Bioinformatics Needs Survey: Report

As part of TrainMALTA Work Package 2 (WP2; Introduction to bioinformatics and statistics for life scientists), a survey was circulated to assess the training needs of researchers within the consortium and tailor the future TrainMALTA training activities to their needs. The survey was designed to specifically assess the background knowledge of researchers across the three project partners (Universities of Malta - UoM, Leuven – KULeuven, and Cambridge), with a focus on UoM researchers and the areas of data analysis and interpretation. These included basic programming, data processing and visualisation, awareness of publicly available resources/tools and knowledge of statistical methods. The survey was created on the SurveyMonkey website and a link was sent out on 8 February 2016. The survey results were collected on 8 March 2016.

Fifty five researchers completed the survey, with the majority of them (46) from UoM (Figure 1). Fifty one of the responders expressed an interest in subscribing to the TrainMALTA mailing list so that they can be kept up-to-date with future events.

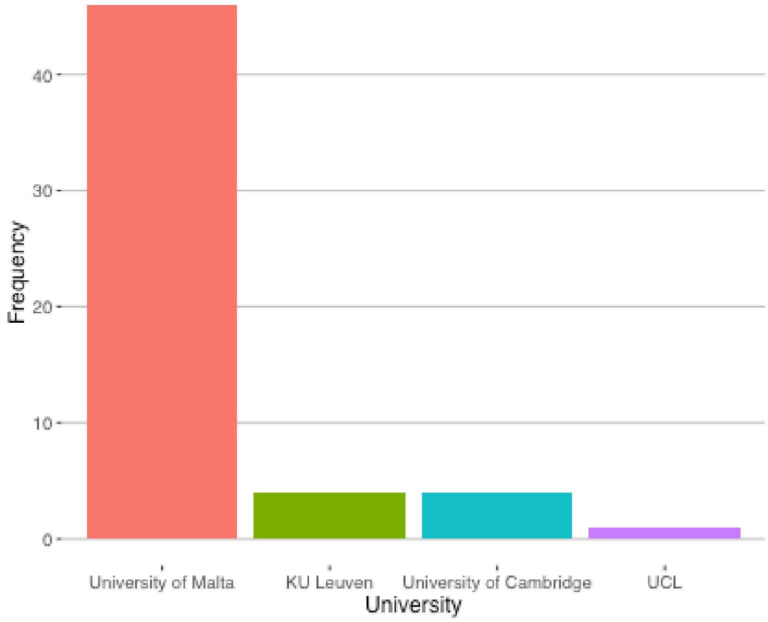

Figure 1: University affiliations of survey responders.

The survey revealed that researchers interested in TrainMALTA activities are at various levels of their academic careers (from undergraduate student to professor) and work in diverse research fields (Figures 2 and 3), with the majority of them interested in genomics/genetics and bioinformatics.

