DeiC
DANISH E-INFRASTRUCTURE COOPERATION

Deff
Danmarks Elektroniske Fag- og Forskningsbibliotek

Final reporting on data management pilot project

Reporting date	2018, August 24

Project:	ActionableBiomarkersDK		
Project start:	2016, August 1	**Project end:**	2018, June 30

1. The project and the infrastructure

The overall aim of the project ActionableBiomarkersDK has been to establish a Danish infrastructure for storing and updating the actionable biomarkers in human disease.

The user communities in Danish healthcare - including the academic and corporate research domains – are in need of interacting manually or via programmatic access with the infrastructure pending approval by the Danish Data Security Agency and relevant ethical committees. Hence, a central resource to handle storage and updates of actionable biomarkers for subsequent use in diagnostics and pharmaco-genomics is needed.
The user community in Danish healthcare is very large including stakeholders in healthcare, in academia and in industry. With ActionableBiomarkerDK the user communities are given the opportunity to test and use the data management workflows (when regulatory measures allow) such that the stakeholder groups gradually will expand and be tested from a number of different disease angles more broadly.

The data foundation for ActionableBiomarkersDK has been a number of already existing repositories worldwide (e.g. existing repositories ClinVar, COSMIC, dbNSFP etc.), which with the project have been aggregated and curated into one major national repository, the ActionableBiomarkersDK. All data used in the project has made use of already existing workflow prototypes for handling sequencing data. Similarly, text-mining workflows for handling full-length papers was already in place.

The project has been rooted in research groups at the universities UCPH, DTU and SDU. It has been coordinated from UCPH (Søren Brunak), who has been head of the boards of directors of the ESFRI ELIXIR infrastructure for biological information in Europe as well as head of the scientific advisory board at the European Bioinformatics Institute.
DTU has the expertise within the supercomputing field, including the handling of large data integration efforts. SDU has expertise also within supercomputing as well as large platforms of experimental equipment within proteomics.
All three involved universities participate in the Danish ELIXIR node.

The project has addressed the needs of the users of the data management infrastructure in various ways – a main task has been to develop the secure private cloud effort that use the two supercomputer facilities Computerome (DTU/KU) and ABACUS 2.0 (SDU) for pilot actions on cloud bursting and remote user access. The private cloud has to handle person-sensitive data as opposed to e.g. Openstack. Hence, meeting regulatory demand across the academic, hospital and company partners has been important.

For URL go to: www.computerome.dk
The computerome wiki can be browsed by topic.
Access to Computerome is open to both internal and external users.

2. Achievements of the pilot project

Milestone (WP1): An updatable workflow for a comprehensive aggregated database of human biomarkers (M24)
Establishing an aggregated, comprehensive database that can be updated on demand brings up a lot of issues with data and metadata management. The issues are not limited to the data within the database, but touch also the computer code involved in extract, processing and depositing of data, database schemas and methods of access to the data within the database. To alleviate these issues, the project focused on creating the workflow not as a one-off black box of activities, but standardized units that can be easily interchanged and customized. The project experienced that some components of the aggregated database would require a fairly standard import procedure that was not sufficient when trying to incorporate another data source. The approach that we have seen other projects take - customizing the original import phase to the requirements of new data source - introduces unnecessary complexity and uncertainty. In this work package we designed the workflow in a modular way by clearly defining what is the expected starting and completion conditions

for each step, which has made the whole process much more resilient and customizable, in case new data providers need to be incorporated or changes have been done to the already existing sources.

