

# A Computational Platform for Gene Expression Analysis

Diogo Teixeira

Supervisors: Rui Camacho<sup>1</sup>, Nuno Fonseca<sup>2</sup>

<sup>1</sup>LIAAD INESC, Porto & DEI FEUP, Universidade do Porto, Porto

<sup>2</sup>EMBL-EBI, Cambridge, UK

July 2014

- 1 Introduction
  - Domain Problem
  - Motivation and Objectives
- 2 Developed Solution
  - Overview
  - RNA-Seq Analysis Pipeline
  - RBP Analysis Pipeline (PBS Finder)
  - Integration
- 3 Case Studies
  - RNA-Seq Analysis Pipeline
  - RBP Analysis Pipeline (PBS Finder)
- 4 Conclusions
  - Objective Fulfilment
  - Future Work

# Domain Problem I

## Introduction

- Molecular biology is a young field of study, with a lot of unknowns and partial knowledge.
- Studying gene expression is crucial to understand the mechanisms that control living organisms.
- We focused on two different areas:
  - differential expression analysis;
  - RNA-binding protein (RBP) discovery and analysis.

# Domain Problem II

## Introduction

Three distinct problems:

- Read alignment against a reference genome and differential expression analysis on the aligned data.
- RBP discovery, analysis and information enrichment.
- Further result analysis using data mining techniques.

# Motivation and Objectives

## Introduction

### **Tools are complex**

Tools for biological data analysis  
often require a very technical set of  
skills.

# Motivation and Objectives

## Introduction

### **Tools are complex**

Tools for biological data analysis often require a very technical set of skills.

### **Tasks are repetitive**

Analysing high quantities of data can be repetitive, especially if executed manually.

# Motivation and Objectives

## Introduction

### **Tools are complex**

Tools for biological data analysis often require a very technical set of skills.

### **Tasks are repetitive**

Analysing high quantities of data can be repetitive, especially if executed manually.

### **Information is scattered**

Information is easy to acquire, but is often scattered through multiple platforms, services and institutions.

# Motivation and Objectives

## Introduction

### **Tools are complex**

Tools for biological data analysis often require a very technical set of skills.

### **Create simpler tools**

Any user should be able to use the tools, with little to no training.

### **Tasks are repetitive**

Analysing high quantities of data can be repetitive, especially if executed manually.

### **Information is scattered**

Information is easy to acquire, but is often scattered through multiple platforms, services and institutions.



# Motivation and Objectives

## Introduction

### **Tools are complex**

Tools for biological data analysis often require a very technical set of skills.

### **Create simpler tools**

Any user should be able to use the tools, with little to no training.

### **Tasks are repetitive**

Analysing high quantities of data can be repetitive, especially if executed manually.

### **Automate tasks**

Automated systems should perform repetitive tasks, so that users can focus on their work.

### **Information is scattered**

Information is easy to acquire, but is often scattered through multiple platforms, services and institutions.

# Motivation and Objectives

## Introduction

### **Tools are complex**

Tools for biological data analysis often require a very technical set of skills.

### **Create simpler tools**

Any user should be able to use the tools, with little to no training.

### **Tasks are repetitive**

Analysing high quantities of data can be repetitive, especially if executed manually.

### **Automate tasks**

Automated systems should perform repetitive tasks, so that users can focus on their work.

### **Information is scattered**

Information is easy to acquire, but is often scattered through multiple platforms, services and institutions.

### **Gather information**

Information should be contextually aggregated, allowing for quick access of relevant information.

# Overview

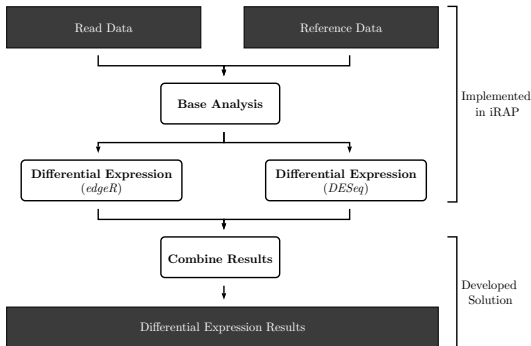
## Developed Solution

- Two distinct problems warrant two different solutions.
- The developed system should be available anywhere, through the internet.
- The system should be as modular as possible, to allow future extensions.