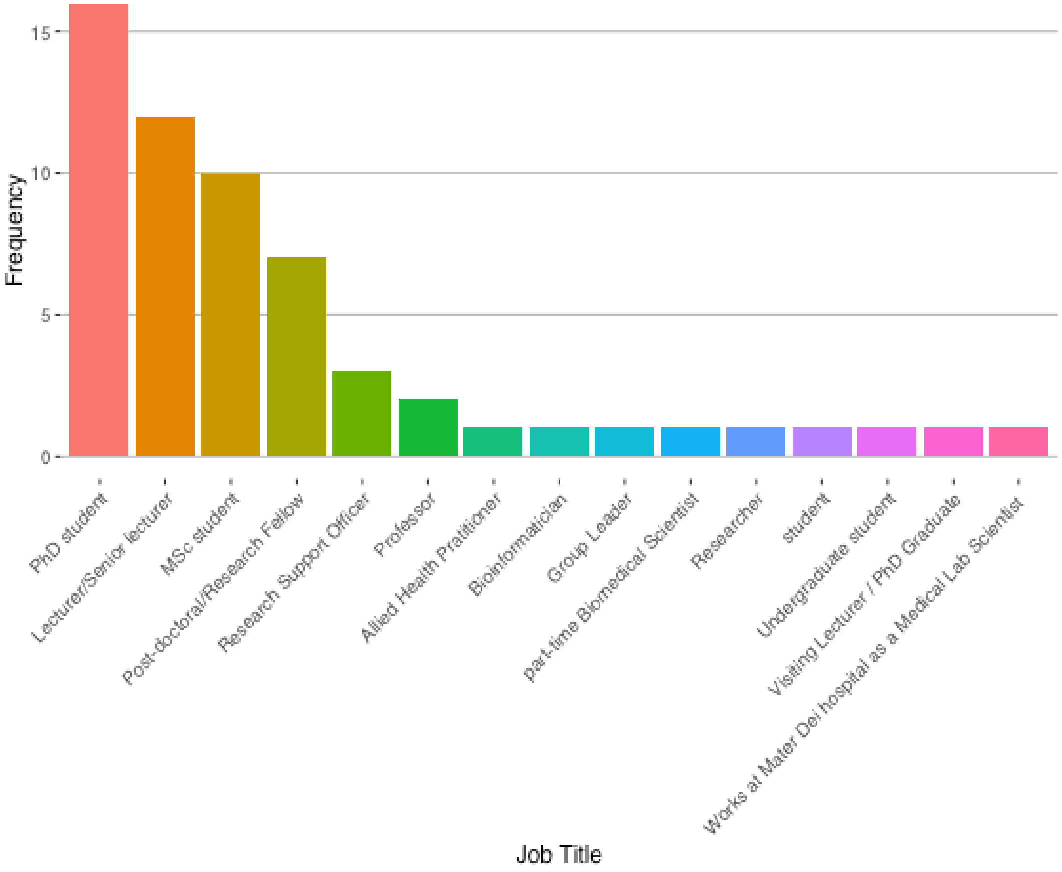

Figure 2: Academic positions/Job titles of survey responders (participants could select multiple options).

