Development of Bioinformatics Resources for the Integrative Analysis of Next Generation Omics Data

Rafael Hernández de Diego

PhD Thesis

Supervisor

Dr. Ana Conesa Cegarra

July 2017

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Development of Bioinformatics Resources for the Integrative Analysis of Next Generation Omics Data

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Rafael Hernández de Diego

Supervisor:
Dr. Ana Conesa Cegarra
Genomics of Gene Expression Lab
Centro de Investigación Príncipe Felipe

Tutor:
Dr. Monserrat Robles Viejo
Instituto Universitario de Tecnologías de la Información y Comunicaciones (ITACA)
Universitat Politècnica de València

A doctoral thesis submitted to
Instituto Universitario de Tecnologías de la Información y Comunicaciones (ITACA)

July 2017

Abstract

Advances in high-throughput sequencing techniques and the technological development accompanying them have favoured the development and popularisation of a new range of genomic research disciplines, collectively known as the *omics*. These technologies are capable of simultaneously measuring thousands of molecules which are essential for life, including DNA, RNA, proteins, and metabolites. Historically, classical genomic research has followed a reductionist approach by studying the structure, regulation, and function of these biological units independently. However, despite being a powerful analytical tool, the reductionist method cannot explain many of the biological phenomena that take place in living systems. This is because these biological events are not represented by the sum of their components, rather, only the interacting dynamics of the different omics elements can explain their complexity.

In recent years Systems Biology has established itself as a multidisciplinary area of research which tries to model the dynamic behaviour of biological systems by holistically studying the interactions between the different omics disciplines; it combines simultaneous measurements of different types of molecules and integrates multiple sources of information in order to identify changing components in a coordinated way and under controlled study conditions. Thus, Systems Biology is an interdisciplinary area that requires biologists, mathematicians, biochemists, and other researchers to work closely together, and in which computer sciences plays a fundamental role because of the volume and complexity of the data handled.

This thesis addresses the problem of data management, integration, and analysis in multi-omics studies. More specifically, this research focused on two of the most characteristic computational challenges in Systems Biology: the development of integrated databases and the problem of integrative visualisation. Therefore, the first part of this work was devoted to designing and creating a bioinformatics resource for managing multi-omics experiments. The resulting platform, known as STATegra Experiment Management System (EMS), offers a complete set of tools that facilitate the storage and organisation of the large datasets generated during omics experiments, and also provides tools for data annotation in the later stages of processing and analysis of the information. The development of this platform required overcoming problems created by the heterogeneity, volume, and high variability of the data. Thus, as part of the solution to these problems, detailed metadata can be recorded within STATegra EMS, allowing dataset discrimination and successful data integration. To aid this process, the platform also offers a collaborative and easy-to-use web interface that combines modern web technologies and well-known community standards to represent the different components of the integrated experiments.

The second part of this thesis examines the current situation and challenges in integrative data visualisation in multi-omic experiments, and presents the PaintOmics 3 web tool which was developed to address these issues. Since the capacity of the human brain for visual processing is highly evolved, integrative visualisation combined with data analysis techniques is probably one of the most powerful tools for interpreting and validating results in Systems Biology. PaintOmics 3 provides a comprehensive framework for performing biological function enrichment analyses in experiments with multiple conditions and data types; it combines powerful tools for integrative data visualisation on Kyoto Encyclopedia of Genes and Genomes (KEGG) molecular-interaction diagrams, biological-process interaction-networks, and statistical analyses. Moreover, unlike similar tools, PaintOmics 3 is interactive and easy to use, and stands out for its flexibility and the variety of omics data types it accepts, which include epigenomics data based on genomic regions, proteomics data, and miRNA-study data.

Resumen

Los avances en las técnicas de secuenciación de alto rendimiento y el posterior abaratamiento tecnológico han favorecido el desarrollo y la popularización de una nueva gama de disciplinas de investigación genómica, conocidas colectivamente como "ómicas". Estas tecnologías son capaces de realizar mediciones simultáneas de miles de moléculas esenciales para la vida, tales como el ADN, el ARN, las proteínas y los metabolitos. Históricamente, la investigación genómica clásica ha seguido un enfoque reduccionista al estudiar la estructura, regulación y función de estas unidades biológicas de manera independiente. Sin embargo, pese a ser una poderosa herramienta analítica, el método reduccionista es incapaz de explicar muchos de los fenómenos biológicos que tienen lugar en un sistema vivo, sugiriendo que la esencia del sistema no puede explicarse simplemente mediante la enumeración de elementos que lo componen, sino que radica en la dinámica de los procesos biológicos que entre ellos acontecen.

La Biología de Sistemas se ha establecido en los últimos años como el área de investigación multidisciplinaria que trata de modelar el comportamiento dinámico de los sistemas biológicos a través del estudio holístico de las interacciones entre sus partes, combinando mediciones simultáneas de diferentes tipos de moléculas e integrando múltiples fuentes de información para identificar aquellos componentes que cambian de manera coordinada en las condiciones bajo estudio. La Biología de Sistemas es un área interdisciplinar que requiere que biólogos, matemáticos, bioquímicos y otros investigadores trabajen en estrecha colaboración, y en la que la informática tiene un papel fundamental dado el volumen y la complejidad de los datos manejados.

Esta tesis aborda el problema de la gestión, integración y análisis de los datos en estudios *multi-ómicos*. Más específicamente, la investigación realizada se ha centrado en dos de los retos computacionales más característicos de la Biología de Sistemas: el desarrollo de bases de datos integrativas y el problema de la visualización integrativa. Así, la primera parte de este trabajo se ha dedicado al diseño y creación de un recurso bioinformático para la gestión de experimentos *multi-ómicos*. La plataforma desarrollada, conocida como STATegra EMS, ofrece un conjunto de herramientas que facilitan el almacenamiento y la organización de los grandes conjuntos de datos que son generados durante estos experimentos, así como la anotación de las etapas posteriores de procesamiento y análisis de la información. La heterogeneidad, el volumen y la alta variabilidad de los datos *ómicos* son algunos de los obstáculos que han sido abordados durante el desarrollo del STATegra EMS, con el fin de alcanzar un registro detallado de la meta-información que permita discriminar cada conjunto de datos y lograr así una integración exitosa de la información. Para ello, la plataforma desarrollada ofrece una interfaz web colaborativa y de fácil manejo en la que se combinan modernas tecnologías web y conocidos estándares comunitarios para la representación de los diferentes componentes del experimento.

En la segunda parte de esta tesis se discuten la situación actual y las dificultades de la visualización integrativa de datos en experimentos *multi-ómicos*, y se presenta la herramienta web desarrollada, PaintOmics 3. Dado que la capacidad del cerebro humano para el procesamiento visual está altamente evolucionada, la visualización integrativa en combinación con técnicas de análisis de datos es probablemente una de las herramientas más poderosa para la interpretación y validación de los resultados en Biología de Sistemas. PaintOmics 3 proporciona un completo marco de trabajo para realizar análisis de enriquecimiento de funciones biológicas en experimentos con múltiples condiciones y tipos de datos *ómicos*, en el que se combinan potentes herramientas de visualización integrativa de datos sobre diagramas de interacción molecular y redes de reacción KEGG, redes de interacción de procesos biológicos, y estudios estadísticos de los datos. Además, a diferencia de otras herramientas desarrolladas, PaintOmics 3 destaca por su facilidad de uso y su gran interactividad, así como por su flexibilidad y variedad de los datos *ómicos* aceptados, incluyendo datos de

epigenómica basados en regiones genómicas, datos de proteómica o estudios de miRNA.

Resum

Els avenços en les tècniques de seqüenciació d'alt rendiment i l'abaratiment tecnològic posterior han afavorit el desenvolupament i la popularització d'una nova gamma de disciplines d'investigació genòmica, conegudes col·lectivament com a "òmiques". Aquestes tecnologies permeten realitzar mesuraments simultanis de milers de molècules essencials per a la vida, com ara l'ADN, l'ARN, les proteïnes i els metabòlits. Històricament, la investigació genòmica clàssica ha seguit un enfocament reduccionista a l'hora d'estudiar l'estructura, la regulació i la funció d'aquestes unitats biològiques de manera independent. No obstant això, tot i ser una eina analítica poderosa, el mètode reduccionista és incapaç d'explicar molts dels fenòmens biològics que tenen lloc en un sistema viu, suggerint que l'essència del sistema no es pot explicar simplement mitjançant l'enumeració d'elements que el componen, sinó que radica en la dinàmica dels processos biològics que tenen lloc entre ells.

La Biologia de Sistemes ha esdevingut els darrers anys l'àrea d'investigació multidisciplinària que tracta de modelar el comportament dinàmic dels sistemes biològics a través de l'estudi holístic de les interaccions entre les seues parts, combinant mesuraments simultanis de diferents tipus de molècules i integrant múltiples fonts d'informació per a identificar aquells components que canvien de manera coordinada en les condicions objecte d'estudi. La Biologia de Sistemes és una àrea interdisciplinar que requereix que biòlegs, matemàtics, bioquímics i altres investigadors treballen plegats i en la qual la informàtica té un paper fonamental, atès el volum i la complexitat de les dades emprades.

Aquesta tesi aborda el problema de la gestió, la integració i l'anàlisi de les dades en estudis *multi-òmics*. Més concretament, la investigació s'ha centrat en dos dels reptes computacionals més característics de la Biologia de Sistemes: el desenvolupament de bases de dades integratives i el problema de la visualització integrativa. Així, la primera part d'aquest treball s'ha dedicat al disseny i creació d'un recurs bioinformàtic per a la gestió d'experiments *multi-òmics*. La plataforma desenvolupada, coneguda com a STATegra EMS, ofereix un conjunt d'eines que faciliten l'emmagatzematge i l'organització dels grans conjunts de dades que són generats durant aquests experiments, així com l'anotació de les etapes posteriors de processament i anàlisi de la informació. L'heterogeneïtat, el volum i l'alta variabilitat de les dades *òmiques* són alguns dels obstacles que han estat abordats durant el desenvolupament de l'STATegra EMS, amb la finalitat d'assolir un registre detallat de la meta-informació que permeta discriminar cada conjunt de dades i aconseguir així una integració reeixida de la informació. Per a aconseguir-ho, la plataforma desenvolupada ofereix una interfície web col·laborativa i fàcil de fer servir que conjumina modernes tecnologies web i coneguts estàndards comunitaris per a la representació dels diferents components de l'experiment.

En la segona part d'aquesta tesi s'hi estudia la situació actual i les dificultats de la visualització integrativa de dades en experiments *multi-òmics* i s'hi presenta l'eina web desenvolupada: PaintOmics 3. Com que la capacitat del cervell humà per al processament visual ha evolucionat en gran manera, la visualització integrativa en combinació amb tècniques d'anàlisi de dades és probablement una de les eines més poderosa per a la interpretació i validació dels resultats en Biologia de Sistemes. PaintOmics 3 proporciona un marc complet de treball per a fer anàlisis d'enriquiment de funcions biològiques en experiments amb múltiples condicions i tipus de dades *òmiques*; s'hi combinen eines potents de visualització integrativa de dades sobre diagrames d'interacció molecular i xarxes de reacció KEGG, xarxes d'interacció de processos biològics i estudis estadístics de les dades. A més, a diferència d'altres eines desenvolupades, PaintOmics 3 és molt interactiva i fàcil d'usar, i destaca per la flexibilitat i varietat de dades *òmiques* que accepta, com ara dades d'epigenòmica basades en regions genòmiques, dades de proteòmica o estudis de miRNA.

Dedicado a Elena, Emilio y Vero

Agradecimientos

En primer lugar me gustaría agradecer a Ana, mi directora y mentora, por la oportunidad de haber formado parte de un grupo tan vivo, capaz y ambicioso, pero a la vez tan cercano, humano y divertido. Gracias por la libertad y la confianza que has tenido siempre en mí a la hora de desarrollar mis ideas, por corregirme, por ayudarme a descubrir mi potencial y por enseñarme tantísimo durante estos años.

Por supuesto también me gustaría agradecer su apoyo a mi grupo, el i52. Gracias a súper Sonia por ser capaz de ayudarme siempre con la mejor de sus sonrisas y con toda la paciencia del mundo aun estando desbordada de trabajo. A Pedro y Lorena mis compañeros de fatiga y de viaje por las Américas, a Eugenia por ser una profesional como una catedral y una persona maravillosa, y a Cristinovka, Patri y Mónica por la buena compañía, las risas y los buenos momentos.

A todos mis amigos, los de aquí y los de allá por animarme en los momentos de bajón y por hacerme, con vuestro cariño, más dulces la victorias. A SiGeCo, porque sois increíbles, únicos, porque sin mi dosis de mails diarios los días me parecen torcidos (y demasiado productivos). Los años han pasado, SiGeCo sigue igual. A Diego, el pelotudo, por ser un gran amigo y el mejor compañero de escalada y de birras. A Pablo, Jaime y Pelle, mis hermanos, porque siempre han creído en mí. Esta tesis lleva un trocito de todos vosotros.

A mi familia. A mi padre por su perseverancia (o cabezonería) y por la fuerza que me ha inculcado y que me ha ayudado a llegar hasta este día sin rendirme. A mi madre por su arte, cuántas veces me habrán dicho lo "bonitos"que son mis programas y siempre pienso "¡se lo debo a mi madre que es una artista!", y por su alegría y espíritu desenfadado que me ayudan a empequeñecer los obstáculos y a pasar por encima de ellos. Y a los dos por trabajar tan duro para que sus hijos tuviéramos la mejor educación; esta tesis es mi manera de daros las gracias. A mis hermanas, Isa y Ele, porque son increíbles y no me fallan nunca. A Susana y Alfonso, por *adoptarme* como su hijo y por ser siempre tan felices, positivos y pacientes conmigo. Qué afortunado soy de teneros a todos.

Y por último, a Vero, la persona más importante de mi vida. Por todas las veces que me has levantado cuando estaba caído, por ayudarme a desconectar, por ser mi templo, por ser tan fuerte, por ser tan crack. Nada de esta tesis habría sido posible sin ti.

Tantas veces he soñado que llegaba el día de acabar la tesis y ahora que termino, resulta que la echaré de menos.

Contents

1

About this Thesis

1.1 Motivation

High-throughput experimental methods provide an outstanding resource for researching the behaviour of complex biological systems, but also represent an enormous challenge when trying to manipulate these large heterogeneous biological data sets and convert them into useful knowledge. The development of integrative methodologies and tools for Systems Biology has significantly increased in recent years, and has been propelled by the proliferation of large international consortia, such as the Encyclopaedia of DNA Elements (ENCODE) project [32] or The Cancer Genome Atlas (TGCA) [153]. Nevertheless, Systems Biology is constantly evolving and further development will still be required to comprehensively integrate the complex multidimensional data generated by different omics platforms. Hence, the work presented in this thesis focuses on two of the most important branches in Systems Biology: the development of integrated databases and the integrative visualisation of omics data.

The field of integrative databases seeks to create single repositories from heterogeneous data sources, establishing interconnections among the different datasets and providing a unified and centralised interface for information retrieval. More specifically, the lack of comprehensive tools for properly storing and organising large datasets and for managing the processing pipelines associated with multi-omics experiments motivated the first part of this thesis (Chapters 3 and 4), which is dedicated to the development and use of a management system for multi-omics experiments.

In contrast, integrative visualisation aims to facilitate the interpretability of different omics-system structures by developing tools that support multiple types of molecular data which are capable of displaying them in different ways. Several resources for integrative visualisation are available in the context of Systems Biology but, in most cases, the integration is not completely effective. Therefore, the second part of this work focuses entirely on integrative visualisation of multi-omics data. Chapter 5 discusses the development of a web-based application for integrative visualisation of multiple biological datasets on Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway diagrams, and Chapter 6 introduces the use of this tool in the context of a Systems Biology research.

1.2 Objectives

- **To develop a user-friendly and integrated system for the management, annotation, and storage of multi-omics experiments.** Specifically, the following tasks will be addressed:

  - The system must include tools for annotating the experimental design, biological material, and subsequent processing and manipulation of the data generated for Systems Biology studies.

  - The application should accept the following omics types: transcriptomics, proteomics, metabolomics, and epigenomics; more specifically, it must include tools for annotating at least the most popular technologies for each omics type (e.g. messenger RNA sequencing (mRNA-seq) for transcriptomics, chromatin immunoprecipitation followed by sequencing (ChIP-seq) for epigenomics, and gas chromatography-mass spectrometry (GC-MS) for metabolomics measurements).

  - The system must be conceived as a centralised service, supporting multiple users and collaborative annotations.

  - To ensure the usability and integration capability of the system, the information must be stored using standards accepted by the research community for each omics data type.

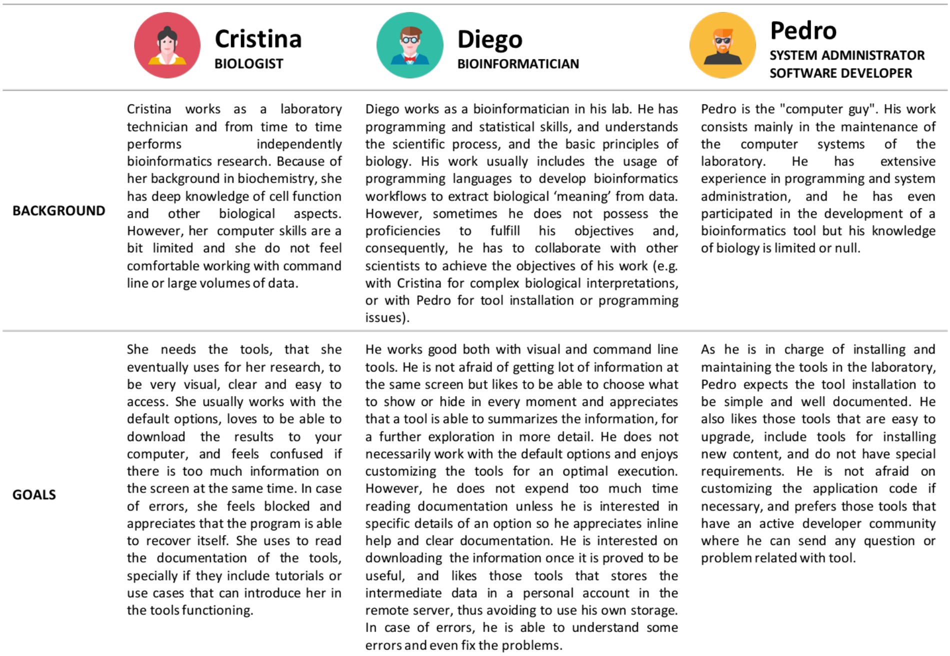
- **To develop a user-friendly tool for integrative visualisation of multiple omics data types based on metabolic pathways.** Specifically, the following tasks will be tackled:

  - The new platform must be flexible enough to accept data for transcriptomics, proteomics, metabolomics, and epigenomics. Tools for data manipulation will be included if necessary.
  - The main visualisation approach will be based on pathway maps, although complementary approaches can be included (networks or genome browsers).
  - The information for pathways should be extracted from a reference database such as the KEGG or the Reactome Pathway Database.
  - The system must provide a complete framework for biological-functions enrichment-analysis for multiple species. Different naming conventions could be accepted for each species.

Moreover, two additional objectives are considered for the proper development of the thesis:

- **As general rule, the new tools developed must be reliable and user-friendly**.This is because potential users for both tools are researchers with medium-low informatics skills. Table 1.1 provides an overview for the target user groups in a typical bioinformatics scenario and their requirements in terms of user-friendliness and accessibility.
- **Good accessibility and wide distribution of the generated software are also objectives for this thesis**. To achieve these goals, development of these tools will include the provision of guidelines and training on their use, as well as the diffusion of the results of this work in scientific journals

and at international conferences. In addition, the use of free and open-source technologies will be prioritised.



	Cristina **BIOLOGIST**	**Diego** **BIOINFORMATICIAN**	**Pedro** **SYSTEM ADMINISTRATOR** **SOFTWARE DEVELOPER**
BACKGROUND	Cristina works as a laboratory technician and from time to time performs independently bioinformatics research. Because of her background in biochemistry, she has deep knowledge of cell function and other biological aspects. However, her computer skills are a bit limited and she do not feel comfortable working with command line or large volumes of data.	Diego works as a bioinformatician in his lab. He has programming and statistical skills, and understands the scientific process, and the basic principles of biology. His work usually includes the usage of programming languages to develop bioinformatics workflows to extract biological 'meaning' from data. However, sometimes he does not possess the proficiencies to fulfill his objectives and, consequently, he has to collaborate with other scientists to achieve the objectives of his work (e.g. with Cristina for complex biological interpretations, or with Pedro for tool installation or programming issues).	Pedro is the "computer guy". His work consists mainly in the maintenance of the computer systems of the laboratory. He has extensive experience in programming and system administration, and he has even participated in the development of a bioinformatics tool but his knowledge of biology is limited or null.
GOALS	She needs the tools, that she eventually uses for her research, to be very visual, clear and easy to access. She usually works with the default options, loves to be able to download the results to your computer, and feels confused if there is too much information on the screen at the same time. In case of errors, she feels blocked and appreciates that the program is able to recover itself. She uses to read the documentation of the tools, specially if they include tutorials or use cases that can introduce her in the tools functioning.	He works good both with visual and command line tools. He is not afraid of getting lot of information at the same screen but likes to be able to choose what to show or hide in every moment and appreciates that a tool is able to summarizes the information, for a further exploration in more detail. He does not necessarily work with the default options and enjoys customizing the tools for an optimal execution. However, he does not expend too much time reading documentation unless he is interested in specific details of an option so he appreciates inline help and clear documentation. He is interested on downloading the information once it is proved to be useful, and likes those tools that stores the intermediate data in a personal account in the remote server, thus avoiding to use his own storage. In case of errors, he is able to understand some errors and even fix the problems.	As he is in charge of installing and maintaining the tools in the laboratory, Pedro expects the tool installation to be simple and well documented. He also likes those tools that are easy to upgrade, include tools for installing new content, and do not have special requirements. He is not afraid on customizing the application code if necessary, and prefers those tools that have an active developer community where he can send any question or problem related with tool.

Table 1.1: Target user groups in a typical bioinformatics scenario and their requirements in terms of user-friendliness and accessibility.

1.3 Organization

This thesis describes the systems biology work performed during the last four years related to the objectives outlined above, more specifically, for the development of integrative databases and tools for integrative visualisation of omics data. The thesis consists of a central section (**Chapter 3 to 6**) which describes the methodologies used to develop both tools as well a use case-study for each one. This is preceded by a general introduction (**Chapter 2, p.7**) which provides some insight into molecular biology and Systems Biology; the main purpose of this chapter is to make the thesis as self-contained as possible.

- Chapter 3 (p.31) describes the state-of-the-art in integrative databases for storing and organising data for multi-omics experiments, and introduces the methodology followed to develop the "STATegra Experiment Management System (EMS)".

- Chapter 4 (p.53) illustrates how the STATegra Experiment Management System (EMS) can be used to annotate a multi-omics experiment.

- Chapter 5 (p.77) focuses on the field of integrative multi-omics data visualisation, and introduces the PaintOmics 3 platform.

- Chapter 6 (p.123) describes a complete use case-study for PaintOmics 3 in the context of real multi-omics biomedical research.

Finally, a general discussion provides an overview of the major contributions of the work and future perspectives.

Appendices A (p.161) and B (p.169) provide some supplementary information that complements the content of the main text body.

1.4 Context

This work was carried out at the Genomics of Gene Expression Laboratory at the *Centro de Investigación Príncipe Felipe* (CIPF), under the supervision of Dr. Ana Conesa (CIPF); Dr. Montserrat Robles (*Universitat Politècnica de València*) provided tuition for this project. The work was developed within the framework of the international *STATegra* research project; the tools developed for this thesis were used to analyse the data generated by the *STATegra* project and therefore, publication of this data analysis also contributed to the dissemination of the results described in this thesis.

This work was funded by the following projects and grants:

STATegra: User-driven Development of Statistical Methods for Experimental Planning, Data Gathering, and Integrative Analysis of Next Generation Sequencing, Proteomics, and Metabolomics data. European Commission (FP7), project number 306000.

The Gerónimo Forteza program for aiding the recruitment of support personnel. Ministry of Education, Research, Culture, and Sport, Valencian Regional Government (*Programa Gerónimo Forteza de Ayudas para la Contratación de Personal de Apoyo*. Conselleria d' Educació, Investigació, Cultura i Esport de la Generalitat Valenciana).

The genomics and transcriptomics of detoxification pathways in Drosophila project. Ministry of Economy and Competiveness, Spanish central Government (*Genómica y transcriptómica de las rutas de detoxificación en Drosophila*. Ministerio de Economía y Competitividad, Gobierno de España). Project number PIB2010AR-00266.

2

Introduction

Part of this chapter have been published in "Conesa A. and Hernández-de-Diego R. *Omics Data Integration in Systems Biology: Methods and Applications*. In: Applications of Advanced Omics Technologies: From Genes to Metabolites, Volume 64 (Comprehensive Analytical Chemistry). Ed. by García-Cañas V, Cifuentes A, and Simó C. Elsevier, 2014".

2.1 Introduction

The objective of this introduction is to provide an overview of the key concepts relevant to the topic of this thesis, namely, the development of software tools for managing and analysing multi-omics data. The content of the chapter has been divided into different sections: the first section describes the molecules and biological processes that are measured by the different omics technologies; the second section is about these technologies themselves, the challenges they currently imply, and the state of the art. Finally, the third section reviews the Systems Biology paradigm and the role multi-omics analysis plays in it.

2.2 Biological background

Deoxyribonucleic acid (DNA), ribonucleic acid (RNA), proteins, and metabolites are present in every cell in multicellular and unicellular organisms. The complex network of interactions between these molecules is vital to all cellular functions and processes; therefore, knowledge of which molecules are required for certain cellular functions, and understanding how these biological processes are regulated, are fundamental to the comprehension of molecular biology.

Metabolism is the term used to describe the set of biochemical reactions that takes place in cells, allowing organisms to grow and reproduce, maintain their structures, and to respond to changes in their environment; it is a dynamic process in which cells are continuously degrading and synthesising most cellular materials [44] (Figure 2.1-A). The molecular transformations of metabolism are organised into metabolic pathways, i.e. coordinated series of interactions between chemical reactions in which one metabolite is transformed into another product through a series of steps [99]. Cellular metabolism can be divided into two broad categories: *catabolism* and *anabolism*. Catabolism is the set of metabolic processes that breakdown and oxidise large molecules; the purpose of catabolic reactions is to provide the energy and components required for anabolic reactions. On the contrary, anabolism is a set of constructive metabolic processes in which cells use the energy released by catabolism to synthesise complex molecules [28]. All biochemical reactions are catalysed by one or more flexible proteins known as *enzymes*, which promote these reactions by reducing the reaction activation energy required, i.e. the minimum amount of energy required to proceed with the reaction. Enzymatic activity is crucial for metabolic processes because it allows cells to respond to environmental challenges and to regulate their metabolic pathways, both processes which are crucial to cell survival.

Proteins are large molecules comprising chains of units called *amino acids* (Figure 2.1-C). An amino acid is an organic compound containing amine ($-NH_2$) and carboxylic acid ($-COOH$) functional groups. There are many types of amino acids; some of them can be synthesised by different organisms while others, the so-called essential amino acids, must be supplied in the organisms' diet.

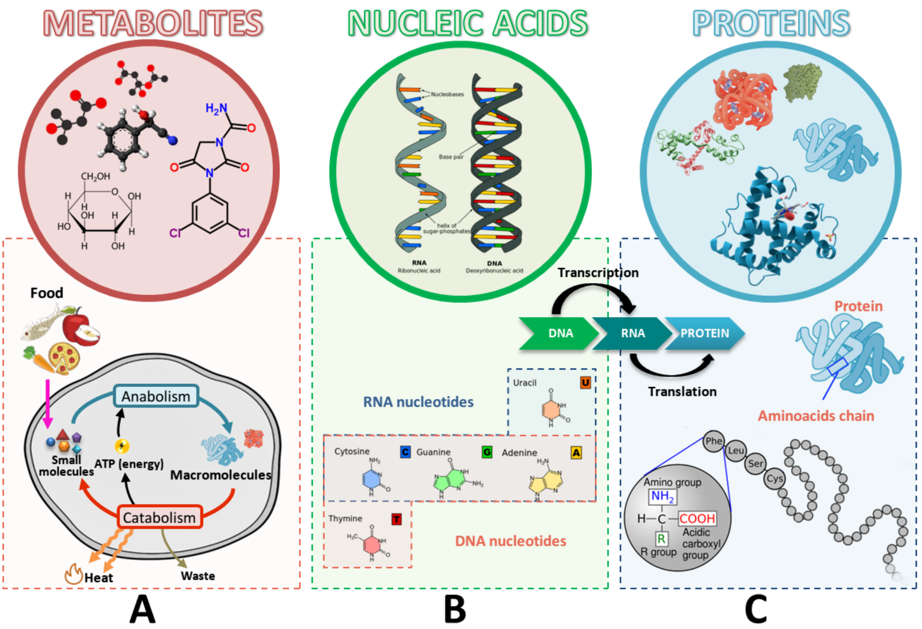


Figure 2.1: The major constituents of organisms: nucleic acids, proteins, and metabolites. (A) - Metabolites are small molecules that work both as intermediate and final products of metabolomic reactions. Primary metabolites are synthesised by cells and are indispensable for their growth. Secondary metabolites are not directly involved in these processes, but usually have an important function on the organism. **(B)** - Nucleic acids are large biomolecules that include DNA (deoxyribonucleic acid) and RNA (ribonucleic acid). The core of nucleic acids consists of nitrogenous bases which are divided into two groups: *pyrimidines*, including cytosine (C), uracil (U), and thymine (T), and *purines*, which include adenine (A) and guanine (G). DNA is formed by two complementary strands of T, G, C, and A bases joined by a phosphate backbone, while RNA molecules are single-stranded and combine T, G, C, and U monomers. DNA sequences are converted into RNA molecules by the action of specific enzymes in a process called transcription. **(C)** - Proteins are macromolecules comprising one or more chains of amino acids, organic compounds containing amine $(-NH_2)$ and carboxylic acid (-COOH) functional groups. There are hundreds of known amino acids but only 22 are used for protein synthesis during translation (proteinogenic amino acids). Some amino acids can be synthesised from scratch by organisms while others must be supplied in the diet (essential amino acids).

Proteins are responsible for almost all cellular functions; for instance, as enzymes, proteins catalyse biochemical reactions (Figure 2.2-8), as transcription factors (Figure 2.2-4) they regulate protein synthesis itself, and as antibodies they are used by the immune system to identify and neutralise pathogens. The information necessary to produce all proteins is encoded in the DNA.

DNA is essentially a storage molecule. It contains all of the instructions a cell needs to sustain itself [90]. DNA molecules form a double helix where each chain is a sequence of four basic building blocks, called nucleotides (Figure 2.1-B). There are four nucleotides in the DNA alphabet: adenine (A), cytosine (C), guanine (G) and thymine (T). By combining these four bases DNA encodes genetic information into *genes* –the basic physical and functional units of genetic heredity which, in turn, determine the instructions for building some of the molecules essential for life. When a gene encodes a protein, it is known as a "protein coding gene". Importantly, not all DNA encodes information for protein synthesis. In fact, only a small part of the whole genome in humans is considered to be "coding", while the remaining DNA sequences are involved in regulatory or structural processes.

In eukaryotic organisms, most DNA is located in the cell nucleus, packaged into thread-like structures called chromosomes [110] (Figure 2.2-1). Inside chromosomes the DNA helix wraps multiple times around histone proteins to form nucleosomes (Figure 2.2-4); these nucleosomes coil tightly to form chromatin loops (Figure 2.2-3) which, in turn, wrap around each other to form chromosomes [6] (Figure 2.2-2).

The central dogma of biology explains how the genetic information coded into certain sections of DNA is decrypted into proteins, via the transcription of individual transportable units of RNA. This critical trio of macromolecules – DNA, RNA, and proteins – is present in all cells [81].

RNA molecules are created when the "instructions" encoded by genes are decoded in a process called *transcription* or *gene expression* (Figure 2.2-6); the step leading from RNA to protein production is called translation (Figure 2.2-7). The chemical structure of RNA transcripts is similar to that of DNA, except that it is single-chained rather than a double helix, and one of the four bases,

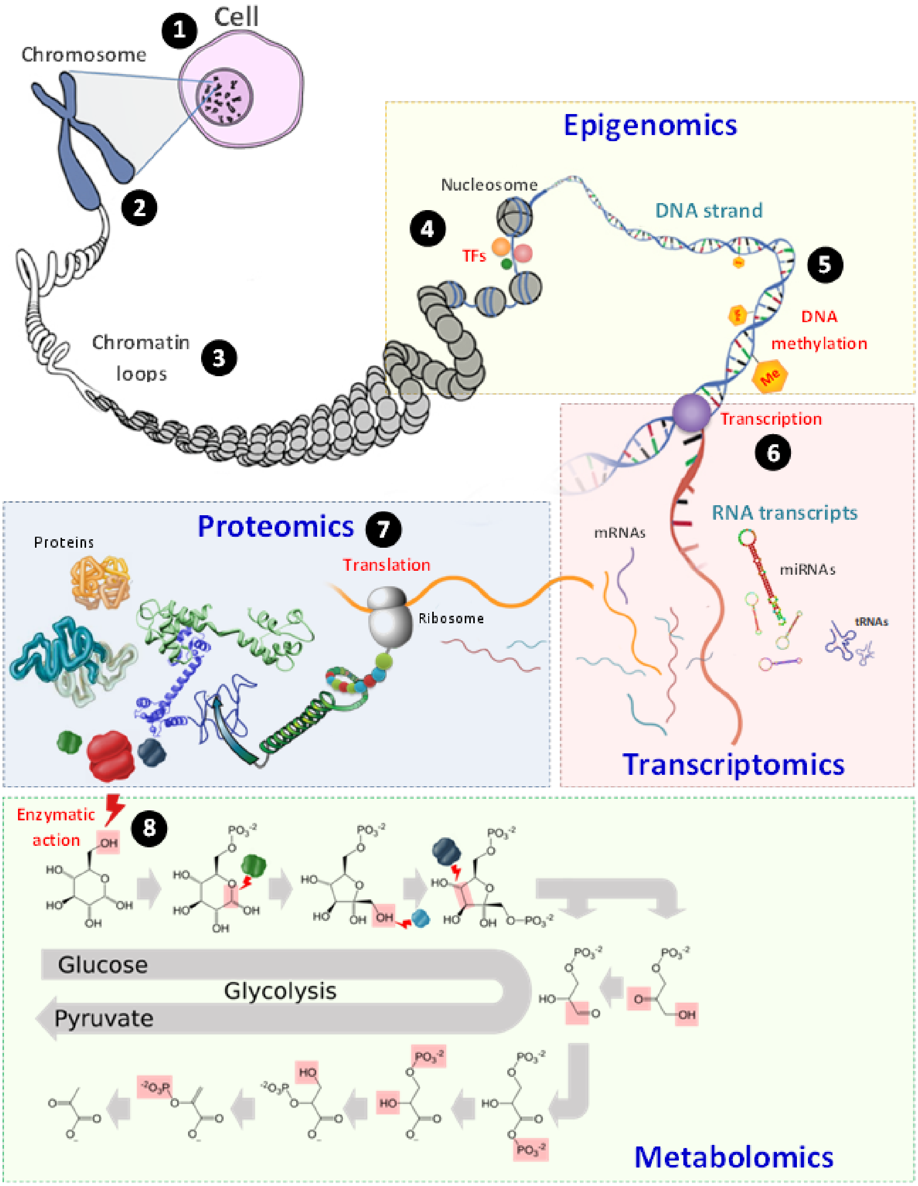


Figure 2.2: A schematic representation for the main branches in functional genomics.

T, is replaced by U (Figure 2.1-B). Several types of RNA are present in cells, but many of them do not code for proteins (non-coding RNAs); *messenger RNAs(mRNAs)* carry the genetic information responsible for protein synthesis to ribosomes –a cellular component where biological protein synthesis occurs (translation). Transfer RNAs(tRNAs) transport amino acids to the protein-synthesis machinery during translation and ribosomal RNAs(rRNAs) then link the amino acids together to form proteins [81].

During transcription, the area around the gene to be transcribed must be unpacked in order to facilitate its access to the transcription machinery. This complex process requires the coordination of multiple proteins, as well as some chromatin-structure modifications to make it more transcription-permissive (by default it is strongly repressive to transcription). Once the gene sequence is accessible, the transcription machinery attaches to the DNA template strand and begins assembling a new chain of nucleotides, thus producing a complementary RNA strand [3]. The more the gene is *read*, the more of the corresponding transcript is produced in the cell. However, gene expression is a highly regulated process: only a subset of genes in the genome are expressed at any particular moment or in a given tissue, and genes turn "on" and "off" during a cell's lifetime in a process known as *gene regulation* [110].

In eukaryotes, the regulation of transcription occurs at many different stages: transcriptional initiation, elongation, splicing, etc. and is the result of the combined effects of structural properties and the interactions of several molecules such as the transcription factors. Moreover, mRNAs can be post-transcriptionally regulated, for example, by microRNAs(miRNAs) or post-translational protein modifications, which add on additional layers of biological regulation.

Transcription factors(TFs) are regulatory proteins whose function is to control which genes are turned "on" or "off" in the genome by binding to DNA and other proteins. Once bound to DNA, these proteins can promote or block the enzyme that controls the reading (transcription) of genes, making genes more or less active [59]. However, even when TFs are present in a cell, transcription does not always occur. For example, in eukaryotes the accessibility of different DNA regions depends on the chromatin structure and the density of its packing may indicate the frequency of transcription. Another important regulation method in eukaryotes is the *DNA methylation* that occurs when a methyl-group is added to cytosine DNA bases (Figure 2.2-5) and whose presence constitutes a common *epigenetic* signalling factor that cells use to lock genes in the "off" position [106].

In addition, *miRNAs* have emerged in the last few years as a major new research focus in molecular biology. These small non-coding RNA transcripts are known for their involvement in a wide range of biological processes, especially as key post-transcriptional gene-expression regulators. Although the mechanisms of miRNA-mediated repression are not yet fully understood, it is thought that they either trigger bound mRNA degradation or hamper the translation of target genes into active proteins [36].

2.3 The omics technologies and bioinformatics

Understanding the behaviour of cells, tissues, organs, and the entire organism at the molecular level is the major objective of molecular biology, and to achieve this goal it is essential to characterise the function, regulation, and interaction of all the biomolecules described above. The advances in sequencing technologies achieved over the last two decades, and the subsequent development of high-throughput technologies, have given rise to a new range of research disciplines, which are collectively referred to as omics. The so-called *omics* technologies allow individuals of the same (or even different) species to be compared at the molecular level. Thanks to these technological advances, it is now practical and affordable to sequence entire genomes to look for variants which might be associated with diseases, measure variations in molecular profiles in response to drugs or other environmental challenges, and even compare the sequences of hundreds of genes between different species in order to create a more detailed and accurate evolutionary tree than ever before possible.

High-throughput instruments are now routinely used in individual laboratories and, as an immediate consequence, scientists must now deal with the resulting massive datasets and subsequent challenges of handling, processing, and analysing this information [13]. It is within this context that *Bioinformatics* has arisen as a key tool for studying the vast amount of data generated by these technologies. Bioinformatics is an interdisciplinary field that combines very different techniques from vastly different fields such as biology, computer sciences, mathematics, statistics, and physics, and is commonly applied in the great variety of omics techniques currently available.

In general terms, each omics approach seeks to provide specific insights into biological systems and relationships between cellular elements and their phenotypic manifestations at the organismal level. Omics disciplines can be roughly divided according to the aspect of biological research of interest: *Genomics* is devoted to comprehensively studying the "static" or "structural" aspects of genomes, including their architecture, origin, and evolution; *Functional genomics* focuses on dynamic processes linked to functionality such as the transcription, translation, and regulation of gene expression.

Genomics has numerous applications, especially in medicine for studying genetic diseases or in the development of more precise drugs, and encompasses many research areas of molecular biology. Some popular areas of genomics are: *comparative genomics*, which studies the similarities and differences between the genomes of different organisms; *metagenomics*, the study of the entire microbial genome in a given environment such as water or soil; and *structural genomics*, which seeks to determine the three-dimensional structure of all proteins encoded by a given genome. This latter aspect of genomics is not considered in this work.

In contrast, functional genomics aims to understand the complex relationship between the genome and phenotypic manifestations, focusing on the dynamic aspects of genes and the biochemical and physiological functions of all gene products [61]. Consequently, functional genomics comprises a wide range of omics disciplines that measure molecular activities, and is the focus of this thesis. In particular, this work deals with transcriptomics, epigenomics, metabolomics, and proteomics.

2.3.1 Transcriptomics

The transcriptome is the complete set of transcripts encoded by the genome of a cell or organism under specific physiological conditions. Studying the transcriptome is essential for understanding the functionality of the genome and revealing the molecular processes of development and disease in cells and tissues. Transcriptomics is the generic term used to describe the set of methods which aim to comprehensively identify the whole catalogue of transcripts, including mRNAs, non-coding RNAs, and small RNAs, to determine how expressed RNAs are processed (including splicing, transport, and editing), and to profile how the expression levels of each transcript change under different conditions [151].

Transcriptomic analysis has been traditionally performed using *microarray* technology. A microarray is a solid support (e.g. a glass slide or a silicon chip) that contains thousands of spotted samples, known as probes, which are coupled to a fluorescent marker; each one represents a specific gene from a known subset of the genes of a cell or organism (Figure 2.3). The key principle of microarrays is that complementary sequences of nucleic acids tend to pair with each other, forming strong hydrogen bonds. RNA samples are treated to convert them into complementary DNA (cDNA), i.e. "synthetic" DNA strands transcribed through the reaction of an enzyme called *reverse transcriptase* and using the RNA sequence as a template. These treated samples are then injected into the microarray so that they come into contact with the probe-spotted support and where some of the cDNA strands bind to their target probes (i.e. the complementary strand). Finally, the array is washed to remove all the unbound strands and is scanned by a machine that uses a laser to excite the fluorescent probes on the target sequences and to measure the intensity of the emissions with a detector. These intensity values are interpreted as an estimation of the expression level of the target genes.

Although microarrays allow the simultaneous measurement of the expression levels for thousands of genes, this technology has a major limitation: previous knowledge of the target genes is required to fabricate a microarray and, consequently, this technology is unable to detect and measure the expression of unknown genes.

RNA sequencing (RNA-seq) is the alternative transcriptomics technology. RNA-seq is a highly sensitive and accurate tool that allows researchers to detect both known and novel transcripts, without the limitation of prior knowledge [94]. RNA-seq, as well as *DNase-seq*, *ChIP-seq*, and *Methyl-seq*, belong to a family of the so-called sequencing-based assays or *Next-Generation Sequencing Technologies*. These methods link particular biochemical assays (used to isolate target-nucleotide sequences) to the use of massively-parallel sequencing instruments and are capable of sequencing millions of base-pairs at an affordable price. Thus, these sequencing machines have created unprecedented possibilities for genomic research.

RNA-seq can be divided into several sub-techniques, depending on the type of RNA being measured; some popular examples are small RNA sequencing (sRNA-seq), microRNA sequencing (miRNA-seq), and messenger RNA-seq (mRNA-seq). In general terms, the methods used for these different branches are similar, and differences are often only obvious at the level of the input material (e.g. for sRNA-seq experiments, transcripts are filtered by size, removing all transcripts whose sequence is longer than a given threshold).

During a typical RNA-seq experiment (Figure 2.4), the RNA population is first converted to smaller cDNA fragments using specific short DNA sequences attached to one or both ends (adaptors). The cDNA fragments are then injected onto a surface (flowcell) that contains short strands (primers) complementary to the adaptors. The complementary strand binds to the adaptor and cDNA fragments are anchored to the flowcell. A complex cycle of amplification and washing is then performed, resulting in a huge increase in the amount of cDNA present. Finally, the resulting strands are reused as a template to synthesise more cDNA molecules, but this time using fluorescently-tagged nucleotides. Every time a nucleotide is added to the new strand, a fluorescent signal is emitted and is detected by a special camera. Each of the tagged bases (A, C, G, and T) gives off a characteristic colour. This massive, parallel process results in the generation of millions of short sequences of characters (reads), which usually must be collected and mapped back to their transcripts of origin before proceeding with gene-expression quantification. The total number of

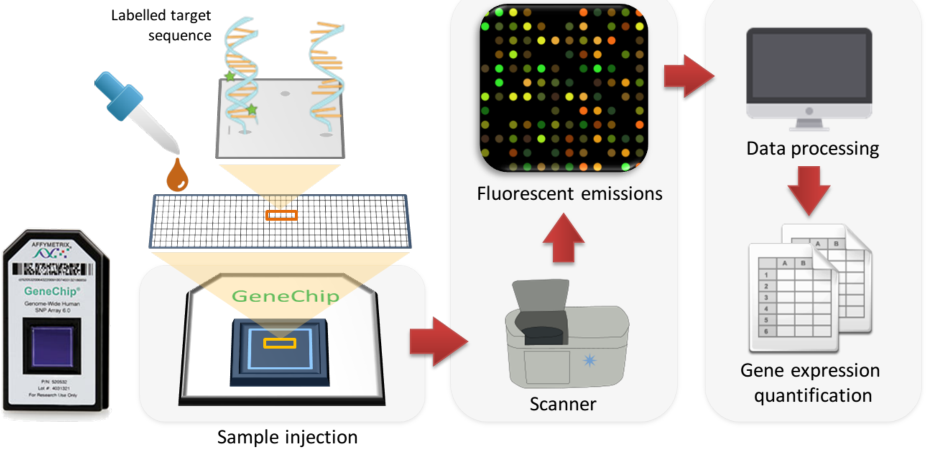


Figure 2.3: Overview of the use of microarrays for gene-expression quantification. The sample is injected into the microarray; after binding takes place it is washed and scanned. Fluorescent emissions are measured and registered, resulting in a matrix of colours where each spot represents a targeted sequence, and the colour intensity indicates the relative abundance of the target. Raw data is then normalised and processed, which for most genes results in an estimation of their absolute abundance.

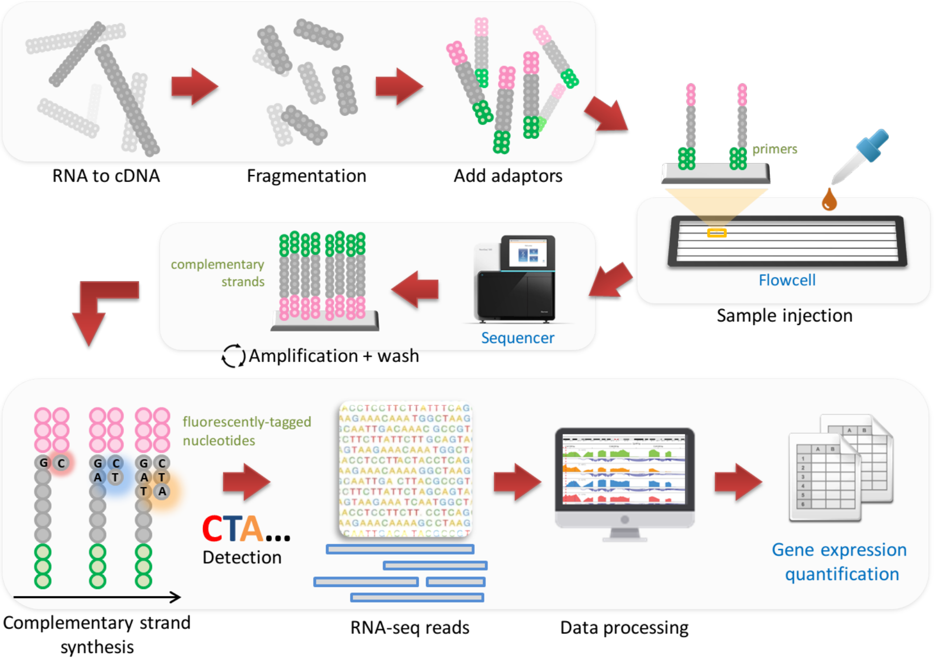


Figure 2.4: Overview of how sequencing for gene-expression quantification (RNA-seq) is implemented.

sequencing reads that map to a given gene is an estimation of the expression level of the gene.

2.3.2 Epigenomics

Epigenomics studies the set of chemical DNA and DNA-associated protein modifications that take place in cells and which alter gene expression [133]. A variety of methods are used for study in epigenomics, among them some of the most popular are:

i. *ChIP-on-chip*. This technique combines chromatin immunoprecipitation (ChIP) and DNA microarrays (chip) and aims to identify the binding sites of transcriptional regulators and other relevant proteins. In a ChIP-on-chip experiment, DNA samples are first treated in order to fix the protein of interest (POI) with the DNA binding sites (crosslinking). The DNA is then fragmented into small double-stranded fragments and an antibody specific to the POI is incorporated, forming antibody-POI-DNA complexes. Next, these complexes are isolated (usually using beads that are mixed into the sample and which specifically capture and immobilise the antibodies in a solid-phase) in procedure known as immunoprecipitation. After that, the DNA fragments are purified, amplified, and labelled using a fluorescent tag. The labelled fragments are injected into a DNA microarray and illuminated with fluorescent light; any probes on the array that contain labelled fragments emit a signal that is captured by a high-sensitivity camera which allows researchers to locate DNA sites containing histone modifications.

ii. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a technique for genome-wide profiling of DNA-binding proteins, histone modifications or nucleosomes [103]. ChIP-seq experiments also use chromatin immunoprecipitation but, instead of microarrays, sequencing technologies are used to locate the binding sites.

iii. Reduced representation bisulfite sequencing (RRBS-seq) is a common high-throughput technique for assessing DNA methylation patterns across the genome at a single-base resolution. During a RRBS experiment, DNA fragments are "bisulfite converted", a key process that converts all the non-methylated cytosines into uracils. Hence, after sequencing, sequences that contain uracils can be identified as methylated DNA regions [77].

iv. DNase I hypersensitive site sequencing (DNase-seq) is a genome-wide sequencing method used to identify regions sensitive to cleavage by the DNase enzyme. These hypersensitive sites are DNA regions that have a less-compact structure which increases the availability of DNA to interact with transcription factors and other regulatory elements, a feature that characterises regions of the genome that likely contain active genes.

2.3.3 Proteomics

By proteomics we understand the comprehensive study of the entire set of proteins produced by an organism or cellular system, also known as the proteome. Typically, mRNA molecules in cells rapidly degrade, become inefficiently translated, or are affected by post-translational modifications such as "alternative splicing" – a process by which certain parts of the transcript are removed, allowing the production of different proteins from the same piece of DNA. Consequently, the proteome has a very dynamic nature and is much larger than the genome, especially in eukaryotes. While measuring the level of gene transcription can provide a rough estimate of the level of translation into proteins, studying changes in the proteome can provide a more accurate snapshot of cellular processes.

