European
Genomic Data
Infrastructure

GDI Milestone MS26

Initial set of questions by the use cases

Table of Contents

1. Introduction 1

2. Initial set of questions 2

2.1. Genome of Europe demonstrator 2

2.2. Building on the experiences of the 1+MG/B1MG use-cases 5

2.3. Computational approaches for infectious diseases with an initial focus in COVID-19 7

2.4. Data-driven models for Cancer Research 9

1. Introduction

Milestone MS26 includes a set of initial questions proposed by the GDI use-cases, serving as catalysts to propel the efforts within Pillar II and III. The proposed questions will undergo validation and mapping to align with the technological requirements of GDI. This iterative process will involve collaboration between WP7 and Pillar II, ensuring comprehensive mappings, and then the work can scale it up to Pillar II and Pillar III conversations. In the case where these defined requirements already exist, they will be forwarded to Pillar II. Conversely, if a requirement is novel, it will be directed to WP8.

The following tables include an initial mapping and description of the technologies that would be needed to solve the questions, to serve as a starting point on the connections between WP7 and the rest of the GDI WPs.

2. Initial set of questions

2.1. Genome of Europe demonstrator

The following questions from the Genome of Europe use-case build-up of each other. So, it is expected to start with the first one and then continue with the rest of them.

Genome of Europe **Question 1:** Lookup of individual genetic variants	
Data Reception (Data preparation & inclusion)	Reference genome (e.g., hg38) , common file format (e.g., vcf/idx). Simple metadata (e.g., sex, age, ancestry)
Storage and Interfaces (Data Storage & Management)	Access to individual variant information. Aggregate carriership on group level.
Data Discovery	Beacon & Central Catalog
Data Access Management Tools	Life Sciences AAI & REMS (to be discussed with Pillar II)
Data Processing (Data use)	Query based on position, summarize on group level.
Do you need to access data transnationally? Are you considering a federated system? Or something else?	Yes, across European countries. Individual participant data should stay within country
Are you already engaged with projects/efforts/initiatives beyond 1+MG/B1MG/GDI that are developing infrastructures for tackling your needs?	No, although many countries have (part of) their own infrastructure.
Are you aware or any other use case requirements relevant for your use case?	Needs to scale/update as more data is added by the GoE partners. GoE needs to take into account future sharing with EHDS
Are you using structured pheno-clinical data? If yes, which one; if no, do you need/plan to use it?	Not at first, age, sex and ancestry is enough. Sex and ancestry can be derived from genetic data. We can do without age if really needed. Later I hope to add phenotypes and/or clinical data.

Genome of Europe

Question 2:

Recalibration of polygenic risk scores

Data Reception (Data preparation & inclusion)	No additional preparation requirements
Storage and Interfaces (Data Storage & Management)	Need to access multiple variants at once (i.e., enough variants that form unique carriership combination per person)
Data Discovery	In addition to which datasets exist, perhaps we need to know how many of the variants in the PRS exist within each dataset.
Data Access Management Tools	The same? From an application perspective I have no preference, not sure about others.
Data Processing (Data use)	Collect a list of variants per person. Multiple carrier status (0, 1 or 2) for each variant with a predefined weight. Return the sum of weights (PRS). Need to aggregate/show distribution at group level (sex, ancestry).
Do you need to access data transnationally? Are you considering a federated system? Or something else?	Same as Q1
Are you already engaged with projects/efforts/initiatives beyond 1+MG/B1MG/GDI that are developing infrastructures for tackling your needs?	Same as Q1, though I think it's unlikely individual partners have already built something like this. (in contrast to Q1, where some solutions probably exist)
Are you aware or any other use case requirements relevant for your use case?	The PRS really needs to be defined within ancestries. If that info is not available, we need to learn it from the genetic data (can be on the fly, removed after)
Are you using structured pheno-clinical data? If yes, which one; if no, do you need/plan to use it?	Not at first, but ultimately we want to get distribution in cases/controls. (e.g., breast cancer PRS distribution in men, women, cases and controls, other cancer cases, and per ancestry). This is for the future, in the first parts of GoE, having it by ancestry or country of origin is already huge.