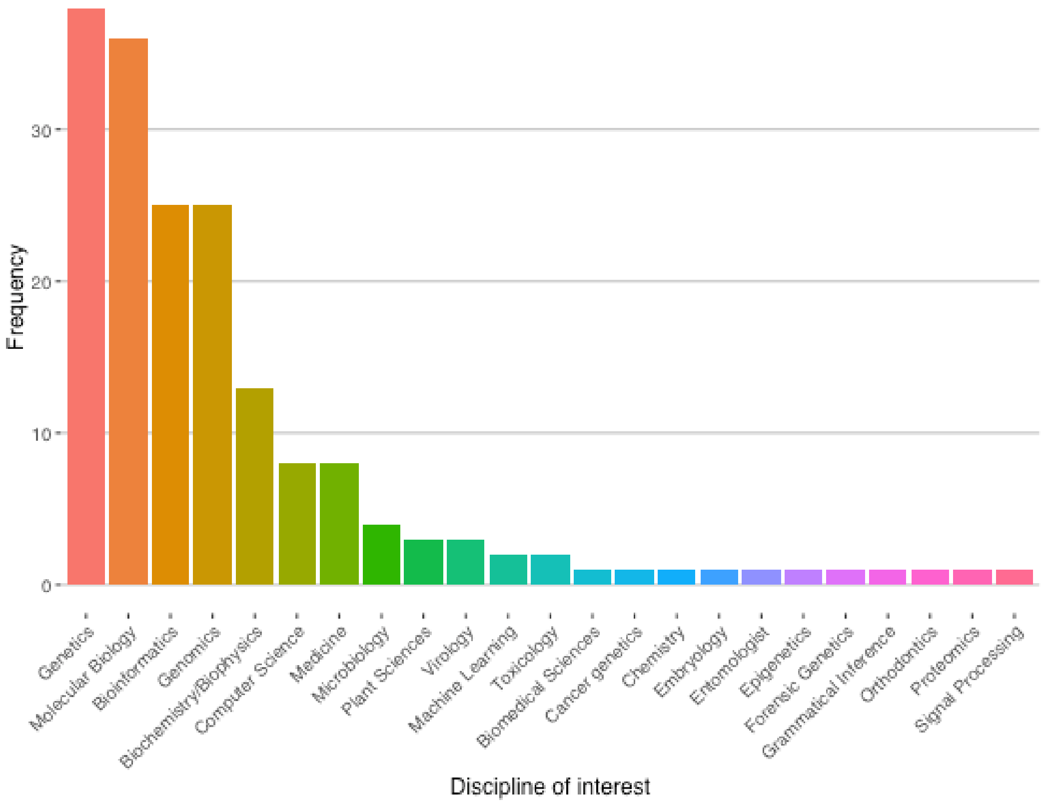

Figure 3: Discipline of interest (participants could select multiple options).

Part of TrainMALTA activities will focus on familiarising researchers with high-throughput sequencing (HTS) applications by providing training courses in both data generation and analysis. To better cater for the research community needs, we asked participants to describe their projects, indicate if they are already using any HTS applications and if so which ones, as well as discuss future plans for incorporating HTS applications in their research.

Responses revealed, as did previous answers, a largely diverse spectrum of research projects, 44% of which already include a HTS assay. Of those the most popular are exome, whole-genome or targeted (gene) sequencing (Figure 4 - Current). In addition a small fraction of the respondents works on projects unrelated to HTS, such as image & signal processing or structural bioinformatics.

The participants future research plans show a shift towards RNA sequencing (Figure 4 - Future), followed by exome and whole-genome sequencing. Some responders stated that they would like to better understand the opportunities offered by HTS to decide if these applications will be applicable to their research field. Finally some expressed interest on more general bioinformatics topics, such as public databases on GO terms and pathways as well as genomic variants.

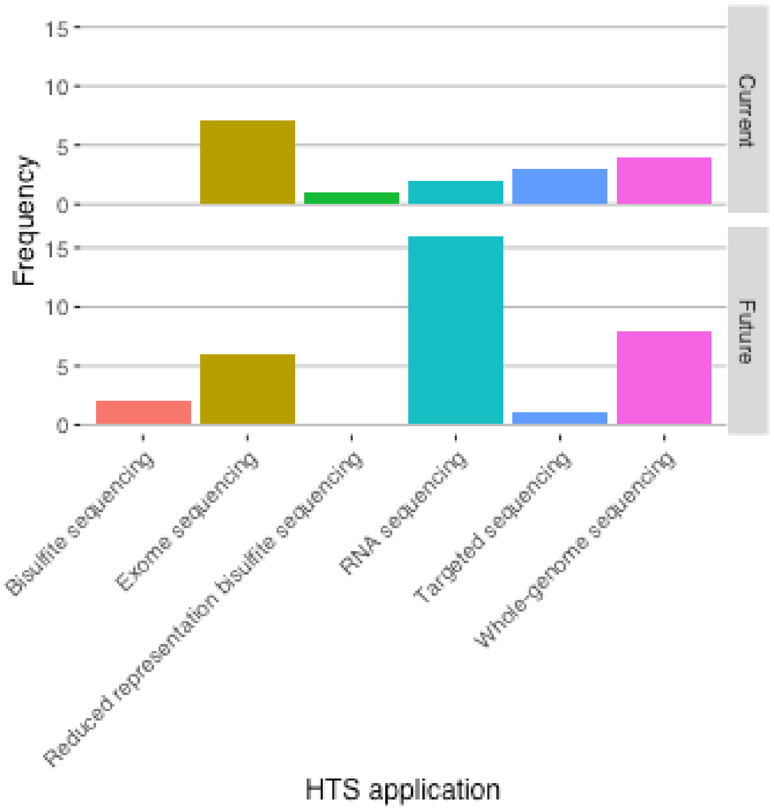

Figure 4: Current and future use of HTS applications by survey responders.

When discussing the tools they currently use, spreadsheets (e.g. Microsoft Excel) come on top of the list, followed by sequence alignment, similarity and homology tools. Examining what is used by the 44% of the participants that already use some HTS assays in their research, 10 use commercial data analysis software packages, none uses workflow tools (i.e. Galaxy), only 5 use HTS alignment or assembly tools and 8 have experience with data analysis environments (e.g. R, Matlab).

