From thousands of mass spectrometry profiles to biomarkers within a day ? '*Arion 4 Omics'*, a novel solution to facilitate and accelerate omics-based decision making.

Doroteya K. Staykova, Matthew E. Lea.
Multicore Dynamics Ltd

MULTICORE DYNAMICS
data illumination accelerated

Motivation

As momentum is growing for the profiling of disease using data generated from population based studies, high-throughput instruments generate staggering amounts of omics data. A number of authors [1, 2] have addressed the problems currently faced with the management, integration and interpretation of the resulting mass of data which together with escalating costs, are recognised as being the main contributing factors adversely affecting the transition from conventional to personalised medicine [2, 3].

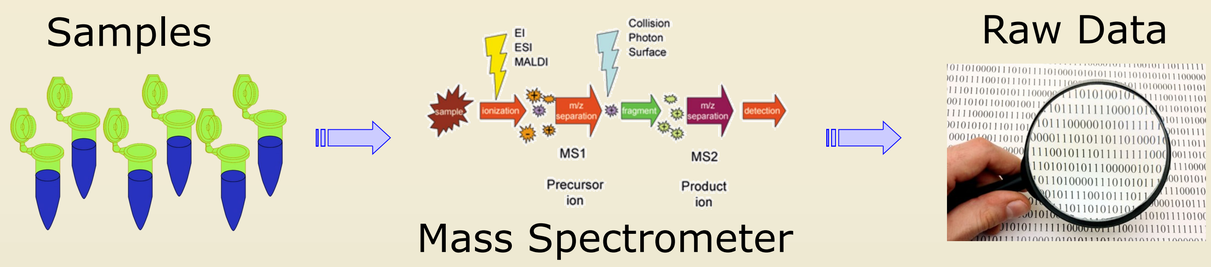
We present 'Arion 4 Omics', an *advanced*, high-performance analytical platform designed for the efficient management and systematic analysis of large-scale proteomics data. Featuring automated quality checks, processing, statistics, machine learning and auditing (Fig.1.) with further modules planned.

Mass Spectrometry in Clinical Studies

Recent advancements in liquid chromatography-tandem mass spectrometry (LC-MS/MS), enable short measurement times and high sensitivity for the detection of low-abundant biomolecules in clinically accessible fluids (e.g. blood, saliva, etc).

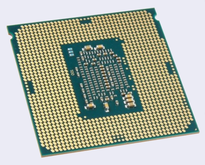
Large-cohort clinical studies can utilise such high-throughput technology to profile disease groups on a molecular level or for the discovery and validation of biomarkers [4, 5].

Process



An 'all in one' solution

'Arion 4 Omics' employs advanced database technologies to process in parallel, key database operations. This provides ultra-rapid queried results to facilitate and accelerate data exploration.



'Arion 4 Omics' uses scalable, parallel hardware containing thousands of processing cores.

Novel domain-specific algorithms are exhaustively benchmarked to ensure accurate reproducible results, consistent performance and ability to maximise data insight.

System strengths include -

- ☑ Flexible choice of pre-defined processing steps
- ☑ Intuitive interface to update and re-process large sample sets with further samples
- ☑ Database driven integration of clinical and omics data
- ☑ Rapid processing of large scale, high dimensional datasets
- ☑ Auditing, full traceability of actions performed
- ☑ Integrity, automated data checks prior to transformation
- ☑ Reproducibility of results
- ☑ Scalable with a no compromise approach to systems reliability
- ☑ Compatible with multiple machine instrument vendors