Proteomics studies a wide variety of aspects of proteins, including their structure, protein-protein interactions, and protein expression [46]. Of particular interest to the work in this thesis is *quantitative proteomics*, a technique used to determine the quantity of different proteins and the differences between a set of samples. There are several methods for detecting and profiling the presence of proteins; some of most popular are:

i. *Protein Microarrays*, a high-through put method which is conceptually similar to DNA microarrays. This technology method allows specific proteins to be targeted, captured, and measured, allowing scientists to isolate and study many potential biomarker proteins.

ii. *Mass spectrometry (MS)* is a high-resolution method that allows scientists to detect and quantify proteins. In a typical MS experiment samples are ionised and molecules in the sample are converted into gas-phase ions which are then sorted and separated according to their mass-to-charge (m/z) ratio using an electric or magnetic field. They are then injected into an ion detection system which generates a "mass spectrum", i.e. a plot that presents the relative abundance of each detected ion against its m/z ratio, which can be used to identify the peptides originally present in the sample. For example, "peptide mass fingerprinting" is a popular analytical technique for protein identification that compares the masses of the detected molecules to the masses predicted based on the digestion of a set of known proteins, available in certain databases [46].

2.3.4 Metabolomics

Metabolomics refers to the comprehensive qualitative and quantitative study of the metabolic content of a cell, tissue, or organism, which usually focuses on metabolites – the intermediate and end products of cellular processes. Three major branches of metabolomics are extensively used, depending on the experimental goals: *Metabolite fingerprinting*, a technology which provides information about the overall composition of metabolites in a sample, without necessarily identifying or quantifying any particular compound. This high-throughput approach is normally used in tissue comparison or discrimination analysis. *Metabonomics*, which focuses on the metabolic response of organisms to certain pathophysiological stimuli or genetic modifications, and is generally restricted to microbiological studies [39]; and *metabolite profiling*, which aims to identify and quantify low molecular-weight metabolites and their intermediates which together reflect dynamic cell responses induced by genetic modifications or external stimuli (e.g., a drug treatment) [25]. Metabonomics can be considered to be an extension of metabolomic profiling which also considers perturbations caused by environmental factors, while metabolomic profiling studies should, by definition, exclude metabolic contributions from extra-genomic sources.

The main platforms developed to detect metabolites are based on MS and on nuclear magnetic resonance (NMR) spectroscopy. MS-based metabolomics follows a similar approach to proteomics: gas or liquid samples for MS are introduced into the spectrometer where they are ionised and separated by their mass-to-charge (m/z) ratio. Ions are then detected to obtain a mass spectrum that allows the metabolomic composition of the sample to be identified. A very different approach is used for NMR-based methods: NMR is a physical property of the nuclei of atoms which absorb and re-emit electromagnetic radiation when located within a magnetic field. By exciting the sample with radio-frequency pulses a NMR response is emitted and detected by a sensitive radio receiver, resulting in a series of peaks known as the NMR spectrum which plot the radio frequency applied to the sample against its absorption. The shift (i.e. the difference from the zero point), shape, and area of the peaks provide information about the chemical structure of the molecules in the sample. This technology is considered to be "non-destructive" because it does not require separation of the sample components, thus allowing the entire sample to be recovered for further analysis.

2.4 From *omics* to Systems Biology: towards a more complete picture of life at the molecular level

Historically, molecular biology, along with other modern sciences, has taken a reductionist approach, dividing (biological) systems into their constituent parts and studying them in isolation [144]. Methodological reductionism has proven to be a very powerful analytical tool that has allowed scientists to investigate many basic molecular and cellular processes; however, this approach is nearing its limits and it has become evident that, by its nature, reductionism will be unable to provide a complete understanding of the behaviour of systems via exclusively reductive explanations. This is because this approach ignores the highly-structured biological networks that are known to operate in complex biological systems such as cells, tissues, diseases, and even human societies.

For instance, the discovery of alternative splicing made it evident that genes are not just linear representations of the information encoded by DNA. On the contrary, an intricate network of regulatory factors, RNA editing events, and post-translational protein modifications exists, which adds several layers of complexity to gene transcription. Indeed, these events are not controlled by genes, but rather, by other molecules such as proteins or small RNAs. Not surprisingly, the main challenges in computational biology now involve understanding biological phenomena *holistically* in a context where complementary genome-wide measurements could be combined to provide an even more complete picture of life at the molecular level.

Thanks to technological advances, it is now practical and affordable for researchers to obtain simultaneous measurements of different types of features from the same set of samples. However, the high dimensionality of omics measurements requires sophisticated analytical methods and computer modelling to enable the search for meaningful interactions between the components of biological systems.

Systems Biology (SB) arose as a discipline that seeks to provide insights into the processes of living systems, by holistically studying the behaviour and relationships of the components forming it. SB usually involves monitoring the responses of genes, proteins, and other particles in controlled biological, genetic, or chemical perturbation conditions, integrating these data and, ultimately, creating mathematical models that describe the system's structure and allow responses to certain stimuli or environmental changes to be predicted [57]. The high dimensionality and massive size of omics data, as well as the inherent variability and heterogeneity of the aggregated datasets, give rise to unique computational and statistical challenges. Thus, scalability and storage bottlenecks, noise accumulation, measurement errors, and heavy computational costs are some of the hurdles faced by researchers and tool developers when trying to achieve effective multi-omics data integration [34]. Basic concepts in SB are comprehensiveness (incorporation of all the molecular elements of the system), interpretability (SB has a clear goal to better understand biology), and predictive power (mathematical modelling is considered to be a fundamental aspect of SB).

2.4.1 Data integration in Systems Biology

With the rise of novel omics technologies, data integration has become a very commonly used idea in the life sciences. Data integration refers to the combination of multiple sources of heterogeneous data with the aim of better understanding a system under study (SuS); it constitutes not only a conceptual challenge but also a practical challenge in terms of every day analysis in SB [42]. Thus, the high dimensionality, inherent variability, and contrasting nature of omics data are some of the hurdles that researchers must overcome before effective integration can be achieved (Figure 2.5).

Many integrative methodologies and tools have proliferated over the last few years, covering the properties of comprehensiveness, prediction power, and interpretability that, to different degrees, define SB and allows them to be classified into three major groups: Predictive power is the main objective for *integrative omics analysis*, which makes use of statistical methods to unravel relationships between diverse molecular entities and to create predictive models of the biological system being studied. Comprehensiveness is the key characteristic of *integrative databases*, which aim to create single repositories from various heterogeneous data sources, to establish interconnections among them, and to provide a unified and centralised interface for information retrieval. Last but not least, many, if not most, studies incorporating multiple omics types address the integration challenge at the level of *visualisation*, by developing tools that support multiple types of molecular entities and display them in different ways; this approach to integration seeks to facilitate the interpretability of the systems structure.

Integrative omics analysis

Integrative analysis of omics data describes the algorithmic and statistical approaches used to pursue the compilation of different omics data into one analysis. Often, integrative analysis is used for two purposes: first, to perform a descriptive analysis designed to find any underlying relationships between the datasets and second, to predict a certain response using one or more explanatory datasets [109].

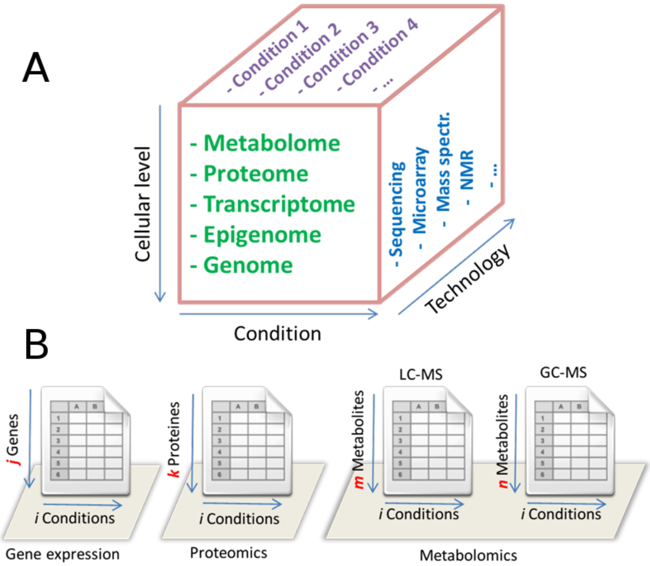


Figure 2.5: Schematic representation of the multi-dimensional complexity of biological data. (A) - The heterogeneity of the data generated depends on multiple factors, such as the nature of the cellular levels measured, the different conditions of the experiment, and the variety of techniques used for the measurements [109]. **(B)** - An example of integrative analysis using four datasets. Although the same number of observations was made for the three cellular levels, each omics approach reports a different number of variables (i.e. genes, proteins, and metabolites) and, even within the same omics experiment-type, different technologies may report a different number of measurements.

In general terms, we can identify two main categories of analytical strategies [53]:

i. Methods that first determine the correlation between the features in each data type and the conditions being studied, and which subsequently combine the results of the different analyses into one interpretative model (*multistage approach*). Most of omics integration reports already published follow this approach.

ii. Methods that initially combine all data and then use computational resources to find mathematical models that allow relationships to be inferred between the different omics variables, and which can, together, better explain the conditions present in the analysis (*simultaneous analysis*).

Furthermore, data integration can be classified into two general types: *horizontal data integration* and *vertical data integration* [140]. Horizontal data integration involves the combination and analysis of different datasets measuring the *same molecular* events under similar experimental conditions; for example, combining gene expression data sets from different breast cancer studies [154]. This approach is frequently used in *meta-analysis*.

Meta-analysis is the combined analysis of multiple datasets (typically collected from public repositories) of one type or biological scope and is one of the most extended forms of integrative omics. Meta-analysis has become a widely used statistical tool and benefits from combining of a large number of observations in order to enhance statistical and discovery power. For example, large-scale correlation meta-analysis across thousands of datasets has been used to identify distinct co-expression modules associated with specific cancer subtypes [78], to dissect the association between gene co-regulation and function, or to discover genetic-risk variants.

In contrast, *vertical data integration*, which is commonly used in SB, combines different data types into one model; for example, by relating transcriptional and metabolic data sets from the same patient cohorts [156]. When building models which can predict biological situations, vertical data integration methods have to deal with different variable sets which may have different properties, e.g. the data might span different dynamic ranges, follow different data distributions, or bear different levels of noise. Moreover, vertical integration may also need to consider the biological relationships between variables; however, in some cases these relationships are unknown and so further investigation is required before precise inferences can be made. Hence, most mathematical SB models restrict themselves to only analysing systems with a reduced number of variables and where the model topology is imposed beforehand. This paradigm, also called "reverse engineering", has been solved by different mathematical approaches, each based on the specific goal of the integrative effort. When the statistical power is insufficient for predictive purposes, vertical integration, such as the afore mentioned genome-wide gene regulatory network, can provide very useful hypotheses for proposing models that can be subsequently validated by experimentation.

In addition to reverse modelling, many attempts have been made to identify key biomolecules in SuS and to create models of molecular function. These are not within the scope of the work presented in this thesis and therefore, more information can be found in some of the excellent reviews in this field, including the one published by Hawkins and colleagues [49], and more recently in 2014 by Gomez-Cabrero et al. [42] and Conesa et al. [26].

Integrated database resources

As consequence of the fast-growing availability of omics biological data, a wide variety of data sources, databases, and web services have been created to facilitate data management, accessibility, and analysis [158]. While these specialised platforms may answer several specific research-community needs, the heterogeneity of the data types, formats, and their contents mean that significant effort is required in order to access and analyse these data from multiple sources [42].

Consequently, substantial effort has been invested in creating organism-specific or field-specific integrative databases that collect different types of omics data and to make them available in an integrated way in order to support SB research. Practically every large international consortium in SB has produced an organism-specific database or repository that acts as a reference source for the genomic information and resources collected for the targeted species; some of these resources contain mostly genomic, annotation, gene expression, and variant information. Representative examples include the UCSC Human Genome Browser [113], which contains a large collection of genomes and annotations, as well as several tools for querying, visualising, and retrieving data; the Ensembl Genome Database [157], a huge repository that collects genes, variations, sequence conservation data, and other types of annotation for hundreds of species, including vertebrates, fungi, bacteria, and plants; and the Encyclopaedia of DNA Elements (ENCODE) Project [32], a worldwide consortium for cataloguing the functional elements encoded in the human genome, including genes, transcripts, and transcriptional regulatory regions, together with their associated chromatin states and DNA methylation patterns.

Alternatively, other solutions focus on building unified environments around distributed data sources. Typically, data warehouses provide interfaces (e.g. web-services) for querying and retrieving the stored data so that third-party services can search data from different sources, assemble the results, and present them in an integrated format [96]. BioMart [120] provides an easy-to-use web-based system for performing complex queries across a variety of genomic databases; in addition, the system includes several tools, application programming interfaces(APIs), and data services that allow researchers to retrieve information directly from their own programs. Pathway Tools [67] and InterMine [121] are both services that enable the creation of biological databases, and which provide tools for integrating data from many common biological data sources and formats as well as for accessing the data using sophisticated web-based queries. Some data warehouses powered by these tools are FlyMine [83], an integrated database for Drosophila and Anopheles genomics, and HumanCyc [112], which provides an encyclopaedic reference on human metabolic pathways and integrates data from different sources.

Integrating such diverse data creates several problems that usually hamper the process. Probably the most obvious integration problem is that, as a consequence of its heterogeneity, data in SB is highly context-dependent [22]. For example, data from gene expression is meaningful only in the context of the conditions under which they were generated, hence, reliable integration requires a detailed record of the *meta-data* that discriminate each dataset; however, this can only be accomplished by adopting community standards for the data schemas, formats, nomenclatures, and protocols. Standardisation is critical for allowing data exchange and interconnection and, consequently, this has become an important field in the context of SB. Thus, numerous attempts have recently been made to define a set of rules widely accepted by the scientific community. The various *minimal information recommendations*, such as the minimum information about a microarray experiment (MIAPE) [18], minimum information about a high-throughput sequencing experiment (MINSEQE) [131] and the minimum information about a proteomics experiment (MIAPE) [129], as well as the ontology efforts, such as that of the Gene Ontology Consortium [8], are good examples of these attempts [41].

Integrative visualization of omics data

Integrated visualisation of different omics data types is probably the most powerful tool for the interpretation of SB results. While mathematical models can reveal significant associations between system components or can predict the behaviour of the system, graphical display of omics data can lead to better insights into its global functional properties. Many software tools are now available that accept diverse omics measurements and generate joint visual representations of the data, which itself can be grouped based on different criteria. Among these tools, the first major distinction is whether it is devoted to displaying genome or network information.

Genomic information is often visualised using genome browsers(GBs). GBs are interactive tools that usually display the different layers of information as customisable tracks which are continuously distributed along the chromosome coordinates (Figure 2.6-A). Examples of displayable information include annotations for gene sequences and genomic variants, as well as dynamic features such as ChIP-seq, RNA-seq, or methylation tracks from a particular biological sample. Some of the most popular databases (mentioned above) incorporate their own web-based GB that allows researchers to explore the stored datasets. For example, the UCSC GB, developed by the University of Californica Santa Cruz [70], collects a wide range of annotation datasets, including hundreds of human and mouse datasets, from the ENCODE Project, and the Ensembl Genome Browser [157] acts as a single point of access to all of these annotated genomes from within the Ensembl databases. Another interesting web-based genome viewer is Genome Maps [88] which allows users to upload large volumes of high-throughput sequencing data and is particularly well suited to the analysis of large data collections, such as cancer or population studies; the data is locally cached and visualised in real time on the client side. Finally, other GBs such as the Integrative Genomics Viewer (IGV) [136] run on the user's computer as standalone applications, allowing offline viewing and analysis of local genomic datasets.

A very different visualisation approach is required to represent the interactions between molecular features, such as genes, proteins, and metabolites. These data are best analysed in the form of graphs, where nodes represent features and

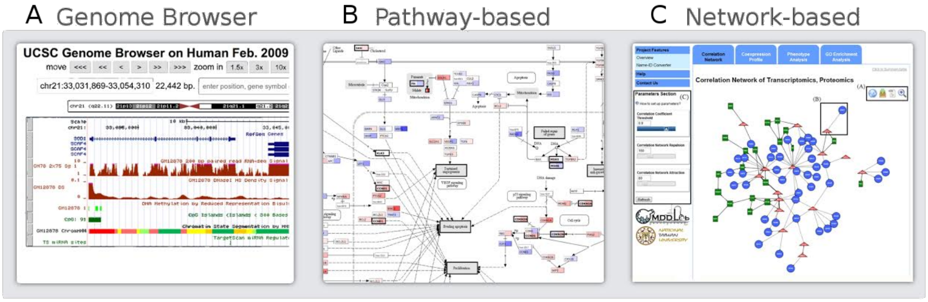

A Genome Browser
B Pathway-based
C Network-based
UCSC Genome Browser on Human Feb. 2009
Correlation Network of Transcriptomics, Proteomics

Figure 2.6: Overview of the three types of tools for integrated omics data visualisation. (A) - Genome browsers display user tracks to display measurements and annotation along genome coordinates (image from UCSC GB [70]). **(B)** - Pathway-based methods use predefined pathways to map multi-omics data and show them jointly (image from PaintOmics [40]). **(C)** - Network-based methods build data-specific networks, integrating relationships between different types of features (image from 3Omics [75]).

edges indicate an interaction between these features (Figure 2.6-C). Popular tools such as Cytoscape [118] provide a general framework for representing such biological networks; they can display virtually any feature, are linked to existing databases, and are visually enhanced with different graphical resources. Additionally, Cytoscape also offers many add-ons that allow further analysis of topological or functional properties of the network.

Another set of tools incorporate extensive analysis options and focus on the discovery of integrated networks. For example, 3Omics [75] can accept a sufficient number of proteomics, metabolomics, and transcriptomics data samples to enable it to compute pairwise correlations and to create correlation networks containing elements of all the three types of molecular data. VANTED v2 [111] is a comprehensive omics integration framework that supports different types of topological, functional, and statistical analysis on the supported data types, and can also run simulation tasks on a predefined network in order to study behavioural aspects of the network. SteinerNet [141] accepts proteomics and transcriptomics data and maps these to a database of protein-protein interaction networks and transcription factor target-gene interactions to create a network that maximises the connectivity of the submitted data, incorporating additional database elements when needed.

Finally, other tools map the data (such as metabolic or signalling pathways) provided by the user onto interaction templates that guide visualisation and facilitate interpretation (Figure 2.6-B). For example, KaPPa-View [138] and MapMan [135] display metabolite and transcript information on predefined pathway blocks. The MassTRIX software [124] translates NMR spectra into metabolic compounds and maps them onto Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways together with genome and transcriptome information, and PaintOmics [40] represents multiple gene expression and metabolomics datasets as KEGG pathways and performs joint pathway analysis, taking both types of data into consideration. However, as these solutions use a known scaffold to analyse omics data, they are limited in their network inference and data mining functionalities, which is usually restricted to some kind of functional enrichment analysis.

3

Managing multi-omics experiments: the STATegra Experiment Management System

Part of this chapter have been published in "Hernández-de-Diego R, Boix-Chova N, Gómez-Cabrero D, Tegner J, Abugessaisa I, and Conesa A. *STATegra EMS: an Experiment Management System for complex next-generation omics experiments.* BMC Systems Biology, 8 Suppl 2:S9, 2014".

3.1 Introduction

The widespread availability of high-throughput sequencing techniques has had a big impact on genome research and has reshaped the way we study genome function and structure, and the rapidly decreasing costs of sequencing have made these technologies affordable to small and medium size laboratories. Furthermore, the continuing development of novel sequencing-based assays, referred to with the suffix *-seq*, has expanded the scope of cell properties which can be analysed by high-throughput sequencing, making sequencing reads the new underlying common data format. Today, virtually all nucleic acid omics methods traditionally based on microarrays have a *-seq* counterpart and many more have recently become available. As a consequence, it has become a lot more practical to run multiple sequencing-based experiments to measure different aspects of gene regulation and to combine these with non-sequencing omics technologies such as proteomics and metabolomics [9, 17, 115, 122, 152]. For example, the ENCODE project [32] combined ten major types of sequencing-based assays in order to unravel the complexity of genome architecture. Many records can be found at the Sequence Read Archive (SRA), a resource which integrates multiple sequencing technologies measured in the same samples; similarly, searching PubMed for "next-generation sequencing" plus "proteomics" or "metabolomics" returns more than a hundred entries. Finally, one of the advantages of sequence-based experiments is that they are equally applicable to the study of both well-annotated model organisms and less-studied non-model organisms because little or no a priori genome knowledge is required.

However, sequencing-based assays also bring new data-processing and storage challenges. The memory requirements for a medium-sized sequencing experiment exceed the capacity of current regular workstations, and at the same time, the analysis steps required to go from raw to processed data are becoming increasingly complex and memory-intensive. As the number of datasets grows, the need to properly store and track the data and its associated metadata is becoming more pressing. For example, a medium-sized RNA-seq experiment ranging from 4 to 20 samples of 20 million reads each may produce 40 GB of raw data and generate multiple quality control and intermediate processing-step files, occupying a total of up to 500 GB of memory. Laboratory information and management systems (LIMS) or sample management systems (SMS) are bioinformatic tools that help experimentalists to organise samples and experimental procedures in a controlled and annotated way. There are several dedicated LIMS, both commercial and free, that have been developed specifically for genotyping labs where thousands of samples are processed by automated pipelines and procedures are tightly standardised [14, 48, 145]. One popular LIMS for genomics is BASE [143]. This software includes a highly structured system for metadata annotation and flexible architecture for defining experiments and incorporating analysis modules. However, BASE is currently limited to microarray experiment annotation.

Several LIMS have been specifically developed and implemented at different sequencing facilities in order to manage the large volume of samples and data they routinely handle. Some of these have been made available to the scientific community or exported to other centres, such as the Leeds University DNA sequencing facility where they have been published as an extension of the Protein Information Management System (PIMS), designed to provide sample tracking both to users and operators [139]. The system allows facility users to place orders and monitor the processing status of their samples while a different interface provides operators with full control of the sequencing pipeline with automated connection to the sequencing robots. The Leeds system supervises the whole procedure from sample submission to generation of FASTQ files but does not track the actual experimental characteristics of the sequenced samples or post-processing of the raw data. Other solutions track the sequencing-samples via analysis-modules and execute some raw-data processing steps such as quality control analysis or reference-genome mapping. For example, QUEST software [21] uses an experiment-resolved configuration file to store experiment metadata and execute predefined processing pipelines. Another example is NG6 [87], an integrated next-generation sequencing (NGS) storage and processing environment where workflows can be easily defined and adapted to different data input formats. NG6 can be used interactively to generate intermediate analysis statistics and downloadable end results. Similarly, Scholtalbers et al. [116] recently published a LIMS for the Galaxy platform that keeps track of input-sample quality and organises flow cells. By working within the Galaxy system, the associated FASTQ files are readily available for processing using the platform's analysis resources. Finally, another interesting package is the MADMAX system that considers multiple omics experiments by incorporating modules for microarrays, metabolomics, and genome annotation [79]. MADMAX uses an Oracle relational database to store sample and raw data, and to facilitate data analysis it links to common bioinformatics tools such as Blast or Bioconductor when they are installed on a computer cluster.

In this chapter, we describe the STATegra Experiment Management System (EMS), which is an information system for storing and annotating complex multi-omics experiments. In contrast to other solutions that put the focus on managing thousands of samples for core sequencing facilities, the primary goal of the STATegra EMS is to annotate the experiments that are designed and run at individual research laboratories. The system contains modules for defining omics experiments, samples, and analysis workflows and it can incorporate data from different analytical platforms and sequencing services with great flexibility; it currently supports mRNA-seq, ChIP-seq, DNase-seq, Methyl-seq, miRNA-seq, proteomics and metabolomics by default, and can be easily adapted to support additional high-throughput experiments.

3.2 Methods

The STATegra EMS is a multiuser web application developed using free, open source software technologies such as Java Servlets, the Sencha ExtJS framework, the MySQL relational database system, and the Apache Tomcat Servlet engine, and is freely available to the scientific community.

3.2.1 STATegra EMS architecture

The architecture of the STATegra EMS follows the Client-Server paradigm, which divides the system into two main components: the SERVER-SIDE application and the CLIENT-SIDE web application (Figure 3.1). While the SERVER-SIDE is responsible for maintaining data consistency and for controlling access to the stored information, the CLIENT-SIDE must request data from server and suitably present it to the user. Clients typically communicate with server using a request-response communication pattern: the client requests data from the server, the server receives and processes the petition, fetches the information required (e.g. from databases), and returns a response to the client (Figure 3.2).

For the STATegra EMS, the communication between the client and server is mainly handled using asynchronous JavaScript and XML (AJAX) techniques and the data is encoded using JavaScript object notation (JSON). With AJAX, web applications can send and receive data asynchronously from the server and update specific parts of a web page which allows them to create very dynamic web interfaces without reloading the whole page, as usually occurs with classic web pages.

Finally, there are some important advantages to using this client-server architecture model compared to more traditional approaches, the most important of which are:

- *Centralisation*, because data are stored and retrieved from a single source it is easier to back up and manage error and, more importantly, data duplication is reduced.
- *Security*, the server-side can control the way clients access data by using different levels of permissions. Centralising data may also require strategies to be developed to handle concurrent access to data, for example, when different users edit the same information at the same time.
- *Accessibility*, server data can be accessed remotely across multiple platforms .

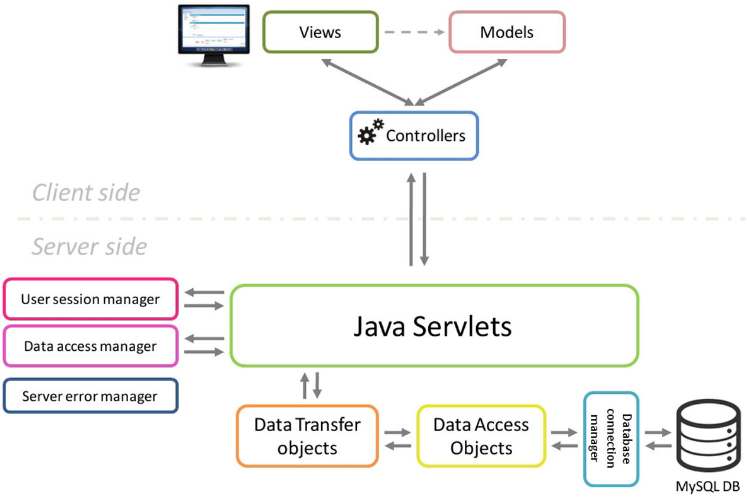


Figure 3.1: Overview of the STATegra EMS architecture.

STATegra EMS server application

The STATegra EMS server-side was built using Java servlets and MySQL relational databases and is unique for all clients. Although the application was primarily designed and tested on UNIX servers, the server EMS code could easily be adapted to work over other architectures (e.g. Windows systems) due to the cross-platform nature of Java. Java servlets provide a *component-based* and *platform-independent* interface for developing web-based applications, keeping all the benefits of the Java language (portability, re-usability, performance and robustness). A servlet is a Java class designed for responding to client requests, usually HTTP requests, with dynamic content, and are deployed and executed in a servlet container (Figure 3.3). Some examples of servlet containers are Apache's Tomcat server, the official reference implementation for a servlet engine developed by the Apache Software Foundation; and Jetty, a web server container developed as a free and open source project as part of the Eclipse Foundation [56, 101, 105].

In terms of the connection with the databases, the server code was implemented using the data access object (DAO) design pattern in conjunction with the data transfer object (DTO) pattern. Using both patterns provides an abstraction layer for interaction with databases, and works as an intermediary between the server application (servlets) and the MySQL database. The main objective for the DAO pattern is to provide a set of data operations (e.g. inserting,

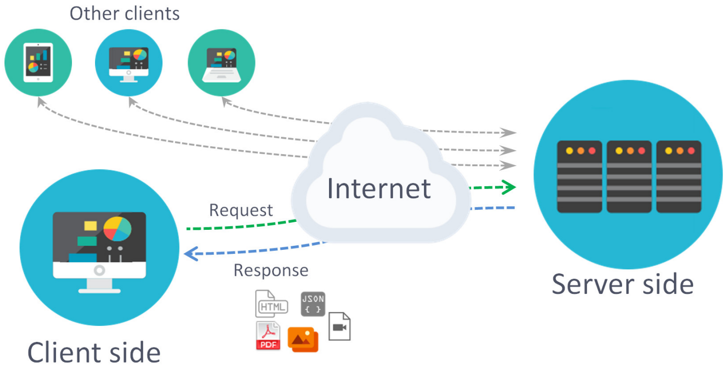


Figure 3.2: An example of a request-response exchange between clients and servers. Multiple client programs may share services from the same server. Responses can include different information such as images, text files, or HTML files.

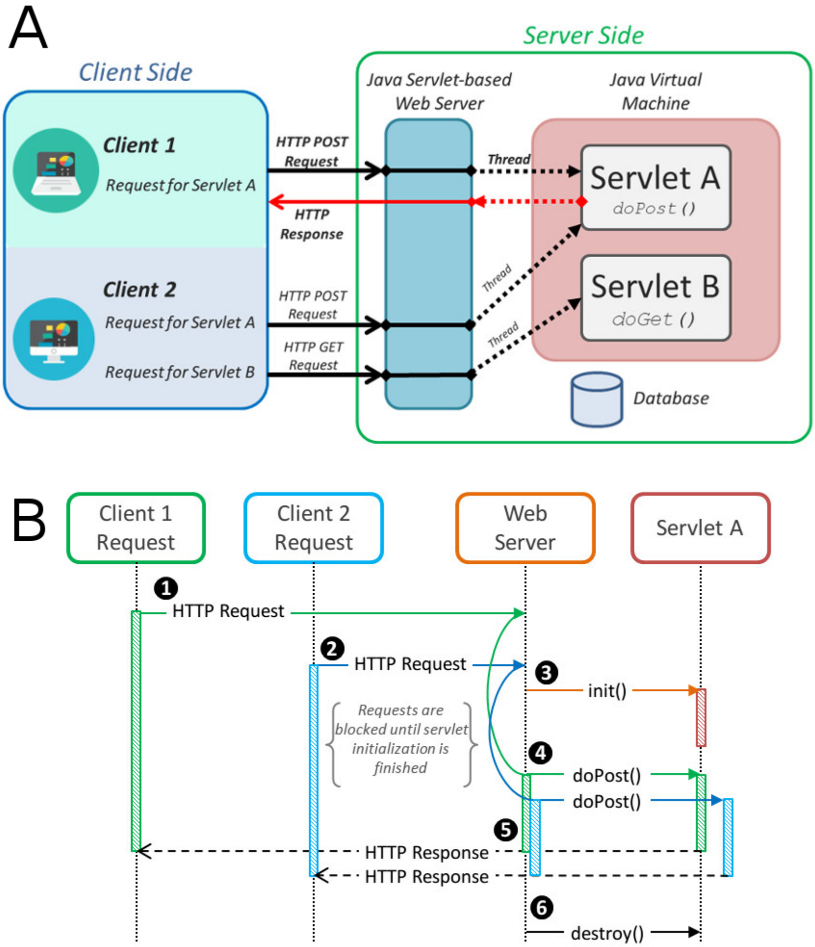


Figure 3.3: Example of a basic interaction between clients, a web server and a servlet registered within the web server. A client application (Client 1) sends a request to the server **(B.1)**. A second client (Client 2) could send another request which would be processed in parallel **(B.2)**. The web server receives the request and identifies the destination servlet, i.e. the servlet that will accept the request from clients and that will return a response (Servlet A). The first time that a server loads a servlet, it runs the servlet's *init* method **(B.3)**. Once the servlet is initialised, it is able to handle the client requests using its *doPost* or *doGet* method within the Java virtual machine (depending on the HTTP method used, i.e. *POST* or *GET*). Each client's request involves a call to the *doPost* or *doGet* method **(B.4)**, which results in a HTTP response that the server returns to the client *(B.5)*. Lastly, servlets run until they are removed from the server by running the servlet's *destroy* method **(B.6)**.

removing, or querying a database) that allows such access to the database without exposing the details of the database schema.

The advantage of the DAO pattern is that by introducing this separation between two important parts of the application (business logic and access to the database) changes in the underlying persistence mechanism only affect the DAO implementation and leaves the rest of the application unaffected [4, 126]. This helps to reduce the workload if there are any changes made to the database model, if the application code is extended with new features, or if alterations are made to the database engine as a consequence of changes in the existing omics approaches or the emergence of new techniques.

The DAO pattern can be complemented using a DTO: a simple object that does not have any behaviour except for holding and retrieving its own data by using *accessors* and *mutators*. The DTO encapsulates the data in an object that is transferred between the server-side processes, thus considerably reducing the number of method calls [37, 89]. In addition to the *accessors* and *mutators*, DTO is also responsible for serialising its data into specific formats such as XML or JSON, which are usually transferred out of the application (e.g. when a client application is used). Figure 3.4 shows an example for the interaction between business and DAO objects in the STATegra EMS.

STATegra EMS client application

The entire STATegra EMS client side was developed using JavaScript and HTML. The core of the application was built using Sencha ExtJS [117], a comprehensive JavaScript framework for building rich cross-platform web applications. ExtJS simplifies the generation of interactive and user-friendly interfaces by including multiple graphical resources such as panels, grids, and form controls, as well as other valuable features including a flexible layout manager that helps to organise how the content is displayed across multiple browsers or devices, and a charting package that allows data to be visually represented with a broad range of chart types. The JavaScript library JQuery [132] was used to complement this framework. JQuery is an easy-to-use API for the manipulation of the HTML documents and simplifying the use of animations,

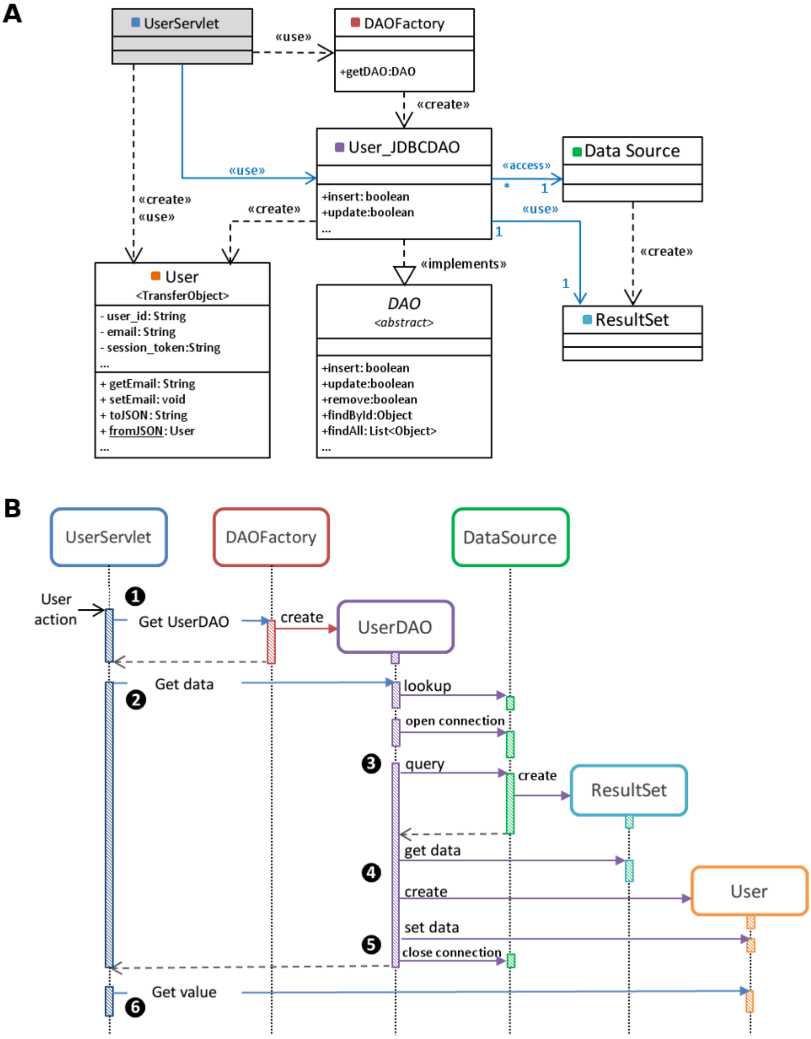


Figure 3.4: Class diagram (A) and sequence diagram (B) for an interaction between Servlets, DAO, DTO, and data sources in the STATegra EMS. When the servlet is required to retrieve data from the database (B.2), it first receives an instance of the corresponding DAO using a "factory", which wraps the creation of the objects (B.1). DAO classes are implementations of the DAO abstract class, which defines the set of methods available for manipulating the information at the data source. For example, in diagram A, the class *UserJDBCDAO* implements the abstract methods in order to manipulate the data for users stored in a SQL-based database, using the JDBC API. The DAO object sends a valid query to the data source (B.3), which usually returns an iterable set of records that represents the results for executing the statement at the data source. Before exchanging information between the servlet and the DAO, it is usually compiled as an implementation of the DTO pattern, in this case, an instance of the User class (B.5 and B.6).

event handling, or AJAX communications. The client-side is built following the model-view-controller (MVC) architectural pattern, which makes it easier to organise, maintain, and extend large client applications. The MVC pattern separates data modeling, visual representation, and the action's handlers into three separate but interconnected parts [19, 102]:

(i) *Model*, stores the application's data in a structured way. A model can be a single object or a more complex structure, and typically includes methods for manipulating, retrieving, and validating the stored data (*accessors* and *mutators* methods).

(ii) *View*, is a visual representation for the application's data. A view usually represents the information for a given model or set of models, and multiple views can represent the same model in different ways. Views usually have an associated *controller* and may also contain tools for manipulating the information for the represented model, such as buttons or text fields. However, a view does not directly change the model's data, rather, it delegates the action to its associated *controller*.

(iii) *Controller*, works as an intermediary between the *models* and *views*; it updates the view when the model changes, includes event handlers for the views, and updates the model when the user manipulates a view.

An important factor in the MVC architectural pattern design is use of the Observer design pattern in which an object (known as a *subject* or *observable*) maintains a list of objects which depend on it (*observers*) [102]. In the context of the MVC pattern, the models will be the observable object and the views will be the observers. When a model changes, it typically notifies its observers that a change has occurred and they need to update their model's representation in order to make the change visible (Figure 3.5).

3.2.2 User system

An important aspect when developing web applications, especially when the data stored by the tool are sensitive in terms of privacy and confidentiality, is the use of security constraints to keep data safe and away from undesired access.

In general terms, there are two levels of security verifications.

- **Authentication**, is probably the single most common requirement of any application. Authentication is the process of verifying that an individual is a valid and trusted application user. In a traditional web application, this is usually done using server-side session tracking where each registered user in the application has a user identifier (ID) and a password [30]. The STATegra EMS included a Session Management system which controls access to the application as follows:

  (i) Before users can start working with the application they must create a new user account.

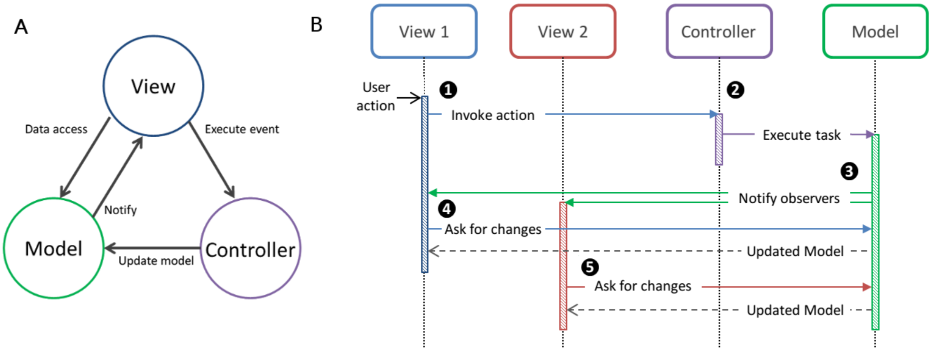


Figure 3.5: Diagram for the Model-View-Controller interaction. When a user acts on a view (e.g. by pressing a button on View 1, **B.1**), the view propagates the event to its associated controller. The controller handles the event and executes the corresponding task (**B.2**) which may have an effect on a model or a set of models. When a model detects changes in its data, it notifies all of its known observers (i.e. views that are showing its data) about the changes (**B.3**). Each observer that receives the notification (**B.4** and **B.5**) asks for an updated version of the model and updates their visual representation accordingly.

(ii) With a valid account, the user can login to system. When logging in, the user is asked for the account credentials (account and password). The username and password combination is passed in an unencrypted form to the server. The server compares the values provided to the encrypted values stored in the database [27].

(iii) If the user is authenticated, the server verifies that they have the privilege to access the application and then generates a session token (a random combination of 15-30 alphanumerical characters). The new session is registered by the *Session Manager* (a singleton instance for the application) which is responsible for controlling and validating the open sessions, as well as for closing the session after a long period of inactivity.

(iv) Finally, if authentication is successful, the server returns the generated session token (stored in the browser as an authentication cookie) to the user.

- **Authorisation**, is the process of verifying that users can only perform the actions they have been given permission for by the system administrator, thus preventing unauthorised actions and limiting data access. Authorisation is managed using two different strategies in the STATegra EMS:

  (i) a privileges scale for user accounts, where the administrator accounts can access much more advanced system options compared to regular accounts, including user management and databases administration.

  (ii) Using ownership and membership as constraints for editing data. As a general rule, the user creating a data element becomes its owner and has exclusive rights to edit or delete it. However, any owner can grant access rights to other users registered on the system.

3.2.3 Data specification

The overall objective of the STATegra EMS is to serve as a logbook for high-throughput genomics projects performed at research labs by providing an easy-to-use tool for annotating experimental designs, samples, and measurements, and for analysing the pipelines applied to the data. Experimental data and metadata are organised in the EMS around three major metadata modules (Figure 3.6): the *Study* module that records experimental design information and associated samples; the *Samples* module that collects all the available information about the biological material used; and the *Analysis* module that contains analysis pipelines and results. Both Sample and Analysis modules have been broadly defined to accommodate data from different types of omics experiments and provide a common annotation framework. Commonly used standards in omics experimental data annotations were used when defining the data specifications in order to facilitate EMS interoperability. In particular, we leveraged minimum information about a proteomics experiment (MIAPE) [129] for proteomics analysis annotation, metabolomics guidelines proposed by [125] and [43], and minimum information about a microarray experiment (MIAME) [18] and minimum information about a high-throughput nucleotide sequencing experiment (MINSEQE) [131] for sequencing experiments.

Sample and *Analysis* modules contain distinct information units(IUs), which are the basic elements of data input into the system and are connected by an experimental or analysis workflow. The *Study* module wraps *Samples* and *Analyses* modules with one single data input form.

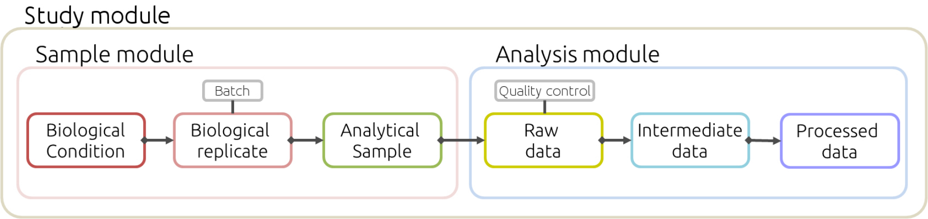


Figure 3.6: Metadata module structure in the STATegra EMS. The *Sample* module stores information about biological conditions, biological replicates, and the associated analytical samples. The *Analysis* module contains all the analysis steps required to process the raw data. Both *Samples* and *Analyses* are associated to one or more studies within the *Study* module.

(i) **Study module**: The study module is the central unit of information in the STATegra EMS. An study is defined by some scientific goals and a given experimental design that addresses these goals. This design requires the use of a number of biological samples and an array of omics measurements, which are assigned to the study.

(ii) **Sample module**. This section hosts information about the biological conditions and their associated biological replicates and analytical samples. The IUs of this module are:

Biological condition. These are defined by the experimental design and consist of a given biological material such as the organism, cell type, tissue, etc. and, when applicable, an experimental condition such as treatment, dose, or time-point for time-series samples.

Biological replicate or *sample*. Each biological condition is assessed by using one or more biological replicate. The biological replicate stems directly from the biological condition by adding a replicate number and, if applicable, a "batch number". When an study comprises a large number of samples, very often only some of them can be generated at the same time; these samples correspond to the same batch. Batch information is relevant to identify systematic sources of noise that might affect all the samples in the batch.

Analytical sample or *aliquout*. Omics experiments analyse the molecular components of biological replicates using the chosen experimental protocol to produce samples ready to be measured by the relevant high-throughput technique. For example, a RNA-seq analytical sample is obtained after using a cytosolic mRNA extraction protocol. Similarly, in metabolomics, different aliquots can be obtained by applying certain extraction protocols that target distinct metabolic compounds.

(iii) **Analysis module**. The analysis module describes the process for obtaining the measurements from the high-throughput platforms, as well as the later steps of processing and analysis of the raw data. In contrast to the sample module where only metadata is stored, the analysis module also

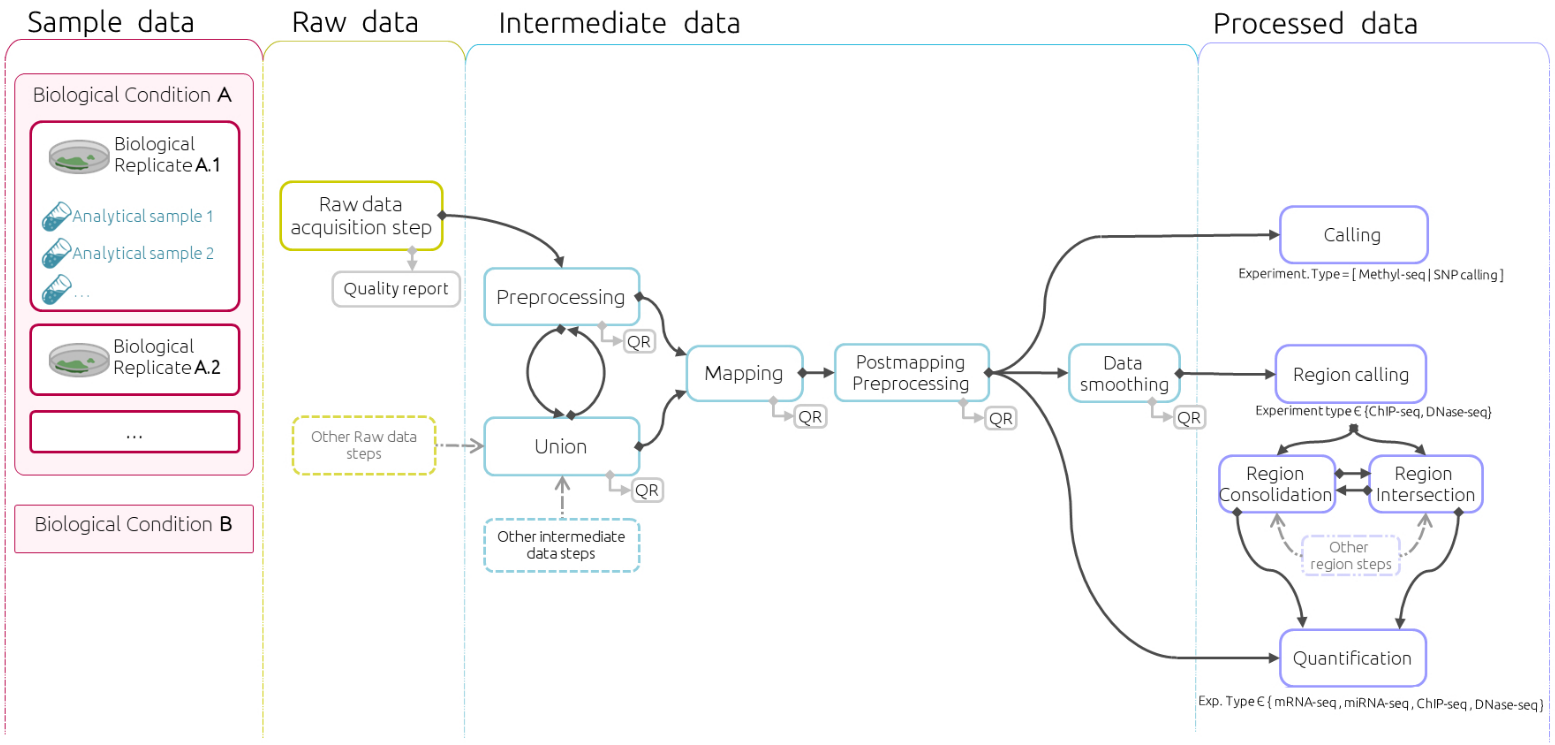


Figure 3.7: Generic workflow for sequencing-based analysis. All the steps in the workflow are optional and can be repeated several times. Steps for quality control are also optional and applied to raw- and intermediate-data steps.

stores the data files. The analysis module consists of one logical and three data IUs:

Raw data. The raw data IU provides details for the generation of the raw data, including the type and configuration of the omics equiment, or any protocol followed to prepare the analytical sample for its measurement.

Intermediate data. This IU covers all the intermediate processing steps from the raw data production to the generation of the final data. Different omics experiments might require zero, one, or several intermediate steps. For example, in the case of RNA-seq, mapping to a reference genome that produces a BAM file constitutes an intermediate step. ChIP-seq generally has two intermediate steps consisting of read-mapping and peak-calling.

Processed data. The processed data IU describes the final processing step that results in certain data files containing the ultimate signal values for the omics assay.

Analysis. The STATegra EMS includes an additional analysis IU constructed by connecting some of the previous data IUs to define a data processing workflow. Figure 3.7 shows a generic representation of the workflow elements used in sequencing data analyses. An analysis starts on a raw data file obtained from a particular analytical sample, continues through one or several intermediate data files covering different processing steps (such as trimming, mapping, filtering, merging, etc.), and finishes with a processed data file that contains the signal values of the omics features. Alternatively, processed data files can be inputted into an analysis which then applies additional processing steps to render higher-level processed data. For example, in DNase-seq analysis, a primary workflow would be to call DNase hypersensitivity regions by applying a peak-calling algorithm to a file of mapped reads (Figure 3.8-A); whereas a secondary *analysis* could involve merging the regions from *N* different samples to obtain a set of consolidated regions, and then counting the number of reads of each sample in the consolidated region set to generate a per-sample signal value file (Figure 3.8-B).

An analysis is always associated with one or more studies and, because the analysis workflow can be traced back to raw data and its associated analytical samples, it provides the link between the study and the sample modules. By default, when a new analysis is created, it is assigned to the active study. Figure 3.9 shows the data input window for the analysis module; the central panel displays the input form for the different analysis steps, while at the bottom a graphical representation of the workflow allows the elements and structure of the analysis to easily be monitored.

