

POSTDOCTORAL RESEARCH ASSISTANT – BIOINFORMATICS
ANALYSIS OF EUKARYOTIC GENE REGULATION

Grade 7 (Salary £28,839 – £31,513 pa)

Closing date for applications: 15th January 2010

FURTHER DETAILS

1. Introduction

This document gives further particulars on the BBSRC project to study alternative polyadenylation by next generation sequencing. The document provides some background on the groups involved and on the computational work that will be carried out by the bioinformatics Postdoctoral Research Assistant (RA).

2. The Project leaders

This is a close collaboration between Geoff Barton's Bioinformatics research group and the group of Gordon Simpson.

Geoff Barton (GJB) is Professor of Bioinformatics and co-director of the Post-Genomics and Molecular Interactions Centre at the University of Dundee College of Life Sciences. GJB has 20 years of research experience in computational molecular biology, with particular emphasis on protein sequence and structure analysis and prediction. His work has included both technical innovations in algorithm development as well as application of these techniques in collaboration with bench scientists. GJB has extensive experience and understanding of the needs of large-scale data management in biology having previously been head of the European Macromolecular Structure Database at the EMBL European Bioinformatics Institute. He has published over 80 peer reviewed papers with >5,900 combined citations and currently holds grants from BBSRC, MRC, Wellcome Trust, EC and SFC. Recent publications include collaborative work on annotation of the protein kinase complement for newly sequenced genomes (two Science publications) as well as the development of novel techniques for the prediction of protein-protein interactions. In addition to conventional publications, GJB's group at UoD provides software and databases that are widely used by the scientific community either through web-access or downloads. For example, the JPred/Jnet secondary structure prediction server and the Jalview multiple alignment analysis tool that are each accessed over 2,000 times per week by researchers in more than 50 different countries. Over the last 2 years GJB's group has collaborated with other UoD PIs on the analysis of Solexa sequence data from human RNA. See www.compbio.dundee.ac.uk for more information and a full publication list.

Gordon Simpson (GGS) is a Principal Investigator and Lecturer in the Division of Plant Sciences at the University of Dundee (UoD) College of Life Sciences. GGS has over 15 years research experience focused on RNA processing and flowering time control. He was the first to clone genes encoding plant splicing factors (EMBO J) and the first to localise them to plant Cajal (coiled) bodies. He revealed that UBP1 could mediate the enhancement of splicing dependent on U-rich sequences characteristic of plant introns (EMBO J). He cloned the flowering time gene, FY, and dissected the functional significance of alternative processing of FCA pre-mRNA, resulting in high-profile publications in Cell and EMBO J. He has published highly cited reviews on pre-mRNA splicing in plants and on flowering time control (Science & Ann Rev Cell & Dev Biol). GGS established his own lab at UoD in July 2004. With funding from the Scottish government and BBSRC, his lab has focused on characterising the role of the RNA binding protein, FPA, in flowering time control and isolating new mutants affecting flowering time and inflorescence

development in cereals. GGSs' lab successfully established a method to cross-link Arabidopsis RNA binding proteins to their target RNAs in vivo - a method known as RNA immunoprecipitation (RIP). Having previously discovered that the RNA binding protein, FCA worked with the pA factor, FY, to control the site of 3' end formation, his lab has now discovered that FPA also controls 3' end formation, which is the basis of this project.

3. Introduction to the project for non-specialists

Our genes are made of DNA, but when they are switched on, copies are made in a related molecule called RNA and this RNA goes on to code for the protein products of our genes. As the gene is copied into RNA, the RNA is cut and a string of Adenine molecules (A for short) are added at the end. This so-called "poly A tail" functions to protect the RNA from being degraded, and helps to transport the RNA around the cell and stimulates the formation of protein from the RNA. The site at which the poly A tail is added is not always the same, even for the same gene. For example, half of all human genes have RNAs with more than one site for adding a poly A tail. Controlling the site at which the poly A tail is added is very important because it ultimately affects how genes function. However, this is a process we know surprisingly little about. It's not just human RNAs that have different poly A tails, other animals and plants do too.

We have been studying how plants control the time at which they flower, a process where genes are very precisely controlled. In the course of this work, we have discovered that three factors called FCA, FY and, most recently, FPA, function to control poly A site selection of some RNAs. Such basic aspects of gene expression are very similar in plants and animals and it turns out that there are human proteins highly related to FY and FPA. It is possible therefore, that these proteins control poly A site selection in humans too, but very little is known about them. As we have found that FCA and FPA don't need each other to control poly A site choice, we think they must be doing this in different ways. This gives us a chance to understand how poly A site choice can be controlled.

