GIO: Standards-compliant Galaxy workflows for proteomics informed by transcriptomics

Jun Fan[1], Shyamasree Saha[1], Adelyne Sue Li Chan[1], David A Matthews[2] and Conrad Bessant[1*]

Introduction

The most common method of identifying proteins in a complex sample is to perform liquid chromatography tandem mass spectrometry (LC-MS/MS) then search the acquired spectra against a reference proteome downloaded from a database such as UniProt. This approach has the major drawback of not being able to identify gene products that are not already known. We recently developed the proteomics informed by transcriptomics (PIT) methodology, which tackles this problem by using RNA-seq to generate sample-specific protein databases that the LS-MS/MS data can be searched against [1]. This allows the detection and quantitation of previously unknown proteins, protein variants and other exotic translated genomic elements. This is of particular utility when studying non-model organisms and samples with very dynamic proteomes, e.g. stem cells, cancer cells and virus-infected cells. The analysis of PIT data is complex and computationally intensive, requiring the integration of multiple third party tools from the proteomics, transcriptomics and genomics communities. To make this analysis tractable and repeatable we have produced GIO (Galaxy Integrated Omics) – a Galaxy-based framework containing the key tools and workflows needed to analyse data from PIT experiments in a reliable and repeatable way.

Proteomics informed by transcriptomics (PIT)

In a PIT analysis the sample of interest is analysed by both RNA-seq and LC-MS/MS, as shown in Figure 1. The RNA-seq short reads are *de novo* assembled into transcripts using a tool such as Trinity [2]. These transcripts are then used to produce a list of open reading frames (ORFs) against which we can attempt to match peptide spectra from the mass spectrometer. A strong match indicates that a transcript is being translated to a polypeptide (typically a protein).

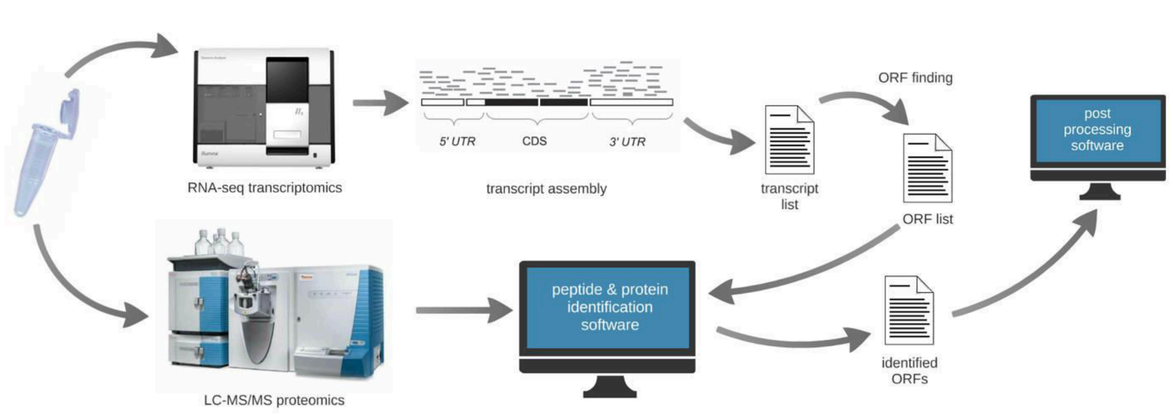


Figure 1: *Schematic representation of PIT analysis. The results are post-processed depending on the application – this may involve determining the likely identity of observed ORFs by homology search, or mapping observed ORFs to a reference genome.*

GIO: A customised version of Galaxy for PIT analysis

Data from early PIT experiments was analysed using a combination of closed source proteomics software packages, tools from the NGS community, and in-house Perl scripts [1]. This combination proved to be challenging to configure and use, impossible to scale, and difficult to repeat. To overcome these problems, we established an in-house Galaxy server, to which we added a range of tools for performing steps in the PIT data analysis protocol. These included pre-existing proteomics tools (some of which had already been wrapped by the Galaxy-P project [3]), pre-existing sequence analysis tools, and in-house tools written specifically for PIT. To avoid confusion with more typical Galaxy servers we named our server GIO, for Galaxy Integrated Omics.

The tools are combined into workflows that support the analysis of data for various applications of PIT. One such workflow, which identifies the proteins present in a sample and uses this information to annotate a reference genome, is shown in Figure 2. A key facilitator for workflow creation has been the adoption within GIO of data formats from the Proteomics Standards Initiative (PSI) [4] which allow tools to be effortless connected together. Standardising output formats has also simplified the visualisation of results within GIO, with in-house tools for viewing peptide and protein identifications and existing tools for showing genome annotation.