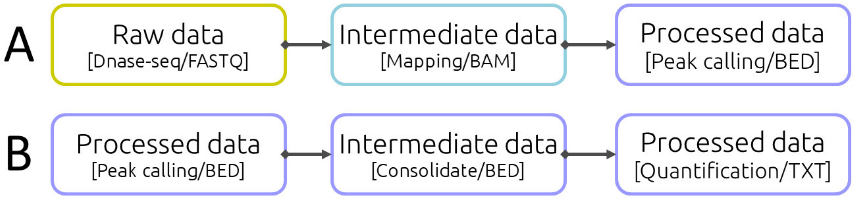
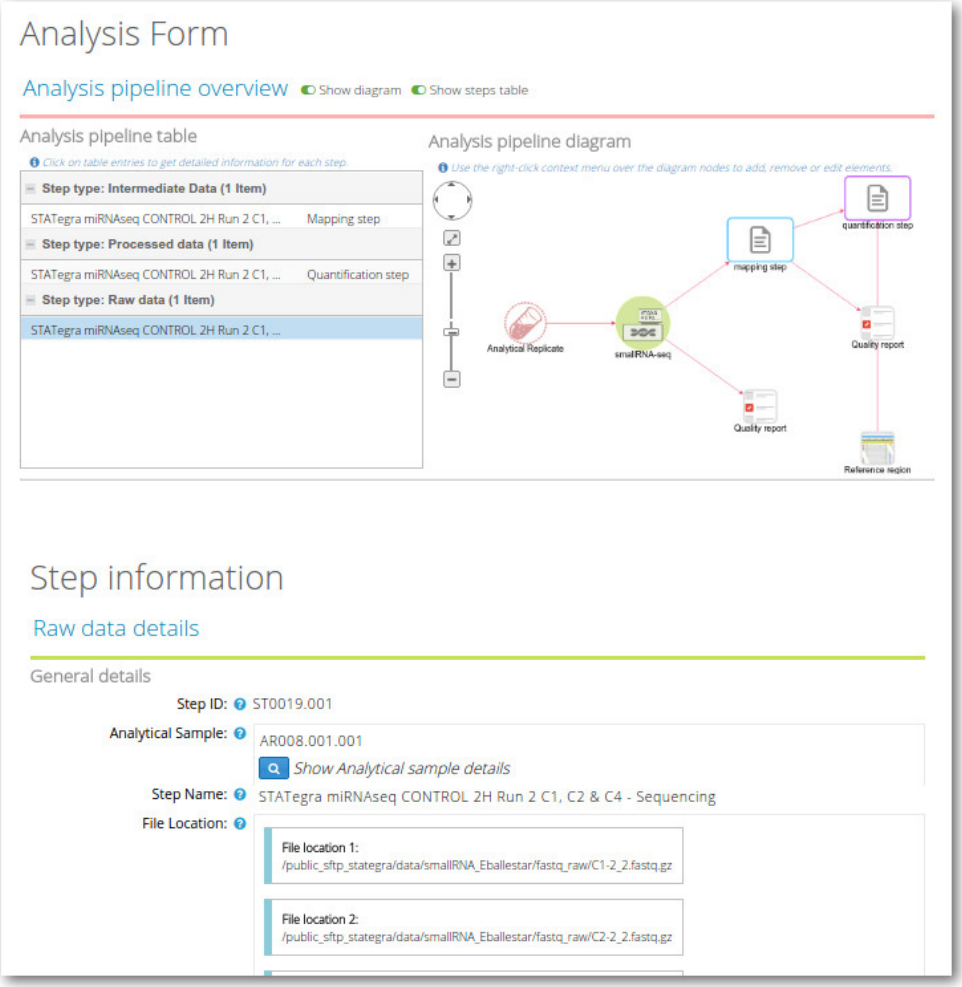


Figure 3.8: Example of a primary and secondary workflow for a DNase-seq analysis. The primary workflow **(A)** involves calling DNase hypersensitivity regions by applying a peak-calling algorithm to a BAM file of mapped reads whereas the secondary workflow **(B)** involves merging region files from different samples to obtain a set of consolidated regions and then counting the number of reads of each sample in the consolidated region set to generate a per-sample signal value file.



Analysis Form

Analysis pipeline overview

Analysis pipeline table

Click on table entries to get detailed information for each step.

Step type: Intermediate Data (1 Item)	
STATegra miRNAseq CONTROL 2H Run 2 C1, ...	Mapping step
Step type: Processed data (1 Item)	
STATegra miRNAseq CONTROL 2H Run 2 C1, ...	Quantification step
Step type: Raw data (1 Item)	
STATegra miRNAseq CONTROL 2H Run 2 C1, ...	

Analysis pipeline diagram

Use the right-click context menu over the diagram nodes to add, remove or edit elements.

Step information

Raw data details

General details

Step ID: ST0019.001

Analytical Sample: AR008.001.001

Step Name: STATegra miRNAseq CONTROL 2H Run 2 C1, C2 & C4 - Sequencing

File Location:

File location 1:
/public_sftp_stategra/data/smallRNA_Eballestar/fastq_raw/C1-2_2.fastq.gz

File location 2:
/public_sftp_stategra/data/smallRNA_Eballestar/fastq_raw/C2-2_2.fastq.gz

Figure 3.9: Input window for the analysis module.

3.3 Results

3.3.1 Availability and requirements

The STATegra EMS application is distributed under the GNU general public license version 3 and can be downloaded from the project site (*see table B.6 below*). This site also provides useful links to the documentation, news about the application, and a link to a test-instance of it. In addition to this website, sources for the application are hosted at *GitHub*, a popular web-based Git repository, which allows anyone to browse and download the code or discuss it, submit contributions, and review the code.

The documentation and user manuals for the application are hosted at *Read the Docs*, a free web platform for generating *fully-searchable* and *easy-to-find* documentation, and are imported automatically from major version-control systems such as Mercurial or Git. Documentation sources were written in Markdown, a lightweight markup language, and are stored on the GitHub repository. As previously mentioned, the STATegra EMS was developed in Java and is therefore platform independent, although it has only been extensively tested in UNIX environments. Installation instructions can be found on the *Read the Docs* site.

Availability and requirements	
Project links	
Project site:	`http://bioinfo.cipf.es/stategraems/`
Sources:	`https://github.com/fikipollo/stategraems`
Documentation:	`http://stategraems.readthedocs.org`
Other information	
Operating system(s):	Platform independent
Programming language(s):	HTML, JavaScript, Java
Other requirements:	Web browser. Google Chrome is recommended.
License:	GNU General Public License Version 3

3.3.2 Discussion

As high-throughput sequencing costs decrease and new sequencing-based molecular assays become available, more research laboratories are incorporating NGS technology as a tool for addressing their scientific goals. This has also been promoted by the fact that high-throughput sequencing is now feasible in organisms for which very little genome information is currently available. In a typical scenario, the researcher plans and outsources their experiments to sequencing facilities that might vary over time or according to the specific NGS assay required. When the sequencing results arrive and start to accumulate over several experiments, the researcher must then find ways to properly store and organise large datasets and their associated processing pipelines. In this chapter, we showed how the STATegra EMS was conceived to provide a management solution in such cases.

The architecture of the system was designed with the current organisation in research labs in mind, i.e. where multiple experiments are run, samples might be replicated or reused in successive experiments, and the same biomaterial source might be used in different types of NGS assays. For this reason, the sample module arranges annotations into three IUs: *biological condition*, *biological replicate* and *analytical sample* and permits one or many relationships between them, which is therefore sufficiently flexible to define complex sampling settings without duplicating information. Similarly, the analysis module divides metadata annotation into steps that can be reused to create alternative analysis workflows. Finally, by allowing samples and analyses to belong to different studies, the STATegra EMS can accommodate possible connections between studies.

This architecture is substantially different from other information management solutions created for NGS data which have been designed for sequencing facilities, such as the Galaxy LIMS [116] which handles requests to the service from users, or the NG6 [87] that controls the sequencing workflow the level of the sequencing providers. In these cases the management system is adapted to the production pipeline at the sequencing centre and applies strong control limits to the facilities wet-lab, including library preparation and sequencer runs. This type of information is absent from the STATegra EMS, which may even be able to accept data from multiple sequencing providers. On the contrary, the STATegra EMS records experimental information and sample metadata that might not be relevant in a production centre. In conclusion, NGS LIMS and the STATegra EMS target different users and needs in sequencing data management. The remaining challenge is how to best optimise the integration of seq data with clinical information, in a similar way to current practices in clinical development centres [1, 2].

The current STATegra EMS supports analysis workflows for five popular sequencing functional assays but can easily be extended to other *-seq* applications because it uses generic processing step forms for DNA and cDNA high-throughput sequencing. Additionally, the system supports the annotation of omics experiments, targeting non-nucleic acid components such as proteomics and metabolomics, for which specific input forms have been incorporated. In summary, the STATegra EMS provides an integrated system for annotation of complex high-throughput omics experiments in functional genomics research laboratories.

4

Experiment annotation using the STATegra EMS: an example of its use in Systems Biology

Part of this chapter have been published in "Hernández-de-Diego R, Boix-Chova N, Gómez-Cabrero D, Tegner J, Abugessaisa I, and Conesa A. *STATegra EMS: an Experiment Management System for complex next-generation omics experiments.* BMC Systems Biology, 8 Suppl 2:S9, 2014".

4.1 Introduction

The objective for this chapter is to illustrate how the STATegra EMS can be used in the context of real multi-omics biomedical research. For this use case we consider the data generated by the STATegra project [134], an 7th Framework Programme (FP7) European Consortium that aims to develop new bioinformatics methodologies and tools for the integrative analysis of multi-omics datasets.

4.1.1 The STATegra project

As discusses by Gomez-Cabrero et al. [42], data integration is becoming very common in life sciences research. In just a few years the rise of new omics technologies, as well as the funding of large-scale consortia projects, has increased biological systems research on an unprecedented scale, generating heterogeneous and often large data sets. Although these multi-level data provide new insights into different aspects of genomic regulation, they also represent an important challenge in the analysis and manipulation of the data. Consequently, the scientific community is investing a lot of time and effort in the design of novel methodologies, approaches, and frameworks that allow meaningful knowledge to be extracted from this overwhelming amount of data [95].

The objectives for the FP7 European STATegra project [134] include the development of a new generation of statistical methods and tools for the integrative analysis of multiple types of omics data. The STATegra consortium comprises 11 teams from 8 countries, each with backgrounds in different aspects of biomedical research such as bioinformatics and biostatistics, omics technologies and experimentation, and commercial software development.

The STATegra project [134] has developed a wide variety of statistical methodologies and software implementations targeting different aspects of the integration of multi-omics data such as the design of multi-omics experiments, integrative variable selection, data fusion, integration of public domain data, and integrative pathway and network analysis. By combining these different approaches researchers can get a more comprehensive view of how the observed element or phenomena, also known as system under study (SuS), behaves at the different molecular layers measured.

4.1.2 Studied system and experimental design

Based on the positive results shown by Ferreiros et al. [35], the STATegra consortium chose the differentiation process of mouse pre-B-cells as a model biological system to generate experimental datasets for developing the methods. The model describes the differentiation of the mouse B3 cell line (cycling pre-B cells, Figure 4.1-A) under the controlled induction of the Ikaros transcription factor (TF). The differentiation is controlled by a tamoxifen-inducible vector of the Ikaros TF (Ikaros-ERt2), while control cells carry an empty vector (Figure 4.1-B). This model is of special clinical interest because the genetic deletion of Ikaros can result in severe disturbances or even completely block B-cell development [35].

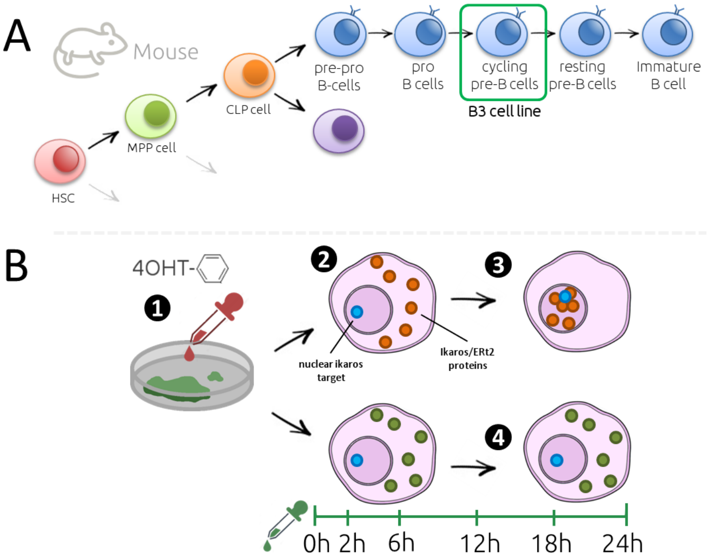


Figure 4.1: The biological model system used by the STATegra project [134]. Figures based on Ferreiros et al. (2013) [35]. **(A)** - B cells develop from hematopoietic stem cells (HSC) that originate in bone marrow. HSC first differentiate into multipotent progenitor (MPP) cells and subsequently into common lymphoid progenitor (CLP) cells, which can either differentiate to T-cells or B-cells. B-cell differentiation occurs in several stages, starting from early pro-B cells, which become pre-B cells, and finally turn into immature B cells. **(B)** - Changes in cells after the addition of 4-hydroxy tamoxifen [4-OHT] (B.1). Ikaros-ERt2 proteins are initially tethered in the cytoplasm (B.2) but the addition of 4-OHT triggers their translocation into the nucleus where they interact with Ikaros' targets (B.3).

The experimental design consisted of a replicated time course using seven different omics platforms: mRNA-seq, miRNA-seq, ChIP-seq, DNase-seq, reduced representation bisulfite sequencing (RRBS-seq), proteomics, and metabolomics. Between three and eight samples were extracted per condition (Ikaros and control), at six time-points after the tamoxifen induction (0h, 2h, 6h, 12h, 18h, and 24h), depending on the omics type (Table 4.1). For each omics type, data was acquired, normalised, and pre-processed.

The selected omics data types allow the system to be studied from different but complementary points of view. As explained in Section 2.3, transcriptomics (mRNA-seq) and miRNA-seq focus on profiling the expression of genes in the different conditions of the experiment. DNase-seq provides insights into the genomic regulatory processes based on the genome-wide sequencing of regions

	0h		2h		6h		12h		18h		24h		Total		
mRNA-seq	3	3	3	3	3	3	3	3	3	3	3	3	18	18	36
miRNA-seq	3	3	3	3	3	3	3	3	3	3	3	3	18	18	36
ChIP-seq	2*	2*									2*	2*	4	4	8
DNase-seq	3	3	3	3	3	3	3	3	3	3	3	3	18	18	36
RRBS-seq	3	3	3	3	3	3	3	3	3	3	3	3	18	18	36
Proteomics	3	3	3	3	3	3	3	3	3	3	3	3	18	18	36
Metabolomics	8	8	8	8	8	8	8	8	8	8	8	8	48	48	96
	25	25	23	23	23	23	23	23	23	23	25	25	142	142	284
	50		46		46		46		46		50		284		

🟥 Ikaros-ERt2 🟦 Control ❑ No data

* Samples from Ferreiros et al. (2013)[35]

Table 4.1: Summary of the samples used in the STATegra project [134]. Samples were grown as a batch and, as general rule, three samples were isolated for each condition (Ikaros and control), time point and omics type. Eight samples were isolated per time point and condition for metabolomics data, and four were used for LC-MS and GC-MS measurements. Data from previous studies [35] were used for the ChIP-seq analysis.

sensitive to cleavage by DNase I, and ChIP-seq identifies the binding sites of TFs and other chromatin-associated proteins interacting with the deoxyribonucleic acid (DNA). RRBS-seq determines the pattern of DNA methylation; when located in a gene promoter, DNA methylation typically represses gene transcription. In contrast, proteomics data identifies and quantifies proteins and, at the quantitative level, proteomics reflects the effects of post-transcriptional regulation in controlling gene expression. Finally, studying changes in metabolite profiles provides information about the physiology of the cell. Metabolites represent the end products of cellular processes and are directly regulated by the presence or absence of specific proteins. Consequently, the integration of all these omics data types provides a comprehensive picture of changes in the SuS at the genomic level for the different conditions.

4.2 Project annotation using the STATegra EMS

The variety of techniques used by STATegra project [134] made necessary the development of a platform for the standardised annotation, storage, and management of the huge amount of data generated during the study, the STATegra EMS. The objective in this section is to describe a real example of the system being used; however, due to the wealth of the information produced, and in order to make this section more readable, some aspects of the data annotation have been summarised or omitted, while still providing enough information to fully understand of the scope of the application.

4.2.1 Installing the STATegra EMS

From version 0.6, the STATegra EMS has included a command-line auto-installer which simplifies system deployment and configuration. Assuming that the host machine meets all of the requirements detailed in Section 3.3.1, the script will automatically perform most of the installation steps, such as checking dependencies, and downloading and deploying the binary files. Once the required files are installed, the process continues via a web-based interface which allows some of the application's options to be configured (such as the database credentials, the location for the data directory, or the administrator password) before it is launched for the first time. Additional information about the installation process is available in `http://stategraems.readthedocs.io/en/latest/installation/install/`.

4.2.2 User registration

By default, the application includes a special user account that corresponds to the administrator role. The administrator has some extra privileges compared to other users, including the ability to back up the databases, delete information, and manage the users in the system. New users can easily register using the sign up form, and are described by an username, an email address, and a password. This user system provides the application with ownership control, which avoids undesired changes to the annotations.

4.2.3 Study annotation

As explained in previous sections, the annotation process for biological studies can be divided into three major levels: annotation of the study, annotation of the biological material, and annotation of the analytical processes. The process starts with the annotation of the STATegra project [134] as a new study. On the side menu, we choose the option "Annotate new study" which opens the form for describing the study. First, we fill in the main details of the study, providing a title (*The STATegra project*), a general description of the objectives, some public references or links, and any other relevant information. The field "Data directory" allows users to specify the location of the project files. These files can be stored on the same machine running the STATegra EMS or in an external file system (e.g. an FTP server or an iRODS system) and the application automatically inspects the directory and shows its contents during later steps in the annotation.

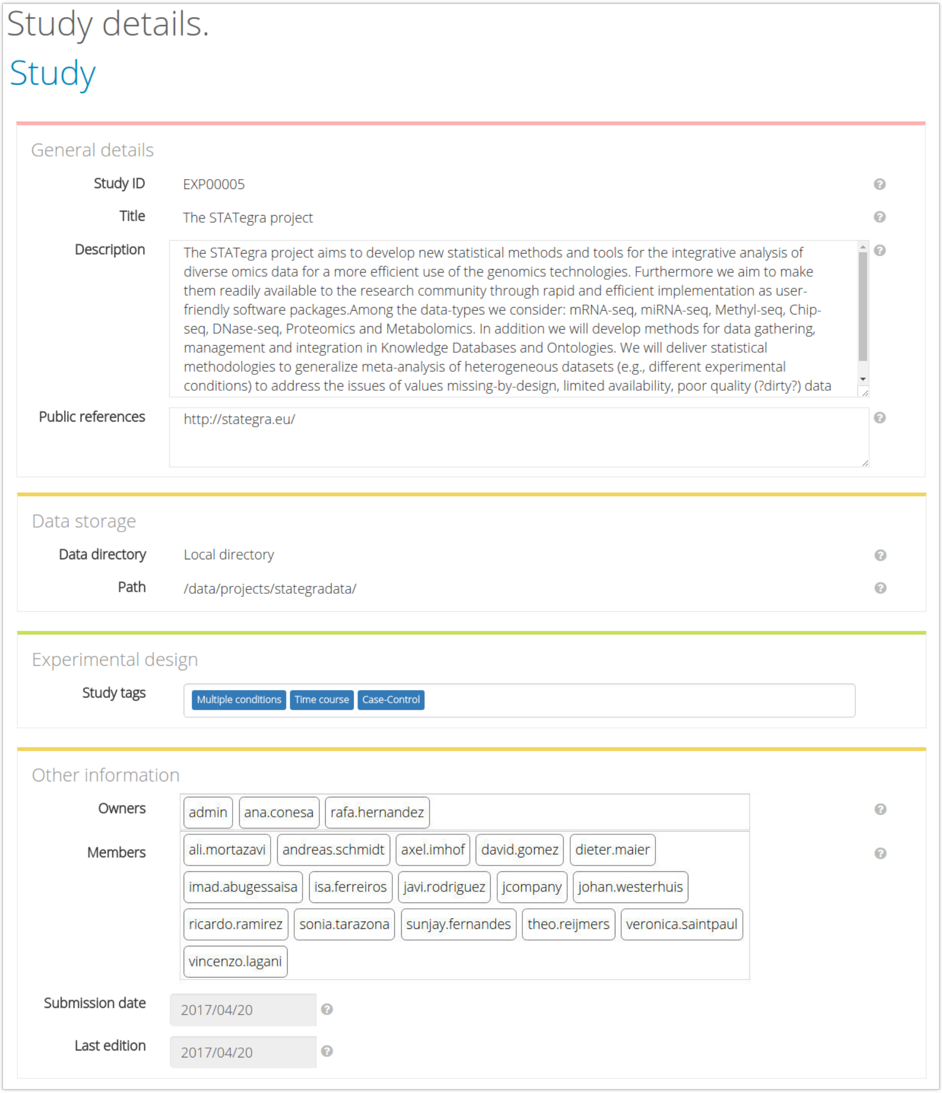
Next, we fill in the "Design summary" fields. For "Type of experiment" we choose *Time course*, *Case-Control*, and *Multiple conditions*, and we indicate the planned omics measurement types (*mRNA-seq, small RNA-seq, ChIP-seq, DNase-seq, Methyl-seq, Proteomics*, and *Metabolomics*). Check boxes next to each planned measurement are available to monitor the progress of the study. These are automatically checked when a matching analysis is annotated and assigned to the study.

Finally, we proceed by defining the users that participate in the study as *owners* (users that can administer all the information in the study), or as *members* (users that can only edit their entries). After saving, a unique identifier is assigned to the new experiment (Figure 4.2).

4.2.4 Sample annotation

Once the new study has been registered we can proceed with the sample annotations. On the side menu we choose the "Browse Samples" option which displays a list of all the annotated samples in the study, grouped by biological condition (BC). From this panel, users can edit, inspect, or annotate the information for the samples.

As previously mentioned, the study comprises six omics measurements for the B3 mouse cell line. In this study, samples are grouped depending on whether they belong to the group of "Case" samples (i.e. samples that contain the tamoxifen-inducible vector of the Ikaros TF); or which belong to the "Control" samples (i.e. samples that contain an empty vector); they are also grouped



Study details.

Study

General details

Study ID	EXP00005
Title	The STATegra project
Description	The STATegra project aims to develop new statistical methods and tools for the integrative analysis of diverse omics data for a more efficient use of the genomics technologies. Furthermore we aim to make them readily available to the research community through rapid and efficient implementation as user-friendly software packages.Among the data-types we consider: mRNA-seq, miRNA-seq, Methyl-seq, Chip-seq, DNase-seq, Proteomics and Metabolomics. In addition we will develop methods for data gathering, management and integration in Knowledge Databases and Ontologies. We will deliver statistical methodologies to generalize meta-analysis of heterogeneous datasets (e.g., different experimental conditions) to address the issues of values missing-by-design, limited availability, poor quality (?dirty?) data
Public references	http://stategra.eu/

Data storage

Data directory	Local directory
Path	/data/projects/stategradata/

Experimental design

Study tags	Multiple conditions, Time course, Case-Control

Other information

Owners	admin, ana.conesa, rafa.hernandez
Members	ali.mortazavi, andreas.schmidt, axel.imhof, david.gomez, dieter.maier, imad.abugessaisa, isa.ferreiros, javi.rodriguez, jcompany, johan.westerhuis, ricardo.ramirez, sonia.tarazona, sunjay.fernandes, theo.reijmers, veronica.saintpaul, vincenzo.lagani
Submission date	2017/04/20
Last edition	2017/04/20

Figure 4.2: Annotation details for the STATegra project in the experiment module.

according to the extraction time for the sample (i.e. 0h, 2h, 6h, 12h, 18h, or 24h). In total, we will consider 12 groups of samples, as follows.

1. Ikaros-ERt2 cells 0h
2. Ikaros-ERt2 cells 2h
3. Ikaros-ERt2 cells 6h
4. Ikaros-ERt2 cells 12h
5. Ikaros-ERt2 cells 18h
6. Ikaros-ERt2 cells 24h
7. Control cells 0h
8. Control cells 2h
9. Control cells 6h
10. Control cells 12h
11. Control cells 18h
12. Control cells 24h

Additionally, samples can be sub-classified by the number of the batch they were cultured in, as well as the protocol followed for sample isolation. Table 4.1 summarises all of the samples generated for the experiment. As a general rule, for each BC, 10 biological replicates(BRs) were cultured in batches, and about 23 analytical samples(ASs) were isolated for the different sequencing assays, except for time points 0 and 24 hours when some extra samples were isolated for ChIP-seq analysis (Figure 4.3).

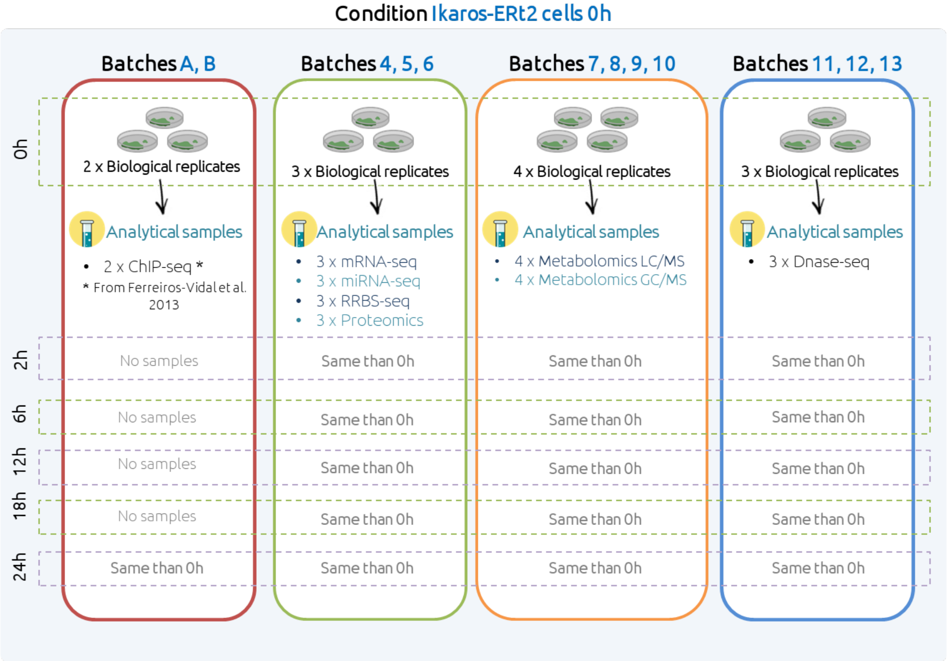
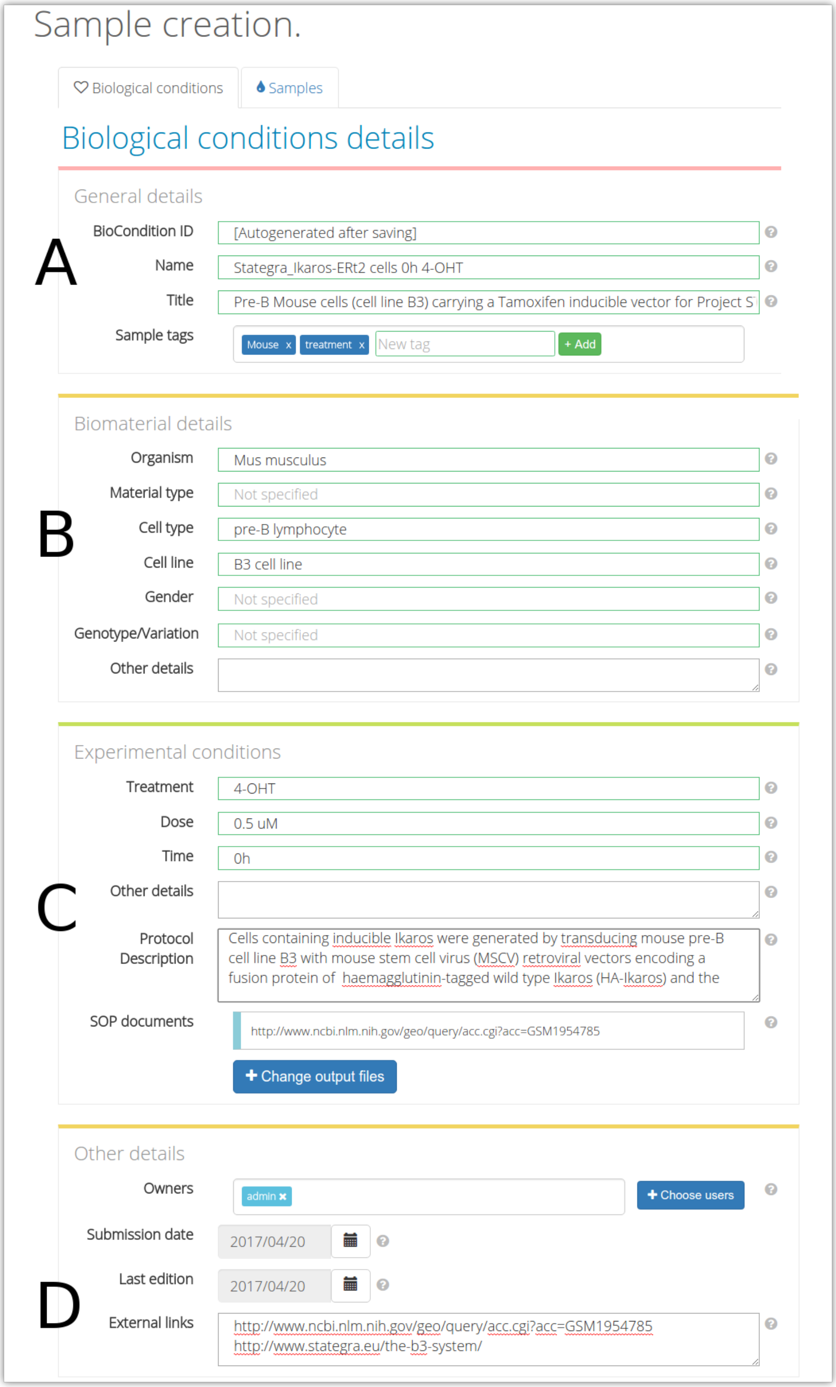


Figure 4.3: Samples for the STATegra project. Summary of sample production for the different omics types for each time point in the study. Samples were grown in batches, with 2-4 replicates for each batch. At each time point, a different number of aliquots were separated that were used for the later omics measurements (approx. 3 aliquots per omics type).

For this example we will focus on annotation of the samples corresponding to the condition *Ikaros-ERt2 cells 0h*, although annotation of the remaining samples would generally be equivalent. The process starts by clicking on the "Add new sample" option, which displays the sample module input form, for describing the samples (Figure 4.4). The first part of this form describes the common details for the BC of all the samples. The second part allows us to annotate the specific details for each BR, such as the batch number, or the ASs isolated from the culture, and the extraction protocols used to extract them.

Focusing on the first part, the first section (Figure 4.4-A), "General details", includes fields such as a human-readable name which identifies the sample (*Stategra-Ikaros-ERt2 cells 0h 4-OHT*) and a more extended description of the sample. The second section (Figure 4.4-B) describes the biological material used for the assay. In this case, we choose *Mus musculus* as the studied organism, *pre-B lymphocyte* as cell type, and *B3 cell line* as cell line, leaving the remaining fields blank because they are not relevant for this use case. The next section (Figure 4.4-C), "Experimental conditions", describes any type of treatment applied to the cells. Some interesting fields would be "Treatment" (*4-hydroxy tamoxifen [4-OHT]*), "Dose" (*0.5 μM*), "Time" (*0h*), and the "Protocol description", a written description for the process of growing and maintaining of the cells. Additionally, it is possible to attach extra files for a more detailed description of the full process. Finally, some extra fields, such as the list of "Owners" (i.e. users that can edit this information), conclude this first part of the annotation (Figure 4.4-D).

After filling in the common biological information, we proceed with the annotation of the individual samples. To illustrate the process we will use the annotation for the BRs grown in batch 4 (which were later used in mRNA-seq, miRNA-seq, RRBS-seq, and proteomics experiments). First, we choose the "Add new Sample" option which displays the form shown in Figure 4.5-A. We type a name that identifies the BR (*STATegra-Ikaros-batch4-0h*), and choose the "Add new aliquot" option. At this point the system shows a new dialog for choosing the process or protocol followed to isolate the ASs (Figure 4.5-B). Protocols are considered to be individual system information units(IUs) which can be manipulated and reused after saving them. In this use case we need to



Sample creation.

Biological conditions | Samples

Biological conditions details

General details

BioCondition ID	[Autogenerated after saving]
Name	Stategra_Ikaros-ERt2 cells 0h 4-OHT
Title	Pre-B Mouse cells (cell line B3) carrying a Tamoxifen inducible vector for Project S
Sample tags	Mouse x, treatment x

Biomaterial details

Organism	Mus musculus
Material type	Not specified
Cell type	pre-B lymphocyte
Cell line	B3 cell line
Gender	Not specified
Genotype/Variation	Not specified
Other details	

Experimental conditions

Treatment	4-OHT
Dose	0.5 uM
Time	0h
Other details	
Protocol Description	Cells containing inducible Ikaros were generated by transducing mouse pre-B cell line B3 with mouse stem cell virus (MSCV) retroviral vectors encoding a fusion protein of haemagglutinin-tagged wild type Ikaros (HA-Ikaros) and the
SOP documents	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1954785

Other details

Owners	admin
Submission date	2017/04/20
Last edition	2017/04/20
External links	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1954785 http://www.stategra.eu/the-b3-system/

Figure 4.4: Annotation details at the Sample module for the condition *Stategra-Ikaros-ERt2 cells 0h 4-OHT*.

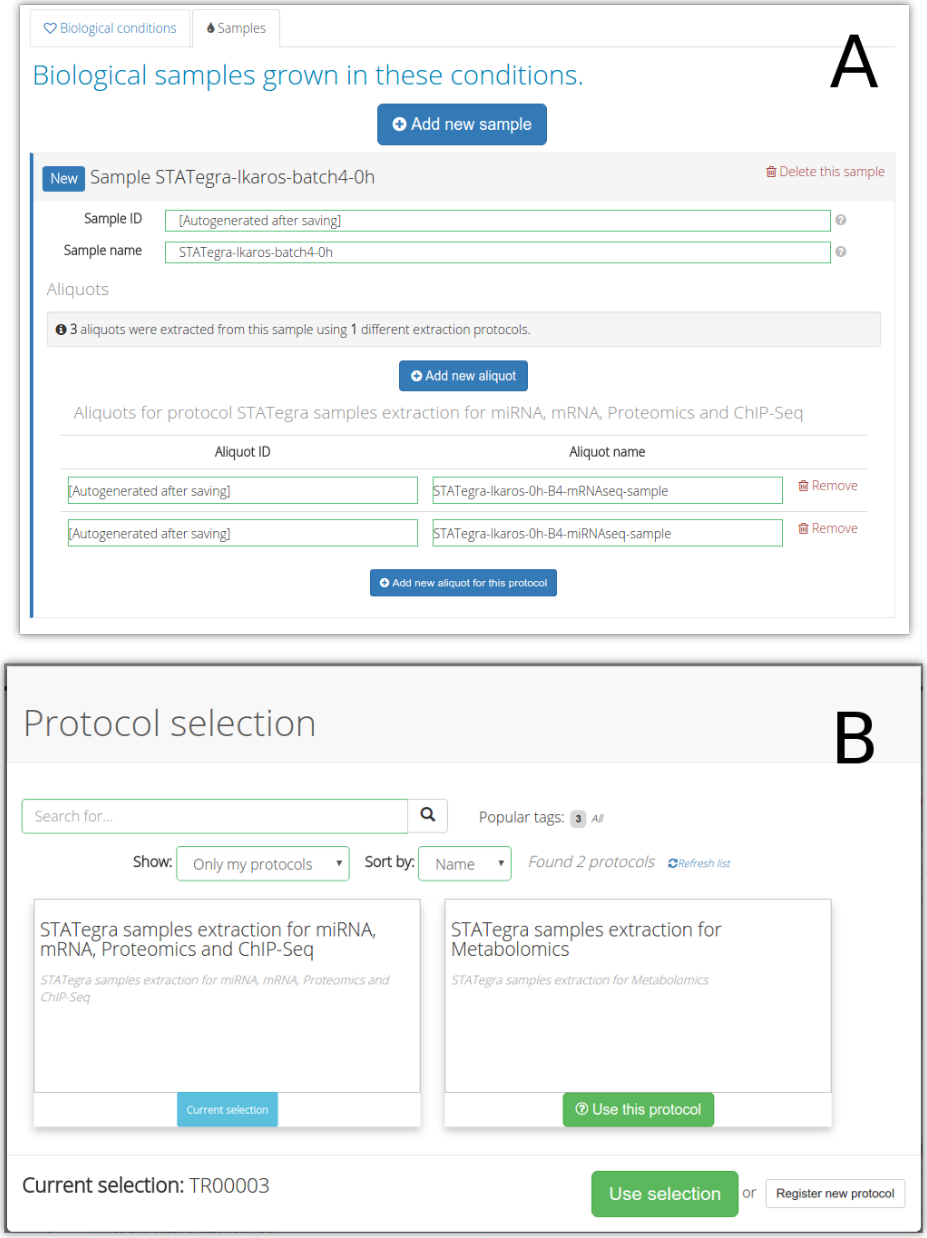
annotate a new protocol. To do this we choose the "Register new extraction protocol" option and fill in the form to describe the process of how the ASs for mRNA-seq, miRNA-seq, RRBS-seq, and proteomics assays were isolated, and define the protocol title and which users can edit the information. After setting up the protocol we add all of the ASs isolated using this protocol, giving each sample a different name (e.g. *STATegra-Ikaros-0h-B4-miRNAseq-sample*). To complete the BR annotation we fill in the information for the remaining analytical samples.

Finally, we save the changes and the new BC is registered in the system which assigns unique identifiers to the BC and all of its subcomponents (BRs and ASs). The remaining BCs are annotated in the same way, adapting the information for each specific case, and the resulting sample panel shows the complete list of samples for the current experiment, grouped by BC (Figure 4.6).

4.2.5 Analysis annotation

The information about the processing workflows is incorporated into the analysis module. While this use case-study involves seven omics types, we will only describe the RNA-seq and proteomics workflows in detail. However, the annotation for the miRNA-seq, ChIP-seq, DNase-seq, RRBS-seq, and metabolomics analysis are conceptually similar.

In the side menu we choose the "Browse analysis" option, which displays the list of all the annotated analyses for the active experiment, and provides users with options for editing, inspecting, and registering the workflows for the analyses performed during the study.



Biological samples grown in these conditions.

Sample STATegra-Ikaros-batch4-0h

Sample ID: [Autogenerated after saving]

Sample name: STATegra-Ikaros-batch4-0h

Aliquots

3 aliquots were extracted from this sample using 1 different extraction protocols.

Aliquots for protocol STATegra samples extraction for miRNA, mRNA, Proteomics and ChIP-Seq

Aliquot ID	Aliquot name	
[Autogenerated after saving]	STATegra-Ikaros-0h-B4-mRNAseq-sample	Remove
[Autogenerated after saving]	STATegra-Ikaros-0h-B4-miRNAseq-sample	Remove

Protocol selection

Search for...

Popular tags: 3 All

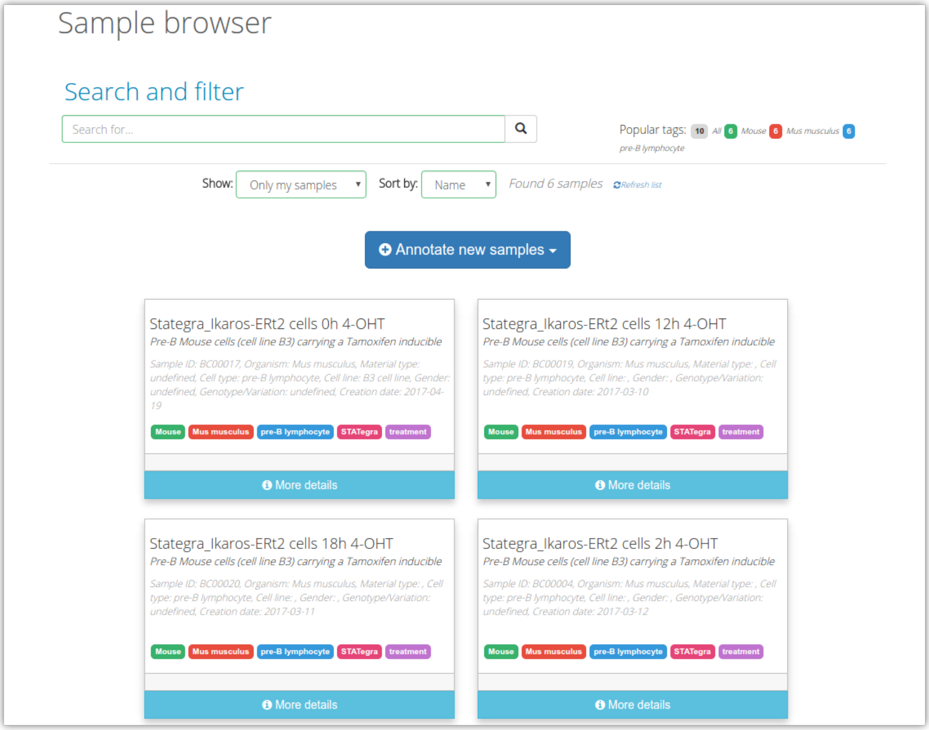
Show: Only my protocols   Sort by: Name   Found 2 protocols   Refresh list

STATegra samples extraction for miRNA, mRNA, Proteomics and ChIP-Seq
STATegra samples extraction for miRNA, mRNA, Proteomics and ChIP-Seq

STATegra samples extraction for Metabolomics
STATegra samples extraction for Metabolomics

Current selection: TR00003

Figure 4.5: Annotation of biological replicates for the "Stategra-Ikaros-ERt2 cells 0h 4-OHT" biological condition. (A)- Details for the annotation of the BR *STATegra-Ikaros-batch4-0h*. **(B)-** Dialog for protocol selection.



Sample browser

Search and filter

Search for...

Popular tags: 10 All 6 Mouse 6 Mus musculus 6 pre-B lymphocyte

Show: Only my samples Sort by: Name Found 6 samples Refresh list

Annotate new samples

Stategra_Ikaros-ERt2 cells 0h 4-OHT

Pre-B Mouse cells (cell line B3) carrying a Tamoxifen inducible

Sample ID: BC00017, Organism: Mus musculus, Material type: undefined, Cell type: pre-B lymphocyte, Cell line: B3 cell line, Gender: undefined, Genotype/Variation: undefined, Creation date: 2017-04-19

Mouse Mus musculus pre-B lymphocyte STATegra treatment

More details

Stategra_Ikaros-ERt2 cells 12h 4-OHT

Pre-B Mouse cells (cell line B3) carrying a Tamoxifen inducible

Sample ID: BC00019, Organism: Mus musculus, Material type: , Cell type: pre-B lymphocyte, Cell line: , Gender: , Genotype/Variation: undefined, Creation date: 2017-03-10

Mouse Mus musculus pre-B lymphocyte STATegra treatment

More details

Stategra_Ikaros-ERt2 cells 18h 4-OHT

Pre-B Mouse cells (cell line B3) carrying a Tamoxifen inducible

Sample ID: BC00020, Organism: Mus musculus, Material type: , Cell type: pre-B lymphocyte, Cell line: , Gender: , Genotype/Variation: undefined, Creation date: 2017-03-11

Mouse Mus musculus pre-B lymphocyte STATegra treatment

More details

Stategra_Ikaros-ERt2 cells 2h 4-OHT

Pre-B Mouse cells (cell line B3) carrying a Tamoxifen inducible

Sample ID: BC00004, Organism: Mus musculus, Material type: , Cell type: pre-B lymphocyte, Cell line: , Gender: , Genotype/Variation: undefined, Creation date: 2017-03-12

Mouse Mus musculus pre-B lymphocyte STATegra treatment

More details

Figure 4.6: List of all the samples for the STATegra Project, grouped by biological condition (BC), after the annotation.

RNA-seq analysis

Workflow overview

In the following sections we describe the general steps required to produce the processed data for RNA-seq in the STATegra project [134]. These steps were applied to each BC and the resulting files were later used in the statistical analysis.

1. For each AS (three for each BC) stranded paired-end mRNA libraries were prepared and sequenced using an Illumina HiSeq 2500 sequencer.

2. The raw data files were evaluated (quality control) and pre-processed in order to remove Illumina primers and low-quality nucleotides.

3. The pre-processed files were mapped to the reference genome using TopHat 2 [72].

4. HTSeq count [5] was used for quantification of the expression of the genes.

5. Counts were normalised using the conditional quantile normalisation (CQN) method [47] and ComBat [62].

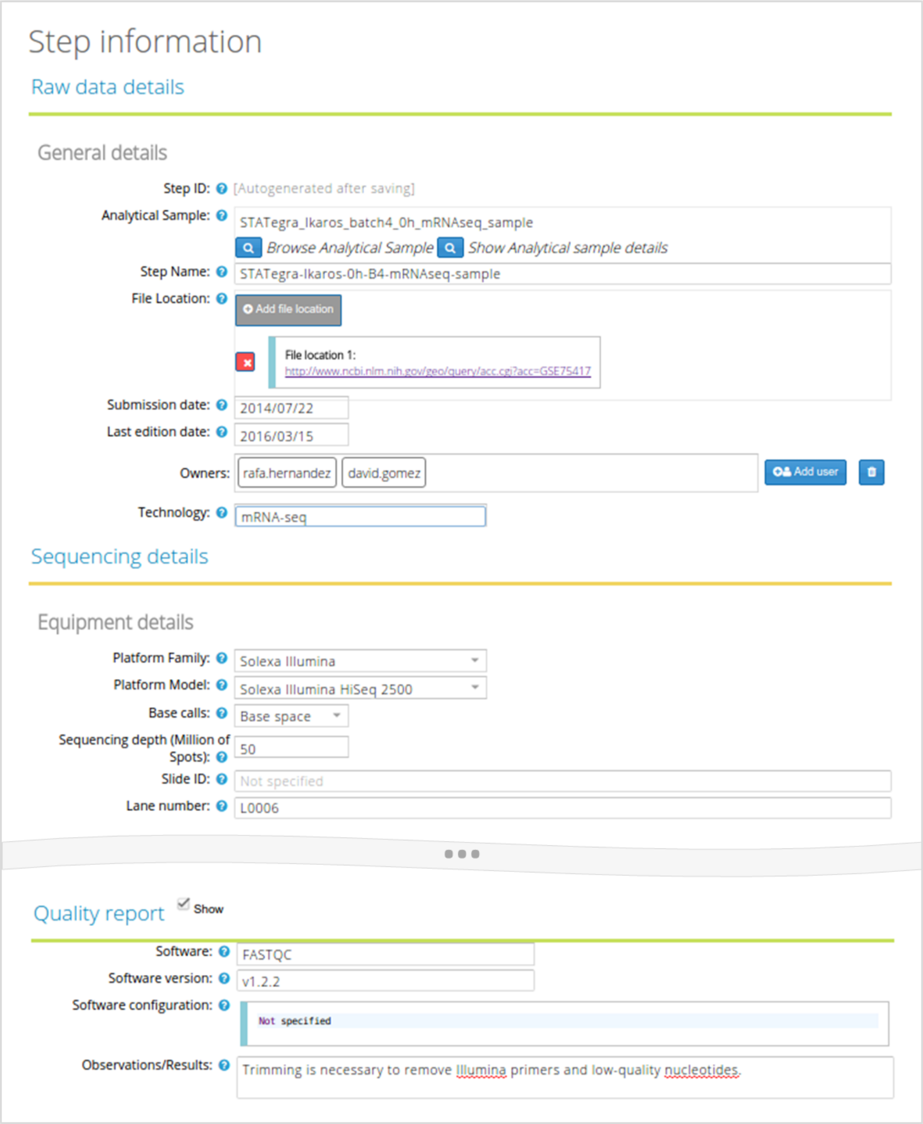
Workflow annotation

Although each AS was processed separately, the resulting data for each BC were combined during the normalisation step, and so we annotate a total of 12 mRNA-seq analyses, grouping the processing workflows by BC. Within the active experiment, we choose the "Annotate a new Analysis" option and select *mRNA-seq* as the analysis type. This option displays the input form for the analysis module. This form is divided into four sections: a toolbar for manipulating the current annotation, a table listing the steps in the current analysis (grouped by analysis type), an interactive diagram which summarises the analysis workflow, and a main section reserved for showing the details for each selected step.

New "Raw data" steps usually need to be added at the beginning of an analysis. Selecting this option opens a new dialog box with the form for annotating the raw data generation-steps (Figure 4.7). At this point we choose an existing AS to start annotating the preparation details and sequencing characteristics of a particular sample in the library. In our example, we choose the AS identified as the *STATegra-Ikaros-0h-B4-mRNAseq-sample* which corresponds to the *Ikaros-ERt2 cells 0h* BC. Additionally we can also indicate the location of the raw data files, whether files are stored in an external database (e.g. a link to a web repository), or in a local directory (using the file browsers). Then, we provide a "Step name", which helps to identify the step (*STATegra mRNAseq IKAROS 0H B4 - Sequencing*), and the "Technology", which describes the method used for data acquisition (*RNA-seq*). Depending on the technology selected, the form may be extended with new fields. Some example fields for this example case are "Platform family" (*Solexa Illumina*) and "Platform model" (*HiSeq 2500*) for describing the sequencing equipment, or "Avg. sequence length" (*75 bp*), "Layout" (*paired-end*), and "Strand specificity" (*stranded*), for describing the library details. Finally, the last section in the form describes any quality evaluation performed on the raw data. Some example fields are the "Software" used (*FASTQC*), "Version" (*v1.2.2*), "Files location" (in case that the generated reports were stored), and "Observations and results".

Once the raw data form is completed and saved, a graphical representation of the "Raw data" step is created on the workflow diagram, which grows as the subsequent analysis steps are completed. The option "Copy step", which duplicates the existing raw data instance, can be used to more quickly annotate the remaining two "Raw data" steps by altering fields such as the used AS or the step name to fit the new instances.