Genome of Europe

Question 3:

Ancestry-specific imputation

Data Reception (Data preparation & inclusion)	GoE data should be turned into reference panels, including learning correlation between genetic variants.
Storage and Interfaces (Data Storage & Management)	Need simultaneous access to multiple samples at once, to learn correlation structure.
Data Discovery	Need to know of samples from same ancestry exist/can be combined across (national) datasets
Data Access Management Tools	The panels would be used by other researchers to "impute" data. They would need access to the aggregated results for that.
Data Processing (Data use)	For details on the steps we need to reach out to experts (Elly Zeggini). In short, take a piece of DNA (few Mb at the time) and correlate all combinations of variants in that area. Then reduce to the smallest number of variants needed to measure and still later guess/impute all variants in the block. Here I would focus on part 1: creating the reference panels.
Do you need to access data transnationally? Are you considering a federated system? Or something else?	Yes, currently a group in Germany holds an imputation server that could perhaps be used as a central hub. tbd.
Are you already engaged with projects/efforts/initiatives beyond 1+MG/B1MG/GDI that are developing infrastructures for tackling your needs?	I think there exist grants/proposals to create the data processing part, but not the underlying infrastructure.
Are you aware or any other use case requirements relevant for your use case?	In part 2: access to aggregated reference panels for imputation of other genetic datasets.
Are you using structured pheno-clinical data? If yes, which one; if no, do you need/plan to use it?	Apart from ancestry, I do not think other phenotypes are needed for this question.

2.2. Building on the experiences of the 1+MG/B1MG use-cases

1+MG/B1MG Use-cases

Question 1:

Why do people with certain disease-specific genes not develop the disease?

Data Reception (Data preparation & inclusion)	**Phenotype data:** 1. medical history of subjects (diseases, lab analyses) 2. subject lifestyle (smoking, alcohol use, obesity etc) 3. subject measurements (weight, height, strength, blood pressure etc) **All called genetic variants:** 1. single nucleotide variants (SNVs); 2. copy number variants (CNVs); 3. phased genotype; 4. CNV data;
Storage and Interfaces (Data Storage & Management)	History of searches and results
Data Discovery	List of available data
Data Access Management Tools	- Simple, user-friendly and low effort process to apply data access. - Access documentation and communication history in one place (CRM, Pipedrive)
Data Processing (Data use)	- Quality metadata (describing sample data collection and quality, what tech/methods were used for WGS, quality references) - Privacy technologies (anonymization, encryption, aggregation) - Distributing query algorithms between federated nodes. - Receiving algorithm results.
Do you need to access data transnationally? Are you considering a federated system? Or something else?	Yes, the bigger the cohort the better, federated analysis would be suitable.
Are you already engaged with projects/efforts/initiatives beyond 1+MG/B1MG/GDI that are developing infrastructures for tackling your needs?	BBMRI-ERIC
Are you aware or any other use case requirements relevant for your use case?	No
Are you using structured pheno-clinical data? If yes, which one; if no, do you need/plan to use it?	OMOP for clinical data

1+MG/B1MG Use-cases

Question 2:

Why do some gene variants cause adverse side effects for medications?

Data Reception (Data preparation & inclusion)	**Phenotype data:** 1. Particular medication usage history (exact medication and compounds - ATC codes, dosages) 2. subject lifestyle (smoking, alcohol use, obesity etc) 3. subject measurements (weight, height, strength, blood pressure etc) **All called genetic variants:** 1. single nucleotide variants (SNVs); 2. copy number variants (CNVs); 3. phased genotype; 4. CNV data;
Storage and Interfaces (Data Storage & Management)	History of searches and results
Data Discovery	List of available data
Data Access Management Tools	- Simple, user friendly and low effort process to apply data access. - Access documentation and communication history in one place (CRM, Pipedrive)
Data Processing (Data use)	- Quality metadata (describing sample data collection and quality, what tech/methods were used for WGS, quality references) - Privacy technologies (anonymization, encryption, aggregation) - Distributing query algorithms between federated nodes. - Receiving algorithm results.
Do you need to access data transnationally? Are you considering a federated system? Or something else?	Yes, the bigger the cohort the better, federated analysis would be suitable.
Are you already engaged with projects/efforts/initiatives beyond 1+MG/B1MG/GDI that are developing infrastructures for tackling your needs?	BBMRI-ERIC
Are you aware or any other use case requirements relevant for your use case?	No
Are you using structured pheno-clinical data? If yes, which one; if no, do you need/plan to use it?	OMOP for clinical data

2.3. Computational approaches for infectious diseases with an initial focus in COVID-19

Infectious diseases

Question 1:

GWAS (validation): risk variants of severe COVID-19 (research)

