Open-source bioinformatic solutions for 'Big Data' analysis

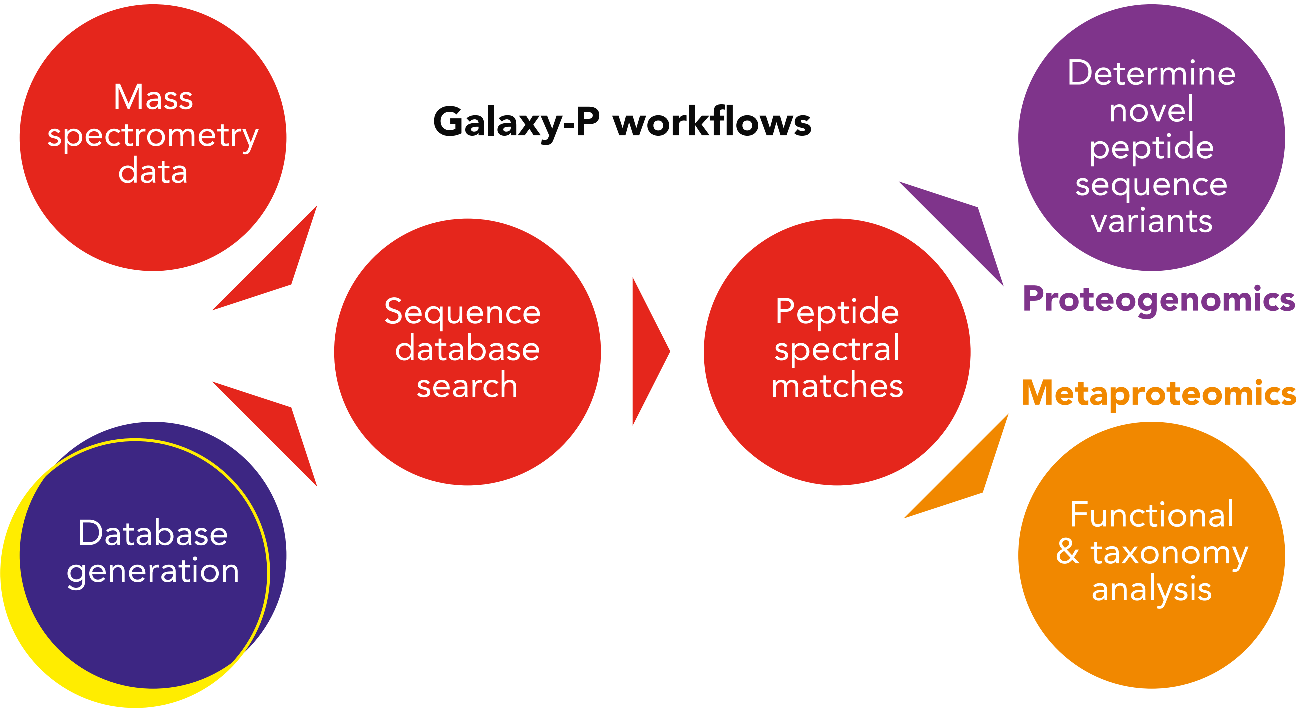
***Drs Tim Griffin** and **Pratik Jagtap** along with the Galaxy-P team from the University of Minnesota are working to develop workflows on an open source platform for the analysis of multi-omic data. They are currently focusing on using a Galaxy-based framework to investigate the integration of genomic datasets with mass spectrometry-based 'omics' data. But in the long term, they aim to expand the platform to cope with many other 'Big Data' domains.*

Currently, a major limitation to what we can discover from complex datasets derived from next-generation technologies is our ability to analyse them. This is where the work of Dr Tim Griffin, Dr Pratik Jagtap and their research team will play an important role.

THE 'BIG DATA' ERA

Moore's Law predicts that computing power will double approximately every two years, and with this, the cost of high-powered machines will also decrease. However, this cannot continue indefinitely and 2017 may be the crunch point at which physical limitations intervene, with the rate of progress becoming ever more saturated.

But what influence has this increase in computer power had on science? One of the major advances has been the ability to generate data using next generation, high throughput techniques, resulting in 'Big Data'. Although 'Big Data' has been used to define many datasets, the term often corresponds to what are now commonly known as 'omics datasets' – genomics, metabolomics, proteomics, transcriptomics and epigenomics to name but a few. For example, in biomedical science, we see large scale, system-wide approaches being used more and more commonly. These include the 1000 Genomes Project, the emergence of personalised medicine – tailored to an individual's needs – and systems biology, examining multiple, interacting pathways concurrently as one giant network.



