Data management pilot project report

Report date	2017 11 21	Report #	3

Project title	ActionableBiomarkersDK		
Grant holder	Prof. Søren Brunak, KU (UCPH)		
Partner institutions	SDU + DTU + Zealand University Hospital		
Project start	01.08.2016	Project end	30.06.2018

Overall assessment at this point in time

The project is on track and momentum is good.

Action points for grant holder, DM Secretariat or others

	Assignee	Deadlines	Who's in the loop
N/A	N/A	N/A	N/A

Project progress so far

- **WP1: Data capture, data harmonization, conversion of unstructured data into structured biomarker formats**

  Milestone (T1.1): An updatable workflow for a comprehensive aggregated database of human biomarkers (M24). In progress.

- **WP2: Data management effort addressing primary data types: genome and proteome sequences**

  Milestone (T2.1): A cloud compatible, implemented workflow for genomics and proteomics data preparing for biomarker extraction (M12).

  Not begun yet. A challenge is lack of people who understand the particular life science data analysis workflows, plus Abacus support lacks. Accomplished. See annex.

- **WP3: DTU text mining effort addressing full length papers for novel biomarker detection**

  Milestone (T3.1): An updatable workflow for controlled vocabularies relevant for biomarker detection in scientific literature (M12). Accomplished, see annex

- **WP4: Data management workflows implementing the condensation of genomic and proteomic data into actionable biomarkers**

  Milestone (T4.1): An updatable workflow for a comprehensive biomarker annotation (M18). Accomplished. See annex.

  Milestone (T4.2): An improved reference for better biomarker identification in the Danish population (M9). Accomplished. See annex.

- **WP5: Secure private cloud effort for biomarker workflows on Computerome (DTU/KU) and ABACUS 2.0**

  Milestone (T5.1): Virtual integration of sensitive data through cloud bursting (M24).

  In progress for genomics - proteomics is not in the loop from the start, but is expected to join shortly. VM containerization down to network level (=> HW agnostic) and APIs for automating the ordering of resources have already been completed. Establishing tunnel and call back between Computerome and Abacus for storing and updating the biomarkers is in progress. Likewise, with resource pooling.

  Discussion: Establish a "joint cloud" of Abacus and Computerome for the life sciences, i.e. application of seamless access to both (mutually exchangeable seen from researcher perspective). Henrik Pedersen has been in contact with Martin Zachariasen, Claudio Pica and Ole Nørregaard Jensen to promote this idea.

Budget and timeline

The budget is geared through other project funding and the project is on track to reach all milestones in time.

Annex WP2

Milestone: *A cloud compatible, implemented workflow for genomics and proteomics data preparing for biomarker extraction (M12).*

We designed and implemented a workflow covering life cycle of data in a biomarker extraction project. We have identified and resolved multiple issues arising from utilizing multiple supercomputing centers simultaneously (Computerome (KU/DTU) and ABACUS 2.0 (SDU)) and have 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 have found that utilizing cloud technologies removes 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 allowed us to identify common patterns and issues in data management workflows and lead to creating abstractions and automate certain parts of routinely performed tasks.

Key point in the process, that should not be understated, is aligning data and metadata management practices and, as soon as possible in the project, putting effort into adopting FAIR-based data sustainability practices. This has allowed us to significantly reduce ambiguity in the downstream analysis of the data and without a doubt has contributed to the ability to achieve the goal of this project.

Over the duration of this task, we have made multiple attempts at adopting already existing standards for the purpose of the workflow. While we have not reached a point where we can say we have concluded our work, on the data schema and principles for the following biomarker extraction, we can with confidence demonstrate a net gain of our workflow for this project and will take lessons learned to help us in similar efforts in the future.

Annex WP3

Milestone: *An updatable workflow for controlled vocabularies relevant for biomarker detection in scientific literature (M12)*

In this milestone we have set up a workflow to gather and process information automatically from 15 million scientific full text articles provided by the DTU Librarym and PubMed Central[1]. For the latter articles we only included those approved for text mining. We have demonstrated that the workflow works on full text articles. The workflow is based on a set of steps that is agnostic to the input source. Our workflow consists of a number of steps, namely

1. Source cleaning
2. Entity identification
3. Post-processing

The full text articles supplied by the DTU Library consist 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. We used pdftotext (v0.47.0, part of the Poppler suite available at poppler.freedesktop.org) to convert the PDF files into text files. We used the Python lxml library to extract text from the XML files. Language identification was done using the Python library langdetect (v1.0.7, https://pypi.python.org/pypi/langdetect). We only included articles in English. Furthermore, we applied a series of steps to filter out articles in which the PDF to TXT conversion had gone bad. In some cases the PDF files were scans of the article, however we did not attempt any optical character recognition conversion (OCR) as the old typesetting fonts often are less compatible with present day OCR programs.

We used a Named Entity Recognition (NER) system to identify entities of interest. A NER system depends heavily on the dictionaries supplied, and is very sensitive to ambiguous words. To combat this we have used dictionaries from well-known and peer-reviewed databases, and we have included other dictionaries to avoid ambiguous terms. In this pilot study we limited ourselves to included genes, diseases and subcellular compartments. However, the addition of a dictionary tailored for biomarkers is trivial.

Our post-processing steps consist mainly of scoring the co-occurrence. In the pilot study we used a metric that takes into account the mention of two entities across the whole document, and penalizes entities that are mentioned in many articles. For the identification of biomarkers it would be more beneficial to look only at the sentences in which a phenotype and biomarker has been identified, and our workflow can easily facilitate this.

In the pilot study we investigated how well NER could retrieve known associations between diseases and genes, proteins and proteins, protein and their subcellular localization. These are three very important type of associations in biomedical research. We compared our findings from the full text articles to the 16 million abstracts found in MEDLINE, and found that using full text articles we can retrieve better associations with a lower false positive rate. Consequently, we strongly believe that the addition of full text articles will provide more information, and a better method for performing novelty screening on phenotype-biomarker associations.

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

Annex WP4

Milestone (T4.2): *An improved reference for better biomarker identification in the Danish population (M9).*

Our group at DTU has been a main contributor to the generation of the Genome Denmark reference[1-4]. We assembled a reference genome representing the population of Denmark with high-quality. Objective measures indicate that the Danish one 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 (T4.1): *An updatable workflow for a comprehensive biomarker annotation (M18)*

We 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, separating the wheat from the chaff, constitutes a non-trivial task that accounts for most of the analysis time in a sequencing project. We have implemented a battery of strategies for alleviating the burden of mutation filtering based on our expert knowledge of the healthy population of Denmark.

In addition to the Danish reference genome our 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 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. KinMut[5] or SIFT).

Our 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 Center for GeoGenetics aiming to the identification of variants in ancient archeological samples dating from Viking and Iron Age times.

1. Besenbacher, S. *et al.* Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat Commun* **6,** 5969 (2015).
2. Maretty, L. *et al.* Sequencing and *de novo* assembly of 150 genomes from Denmark as a population reference. *Nature*
3. 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).
4. 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
5. 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).