Data Reception (Data preparation & inclusion)	In 1+MG, this use case requires synthetic data to get started. A minimal metadata structure has been developed. We are in discussions with WGs to arrange the generation of the data. Synthetic data are used because of the absence of suitable, easy-to-access real-life data. Synthetic data should allow us to simulate at least the data storage and analysis options. We assume at this point that we may need some 50000 samples at least.
Storage and Interfaces (Data Storage & Management)	Can and should we simulate data being sensitive in nature? If so, how?
Data Discovery	What system (nomenclature) should we adhere to to make the data files findable?
Data Access Management Tools	We should implement a data access committee.
Data Processing (Data use)	Data generation / QC, TBD > WP8
Do you need to access data transnationally? Are you considering a federated system? Or something else?	Distributed storage of the synthetic data across countries will be necessary to reflect the real-life scenario. For more reliable results, GWAS requires large datasets. Sometimes datasets need to be merged to obtain larger ones. Analysis pipelines should be able to merge datasets from different countries. Are there options in place already that allow GWAS analysis?
Are you already engaged with projects/efforts/initiatives beyond 1+MG/B1MG/GDI that are developing infrastructures for tackling your needs?	
Are you aware or any other use case requirements relevant for your use case?	
Are you using structured pheno-clinical data? If yes, which one; if no, do you need/plan to use it?	

Infectious diseases

Question 2:

Variants that may guide prognosis and/or treatment (healthcare)

Data Reception (Data preparation & inclusion)	In 1+MG, this use case requires synthetic data to get started. A minimal metadata structure has been developed. We are in discussions with WGs to arrange the generation of the data. Synthetic data are used because of the absence of suitable, easy-to-access real-life data. Synthetic data should allow us to simulate at least the data storage and analysis options. For this use case, summary results accompanied by some healthcare information (e.g. treatment) will be the outcome of the analysis.
Storage and Interfaces (Data Storage & Management)	At this early stage, can and should we simulate data being sensitive in nature? If so, how?
Data Discovery	What system (nomenclature) should we adhere to to make the data files findable?
Data Access Management Tools	We should implement a data access committee.
Data Processing (Data use)	Data generation / QC, TBD > WP8
Do you need to access data transnationally? Are you considering a federated system? Or something else?	Distributed storage of the synthetic data across countries will be necessary to reflect the real-life scenario. In this healthcare use-case, a physician needs access to as many datasets as possible, to obtain the most reliable information. It would be tedious to have to check one dataset after the other by country. Therefore, merged datasets would be beneficial.
Are you already engaged with projects/efforts/initiatives beyond 1+MG/B1MG/GDI that are developing infrastructures for tackling your needs?	
Are you aware or any other use case requirements relevant for your use case?	Not right now.
Are you using structured pheno-clinical data? If yes, which one; if no, do you need/plan to use it?	Yes, snomed.

2.4. Data-driven models for Cancer Research

Cancer research

Question 1:

We were inspired by a clinical case report [https://doi.org/10.1002/onco.13979] : A patient diagnosed with NSCLC ROS1 fusion-positive cancer (T0) was treated with standard of care TKI Crizotinib (T1). Remission after T2 months. Recurrence after T3 months and NGS sequencing. Can the tumor relapse be explained by the NGS molecular profile at recurrence? In other words, can we match the mutations found at T3 against a database with known variants responsible for tumor regrowth or therapy resistance (e.g. BRAF V600E). Is there a second line of treatment against this second mutational hit?

Data Reception (Data preparation & inclusion)	To simulate T3 sequencing, we generated in-silico genomic data starting from GIAB NA12878 'germline' DNA. Random mutations were simulated (BamSurgeon) and added including BRAF and other common cancer variants at variable VAF (got selected VCFs from ICGC of NSCLC BRAF V600 positive cases). We developed and populated the minimal data set for cancer according to the case report.
Storage and Interfaces (Data Storage & Management)	We need simulated data because of the restriction of usage of sensitive data. More simulations are needed to cover other cases and cohort studies.
Data Discovery	The query must support genomic coordinates (intervals for structural variants), gene fusions terms. Access to treatment and follow ups of the matched cases would be necessary to help make a decision. Query filters must include tumor type, tissue, site, stage etc.
Data Access Management Tools	
Data Processing (Data use)	A tool to visualize aggregated data (cBioportal-like). For example we can ask which are the most common co-occurring (or exclusive) mutations of BRAF V600E. Other external sources like approved therapies and ongoing trials can be linked to hosted data.
Do you need to access data transnationally? Are you considering a federated system? Or something else?	A comprehensive database (transnational) is needed.
Are you already engaged with projects/efforts/initiatives beyond 1+MG/B1MG/GDI that are developing infrastructures for tackling your needs?	We do. Alliance Against Cancer, Health Big Data project, DIGICORE.
Are you aware or any other use case requirements relevant for your use case?	
Are you using structured pheno-clinical data? If yes, which one; if no, do you need/plan to use it?	The minimal dataset for cancer uses clinical and phenotypic structured terms (ICGC-ARGO data dictionary, OSIRIS and mCODE)