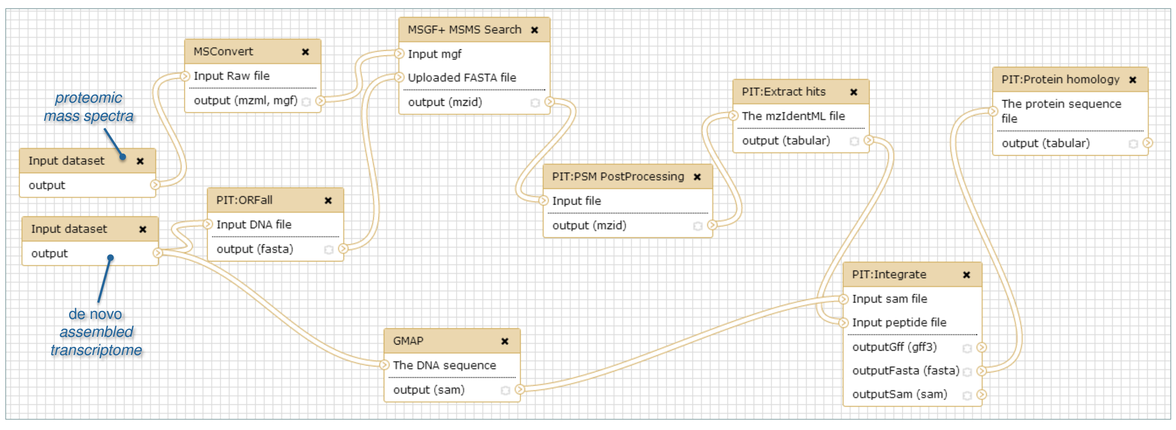


Figure 2: *Example of a genome annotating PIT workflow implemented within GIO. The workflow is comprised of pre-existing tools from the proteomics and genomics communities, integrated using in-house tools developed in various programming languages.*

Conclusion and future plans

GIO has proven to be an excellent framework within which to develop, use and share the complex data processing workflows needed to analyse data from PIT experiments. It is already being used by groups who would not otherwise be able to perform such analysis without specialised bioinformatics support. Development of more advanced workflows, particularly for the downstream processing and visualisation of results, is a priority for future work. Support for quantitative analysis is another obvious direction for development.

Funded by BBSRC grants BB/K016075/1 and BB/L018438/1.

[1]School of Biological and Chemical Sciences, Queen Mary University of London, UK.
[2]School of Cellular and Molecular Medicine, University of Bristol, Bristol, UK.
*Correspondence should be addressed to **c.bessant@qmul.ac.uk**

Results

GIO has been tested on various PIT datasets, including one acquired from HeLa cells infected with adenovirus [1]. The results, summarised in Figure 3, show that a PIT analysis in which no reference genome is used identifies over 90% of the proteins that can be found using a traditional search against the UniProt reference proteomes from human and adenovirus. Additionally, PIT finds a number of proteins that were missed in the UniProt search. These results concur with those obtained in the original non-Galaxy analysis.