In this project we will build on what we know about FCA and FPA in plants, but this knowledge should be of much more general interest. We want to know two things: (1) How do FCA and FPA control the site at which a poly A tail is added (2) What genes do FCA and FPA regulate by controlling alternative poly A site choice? We will work out how FCA and FPA control poly A sites by identifying the features of the RNA required. This should be quite straightforward. We will make test genes containing different parts of the target gene and see how they affect poly A site selection when placed back in plants. In order to find the other genes whose normal poly A tail depends on FCA and FPA, we will look at where RNAs are polyadenylated in normal plants and in mutant plants that lack FCA or FPA. It is now possible for us to look at nearly all the RNAs in a cell thanks to Next Generation Sequencing, a technology that is revolutionizing modern biology by giving us huge amounts of sequence data, very quickly and at a fraction of the cost to before. This technology has been developed to look at RNA by sequencing a short part of every RNA, sufficient to identify it, called a "tag". To find the tag, we use the poly A tail and sequence what is next to it. This is a happy coincidence for us, because it means that in addition to tagging a particular RNA, this method also tells us where a poly A tail has been added to RNA. To analyse the large amounts of data and make comparisons, we will need to develop specialized computational tools. Because we already know genes where FCA and FPA control poly A site selection, we should be able to find changes in these "tags" if our tools are working well. Once we are sure they are, we can look for other shifts in "tags" to identify other genes controlled by FCA and FPA. As lots of other scientists are also using this sequencing technology, but for completely different reasons, we can use our analysis tools to look at changes in polyadenylation in their data too. In this way we will be able to identify cell-types and situations where alternative polyadenylation is an important part of gene regulation.

4. A more technical summary

Alternative polyadenylation (pA) is a commonplace, but surprisingly poorly characterised aspect of eukaryotic gene regulation. We have discovered that the Arabidopsis RNA binding proteins, FCA

and FPA (regulators of flowering and RNA silencing) function genetically independently to control the site of RNA 3' end formation. The work in Gordon's lab is unique because no other trans-acting regulators of RNA 3' end formation, that are not components of the splicing or polyadenylation machinery, have been identified. His discovery therefore provides a genetically tractable system to dissect alternative pA. We will first define the level at which FCA and FPA control pA site selection through cis-element analysis, which we will couple with PolIII ChIP and transcription run-on assays. A combination of RNA immunoprecipitation (which we have successfully developed) and RNA specificity-swap assays will be used to determine how directly these proteins regulate 3' end formation. Next, a genome-wide identification of pA sites regulated by FCA and FPA will be made by utilising recent developments in next generation sequencing: Fortunately for us, Digital Transcriptomics (DT) involves sequencing short "tags" of RNA adjacent to pA tails, thereby providing positional information on pA sites. We will develop a bioinformatics pipeline to analyse DT data to quantify changes in pA site selection, working first with different FCA and FPA genetic backgrounds that we know have contrasting patterns of pA. With these tools in place, we will be able analyse other DT data releases and associate alternative pA with diverse backgrounds, cell types or treatments. Our work will have widespread impacts in understanding gene regulation because it will define mechanisms by which alternative pA can be controlled, clarify the connections between 3' end formation and RNA silencing and establish generic bioinformatics tools to identify alternative pA in next generation sequencing data.

5. Outline of computational work required

The following paragraphs are a modified extract from the research proposal. The "RA" is the bioinformatics RA that we are seeking in this job advertisement.