After the "Raw data" steps are annotated, intermediate steps can be added: in this use case, the first intermediate step corresponds to the "Data preprocessing" step. First, we choose the option for adding new "Intermediate steps"; in the new window (Figure 4.8-A) the type of step (*Data preprocessing*) is selected, and then the form is filled in with the description of the step. This form is divided into four sections: the first section contains general fields for the step, such as the "Step name" (*STATegra mRNAseq IKAROS 0H B4 - Trimming*), the "Files location" (i.e. the location for the



Step information

Raw data details

General details

Step ID: [Autogenerated after saving]

Analytical Sample: STATegra_Ikaros_batch4_0h_mRNAseq_sample

Step Name: STATegra-Ikaros-0h-B4-mRNAseq-sample

File Location:

File location 1: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75417

Submission date: 2014/07/22

Last edition date: 2016/03/15

Owners: rafa.hernandez david.gomez

Technology: mRNA-seq

Sequencing details

Equipment details

Platform Family: Solexa Illumina

Platform Model: Solexa Illumina HiSeq 2500

Base calls: Base space

Sequencing depth (Million of Spots): 50

Slide ID: Not specified

Lane number: L0006

Quality report ☑ Show

Software: FASTQC

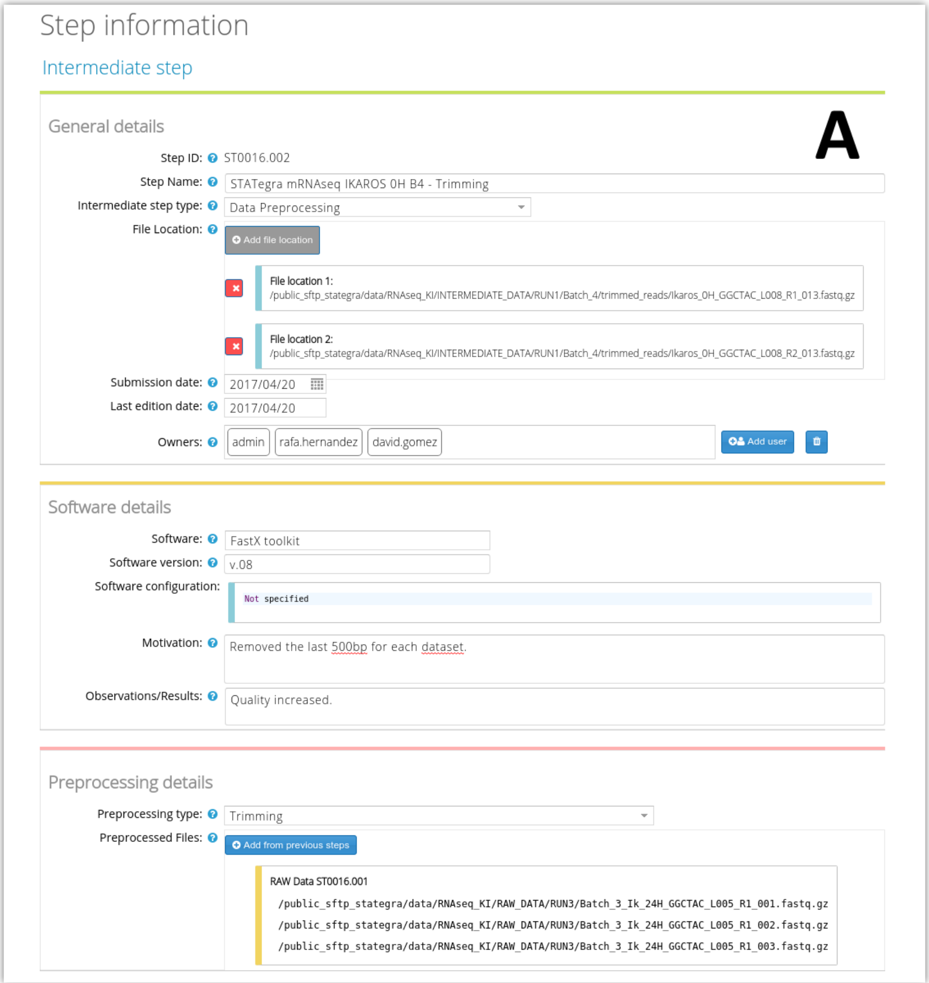
Software version: v1.2.2

Software configuration:

```
Not specified
```

Observations/Results: Trimming is necessary to remove Illumina primers and low-quality nucleotides.

Figure 4.7: Annotation form for the raw data acquisition steps for a RNA sequencing step in the STATegra project.



Step information

Intermediate step

General details

Step ID: ST0016.002

Step Name: STATegra mRNAseq IKAROS 0H B4 - Trimming

Intermediate step type: Data Preprocessing

File Location:

File location 1:
/public_sftp_stategra/data/RNAseq_KI/INTERMEDIATE_DATA/RUN1/Batch_4/trimmed_reads/Ikaros_0H_GGCTAC_L008_R1_013.fastq.gz

File location 2:
/public_sftp_stategra/data/RNAseq_KI/INTERMEDIATE_DATA/RUN1/Batch_4/trimmed_reads/Ikaros_0H_GGCTAC_L008_R2_013.fastq.gz

Submission date: 2017/04/20

Last edition date: 2017/04/20

Owners: admin rafa.hernandez david.gomez

Software details

Software: FastX toolkit

Software version: v.08

Software configuration:

```
Not specified
```

Motivation: Removed the last 500bp for each dataset.

Observations/Results: Quality increased.

Preprocessing details

Preprocessing type: Trimming

Preprocessed Files:

RAW Data ST0016.001

```
/public_sftp_stategra/data/RNAseq_KI/RAW_DATA/RUN3/Batch_3_Ik_24H_GGCTAC_L005_R1_001.fastq.gz
/public_sftp_stategra/data/RNAseq_KI/RAW_DATA/RUN3/Batch_3_Ik_24H_GGCTAC_L005_R1_002.fastq.gz
/public_sftp_stategra/data/RNAseq_KI/RAW_DATA/RUN3/Batch_3_Ik_24H_GGCTAC_L005_R1_003.fastq.gz
```

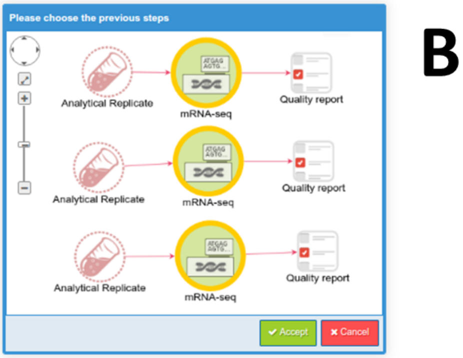


Figure 4.8: Annotation for an intermediate step in the STATegra project (A). The input for a new step usually corresponds to the output for the previous steps. Users can define that relationship by choosing the previous steps in the interactive workflow diagram **(B)**.

files it produces, if any), and the "Owners" of the information. The second section describes the "Software" used (*FastX toolkit*), as well as the "Objectives" and "Results" for the step. A third section displays specific details for the selected step type; for example, the fields available for *Data pre-processing* are "Pre-processing type" (*Trimming*) and "Pre-processed files", which specifies the input files for the step (i.e. the raw files generated in the previous steps). As Figure 4.8-B shows, the input files can be selected from the existing workflow and the system automatically determines the location for the files. Again, the last section in the form is used to describe any quality control evaluations later applied to the resulting files.

Next, the pre-processed sequences are mapped to the mouse reference genome. The annotation process for this step is equivalent to the explained before: first, we choose *Data Mapping* as the "Step type" and then fill in the specific fields for the mapping steps, such as "Genome species" (*Mus musculus*), "Genome version" (*mm10/GRCm38*) and "Source" (*University of California Santa Cruz [UCSC]*).

As general rule, an analysis ends with the generation of some "Processed data". For this RNA-seq analysis the generation of processed data corresponds to a gene-expression quantification and normalisation. Annotating "Processed data" steps is similar to intermediate step annotation. First, we fill in the general details, such as the "Step name" (*STATegra mRNAseq IKAROS 0H B4 - Quantification*), or the type of step (*Features quantification*). Then, we describe the software used, and lastly, we fill in the specific details for the quantification steps, i.e. the "Quantified files" (the output for the three previous mapping steps) and the "Reference file". The reference file usually contains the location for the genes in the genome, and can be an external resource (e.g. the reference genome downloaded from the University of California Santa Cruz (UCSC) repositories), or the result of another analysis already annotated in the system.

After saving the new step, the diagram shows the complete workflow for the analysis (Figure 4.9). Once all the steps have been properly annotated, the changes are saved and the new analysis is listed as part of the current experiment. The annotation for the remaining mRNA-seq analysis is similar and

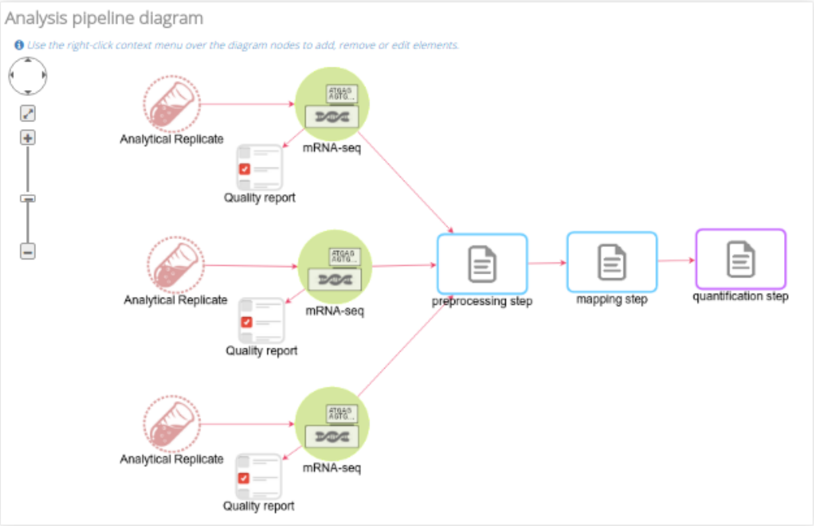


Figure 4.9: Diagram for the RNA-seq analysis after annotation of all the steps involved. Users can interact with this diagram and retrieve the complete description for each step by clicking on the corresponding node in the tree.

therefore can be performed faster using the tools provided for copying the analysis. For example, to annotate the data analysis for the *STATegra mRNAseq IKAROS 2H* condition, we can duplicate the previous analysis and adapt its content to correct the name for the steps, update the files location, and fill in any new fields or steps where the workflow varies.

Proteomics analysis

Workflow overview

In the following sections we describe the general workflow for producing processed proteomics data. These steps were applied for each BC and the resulting files were used later in the statistical analysis.

1. For each analytical sample (three for each BC) protein extracts were prepared and measured, and then injected onto the liquid chromatography (LC)-mass spectrometry (MS) system.

2. Raw LC-MS were processed using the quantitative proteomics software package MaxQuant [142] and mapped against a SwissProt mouse protein sequence-database containing canonical mouse protein sequences and common contaminants.

3. Peptide counts were normalised with using the CQN method [47].

Workflow annotation

Despite being quite similar to the annotation of sequencing omics, such as RNA-seq or ChIP-seq, the annotation for proteomics data slightly differs when registering the "Raw data" steps. Proteomics, as well as metabolomics, use different technologies for measuring the samples, such as LC-MS or gas chromatography (GC)-mass spectrometry; and consequently require different fields to report the process and the results of the experiments.

As we have previously seen, the annotation process usually starts by the annotation of raw data. The "Raw data" input form includes some general fields for setting the "Step name" (*STATegra Proteomics IKAROS 0H B4 - LCMS*), the AS measured (*STATegra-Ikaros-0h-B4-Proteomics-sample*), or the location for the resulting files. However, defining the "Technology" as *LC-MS* extends the form by adding multiple sections as recommended by the minimum information about a proteomics experiment (MIAPE) guidelines [129].

After adding the general details, the next section in the form describes the "Liquid chromatography" process, which is also divided into several blocks. For example, the first block describes the protocols for sample processing and its injection in the chromatographer; the second and third blocks describe the equipment used for separation and detection (e.g. the column and chromatography system models and manufacturer); the remaining blocks describe the details for the static and mobile phases in the columns, and the pre- and post-processing of the signals, among other information.

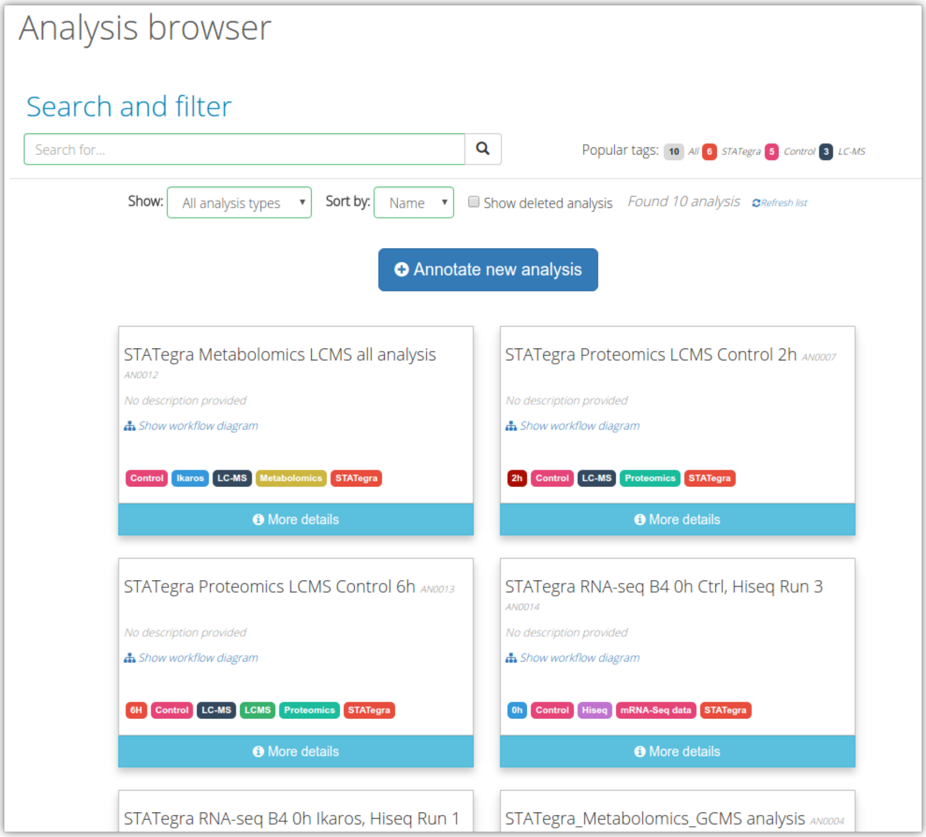
The next section describes the subsequent MS with fields for describing the manufacturer, model, and configuration of the mass spectrometer. Following MIAPE standards [129], there are also fields for describing the data acquisition process and analysis of the resulting data. Lastly, a final section allows us to describe any quality control evaluations performed on the resulting data.

After saving the raw data step, the remaining two measurements can easily be annotated by copying and adapting the previous step. We then annotate the subsequent step in the workflow, involving peptide mapping, quantification, and normalisation. These processes are compiled into a special type of "Processed data" step, known as "Proteomics MS quantification", which provides details about the procedures, software, and settings used for quantifying and identifying the proteins and peptides. Adding this step to the workflow completes the MIAPE [129] annotation for the proteomics analysis. Figure 4.11 shows an extract of the LC-MS input form and Figure 4.10 shows the "Browse analysis" panel after annotating the remaining analysis for the current experiment.

4.3 Results

As result of this previous process, the complete description of the STATegra project was stored in the system, registering numerous details about the experimental design, the main objectives, and the biological material and analytical procedures used for the study. The annotation of the biological material includes the detailed descriptions for over 280 analytical samples, extracted from a total of 128 biological replicates by applying a total of 10 different extraction protocols. Finally, an extensive description for the 12 biological conditions in the experiment completed the annotation for the biological material.

In addition to the samples metadata, 120 bioinformatics pipelines were registered in the system for the seven omics data types that were measured in the experiment. As described in previous sections, the annotations for the analysis include details that are essential for the reproducibility of the experiment, such as the name and version of the used bioinformatics tools, and the selected parameters and the inputs for each step in the workflow.



Analysis browser

Search and filter

Search for...

Popular tags: 10 All 6 STATegra 5 Control 3 LC-MS

Show: All analysis types ▾ Sort by: Name ▾ □ Show deleted analysis *Found 10 analysis* *Refresh list*

Annotate new analysis

STATegra Metabolomics LCMS all analysis *AN0012*

No description provided

Show workflow diagram

Control Ikaros LC-MS Metabolomics STATegra

More details

STATegra Proteomics LCMS Control 2h *AN0007*

No description provided

Show workflow diagram

2h Control LC-MS Proteomics STATegra

More details

STATegra Proteomics LCMS Control 6h *AN0013*

No description provided

Show workflow diagram

6H Control LC-MS LCMS Proteomics STATegra

More details

STATegra RNA-seq B4 0h Ctrl, Hiseq Run 3 *AN0014*

No description provided

Show workflow diagram

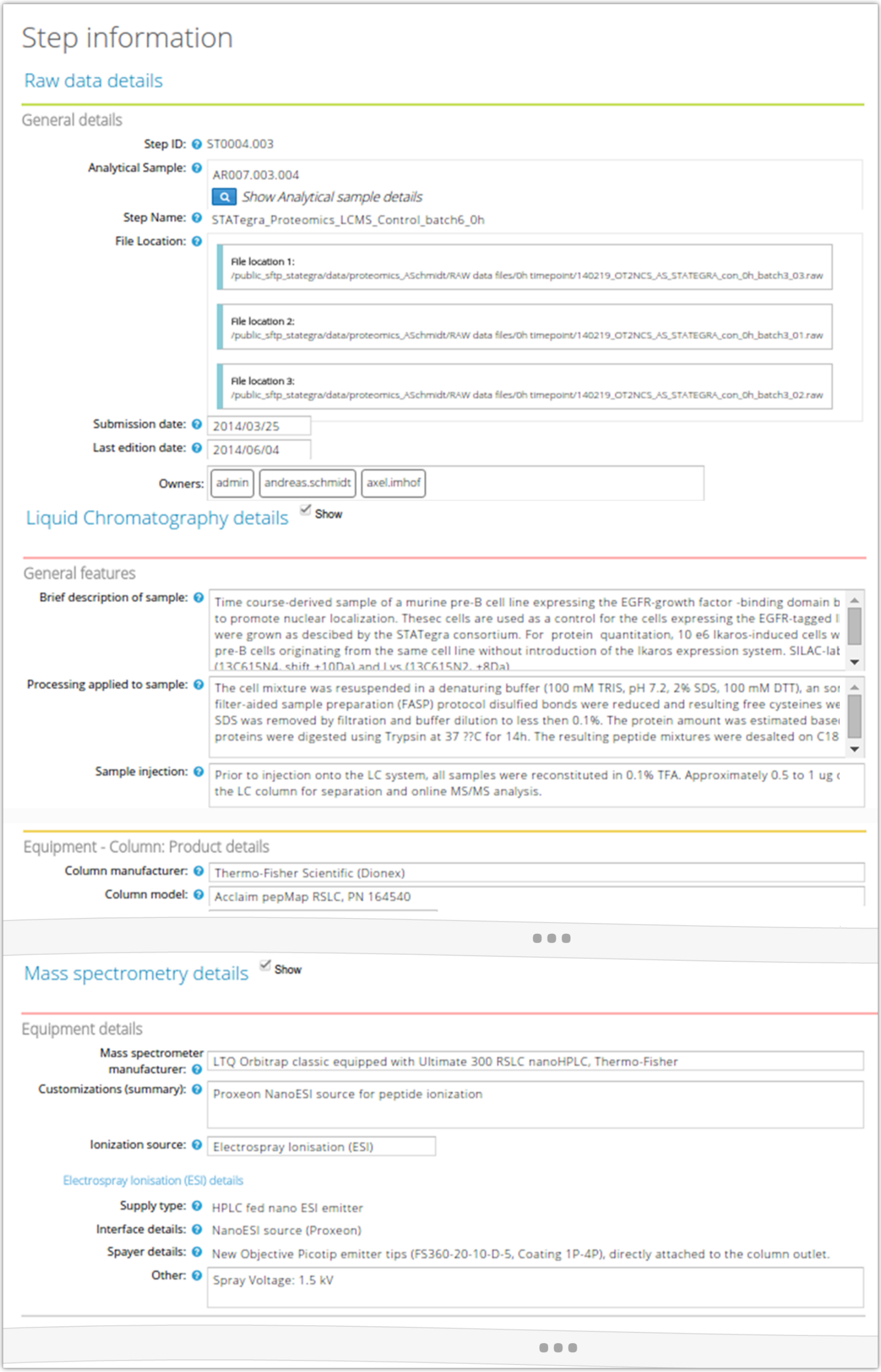
0h Control Hiseq mRNA-Seq data STATegra

More details

STATegra RNA-seq B4 0h Ikaros, Hiseq Run 1

STATegra_Metabolomics_GCMS analysis *AN0004*

Figure 4.10: List of all analyses for the STATegra project after annotation.



Step information

Raw data details

General details

Step ID: ST0004.003

Analytical Sample: AR007.003.004
Show Analytical sample details

Step Name: STATegra_Proteomics_LCMS_Control_batch6_0h

File Location:

File location 1:
/public_sftp_stategra/data/proteomics_ASchmidt/RAW data files/0h timepoint/140219_OT2NCS_AS_STATEGRA_con_0h_batch3_03.raw

File location 2:
/public_sftp_stategra/data/proteomics_ASchmidt/RAW data files/0h timepoint/140219_OT2NCS_AS_STATEGRA_con_0h_batch3_01.raw

File location 3:
/public_sftp_stategra/data/proteomics_ASchmidt/RAW data files/0h timepoint/140219_OT2NCS_AS_STATEGRA_con_0h_batch3_02.raw

Submission date: 2014/03/25

Last edition date: 2014/06/04

Owners: admin andreas.schmidt axel.imhof

Liquid Chromatography details ☑ Show

General features

Brief description of sample: Time course-derived sample of a murine pre-B cell line expressing the EGFR-growth factor -binding domain b
to promote nuclear localization. Thesec cells are used as a control for the cells expressing the EGFR-tagged I
were grown as descibed by the STATegra consortium. For protein quantitation, 10 e6 Ikaros-induced cells w
pre-B cells originating from the same cell line without introduction of the Ikaros expression system. SILAC-lat
(13C615N4, shift +10Da) and Lys (13C615N2, +8Da)

Processing applied to sample: The cell mixture was resuspended in a denaturing buffer (100 mM TRIS, pH 7.2, 2% SDS, 100 mM DTT), an so
filter-aided sample preparation (FASP) protocol disulfied bonds were reduced and resulting free cysteines we
SDS was removed by filtration and buffer dilution to less then 0.1%. The protein amount was estimated base
proteins were digested using Trypsin at 37 ??C for 14h. The resulting peptide mixtures were desalted on C18

Sample injection: Prior to injection onto the LC system, all samples were reconstituted in 0.1% TFA. Approximately 0.5 to 1 ug
the LC column for separation and online MS/MS analysis.

Equipment - Column: Product details

Column manufacturer: Thermo-Fisher Scientific (Dionex)

Column model: Acclaim pepMap RSLC, PN 164540

Mass spectrometry details ☑ Show

Equipment details

Mass spectrometer manufacturer: LTQ Orbitrap classic equipped with Ultimate 300 RSLC nanoHPLC, Thermo-Fisher

Customizations (summary): Proxeon NanoESI source for peptide ionization

Ionization source: Electrospray Ionisation (ESI)

Electrospray Ionisation (ESI) details

Supply type: HPLC fed nano ESI emitter

Interface details: NanoESI source (Proxeon)

Spayer details: New Objective Picotip emitter tips (FS360-20-10-D-5, Coating 1P-4P), directly attached to the column outlet.

Other: Spray Voltage: 1.5 kV

Figure 4.11: Annotation for the proteomics LC-MS. The form is divided into two main sections: a separate description of the LC and MS, following the MIAPE [129] guidelines.

]

4.4 Discussion

Reproducibility is one of the fundamental principles of science and, without it, a scientific discovery could not be validated or even distinguished from error or chance. The experimental provenance of data has been usually recorded in form of laboratory notebooks; however, as big data has become a standard in biomedical research, encapsulate the experimental metadata along with the data in a digital manner is now an imperative [69]. In a typical scenario in bioinformatics research, sample production is performed by a biologist without extensive bioinformatics skills. The biologist often collaborates with bioinformaticians who process the data and carry out statistical analyses that help interpreting it [146]. From the user's point of view, keeping a track for all these steps is not always easy and makes indispensable the usage of a dedicated and collaborative information management system.

In this chapter we have seen a use of case of the STATegra EMS for the annotation of a real experiment in the field of Systems Biology. As it was discussed in previous sections, the usage of this tool for storing the experimental information and sample metadata provides researchers with a valuable resource for tracking the location and origin of the produced data, as well as the current state of their experiments, samples and analysis. The availability of this information, besides the usage of standards for structuring information and the tools for exporting this information to common formats, are some of the features that situates the STATegra EMS as a powerful tool for ensuring the reproducibility the biological experiments.

5

Integrative visualization of multi-omics data: The PaintOmics 3 Platform

Parts of this chapter are contents of the manuscript: "Hernández-de-Diego R, Tarazona S, Furió-Tarí P, and Conesa A. *Paintomics: a web resource for the pathway level visualisation of multiomics data*. (In preparation)".

5.1 Introduction

The heterogeneity, high dimensionality, and interconnectivity of the multi-omics data are hurdles which must be overcome in order to extract comprehensive knowledge about the system's responses. Within this scenario, statistical models have proven to be a powerful approach for integrative analysis. However, a deep understanding of statistical concepts as well as statistical computing is required for the proper use of this approach [156]. Alternatively, considering that the capacity of human brains for visual processing is highly evolved, graphical visualisation in combination with data analysis techniques are a valuable way of simplifying data interpretation and assisting its comprehension [104].

Several resources for integrative visualisation are already available in the context of Systems Biology. As previously discussed, visualisation tools can be classified following different criteria such as being web-based or stand-alone applications, whether they provide interactive or static images or, more importantly, whether the tool is devoted to displaying genome or network information (see Section 2.4.1). Biological networks can reveal hidden connections between molecular features by representing the data in the form of graphs, where *nodes* represent biological features, such as genes, gene products, phenotypic expressions, or biological processes, and *edges* indicate an existing interaction between pairs of features. Two well-known tools for graph and network analysis and visualisation are Cytoscape [118] and Gephi [10]. Both applications are open-source and desktop-based, and provide numerous tools for exploring, manipulating and analysing complex networks (whether biological or not). Additionally, many plugins are available for both applications, enabling more specialised analysis of the networks and molecular profiles [104]. Similarly, the web-based workbench VisANT [54] includes several tools for drawing and analysing large biological networks; some examples are network structure analysis, expression enrichment analysis, integration of several external databases, and the ability to combine multiple types of networks to systematically analyse the correlations between disease, therapy, genes, and drugs. Another interesting tool is 3Omics [75] a web application specifically designed for the analysis of human data, which supports datasets for transcriptomics, metabolomics, and proteomics. Using 3Omics, users can perform correlation analysis, coexpression profiling, phenotype mapping, pathway enrichment analysis and GO enrichment analysis on each dataset, and visualise results graphically.

Alternatively, other tools determine the interaction between biological features based on existing knowledge, usually curated, of specific biological processes such as metabolic or signalling pathways. Pathways are a fundamental part of interpreting omics data, as they provide the biological context for a given observation [147]. One popular tool for pathway-based visualisation is MapMan [135]. This tool is available both as a desktop and a web application and allows large datasets, including multiple conditions or time-series experiments, to be displayed as pathway diagrams. Another example is KaPPa-View [138] a web-based tool for integrating transcript and metabolite data into pathway maps. The latest version of the tool (version 4) allows multiple-condition datasets to be incorporated, and includes several resources for data analysis. Garcia-Alcalde et al. developed PaintOmics 2, a web-based tool for integrated visualisation using the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways as a template [40]. Interesting features of PaintOmics 2 are the large range of available organisms and the use of interactive and downloadable images complemented with both experimental and KEGG information. Finally, Luo and Brouwer introduced Pathview, an R/Bioconductor package for data integration and visualisation using KEGG pathways [82]. Pathview allows integration of a wide variety of biological data and, because it is an R package, can easily be integrated into the user's analysis workflow. Last but not least, some tools are devoted to displaying the data along the entire genome. Genome browsers are especially useful when working with region-based omics data types, such as chromatin immunoprecipitation sequencing (ChIP-seq) or DNase I digestion and high-throughput sequencing (DNase-seq) data, although it is possible to visualise other types of data, such as RNA sequencing (RNA-seq), they are usually displayed as coverage graphs. One popular genome browser with integrative capabilities is the Integrative Genomics Viewer (IGV) [136]. This browser, available as a high-performance desktop application, supports a large variety of genome-wide data — including data from exome and whole-genome sequencing, epigenomics, RNA expression profiling, and single nucleotide polymorphisms(SNPs) — and incorporates useful tools for exploring the data and for retrieving datasets from popular projects such as ENCODE or 1000 Genomes. In a similar way, the Integrated Genome Browser (IGB) [97] also integrates a wide variety of omics data types and includes some interesting features such as the possibility of running basic local alignment search tool (BLAST) searches for sequences, retrieving information from the web for genes and other biological entities, or exporting and saving high quality images.

Despite being useful, these tools do have some limitations in terms of effective data integration and visualisation. As a general rule, network-based tools are useful for identifying the interconnections between multiple biological features, but the size and complexity of the network, and the lack of similarity with existing knowledge (e.g. with pathway diagrams) often hamper the interpretation. Pathway-based solutions reduce the size of the displayed data by grouping the information based on biological insights, but they do not allow new knowledge to easily be inferred. Moreover, neither type of tool allows researchers to easily integrate data from chromatin profiling experiments, which are usually explored using genome viewers. In contrast, genome browsers, which are able to display this data, make it difficult to visualise changes in multi-condition experiments and to infer possible relationships between different omics data types.

In this chapter, we introduce PaintOmics 3, a web-based application for integrative visualisation of multiple biological datasets on KEGG pathway diagrams. As opposed to other visualisation tools, the system covers a complete multi-omics pathway analysis workflow, including automatic feature name/identifier conversion, pathway enrichment analysis, network analysis and integrative visualisation; and supports data for a wide range of omics data types and organisms, as long as they are included in KEGG databases. Data visualisation in PaintOmics 3 is implemented using the latest technologies in web-based visualisation, providing powerful exploration tools and strong explanatory capabilities. Finally, this system includes other features which are typical in modern web-applications such as cloud storage for user data and cloud computing which results in an enhanced user-experience.

5.2 Methods

5.2.1 The PaintOmics 3 architecture

In order to maintain consistency and reduce the work necessary for future maintenance, the PaintOmics 3 architecture (Figure 5.1) uses the same approach as the STATegra EMS application. Hence, the platform is divided following the Client-Server paradigm, keeping the server-side in charge of processing the client data as well as managing access to the stored information; and client-side responsible for correctly presenting the data to the users, as well as providing the necessary tools for their manipulation (see section 3.2.1). Communication between the client applications and the server-side is handled by asynchronous JavaScript and XML (AJAX) mechanisms where data are exchanged encoded using JavaScript object notation (JSON).

PaintOmics 3 was entirely developed using open-source resources: Python 2.7 [107], R [108] and MongoDB [92] for building the server-side application, and JavaScript for building the client-side application. Consequently, the resulting application is also available as a free-to-use and open-source web-based application.

PaintOmics 3 server application

As mentioned above, the PaintOmics server application was mainly developed using Python, a general-purpose, high-level programming language which supports multiple programming paradigms, including object-oriented, imperative, and functional programming, among others. Python is free and open-source software and is available for installation on many operating systems. An interesting aspect of Python is the large number of modules and packages available for it and which easily extend the functionality of the language, providing new methods and functions for multiple purposes: scientific computing (SciPy [63]), image manipulation (Python Imaging Library), game development (PyGame), database manipulation (SQLAlchemy [123]), etc. Additionally, Python can serve as a scripting language for web applications through Web application frameworks such as Django [130], Flask [7], or Bottle [86]. PaintOmics 3 makes use of several python modules for processing and manipulating the user data. Some key modules for the development were:

1. **Flask** [7], a simple, extendible, and light framework written in Python for developing python-based web applications. This module determines the application structure and provides PaintOmics 3 with a routing system which maps URLs to the specific code blocks that handle them, responding with dynamic content to client requests, usually HTTP requests (Figure 5.2).

2. **SciPy** [63], a Python-based ecosystem of open-source software for mathematics, science, and engineering. This module extends Python by adding support for large, multi-dimensional lists and matrices, along with high-level mathematical functions to operate on these data structures. In PaintOmics 3 this module is essential for all statistical and mathematical estimations.

3. **PyMongo** [93], is a Python module containing tools for working with MongoDB.

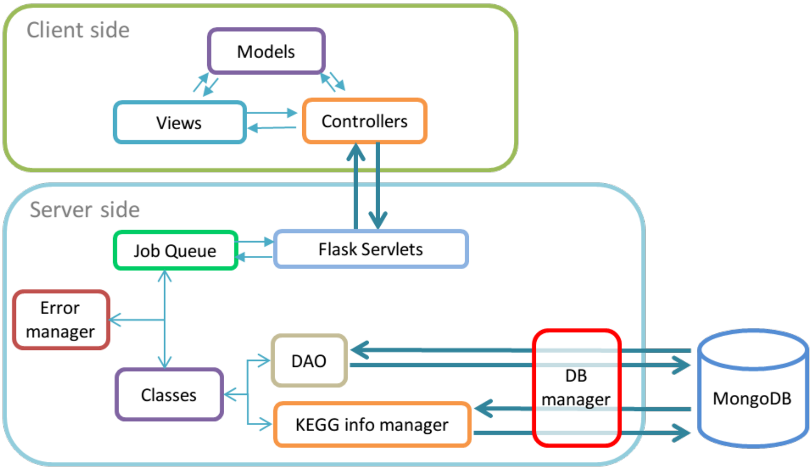
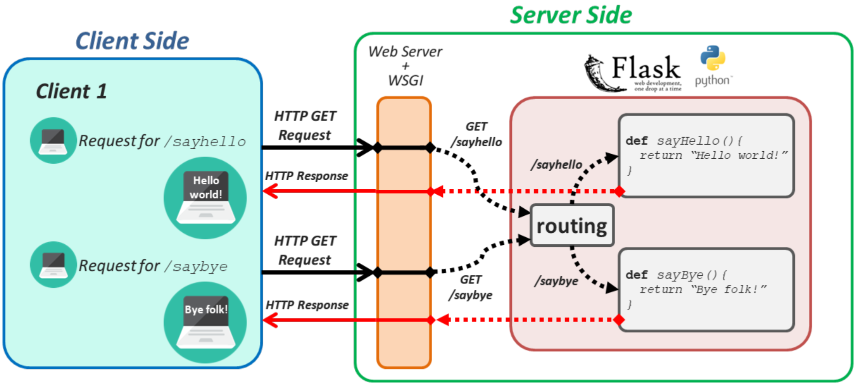


Figure 5.1: Overview of the PaintOmics 3 architecture. The platform is divided following the Client-Server paradigm. Client side implemented the Model-View-Controller pattern (see section 3.2.1 for more info). Server side is implemented in Python and is divided in several subcomponents. The main entry point for the client's requests are the servlets implemented using the Flask module. Requests are processed and delivered to the corresponding servlet (e.g. requests related with user management are managed by the Users servlet). Most of the requests requires heavy computational processing. Hence requests are encapsulated in *job* objects and enqueued for being executed as soon as enough resources are available. As a general rule, all information in the application is encapsulated in Python classes (e.g. information for user is kept in object of the class *User*). Interaction with the database is made through Data Access Objects and connections are controlled by a database manager.

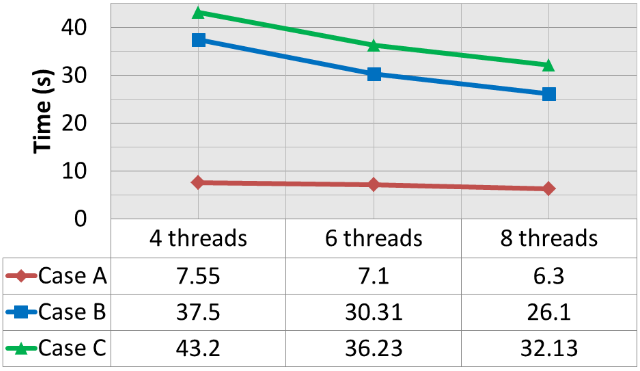


```
def sayHello(){
  return "Hello world!"
}
```

```
def sayBye(){
  return "Bye folk!"
}
```

Figure 5.2: An example of a request-response exchange between clients and servers using Flask routes. Different requests may follow different routes when arriving to the server. Responses include different information such as text, images, files, or HTML code.

4. **Multiprocessing**, a package which enables concurrent computing. This module allows PaintOmics to use parallelisation, fully leveraging the multiple processors on the host machine and dramatically reducing the times for input data processing. Figure 5.3 displays a comparative study of the performance for the ID/name translation tool for a selection of cases where the number of available parallel threads is changed.



	4 threads	6 threads	8 threads
Case A	7.55	7.1	6.3
Case B	37.5	30.31	26.1
Case C	43.2	36.23	32.13

Figure 5.3: A comparative analysis of the performance for the ID/name translation tool by changing the available number of threads for the translation. Case A (red line) uses the data from a Proteomics comparative study in mouse. Input data consists on a quantification file with 1050 proteins, and a relevant protein file containing 125 differentially expressed (DE) proteins. Case B (blue line) combines transcriptomics and metabolomics measurements for mouse. Input data for transcriptomics includes the quantification for 12763 genes and 5618 DE genes. Finally, Case C (green line) includes a more developed example that combines measurements for gene expression (6337 genes, 5524 DE genes), proteomics (1110 proteins, 148 DE proteins), metabolomics (59 compounds, 41 relevant compounds), DNase-seq (5101 regions, 3596 relevant regions) and miRNA-seq data (998 genes with miRNA values, 605 relevant genes). In all cases the files, the measured time includes the time for request processing, name translation, job storage and response processing at client side. The time for uploading the files to the server was not considered due to files were previously uploaded. All tests were made in a local instance of PaintOmics 3 in a workstation with an Intel(R) Core(TM) i7 CPU (2.67GHz , 4 cores and hyperthreading enabled) and 12GB of RAM.

In addition to Python, PaintOmics makes use of R, a powerful programming language for statistical computing. Data storage for PaintOmics 3 is managed using MongoDB databases, a scalable, high-performance, open source NoSQL database. MongoDB is document-oriented: instead of breaking up the information into multiple relational structures (as typically happens with other systems such as MySQL, where data are stored in rows and grouped in tables), information is stored in MongoDB in the form of BSON documents, a binary representation for JSON objects, which are grouped in collections. This structure makes MongoDB an effective storage system when working with Python and JavaScript because data exchange requires little or no adaptation, significantly reducing the workload when managing large amounts of data. In terms of connection with the database, PaintOmics makes use of PyMongo, a python module for working with MongoDB, as well as an implementation for the DAO pattern (see section 3.2.1) that simplifies the interaction between the server functions and the MongoDB database.

Finally, an important consideration for the development of PaintOmics was the need for a queuing system for executing user jobs. The typical workflow for PaintOmics implies processing input data that can contain thousands of biological features for multiple omics data types. PaintOmics 3 is meant to be a multi-user application, and despite the optimisation of other concurrent computing code that it uses, processing such large amounts of data can be time-consuming, thus a system for managing the server resources is essential for maintaining system stability and for improving the user experience. Therefore, PaintOmics includes *PySiQ*, a simple but effective task management system, developed as a reusable, configurable, and open-source Python module (see Appendix B for more details).

PaintOmics 3 client application

The PaintOmics 3 client application was entirely developed using JavaScript and HTML5 technologies and is divided into two separate layers: a *front-end* application oriented to users and data visualisation and a *back-end* application for administrative purposes. Both applications were developed as independent projects using different technologies. The *back-end* application, which is only accessible by the administrator users, was developed using AngularJS [45] and Bootstrap [16], both popular HTML, CSS, and JavaScript frameworks for developing responsive web applications. The *front-end* application was built using the Sencha ExtJS framework [117] which provides sotisficated tools for controlling the layout and the web components that shape the application. The Sencha ExtJS framework was complemented by using the open-source

JavaScript library JQuery [132], which makes it easier to navigate through the HTML documents, manipulate the web components, or to add a wide variety of effects and animations. As visualisation was the final objective of the application, a large number of JavaScript resources for data representation were studied. As a result, the following libraries were selected as the best options in terms of representative effectiveness (visual resources available, configurability, extensibility, etc.), robustness, user friendliness, and maintenance.

1. **HighCharts** [52], a charting library written in pure JavaScript. HighCharts is cross-platform and fully compatible with most of the modern browsers. This library is also free (for non-commercial use) and is open-source. HighCharts supports a many interactive graph types, including line, area, and pie charts, as well as more atypical options such as heatmaps or gauges. Another interesting characteristic for this library is its extendibility, which allowed us to develop new features for the heatmap diagrams such as the usage of clustering strategies and dendrograms.

2. **Linkurious.js** [80], a cross-browser JavaScript library for interactive network visualisation. This library is based on Sigma.js [119] a powerful JavaScript library dedicated to graph drawing, which extends with new HTML5 features. Linkurious provides a lot of plugins and extensions and is distributed as open-source, which makes this library easy to extend and customise.

3. **SVGjs** [127], is a lightweight open-source library for manipulating and scalable vector graphics(SVGs). This library was especially useful for generating interactive diagrams for the KEGG pathways.

Regarding the application architecture, the client-side is built following the model-view-controller (MVC) architectural pattern, which simplifies the organisation, maintenance and extension of large client applications (see Section 3.2.1).

5.2.2 The KEGG pathways database

The KEGG is a collection of databases and resources for studying the high-level functions of biological systems [64, 65]. The KEGG database project was started in 1995 at the Institute for Chemical Research, Kyoto University, who were looking for a computerised representation for the links between genomic information and higher-level systemic functions of cells, organisms, and ecosystems (Figure 5.4).

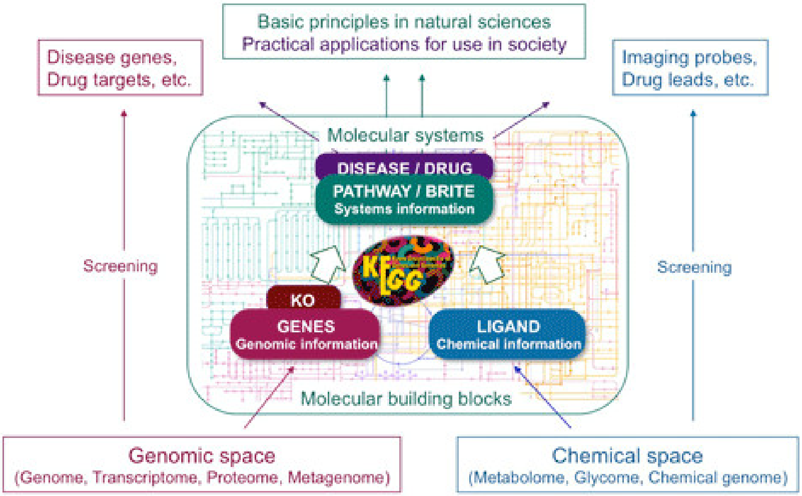


Figure 5.4: Overview for the integrated information in KEGG database.

The most recent release of the KEGG database (Release 81.0, January 1, 2017) includes 17 main databases maintained in an internal Oracle database [64, 65], which contain large amounts of genomic and molecular-level information for up to 5014 species (360 eukaryotes, 4090 bacteria, 247 archaea, and 317 viruses). These databases are broadly categorised into *Systems information*, *Genomic information*, *Chemical information*, *Health information*, and *Drug labels* (Table 5.1).

The KEGG PATHWAY database is a collection of graphical diagrams, usually known as pathway maps, which represent molecular interaction and reaction networks within a cell during specific biochemical processes, which usually leads to the output of a product or a change in the cell. KEGG PATHWAY contains about 508 reference pathways (Release 81.0, January 1, 2017), which are manually drawn and continuously updated according to biochemical evidence,

Category	Database	Content
Systems information	KEGG PATHWAY KEGG BRITE KEGG MODULE	KEGG pathway maps BRITE functional hierarchies KEGG modules of functional units
Genomic information	KEGG ORTHOLOGY KEGG GENOME KEGG GENES KEGG SSDB	KEGG Orthology (KO) groups KEGG organisms with complete genomes Gene catalogues of complete genomes Sequence similarity database for GENES
Chemical information	KEGG COMPOUND KEGG GLYCAN KEGG REACTION KEGG RPAIR KEGG RCLASS KEGG ENZYME	Metabolites and other small molecules Glycans Biochemical reactions Reactant pair chemical transformations Reaction class defined by RPAIR Enzyme nomenclature
Health information	KEGG DISEASE KEGG DRUG KEGG DGROUP KEGG ENVIRON	Human diseases Drugs Drug groups Crude drugs and health-related substances
Drug labels	KEGG MEDICUS	Drug labels from different sources.

Table 5.1: The main KEGG database categories.

and are categorised by a hierarchical classification as shown in Figure 5.5. In addition to reference maps, the KEGG PATHWAY database contains over than 490,000 organism-specific pathways inferred by automatic-mapping based on existing orthologies between species.

Each pathway in the KEGG is identified by a 5 digit number (referred to as the entry name or the accession number) proceeded by a 2-4 letter code that indicates the organism or databases to which it belongs. Some examples of valid identifiers are *map03060* (Reference protein export pathway, Figure 5.6), *mmu03060* (Mus musculus protein export pathway), or *hsa03060* (Homo sapiens protein export pathway).

Pathways are manually drawn using in-house software called KegSketch which uses different graphic resources to visualise the information. For example, boxes represent gene products, mostly proteins but also RNA, while circles represent other molecules such as chemical compounds (Figure 5.7-A). Interactions between biomolecules or other pathways are drawn using different arrows (Figure 5.7-B), and the combination of multiple shapes can be interpreted as different biochemical processes or molecular interactions (Figure 5.7-C). Colouring is also another resource for diagram interpretation: as a general rule, reference pathways are not coloured while variations of pathways for the KEGG ENZYME database are coloured blue, and organism-specific pathways are coloured green, where colouring indicates that the biological feature (i.e. the gene or metabolite) exists in the corresponding database (Figure 5.7-D).

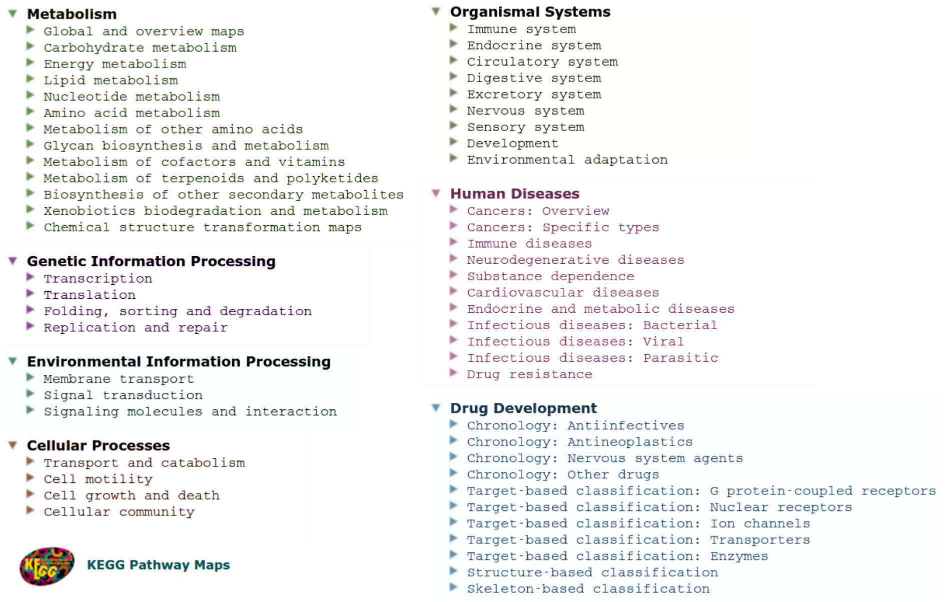
Finally, it is interesting to highlight the KEGG markup language (KGML), an exchange format for KEGG pathway maps which contains computerised information about the graphical objects it represents and their relationships in the KEGG pathway, i.e. coordinates for shapes, dimensions, colours, or links to databases, among other information (Code fragment 5.1).

KEGG is the main biological resource for PaintOmics. The application uses the information from KEGG Pathways in two different ways:

1. Pathway Diagrams: PaintOmics requires both the static images in PNG format and the KGML files, to generate the customised KEGG diagrams.

2. Mapping files: using these files the application is able to associate the different biological features with the pathways in which they are involved, and vice versa. This allows us to infer the set of pathways of interest and to perform their subsequent enrichment analysis.

All the required information is downloaded and processed from KEGG using the Administrator tools (see Section 5.2.7) and are stored locally in MongoDB.



▼ **Metabolism**
- ▶ Global and overview maps
- ▶ Carbohydrate metabolism
- ▶ Energy metabolism
- ▶ Lipid metabolism
- ▶ Nucleotide metabolism
- ▶ Amino acid metabolism
- ▶ Metabolism of other amino acids
- ▶ Glycan biosynthesis and metabolism
- ▶ Metabolism of cofactors and vitamins
- ▶ Metabolism of terpenoids and polyketides
- ▶ Biosynthesis of other secondary metabolites
- ▶ Xenobiotics biodegradation and metabolism
- ▶ Chemical structure transformation maps

▼ **Genetic Information Processing**
- ▶ Transcription
- ▶ Translation
- ▶ Folding, sorting and degradation
- ▶ Replication and repair

▼ **Environmental Information Processing**
- ▶ Membrane transport
- ▶ Signal transduction
- ▶ Signaling molecules and interaction

▼ **Cellular Processes**
- ▶ Transport and catabolism
- ▶ Cell motility
- ▶ Cell growth and death
- ▶ Cellular community

KEGG Pathway Maps

▼ **Organismal Systems**
- ▶ Immune system
- ▶ Endocrine system
- ▶ Circulatory system
- ▶ Digestive system
- ▶ Excretory system
- ▶ Nervous system
- ▶ Sensory system
- ▶ Development
- ▶ Environmental adaptation

▼ **Human Diseases**
- ▶ Cancers: Overview
- ▶ Cancers: Specific types
- ▶ Immune diseases
- ▶ Neurodegenerative diseases
- ▶ Substance dependence
- ▶ Cardiovascular diseases
- ▶ Endocrine and metabolic diseases
- ▶ Infectious diseases: Bacterial
- ▶ Infectious diseases: Viral
- ▶ Infectious diseases: Parasitic
- ▶ Drug resistance

▼ **Drug Development**
- ▶ Chronology: Antiinfectives
- ▶ Chronology: Antineoplastics
- ▶ Chronology: Nervous system agents
- ▶ Chronology: Other drugs
- ▶ Target-based classification: G protein-coupled receptors
- ▶ Target-based classification: Nuclear receptors
- ▶ Target-based classification: Ion channels
- ▶ Target-based classification: Transporters
- ▶ Target-based classification: Enzymes
- ▶ Structure-based classification
- ▶ Skeleton-based classification

Figure 5.5: The hierarchical classification of KEGG Pathways: primary and secondary categories.