Milestone (WP2): A cloud compatible, implemented workflow for genomics and proteomics data preparing for biomarker extraction (M12)
The project has designed and implemented a workflow covering life cycle of data in a biomarker extraction project. It has identified and resolved multiple issues arising from utilizing multiple supercomputing centres simultaneously (Computerome (KU/DTU) and ABACUS 2.0 (SDU)) and has demonstrated advantages of collaborative approach to the cloud-computing model.
The challenges stemming from having researchers using datasets across multiple computing infrastructures are rooted in a siloed approach to using supercomputer installations (the traditional HPC model). We found that cloud technologies remove limitations of any one site and significantly speeds up the process of preparing data for analysis (from initial QA/QC to validated inputs for extraction pipelines). Working with multitude of expert analysts at participating research institutions allows for the identification of common patterns and issues in data management workflows, and leads to creating abstractions and automate certain parts of routinely performed tasks.
Key point in the process is aligning data and metadata management practices, and putting effort into adopting FAIR-based data sustainability practices. This reduces the ambiguity in the downstream analysis of the data, and has without a doubt contributed to the successful completion of this project.
Multiple attempts at adopting already existing standards for the purpose of the workflow have been made. This work has yet to be concluded, but has already demonstrated a net gain of workflows and lessons learned will help in similar efforts in the future.

Milestone (WP3): An updatable workflow for controlled vocabularies relevant for biomarker detection in scientific literature (M12)
The project has set up a workflow to gather and process information automatically from 15 million scientific full text articles provided by the DTU Library and PubMed Central. The workflow was applied on full text articles, in English only. It is based on a set of steps that is agnostic to the input source, namely 1) source cleaning, 2) entity identification, and 3) post-processing.
The full text articles supplied by the DTU Library consists of PDF files, whereas the PMC articles comes in an XML format. In both cases, the text must be extracted using automated measures, as manual extraction is not feasible. The project used pdftotext to convert the PDF files into text files. The XML processing has been done with lxml library and langdetect for language identification.
A Named Entity Recognition (NER) system was used to identify entities of interest. A NER system depends heavily on the dictionaries supplied, and is very sensitive to ambiguous words. To combat this, dictionaries from well-known and peer-reviewed databases were used with inclusion of other dictionaries. The project limited itself to included genes, diseases and subcellular compartments. The post-processing steps focuses on scoring the co-occurrence; using a metric that takes into account the mention of two entities across the whole document, and penalizing entities mentioned in many articles. To identify biomarkers it would be more beneficial to look only at the sentences in which a phenotype and biomarker has been identified, and the workflow can easily facilitate this.
The project has investigated how well NER can retrieve known associations between three very important associations in biomedical research: diseases and genes, proteins and proteins, protein and their subcellular localization. The findings from the full text articles were compared to the 16 million abstracts found in MEDLINE, and the use of full text articles gives better associations with a lower false positive rate. It is strongly believed that the addition of full text articles will provide more information, and a better method for performing novelty screening on phenotype-biomarker associations.

Milestone (WP4): An improved reference for better biomarker identification in the Danish population (M9)
The group at DTU has been a main contributor to the generation of the Genome Denmark reference. It assembled a reference genome representing the population of Denmark with high quality. Objective measures indicate that it constitutes the best population-specific reference genome to date. In turn, the quality of the reference genome to map the sequencing reads against is a major determinant of the quality of the variant calling and an extremely valuable resource in the filtering of common variants private to the population on focus.

Milestone (WP4): An updatable workflow for a comprehensive biomarker annotation (M18)
The project designed and implemented a pipeline for the identification of high confidence variants that might likely play a role in the onset of a number of human diseases. Current state-of the-art sequencing technologies are capable at producing sequencing data at astonishing throughput at affordable costs. Similarly, in the advent of life sciences dedicated supercomputing, the comparison of the millions of short sequencing reads produced by the sequencers to the reference genomes of choice and the identification of variants with respect to these reference genomes constitutes a routinely automated task. A typical sequencing project would identify several millions of these differences, of which only a relatively small fraction would have a causative role in the onset of diseases such as diabetes, cancer, congenital heart disease, etc. The screening for the relevant variants constitutes a non-trivial task that accounts for most of the analysis time in a

sequencing project. A battery of strategies for alleviating the burden of mutation filtering based on expert knowledge of the healthy population of Denmark was implemented in the project.

In addition to the Danish reference genome, the pipeline incorporates additional resources that provide information on the expected allele frequencies of variants commonly found in the healthy population. Some examples are the variants stemming from international consortium projects such as the 1000 Genomes, HapMap, GnoMAD, and the exome aggregation consortium (ExAC), among many others. Further filtering of non-causative variants is exerted by the implementation of a number of modules providing different levels of functional annotation. This ranging from an evaluation of the function of the genes harbouring the mutations (e.g. Gene Ontology enrichment analyses or InWeb-based identification of mutated protein-protein interactions) to an assessment of the evolutionary constraints prohibiting the mutation of certain conserved genomic positions (e.g. KinMut5 or SIFT).