# RNA-Seq Analysis Pipeline I

## Developed Solution

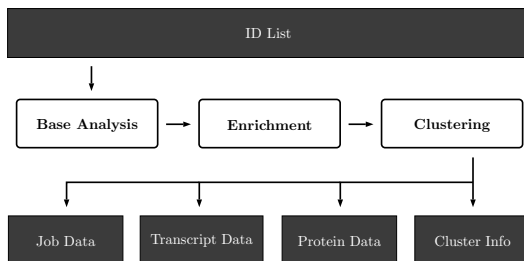
- Uses iRAP as the analysis pipeline.
- Conducts multiple differential expression analyses with different tools.
- Combines results from multiple tools.



# RBP Analysis Pipeline (PBS Finder) I

## Developed Solution

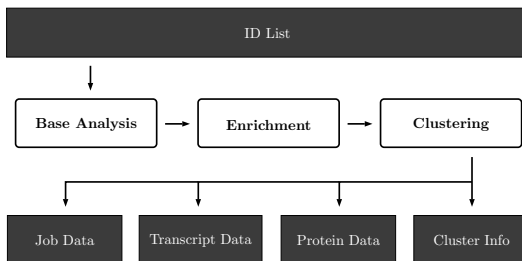
- Uses Ensembl and NCBI to identify gene species, obtain basic information and extract genetic sequences (5' UTR, 3' UTR, 3' UTR downstream).



# RBP Analysis Pipeline (PBS Finder) II

## Developed Solution

- Uses RBPDB to discovery RNA binding proteins based on the obtained sequences.
- Uses UniProt to enrich the obtained results and performs clustering analysis on those results.



# RBP Analysis Pipeline (PBS Finder) III

## Developed Solution

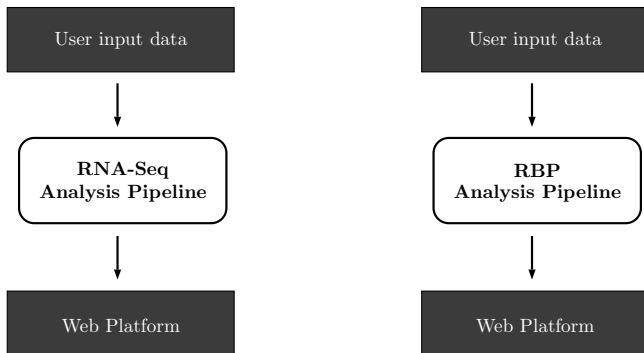
Clustering analysis:

- Uses  $k$ -medoids and hierarchical clustering, both with Jaccard and binary distance matrices.
- Executes every possible combination of clustering setups (alternates algorithms, distance matrices, used features, etc.).
- Results are filtered (acceptable solutions must have a minimum percentage of entries per cluster, clusters must have defining features, etc.).
- Solution quality internally determined based on the average silhouette.

# Integration

## Developed Solution

While focusing on aggregation and quick access to information, does it make sense to separate the results into two different platforms?