Our objectives are to develop the necessary database, software and statistical analysis procedures to identify and quantify changes in RNA 3' end formation. A secondary objective is to make our data, and generic tools and techniques widely available by establishing a web resource for alternative pA. This resource will initially focus on the experimental findings generated in this project for *Arabidopsis*, here, but our aim is to make it flexible enough to accommodate data of this type from any species. A huge asset in our analysis is the existence of positive internal controls in our RNA samples. We already know that pA site selection at FPA and FCA is affected in the backgrounds under study, so this change in FPA and FCA will provide the baseline for our analysis. Over the last 2 years, we (GJB lab) have gained experience of Solexa sequence analysis in collaboration with a number of groups. For example, our recent work based on a single lane Solexa run (3×10^6 reads) with Dr Gyorgy Hutvagner, an expert in miRNA analysis, has led to the discovery of novel RNA processing in human cells (Cole et al, *RNA*, 2009). In order to carry out these studies, we have established procedures for consistent quality filtering and clipping of reads as well as developing a relational database implemented in MySQL to store, organise and cross-reference the reads. This experience will give the RA on the current project a strong foundation on which to build a system appropriate to alternative pA analysis. Each Solexa flow cell or slide has 7 project lanes and with current technology, each lane produces approximately 5×10^6 reads giving 3.5×10^7 reads per slide. In this proposal we will be analysing full slide Solexa runs from 3 biological replicates of 8 strains from 2 different sample types giving a total of 16 Solexa runs of 7 lanes each. This will give a total of at least 6.3×10^8 reads generated for this project. In fact, the numbers will be higher, as recent developments to the technology will allow around double the number of reads and reads that are significantly longer.

Since genome annotation is imperfect and high-throughput sequencing allows rarer transcripts to be identified than previously possible, the most straightforward method to identify the source of each read is to compare reads directly to the *A. thaliana* genome. For this task, we have recently migrated from Vmatch to the ultra-fast memory-efficient short read aligner Bowtie for most short-read versus genome matching problems. Our own benchmarks suggest that Bowtie will allow approx. 7.5×10^5 reads per hour to be compared to the *A. thaliana* genome on a single CPU when allowing for mismatches. Thus, the results from each Solexa slide will take on the order of 47 hours to scan against the genome on a single CPU, or around 6 hours on an 8 CPU cluster node. Reads from Solexa technology degrade as the read length increases. Sequences that have few

errors in the 5' end, but accumulate errors towards the 3' end lead to fast sequence matching methods (such as Bowtie) failing to find matches against the genome. In order to minimise this problem, in our previous work we clipped reads based on a moving-window filter of the quality scores reported by the Illumina Eland software pipeline. The clipping was optimised to obtain a balance between the read length remaining and the number of orphan reads (those that did not match to the genome). In this project a similar procedure will be followed. However, since we are also interested to identify alternative pA sites in previously uncharacterised genes, we will experiment with a two-step clipping strategy. Firstly, we will use aggressive quality clipping to identify the genome location unambiguously, then relax this to identify where possible, the start of pA. Alternative strategies will also be investigated in order to identify the most effective method for these data.

The core of the bioinformatics analysis lies with classification of genes with expression correlated with FPA/FCA alternative pA and will be open-ended since it will only be once the data are available that the number of possible avenues for research will become apparent. While the detailed procedures will be developed following initial analysis, the broad steps are as follows: Following quality filtering and genome scanning, our database will enable us to quantify the number of unique reads to each location of the genome. Reference to the annotated *A. thaliana* genome will classify these reads by gene, or as unidentified locations in the genome. Given this initial screening and linking of our DT (Digital transcriptomics) results with the known gene locations we will be able to identify which genes are represented by alternative pA. Reference to the FPA/FCA genes will also enable the relative levels of expression to be established within the context of our DT experiments. Given this baseline we will be able to identify which genes follow correlated expression patterns with the FPA/FCA genes and further, identify which genes have correlations in alternative pA levels. Technically we will achieve this analysis by a combination of direct SQL database queries and analysis by code in the R-programming statistical language. These are tools that we exploit routinely in our laboratory (GJB) and we feel are most appropriate to this problem.

This project will generate significant new information on pA in *A. thaliana* and a key result of our work will be to make this generally available, not just to our own laboratory but world-wide. For basic annotation of new pA sites this is likely to be best served by establishing a DAS server (Distributed Annotation Service) for the newly identified sites. A DAS server allows a DAS client (e.g. the Ensembl genome browser) to add the annotation to its view of the genome. Thus, a user can see the new annotation in context with all other annotation available for that genome. The *A. thaliana* genome is being maintained in Ensembl by the NASC and the Ensembl genome browser makes extensive use of DAS to present annotations so it should prove feasible to publish our new observations by this method. The genome browser will provide a valuable way to analyse and present alternative pA as a regulator of *A. thaliana* gene expression. Our aim is for this browser and the analysis pipeline developed in this study to be able to deal with other data sources. DT is likely to be adopted more widely as *Arabidopsis* researchers use it as a substitute for microarrays. These analyses will likely involve different developmental stages, stress treatments and different genetic backgrounds, but we could reanalyse their data to provide quantitative analysis of changes in 3' end formation. In the future, this analysis could accommodate changes in alternative pA in other organisms. With a growing focus on regulated 3' end formation in GGS's lab, we have started to investigate the role that human proteins related to Arabidopsis FY and FPA might play in 3' end formation. This coupled with the experience of GJB's group on developing bioinformatics resources for a wide-range of organisms, from protozoa to plants and metazoans, means that we are in a position to develop such a site for 3' end processing in general. An important output then would be a web presence for a focus on genome-wide RNA 3' end formation. Despite the frequency and importance of alternative pA in gene expression control, no such site currently exists.

