thus mean: how many RNA copies are made from the DNA during lifecycle of the cells in the human body.

On DNA level, NGS enables to measure multiple or even all genes at once (also called ‘exome’) or even to measure whole of the DNA (whole genome). These enable fine grained characterisation of all DNA variation points between individuals, i.e., differences in ‘genotype’. This is a basis for studies that in case of traits that are heritable to pinpoint what DNA differences are relevant for predicting why some individuals get sick and other stay healthy and as a basis to diagnose disease in ‘rare disease’ where one DNA mutation can be enough to get ill. On 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 on RNA level. Much of the focus of this research is now on ‘personalized medicine and health’, that is, for example on how can we predict what (expensive) medicine will be most effective while having no (lethal) side effects in what groups of patients when considering these DNA and RNA profiles.

To analyse these NGS data much computational analysis is needed which are typically called ‘pipelines’. First the data need to be made ready for analysis (what is called ‘processing’) because the NGS machines don’t produce complete sequences of whole chromosomes but instead measure small fractions of 100–200 characters. Therefore analysis starts with large text files (GBs) that contains thousands of lines (‘reads’), i.e., short text sequences of CATG. The first processing is to reconstruct these fragments into complete chromosomes, which is typically done by alignment the fragments to the known human ‘reference’ genome. Subsequently, the differences are assessed between the experimentally derived sequence and the reference by variant calling, resulting in a listing of the differences which are on average 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, also the number of reads is quantified as a proxy for quantifying ‘gene expression’, that is 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 all above 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 we have 2 research clusters and 2 diagnostic clusters next to BBMRI.nl central clusters where it is essential that data and pipelines can be moved or replicated between them 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 integrated in order to reach sufficient statistical power. Typically, such large analysis are implemented as a large multi-center (and often even multi-national) consortium where researchers from many institutes need to collaborate around the data, requiring central access to data and analysis procedures. Such large scale facilities are beyond what individual institutes can provision therefore there is a demand for 'cloud' solutions (IaaS, SaaS) that enable research consortia to have a 'digital research environment for human subject research data' to conduct their analysis with suitable facilities. Therefore there is a large demand from BBMRI-ERIC Members for scale-out facilities.

New research methods such as NGS and 'personalized medicine' are also rapidly uptaken in the context of healthcare. In addition the speed in 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 requested to also make analysis pipelines available to diagnostics labs, e.g., diagnostics up to a high standard in terms of operation (Standard Operating Procedures = SOPs) and privacy constraints. However, next to the requirement of pipeline portability as mentioned above, 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). However, hospitals also access to suitable storage and compute facility including fail-over scenarios while considering ethical, legal and societal considerations such as privacy which are much harder.

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 are planned to be implemented in the BBMRI Competence Centre of EGI-Engage. We aim at two types of the workflows:

- workflows that are not including particularly 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 provide it to the biobanks as logically private cloud infrastructure).

First part covers several different proteomics worflows, which are main focus of 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 medical research of 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. Data formats used in the proteomics applications as 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 largest analysis step before biobank data analysis can commence. This pipeline is an example for many more pipelines and when this pilot is succesfull 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 property of aformentioned pipelines) and information security (which is essential before sensitive data from biobanks and hospitals will be allowed to sent to cloud providers).

3.1 Workflow: Identification of protein

The objective of protein identification is identify unknown protein (or probability of protein identification). Mass spectrometry is able to measure the mass of charged peptides, i.e., protein consists of peptides - chains of amino acid monomers, in unknown sample. Measured mass of peptide (raw outcome from mass spectrometry in .raw format) is used as the input to an analytic SW (Proteome discoverer) which searches database of known proteins and matches the hit. Search engine compares the spectrum of product ion masses originating from a peptide is compared with database. The result is list of proteins and probability of hit.

- GOAL: identify unknown protein
- INPUT:

  - .raw file (app 1 GB), mass of found peptides

- PROCESS:

  - Find the best probable hit in database of known proteins (free downloadable, app 300 MB, text file: name and sequence, .fasta)
  - App. 1–6 hours on an office PC (depends on .fasta database size and on parameters)
  - Used SW - Proteome Discoverer[2]

- OUTPUT:

  - .msf
  - List of proteins which match the unknown protein the best and 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 sample. The obtained information can be further used to understand cancer behaviour by comparing the level of protein in healthy tissue with the tumor because the change of protein level might be an important disease biomarker.