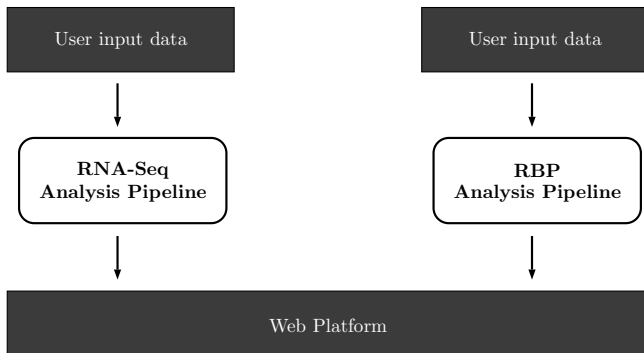




# Integration

## Developed Solution

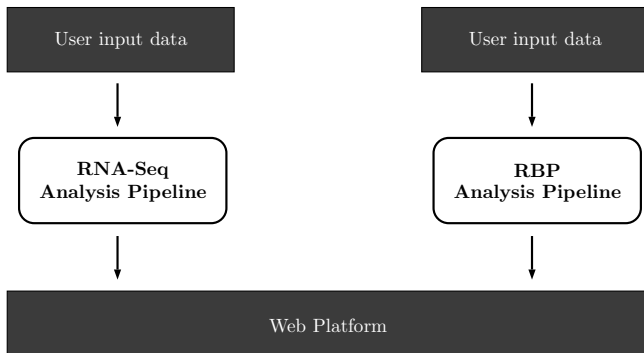
While focusing on aggregation and quick access to information, does it make sense to separate the results into two different platforms?



# Integration

## Developed Solution

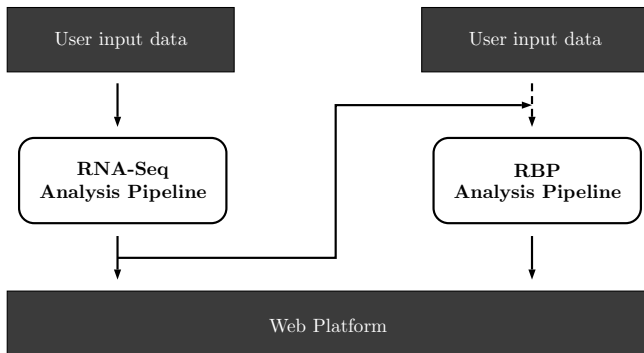
A list of differentially expressed genes is not very useful without further information about those genes. Does it make sense for a user to launch a new gene enrichment task by hand?



# Integration

## Developed Solution

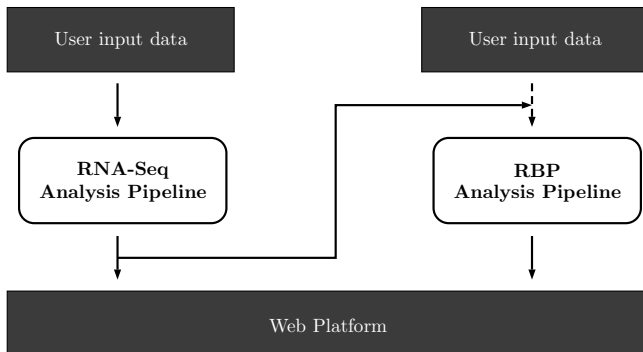
A list of differentially expressed genes is not very useful without further information about those genes. Does it make sense for a user to launch a new gene enrichment task by hand?



# Integration

## Developed Solution

A fully integrated solution: the analysis pipelines can be used separately or automatically executed in sequence; result visualization for both pipelines is isolated.



# RNA-Seq Analysis Pipeline I

## Case Studies

### Objective

- Ascertain if combining the results of multiple tools has impact on the set of differentially expressed genes.

### Data set

- Reproduction of ArrayExpress experiment E-GEOD-48829 (*Escherichia coli*).
- Reference genome obtained from Ensembl Genomes and read data obtained from ENA Sequence Read Archive.

# RNA-Seq Analysis Pipeline II

## Case Studies

Results (number of differentially expressed genes)

	<i>Raw results</i>	<i>Filtered results</i>	<i>Combined results</i>
<i>edgeR</i>	4494	386	191
<i>DESeq</i>	4494	204	

## Conclusions

- Combining results impacts the final differentially expressed gene list by reducing its size.
- The combined results will hopefully give researchers an higher confidence in the experimental results.

# RBP Analysis Pipeline (PBS Finder) I

## Case Studies

### Objectives

- Assess the general usefulness of PBS Finder.
- Compare PBS Finder with the existing techniques of manual analysis.
- Assess the impact of differences in hardware performance in the overall performance of the platform.