Changes in pA site selection identified after bioinformatics analysis will be validated by the molecular biologist RA. With our experience to date, this can involve the identification of previously unannotated RNAs. We will use 5' and 3' RACE kits to define the RNAs we see change. 3' RACE will be used to precisely identify pA sites. We will use Northern analysis of polyA+ RNA to detect

full-length pA RNAs. Where the RNAs are weakly expressed, we will use RT-PCR and RNase AT1 to detect changes in pA site selection.

6. Computer Hardware Facilities

The bioinformatics group at Dundee has state-of-the-art computing facilities for research. Computational power is provided by a dedicated 500 core Linux cluster managed by Sun Grid Engine with a range of queuing options that permit both true parallel and farm-type computing. Individual cluster nodes have a minimum of 16GBytes RAM with many on 32GBytes. A proportion of the cluster has infiniband connections for optimum performance in parallel applications. Structured data storage is provided by dedicated MySQL and Oracle servers which give flexibility in database application options and allow members of the group to exploit the most appropriate technology for each research problem. The cluster, general-purpose computing and databases are supported by high-speed disk arrays connected over a fiberchannel network. Most importantly, this core infrastructure is maintained and supported by IT professionals who work closely with the group to match the hardware to the specific demands of bioinformatics research.

7. Working Environment

The bioinformatics group enjoys excellent office space with good opportunities for face-to-face interactions with colleagues and communication with other scientists throughout the College. The group has weekly meetings at which members of the group present recent results, or review a paper or technical subject. We also encourage and facilitate informal contact and meetings and to aid this we maintain an active group mailing list and Wiki that act as conduits for scientific discussion as well as social chat and planning. We also have strong links and a shared seminar programme with the Division of Mathematics with its strong research grouping in mathematical modeling of biological systems. Accordingly, the person appointed to the position advertised here will benefit from a supportive network of colleagues with deep knowledge in many areas of computing, bioinformatics, statistics and mathematics.

8. The Job Opportunity

This project is a unique opportunity for a scientist with strong computing and data analysis skills to develop experience of working with next generation sequence data on a tractable biological system. It is highly likely there will be significant scientific outputs which will help the appointed bioinformatics RA in the development of their scientific career.

Unlike many bioinformatics positions, this job will allow the successful candidate to interact daily with colleagues both in bioinformatics and in molecular biology and will expect them to take a significant role in developing the scientific outcomes of the project. There will be a tight cycle of wet and dry research carried out throughout the project, so there will be ample opportunities to make impact on the project's success at the interface between computational and "wet-lab" experimental research. In addition, there will be ample opportunities to create novel bioinformatics resources that will be publishable in their own right.

The position requires the dedicated efforts of individual who has demonstrated experience of analysing large biological datasets with their own computer code and databases, or has developed these skills in a different scientific discipline, but who is highly motivated to make the transition into biological/bioinformatics research. As a consequence, we are looking for someone with the following profile:

Essential: Flexibility and demonstrated ability to learn and apply new computational tools and techniques as appropriate.

Essential experience: 3-years minimum demonstrated programming/development experience with at least one of – Perl, Python, Java, C/C++ in a Unix/Linux environment; Knowledge of statistics.

Desirable experience: Familiarity with molecular biology/structural biology concepts and principles; Knowledge of R-programming; SQL database design and development; Knowledge of standard bioinformatics tools, algorithms and techniques.

While data interpretation will undoubtedly require the development of novel algorithms and software, algorithm development and testing is not the primary focus of the project.

Further information

Informal enquiries about this position may be directed by email to Prof. Geoff Barton (g.j.barton@dundee.ac.uk). Please clearly identify your email with the words: "BBSRC NGS RA" to help me spot it in the flood of daily email. However, please do not make full applications to this email address. See the instructions on how to apply in the accompanying job advert.