Finally, the assessment of the participants' confidence with computational and statistical tools revealed a lack of confidence with the UNIX environment and command-line tools for the majority, an equal split between 'confident' and 'not so confident' groups in regards to statistics and no familiarity or confidence with workflow tools, as expected based on the answers to the previous question.

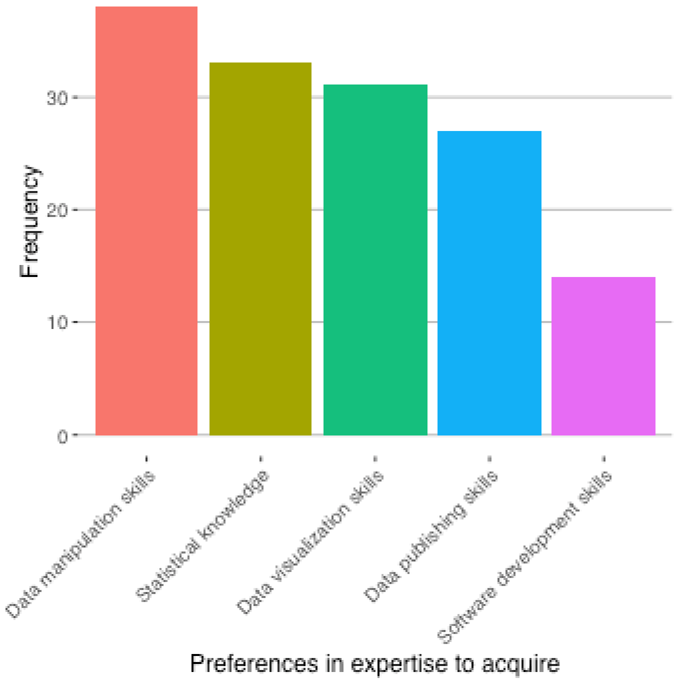

Figure 5: Expressed preference for training areas

The survey also highlighted the different areas in which training and expertise are sought (Figure 5). Taking into consideration these results, TrainMALTA has decided to offer the following courses and events in the first year.

- Our first networking event takes place at the end of April, where the focus will be on the formal presentation of the 3-year training project to the Maltese research community and the dissemination of the bioinformatic survey results. In addition we have arranged for a series of talks in order to showcase the expertise that KU Leuven and University of Cambridge bring into the project. The highlight will be the keynote talk by Dr Hana Lago Allen on precision medicine through her involvement with the Genomics England pilot project on rare immune diseases. The one-and-a-half day event will close with several shortlisted presentations dedicated to current and future research questions addressed by the Maltese participants.

- Regarding the training activities offered, the first year programme will kick off with a 2-day Data Carpentry workshop scheduled for 16th & 17th May 2016. This will allow participants to get a formal training on spreadsheets usage, the most popular tool used by our target audience according to the bioinformatics needs survey, best practices and a first feeling about the advantages of data analysis environments, such as R. To enable the highest participation possible, the workshop, delivered by Laurent Gatto, Romina Petersen and

Hugo Tavares, will be streamed live from Cambridge to Training Room 3, IT services building, University of Malta. Details of the workshop can be found at: http://www.training.cam.ac.uk/event/1572489.

- Through TrainMALTA, staff and students at UoM and KULeuven can also participate in the Courses offered by Bioinformatics Training. Details of upcoming courses may be found at: http://www.training.cam.ac.uk/programme/1572176 http://bioinfotraining.bio.cam.ac.uk/.

- The second half of the year will include the first summer school during which we will present certain computational concepts on which participants have expressed little to no confidence, such as UNIX scripting and command-line tools. We will address this by introducing some basic modules. We are also discussing the possibility of including a session on the use of Galaxy.

- Given that a lot of interest was shown in scientific databases and data manipulation we also aim to include an EBI roadshow session highlighting all available EBI online resources.

- Finally attention will also be given to the data publishing part through the delivery of a training course on best ways to present results and which tools to use to this direction.

A copy of the full survey results and of this report can be found on the TrainMALTA website www.um.edu.mt/project/trainmalta