However, the analysis of these large and complex datasets requires an analytical platform which can cope with the intense informatics requirements, as well as the ability to access disparate software from different 'omics' domains. Many wet-bench researchers will not have access to this level of compute-power or expertise locally, and therefore there is an increase in remote, or cloud, open-access platforms being used to access the necessary bioinformatic tools needed to cope with the complex results that researchers are obtaining.

> One of the major advancements is the ability to generate data using next generation, high throughput techniques, resulting in 'Big Data'

ONE SOLUTION FOR ALL

At the University of Minnesota, Drs Tim Griffin, Pratik Jagtap and team are working on solutions to analyse these complicated datasets. This is a multi-disciplinary, collaborative project between Dr Griffin's lab and the Minnesota Supercomputing Institute, which involves software developers, data scientists and wet-bench biological researchers. Specifically, the team are focusing on mass spectrometry (MS)-based 'omics' data (metabolomics and proteomics) and how they can harness an existing open-source framework, called Galaxy.

Put simply, mass spectrometry represents a high throughput technique that sorts ions based on their mass to charge ratio. Once certain signatures have been recorded for individual ions, this information can, for example, be extrapolated to identify peptides, the building blocks of proteins. Tandem mass spectrometry (MS/MS) further expands on this by using at least two stages of mass analysis.

GALACTIC PLATFORM

Galaxy was originally developed over a decade ago to solve problems in genomic informatics. It can be hosted on a scalable compute infrastructure, helping to cope with the problem of large data volume, and can be accessed remotely by researchers across the globe. Supported by a team of experts and software developers, Galaxy integrates many individual 'omics tools in a single environment, and also has many functionalities that promote workflow sharing and reproducibility. The latter is particularly important, as there may be multiple research projects that can utilise one particular dataset or workflow. Data sharing and transparency also encourages collaboration, and increases the number of expert approaches that can be combined to maximise novel findings.

In particular, the Galaxy for proteomics (Galaxy-P) team investigates ways in which genomic and transcriptomic data can be integrated with MS-based proteomics data. From here, they aim to verify the expression of

Members of the Galaxy-P team, (http://galaxyp.org/people/)

protein sequence variants that result from sequence variations at the DNA or RNA level. This approach, known as proteogenomics, commonly uses transcriptomic data translated *in silico* to produce a customised protein sequence database. This database is subsequently used to match proteins obtained through MS technologies. The major advantage of this approach is that no existing reference sequence is required, and so novel protein sequence variants, which may previously have gone undetected, can be identified. The analysis can also be extended to compare expression levels of genes and proteins.

Similar to proteogenomics, metaproteomics is also based on integration of metagenomic data with MS-derived proteomics data. However, unlike the previous approach, this concentrates on integrating these with sequence data derived from bacterial communities (microbiomes). As before, metagenomic data are translated *in silico* to create a protein sequence database. MS/MS peak lists, derived from the raw data, are matched against the database. Once peptides of interest have been identified, they are assigned to taxonomies and verified. Additional analysis using tools for functional analysis such as MEGAN, provide information about the functional categories of microbial protein expression. Metaproteomics can provide us with functional data to complement the taxonomical findings of a metagenomic approach. The main draw of this approach is that it can potentially be used to analyse data from diverse sample types – ranging from clinical to environmental samples.

An example of where Galaxy-P (galaxyp.org) provides ideal tools could be in helping cancer researchers identify which protein sequences may have a functional role in causing a specific cancer. Not only does Galaxy-P provide the necessary tools required for complex analyses, it can also potentially train non-expert, bench scientists through public Galaxy platforms (tiny.cc/galaxyp-proteogenomics; z.umn.edu/metaproteomicsgateway). This platform provides small-scale data for users to access and use with already published workflows. Existing studies have already used the Galaxy-P platform successfully to look at a range of topics, from proteogenomic analysis of hibernating mammals, to protein expression in the lungs of patients with acute respiratory distress syndrome.

TO INFINITY AND BEYOND

Drs Griffin and Jagtap hope their work will provide a novel environment to integrate multiple 'omics' datasets, and that this approach will provide unique opportunities for future discovery.

So far, the Galaxy-P team has advanced the abilities of Galaxy to cope with the many challenges of multi-omics informatics. An accessible, unified environment now exists to help non-experts navigate the analysis of MS-based proteomics and metabolomics data, in addition to a platform with the potential to develop workflows for proteogenomic and metaproteomic analyses.

The next steps will continue to involve the consultation of biological researchers to help the team translate their informatics findings into basic biological contexts, and to aid projects which address human diseases. The team will also continue to develop visualisation tools that can help with the interpretation of outputted data.

There is also potential to add extra layers of omics to the analysis. So, for example, metabolomics could be included in the mix. Using this approach, the possibilities for new discoveries are endless.