There are two basic division of quantitation: label or label-free. In label-free method all samples are measured separately and then analyzed so that more files are produced and the method is more time consuming. In label method the isotopes are used to mark sample. Two or more samples are mixed together and put into mass spectrometry device. Due to isotopes the device can distinguish among peptides from different sample and it can measure all of them in one single run. 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 could be used in comparison of proteins in tumor and in healthy tissue.

- INPUT:

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

- PROCESS:

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

- Find the best probable hit in database of known proteins (free downloadable, app 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)
- App. 1 day on office PC
- Used SW:
  * Label free method: Protein Pilot[3] (identification), PeakView[4] (quantitation), MarkerView[5] (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 (app. 0.5 GB)

- PROCESS:
  - Computationally intensive (?)
  - Used SW: HDExaminer

- OUTPUT:
  - 5–15 GB of output data per experiment (size depends on experiment extent)

- PARALLELISM

[3] http://sciex.com/products/software/proteinpilot-software

[4] http://sciex.com/products/software/peakview-software

[5] http://sciex.com/products/software/markerview-software

  - at least on the level of individual analyses

3.4 Workflow: DNA alignment and genotyping

MMCI summary: 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 you have various capacity needs: 1–20 CPU cores, RAM memory ranges from 1GB to 32 GB, walltime ranges from 1 minute to 64 hours. Ideally we would want to evaluate small (diagnostics) and large (research) analysis using this pipeline in the cloud.

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. Essential is 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[6] @ compile / installation time and Lmod[7] @ 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 (for example images with colored dots from fluorescent probes) into sequence reads (strings of nucleotides A, T, C or G and associated quality scores in a text based format) per sample. De facto standard file format for sequence reads is gzip compressed FastQ (*.fq.gz).
2. Alignment of sequence reads to a reference genome. De facto standard file format for aligned reads: Binary SAM (*.bam)

[6] `https://github.com/hpcugent/easybuild`

[7] `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, but 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 many and small files, which makes them inefficient to transfer over networks or store on large parallel storage systems). Once the data is converted to 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) $\ll$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 1GB to 32 GB

Implementation of this pipeline consists of various applications developed by various research institutes / projects and written in various languages. The core apps include:

- FastQC[8] for quality control of FastQ files (written in Java)
- BWA[9] for alignment (written in C)
- Sambamba[10] for BAM file processing (written in D)
- GATK[11] "best practices" for variant calling (written in Java)
- R[12] for statistics (written in R, C, C++ and Fortran)
- Molgenis Compute[13] for job orchestration (written in Java)
- Various 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.

[8] `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`
[9] `http://bio-bwa.sourceforge.net/`
[10] http://lomereiter.github.io/sambamba/
[11] `https://software.broadinstitute.org/gatk/`
[12] `https://www.r-project.org/`
[13] `http://molgenis.github.io/software/compute`

