**OVERVIEW**

**Keywords: pseudogene, annotation, plant, transcriptomics, cloud computing** Understanding plant genomes is key to boosting food production and generating clean energy and new drugs. The past decade has seen a burst of large-scale plant sequencing projects. While hundreds of plant genomes have been sequenced their content is yet to be annotated and characterised. Thus, a fundamental challenge in bioinformatics is to provide accurate, and reliable resources for their genome-wide analysis. In this project we focus on a systematic identification and analysis of pseudogenes in representative plant species. We aim to build cloud deployable pipelines that given a genome sequence can detect and characterise pseudogene sequences. Furthermore, by integrating data on evolutionary, functional, and phenotypic levels our computational resources will provide a detailed picture of pseudogenes’ role in plant biology.

More specifically, our first aim is dedicated to the development of a comprehensive annotation pipeline prototype that uses protein and transcript information to identify processed, unitary, and duplicated pseudogenes. Furthermore, we will use homology searches and global alignment methods to annotate specific disablements that resulted in loss or gain of function events.

Our second aim will provide a biological activity description of pseudogenes by integrating annotations with data from functional genomics, phenotype ontologies, and evolutionary analyses. For this, we propose to prototype the development of specialised and stand-alone modules that mine the available data to characterise pseudogene activity and potentially infer functional roles.

Finally, our third aim will synergistically combine the results obtained from aims 1 and 2 into a comprehensive database for plant pseudogenes, as well as industrializing the prototype pipelines and deploying them online. All the tools and resources will be made available as stand-alone and cloud compatible packages and will address a gap of knowledge in the current plant pseudogene research.

The project will be developed as a bilateral collaboration between the groups of Dr Mark Gerstein (US NSF PI) at Yale University and Dr Cristina Sisu (UK BBSRC PI) at Brunel University London. The two PIs have a long history of successful collaborations on many pseudogene annotation and analysis projects for human and a variety of model organisms.

**INTELECTUAL MERIT**

The proposal’s intellectual merit lies in its systematic and robust approach for the identification and annotation of pseudogenes in plant genomes. Furthermore, we will develop a second pipeline to integrate data on functional, evolutionary, and phenotypic level to provide a comprehensive characterization of pseudogenes. The two independent pipelines bring valuable contributions to the fields of genome annotation and plant science. Another key aspect of the proposal is the development of robust freely available cloud compatible computational resources and an online database to inform the identification and characterization of plant pseudogenes. In particular, the proposed database will focus on key model plant organisms such as Arabidopsis, essential crops such as maize, rice, wheat, soybean, and red clover, and economically important plants such as poplar. Moreover, we will use available gold standard datasets and develop *“in-silico”* experiments to validate and improve our proposed computational methods.

**BROADER IMPACTS**

The proposed work will lead to impact beyond the genome sequencing community – in particular, in the field of plant biology. In recent years, numerous large-scale plant genome sequencing projects have come to fruition, providing a need for large-scale system-level annotation and characterization of pseudogenes. The. project will shed light on gene regulation in plants by highlighting pseudogenes’ regulatory capacity through potential siRNA transcription. Our approach to develop specialised pipelines for plant pseudogene analysis will also have direct relevance for functional genomics community. Moreover, the proposed methods will be universally applicable to a large variety of plant genomes not only the high-quality model organisms. They can be used to provide a fast screening for disabled sequences and capture key loss and gain of function events. Finally, we outline our plan to use the developed tools as devices for education in bioinformatics, as well as our strategy to create workshops dedicated to computer scientists and experimental biologists that would like to learn more about *“in-silico”* analysis of biological data.