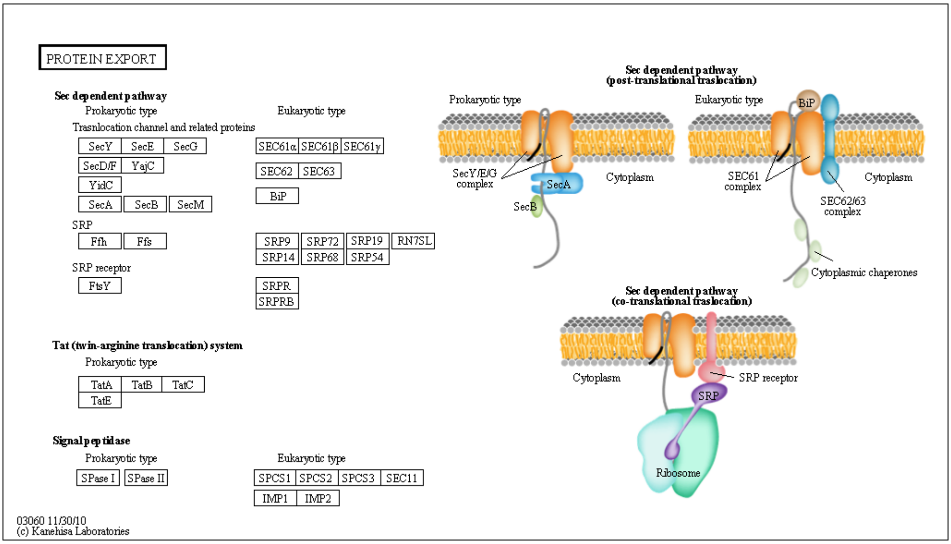


Figure 5.6: KEGG diagram for reference Protein export pathway (KEGG id map03060).

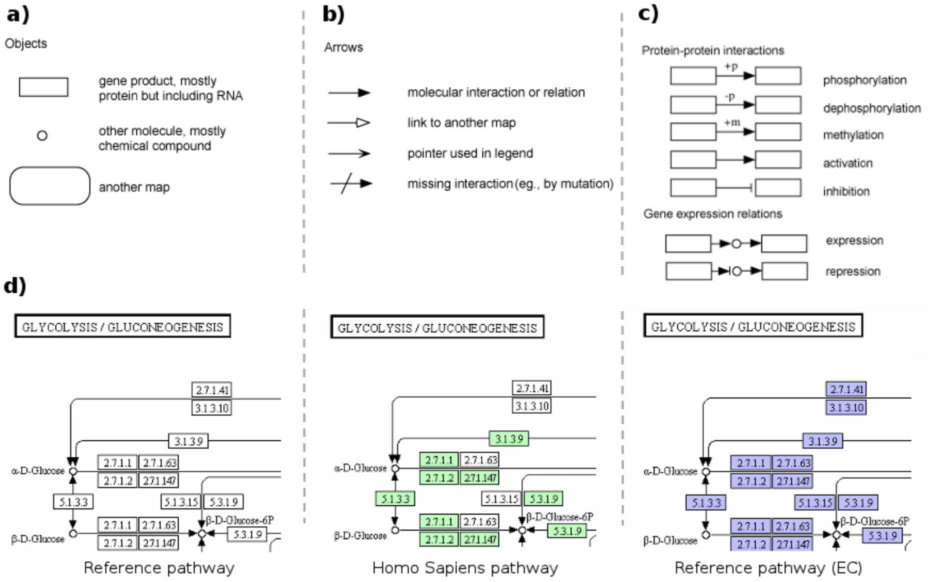


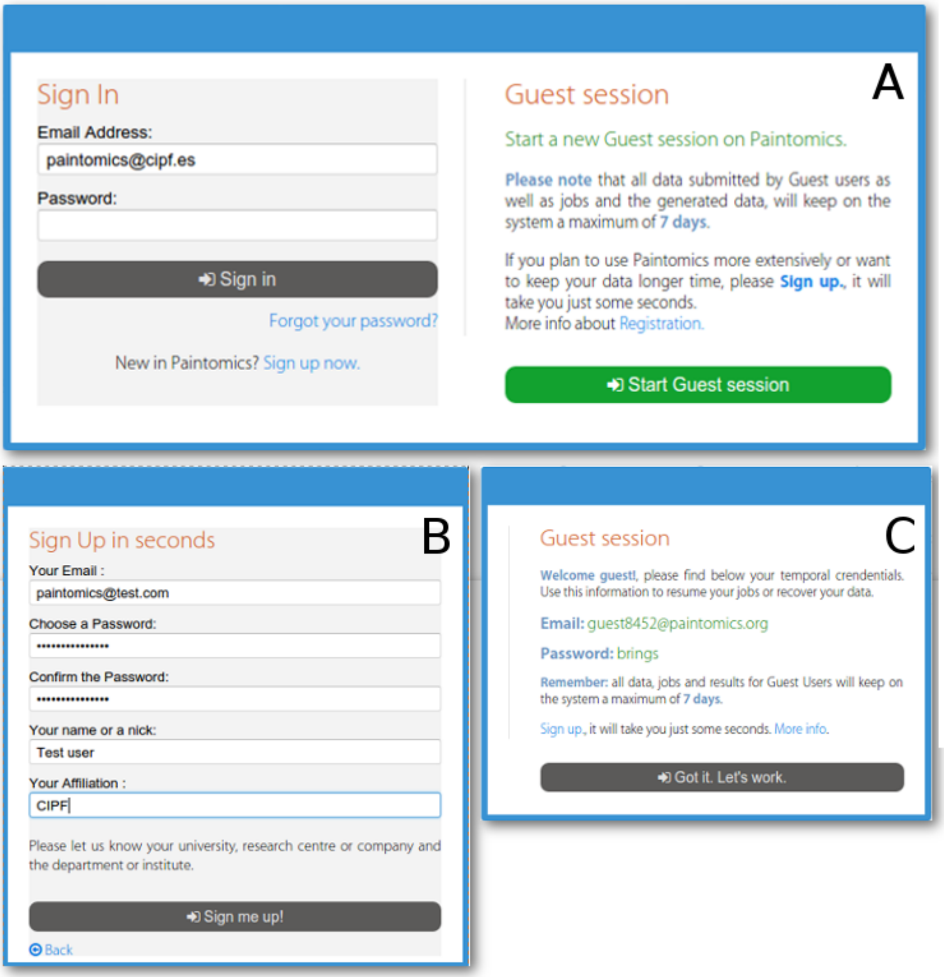
Figure 5.7: Some examples of the graphic resources used in KEGG diagrams to visualise the information. Graphical representation of genes and molecules and its interactions (a and b); Representation in shapes of biochemicals processes (c); Colouring representations of the different databases available in KEGG (d).

```
<?xml version="1.0"?>
<!DOCTYPE pathway SYSTEM "http://www.kegg.jp/kegg/xml/KGML_v0.7.1_.dtd">
<pathway name="path:hsa00010" title="Glycolysis/Gluconeogenesis" image="http://www.kegg.jp/.../hsa00010.png"
      link="...">
  <entry id="13" name="hsa:226 hsa:229 hsa:230" type="gene" reaction="rn:R01070" link="...">
    <graphics type="rectangle" x="483" y="407" width="46" height="17" name="ALDOA, ALDA,..." fgcolor="#000"
          bgcolor="..." />
  </entry>
  <entry id="37" name="hsa:217 hsa:219 hsa:223 hsa:224 hsa:501" type="gene" reaction="rn:R00710">
     [...]
  </entry>
      [...]
  <entry id="113" name="cpd:C00036" type="compound" link="http://www.kegg.jp/dbget-bin/www_bget?C00036">
    <graphics name="C00036" fgcolor="#000" bgcolor="#FFF" type="circle" x="146" y="736" width="8" height="8"/>
  </entry>
  [...]
  <relation entry1="68" entry2="70" type="ECrel">
     <subtype name="compound" value="86"/>
  </relation>
  [...]
  <reaction id="47" name="rn:R00014" type="irreversible">
    <substrate id="98" name="cpd:C00022"/>
    <substrate id="136" name="cpd:C00068"/>
    <product id="99" name="cpd:C05125"/>
  </reaction>
  [...]
</pathway>
```

Code fragment 5.1: Fragment of the KGML file for the Homo Sapiens Glycolysis/Gluconeogenesis KEGG pathway. In the KGML the *pathway* element specifies an object with *entry* elements as its nodes and the *relation* and *reaction* elements as its edges. The *relation* and *reaction* elements indicate the connection patterns of rectangles (gene products) and the connection patterns of circles (chemical compounds), respectively, in the KEGG pathways [66].

5.2.3 User system in PaintOmics 3

By default, all data uploaded by users, as well as the results from running the application, are stored on the server-side for future use, which avoids users having to re-submit files and provides them with tools for resuming previous executions of the application. Because data in PaintOmics 3 are potentially sensitive in terms of privacy and confidentiality, the application includes some security constraints in order to keep the data safe from undesired accesses. As explained in Section 3.2.2, the current PaintOmics 3 version follows two levels of security verifications: **Authentication**, which verifies that the stored data can be accessed and manipulated only by the owners; and **Authorisation**, which verifies that users are allowed to perform an action which affects or makes use of certain data. Therefore, PaintOmics 3 implements a User System based on sessions, which support typical features such as log in (open new session), log out (close session), and sign in (registration), with email confirmation for account activation (Figure 5.8).



Sign In

Email Address: paintomics@cipf.es

Password:

Forgot your password?

New in Paintomics? Sign up now.

Guest session

Start a new Guest session on Paintomics.

Please note that all data submitted by Guest users as well as jobs and the generated data, will keep on the system a maximum of 7 days.

If you plan to use Paintomics more extensively or want to keep your data longer time, please Sign up., it will take you just some seconds.
More info about Registration.

Sign Up in seconds

Your Email : paintomics@test.com

Choose a Password: •••••••••••••••

Confirm the Password: •••••••••••••••

Your name or a nick: Test user

Your Affiliation : CIPF

Please let us know your university, research centre or company and the department or institute.

Guest session

Welcome guest!, please find below your temporal crendentials. Use this information to resume your jobs or recover your data.

Email: guest8452@paintomics.org

Password: brings

Remember: all data, jobs and results for Guest Users will keep on the system a maximum of 7 days.

Sign up, it will take you just some seconds. More info.

Figure 5.8: Session management in PaintOmics 3. When accessing the application, the user can choose between three options: log in using a valid user account **(A)**, create a new regular account **(B)** or work as a Guest user **(C)**. Guest user accounts are temporary and the account, as well as the data uploaded and produced, are automatically deleted after 7 days. These accounts are meant for testing or for educational purposes (e.g. courses or workshops).

Additionally, PaintOmics 3 creates a cloud storage space for each user account where uploaded files and results are stored. Users can manipulate the content of their personal storage through the web application, upload or remove files, add textual descriptions and other annotations to the uploaded content, and resume or clean previous executions of the application (Figure 5.9).

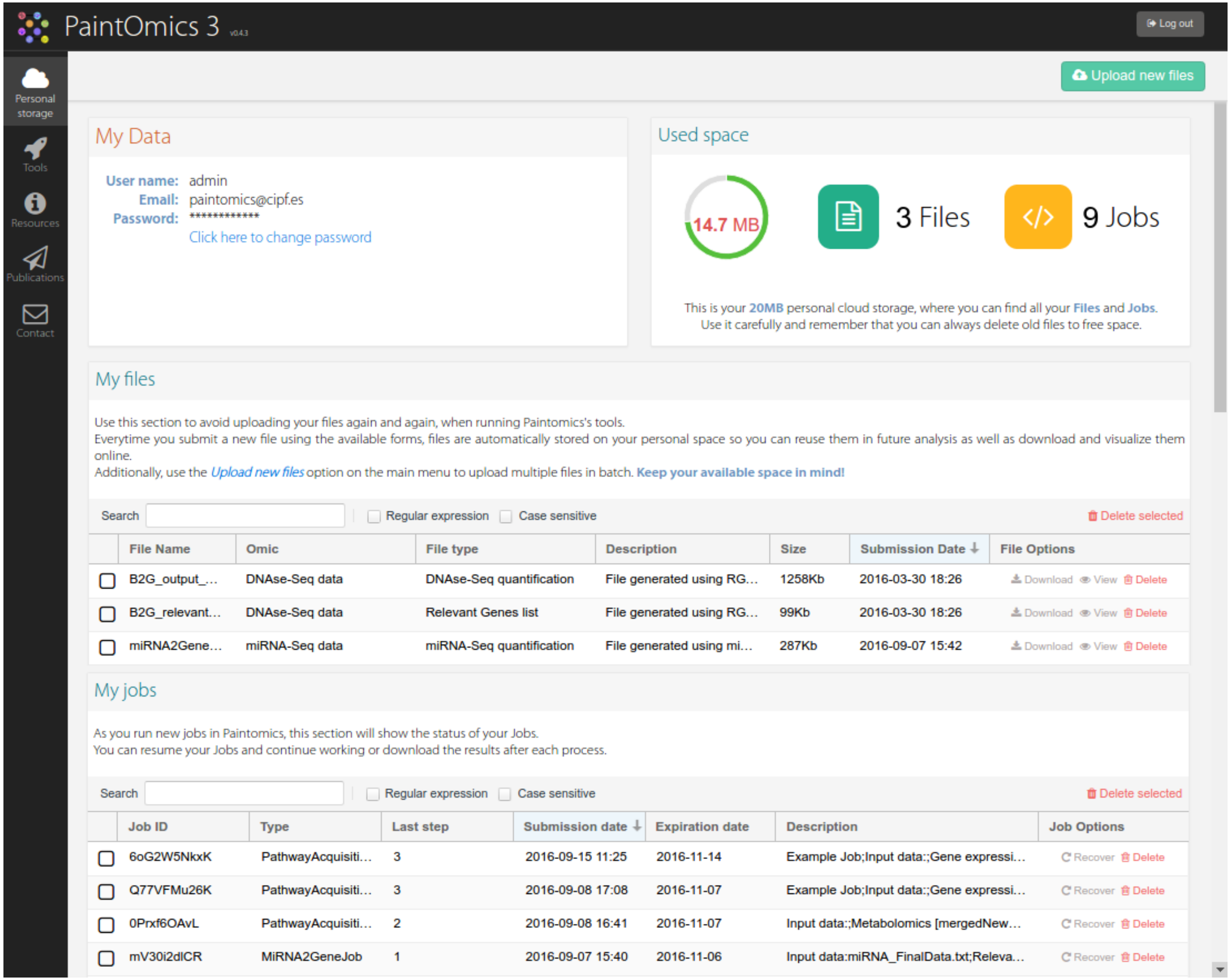
5.2.4 Accepted input data

As previously mentioned, the use of multiple complementary genome-wide measurements is becoming a powerful tool for better understanding the complexity of biological systems. Within this scenario, new tool for Systems Biology tools cannot ignore this emerging trend and should be able to support multi-omics experiments. Following this idea, PaintOmics 3 has been developed to accept diverse data types, including those from common techniques such as Transcriptomics, Metabolomics, and Proteomics, as well as emerging approaches such as DNase-seq, ChIP-seq, miRNA-seq, and methylation sequencing. As result, the data input accepted by PaintOmics 3 can be broadly classified into four categories, depending on their nature.

(i) **Gene-based omics:** this category covers omics data types where the biological features measured are, or can be, translated into genes. Some typical examples are mRNA-seq or microarrays, where measurements are made at the gene or transcript level, and proteomics, where protein quantification can be imputed to the codifying gene.

(ii) **Metabolite-based data:** here we include omics types where the studied biological features are, or can be, assigned to metabolites. This category would include metabolite quantification using, for example liquid chromatography coupled with mass spectrometry (LC-MS).

(iii) **Region-based omics:** this category includes all omics data types where the information is grouped around a set of genomic locations of interest (genomic regions). Some examples of region-based omics are ChIP-seq data which analyses protein interactions with DNA and generally results in a set of genome regions where the target protein may be bound; or DNase-seq, where regions indicate the location of regulatory sites



My Data

User name: admin
Email: paintomics@cipf.es
Password: *************
Click here to change password

Used space

This is your 20MB personal cloud storage, where you can find all your Files and Jobs.
Use it carefully and remember that you can always delete old files to free space.

My files

Use this section to avoid uploading your files again and again, when running Paintomics's tools.
Everytime you submit a new file using the available forms, files are automatically stored on your personal space so you can reuse them in future analysis as well as download and visualize them online.
Additionally, use the *Upload new files* option on the main menu to upload multiple files in batch. **Keep your available space in mind!**

Search □ Regular expression □ Case sensitive

	File Name	Omic	File type	Description	Size	Submission Date ↓	File Options
□	B2G_output_…	DNAse-Seq data	DNAse-Seq quantification	File generated using RG…	1258Kb	2016-03-30 18:26	Download View Delete
□	B2G_relevant…	DNAse-Seq data	Relevant Genes list	File generated using RG…	99Kb	2016-03-30 18:26	Download View Delete
□	miRNA2Gene…	miRNA-Seq data	miRNA-Seq quantification	File generated using mi…	287Kb	2016-09-07 15:42	Download View Delete

My jobs

As you run new jobs in Paintomics, this section will show the status of your Jobs.
You can resume your Jobs and continue working or download the results after each process.

Search □ Regular expression □ Case sensitive

	Job ID	Type	Last step	Submission date ↓	Expiration date	Description	Job Options
□	6oG2W5NkxK	PathwayAcquisiti…	3	2016-09-15 11:25	2016-11-14	Example Job;Input data:;Gene expressi…	Recover Delete
□	Q77VFMu26K	PathwayAcquisiti…	3	2016-09-08 17:08	2016-11-07	Example Job;Input data:;Gene expressi…	Recover Delete
□	0Prxf6OAvL	PathwayAcquisiti…	2	2016-09-08 16:41	2016-11-07	Input data:;Metabolomics [mergedNew…	Recover Delete
□	mV30i2dICR	MiRNA2GeneJob	1	2016-09-07 15:40	2016-11-06	Input data:miRNA_FinalData.txt;Releva…	Recover Delete

Figure 5.9: Personal storage management in PaintOmics 3.

sensitive to cleavage by DNase I. In some cases, region-based data can be also translated to gene-based domains for regions that totally or partially match a gene to the genome.

(iv) **Micro-RNA-based omics:** this category can be considered a special case of category (i) where the measured molecules are miRNAs. Data files for this category need to be pre-processed before being treated as a gene-based omics type (see below).

For the current version, acceptable data files must be saved as tab-separated plain text files. As general rule, users should provide two files for each omics data type: a quantification file containing measurements for each biological entity (e.g. gene expression quantification) and a second file with a list of features that the user considers relevant for the experiment (usually the list of differentially-expressed (DE) features). PaintOmics 3 is meant to work with log-scale quantification values where positive values indicate overexpression or the increased presence of the features, and negative values indicate repression or the reduced presence of these features, with regard to a reference or control condition.

Table 5.2-A shows an example of a quantification file for categories (i) and (ii): the first column must contain the feature name or identifier. As a general rule PaintOmics accepts Entrez Gene IDs as feature identifiers, although for some species other identifier/name domains are supported (see below identifier and name conversion). The remaining columns contain the quantification values for each sample in the experiment, preferably in a logarithmic scale format. Table 5.2-B shows an example of a relevant features file for categories (i) and (ii). This file must contain a single column with the identifiers or names for each of the significant features in the experiment (e.g. the differentially expressed genes). The accepted input formats for categories (iii) and (iv) are extensively described in the following sections.

(A)

# Feature name	Treated 0h	Treated 2h	Treated 6h
ENSMUSG000000270914	-0.82545	-0.13123	0.65332
ENSMUSG000000271116	0.31252	0.23123	0.11124
ENSMUSG000000271315	1.25544	-0.00123	-0.02265
...	...	...	...

(B)

Feature name
ENSMUSG000000270914
ENSMUSG000000271116
ENSMUSG000000271641
...

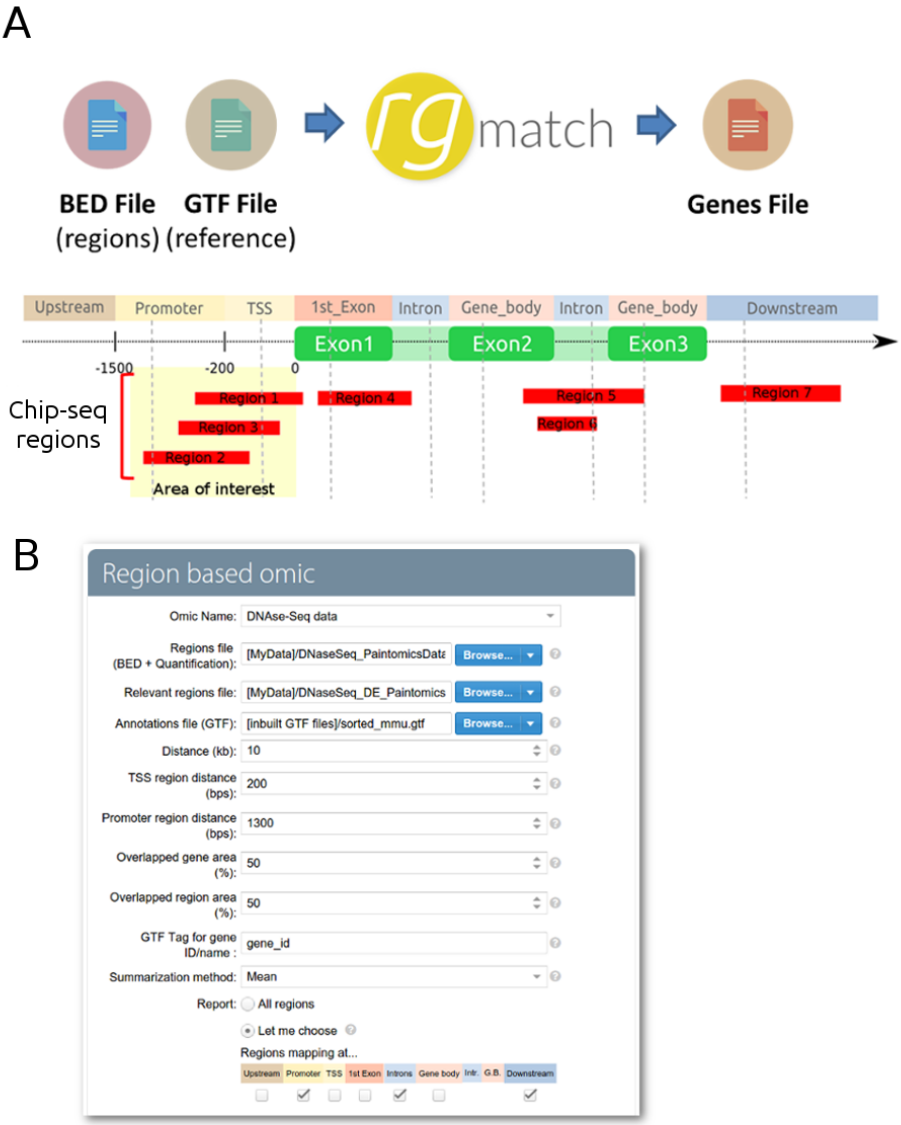
Table 5.2: Example of the input for a gene-based omics type dataset. (A) - The quantification file contains the gene name or identifier (first column) followed by the quantification values for 3 different time points, on a logarithmic scale. **(B)** - Differentially-expressed genes are provided as a list.

Converting region-based data to genes

Omics approaches to studying regulatory aspects of gene expression such as ChIP-seq, DNase-seq, assay for transposase-accessible chromatin using sequencing (ATAC-seq), or Methyl-seq, typically return potentially functional regions, defined by genomic coordinates, which must then be related to proximal genes in order to gain any biological meaning. Therefore, integration of these region-based "omics" requires an extra step where regions are associated with genes based on their relative position with respect to specific areas of the gene (i.e. the promoter region, the first exon, intronic areas, etc.). For example, in a ChIP-seq experiment, the predicted transcription factor binding sites are generally expected to be located in the transcription start site (TSS) or promoter regions of the gene that is being regulated (Figure 5.10-A).

To achieve this objective, PaintOmics 3 incorporates RGmatch [38], a rule-based and highly configurable method for computing region-gene associations, annotating each association with the area of the gene where the region overlaps. As RGmatch was developed as a command-line algorithm, PaintOmics 3 provides a web interface to run the tool which is fully integrated within the application workflow (Figure 5.10-B).

The format for input files for region-based data is slightly different from the default because in this case the features are genomic regions. For this category of data, PaintOmics uses a modification of the BED format in which the first three columns indicate the name of the chromosome or scaffold, the start position of the feature in standard chromosomal coordinates (i.e. first base is 0), and the end position of the feature in standard chromosomal coordinates,



Region based omic

Omic Name:	DNAse-Seq data
Regions file (BED + Quantification):	[MyData]/DNaseSeq_PaintomicsDatε
Relevant regions file:	[MyData]/DNaseSeq_DE_Paintomics
Annotations file (GTF):	[inbuilt GTF files]/sorted_mmu.gtf
Distance (kb):	10
TSS region distance (bps):	200
Promoter region distance (bps):	1300
Overlapped gene area (%):	50
Overlapped region area (%):	50
GTF Tag for gene ID/name :	gene_id
Summarization method:	Mean
Report:	○ All regions ● Let me choose

Regions mapping at...

Upstream	Promoter	TSS	1st Exon	Introns	Gene body	Intr.	G.B.	Downstream
□	☑	□	□	☑	□			☑

Figure 5.10: RGmatch in PaintOmics 3. **(A)** - Associations between genes and ChIP-seq regions, valid regions should match to the transcription start site (TSS) or promoter regions of the gene (area of interest). **(B)** - The input web-form for RGmatch.

respectively. The remaining columns contain the quantification values for each sample in the experiment, again, on a logarithmic scale (Table 5.3-A). Following this idea, the file with relevant regions (e.g. the DE regions) for this category must contain three columns, corresponding to the chromosome, the start, and the end position (Table 5.3-B).

(A)

#Chr	Start	End	Treated 0h	Treated 2h	Treated 6h
10	100487291	100487483	0.514722	0.938385	0.43417
10	100487717	100487888	0.785665	0.679	0.723835
...	...	...	...	...	...

(B)

#Chr	Start	End
8	66438023	66438216
8	66913732	66913929
...	...	...

Table 5.3: Example of input for a region-based "omics" type dataset. (A) - Quantification file contains the chromosome number, start and end positions (first, second, and third columns, respectively) for each region, followed by the quantification values for three different time points, on a logarithmic scale. **(B)** - Relevant regions are provided as a list of genomic coordinates.

For each genomic region in the input file, RGMatch computes all the possible associations between the region and closest genes, and reports the areas of the gene that the region overlaps. Parameters such as the minimum percentage of the area of the gene that should be overlapped to accept an association, or the distance of TSS and promoter areas of the gene, can be configured for more accurate associations. By default, the output for RGMatch includes all the found associations for the given parameters, including regions that could match to undesired areas of the gene. Besides, multiple regions can be reported for the same area of a certain gene. For that reason, PaintOmics 3 incorporated a post-processing of the RGMatch output as follows.

- First, reported pairs "region-gene" are filtered based on the gene areas selected by the user. Gene areas can be easily chosen using the selector at the bottom of the web interface (Figure 5.10-B).

- Then PaintOmics 3 examines the list of associations and applies the selected strategy for resolving them. Available strategies are: "Do nothing" (i.e. leave the list of associations as it is), "Mean" (i.e. calculate the mean of the quantification values for all the regions matching the same gene area) and "Maximum" (choose the region with a maximum fold-change).

Converting miRNA-based data to genes

Micro RNAs (miRNA) are small, non-coding RNA molecules that can bind to mRNA transcripts of certain protein-coding genes (known as target genes) and negatively control their translation or even cause their degradation. Identifying the miRNA targets accurately is crucial for the better understanding of cellular functions [137]. Although many miRNA target prediction algorithms, as well as methods for experimental validation, have been developed during the last years, an effective prediction of miRNA-mRNA interactions remains challenging due to the interaction complexity and a limited knowledge of rules governing these processes [155]. Nevertheless, the inferred miRNA-mRNA interactions are usually gathered in dedicated miRNA-mRNA repositories. Some examples of popular databases are miRWalk2.0 [31], a comprehensive archive, supplying the largest available collection of predicted and experimentally verified miRNA-target interactions, miRTarBase [23], an experimentally validated microRNA-target interactions database, and miRBase database [74], a central online repository that stores miRNA nomenclature, sequence data, annotation and target prediction.

For a fully support of miRNA data, PaintOmics 3 uses "miRNA2Genes", a rule-based python script that automatically match miRNAs to their target protein-coding genes. This tool processes the input miRNA quantification data and assigns the expression values to the known list of target genes for each miRNA. The tool includes many options to customize the resulting gene list and is accessible through a user-friendly web interface.

In order to compute the associations, the required input for the tool is:

1. A tabulated file containing the quantification values for all the miRNAs (Figure 5.11-A).
2. A tabulated file containing the list of miRNA → target gene associations (Figure 5.11-B). This file can be downloaded from any of the public databases described above.

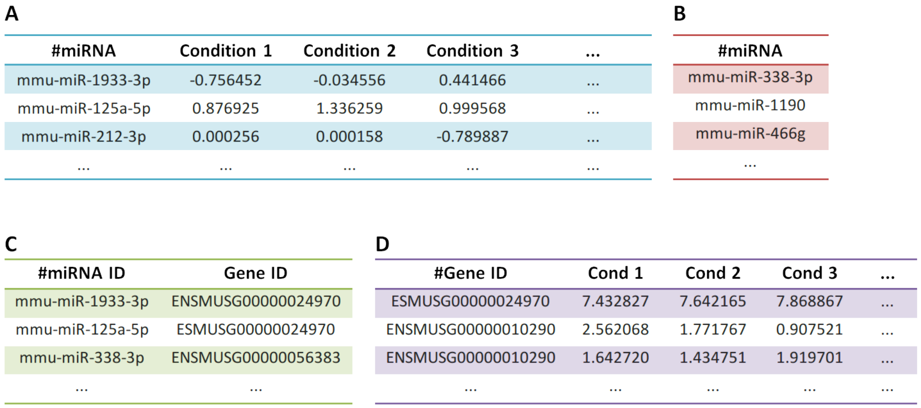
Both input files must use the same convention for the identifiers or names for the miRNAs. Additionally, two secondary files can be provided for more accurate results.

3. A list of relevant miRNAs, usually the DE miRNAs (Figure 5.11-C).
4. A transcriptomics file with the quantification values for the expression of genes for the same experiment. This file is necessary for filtering the reported target genes based on the correlation between the target gene expression and the miRNA expression. Naming convention must be the same that the used in the miRNA → target gene file (Figure 5.11-D).

As a general rule, each miRNA has numerous known target genes. Nevertheless, the presence of a miRNA does not necessary mean that a certain target gene is being regulated for that miRNA. Hence, it is necessary to discriminate those genes that may be affected by the action of a miRNA from the complete list of potential target genes for that miRNA. Assuming that all files explained above are provided, *miRNA2Genes* includes the following selection strategies.

(i) If the list of relevant miRNAs is provided, the user can choose between reporting the target genes for all miRNAs in the input files, or just those target genes being regulated by a relevant miRNAs (e.g. the DE miRNAs), ignoring the rest.

(ii) If the transcriptomics quantification file is provided, reported target genes can be discriminated based on the existing correlation between the quantification for a miRNA and the codified by target genes. Usually, it is expected a negative correlation between miRNA and target genes being regulated. The tool calculates the correlation for each pair miRNA-target gene and the usage of a cut-off for correlation value determines the selection of the genes that are finally reported (Figure 5.12). If the transcriptomics file is not available, the default score methods is the value of the fold-change for the expression of each miRNA.



A

#miRNA	Condition 1	Condition 2	Condition 3	...
mmu-miR-1933-3p	-0.756452	-0.034556	0.441466	...
mmu-miR-125a-5p	0.876925	1.336259	0.999568	...
mmu-miR-212-3p	0.000256	0.000158	-0.789887	...
...	...	...	...	...

B

#miRNA
mmu-miR-338-3p
mmu-miR-1190
mmu-miR-466g
...

C

#miRNA ID	Gene ID
mmu-miR-1933-3p	ENSMUSG00000024970
mmu-miR-125a-5p	ESMUSG00000024970
mmu-miR-338-3p	ENSMUSG00000056383
...	...

D

#Gene ID	Cond 1	Cond 2	Cond 3	...
ESMUSG00000024970	7.432827	7.642165	7.868867	...
ENSMUSG00000010290	2.562068	1.771767	0.907521	...
ENSMUSG00000010290	1.642720	1.434751	1.919701	...
...	...	...	...	...

Figure 5.11: Example of the input for a miRNA-based omics type dataset. (A) - The quantification file contains the miRNA name or identifier (first column) followed by the quantification values for the different time points, on a logarithmic scale. **(B)** - Differentially-expressed genes are provided as a list. **(C)** - Example for the miRNA → target gene associations file. **(D)** - Example for transcriptomics file.

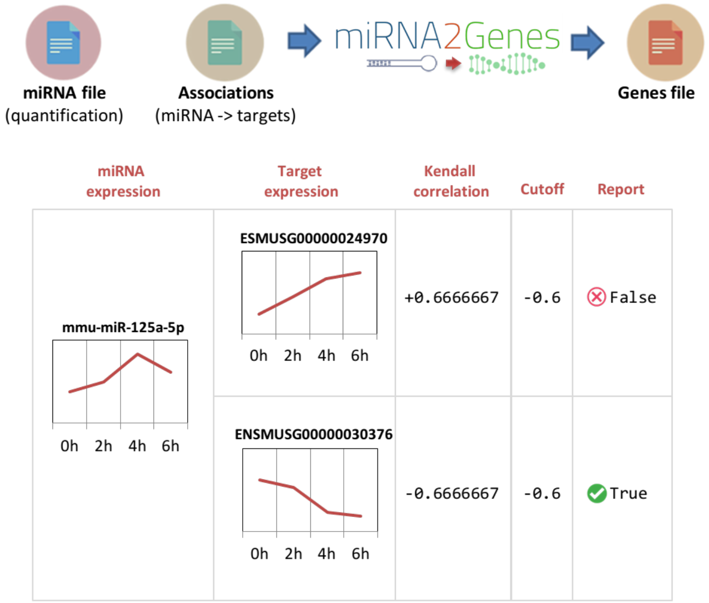


Figure 5.12: Example for the filtering by correlation for the miRNA2Genes results. The image displays the profiles for the expression for certain miRNA and two of its targets genes. Using Kendall correlation and a cut-off of -0.6 we can discriminate the genes that are included in the results for the tool. For this example, the first target gene would be considered as "not regulated by the miRNA" and, consequently, it would be rejected.

Identifier and name conversion for gene-based data

The lack of standard naming conventions for the biological features is a common hurdle when trying to integrate or extract information from multiple data sources [91]. Although some effort has been made to move to a non-redundant standardised naming domain [76], most of the databases and bioinformatics resources are independent from each other and assign custom naming conventions to the biological features they incorporate. For example, while some of the major public databases such as GenBank [12], NCBI RefSeq [85] or UniProt [11] organise the stored data based on accession numbers, other resources use organism-specific naming conventions (e.g. the Gene Ontology database [15]) or numerical codes (Entrez database [84]). The KEGG database uses different gene name conventions for each species, generated from publicly available resources, mostly NCBI RefSeq and GenBank. This usually results in extra work for the researchers who need to move from their usual set of identifiers to the set of identifiers accepted by the resource to be used. Historically, multiple resources have been developed to address this problem of identifier mapping. These tools provide web-based applications for converting the user's input gene or gene product identifiers, usually complemented by application programming interfaces(APIs) and web services. Examples are the DAVID Gene identifier (ID) Conversion Tool (DICT) [55], CRONOS [148], MatchMiner [20], and BridgeDB [58]. Both CRONOS and MatchMiner have important limitations in terms of the number of available species and because their databases have been outdated since 2011. On the other hand, DICT supports over 65,000 species and its databases are periodically updated. Additionally, DICT can be accessed programmatically using the DAVID web services and the DAVID API. Nevertheless, both tools are limited by the number of genes that they can process per job and the number of jobs each user can perform per day. Finally, BridgeDB is a software library intended to provide a standardised interface layer through which bioinformatics tools can be connected to different identifier mapping services using short and simple code. The BridgeDB API takes two different forms: a Java API and a REST-based API that can be embedded into non-Java applications. Although the BridgeDB approach is significantly different, we found common limitations such as the number of available species, the limited number of genes per job that the tool can process, and greatly reduced tool performance when querying large lists of identifiers. Consequently, as part of the PaintOmics 3 development, a Python module for Name/ID translation was implemented in order to extend the scope of the application, allowing users to input their data regardless of the naming domain it uses.

This Python module fetches the translation information from public databases such as Ensembl, PDB, NCBI RefSeq, and KEGG, processes the downloaded files, generates the translation tables, and stores them in MongoDB collections. The central pillars for translation process are the transcripts. For example, given a feature ID (gene, protein, or transcript) for database A, which we want to translate into a valid gene name for database B, first the system retrieves the list of transcripts associated with the feature (if any). Then, for each transcript ID in database A, it searches for the equivalent transcript identifier for database B. Finally, as we requested the gene name, the system finds the gene name associated for each transcript found (Figure 5.13). Although this method has some limitations, mainly because the intersection between databases is not complete (which means that some biological features in database A do not exist in database B), in general terms the percentage of translated features is good and is sufficient to work properly with PaintOmics. Alternatively, users can translate their data using third-party tools and input them into the corresponding KEGG name domain for the studied organism.

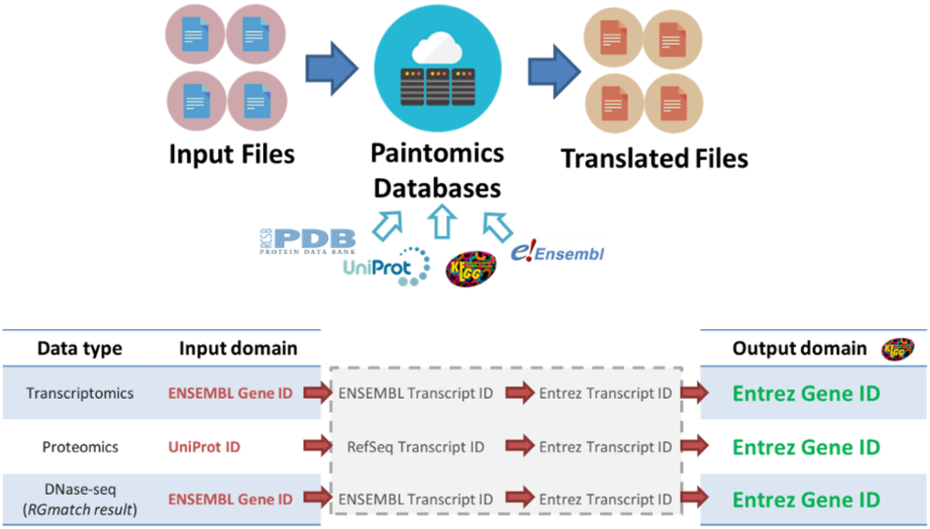
In both cases, PaintOmics 3 processes the input and presents the user with some statistics summarising the distribution for the data for each omics data type as well as the percentage of features translated to valid KEGG feature names, as shown in Figure 5.14. These results can be downloaded as text files to manually evaluate the translation.

Resolving metabolite names

As seen in previous chapters, high-throughput techniques for metabolomics use certain properties of the particles, such as the mass or the chemical composition, for the detection and identification of the metabolites in a sample. Although these technologies are extensively used in the biomolecular researching, identifying a metabolite unambiguously and confidently is yet a bottle-neck in metabolomics studies, especially when the differences between two metabolites are almost undetectable as in the case of the isomerism. Due to the results in later stages can be affected by the presence or absence of certain metabolites, PaintOmics 3 includes some tools for resolving any existent ambiguity during the translation of the supplied metabolites to KEGG metabolite names. Hence, for each input metabolite, PaintOmics 3 generates a list of potentially related metabolites based on the similarities in their names as follows:

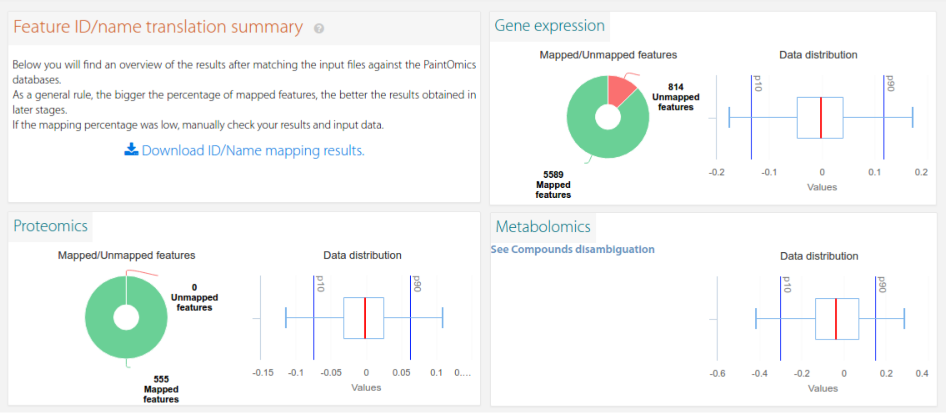
- Identical names receive a score of 1.
- Known names for isomers and other structural variations of the metabolite (e.g. beta-Alanine, L-Alanine or D-Alanine) receive a score of 0.9.
- For the remaining metabolites, similarity is calculated using a python algorithm includes that returns a measure of the similarity between two sequences.

The resulting list of metabolites is then displayed to the user (Figure 5.15). By default, metabolites with a similarity score of 0.9 or more are automatically selected, but the user can change this selection manually. To avoid duplicated



Data type	Input domain			Output domain
Transcriptomics	ENSEMBL Gene ID	ENSEMBL Transcript ID	Entrez Transcript ID	Entrez Gene ID
Proteomics	UniProt ID	RefSeq Transcript ID	Entrez Transcript ID	Entrez Gene ID
DNase-seq (*RGmatch result*)	ENSEMBL Gene ID	ENSEMBL Transcript ID	Entrez Transcript ID	Entrez Gene ID

Figure 5.13: Identifier/name conversion in PaintOmics 3.



Feature ID/name translation summary

Below you will find an overview of the results after matching the input files against the PaintOmics databases.
As a general rule, the bigger the percentage of mapped features, the better the results obtained in later stages.
If the mapping percentage was low, manually check your results and input data.

Download ID/Name mapping results.

Figure 5.14: Results for the ID/name conversion step in PaintOmics 3. For each input data type, an interactive chart indicating the percentage of translated features is shown. Additional statistics about the data distribution are also reported.

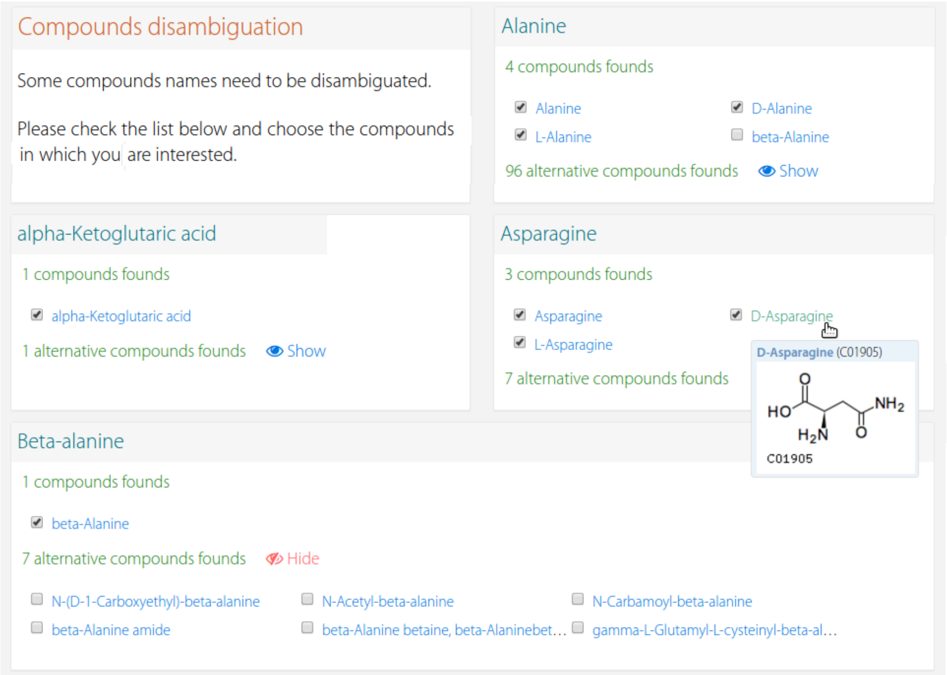
metabolites, the user will be notified in case of selecting a metabolite twice and only the metabolite with higher similarity score will be selected by default.

5.2.5 Pathway enrichment analysis

Pathway analysis is a powerful tool for understanding the biology underlying the data contained in large lists of differentially-expressed genes, metabolites, and proteins resulting from modern high-throughput profiling technologies. The central idea of this approach is to group these long lists of individual features into smaller sets of related biological features (genes and metabolites), usually based on biological processes or cellular components in which genes, proteins, and metabolites are known to be involved [71].

One popular method for pathway analysis is pathway enrichment analysis (PEA), which is the approach adopted by the current version of PaintOmics 3. Figure 5.16 depicts the process followed to determine the set of significantly enriched pathways for the input data. First, the tool identifies the subset of genes, proteins, and metabolites that participate in a particular KEGG pathway for the input. Then, it evaluates the fraction of those biological features which overlaps with the set of features that the researcher considered significant, usually features showing significant changes in expression or concentration (e.g. the differentially-expressed genes). In the final step, the tool computes the significance of the overlap using the Fisher exact test. The p-value obtained can be interpreted as a measurement of the confidence that this overlap is due to chance (null hypothesis). The smaller the p-value, the more likely that the association between the features of interest and the pathway is not random, i.e. an overrepresentation of the significant biological features of that pathway may exist. As a general rule, a p-value of 0.05 is accepted as the threshold indicating a statistically significant association.

From this analysis the application ranks the pathways for each omics data type and sorts them from higher to lower statistical significance values. However, extracting meaning from multiple significance values can be complicated, especially when individually evaluating very different omics data types. Therefore, PaintOmics 3 incorporates an additional step in the process in order to obtain a



Compounds disambiguation

Some compounds names need to be disambiguated.

Please check the list below and choose the compounds in which you are interested.

Alanine

4 compounds founds

☑ Alanine ☑ D-Alanine

☑ L-Alanine □ beta-Alanine

96 alternative compounds founds Show

alpha-Ketoglutaric acid

1 compounds founds

☑ alpha-Ketoglutaric acid

1 alternative compounds founds Show

Asparagine

3 compounds founds

☑ Asparagine ☑ D-Asparagine

☑ L-Asparagine

7 alternative compounds founds

D-Asparagine (C01905)

Beta-alanine

1 compounds founds

☑ beta-Alanine

7 alternative compounds founds Hide

□ N-(D-1-Carboxyethyl)-beta-alanine □ N-Acetyl-beta-alanine □ N-Carbamoyl-beta-alanine

□ beta-Alanine amide □ beta-Alanine betaine, beta-Alaninebet... □ gamma-L-Glutamyl-L-cysteinyl-beta-al...

Figure 5.15: Compound disambiguation in PaintOmics 3. For each input metabolite, a panel is displayed containing all the possible disambiguation for the metabolite name. Metabolites with a similarity score over 0.9 are selected by default, while remaining options are displayed in a secondary section (see Beta-alanine in the figure). The user can get more details for each metabolite moving the mouse cursor over the metabolite name, and clicking in the name, a new window is opened in the browser with the entry in the KEGG database for the selected metabolite.

joint significance value which indicates its relevance in the context of the biological system for each pathway. Hence, it applies the Fisher combined probability test, a statistical method which allows the results from several independent tests for similar null hypotheses to be combined. This method combines the p-values for each test into one test statistic (X) using the formula:

$$X = -2\sum_{i=1}^{k} \ln(p_i) \tag{5.1}$$

where p_i is the p-value for the i^{th} hypothesis test, k is the number of tests being combined and with X following a χ^2 distribution with $2k$ degrees of freedom, from which a p-value for the global hypothesis can be easily obtained.

Figure 5.17 shows an example of the visual representation of the ranking as a table. Note that the upper positions in the table correspond to the most relevant pathways, based on the combined p-value. Each row in the table represents a Pathway. The first column displays the name of the pathway, while the second and third columns indicate the total number of genes and metabolites found in the pathway. A colour label close to the pathway name identifies the main classification for the pathway, using the same colour code as that used in the "Pathway classification" section. The last column provides some useful links to external sources, such as KEGG or PubMed. The remaining columns indicate the significance value for each omics data type for each pathway, and a colour scale is used to highlight the level of enrichment for each one. When the mouse is moved over each cell the application displays the contingency table used for obtaining the significance value.

5.2.6 Exploring data with PaintOmics 3

Hierarchical classifications for KEGG Pathways

As mentioned in previous sections, KEGG Pathways are organised in a hierarchy around seven main classifications (*Cellular Processes*, *Drug Development*, *Environmental Information Processing*, *Genetic Information Processing*, *Human Diseases*, *Metabolism*, and *Organismal Systems*) and over 50 secondary

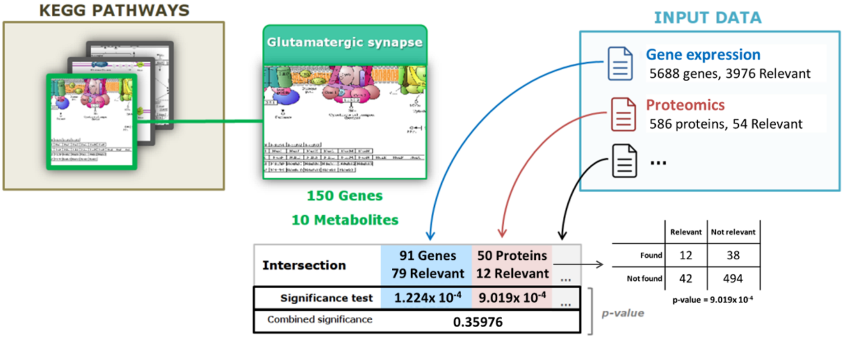
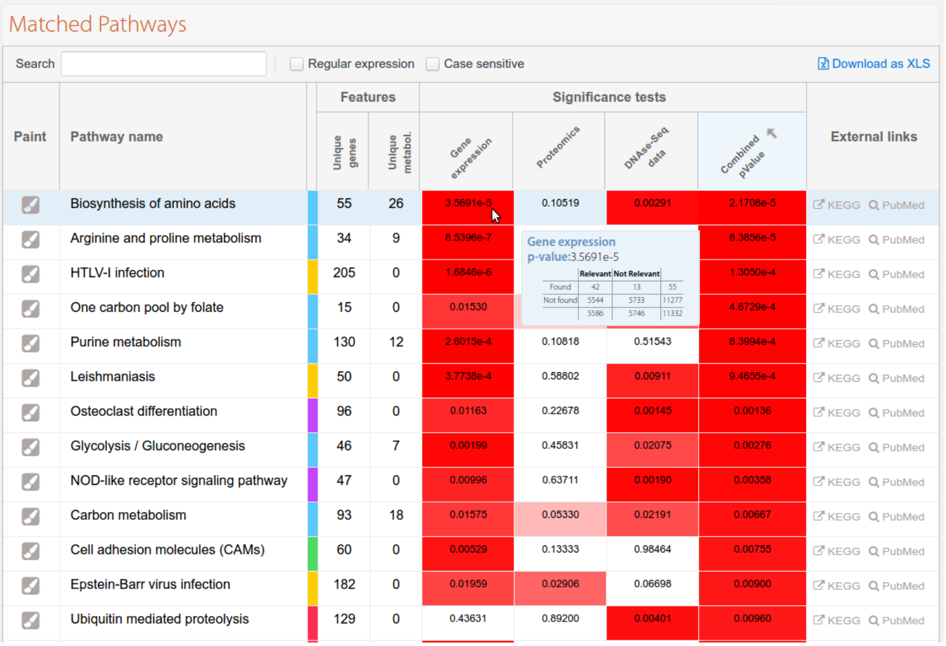


Figure 5.16: Significance evaluation for the Glutamatergic synapse KEGG pathway. The Glutamatergic synapse pathway contains a total of 150 genes and 10 metabolites that are known to participate in the biological processes. First, PaintOmics 3 finds the intersection between the features in the pathway and the features at each input data type. Next, the tool evaluates the fraction of relevant features that fall into the intersection. PaintOmics 3 uses these values to calculate a significance value for each omics data type. Lastly, a combined p-value is calculated using the Fisher combined probability test.