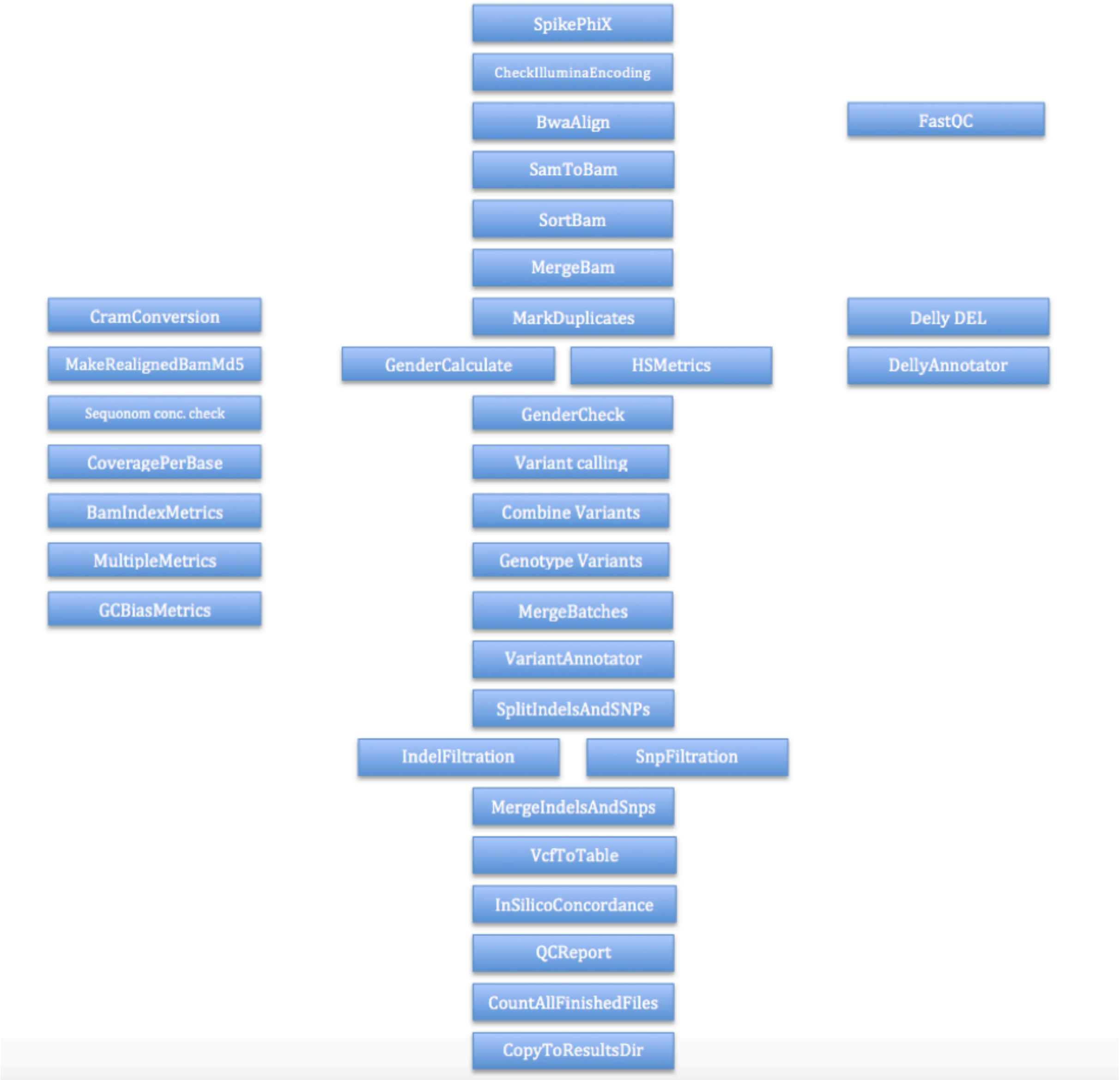

Figure 2: Summary of the NGS_DNA pipeline. Dataflow from top to bottom; parallel tasks are shown next to each other. At various steps the analysis is run split in smaller parallel jobs (not shown), e.g., alignment is parallelised per NGS machine run, variant calling is run for each of the 23 chromosomes seperately. 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 + the intermediary who implements the pipelines must demonstrate that the facility sufficiently 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 the risks with regard to data integrity as well as data availability are high. In case of diagnostics, also the operational risks are classified as 'high', i.e., it cannot be that diagnostics is too late because the 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 [2] (H5-15), while national regulation might apply as described by BBMRI-ERIC in the report "Security and Privacy Architecture of BBMRI-ERIC IT" [3]. Below we will 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 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, e.g., by using input validation of files and parameters.

- Separated of development, acceptance and production environments.

- Version management of the software (github, releases) and the deploy configurations (e.g., Ansible playbooks) to ensure.

- Validation of analysis pipeline after development, automated testing and verification when a previously validated pipeline is deployed in the acceptance or production environment.

- SOPs and checklists up to diagnostics standards.

- Minimal the contact interfaces to other systems (in particular those having lower information security levels).

4.2 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 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 such data data cannot be shared incidentally
- Chain of trust, federated identity and authorization (e.g., OpenConext + COmanage[14]) to ensure that when employees change function or leave the institute that their access is revoked within a reasonable time frame.
- Secure upload/download data transfer channel
- 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 end of life hardware (e.g., use self encrypting hard drives or put harddisks in shredder).

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

4.3 Demonstrate measures to ensure reliable operations/runtimes

In case of diagnostics, also the operational risks are classified as 'high', i.e., it cannot be that diagnostics is too late because the system failed. 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.
- Redundant power supply.
- Ensure quality of server hosting, e.g., , climate control, emergency power to survive outages, etc.
- Quota.
- Change management $\Longrightarrow$ validation/verification or support contracts for the software used. Configuration management to guarantee reproducible analysis environments such that these don't lead to accidental changes that may impact diagnostic outcome.
- Inventory management to ensure services are not running on the wrong or deprecated machines.
- Monitoring on availability and traceability, e.g., online status (e.g., Nagios,[15] node health check), logging plus proving immutability of these logs and enabling filtering of these logs to derive knowledge.
- Restart of a running pipeline because of long runtimes.
- 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.

[15] `https://www.nagios.org/`

References

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

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

[3] 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.