Behind the Bench

Professor
Tim Griffin

Professor
Pratik Jagtap

E: tgriffin@umn.edu **E:** pjagtap@umn.edu **T:** +1 612 624 5249 **W:** http://galaxyp.org/ @usegalaxyp

Research Objectives
Drs Griffin and Jagtap's research focuses on the Galaxy-P project – developing, testing, optimising and applying multi-omics software tools to a variety of biological questions, including cancer and big data research.

Funding
- National Science Foundation (NSF)
- National Institutes of Health (NIH)

Collaborators
- Minnesota Supercomputing Institute
- Galaxy software platform developers
- Jetstream research computing resource

Bio
Professor Tim Griffin serves as the Principal Investigator on the Galaxy for proteomics (Galaxy-P) project, as well as the Faculty Director for the Center for Mass Spectrometry and Proteomics at the University of Minnesota.

Research Assistant Professor Pratik Jagtap has been the co-leader of the Galaxy-P project since its inception in 2012, helping to develop and apply software and workflows in metaproteomics, proteogenomics and more recently data-independent acquisition methods.

Contact
Dr Tim Griffin PhD
Professor and Director, Center for Mass Spectrometry and Proteomics
University of Minnesota
Dept. of Biochemistry, Molecular Biology and Biophysics
6-155 Jackson Hall
321 Church Street SE
Minneapolis, MN 55455
USA

Q&A

If your research were awarded a considerable amount of money and granted access to the world's most powerful computer – which informatics tool would you develop?

A tool that integrates outputs from all 'omics' platforms and provides a 'Google earth' like interactive visual data. Such a tool would be extremely useful to a biological researcher in both providing an overview of 'data landscape' for biological interpretation while providing opportunities to dive-in into regions of interest for validation and actionable intervention/follow-up. We continue to be amazed and fascinated by the depth of analyses that the Galaxy platform offers in challenging fields of research. Another avenue might be to use such a powerful compute platform to re-analyse existing publically available proteomic and transcriptomic datasets using newer multi-omic tools, and develop tools to mine for new discoveries.

What was the biggest challenge you had to overcome when developing Galaxy-P?

The development of tools and workflows for multi-omic analysis of mass spectrometry data provided challenges at many levels. Be it at the conceptualisation stage, or at grant seeking stage, or at tool selection or workflow stage, we looked at all the challenges as opportunities. Deciding which of the many effective software tools to implement in Galaxy has been a challenge, as well as understanding the many different 'omic sub-fields and how different software tools work, as well as which are at the forefront in terms of functionalities. However, the biggest challenge and priority in efforts has been to maintain the relevance of workflows in a constantly emerging environment where the inputs are diverse and outputs offer deeper and newer interpretations.

What is the most niche/unexpected dataset that you've been asked to analyse?

The breadth of biological research and flexibility of the Galaxy-P workflows has exposed us to many interesting datasets. These range from human salivary datasets for metaproteomics and proteogenomics, to dental plaque metaproteomes in presence of sugar to the study of metaproteomes from the North Pacific Oceans. But the most unexpected dataset has been the study of cardiomuscular protein expression in hibernation of ground squirrels. Human hearts lose the ability to function at temperatures of 20°C and below. The study tried to shed light on how the heart of hibernating animals can withstand these low temperatures. We are certain that we will continue to see more of these interesting datasets as we continue our research work.

In the future, do you see Galaxy-P becoming a desk-based tool that can easily and universally used by anyone, anywhere in the world?

The research community has been using Galaxy platform for genomics studies for quite some time now and there is a stable ecosystem of developers and users, which makes this sustainable. We have seen a gradual increase in interest in using Galaxy-P amongst researchers as we have promoted it via research publications, workshops and presentations worldwide. Along with the Galaxy community of developers and researchers, we have been working on making the workflows available via downloadable tool containers or by making public instances available so that researchers can access pre-installed tools and workflows for the research areas of their interest. The vision for the future is that researchers will access these software tools remotely, where they are housed on powerful cloud based hardware.

Leading on from this, do you think that younger students and early career researchers should be given compulsory bioinformatic training as part of their studies?

Absolutely! Bioinformatics has become a necessary research skillset for experimental researchers. Programming skills enable young researchers to perform novel analyses of previously acquired data. For users, analytical and data interpretation skills expand their ability to seek newer avenues in their research fields. We strongly believe that bioinformatics training will help in introducing and honing skills in programming and data processing, and helps in continuing to expand the breadth and depth of questions that can be sought by the future generation of scientists. 'Big Data' will only continue to be generated in biological research, and having the ability to speak both languages in terms of biology and computational science will be a critical skill, and one that is very much in demand in years to come.