Matched Pathways

Search □ Regular expression □ Case sensitive Download as XLS

Paint	Pathway name	Features		Significance tests				External links
		Unique genes	Unique metabol.	Gene expression	Proteomics	DNAse-Seq data	Combined pValue	
	Biosynthesis of amino acids	55	26	3.5691e-5	0.10519	0.00291	2.1708e-5	KEGG PubMed
	Arginine and proline metabolism	34	9	8.5396e-7			8.3856e-5	KEGG PubMed
	HTLV-I infection	205	0	1.6846e-6			1.3050e-4	KEGG PubMed
	One carbon pool by folate	15	0	0.01530			4.6729e-4	KEGG PubMed
	Purine metabolism	130	12	2.6015e-4	0.10818	0.51543	8.3994e-4	KEGG PubMed
	Leishmaniasis	50	0	3.7738e-4	0.58802	0.00911	9.4655e-4	KEGG PubMed
	Osteoclast differentiation	96	0	0.01163	0.22678	0.00145	0.00136	KEGG PubMed
	Glycolysis / Gluconeogenesis	46	7	0.00199	0.45831	0.02075	0.00276	KEGG PubMed
	NOD-like receptor signaling pathway	47	0	0.00996	0.63711	0.00190	0.00358	KEGG PubMed
	Carbon metabolism	93	18	0.01575	0.05330	0.02191	0.00667	KEGG PubMed
	Cell adhesion molecules (CAMs)	60	0	0.00529	0.13333	0.98464	0.00755	KEGG PubMed
	Epstein-Barr virus infection	182	0	0.01959	0.02906	0.06698	0.00900	KEGG PubMed
	Ubiquitin mediated proteolysis	129	0	0.43631	0.89200	0.00401	0.00960	KEGG PubMed

Gene expression
p-value:3.5691e-5

	Relevant	Not Relevant	
Found	42	13	55
Not found	5544	5733	11277
	5586	5746	11332

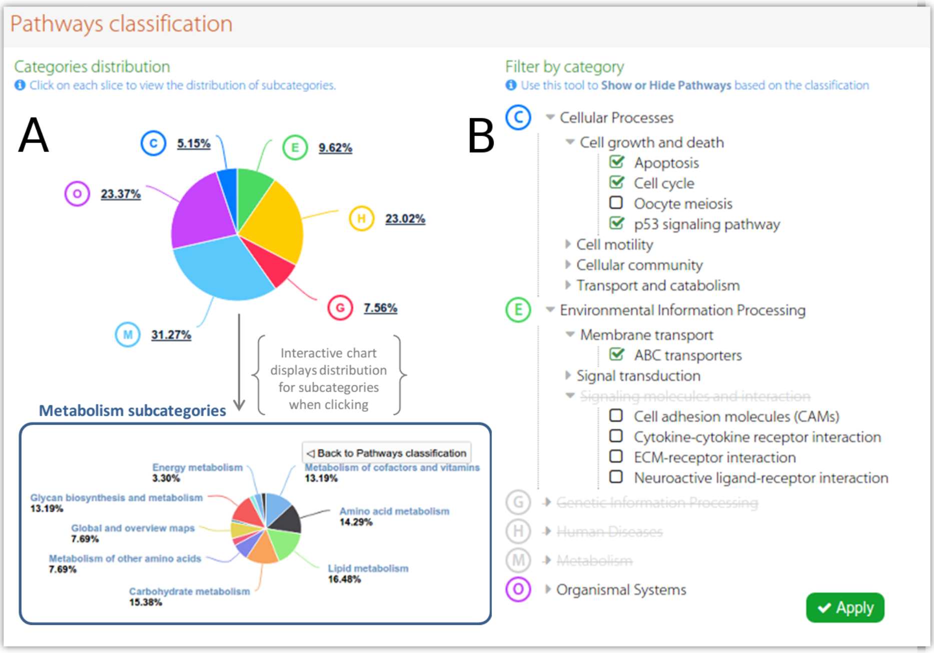
Figure 5.17: Pathways found for an experiment combining data from Gene Expression, Proteomics, and DNase-seq.

classifications. The PEA performed by PaintOmics may return a high number of pathways, however, sometimes not all of these are of interest to the user in functional terms. For example, in the study of certain biological processes in murine cells the PEA may return pathways from the *Human Diseases* classification. Hence, the presence of these undesired pathways can hamper interpretation of the results. For this reason, PaintOmics 3 includes a *Pathway Classification* tool which organises the reported pathways based on their KEGG classifications and provides some details about their hierarchical distribution (Figure 5.18-A). This tool allows the user to browse the results and hide any undesired pathways based on their main or secondary classifications (Figure 5.18-B).

The pathways interaction network

Although tables and tree diagrams are very helpful for visualising hierarchical information and understanding the PEA results, they do not always sufficiently represent the underlying information or relationships that exist between the KEGG pathways. Examples of this type of hidden information may include genes or metabolites shared between biological processes, belonging to the same classification of two or more pathways, or even that two or more pathways show a similar behaviour or trend at similar biological agent (gene or metabolite) concentration levels, under the same conditions.

In this scenario network-based approaches become a valid option for representing such biological interactions in combination with the results obtained from the PEA. PaintOmics 3 includes an interactive network where nodes represent pathways and the existing relationships are displayed by drawing edges (Figure 5.19). In the current version of PaintOmics, the existence of an edge between two nodes indicates that both pathways share an important percentage of biological features, suggesting that both biological processes may be somehow related. Additional information is represented by using different visual resources, as follows: the presence of nodes in the network depends on the percentage of input biological features that participate in the biological process. By setting this threshold we are able to discriminate pathways that do not contain enough input information to be considered by the user. The



Pathways classification

Categories distribution

Click on each slice to view the distribution of subcategories.

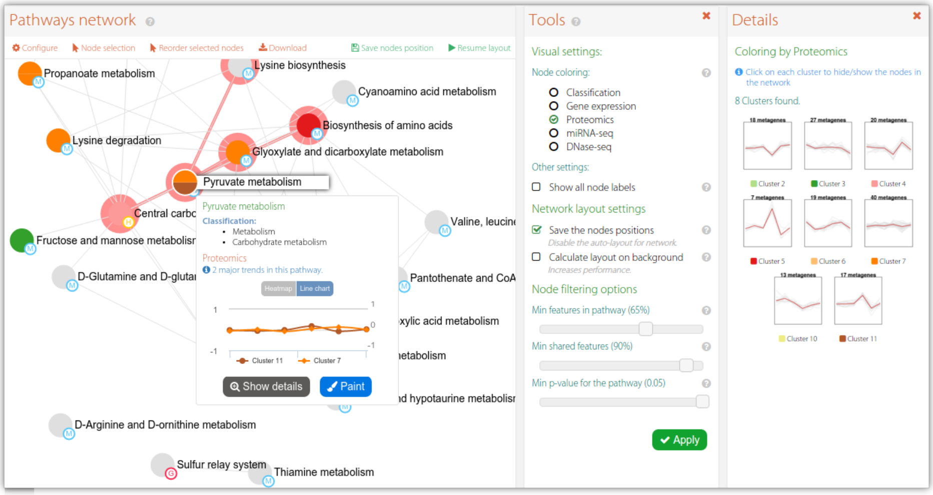
Filter by category

Use this tool to **Show or Hide Pathways** based on the classification

- Cellular Processes
  - Cell growth and death
    - ☑ Apoptosis
    - ☑ Cell cycle
    - ☐ Oocyte meiosis
    - ☑ p53 signaling pathway
  - Cell motility
  - Cellular community
  - Transport and catabolism
- Environmental Information Processing
  - Membrane transport
    - ☑ ABC transporters
  - Signal transduction
  - ~~Signaling molecules and interaction~~
    - ☐ Cell adhesion molecules (CAMs)
    - ☐ Cytokine-cytokine receptor interaction
    - ☐ ECM-receptor interaction
    - ☐ Neuroactive ligand-receptor interaction
- ~~Genetic Information Processing~~
- ~~Human Diseases~~
- ~~Metabolism~~
- Organismal Systems

A
B
C 5.15%
E 9.62%
O 23.37%
H 23.02%
G 7.56%
M 31.27%
Interactive chart displays distribution for subcategories when clicking
Metabolism subcategories
◁ Back to Pathways classification
Energy metabolism 3.30%
Metabolism of cofactors and vitamins 13.19%
Glycan biosynthesis and metabolism 13.19%
Amino acid metabolism 14.29%
Global and overview maps 7.69%
Metabolism of other amino acids 7.69%
Lipid metabolism 16.48%
Carbohydrate metabolism 15.38%
C
E
G
H
M
O
✔ Apply

Figure 5.18: Pathway classification (A) and Classification filters (B) in PaintOmics 3. Both tools allow users to discriminate pathways which are less important for their particular experiment.



Pathways network

Pyruvate metabolism

Classification:
- Metabolism
- Carbohydrate metabolism

Proteomics

2 major trends in this pathway.

Tools

Visual settings:

Node coloring:
- ○ Classification
- ○ Gene expression
- ● Proteomics
- ○ miRNA-seq
- ○ DNase-seq

Other settings:
- ☐ Show all node labels

Network layout settings
- ☑ Save the nodes positions
  Disable the auto-layout for network.
- ☐ Calculate layout on background
  Increases performance.

Node filtering options

Min features in pathway (65%)

Min shared features (90%)

Min p-value for the pathway (0.05)

Details

Coloring by Proteomics

Click on each cluster to hide/show the nodes in the network

8 Clusters found.

Figure 5.19: Example of an interactive network diagram showing pathway-pathway interactions in PaintOmics 3. Nodes represent pathways with a combined p-value lower than 0.05. In this case the node colour indicates the classification for the pathway based on the trend of the biological features for Proteomics data. Finally, an existing edge between two nodes indicates that both biological processes are closely related in biological terms, e.g. the reaction described by pathway A fires another biological process described by pathway B.

sizes for the nodes are directly proportional to the statistical significance of the represented pathway (i.e. inversely proportional to the computed p-value). Spatial arrangement for nodes is also informative. By default, PaintOmics 3 uses ForceAtlas2 [60], a forced-direct layout algorithm which distributes the nodes based on repulsion and attraction forces until it reaches a steady state. While all nodes are affected by repulsion forces, the attraction forces act on pairs of nodes connected by edges in proportion to the weight associated with the edge (i.e. the percentage of shared biological features). Hence, spatial distribution will likely show nodes grouped into sets of related nodes, allowing users to rapidly identify potential pathways of interest for further research. Alternatively, users can manually distribute the nodes in the pathway using the tools provided in PaintOmics 3.

Last but not least, PaintOmics 3 makes use of different approaches for colouring the nodes in order to discriminate whether a pathway belongs to a KEGG classification or follows a certain regulatory trend for the biological features involved. More specifically, after the PEA step, PaintOmics 3 calculates the major regulatory trend(s) for the features in each ranked pathway, for each omics type. The strategy adopted, implemented as an external R routine, consists of three key steps: (i) for each omics data type it evaluates the regulatory profile of the biological features in different conditions or samples, based on the method proposed by Nueda et al. [98]. Essentially, the strategy applies principal component analysis (PCA) to the data matrices and obtains a set of "synthetic genes", also called *metagenes*, that depict the most representative regulation patterns for each pathway, (ii) it then clusters the resulting profiles and groups the pathways based on similar behaviour for each omics data type (iii), and finally it assigns colours to the different clusters and renders the nodes in the network accordingly. The optimal number of clusters is calculated automatically using the average silhouette approach. The silhouette of an observation is a measure of how similar it is to its own cluster, compared to the other clusters (ranging from -1 to 1), providing an appreciation of the relative quality of the clusters. Obtaining the average silhouette of observations for different values of k, it is possible to determine the optimal number of clusters k, as the one that maximizes the average silhouette [114]. After that, users can switch the rendering strategy for the network by choosing the omics type they are interested in. Figure 5.20 shows an example of a network painted using 2 different approaches: colouring by classification (A) and colouring by RNA-seq trends (B). Finally, is important to note that each node in the network is extended with a pop-up showing the pathway trend (for the chosen omics type, Figure 5.21-A), as well as a supplementary panel summarising the trends for each cluster (Figure 5.21-B) and an advanced view showing the trends for each omics type for a chosen pathway (Figure 5.21-C). This tool introduces a valuable characteristic to the application: researchers can easily switch between the different complexity levels in the biological system of interest, moving from the pathway network to the single pathway level and, from there, down to the level of individual genes, with a single mouse click.

Multi-omic pathway-based visualisation

One of the core features of PaintOmics 3 is that it allows individual KEGG pathways to be studied in combination with user input data. After filtering the initial set of pathways based on their classification and/or behaviour criteria, users can then explore the pathways. Figure 5.22 illustrates an example of the typical workspace for pathway exploration. An important decision when designing visualisation tools is the layout of the application as it should be able to exploit the space available in the window, which is often limited by the size of the screen. Such limitations in the usable space usually result in restrictions in the displayable data (e.g. only being able to use a single view at a time) which become a hurdle for analysing and comparing information [159]. In contrast, overuse of simultaneous views may become confusing and thus dramatically reduce user-friendliness.

Considering the characteristics described above, the layout for the pathway exploration section was divided into three simultaneous, closable, resizable, and window-based panels which allow users to visualise the optimal amount of information they require to conduct their research. The main panel (Figure 5.22-A) contains an interactive diagram representing the KEGG pathway combined with the input data. As mentioned in previous sections (see section 5.2.2), several graphical resources are used for pathway diagrams, for example, boxes to represent gene products, circles for chemical compounds, etc. Re-

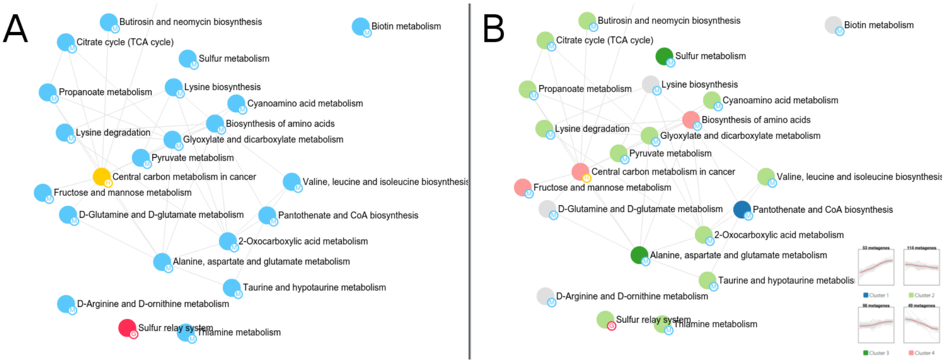
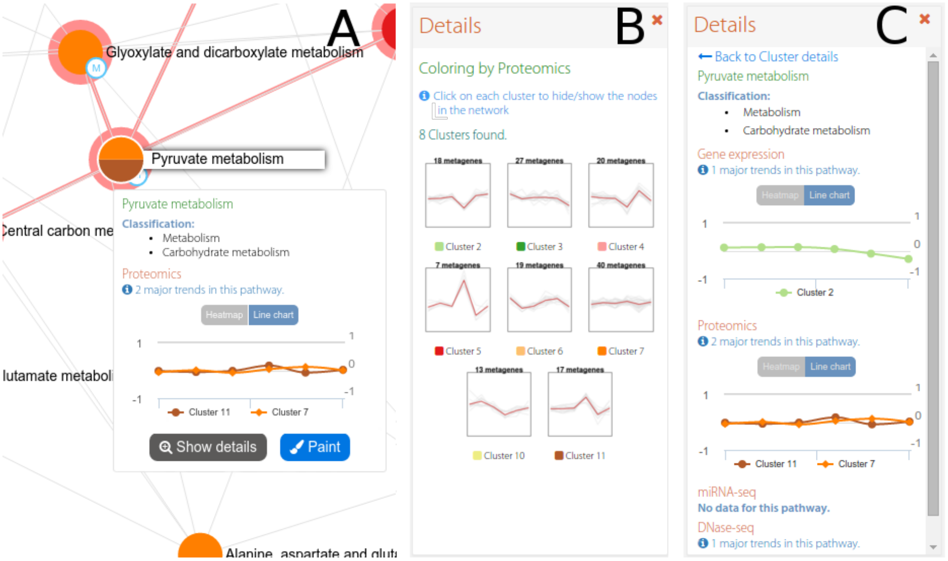

A
Butirosin and neomycin biosynthesis
Biotin metabolism
Citrate cycle (TCA cycle)
Sulfur metabolism
Propanoate metabolism
Lysine biosynthesis
Cyanoamino acid metabolism
Biosynthesis of amino acids
Lysine degradation
Glyoxylate and dicarboxylate metabolism
Pyruvate metabolism
Central carbon metabolism in cancer
Valine, leucine and isoleucine biosynthesis
Fructose and mannose metabolism
D-Glutamine and D-glutamate metabolism
Pantothenate and CoA biosynthesis
2-Oxocarboxylic acid metabolism
Alanine, aspartate and glutamate metabolism
Taurine and hypotaurine metabolism
D-Arginine and D-ornithine metabolism
Sulfur relay system
Thiamine metabolism
B
Butirosin and neomycin biosynthesis
Biotin metabolism
Citrate cycle (TCA cycle)
Sulfur metabolism
Propanoate metabolism
Lysine biosynthesis
Cyanoamino acid metabolism
Biosynthesis of amino acids
Lysine degradation
Glyoxylate and dicarboxylate metabolism
Pyruvate metabolism
Central carbon metabolism in cancer
Valine, leucine and isoleucine biosynthesis
Fructose and mannose metabolism
D-Glutamine and D-glutamate metabolism
Pantothenate and CoA biosynthesis
2-Oxocarboxylic acid metabolism
Alanine, aspartate and glutamate metabolism
Taurine and hypotaurine metabolism
D-Arginine and D-ornithine metabolism
Sulfur relay system
Thiamine metabolism
Cluster 1
Cluster 2
Cluster 3
Cluster 4

Figure 5.20: Example of three different approaches for node colouring systems in PaintOmics 3. Rendering by pathway classification **(A)** and by RNA-seq trends **(B)**.



Pyruvate metabolism
Classification:
- Metabolism
- Carbohydrate metabolism

Proteomics
2 major trends in this pathway.

Details
Coloring by Proteomics
Click on each cluster to hide/show the nodes in the network
8 Clusters found.

Details
← Back to Cluster details
Pyruvate metabolism
Classification:
- Metabolism
- Carbohydrate metabolism

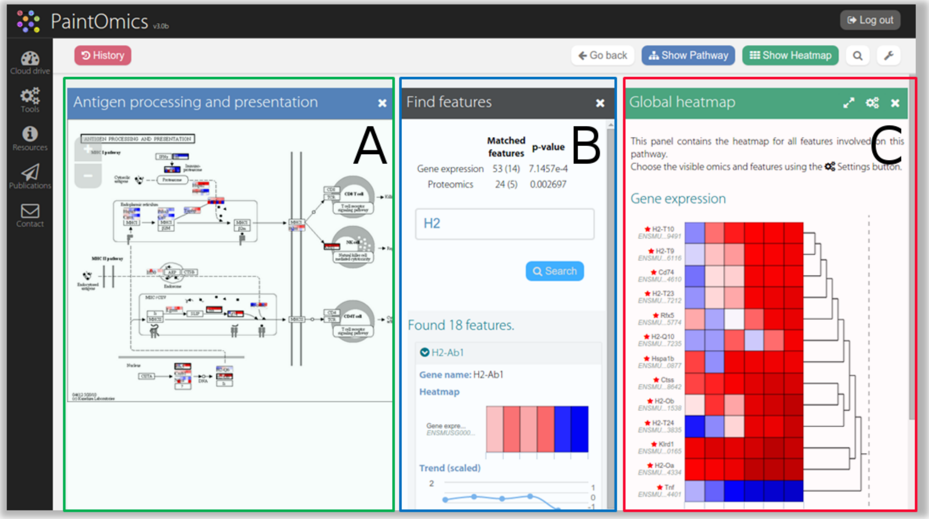
Gene expression
1 major trends in this pathway.

Proteomics
2 major trends in this pathway.

miRNA-seq
No data for this pathway.
DNase-seq
1 major trends in this pathway.

Figure 5.21: An example of an interactive network in PaintOmics 3. The interactive network panel **(A)** is complemented by a secondary panel which shows the trends for all the clusters in a given omics type **(B)**, or the trends for each omics type for a chosen pathway **(C)**.

sults from previous steps in PaintOmics which supply input data (omics measurements) are mapped to KEGG biological features (genes and metabolites). Consequently, users can easily navigate through the diagram and visualise the different values associated with each biological feature. As depicted in Figure 5.23-A, for each matched feature, a gridded box is overlaid on the KEGG image which shows an equal number of sections to the number of columns present in the input files, and as many rows as there are omics data-types supplied (this option is configurable by the user). Each section is coloured using as a heatmap approach, according to the corresponding ratio of the expression or concentration value. By default, the 10th and 90th percentiles of the ratio values determine the min and max tonalities used for heatmaps, thus preventing incorrect colouring due to the presence of outliers. Additionally, any biological features considered as significant for any of the omics data types (based on the input data), are highlighted by a thicker border and a special mark at the top right corner. This approach helps users to get an at-a-glance overview of the behaviour of all the biological features involved in the biological process of interest in the different conditions analysed.



	Matched features	p-value
Gene expression	53 (14)	7.1457e-4
Proteomics	24 (5)	0.002697

This panel contains the heatmap for all features involved on this pathway.
Choose the visible omics and features using the Settings button.

Figure 5.22: Example of a workspace for pathway exploration in PaintOmics 3. The layout for pathway exploration is divided into three panels. The main panel **(A)** contains the interactive pathway diagram, the auxiliary panel **(B)** shows some useful tools for analysis, and the secondary panel **(C)** contains information complementary to the KEGG diagram.

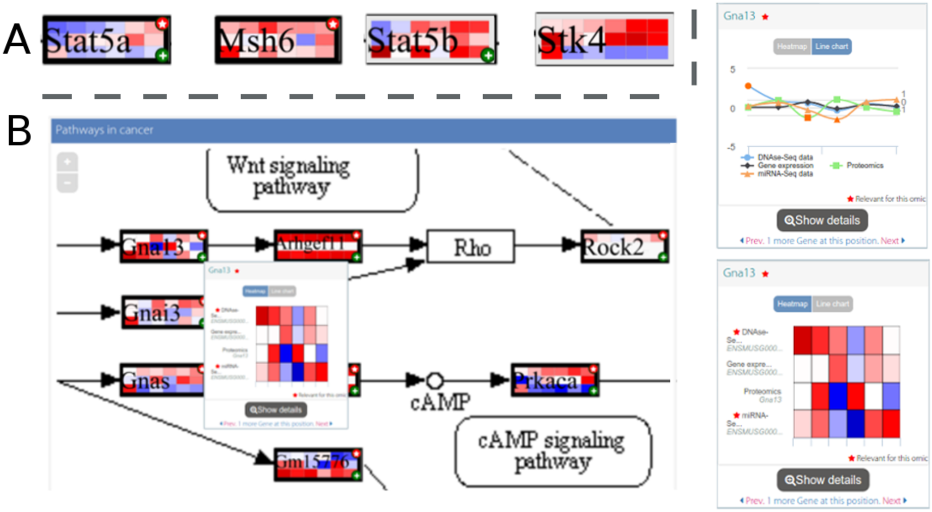


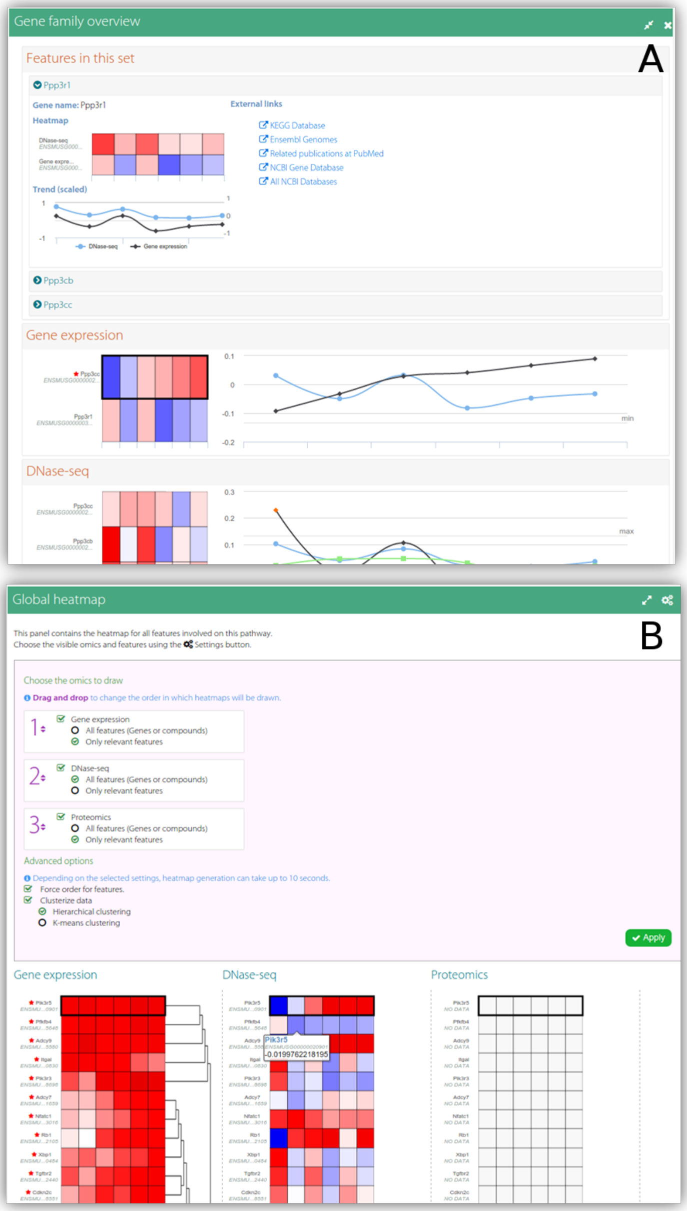
Figure 5.23: Heatmaps in PaintOmics 3. (A) - Concentration values for the features in the pathway are represented using heatmaps. **(B)** - When the user hovers the mouse over a heatmap box in the diagram, PaintOmics shows a floating panel with an interactive heatmap or a line chart depicting the concentration values for all the omics data types for the different conditions in the experiment.

As part of the interactivity of the diagram, users can hover the mouse over the biological features to get extended views of the feature values as an interactive heatmap (by default in blue-white-red scale, where blue indicates downregulation and red represents upregulation) or as a line chart depicting the expression or concentration trend for all the omics data types for the different conditions in the experiment (Figure 5.23-B). It is noteworthy that features which share functions or somehow both contribute equally to the biological process will also share the position in the diagram (i.e. a single box can represent one or more biological feature). For these sets of features (designated as feature families), by default the gridded box shows the values for the most significant features in the set (namely, the feature most frequently marked as significant for the uploaded omics types) and the existence of hidden features is indicated by a special mark in the bottom right corner (Figure 5.23-A). Using the extended view, users can switch between the features in the set and change the feature drawn in the diagram. In a secondary panel (Figure 5.22-C) users can find complementary information for the KEGG diagram. This panel shows a detailed view for feature families or individual features, including links to external databases or resources and charts describing the behaviour of the features for each supplied omics type (Figure 5.24-A). Alternatively, users can visualise a set of heatmaps which provide a global idea of the concentration levels for all the features involved in the biological process on this panel (Figure 5.24-B). Finally, an auxiliary panel (Figure 5.22-B) provides access to useful tools such as a search panel for quickly locating features in the pathway and a settings panel which offers several options for adjusting the visualisation to the user's requirements.

5.2.7 Administrator tools for PaintOmics

As mentioned in previous sections, the PaintOmics 3 client-side consists of two independent applications: the front-end application, oriented towards research and data analysis; and a back-end application that provides administrators with tools for internal management of the system. The current version of the back-end application includes tools for user management, for uploading auxiliary files (e.g. reference files for converting region-based data to genes), and for managing the data for the organisms installed. Despite some of the most commonly studied species being pre-installed in PaintOmics, as discussed in section 5.2.2, the KEGG database contains genomic and molecular-level information for up to 5000 species, which are also regularly updated, meaning that PaintOmics 3 must provide administrators with tools for installing and updating the organisms they use.

Although from 2011 access to the KEGG FTP site for organism data download was made available only to paying subscribers, the database includes a representational state transfer (REST)-style API for academic use [68]. Using the KEGG API, users can access the up-to-date databases on the KEGG server and retrieve specific information programmatically. The APIs available for the PATHWAY database provide methods to search genes, metabolites, and enzymatic reactions in the pathway as well as for retrieving the image file and the KGML representation for the diagram. Hence, a complete pipeline for retrieving, processing and storing the information in the MongoDB databases was developed in Python and included as part of the administrator tools. This pipeline can be executed as a stand-alone command-line program or through the administrator back-end application.



Gene family overview

Features in this set

Ppp3r1

Gene name: Ppp3r1

Heatmap

Trend (scaled)

External links

- KEGG Database
- Ensembl Genomes
- Related publications at PubMed
- NCBI Gene Database
- All NCBI Databases

Ppp3cb

Ppp3cc

Gene expression

DNase-seq

Global heatmap

This panel contains the heatmap for all features involved on this pathway.
Choose the visible omics and features using the Settings button.

Choose the omics to draw

Drag and drop to change the order in which heatmaps will be drawn.

1 ☑ Gene expression
- ○ All features (Genes or compounds)
- ● Only relevant features

2 ☑ DNase-seq
- ● All features (Genes or compounds)
- ○ Only relevant features

3 ☑ Proteomics
- ○ All features (Genes or compounds)
- ● Only relevant features

Advanced options

Depending on the selected settings, heatmap generation can take up to 10 seconds.

☑ Force order for features.
☑ Clusterize data
- ● Hierarchical clustering
- ○ K-means clustering

✓ Apply

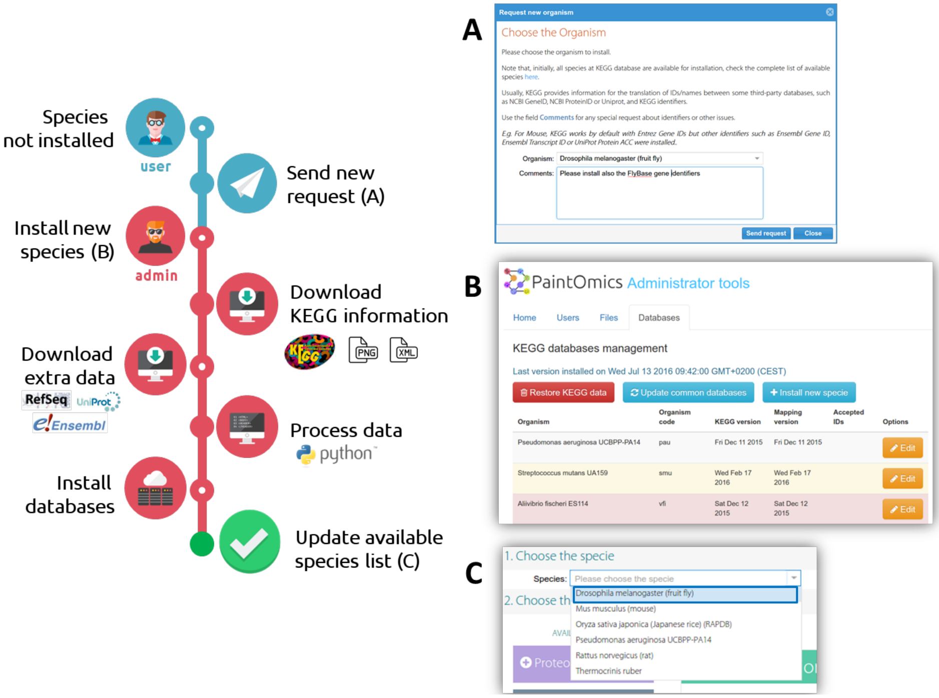
Gene expression

DNase-seq

Proteomics

Figure 5.24: Example of a secondary panel for pathway visualisation. (A) - Using this panel, users can visualise the concentration values for the features in a feature family together, grouped by omics type. **(B)** - Alternatively, this panel can be also used to inspect the concentration values for all the features that participate in the biological process, grouped by omics type. These global heatmaps can be customised by the user by applying different clustering methods, by forcing the order of the features, or by choosing between showing all or just a subset of the relevant features for each omics data type.

Figure 5.25 summarises the main steps in the installation process. Typically, the process starts when researchers contact the administrator requesting the installation of new organisms (if available in KEGG). The administrator can choose between installing a new species and updating an existing one. In both cases, PaintOmics will retrieve the required information (e.g. plaintext files describing the connections between biological features and pathways, files containing the names for pathways, etc.) using the KEGG API. Additionally, the administrator can download the feature ID/name translation information from third-party databases such as the ENSEMBL database or the RefSeq databases, by setting up the corresponding configuration files. After downloading, PaintOmics proceeds with the data processing and the generation of the MongoDB collections. Finally, the databases are installed and indexed to allow them to better perform when queried, and the new species are included in the available organisms chooser.



A

Request new organism

Choose the Organism

Please choose the organism to install.

Note that, initially, all species at KEGG database are available for installation, check the complete list of available species here.

Usually, KEGG provides information for the translation of IDs/names between some third-party databases, such as NCBI GeneID, NCBI ProteinID or Uniprot, and KEGG identifiers.

Use the field Comments for any special request about identifiers or other issues.

E.g. For Mouse, KEGG works by default with Entrez Gene IDs but other identifiers such as Ensembl Gene ID, Ensembl Transcript ID or UniProt Protein ACC were installed.

Organism: Drosophila melanogaster (fruit fly)

Comments: Please install also the FlyBase gene identifiers

Send request | Close

B

PaintOmics Administrator tools

Home | Users | Files | Databases

KEGG databases management

Last version installed on Wed Jul 13 2016 09:42:00 GMT+0200 (CEST)

Restore KEGG data | Update common databases | + Install new specie

Organism	Organism code	KEGG version	Mapping version	Accepted IDs	Options
Pseudomonas aeruginosa UCBPP-PA14	pau	Fri Dec 11 2015	Fri Dec 11 2015		Edit
Streptococcus mutans UA159	smu	Wed Feb 17 2016	Wed Feb 17 2016		Edit
Aliivibrio fischeri ES114	vfi	Sat Dec 12 2015	Sat Dec 12 2015		Edit

C

1. Choose the specie

Species: Please choose the specie

- Drosophila melanogaster (fruit fly)
- Mus musculus (mouse)
- Oryza sativa japonica (Japanese rice) (RAPDB)
- Pseudomonas aeruginosa UCBPP-PA14
- Rattus norvegicus (rat)
- Thermocrinis ruber

2. Choose th

AVAIL

Proteo

Ol

Figure 5.25: Installation process for new species in PaintOmics 3. Typically, the process starts when researchers request the installation of new organisms using the channels provided and finishes with the installation of the new species and its inclusion in the list of available organisms.

5.3 Results

5.3.1 Availability and requirements

PaintOmics 3 is free to use and is distributed under the GNU General Public License Version 3. A public copy of the application is hosted at the CIPF facilities (see table 5.3.1) and sources are available at GitHub, a popular web-based Git repository, allowing other laboratories to browse, propose code reviews, and even download the code in order to set up their own instance of the application. The documentation and guides for users and administrators are available at the free web platform Read the Docs (*see table 5.3.1 below*), which provides fully-searchable and easy-to-find documentation. All the documentation was written in Markdown markup language and are stored at the GitHub repository. As previously mentioned, the PaintOmics 3 server-side application was developed in Python and R, and has been extensively tested on Ubuntu and Debian Linux servers, although installation on other platforms (e.g. Windows-based systems) has not yet been tested.

Availability and requirements	
Project links	
Public Instance:	`https://bioinfo.cipf.es/paintomics`
Sources:	`https://github.com/fikipollo/paintomics`
Documentation:	`https://paintomics.readthedocs.org`
Other information	
Operating system(s):	Platform independent
Programming languages:	HTML, JavaScript, Python, R
Other requirements:	Web browser. Recommended Google Chrome.
License:	GNU General Public License Version 3

5.3.2 Discussion

The current trend towards the development of outstanding high-throughput technologies has boosted the scope and the ambitions of the biological projects associated with them. Thanks to these technological advances, a new range of research disciplines and measurement techniques have emerged, and as a natural consequence the combination of these, the trend towards compiling genome-wide measurements into the same experiments has arisen in Systems Biology. Nevertheless, the increasing use of high-throughput technologies has also multiplied the magnitude and the complexity of the data generated. Thus the exceptional amount of data created and its heterogeneity poses new challenges for effective data integration and analysis.

As discussed in this chapter, integrative visualisation can be both a powerful tool and an important bottleneck in data analysis. The main goal of visualisation tools is to provide intuitive data representation that allows researchers to validate their hypothesis or to interpret the results of their experiments. Although network-based tools, such as Cytoscape or Gephi, have been successfully used in many biological domains, the extraction of knowledge is hampered by two factors: the high levels of complexity of the knowledge involved in biological networks, where thousands of nodes are densely connected to each other; and the lack of underlying biological contexts that can increase the explanatory power of the relationships observed. Pathway-based visualisation tools such as MapMan and Pathview solve the problem of this lack of biological background, but they make it hard to expose hidden or new knowledge. In order to provide a more complete visualisation tool, PaintOmics 3 combines both network-based and pathway-based approaches. On the one hand, the platform provides fully interactive pathway visualisation, implemented using modern web-based technologies and complemented with useful information and links to external resources, which distinguishes it from other more static solutions such as Pathview. On the other hand, its use of pathway interaction networks in combination with different colouring and clustering strategies provides a valuable tool for revealing new associations and for results interpretation.

This integrative capability is also an important criterion when classifying these visualisation tools. While some tools, such as PaintOmics 2, KaPPa-View, and 3Omics restrict their accepted input to a closed set of omics data types (usually transcriptomics, metabolomics, and in some cases proteomics), the actual trend in integrative analysis indicates that this reductionist approach may not be enough for the incoming multi-omics era. Other tools such as IGV, IGB, MapMan, and Pathview allow users to upload and simultaneously visualise various types of data. Nevertheless, the integration capability for these tools is not complete as they do not solve existing issues with the visualisation of data from chromatin-profiling (e.g. ChIP-seq or DNase-seq) experiments. While MapMan and Pathview are able to display many omics types, they lack a method for visualising changes at the chromatin level. In contrast, genome browsers such as IGV and IGB, only support data which can be described using genomics coordinates and lack resources for effective visualisation of the changes in, for example, gene expression or protein concentration in the different conditions in an experiment. In opposition to these software solutions, PaintOmics 3 incorporates RGmatch, a method for computing the region-gene associations, that allows researchers to easily visualize their data for chromatin features in the context of the explored biological process.

This chapter described a novel method for integrative visualisation of many different types of omics data: PaintOmics 3, which works as a one-click web tool, and allows the effective and complete analysis and exploration of multi-omics data. Using this tool, researchers can easily move through the different complexity levels of several biological systems, from individual pathways to pathway networks, and from there, down to the level of individual genes and metabolites. This proposed approach supports joint visualisation of a wide variety of omics types, even where no other solutions for display at the pathway level currently exist. This valuable feature bridges the gap that exists in visualising changes which occur at the chromatin level, and those that happen at different levels of biochemical activity such as at the level of gene expression, protein activity, and metabolite concentration. It provides an effective method for visualising and understanding the underlying associations and dynamics of the regulation of cellular processes.

The current version of PaintOmics 3 has been successfully tested in the European 7th Framework Programme (FP7) STATegra Project [134] and has also been used in additional studies on numerous other organisms (mammal, plant, and bacteria models). The variety of organisms that PaintOmics 3 supports highlights its scope in comparison to other solutions such as MapMan and KaPPa-View, which are plant-specific tools, or 3Omics, which is only available for human-specific analyses. PaintOmics 3 supports a wide range of species in different biological kingdoms and offers the user the possibility of requesting the addition of any other organism available in the KEGG database. Furthermore, the inclusion of automatic translation for the feature name/identifier improves the usability of the tool, allowing users to work with their existing datasets without requiring them to first convert their data to the identifier domains used by KEGG for the selected organisms.

All of these aforementioned features, in addition to the ability of PaintOmics 3 to identify the set of significantly-enriched pathways, as well as its use of modern web resources, confirms PaintOmics 3 as an effective platform for full integration, analysis, and visualisation of multi-omics datasets.

6

Visualizing multilevel integration models using PaintOmics 3: a use case in Systems Biology

Parts of this chapter are contents of the manuscript: "Hernández-de-Diego R, Silberberg G*, Ferreiros I, van del Kloet F, Ramirez R, Schmidt A, Marabita F, Lagani V, Papoutsoglou G, Hankemeier T, Westernhuis J, Imhof A, Ballestar E, Meier D, Lappe M, Tsamardinos I, Mortazavi A, Merkenschlager M, Tenger J, Gomez-Cabrero D, and Conesa A. *A comprehensive guideline to the multiomics data analysis paradigm in time course perturbation studies*. (In preparation)".

6.1 Introduction

The objective for this chapter is to illustrate how PaintOmics 3 can be used in the context of real biomedical research for visualising and analysing the multi-omics data. For this purpose, we consider the data from the STATegra project [134], whose main objectives and experimental design were described in detail in Chapter 4. PaintOmics 3 was used extensively as part of the validation process for the conclusions for this study.

6.2 Running PaintOmics 3

Paintomics 3 is available as a web-service in the CIPF facilities. Nevertheless, users can choose to deploy their own Paintomics server by following the installation guidelines available in `http://paintomics.readthedocs.io/en/latest/0_install/`.

6.2.1 Data preparation for PaintOmics 3

Bearing in mind that the goal of PaintOmics is to visualize multiomics data and infer pathways with significant changes, a number of characteristics are expected for the input data.

- Experimental design should be consistent through omics. This implies that the same experimental conditions are measured for each omics and that the omics values file has the same number of columns for all omics.
- Ideally, each column should represent the value of one experimental condition. In other words, replicated measurements should be averaged to one value per experimental condition. Although this is not a hard requirement (the user could choose to submit replicates separately), PaintOmics does not include tools to further process replicates, and data will be treated completely independently.
- To maximally benefit from the coloring rules implemented in PaintOmics, data should be provided as log fold change values meaning that a value of 0 means no change in expression, a positive value is interpreted as up-regulation and a negative value means down-regulation.

For the STATegra collection, three biological replicates were obtained per time point and condition. Therefore, the first pre-processing requirement is to average replicates and compute log-fold change values between Ikaros and Control samples, at each time point. Equations 6.1 shows the first rows of feature values matrix for RNA-seq before and after data preprocessing. Next to feature values file, PaintOmics expects a list of relevant or significant features for each omics data type, which is simply a text file with features ID, one per row.

$$v_t = \frac{\sum_{i=1}^{n_t} ratio_{t,i}}{n_t} \tag{6.1}$$

with *t* signifying a valid time-point for the experiment, and *nt* representing the number of replicates for time-point t.

$$ratio_{t,i} = \log_2(\frac{Ikaros_{t,i}}{Control_{t,i}}) \tag{6.2}$$

This transformation was applied to the mRNA-seq, miRNA-seq, DNase-seq, proteomics and metabolomics STATegra data. ChIP-seq data were excluded since only 2 time points were measured with this technology, while methylation (RRBS-seq) data were not included because no differential methylation sites were detected in this study.

A

# Feature name	Batch-4-Ctr-0H	Batch-5-Ctr-0H	Batch-6-Ctr-0H	...	Batch-4-Ik-24H	Batch-5-Ik-24H	Batch-6-Ik-24H
ENSMUSG00000000001	14.79793	14.81117	15.12762	...	15.27448	15.04589	15.32687
ENSMUSG00000000085	14.21789	14.17325	13.80345	...	14.35347	14.64078	14.25783
ENSMUSG00000000093	8.266099	7.973803	8.260593	...	11.10137	11.63582	11.27248
ENSMUSG00000000148	12.21063	12.32635	12.40045	...	12.84333	12.89742	12.80058
...	...	...	...	...	...	...	...

B

# Feature name	Ikaros vs Ctr 0H	Ikaros vs Ctr 2H	Ikaros vs Ctr 6H	Ikaros vs Ctr 12H	Ikaros vs Ctr 18H	Ikaros vs Ctr 24H
ENSMUSG00000000001	0.04420	-0.08201	0.62356	0.11461	0.66559	0.55371
ENSMUSG00000000085	0.03339	0.369412	0.00021	0.58192	0.10794	0.29233
ENSMUSG00000000093	-1.1108	-1.03058	0.32627	1.01830	1.98777	2.49369
ENSMUSG00000000148	0.03154	0.292241	0.82802	0.54109	0.65787	0.42011
...	...	...	...	...	...	...

Table 6.1: First five rows for the RNA-seq data before (A) and after processing for use in PaintOmics 3 (B).

6.2.2 Data submission

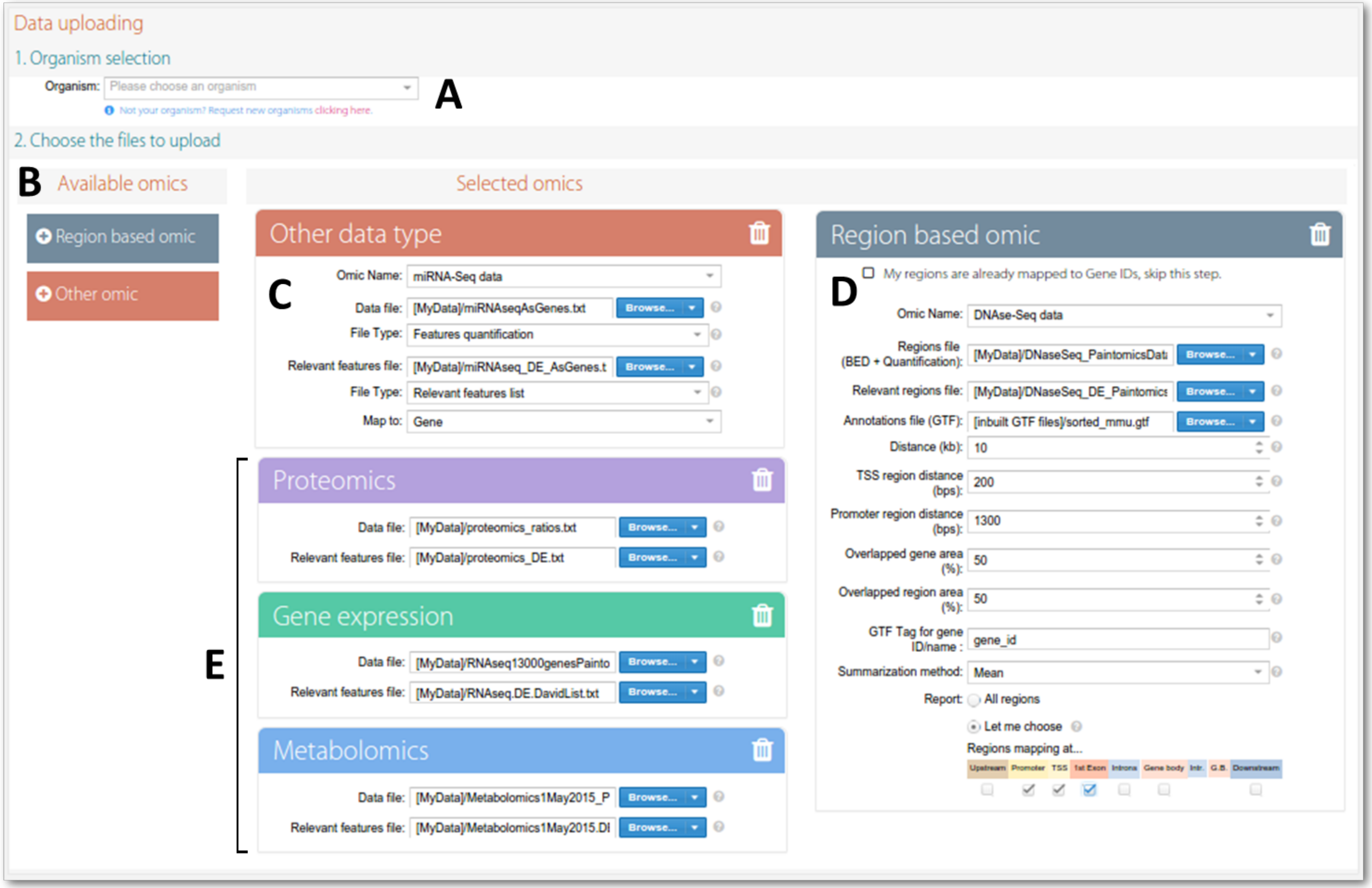
After preparing the files, we can proceed with the file submission. First, we choose the studied organism (Mus musculus, Figure 6.1-A) and we select as many omics types as the experiment involves. By default, *Gene Expression* and *Metabolomics* are selected; however, new omics types can easily be added or removed using the available tools. For this use case, we choose *Proteomics*, *Region-based omics*, and *miRNA-based omics* on the "Available omics" panel (Figure 6.1-B) which will then include three new sections on the submission form.

Once all the omics types are selected, we proceed to fill in the form. Gene expression (RNA-seq), proteomics and metabolomics are straight forward as they only require the features values file and the file with significant features (Figure 6.1-E). Input data for miRNA-seq is slightly different as we need first to execute the data transformation routine present in the Tools utility of the application. The form requires microRNA-target gene mapping file that is utilized to move from microRNA to gene IDs. Additionally we select to consider only microRNAs with a gene expression correlation value with their target mRNAs lower than -0.7. After this transformation has been computed we can use the Other Omics dialog to select and submit microRNA data (Figure 6.1-C).

Finally, we select "Region based omics" to configure the DNase-seq submission (Figure 6.1-D). First, we set the name for the data type (*DNase-seq*) and choose the main input file (normalised coverage for the regions), and the relevant features file with significant changing regions. Some interesting settings pertinent to this use case are:

- Reference genome annotations file: the latest build for mouse genome downloaded from the Ensembl website [157]. This file contains information about the gene structure which is needed in order to calculate the intersections between regions and genes.
- Summarisation method: determines how PaintOmics 3 resolves the gene areas with multiple matched regions. For this use case we choose the mean of the coverage values for all the regions.
- Report: discriminate the reported regions based on the overlapped area of the gene. For this use case we are interested only in regions that match the gene promoter area, the transcription start site (TSS), or the first exon.