Cancer research

Question 2:

We were inspired by a scientific report [https://doi.org/10.1200/po.16.00054] to describe a possible real word case: A patient diagnosed with metastatic melanoma. Molecular standard analysis revealed a BRAF V600E point mutation (T0). The patient was treated with standard of care BRAF inhibitor Vemurafenib (T1). Remission complete after T2 months. Recurrence after T3 months and NGS sequencing. Can the tumor relapse be explained by the NGS molecular profile at recurrence? In other words, can we match the mutations found at T3 against a database with known variants responsible for tumor regrowth or therapy resistance (e.g. PTEN *c.493_634del142*). Is there a second line of treatment against this second mutational hit?

Data Reception (Data preparation & inclusion)	To recreate T3 sequencing, we took advantage of a patient-derived metastatic melanoma cell-line (COLO-829), comprehensively characterized by whole-genome sequencing, together with its normal counterpart. We developed and populated the minimal data set for cancer according to the case report.
Storage and Interfaces (Data Storage & Management)	We need cell-line data because of the restriction of usage of sensitive data.
Data Discovery	The query must support genomic coordinates (intervals for structural variants), gene fusions terms. Access to treatment and follow ups of the matched cases would be necessary to help make a decision. Query filters must include tumor type, tissue, site, stage etc.
Data Access Management Tools	Life Sciences AAI & REMS
Data Processing (Data use)	A tool to visualize aggregated data (cBioportal-like). For example we can ask which are the most common co-occurring (or exclusive) mutations of BRAF V600E. Other external sources like approved therapies and ongoing trials can be linked to hosted data.
Do you need to access data transnationally? Are you considering a federated system? Or something else?	A comprehensive database (transnational) is needed.
Are you already engaged with projects/efforts/initiatives beyond 1+MG/B1MG/GDI that are developing infrastructures for tackling your needs?	We do. Alliance Against Cancer, Health Big Data project, DIGICORE.
Are you aware or any other use case requirements relevant for your use case?	
Are you using structured pheno-clinical data? If yes, which one; if no, do you need/plan to use it?	The minimal dataset for cancer uses clinical and phenotypic structured terms (ICGC-ARGO data dictionary, OSIRIS and mCODE)

Cancer research

Question 3:

Determine MicroSatellite Instability score in ColoRectal Cancer. MSI score is used in tumors for prognostic and therapeutic choices, but is also important to study defective DNA repair mechanisms.

Data Reception (Data preparation & inclusion)	MSI are Short Tandem Repeats (STRs) are small (1-6 base pairs) repeating stretches of DNA scattered throughout the entire genome and account for approximately 3 % of the human genome. Methods to measure somatic MSI imply a comparison between the extent of such repeats in normal and pathological conditions (tumor vs control DNA). The quantification is made via specific tools looking for the exa-mers repeats directly in the sequenced reads (FASTQ) or extracted from the alignment files (BAM, CRAM). The MSI status is expressed as a 'MSI-H' or with a single score (a ratio of the satellite in tumor versus normal). An MSI-high sample can have a score of 40/50 (for example) or more, and MS-stable (MSS) samples usually have a lower value (0-5). Still there is no consensus on the scale to use for these values as different tools use different ranges. Synthetic data could mimic MS, since the most advanced simulation tool can also simulate MSI.
Storage and Interfaces (Data Storage & Management)	The MSI score itself is not sensitive data, but to extrapolate it from raw data we must use both tumor samples and matched control DNA (WXS or WGS).
Data Discovery	A common scenario often requires the concomitant information about MSI score and the mutational status of Mismatch Repair Genes, and/or their RNA level.
Data Access Management Tools	Life Sciences AAI & REMS
Data Processing (Data use)	Data generation / QC, TBD > WP8
Do you need to access data transnationally? Are you considering a federated system? Or something else?	
Are you already engaged with projects/efforts/initiatives beyond 1+MG/B1MG/GDI that are developing infrastructures for tackling your needs?	
Are you aware or any other use case requirements relevant for your use case?	
Are you using structured pheno-clinical data? If yes, which one; if no, do you need/plan to use it?	We usually report MS status as a biomarker field in our metadata model