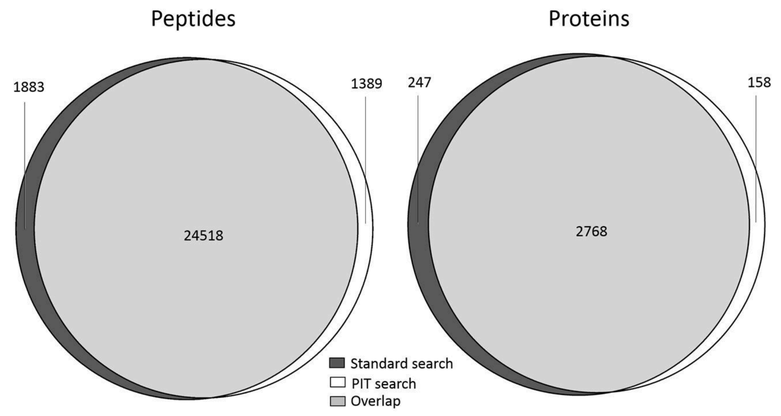
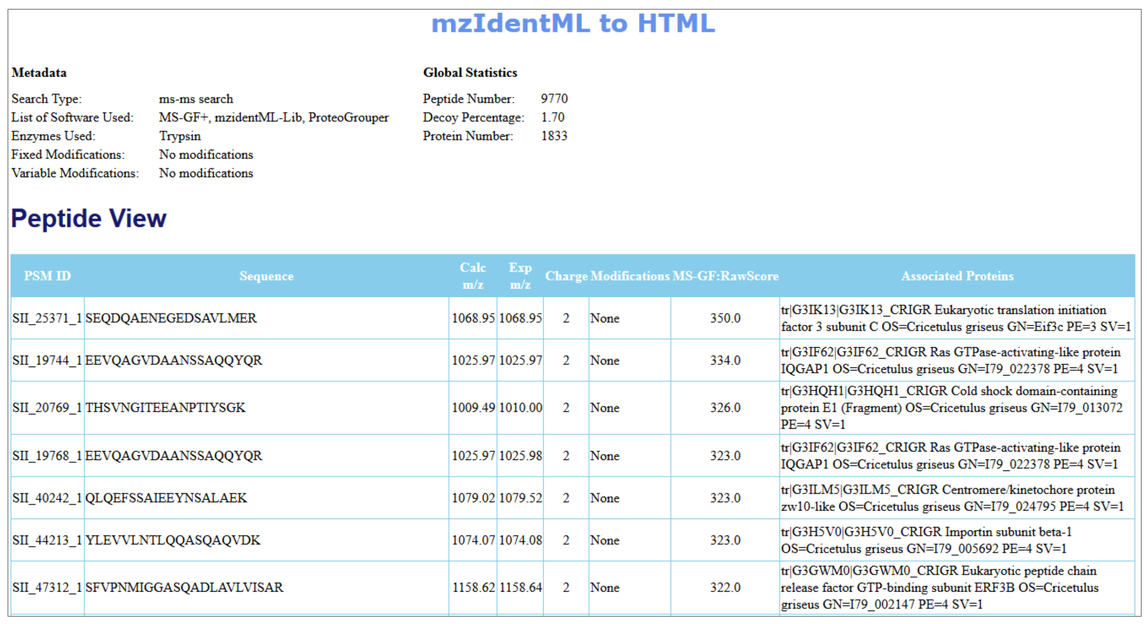


Figure 3: *Comparison of peptides and protein groups identified by a standard protein identification workflow in which mass spectra were searched against Uniprot, and a PIT workflow using a* de novo *assembled transcriptome, for HeLa cells infected with adenovirus.*

By implementing HTML renderers for the PSI data standards, we are able to view the results of proteomics analysis directly within Galaxy, allowing instant access to results without the need to download files and view them in a locally installed application. An example of such output is shown in Figure 4.



mzIdentML to HTML

Metadata

Search Type:	ms-ms search
List of Software Used:	MS-GF+, mzidentML-Lib, ProteoGrouper
Enzymes Used:	Trypsin
Fixed Modifications:	No modifications
Variable Modifications:	No modifications

Global Statistics

Peptide Number:	9770
Decoy Percentage:	1.70
Protein Number:	1833

Peptide View

PSM ID	Sequence	Calc m/z	Exp m/z	Charge	Modifications	MS-GF:RawScore	Associated Proteins
SII_25371_1	SEQDQAENEGEDSAVLMER	1068.95	1068.95	2	None	350.0	tr|G3IK13|G3IK13_CRIGR Eukaryotic translation initiation factor 3 subunit C OS=Cricetulus griseus GN=Eif3c PE=3 SV=1
SII_19744_1	EEVQAGVDAANSSAQQYQR	1025.97	1025.97	2	None	334.0	tr|G3IF62|G3IF62_CRIGR Ras GTPase-activating-like protein IQGAP1 OS=Cricetulus griseus GN=I79_022378 PE=4 SV=1
SII_20769_1	THSVNGITEEANPTIYSGK	1009.49	1010.00	2	None	326.0	tr|G3HQH1|G3HQH1_CRIGR Cold shock domain-containing protein E1 (Fragment) OS=Cricetulus griseus GN=I79_013072 PE=4 SV=1
SII_19768_1	EEVQAGVDAANSSAQQYQR	1025.97	1025.98	2	None	323.0	tr|G3IF62|G3IF62_CRIGR Ras GTPase-activating-like protein IQGAP1 OS=Cricetulus griseus GN=I79_022378 PE=4 SV=1
SII_40242_1	QLQEFSSAIEEYNSALAEK	1079.02	1079.52	2	None	323.0	tr|G3ILM5|G3ILM5_CRIGR Centromere/kinetochore protein zw10-like OS=Cricetulus griseus GN=I79_024795 PE=4 SV=1
SII_44213_1	YLEVVLNTLQQASQAQVDK	1074.07	1074.08	2	None	323.0	tr|G3H5V0|G3H5V0_CRIGR Importin subunit beta-1 OS=Cricetulus griseus GN=I79_005692 PE=4 SV=1
SII_47312_1	SFVPNMIGGASQADLAVLVISAR	1158.62	1158.64	2	None	322.0	tr|G3GWM0|G3GWM0_CRIGR Eukaryotic peptide chain release factor GTP-binding subunit ERF3B OS=Cricetulus griseus GN=I79_002147 PE=4 SV=1