ARION 4 OMICS

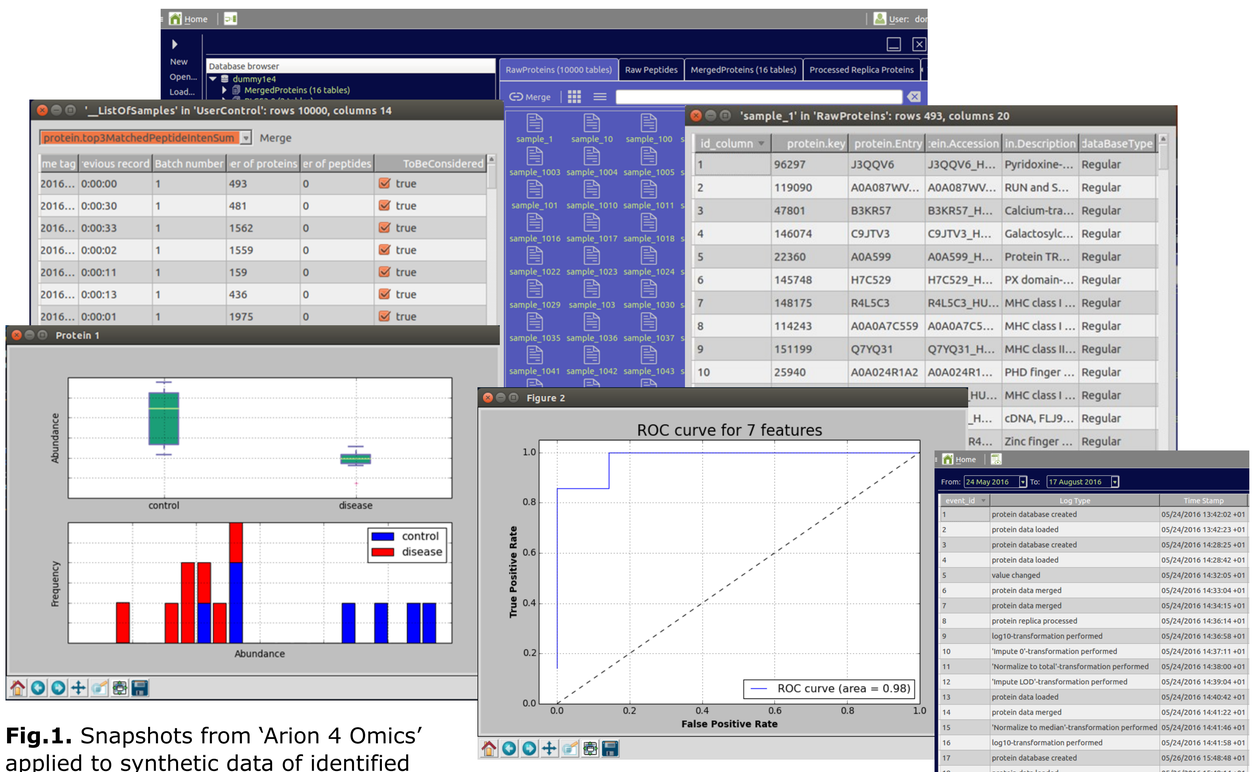


Fig.1. Snapshots from 'Arion 4 Omics' applied to synthetic data of identified proteins and sample groups.

Top left—Following data import and quality checks, the user may select which samples should be processed.

Bottom left—Box plots and histograms offer visual exploration of biomarker results.

Bottom centre—Receiver operating characteristics (ROC) curves associated with machine learning provide excellent diagnostic tools for medical research.

Data Driven Intelligence

Additional methods of analysis such as the machine learning module, provide an alternative and fresh approach to the analysis process. Machine learning is quickly positioning itself as a primary diagnostic tool in modern healthcare as it promises to not only deliver biomarkers but provide data-driven, clinical decision making. Besides offering cost-effective techniques for the utilisation of biomolecular data, it can assist physicians in their diagnosis and propose a suitable course of medication. As a predictive tool, it can quickly identify high risk patient groups and those likely to be re-admitted.

Benchmarking & performance

'Arion 4 Omics' was benchmarked using a prototype hardware configuration using synthetic protein data (Fig.1). Results consistently show that processed datasets and biomarkers can be delivered in less than a day.

Execution times for the pipeline routines e.g. data merging, normalization, imputation etc. were observed for both intensive and less intensive database and processing operations, with result times ranging from a couple of minutes to a couple of hours (Fig.2).

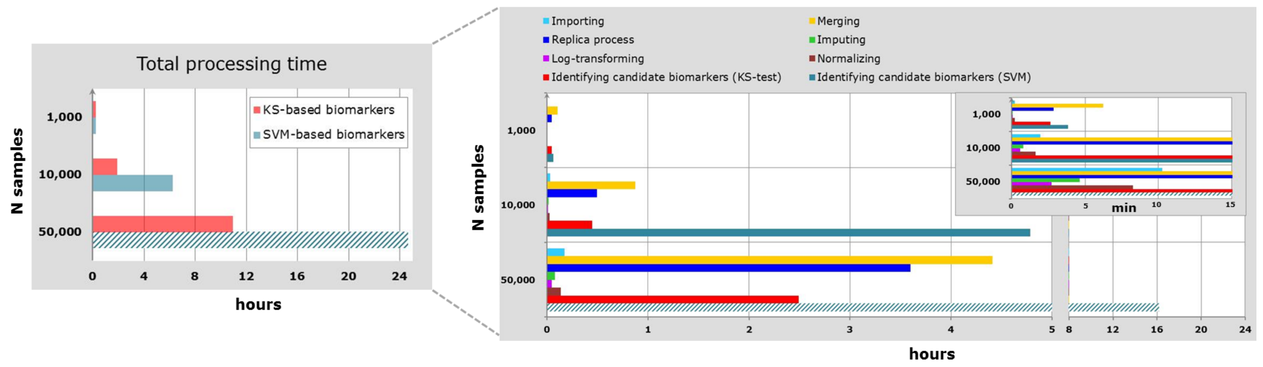


Fig.2. The view on the left shows the total processing time for 1,000 to 50,000 samples of synthetic protein data generated from an entire set of 152,493 human proteins from the UniProt database [6]. The view on the right, shows benchmarked times for the individual steps.

Conclusion

Before translational medicine can become a viable and practical reality, researchers and analysts require the seamless integration of advanced technologies to support the analysis of increasingly large datasets with near instantaneous, quantitative results.

'Arion for Omics' incorporates carefully selected, benchmarked technologies providing supercomputing power in a sensibly priced, integrated solution. It is a state of the art, pre-processing, analysis and diagnostic pipeline, designed for managing both small and large scale, structurally complex, omics-based datasets.

This future proof solution, will complement clinical research teams and lend itself to the development of personalised medicine by encouraging data driven, clinical decision making that until now, has largely been held back.

References

[1] Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. (2015) Big Data: Astronomical or Genomical?. *PLoS Biol* 13(7):e1002195. doi:10.1371/journal.pbio.1002195
[2] Alyass A, Turcotte M, Meyre D. (2015) From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Medical Genomics* 8(33) doi:10.1186/s12920-015-0108-y
[3] Mardis ER (2010) The $1,000 genome, the $100,000 analysis?. *Genome Medicine*, 2(84), http://genomemedicine.com/content/2/11/84
[4] Lesur A, Gallien S, Domon B. (2016) Hyphenation of fast liquid chromatography with high-resolution mass spectrometry for quantitative proteomics analyses. *Trends in Analytical Chemistry*, In Press
[5] Wheelock CE, Goss VM, Balgoma D, Nicholas B, Brandsma J, Skipp PJ, et al. (2013) Application of 'omics technologies to biomarker discovery in inflammatory lung diseases. *Eur Respir J*, 42(3):802-25
[6] http://www.uniprot.org/uniprot/?query=human&fil=organism%3A%22Homo+sapiens+%28Human%29+[9606]%22&sort=score