In summary the following table represents the number of total and significant features submitted for each omics data type to Paintomics 3.



Data uploading

1. Organism selection

Organism: Please choose an organism

Not your organism? Request new organisms clicking here.

2. Choose the files to upload

Available omics

- Region based omic
- Other omic

Selected omics

Other data type

Omic Name:	miRNA-Seq data
Data file:	[MyData]/miRNAseqAsGenes.txt
File Type:	Features quantification
Relevant features file:	[MyData]/miRNAseq_DE_AsGenes.t
File Type:	Relevant features list
Map to:	Gene

Proteomics

Data file:	[MyData]/proteomics_ratios.txt
Relevant features file:	[MyData]/proteomics_DE.txt

Gene expression

Data file:	[MyData]/RNAseq13000genesPainto
Relevant features file:	[MyData]/RNAseq.DE.DavidList.txt

Metabolomics

Data file:	[MyData]/Metabolomics1May2015_P
Relevant features file:	[MyData]/Metabolomics1May2015.DI

Region based omic

☐ My regions are already mapped to Gene IDs, skip this step.

Omic Name:	DNAse-Seq data
Regions file (BED + Quantification):	[MyData]/DNaseSeq_PaintomicsDat
Relevant regions file:	[MyData]/DNaseSeq_DE_Paintomics
Annotations file (GTF):	[inbuilt GTF files]/sorted_mmu.gtf
Distance (kb):	10
TSS region distance (bps):	200
Promoter region distance (bps):	1300
Overlapped gene area (%):	50
Overlapped region area (%):	50
GTF Tag for gene ID/name :	gene_id
Summarization method:	Mean

Report: ○ All regions

● Let me choose

Regions mapping at...

Upstream	Promoter	TSS	1st Exon	Introns	Gene body	Intr.	G.B.	Downstream
☐	☑	☑	☑	☐	☐			☐

Figure 6.1: The input data form for PaintOmics 3. (A) - Available species are listed in a selector input. Users can request the installation of new species using the available tools. **(B)** - New sections can be added to the form by choosing the omics type on the "Available omics" panel. Each section of the form is adapted to the selected omics type. For example, miRNA-based **(D)** and region-based omics **(C)** displays different settings for customising the execution of the secondary tools.

Technology	Total features	Significant features
mRNA-seq	12762	5864
miRNA-seq	4998	605
DNase-seq	10274	5101
Proteomics	1110	148
Metabolomics	61	41

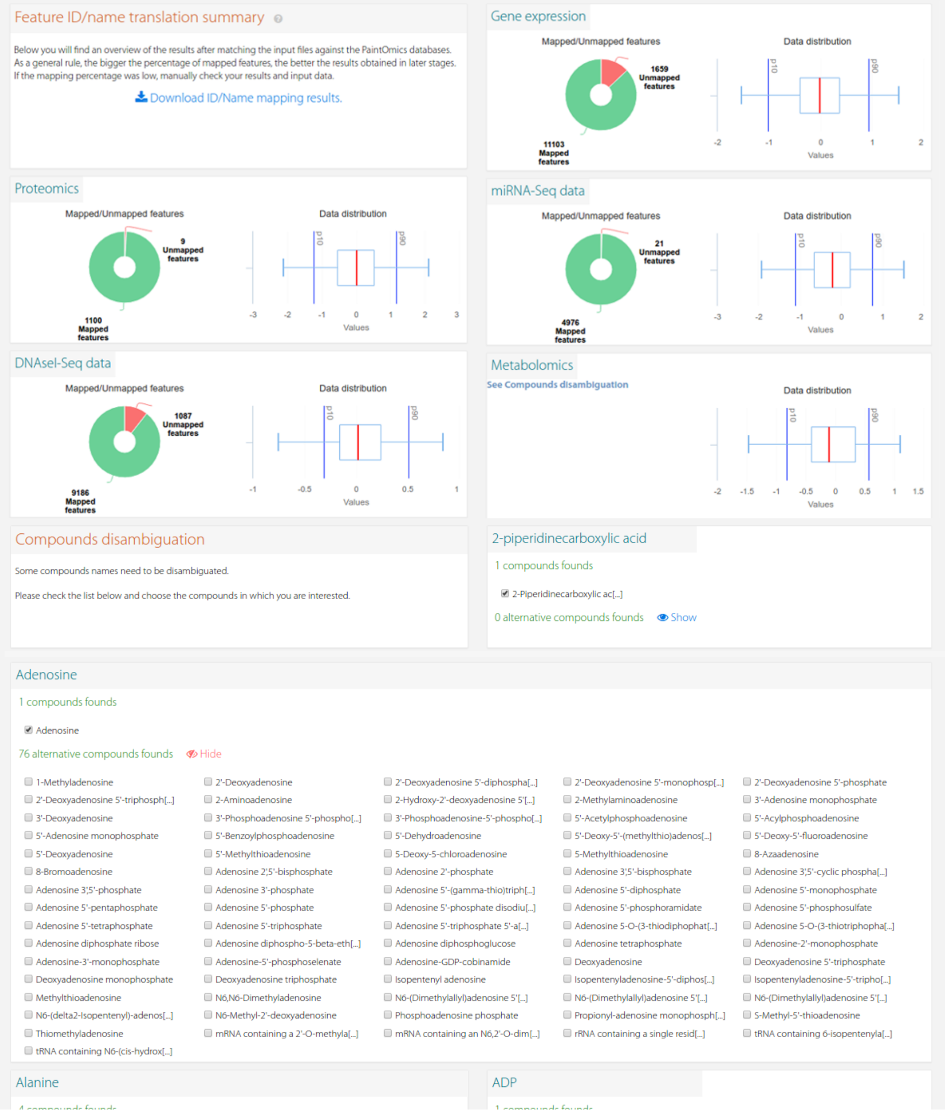
Table 6.2: Total features and significant features for each omics data type in the the STATegra data, after data processing.

6.3 Results

6.3.1 Identifiers and names conversion

After filling in the submission form, the process continues with the conversion of the input data names and identifiers into the accepted Kyoto Encyclopedia of Genes and Genomes (KEGG) name domains for the selected species, and a summary is presented of the number of features successfully found in the KEGG database and the dynamic range of the data.

Figure 6.2 shows the results of the mapping step for this use case. It can be observed that the faction of successfully mapped features is high for all omics and nearly complete for Proteomics and microRNA-seq. Completion for Proteomics data is explained by the fact that proteins detected by proteomics experiments, normally abundant proteins, tend also to be well annotated proteins. In the case of microRNAs, the high matching rate is probably related to the fact that microRNAs tend to target many different mRNAs. On the contrary, about 10% of the gene expression and DNase-seq features were not matched in KEGG. These might be non-coding genes or genes not yet placed in biological pathways. Finally, all metabolites had a match. Inspection of the disambiguation options did not revealed any metabolite matching that should be modified, and hence we can submit the job and continue to the next step.



Feature ID/name translation summary

Below you will find an overview of the results after matching the input files against the PaintOmics databases. As a general rule, the bigger the percentage of mapped features, the better the results obtained in later stages. If the mapping percentage was low, manually check your results and input data.

Download ID/Name mapping results.

Compounds disambiguation

Some compounds names need to be disambiguated.

Please check the list below and choose the compounds in which you are interested.

2-piperidinecarboxylic acid

1 compounds founds

☑ 2-Piperidinecarboxylic ac[...]

0 alternative compounds founds Show

Adenosine

1 compounds founds

☑ Adenosine

76 alternative compounds founds Hide

☐ 1-Methyladenosine
☐ 2'-Deoxyadenosine
☐ 2'-Deoxyadenosine 5'-diphospha[...]
☐ 2'-Deoxyadenosine 5'-monophosp[...]
☐ 2'-Deoxyadenosine 5'-phosphate
☐ 2'-Deoxyadenosine 5'-triphosph[...]
☐ 2-Aminoadenosine
☐ 2-Hydroxy-2'-deoxyadenosine 5'[...]
☐ 2-Methylaminoadenosine
☐ 3'-Adenosine monophosphate
☐ 3'-Deoxyadenosine
☐ 3'-Phosphoadenosine 5'-phospho[...]
☐ 3'-Phosphoadenosine-5'-phospho[...]
☐ 5'-Acetylphosphoadenosine
☐ 5'-Acylphosphoadenosine
☐ 5'-Adenosine monophosphate
☐ 5'-Benzoylphosphoadenosine
☐ 5'-Dehydroadenosine
☐ 5'-Deoxy-5'-(methylthio)adenos[...]
☐ 5'-Deoxy-5'-fluoroadenosine
☐ 5'-Deoxyadenosine
☐ 5'-Methylthioadenosine
☐ 5-Deoxy-5-chloroadenosine
☐ 5-Methylthioadenosine
☐ 8-Azaadenosine
☐ 8-Bromoadenosine
☐ Adenosine 2',5'-bisphosphate
☐ Adenosine 2'-phosphate
☐ Adenosine 3',5'-bisphosphate
☐ Adenosine 3',5'-cyclic phospha[...]
☐ Adenosine 3',5'-phosphate
☐ Adenosine 3'-phosphate
☐ Adenosine 5'-(gamma-thio)triph[...]
☐ Adenosine 5'-diphosphate
☐ Adenosine 5'-monophosphate
☐ Adenosine 5'-pentaphosphate
☐ Adenosine 5'-phosphate
☐ Adenosine 5'-phosphate disodiu[...]
☐ Adenosine 5'-phosphoramidate
☐ Adenosine 5'-phosphosulfate
☐ Adenosine 5'-tetraphosphate
☐ Adenosine 5'-triphosphate
☐ Adenosine 5'-triphosphate 5'-a[...]
☐ Adenosine 5-O-(3-thiodiphophat[...]
☐ Adenosine 5-O-(3-thiotriphopha[...]
☐ Adenosine diphosphate ribose
☐ Adenosine diphospho-5-beta-eth[...]
☐ Adenosine diphosphoglucose
☐ Adenosine tetraphosphate
☐ Adenosine-2'-monophosphate
☐ Adenosine-3'-monophosphate
☐ Adenosine-5'-phosphoselenate
☐ Adenosine-GDP-cobinamide
☐ Deoxyadenosine
☐ Deoxyadenosine 5'-triphosphate
☐ Deoxyadenosine monophosphate
☐ Deoxyadenosine triphosphate
☐ Isopentenyl adenosine
☐ Isopentenyladenosine-5'-diphos[...]
☐ Isopentenyladenosine-5'-tripho[...]
☐ Methylthioadenosine
☐ N6,N6-Dimethyladenosine
☐ N6-(Dimethylallyl)adenosine 5'[...]
☐ N6-(Dimethylallyl)adenosine 5'[...]
☐ N6-(Dimethylallyl)adenosine 5'[...]
☐ N6-(delta2-Isopentenyl)-adenos[...]
☐ N6-Methyl-2'-deoxyadenosine
☐ Phosphoadenosine phosphate
☐ Propionyl-adenosine monophosph[...]
☐ S-Methyl-5'-thioadenosine
☐ Thiomethyladenosine
☐ mRNA containing a 2'-O-methyla[...]
☐ mRNA containing an N6,2'-O-dim[...]
☐ rRNA containing a single resid[...]
☐ tRNA containing 6-isopentenyla[...]
☐ tRNA containing N6-(cis-hydrox[...]

Alanine

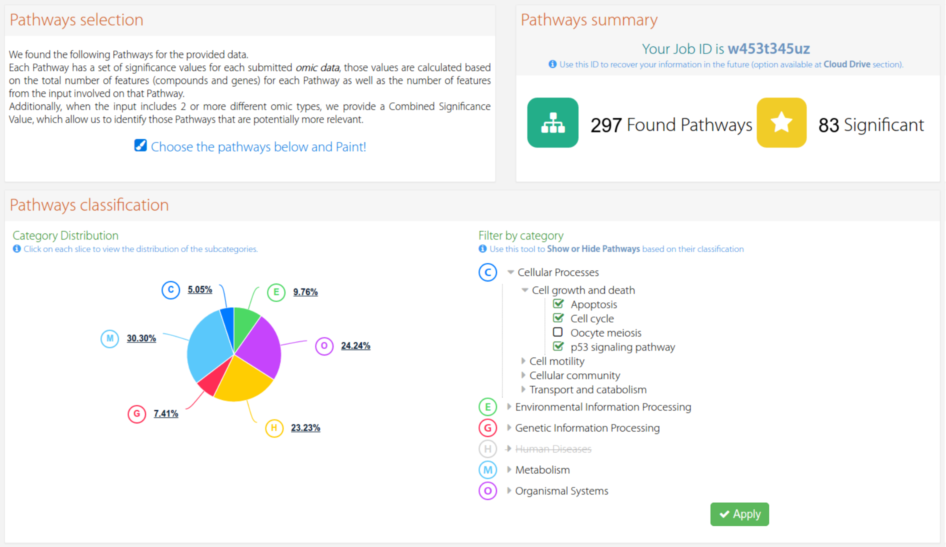
ADP

Figure 6.2: Results for the mapping step for the STATegra data. (A) - The first part of the panel displays the total translated features for each omics data type as a pie chart. Box-plots are used to provide an overview for the distribution of the quantification values. Moving the mouse cursor over the box-plots provides more details about the data distribution. **(B)** - The second part of the panel is devoted to metabolite disambiguation. The user can choose which metabolites should be used for subsequent analysis stages. When the mouse is moved over each metabolite name more details are provided for the metabolite along with its structural diagram. Clicking on the metabolite name causes a new window to open which displays the complete description for the metabolite on the KEGG website.

6.3.2 Obtaining pathways

After feature mapping, PaintOmics 3 obtains the list of pathways that contain genes or metabolites from the input and evaluates their significance level by applying an enrichment analysis. As shown in Figure 6.3, the application reports 297 pathways, of which, 83 are considered to be enriched (combined p-value ≤ 0.05).

Using the tools in the "Pathways classification" section, the displayed pathways can be individually hidden or shown or they can be customised by category or subcategory. Based on the consortium experts' knowledge, several pathways were considered "not relevant" to the study and thus, were filtered out before proceeding with the evaluation of results. For example, "Human Diseases" KEGG category does not contain any pathways directly related to the studied organism (mouse)and therefore the whole category can be excluded from downstream analysis. Appendix A contains the complete list of the 116 pathways which the consortium experts considered to be closely related or interesting for the system under study (SuS).



Pathways selection

We found the following Pathways for the provided data.
Each Pathway has a set of significance values for each submitted *omic data*, those values are calculated based on the total number of features (compounds and genes) for each Pathway as well as the number of features from the input involved on that Pathway.
Additionally, when the input includes 2 or more different omic types, we provide a Combined Significance Value, which allow us to identify those Pathways that are potentially more relevant.

Choose the pathways below and Paint!

Pathways summary

Your Job ID is **w453t345uz**

Use this ID to recover your information in the future (option available at **Cloud Drive** section).

297 Found Pathways

83 Significant

Pathways classification

Category Distribution

Click on each slice to view the distribution of the subcategories.

Filter by category

Use this tool to **Show or Hide Pathways** based on their classification

- C Cellular Processes
  - Cell growth and death
    - ☑ Apoptosis
    - ☑ Cell cycle
    - ☐ Oocyte meiosis
    - ☑ p53 signaling pathway
  - Cell motility
  - Cellular community
  - Transport and catabolism
- E Environmental Information Processing
- G Genetic Information Processing
- H ~~Human Diseases~~
- M Metabolism
- O Organismal Systems

Figure 6.3: Results for the PaintOmics pathway enrichment analysis for the STATegra data. A total of 297 pathways were reported, 83 of them were considered to be significantly enriched. Using the filtering tools, we can exclude from the results pathways that are not interesting for our specific study.

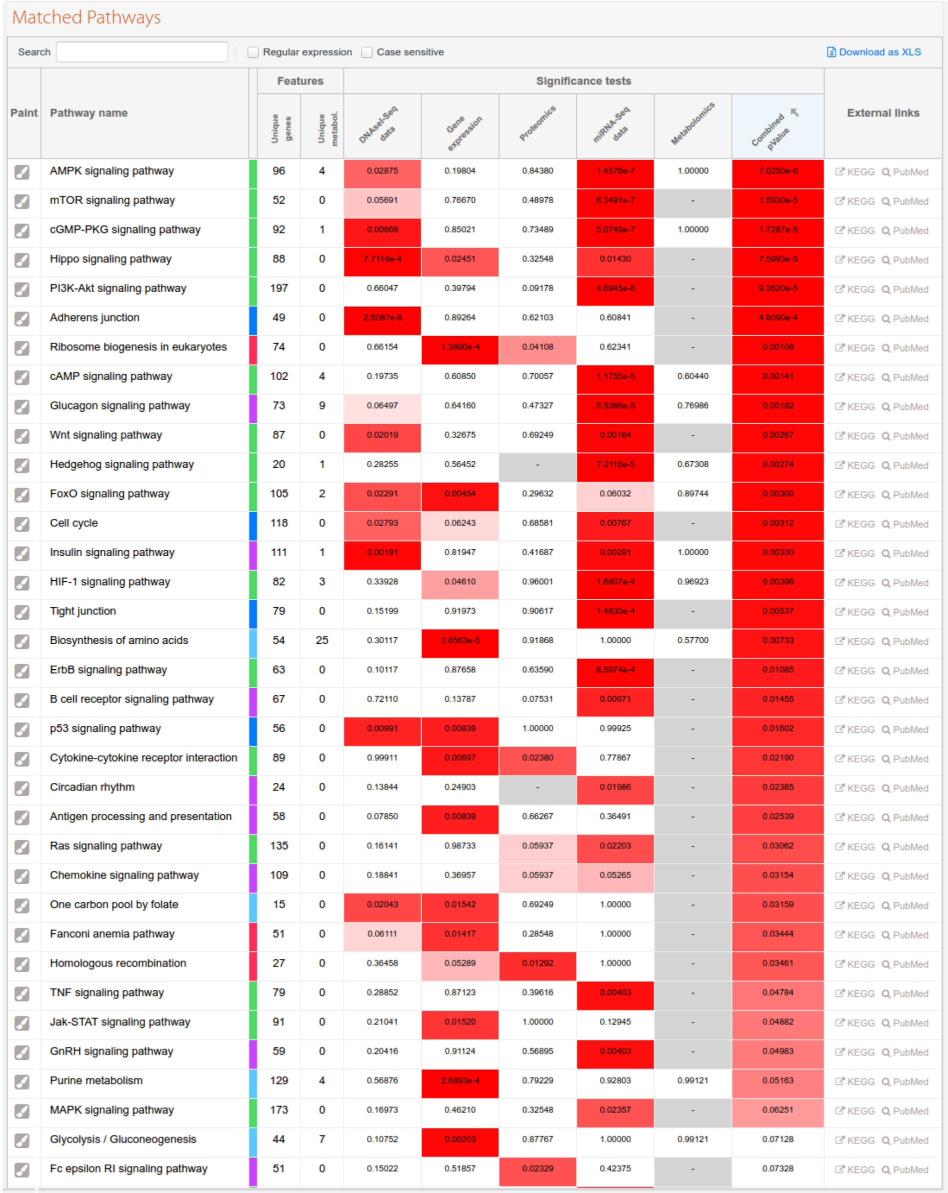
6.4 Discussion

Due to the fact that the main framework of this thesis is computer engineering, the following section intends to be an exemplification of the what can be elucidate from a biological potin of view by using Paintomics 3. Further and more complexes analysis can be found in Tarazona et al. [128].

6.4.1 Significant pathways

After setting up the set of pathways of interest for the study, we can proceed to evaluate the enrichment analysis results. By default, the table lists all the reported pathways sorted by their combined *p*-value, however, the user can reorder or hide columns and add filters to the table content. The upper positions correspond to the most significant pathways (i.e. lower *p*-values).

Paintomics3 returned a total of 31 significant pathways for the B3 cell differentiation course (p-value ≤ 0.05). 27 pathways were significant at the Gene Expression level, 5 for Proteomics, 23 for microRNAs, 15 for DNase-seq and none for metabolomics (Figure 6.4). Interestingly there were not many pathways that were significant across different omics. On the contrary, it appears that each omics layer will reveal as significant a different subset of processes. Gene expression was significant for carbon and amino-acid metabolic pathways and for a number of signaling pathways such as FoxO signaling, Jak-STAT, Notch, p53, HIF-1 signaling. FoxO and Jak-STAT have been extensively reported the transition from proliferating B cell progenitors towards quiescence and differentiation [35]). On the contrary, miRNA-seq returns almost exclusively signally pathways as significant, including AMPK signaling pathway, HIF-1 signaling pathway, or B cell receptor signaling pathway, which may suggest a role of microRNAs in this system to fine-tune the control of the molecular signal processing. The DNase-seq data indicates significant chromatin accessibility changes for genes in relevant signaling pathways (FoxO, p53, T-cell receptor, Wnt, etc.) pathways related to cell division (cell cycle, base excision repair, p53 signaling) and metabolism (Lysine degradation, One carbon pool by folate, etc.), which mimics the transcriptional changes observed at the transcriptomics data. Finally, a low number of pathways are significant either for proteomics or metabolomics, which might be explained by the more reduced number of features in these datasets, the lower coverage of the pathway space and possibly a higher nose level in the measurements.



Matched Pathways

Search □ Regular expression □ Case sensitive

Paint	Pathway name	Features		Significance tests						External links
		Unique genes	Unique metabol.	DNAseI-Seq data	Gene expression	Proteomics	miRNA-Seq data	Metabolomics	Combined pValue	
	AMPK signaling pathway	96	4	0.02875	0.19804	0.84380	1.4576e-7	1.00000	7.0250e-6	KEGG PubMed
	mTOR signaling pathway	52	0	0.05691	0.76670	0.48978	6.3491e-7	-	1.5930e-5	KEGG PubMed
	cGMP-PKG signaling pathway	92	1	0.00666	0.85021	0.73489	5.0749e-7	1.00000	1.7287e-5	KEGG PubMed
	Hippo signaling pathway	88	0	7.7116e-4	0.02451	0.32548	0.01430	-	7.5993e-5	KEGG PubMed
	PI3K-Akt signaling pathway	197	0	0.66047	0.39794	0.09178	4.6945e-6	-	9.3620e-5	KEGG PubMed
	Adherens junction	49	0	2.5087e-6	0.89264	0.62103	0.60841	-	4.8090e-4	KEGG PubMed
	Ribosome biogenesis in eukaryotes	74	0	0.66154	1.3890e-4	0.04108	0.62341	-	0.00108	KEGG PubMed
	cAMP signaling pathway	102	4	0.19735	0.60850	0.70057	1.1755e-5	0.60440	0.00141	KEGG PubMed
	Glucagon signaling pathway	73	9	0.06497	0.64160	0.47327	5.5386e-5	0.76986	0.00182	KEGG PubMed
	Wnt signaling pathway	87	0	0.02019	0.32675	0.69249	0.00164	-	0.00267	KEGG PubMed
	Hedgehog signaling pathway	20	1	0.28255	0.56452	-	7.2116e-5	0.67308	0.00274	KEGG PubMed
	FoxO signaling pathway	105	2	0.02291	0.00454	0.29632	0.06032	0.89744	0.00300	KEGG PubMed
	Cell cycle	118	0	0.02793	0.06243	0.68581	0.00767	-	0.00312	KEGG PubMed
	Insulin signaling pathway	111	1	0.00191	0.81947	0.41687	0.00291	1.00000	0.00330	KEGG PubMed
	HIF-1 signaling pathway	82	3	0.33928	0.04610	0.96001	1.6807e-4	0.96923	0.00396	KEGG PubMed
	Tight junction	79	0	0.15199	0.91973	0.90617	1.4835e-4	-	0.00537	KEGG PubMed
	Biosynthesis of amino acids	54	25	0.30117	3.6563e-5	0.91868	1.00000	0.57700	0.00733	KEGG PubMed
	ErbB signaling pathway	63	0	0.10117	0.87658	0.63590	8.5974e-4	-	0.01085	KEGG PubMed
	B cell receptor signaling pathway	67	0	0.72110	0.13787	0.07531	0.00971	-	0.01455	KEGG PubMed
	p53 signaling pathway	56	0	0.00991	0.00839	1.00000	0.99925	-	0.01602	KEGG PubMed
	Cytokine-cytokine receptor interaction	89	0	0.99911	0.00697	0.02380	0.77867	-	0.02190	KEGG PubMed
	Circadian rhythm	24	0	0.13844	0.24903	-	0.01986	-	0.02385	KEGG PubMed
	Antigen processing and presentation	58	0	0.07850	0.00839	0.66267	0.36491	-	0.02539	KEGG PubMed
	Ras signaling pathway	135	0	0.16141	0.98733	0.05937	0.02203	-	0.03062	KEGG PubMed
	Chemokine signaling pathway	109	0	0.18841	0.36957	0.05937	0.05265	-	0.03154	KEGG PubMed
	One carbon pool by folate	15	0	0.02043	0.01542	0.69249	1.00000	-	0.03159	KEGG PubMed
	Fanconi anemia pathway	51	0	0.06111	0.01417	0.28548	1.00000	-	0.03444	KEGG PubMed
	Homologous recombination	27	0	0.36458	0.05289	0.01292	1.00000	-	0.03461	KEGG PubMed
	TNF signaling pathway	79	0	0.28852	0.87123	0.39616	0.00403	-	0.04784	KEGG PubMed
	Jak-STAT signaling pathway	91	0	0.21041	0.01520	1.00000	0.12945	-	0.04882	KEGG PubMed
	GnRH signaling pathway	59	0	0.20416	0.91124	0.56895	0.00403	-	0.04983	KEGG PubMed
	Purine metabolism	129	4	0.56876	2.6893e-4	0.79229	0.92803	0.99121	0.05163	KEGG PubMed
	MAPK signaling pathway	173	0	0.16973	0.46210	0.32548	0.02357	-	0.06251	KEGG PubMed
	Glycolysis / Gluconeogenesis	44	7	0.10752	0.00203	0.87767	1.00000	0.99121	0.07128	KEGG PubMed
	Fc epsilon RI signaling pathway	51	0	0.15022	0.51857	0.02329	0.42375	-	0.07328	KEGG PubMed

Figure 6.4: Enriched pathways for the STATegra data ordered by combined p-value. Upper positions correspond to the most significant pathways. A color scale is used to highlight the level of enrichment for each pathway where the higher intensity of red, the higher significance is it. Gray cells indicate that the corresponding omics is not present on the pathway.

6.4.2 Pathway network

The pathway network shows the underlying interactions among significant pathways what aids in the interpretation of the experiment in terms of global functional and regulatory relationships at the cellular level. Figures 6.5 - 6.9 displays the network using the different color strategies available on the system. The spatial distribution of the nodes is calculated based on the number of shared genes between pathways and is therefore independent of the omics data. The "a priori" pathway network is colored based on the main pathway categories in the KEGG database such as "Metabolism", "Information Processing", etc. This representation choice creates a uniform network template that is then modified in color and of number of displayed pathway as a function of the represented omics dataset. In this way we facilitate the interpretation of the impact each omics layer has on the underlying pathway network. Figure 6.6 displays the network after imposing the gene expression data. The color and number of displayed pathways change. To interpret the network we need to first turn our attention to the *metagene* patterns at the lower right corner. The pathway metagene analysis indices two clusters of pathways, one of ascent and one of descent, both showing a sharp change at the middle of the series that seems to correspond to the change from proliferation to quiescence state. By matching the color of the metagene with the color pathways displayed for the gene expression data we can appreciate which pathways are either a up or down regulated. We can conclude that in our experiment there is a global downregulation of "Metabolism" and "Genetic Information Processing" (cell cycle, DNA repair, transcription, splicing, etc.) representing an stop in the cellular division process, that occurs fast at 12 hours and slows down at 18 hours (cluster #1). On the other hand, the KEGG Environmental Signal Processing category, i.e, the signaling pathways, are up-regulated suggesting that the differentiation processes is the result of a generalize activation of the whole cellular signaling machinery. This global activation is the one that may control all the changes that can be observed: the cellular cycle arrests, a specific nuclear rearrangement occurs and the metabolism stops. We can highlight the great synchronization that takes place in this process.

The pathway network analysis of miRNAs gives also interesting results. We can observe that a great majority of significant pathways according to the microRNA data have a pattern of down regulation at late time points that specially impacts signaling pathways, suggesting that the activation of signaling in the differentiation course is coupled by a released of microRNA repression of the genes present in these processes.

Interestingly, the DNase-seq data, also clusters pathways in 2 major and symmetric trends. Cluster 1 represents Metabolic and Genetic Information processing (the same as Gene Expression) and is characterized by a more-open to more-close chromatin accessibility at 6 h and 12 h followed by stabilization of the signal. Cluster 2 has the opposite pattern and represents (as in RNA-seq) mostly Environmental Signal Processing pathways. This pattern of activity could be interpreted as the changes in chromatin conformation that are required previous to the changes in gene expression observed in these pathway groups.

Finally, we can appreciate in the pathway network analysis the noisiness of the proteomics data. The metagene clusters of proteomics data reveal a quite erratic behavior of protein signals at the beginning of the experiment that has as a result a quite "flat" mean profile at these time points. This is followed by strong and opposite expression changes at 12 and 18 hours to conclude with a stabilization of the expression profiles as down (Cluster 1) or upregulated (Cluster 2) proteins. Most metabolic pathways are in Cluster 1 (ending in protein down-regulation) what is in agreement with the gene expression data, while signaling pathways are both in Cluster 1 and 2, which is in less agreement with the gene expression. This more noisy behavior fits the overall observation of STATegra researchers of lower signal to noise ratio and replicability in the proteomics data compared to the RNA-seq.

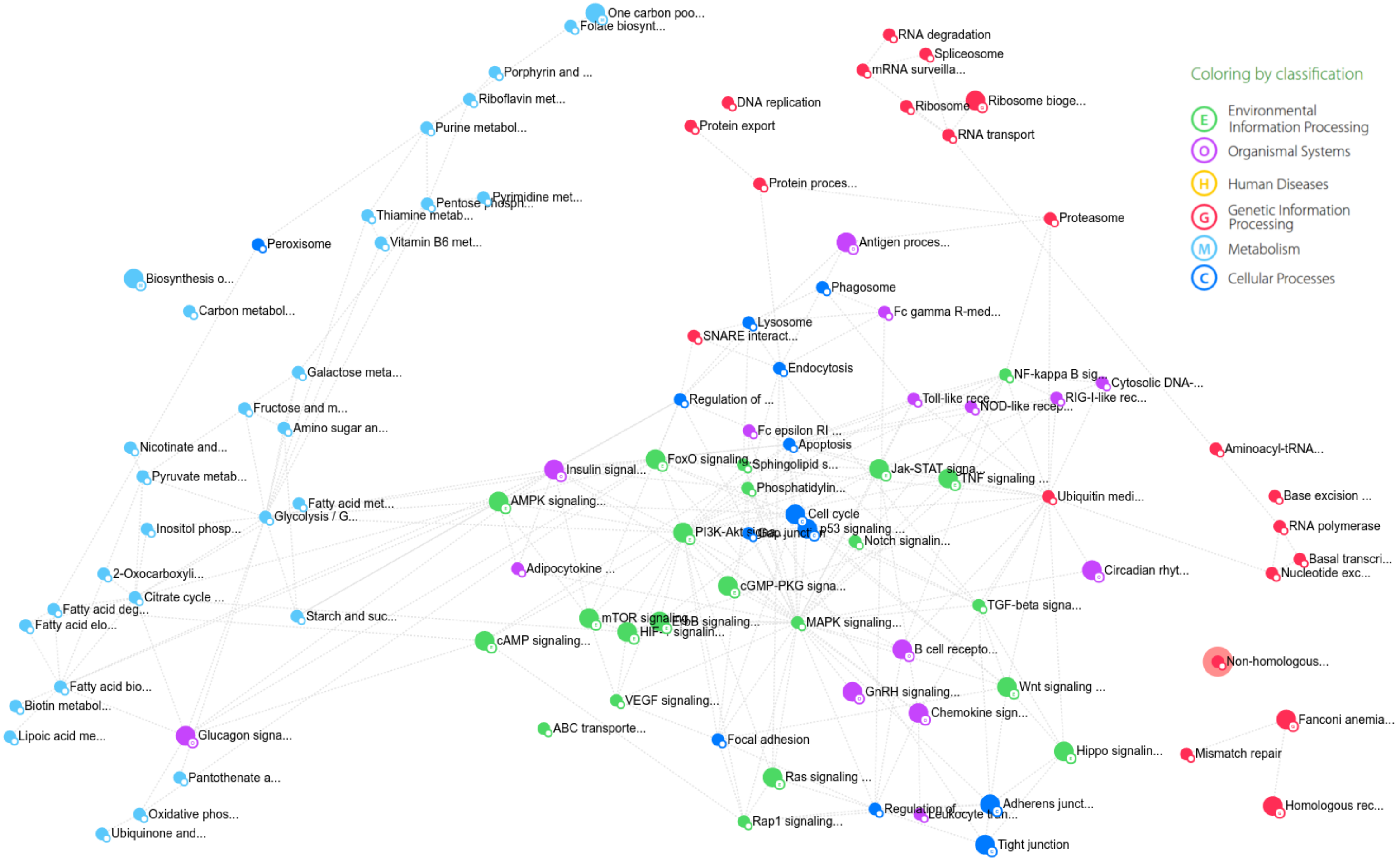


Figure 6.5: Pathway network coloured by pathway classification.

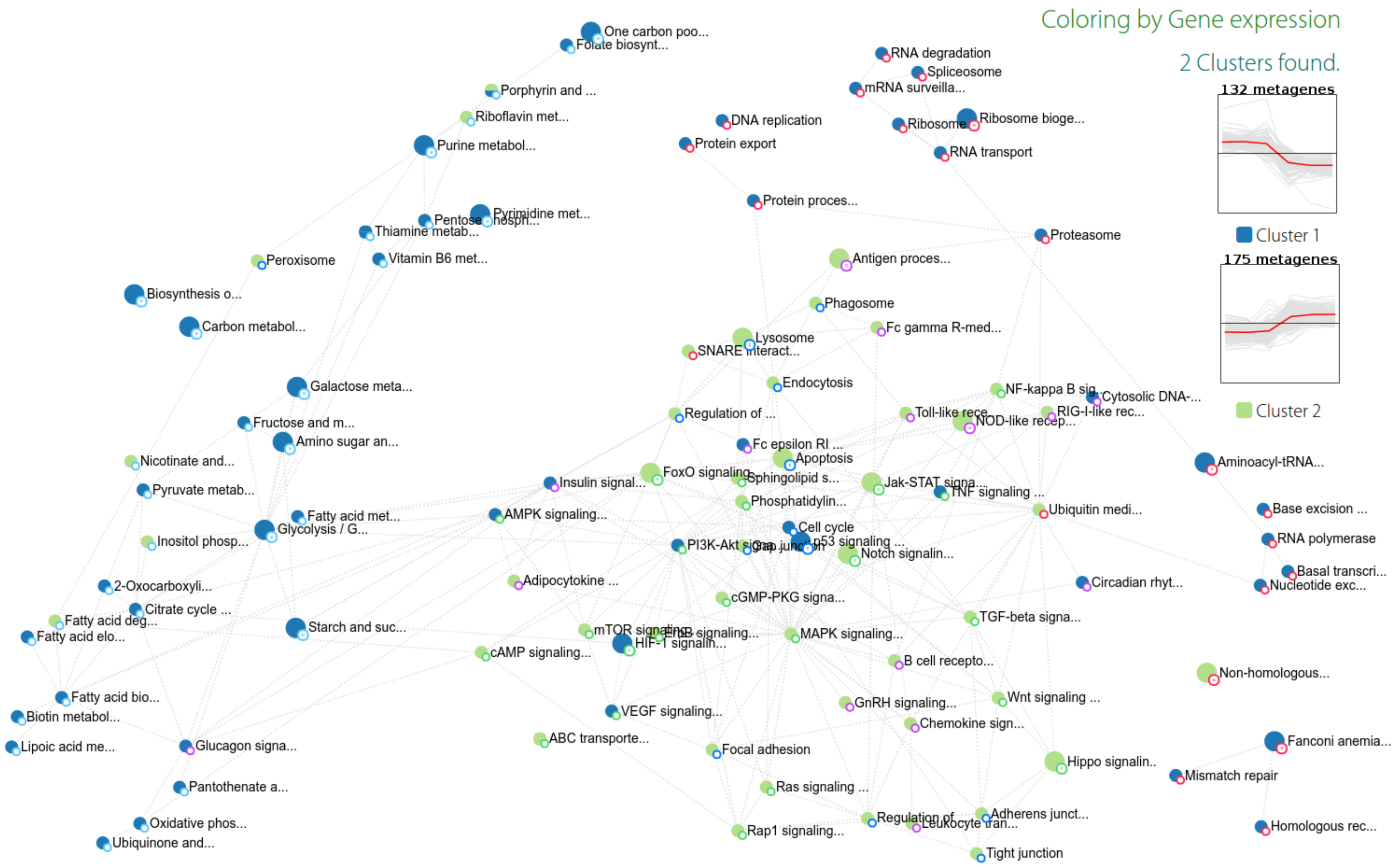


Figure 6.6: Pathway network coloured by transcriptomics clusters.

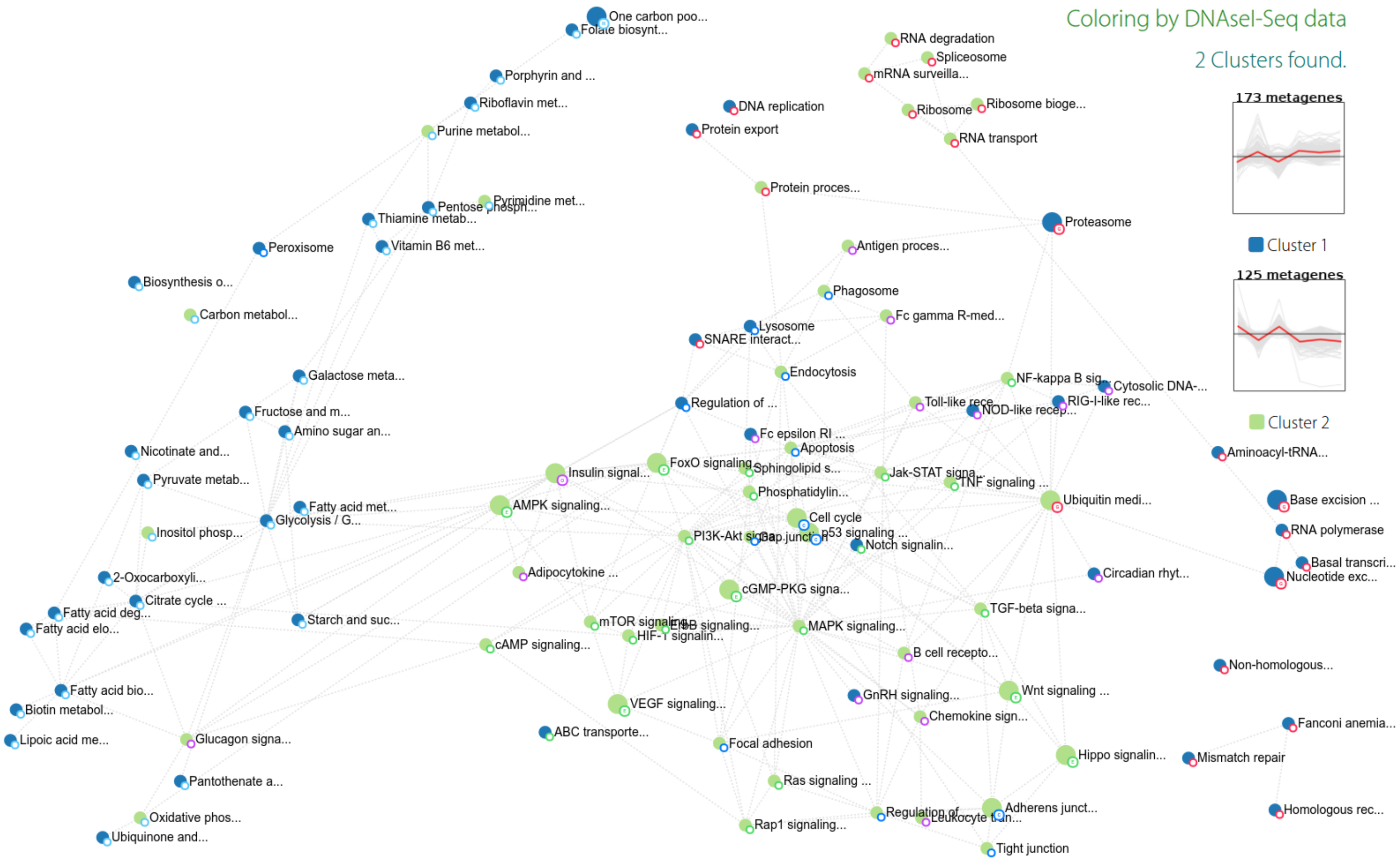


Figure 6.7: Pathway network coloured by DNase-seq clusters.

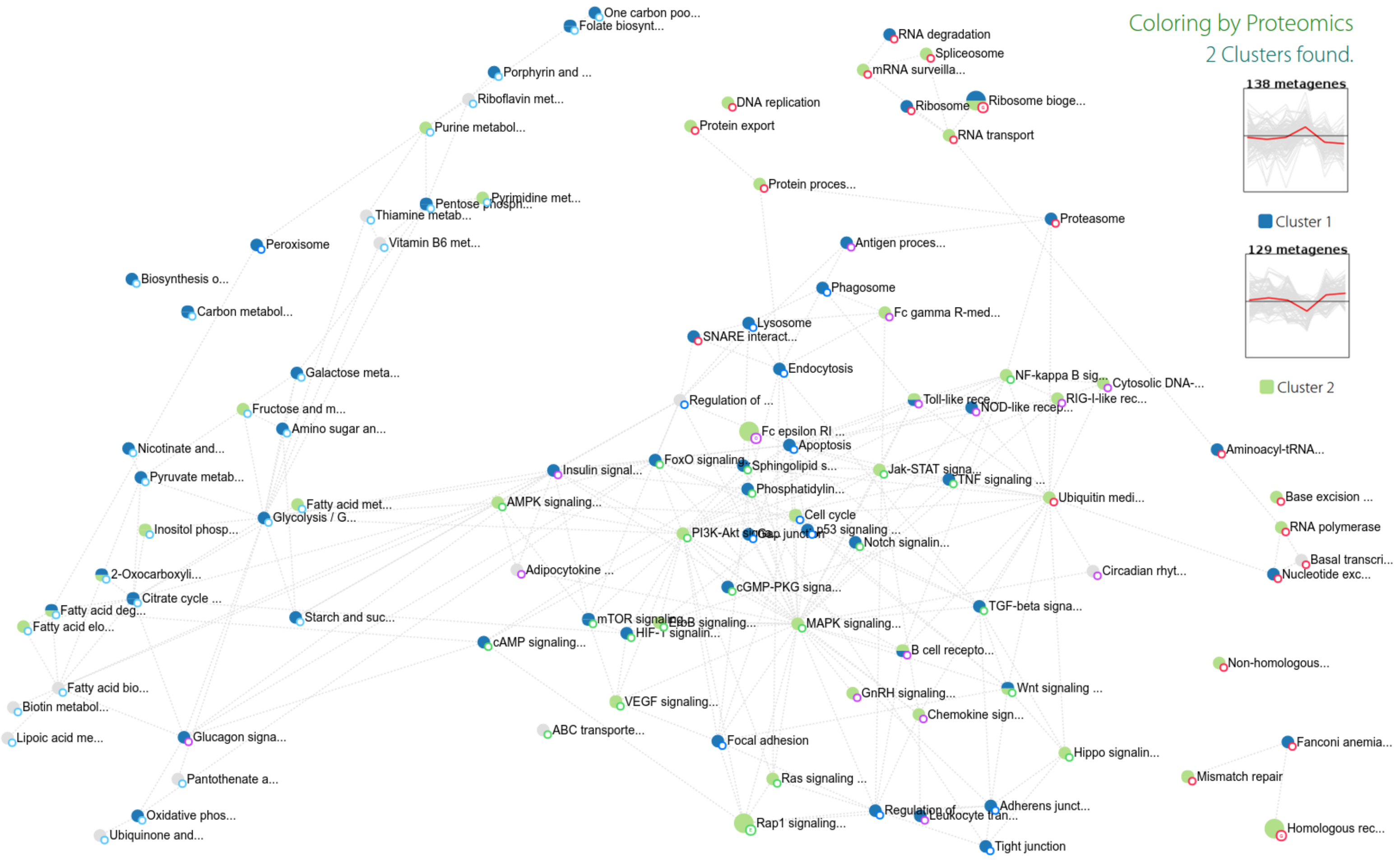


Figure 6.8: Pathway network coloured by Proteomics clusters.

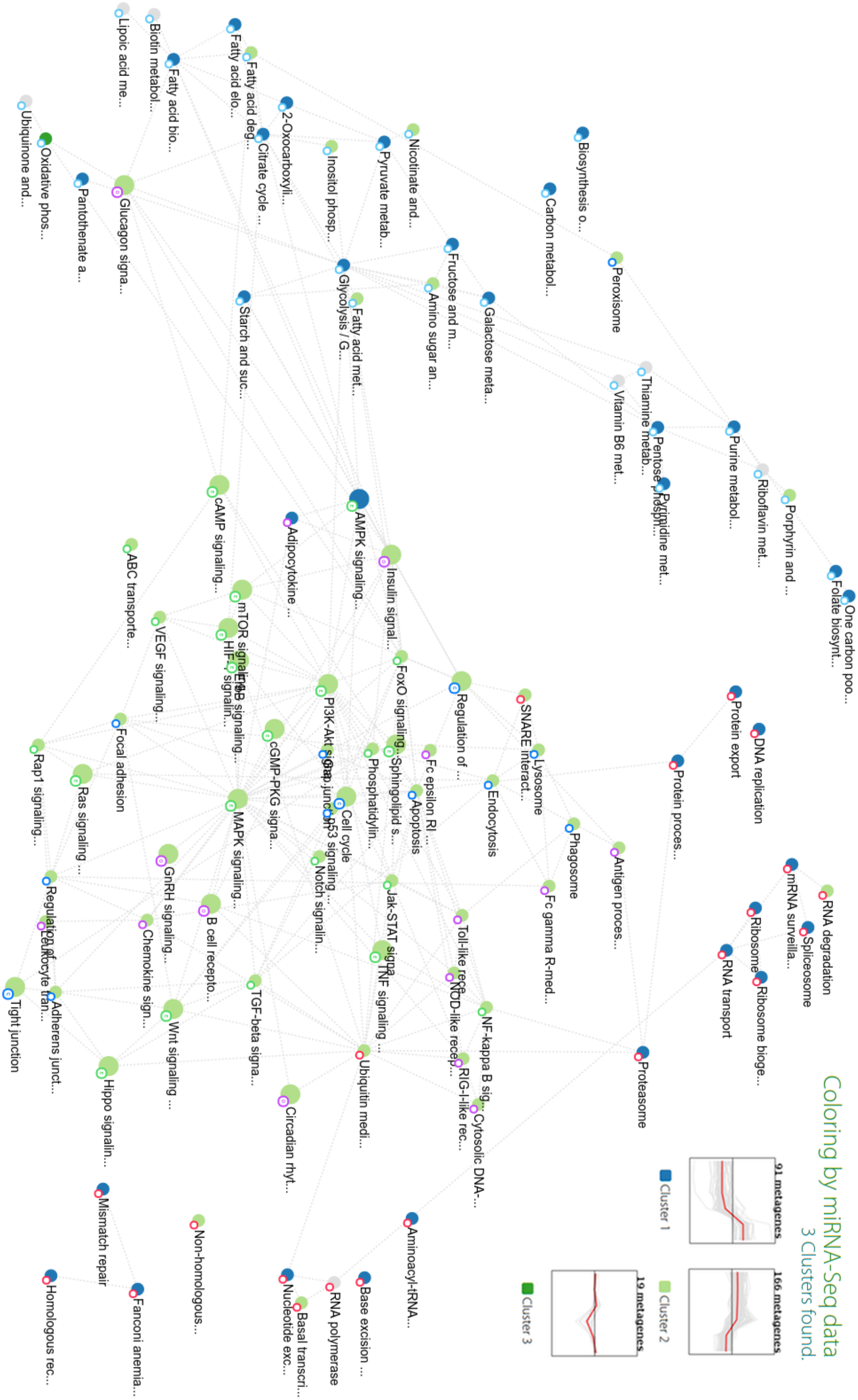


Figure 6.9: Pathway network coloured by miRNA-seq clusters.

6.4.3 Pathway analysis

To date, there numerous publications that have studied the processes of cell differentiation. However, little is known about the dynamics that occur within such processes, including changes in gene expression, proliferative state, biosynthetic capacity, metabolic state and gene regulatory networks. In the system under study we focus on the dynamics and the regulatory control of a critical step in B cell progenitor differentiation, namely the progression of proliferating B cell progenitors towards resting pre-B cells. Following a phase of expansion, cell cycle exit is critical, to avoid the accumulation of genetic lesions that could otherwise result in several diseases. More specifically, when talking about the regulatory control of B cell progenitor differentiation, the transition from a proliferative state to quiescence may inhibit leukemic transformation, and enables immunoglobulin light chain rearrangement, the expression of a functional BCR, and B cell differentiation [24, 29, 51]. For that reason, the stimulus-driven transition from quiescence to an activated state is central to a functional immune system and has been studied extensively. One of the genes that participate in the transition from proliferation to quiescence during haematopoiesis is FOXO1, a direct target of IKAROS [134] with opposite function to MYC [134].

If we want to focus on the different omics processes related to the regulation of FOXO1, we could explore the FoxO signaling pathway individually, visualizing the input data over the KEGG diagram. As commented in previous chapters, the first time that we open a pathway, PaintOmics shows two panels:

- The interactive KEGG diagram, where each feature (significant or not) is displayed as a box divided in several heatmaps, colored according to the concentration values at each particular sample.
- An auxiliary panel showing some interesting information and a "Search" tool.

Within the information displayed at the auxiliary panel, we can find line charts representing the most significant trends for each omics data type (i.e. the metagenes) for current pathway. As shown in Figures 6.10 and 6.11, in FOXO signaling pathway, in a general form, we can observe an increase of gene expression accompanied by a decrease in miRNA expression. More specifically, an up-regulation of several genes associated with cell growth and apoptosis – such as SGK1, STAt3 and AKT1 – that appears early in the process induces an over expression of FOXO1. The progressive transcriptional up-regulation of FOXO1 promotes the establishment of the quiescence state. This fact is supported by the up-regulation of genes like RAG1/2 which are known to be a key step of B-Cell differentiation.