### Data set

- 23 genes from the *RhoGTPase* family (*Rattus norvegicus*) provided by IBMC.

# RBP Analysis Pipeline (PBS Finder) II

## Case Studies

Results (expert estimation of 30 minutes per gene analysed)

<i>Number of IDs</i>	<i>Machine1</i>	<i>Machine2</i>	<i>Manual method</i>
100	9m 56s	11m 1s	$\approx 50h$
500	41m 47s	55m 51s	$\approx 250h$
900	1h 33m 32s	2h 7m 4s	$\approx 450h$

## Conclusions

- PBS Finder can reproduce the results an expert would get.
- Months worth of an expert's manual work can be accomplished in a few hours.
- While hardware performance has a significant impact on analysis time, the platform achieves satisfactory performance on personal computer-level hardware.

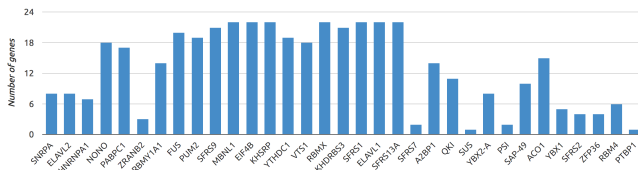


# RBP Analysis Pipeline (PBS Finder) III

## Case Studies

### Results (viewed in PBS Finder)

RBP FREQUENCY



Function	Proteins	Species	Gene ( <a href="#">view original</a> )	Transcript	Protein	SNRPA	ELAVL2	HNRNPA1	NONO	PABPC1	ZKSCAN2	RBMY1A1	FUS	PUM2	SFRS9	MBNL1	EIF4B
		Rattus norvegicus (Rat)	ENSRNOG00000013536 (Cdc42)	ENSRNOT00000029025 (Cdc42-201)	Q8CFN2 (CDC42)												
		Rattus norvegicus (Rat)	ENSRNOG00000013536 (Cdc42)	ENSRNOT00000018118 (Cdc42-202)	Q8CFN2 (CDC42)												
		Rattus norvegicus (Rat)	ENSRNOG00000021919 (Rhoj)	ENSRNOT00000031979 (Rhoj-201)	Q5RJS2 (RHQJ)												
		Rattus norvegicus (Rat)	ENSRNOG00000015415 (Rhoq)	ENSRNOT00000020822 (Rhoq-201)	Q9JUL4 (RHOQ)												
		Rattus norvegicus (Rat)	ENSRNOG00000020393 (Rhog)	ENSRNOT00000027641 (Rhog-201)	Q32PX6 (RHOG)												

# Objective Fulfilment

## Conclusions

- RBP analysis pipeline and web platform (PBS Finder) implemented and tested. PBS Finder has been in production for several months; during this time it was thoroughly tested by IBMC experts.
- RNA-Seq analysis pipeline implemented and tested (iRAP deployed and result consolidation tool implemented).
- Integration of both tools could not be accomplished.

- Fully integrate the RNA-Seq analysis pipeline with the web platform (automatic job configuration, result visualization, etc.).
- Study the requirements for deploying the platform in large scale, and assess the feasibility of making it available internet-wide.

- 1 Introduction
  - Domain Problem
  - Motivation and Objectives
- 2 Developed Solution
  - Overview
  - RNA-Seq Analysis Pipeline
  - RBP Analysis Pipeline (PBS Finder)
  - Integration
- 3 Case Studies
  - RNA-Seq Analysis Pipeline
  - RBP Analysis Pipeline (PBS Finder)
- 4 Conclusions
  - Objective Fulfilment
  - Future Work

# A Computational Platform for Gene Expression Analysis

Diogo Teixeira

Supervisors: Rui Camacho<sup>1</sup>, Nuno Fonseca<sup>2</sup>

<sup>1</sup>LIAAD INESC, Porto & DEI FEUP, Universidade do Porto, Porto

<sup>2</sup>EMBL-EBI, Cambridge, UK

July 2014