The pipeline has been instrumental in a number of ongoing projects aiming to the characterisation of the aetiology of diseases such as cancer or congenital heart disease. Similarly, the pipeline has been used in a number of N=1 discovery approaches preparatory for the advent of precision medicine. Ramifications of the pipeline led to ongoing collaborations with the Centre for GeoGenetics aiming to the identification of variants in ancient archaeological samples dating from Viking and Iron Age times.

Milestone (WP5): Virtual integration of sensitive data through cloud bursting (M24)
The process of analysis and integration of sensitive data must prioritize the security at every step of the analysis. When creating pipelines and sequential processing of any kinds of sensitive data, it has been discovered that it is disturbingly easy to leak data from within the pipeline by crashing components of the analysis (e.g. typo in a script). We have evaluated the environments and computational capabilities of DTU and SDU and confirmed that existence of any architectural differences of supercomputing environments makes it difficult to safeguard against this form of attack on the data – it is simply not feasible to create custom wrappers for each and every analysis ran in any configuration of distributed computing. We have turned our attention to abstracting this problem and container technologies as a method of accomplishing it – we have evaluated LXC, Docked, Singularity and Kubernetes. The container technology allowed us to "plug and play" various tools and even whole pipelines, as the underlying frameworks are taking care of translating hardware differences into safe and easily re-creatable environments. Support of the biggest software companies in the world - Kubernetes has been designed by Google and is widely used in their infrastructure to provide advanced processing capabilities – means that there are thousands of highly skilled people continuously testing and patching any potential security vulnerabilities.
We already see the container technologies emerging as the technology of choice when dealing with distributed computing and the fact that the containers themselves can be "frozen" and audited makes them a perfect mechanism for creating reusable pipelines for processing of sensitive data.

3. Sustainability and vision for permanent infrastructure

The established infrastructure is intended to be sustainable by design. All components, tools, workflows, plans, technologies, frameworks and results have been designed to be as reusable as a whole and in parts. The pipelines, complex computer code and other in-silico artifacts are available in the software catalogues of the supercomputing infrastructures hosting them. It is our intention to reach out to as much of the community of researchers that could use the deliverables as possible. In order to achieve this we are planning on adopting FAIRification of the data (making it Findable, Accessible, Interoperable and Reusable).

Please note that some of the datasets and code, e.g. database of articles used for text mining, are under special licensing agreements and in these situations access to the copy might not be possible.
The tools and methods, once they mature, will optionally be made available via the NeIC's code refinery registry allowing for a broader reach especially in conjunction with additional training materials in the registry to reduce "time to research" for its users.

Appendix 1: financial report

Financial report has been sent separately to DeiC.

Appendix 2: List of publications (with reference in full) or papers in review (working title)

Westergaard D, Stærfeldt H-H, Tønsberg C, Jensen LJ, Brunak S. Text mining of 15 million full-text scientific articles. bioRxiv 2017; : 162099.

Besenbacher, S. et al. Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. Nat Commun 6, 5969 (2015).

Maretty, L. et al. Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. Nature

Skov, L., Consortium, T. D. P.-G. & Schierup, M. H. Analysis of 62 hybrid assembled human Y-chromosomes exposes rapid structural changes and high rates of gene conversion. PLoS Genet. 13, e1006834 (2017).

Jensen, J. M. et al. Assembly and analysis of 100 full MHC haplotypes from the Danish population. Genome Res. gr.218891.116 (2017). doi:10.1101/gr.218891.116

Vazquez, M., Pons, T., Brunak, S., Valencia, A. & Izarzugaza, J. M. G. wKinMut-2: Identification and Interpretation of Pathogenic Variants in Human Protein Kinases. Hum. Mutat. 37, 36–42 (2016).

Appendix 3: List of dataset publications (with reference in full)

List of data set publications is not applicable for the project.