Figure 4: *The top portion of a HTML rendering of a list of identified peptides stored in a PSI identification format (mzIdentML) file.*

For species for which a genome assembly exists, a popular way to visualise the results of PIT analysis is to show them in their genomic context. We are working to provide this functionality in GIO by embedding the Genoverse genome browser [5], as shown in Figure 5.

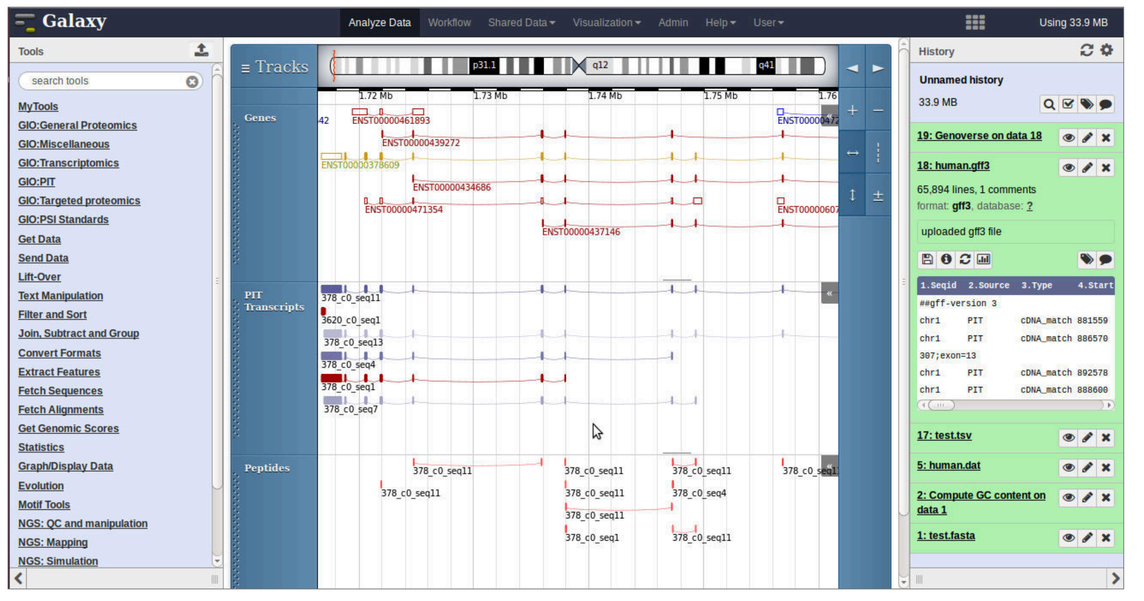


Figure 5: *Example of Genoverse view of a region of the human genome annotated with transcripts and peptides identified using GIO's genome annotating PIT workflow. Here, peptides and transcripts both serve to confirm gene structures that have already been annotated in the reference genome.*

References

1. Evans, V.C.; Barker, G.; Heesom, K.J.; Fan, J.; Bessant, C.; Matthews, D.A., *De novo* derivation of proteomes from transcriptomes for transcript and protein identification. *Nature Methods* **2012,** *9* (12), 1207-11.
2. Grabherr, M. G.; Haas, B. J.; Yassour, M.; Levin, J. Z.; Thompson, D. A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; Chen, Z.; Mauceli, E.; Hacohen, N.; Gnirke, A.; Rhind, N.; di Palma, F.; Birren, B. W.; Nusbaum, C.; Lindblad-Toh, K.; Friedman, N.; Regev, A., Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **2011,** *29* (7), 644-52.
3. Jagtap, P.D.; Johnson, J.E.; Onsongo, G.; Sadler, F.W.; Murray, K.; Wang, Y.; Shenykman, G.M.; Bandhakavi, S.; Smith, L.M.; Griffin, T.J. Flexible and Accessible Workflows for Improved Proteogenomic Analysis Using the Galaxy Framework. *Journal of Proteome Research* **2014**, *13* (12), 5898-5908.
4. www.psidev.info
5. www.genoverse.org