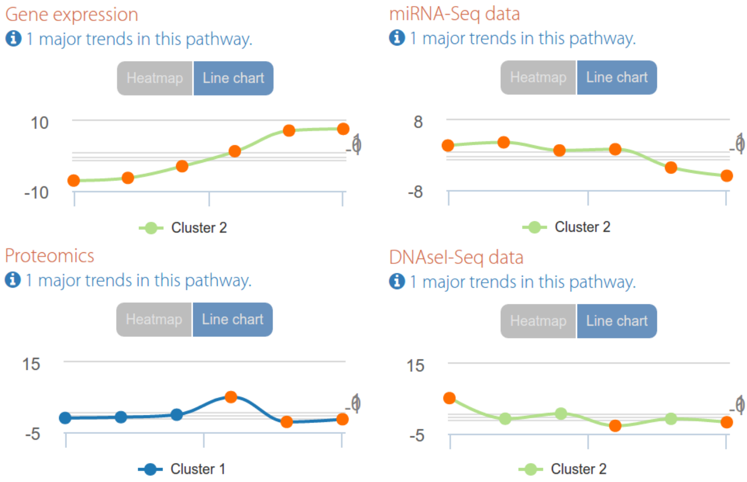


Figure 6.10: Most significant trends for each omics in FoxO signaling pathway.

The overexpression of FOXO1 also induces changes in other metabolic and cellular processes. For example, it is well known that a reverse "metabolic reprogramming" is needed for the activation of quiescent lymphocytes [100, 150] and, as it can be observed in the graph, genes that inhibit glucose metabolism such as PCK2 and G6pc3 are down-regulated. This fact is better understood if we overview the process from a global sight. When cells are proliferating, great bioenergetics consumption are needed and, in order to maintain a minimal content of adenosine triphosphate (ATP) to satisfy energy requirements, a change in cellular pathways is required. On the contrary, after the transition from "cycling" stage to "resting", the high metabolic demands are reduced and only basal metabolic activity is required

At this point, the Glycolysis pathway serves as a prominent example to continue with this analysis validation and we will focus our next steps on the study of this biological process.

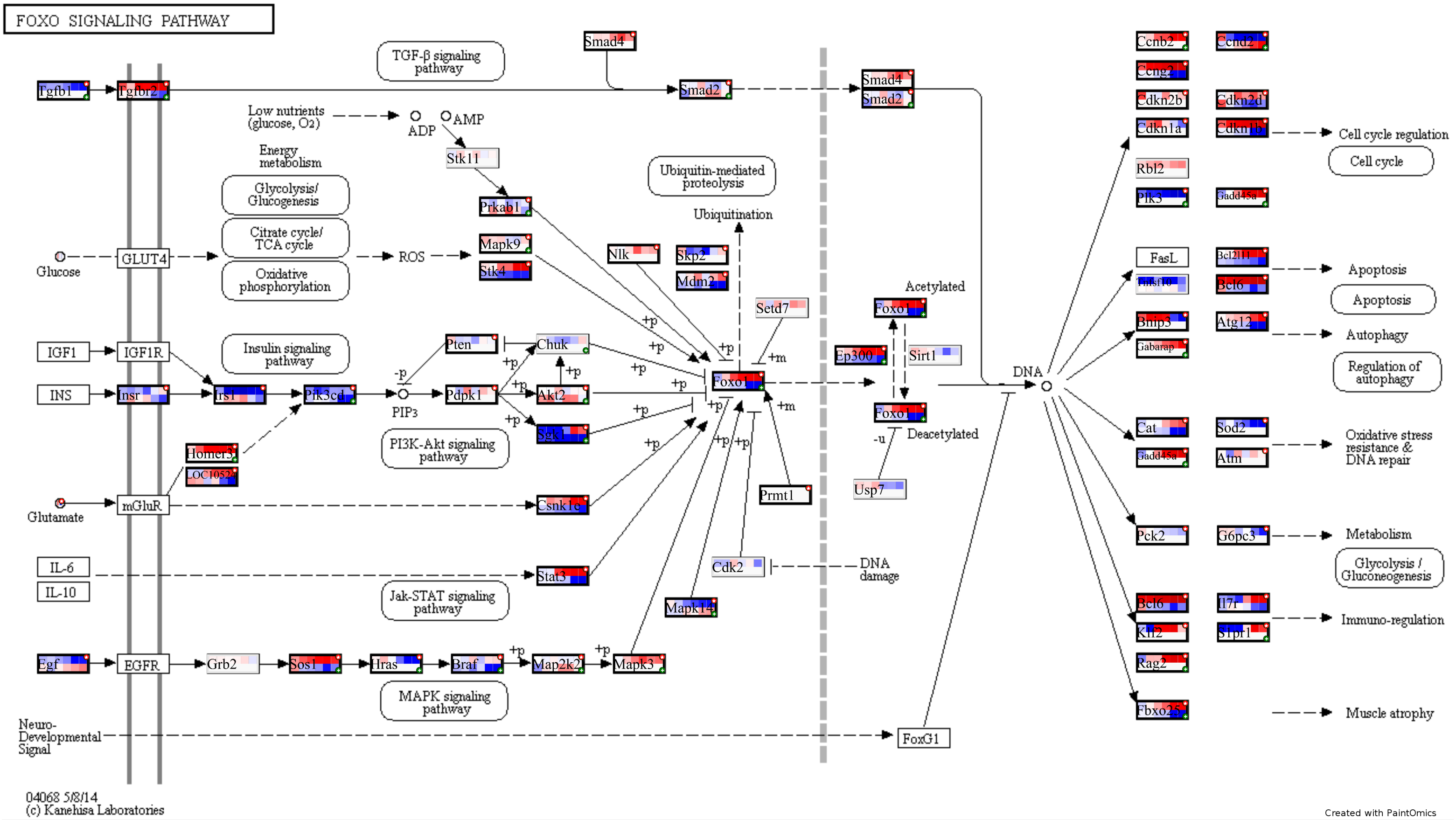


Figure 6.11: The interactive KEGG diagram for FoxO signaling pathway displaying transcriptomics and miRNA-seq expression.

As a rule, glucose works as a major source of cellular energy and new cell mass. Glucose is metabolized via Glycolysis to pyruvate, which can be oxidatively metabolized into the mitochondria producing large amounts of ATP through the process of oxidative phosphorylation. Alternatively, when limited amounts of oxygen are available, glucose can be transformed to lactate (anaerobic glycolysis), usually carried out through fermentation. Glucose fermentation generates ATP with far lower efficiency than oxidative phosphorylation but at a faster rate. Nevertheless, the con-version of glucose to lactate have been observed as a common phenomenon among proliferating animal cells, even in the presence of sufficient oxygen to support oxidative phosphorylation, suggesting that this enhanced rate of ATP generation may be beneficial for rapidly proliferating cells. In addition, some authors suggest that glucose degradation is preferable because produces several intermediate chemical constituents (e.g. nucleotides, amino acids, and lipids) needed to support biosynthesis.

As shown in Figure 6.12, metagenes for Glycolysis support the previous observations: genes involved on Glycolysis undergo a down-regulation at transcriptional level and chromatin accessibility when the transition from proliferation to quiescence takes part. Interestingly, the pattern profile for DNase-seq shows that chromatin conformation changes seems to occur earlier in time than gene expression changes. As, previously know, proteomics data does not follow the same clear trend, and although protein levels in the pathway seems to be generally lower at 24 h than at 0h, the time course profile looks more variable and erratic.

Figure 6.13 describes in more detail the dynamic changes for the significant genes that participate at Glycolysis. If we have a look to Hk2 and Ldha genes, a general agreement between genes - proteins - metabolites and microRNAs can be seen. In both cases, a correlation between the genes' expression and changes in chromatin in the TSS, given by DNase-seq, is clearly visible. In addition, an upregulation of microRNAs that targets the genes takes place and may represent a mechanism of controlling the genes' expression. On the opposite, as was already known, the metabolic data are noisier, but clearly shows how the value of both Hk2 and Ldha proteins is reduced at 24h, corresponding with the halting in the glycolysis at the end of the differentiation time course.

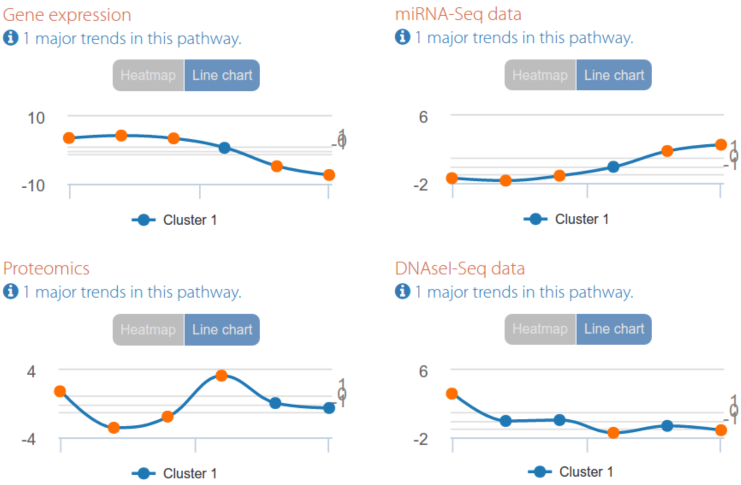


Figure 6.12: Most significant trends for each omics in Glycolysis / Gluconeogenesis pathway.

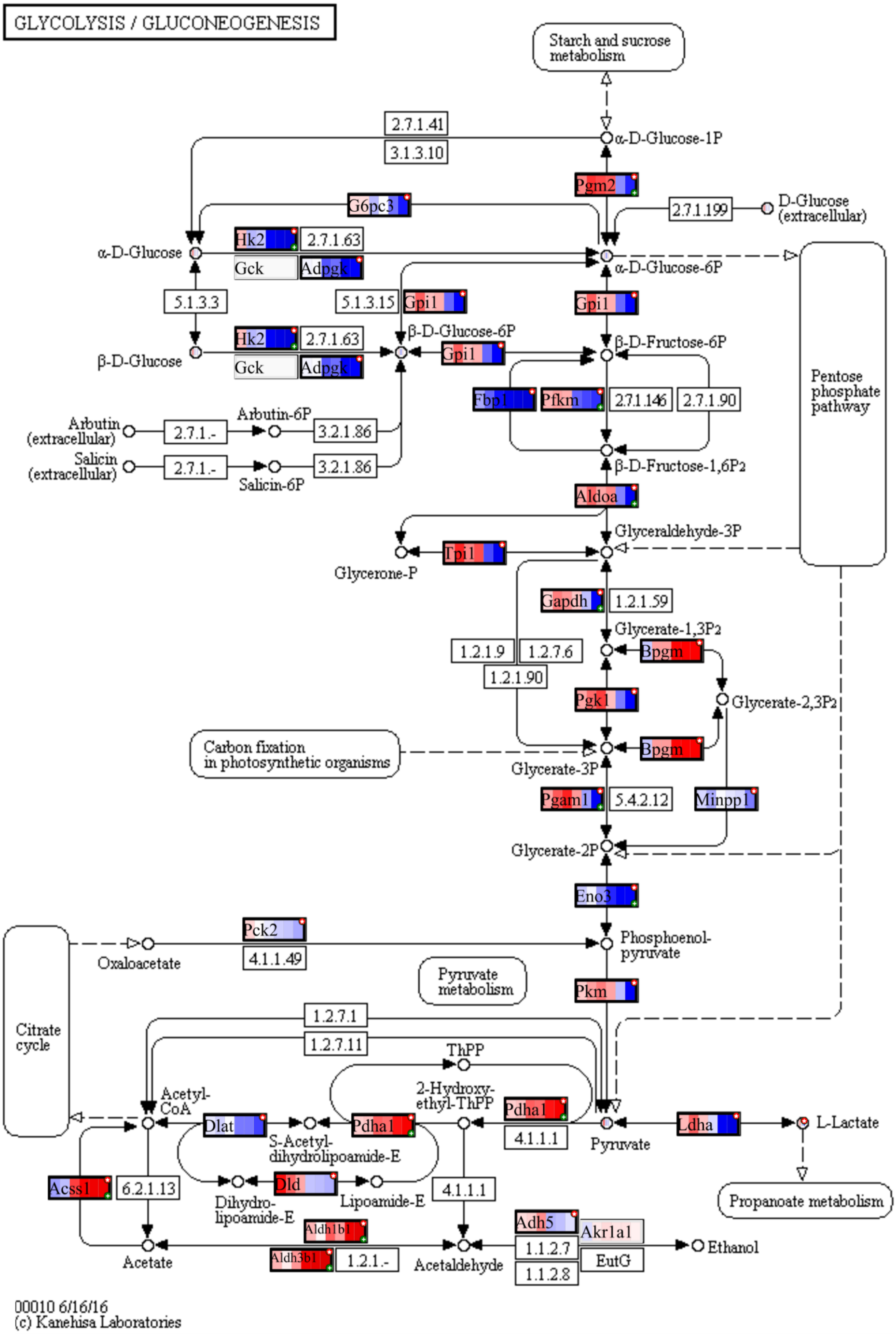


Figure 6.13: The interactive KEGG diagram for Glycolysis / Gluconeogenesis pathway displaying transcriptomics expression.

6.5 Conclusions

We can conclude that PaintOmics 3 can be used in the context of real biomedical research for visualising and analysing data from different omics. It provides us with a very useful and complete tool that shows us different points of view of the same process. On the one hand, when we rely on the part of the network, we can perceive the different general patterns that occur in each metabolic process. However, the possibility of having a focus on each independent pathway, constitutes a very useful strategy in biomedical research as it allow us the ability of visualizing the small details that may be biologically significant in each different pathway.

Considering all these observations, we can assume that, after Ikaros induction, pre-B3 cells undergo different processes which may induce the transition of cycling pre-B cells to the quiescence stage. This change is transcriptionally driven, as implies the activation of major signaling pathways by chromatin accessibility changes and suppression of microRNA downregulation of target genes. On the other hand the process implies a metabolic shut down, observed both at the gene expression and at the metabololic level, as well as the arrest of genetic information processing pathways such as cell cycle and DNA repair that also accordingly modify chromatin accessibility. More specifically, the progressive transcriptional up-regulation of FOXO1 supported by the up-regulation of genes like RAG1/2 promotes the establishment of the quiescence state. However a well-orchestrated reprogramming of metabolism is required and changes in the Glycolysis pathway are mandatory for the transition. Both figures 6.11 and 6.13, as well as the detailed study of the Glycolysis process, point that this process of cell differentiation may be primarily driven by transcriptional regulation. This fact was also observed during other analysis of the study and motivated one of the global strategies adopted in the project: the usage of RNA-seq as the central data type in the study.

7

General discussion and Conclusions

7.1 General discussion

This thesis addresses the problem of data integration for multi-omics experiments. More specifically, this research focused on two of the most characteristic computational challenges of Systems Biology (SB): the development of integrative databases and the problem of integrative visualisation.

From the standpoint of developing integrated database resources, the wide variety of data sources, formats, and content requires a lot of work to store and organise large datasets. In addition, the highly context-dependent nature of omics data requires detailed meta-data to be recorded to help the identification of each dataset; this can only be accomplished by adopting community standards for the data schemas, formats, nomenclatures, and protocols. However, to take full advantage of the potential of visual analytics, visualisation tools for SB must maintain the balance between clarity, precision, and efficiency. The high volume and wide variety of omics data, and the complexity of interactions between biological components, makes choosing data representations that enable the extraction of meaningful knowledge, and minimise information and performance loss challenging [149]. This thesis focuses on some of the problems in multi-omics data integration that were relevant to the state of the art technologies in SB when the work started.

Integrative analysis is usually performed by scientists with a limited computational background and SB approaches are being increasingly adopted by medium and small-scale laboratories with limited access to powerful bioinformatics infrastructures. Thus, the methods proposed had to be easy to use and applicable to the majority of such research. Therefore, we made our tools web-based, open-source, and freely available, ensuring accessibility and free-distribution of the software and also providing a high-throughput and reliable research resource. We also considered user-friendliness and flexibility to be fundamental to the success of our tools. The usability of web tools is hampered by the conceptual complexity of the methods they use and the size of the data displayed, and requires aspects such as efficiency, clarity, and learnability to be balanced. Thus, our strategies focused on creating an effective user experience; for instance, we adopted intuitive layouts and colour schemes and implemented interactive chart designs so that users can easily navigate the system and quickly identify the information relevant to them. Similarly, the design of our simple input forms, complemented with default and optimal options, content validators and help tips, reduces the user's workload.

Performance also constitutes an important challenge for the development of web applications in the context of large datasets which increase data-processing and loading times, significantly degrading the user experience. Key user operations such as interactive real-time data browsing (logical and fluid application performance) are imperative in attracting and retaining users. Different methods can help to increase the application's responsiveness; for example, code optimisation (by introducing changes to data structure design, access patterns and code layouts) can significantly reduce resource use and computation times, but must be weighed against the potential negative effects of reduced code readability and higher maintenance. Thus, we implemented code optimisation techniques in loops and repeated tasks using high-performance parallel computing techniques by dividing large complex tasks across multiple processors. Similarly, good database design and the creation of efficient indexes can dramatically reduce data-query and update response times and increase application performance and so we devoted special attention to optimising resource allocation and retaining data flexibility in our tools. Finally, we also reduced loading times by using data compression, managing the load-distribution of user-client data exchange, and using temporary user data-caching.

STATegra EMS

The first part of this thesis (Chapters 3 and 4) addressed the absence of comprehensive tools for properly storing and organising large datasets and processing pipelines in multi-omics experiments. The STATegra Experiment Management System (EMS) was designed to store experimental information for entire integrative-experiment workflows – from biomaterial production to results generation – including the experimental design and protocols for sample measurements and data processing. This architecture is substantially different from other information management solutions created for next-generation sequencing (NGS) data that focus on the production on the samples. Nevertheless, the EMS goes beyond and provides tools for the further analysis and data generation. As a result, the system provides researchers with a valuable tool for tracking data provenance and ensuring experimental reproducibility. In addition, the tools included for exporting annotations to popular formats allow users to easily distribute their methods and results with the research community. Obtaining an accurate description for each information unit in a given study is not trivial because of the heterogeneity of omics data; thus, before creating this system we extensively studied the data collection techniques commonly used in SB so that we could include the most useful standards for each data type. Moreover, we also put significant effort into conceptualising the analysis workflows for each omics discipline while also retaining sufficient flexibility to allow users to implement alternative analysis flows. Thus, we identified each step involved and the minimum meta-data required to comprehensively describe it.

Another characteristic that distinguishes the proposed tool from other solutions is its collaborative nature, developed to accommodate the fact that in most research labs the samples are usually prepared, measured, and the data analysed by different people. Enabling collaboration in web applications means that the system information synchronisation must be tightly controllable (so as to avoid any data conflicts or loss) and means that several layers of complexity must be introduced when designing the system logic. The STATegra EMS use case-study (Chapter 4) gives detailed insights into how we organised the annotation process in this tool and uses the context of a real integrative study to show the logic behind the incorporation of the different layers of information the application manages.

PaintOmics 3

The second part of this thesis focused on the challenges associated with integrative data visualisation and proposed solutions that can be used in multi-omics experiments. As seen in Chapter 3, there are three major visualisation-tool categories in SB, but none of these can achieve comprehensive integration. Networks are useful for studying the connections between the system components, but the increasing size and complexity of these data and the lack of a biological context usually hamper their interpretation. Alternatively, solutions based on dissecting the pathway data sets into smaller subsets (based on previously established knowledge) adds biological meaning and simplifies interpretation, but means foregoing the ability to gain new insights into the system's functioning. Moreover, neither approach can easily accept omics data types derived from genomic regions such as ChIP-seq or DNase-seq data types. In contrast, genome browsers can handle these data types but fail to provide an appropriate framework for comparative studies in multi-condition experiments.

As discussed in Chapter 5, the proposed solution – PaintOmics 3 – uses pathway diagrams as the major analysis and visualisation method but also takes advantage of the other two approaches. Networks are used to display the pathway interactions and major omics expression data-trends and these are combined with pathway-based visualisation and statistical methods for studying individual pathways. This allows detailed analysis, even at the individual gene or metabolite level, and provides researchers with a comprehensive tool for navigating multiple biological system levels and for observing any underlying associations or regulation dynamics in these biological processes.

One of the key features of PaintOmics 3 is its flexibility. As opposed to the other available systems, which are restricted to specific species subsets, PaintOmics3 supports all of the organisms available in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. Furthermore, in contrast to other pathway-based solutions, PaintOmics 3 does not limit the number or variety of input omics data-types; instead it includes an effective method for translating the different naming conventions and feature identifiers into the corresponding KEGG name domain for the studied organism, including region-based and microRNA-based omics types. These features automatically translate the feature names and identifiers in order to improve the tool's usability and allow users to work with their datasets, regardless of the nomenclature of the identifiers or names used. PaintOmics 3 can be used for simple "Case-Control" experiments with a single omics type, but also works with complex experimental designs that may include several conditions, time-series, and multiple omics measurements. The use case-study, which utilised data from the STATegra project [134] (Chapter 6), reflects this flexibility and demonstrates that PaintOmics 3 represents a powerful minimal-effort tool for data enrichment analysis and results validation.

7.2 Conclusions

The following text summarises the conclusions of this work and is organised according to the goals defined in Chapter 1.

1. **To develop a user-friendly and integrated system for the management, annotation, and storage of multi-omics experiments.**

   - We proposed a novel tool for managing experiments (*STATegra EMS*) that provides an integrated system for annotating complex high-throughput omics studies; this includes descriptions of the biological material used in the experiment, its processing, and data-set analysis.

   - This system accepts multiple types of omics data, including five popular sequencing functional-assays and non-nucleic acid components such as proteomics and metabolomics, and uses accepted standards for storing the information for each specific omics type.

   - Web technologies are used in order to provide the key tools for running the system (i.e. by making it a centralised and collaborative service), and the accounts system allows multiple users to protect their information.

- The proposed solution was successfully applied in the context of a multi-omics study.

2. **To develop a user-friendly tool for integrative visualisation of multiple omics data types based on metabolic pathways.**

    - We developed a novel methodology for integrative data visualisation; its most important features are its flexibility and capacity for inter-user communication. The system provides a complete framework for biological function enrichment analysis and supports multiple species and many different omics types, including epigenomics data based on genomic regions and microRNA (miRNA) data.

    - The main visualisation approach is based on KEGG pathway diagrams but secondary tools extend the tool's scope by adding interaction networks. Other useful features are automatic identifier and name conversion, the interactivity of the displayed information, and the use of parallelisation to minimise back-end calculations.

3. **As general rule, the new tools developed must be reliable and user-friendly.**

    - Biologists and other researchers participated in designing the interfaces and defining the requirements of both tools in order to improve their user-friendliness. The tools were also extensively tested to ensure they are secure, reliable, and error-proof.

4. **Good accessibility and wide distribution of the generated software are also objectives for this thesis.**

    - Both systems are distributed as open-source software and public instances of the tools guarantees their free use. Moreover, both tools have also been presented at national and international conferences and in popular international scientific print journals. Finally, extensive user documentation has been developed for both systems.

7.3 Reach and relevance

- The methodologies described in this thesis have been implemented as open-source and freely-available web applications. On the one hand, the STATegra EMS provides the scientific community with a user-friendly bioinformatics management suite capable of collaboratively annotating multi-omics experiments. On the other hand, PaintOmics 3 provides researchers with an intuitive and flexible web platform for data enrichment analysis and integrative visualisation of biological functions for Systems Biology studies.

- This thesis was developed within the framework of the STATegra international research project [134], and the resulting methods have been extensively used to analyse the data generated within this project, also contributing to the dissemination of the results of the thesis.

- Both applications have been presented at international and national conferences and courses, demonstrating their usefulness to end-users. In addition, descriptions of the main features of both these tools, providing use case-examples in the context of multi-omics studies, have been published in popular open-access bioinformatics journals.

- The STATegra EMS has been incorporated into the eBioKit system [50], an educational and analytical bioinformatics platform developed by the SLU Global Bioinformatics Centre (Swedish University of Agricultural Sciences). This kit is used in more than 20 research centres and universities in Africa, South America, and Asia. Additionally, PaintOmics 3 has been also included in the new version of the eBioKit, currently on development.

- Additionally, the STATegra EMS has been selected as the main experiment management system for the eB3Kit, a bio-banking platform developed by the B3Africa project [73], funded by European Horizon 2020.

This project aims to implement a cooperation platform and technical informatics framework for integrating biobanks in Africa and Europe.

- PaintOmics 3 has been officially installed in two different locations: at the Principe Felipe Research Centre in Spain, and at the University of Florida bioinformatics facilities. The system has over 200 active accounts.

7.4 Major scientific contributions

During the development of this thesis I have had the opportunity to spread my work and the state of the art of omics data integration to the scientific community in different events. Below are some of the most relevant.

Publications

- Hernández-de-Diego R, Boix-Chova N, Gómez-Cabrero D, Tegner J, Abugessaisa I, and Conesa A. *STATegra EMS: an Experiment Management System for complex next-generation omics experiments.* **BMC Systems Biology**, 8 Suppl 2:S9, **2014**.

- Conesa A, and Hernández-de-Diego R. *Omics Data Integration in Systems Biology: Methods and Applications*. In: Applications of Advanced Omics Technologies: From Genes to Metabolites, Volume 64 (**Comprehensive Analytical Chemistry**). **2014**.

- Hernández-de-Diego R, de Villiers EP, Klingström T, Gourlé H, Conesa A, and Bongcam-Rudloff E. *The eBioKit, a stand-alone educational platform for bioinformatics*. **PLoS Computational Biology**. (In revision)

- Tarazona S*, Hernández-de-Diego R, Silberberg G*, Ferreiros I, van del Kloet F, Ramirez R, Schmidt A, Marabita F, Lagani V, Papoutsoglou G, Hankemeier T, Westernhuis J, Imhof A, Ballestar E, Meier D, Lappe M, Tsamardinos I, Mortazavi A, Merkenschlager M, Tenger J, Gomez-Cabrero D, and Conesa A. *A comprehensive guideline to the multiomics data analysis paradigm in time course perturbation studies*. (In preparation)

- Hernández-de-Diego R, Tarazona S, Furió-Tarí P, and Conesa A. *Paintomics: a web resource for the pathway level visualisation of multiomics data*. (In preparation)

Conferences

- HiTSeq ISCB/ECCB 2013. Berlin, Germany. July 2013. Hernández-de-Diego R, Boix-Chova N, Gómez-Cabrero D, Tegner J, Abugessaisa I, and Conesa A. *The STATegra NGS Experiment Management System* (Poster).
- XII Symposium on Bioinformatics. Sevilla, Spain. September 2014. Hernández-de-Diego R, Boix-Chova N, Gómez-Cabrero D, Tegner J, Abugessaisa I, and Conesa A. *STATegra EMS: an Experiment Management System for complex next-generation omics experiments* (Poster).
- SMODIA 2014. Heraklion, Greece. November 2014. Hernández-de-Diego R, Boix-Chova N, Gómez-Cabrero D, Tegner J, Abugessaisa I, and Conesa A. *The STATegraEMS, an Experiment Management System for multi-omics experiments* (Oral presentation).
- SMODIA 2015. Valencia, Spain. September 2015. Hernández-de-Diego R, Tarazona S, Furió-Tarí P, and Conesa A. *Paintomics 3.0: Integrated visualization of multi omics data on KEGG pathways* (Oral presentation).
- XIII Symposium on Bioinformatics. Valencia, Spain. May 2016. Hernández-de-Diego R, Tarazona S, Furió-Tarí P, and Conesa A. *Integrative visualization of multi omics data: the PaintOmics 3 platform* (Poster).

Teaching activities

- X International Course of Massive Data Analysis in Transcriptomics. Centro de Investigación Príncipe Felipe. Valencia, Spain. March 2014.
- The Genomics of Gene Expression RNA-Seq Course. Centro de Investigación Príncipe Felipe. Valencia, Spain. September 2014.

- The H3ABioNet Advanced Systems Administration workshop: Introduction to the eBioKit system: services, architecture and administration. University of Pretoria, Pretoria, South Africa. February 2015.
- The Genomics of Gene Expression RNA-Seq Course. Centro de Investigación Príncipe Felipe. Valencia, Spain. March 2015.
- STATegra Summer School in Omics Data Integration. Benicassim, Spain. September 2015.

7.5 Major future perspectives

Our current and future lines of research are defined by the requests the application users. As usual in software development, the applications are now in the "transition" stage in which new features and tools are added to the system to provide a more complete user-experience. Some of the most important features scheduled for inclusion in their forthcoming versions are:

- Improve the usability for the STATegra EMS adding new features to the system. Some examples would be:
  - Tools for uploading, downloading and manage files into the system (on development).
  - Support for different storage systems (e.g. FTP servers or iRODS storage system, on development).
  - Wizards for creating multiple analysis or samples.
  - New tools for querying the system.
  - Controlled vocabulary and ontologies for use in the experiment annotations.
- Increase the scope of the PaintOmics 3 application by adding new features such as the following:
  - Inclusion of alternative methods for pathway enrichment analysis.

- Inclusion of new information available in KEGG, such as enzymes or Gene Ontology (GO) terms.
- Support for alternative databases of pathway information such as Reactome [33].
- Include new types of visualization tools such as genome browsers.
- Reduces the installation time of new databases by providing database dumps.

Appendices

A

Selection of pathways for the PaintOmics 3 use case

A.1 Introduction

This section enumerates the selected KEGG pathways that were considered as "closely related" or "interesting" for the model biological system studied in the STATegra project [134]. The model describes the differentiation of the mouse B3 cell line (cycling pre-B cells) under the controlled induction of the Ikaros transcription factor (TF). The differentiation is controlled by a tamoxifen-inducible vector of the Ikaros TF (Ikaros-ERt2), while control cells carry an empty vector. This model is of special clinical interest because the genetic deletion of Ikaros can result in severe disturbances or even completely block B-cell development. Based on the knowledge of the consortium experts, the following subset of the mouse pathways was considered for the evaluation of the study results.

A.2 Selection of pathways

- Cellular Processes
  - Cell growth and death
    - Apoptosis
    - Cell cycle
    - p53 signaling pathway
  - Cell motility
    - Regulation of actin cytoskeleton
  - Cellular community
    - Adherens junction
    - Focal adhesion
    - Gap junction
    - Tight junction
  - Transport and catabolism
    - Endocytosis
    - Lysosome
    - Peroxisome
    - Phagosome
    - Regulation of autophagy
- Environmental Information Processing
  - Membrane transport
    - ABC transporters
  - Signal transduction
    - AMPK signaling pathway

        - Calcium signaling pathway
        - cAMP signaling pathway
        - cGMP-PKG signaling pathway
        - ErbB signaling pathway
        - FoxO signaling pathway
        - Hedgehog signaling pathway
        - HIF-1 signaling pathway
        - Hippo signaling pathway
        - Jak-STAT signaling pathway
        - MAPK signaling pathway
        - mTOR signaling pathway
        - NF-kappa B signaling pathway
        - Notch signaling pathway
        - Phosphatidylinositol signaling system
        - PI3K-Akt signaling pathway
        - Rap1 signaling pathway
        - Ras signaling pathway
        - Sphingolipid signaling pathway
        - TGF-beta signaling pathway
        - TNF signaling pathway
        - VEGF signaling pathway
        - Wnt signaling pathway

    - Signaling molecules and interaction

        - Cell adhesion molecules (CAMs)
        - Cytokine-cytokine receptor interaction
        - ECM-receptor interaction
        - Neuroactive ligand-receptor interaction

- Genetic Information Processing

    - Folding, sorting and degradation

        - Proteasome
        - Protein export
        - Protein processing in endoplasmic reticulum
        - RNA degradation
        - SNARE interactions in vesicular transport
        - Ubiquitin mediated proteolysis

    - Replication and repair

        - Base excision repair
        - DNA replication
        - Fanconi anemia pathway
        - Homologous recombination
        - Mismatch repair
        - Non-homologous end-joining
        - Nucleotide excision repair

    - Transcription

        - Basal transcription factors
        - RNA polymerase
        - Spliceosome

    - Translation

        - Aminoacyl-tRNA biosynthesis
        - mRNA surveillance pathway
        - Ribosome
        - Ribosome biogenesis in eukaryotes
        - RNA transport

- Metabolism

    - Carbohydrate metabolism

        - Amino sugar and nucleotide sugar metabolism

    - Ascorbate and aldarate metabolism
    - Citrate cycle (TCA cycle)
    - Fructose and mannose metabolism
    - Galactose metabolism
    - Glycolysis / Gluconeogenesis
    - Inositol phosphate metabolism
    - Pentose and glucuronate interconversions
    - Pentose phosphate pathway
    - Pyruvate metabolism
    - Starch and sucrose metabolism

  - Energy metabolism

    - Oxidative phosphorylation

  - Global and overview maps

    - 2-Oxocarboxylic acid metabolism
    - Carbon metabolism
    - Fatty acid metabolism
    - Metabolic pathways
    - Biosynthesis of amino acids

  - Lipid metabolism

    - Fatty acid biosynthesis
    - Fatty acid degradation
    - Fatty acid elongation

  - Metabolism of cofactors and vitamins

    - Biotin metabolism
    - Folate biosynthesis
    - Lipoic acid metabolism
    - Nicotinate and nicotinamide metabolism

    - One carbon pool by folate
    - Pantothenate and CoA biosynthesis
    - Porphyrin and chlorophyll metabolism
    - Retinol metabolism
    - Riboflavin metabolism
    - Thiamine metabolism
    - Ubiquinone and other terpenoid-quinone biosynthesis
    - Vitamin B6 metabolism

  - Nucleotide metabolism

    - Purine metabolism
    - Pyrimidine metabolism

  - Xenobiotics biodegradation and metabolism

    - Drug metabolism - cytochrome P450

- Organismal Systems

  - Endocrine system

    - Adipocytokine signaling pathway
    - Glucagon signaling pathway
    - GnRH signaling pathway
    - Insulin signaling pathway
    - PPAR signaling pathway

  - Environmental adaptation

    - Circadian rhythm

  - Immune system

    - Antigen processing and presentation
    - B cell receptor signaling pathway
    - Chemokine signaling pathway

- ∗ Cytosolic DNA-sensing pathway
- ∗ Fc epsilon RI signaling pathway
- ∗ Fc gamma R-mediated phagocytosis
- ∗ Hematopoietic cell lineage
- ∗ Leukocyte transendothelial migration
- ∗ NOD-like receptor signaling pathway
- ∗ RIG-I-like receptor signaling pathway
- ∗ Toll-like receptor signaling pathway

B

PySiQ, a Python Simple Queue system

B.1 Introduction

PySiQ (Python Simple Queue) is a job queue or task queue implemented for Python applications. The main objective of task queues is to avoid running resource-intensive tasks immediately and wait for them to complete. Instead, tasks are scheduled by adding them to a queue, where they will wait until eventually a *worker*, i.e. a special process running in separate thread, takes them out of the queue and execute the job. This concept is especially necessary for web applications where it is not possible to handle a heavy task during a short HTTP request window.

B.2 Features

PySiQ has been entirely implemented in Python and provides the following features:

- Multi-process execution on tasks, configurable number of workers.
- The status of the queued tasks can be easily checked at any time.

- Dependencies between tasks can be specified executing them in the appropriate order.
- The results for tasks are stored until the client asks for them.
- Lightweight module. The code takes less than 300 lines of code.
- Easy to use and to install in your application.
- It does not depend on other libraries or tools.

B.3 Design

The main component of PySiQ is the **Queue**, a Python object that works as a task dispatcher and worker-pool manager. The Queue is a *singleton* instance that is listening to the other components in the application (Figure B.1-1). When the queue is instantiated, a certain number of **Workers** are created depending on the user's settings. Workers are special threads that extract tasks from the queue and execute them. By default, workers are idle, waiting for new tasks are sent to the queue (Figure B.1-2). When a client needs to execute certain time-consuming job, it is encapsulated in a **Task** instance, defining the function or code to be executed and the parameters for its execution (Figure B.1-3). Some additional parameters can be specified such as a timeout that will abort the execution of the task if it does not finish after a determined amount of time, and a list of dependencies, i.e. the identifiers for the tasks that must be completed before launching the execution of the new task. The task instance is sent to the queue and workers are notified that a new task is waiting for being executed. As soon as a worker is idle, it takes the next task at the queue and starts the execution, provided that all its dependencies already finished (Figure B.1-4).

The queue contains an internal table that keeps the status for all the tasks in the queue. Possible statuses are: "waiting", "running", "finished", and "error" (Figure B.1-5). When a task is finished, it is kept in this table in addition to the results of the execution, until someone asks for the results. Similarly, failed tasks are kept in the table with the information of the error (Figure B.1-6).

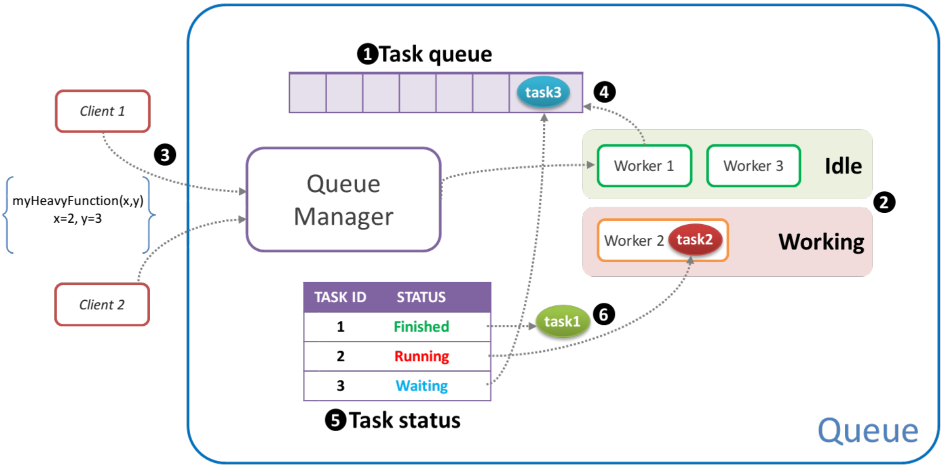


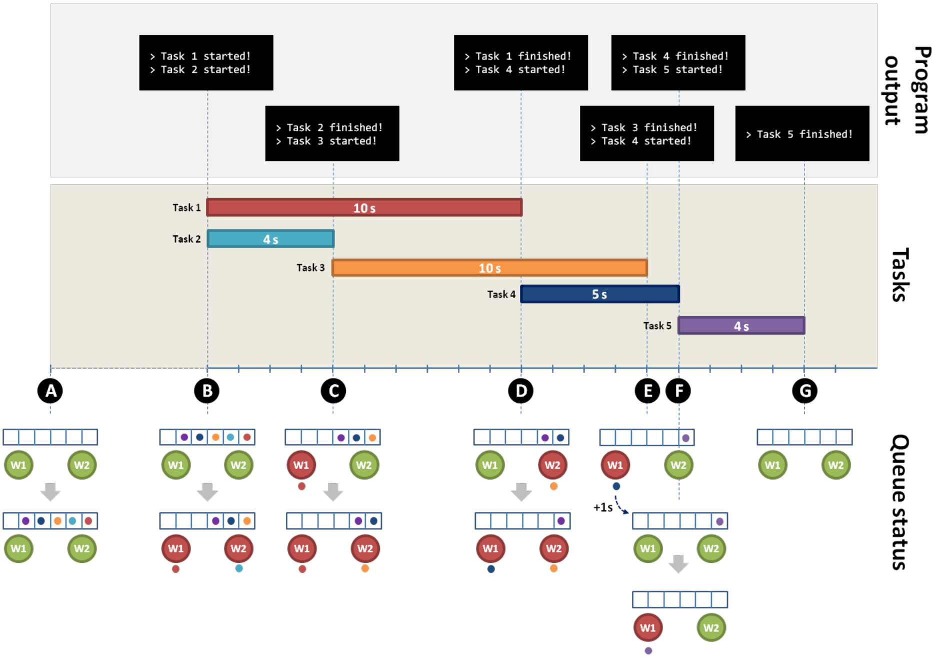
Figure B.1: Overview of the design of the queue system.

B.4 Example of use

The following code fragment exemplifies the usage of the developed module. For this use case an instance of queue is initialized with two workers. A total of five tasks are sent to the queue. All tasks will execute the same function called *foo* which displays a message indicating that execution has started, then waits *N* seconds, and displays a message announcing the end of execution. Both the displayed message and the duration of the delay (the *N* value) are provided as parameters for the *foo* function. Tasks 1 and 3 will take 10 seconds for execution, while tasks 2 and 4 will take less than 5 seconds. Task 5 takes 4 seconds but it won't start until tasks 3 and 4 are completed. Figure B.2 shows the temporal line for the execution of the tasks, as well as the status for the queue and the workers at different time-points.

```
from PySiQ import Queue

# ************************************************************************
# Initialize queue
# ************************************************************************
N_WORKERS = 2

queue_instance = Queue()
queue_instance.start_worker(N_WORKERS)

# ************************************************************************
# Queue tasks
# ************************************************************************

def foo(N, message):
    print message + " started..."
    from time import sleep
    sleep(N)
    print message + " finished"

queue_instance.enqueue(
    fn=foo,
    args=(10, "Task 1"),
    task_id= "Task 1"
)

queue_instance.enqueue( fn=foo, args=(4, "Task 2"), task_id= "Task 2")

queue_instance.enqueue( fn=foo, args=(10, "Task 3"), task_id= "Task 3")

queue_instance.enqueue( fn=foo, args=(5, "Task 4"), task_id= "Task 4")

queue_instance.enqueue(
    fn=foo,
    args=(4, "Task 5"),
    task_id= "Task 5",
        depend= ["Task 3", "Task 4"]
)
```

Code fragment B.1: Example of usage of the PySiQ module.



```
> Task 1 started!
> Task 2 started!
```

```
> Task 2 finished!
> Task 3 started!
```

```
> Task 1 finished!
> Task 4 started!
```

```
> Task 3 finished!
> Task 4 started!
```

```
> Task 4 finished!
> Task 5 started!
```

```
> Task 5 finished!
```

Figure B.2: Output for the example program, overview for the tasks' execution, and status of the queue and workers. (A) - When the program execution starts, the queue is empty and both workers are idle. The five tasks are then sequentially added to the queue and workers are notified to start working. **(B)** - The first two tasks are extracted from the queue and executed by the workers. Each task shows a message in the terminal. **(C)** - After four seconds, task 2 finishes, and worker 2 becomes idle. Worker 2 then asks for a new task (task 3) and executes it. **(D)** - After four seconds, task 1 finishes and task 4 starts. **(E)** - When task 3 is done, worker 2 become idle and waits for a new task; however, task 5 cannot start because it depends on task 4, which still running on worker 1. **(F)** - Once task 4 finished, task 5 is valid for being executed by any worker. **(G)** - Finally, after 4 seconds, task 5 is finished.

B.5 Advanced use

Figure B.3 shows a more complex example of using. For this use of case two different users interact with the queue by sending requests to a server-side program. More specifically, each user sends first a request to the server in order to launch certain time-consuming task. For each client, a task instance is created and sent to the queue and the identifier for the new job is returned to the client application. While waiting for the end if the task, the clients send periodically a request in order to query the execution status. Finally, when the task if finished, the user can retrieve the result for the execution of the task.

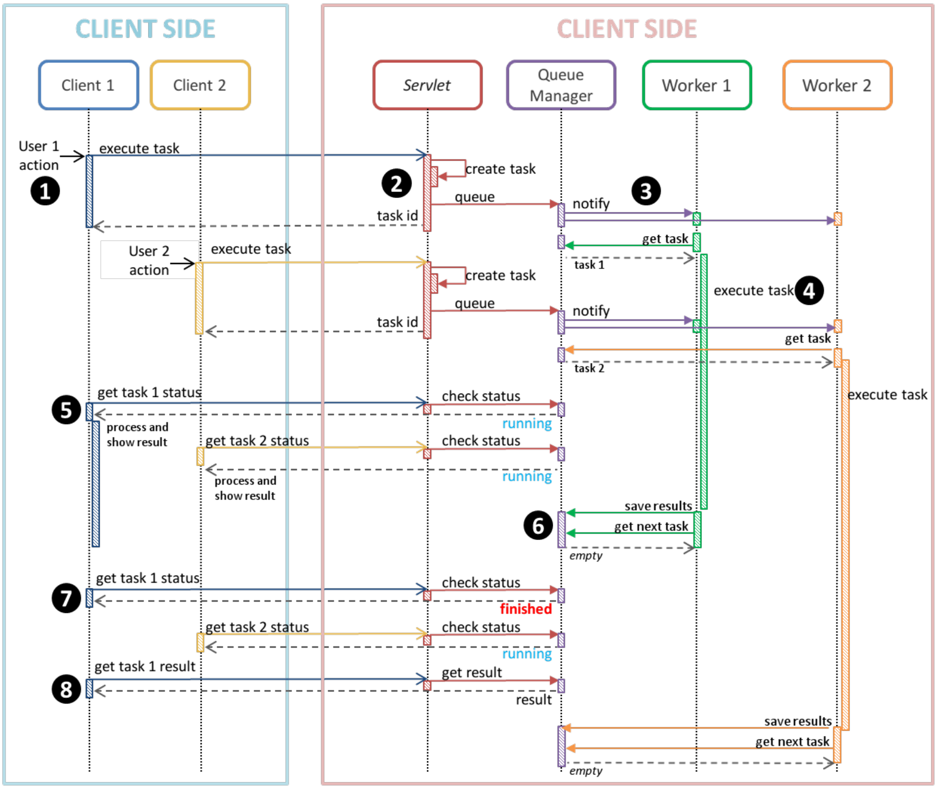


Figure B.3: Using the queue system in a Client-Server environment. From client side, two different users send a request to execute some time-consuming task at server side (1). When a request is received, the server program wraps the job to be executed, and its parameters into an instance of task, and sends the new object to the queue system (2). When a task is queued, an identifier is returned that uniquely identifies the task in the queue system. Task identifiers are then returned to the corresponding client that can use it later to check the execution status or retrieve the results. Every time that a new task is received, the queue system notifies the workers (3), and those that are idle extract and execute the next tasks in the queue (4). While the task is being executed, clients can check periodically the status of the job by sending a request to the server (5). Possible statuses are "queued", "running", "finished" and "failed". When a worker finishes the execution of a task, the result is kept in the queue system until it is requested, and the worker proceeds to execute the next task in queue (6). If no new tasks are available the worker becomes idle. Finally, when the client detects that the task is done - i.e. it receives a 'finished' status (7), a new request is sent in order to retrieve the results of the execution (8).

B.6 Availability

PySiQ module is distributed under MIT license and can be downloaded from the project repository (*see table below*). Sources are hosted at *GitHub*, a popular web-based Git repository, which allows anyone to browse and download the code or discuss it, submit contributions, and review the code.

Availability and requirements	
Project links	
Sources:	`https://github.com/fikipollo/PySiQ`
Other information	
Operating system(s):	Platform independent
Programming language(s):	Python
License:	MIT license

Acronyms

A adenine.

AJAX asynchronous JavaScript and XML.

API application programming interface.

AS analytical sample.

ATP adenosine triphosphate.

BC biological condition.

BLAST basic local alignment search tool.

BR biological replicate.

C cytosine.

cDNA complementary DNA.

ChIP chromatin immunoprecipitation.

ChIP-seq chromatin immunoprecipitation followed by sequencing.

CLP common lymphoid progenitor.

CQN conditional quantile normalisation.

CSS cascading style sheets.

DAO data access object.

DE differentially-expressed.

DNA deoxyribonucleic acid.

DNase-seq DNase I hypersensitive site sequencing.

DTO data transfer object.

EMS Experiment Management System.

ENCODE Encyclopaedia of DNA Elements.

FP7 7th Framework Programme.

FTP file transfer protocol.

G guanine.

GB genome browser.

GC gas chromatography.

GO Gene Ontology.

HSC hematopoietic stem cells.

HTML hypertext markup language.

HTTP hypertext transfer protocol.

ID identifier.

IGB Integrated Genome Browser.

IGV Integrative Genomics Viewer.

IU information unit.

JDBC Java database connectivity.

JSON JavaScript object notation.

KEGG Kyoto Encyclopedia of Genes and Genomes.

KGML KEGG markup language.

LC liquid chromatography.

LIMS laboratory information and management systems.

MIAPE minimum information about a microarray experiment.

MINSEQE minimum information about a high-throughput sequencing experiment.

miRNA microRNA.

miRNA-seq microRNA sequencing.

MPP multipotent progenitor.

mRNA messenger RNA.

mRNA-seq messenger RNA sequencing.

MS mass spectrometry.

MVC model-view-controller.

NGS next-generation sequencing.

NMR nuclear magnetic resonance.

PCA principal component analysis.

PEA pathway enrichment analysis.

POI protein of interest.

REST representational state transfer.

RNA ribonucleic acid.

RNA-seq RNA sequencing.

RRBS-seq reduced representation bisulfite sequencing.

rRNA ribosomal RNA.

SB Systems Biology.

SMS sample management systems.

SNP single nucleotide polymorphism.

SRA Sequence Read Archive.

sRNA-seq small RNA sequencing.

SuS system under study.

SVG scalable vector graphic.

T thymine.

TF transcription factor.

TGCA The Cancer Genome Atlas.

tRNA transfer RNA.

TSS transcription start site.

U uracil.

UCSC University of California Santa Cruz.

XML extensible markup language.

Bibliography

[1] Abugessaisa, I., Gomez-Cabrero, D., Snir, O., Lindblad, S., Klareskog, L., Malmström, V., and Tegnér, J. "Implementation of the CDC translational informatics platform–from genetic variants to the national Swedish Rheumatology Quality Register". In: *J Transl Med* 11 (2013), p. 85 (cit. on p. 51).

[2] Abugessaisa, I., Saevarsdottir, S., Tsipras, G., Ståhle, M., Malmström, V., Klareskog, L., and Tegnér, J. "Implementation of the CDC translational informatics platform–from genetic variants to the national Swedish Rheumatology Quality Register". In: *J Transl Med* 11 (2013), p. 85 (cit. on p. 51).

[3] Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., and Walter, P. *Molecular Biology of the Cell*. Garland Science, 2014. ISBN: 0815344325 (cit. on p. 12).

[4] Alur, D., Malks, D., and Crupi, J. *Core J2EE Patterns: Best Practices and Design Strategies (2nd Edition)*. Prentice Hall, 2003. ISBN: 9780131422469 (cit. on p. 38).

[5] Anders, S., Pyl, P. T., and Huber, W. "HTSeq–a Python framework to work with high-throughput sequencing data". In: *Bioinformatics* 31.2 (Jan. 2015), pp. 166–169 (cit. on p. 66).

[6] Annunziato, A. T. “DNA Packaging: Nucleosomes and Chromatin”. In: *Nature Education* 1.1 (2008), p. 26 (cit. on p. 10).

[7] Armin Ronacher and contributors. *Flask, a Python Microframework*. [Online; accessed 11-December-2015] (cit. on p. 81).

[8] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. “Gene ontology: tool for the unification of biology. The Gene Ontology Consortium”. In: *Nat. Genet.* 25.1 (2000), pp. 25–29 (cit. on p. 27).

[9] Baltz, A. G., Munschauer, M., Schwanhäusser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M., Wyler, E., Bonneau, R., Selbach, M., Dieterich, C., and Landthaler, M. “The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts”. In: *Mol. Cell* 46.5 (June 2012), pp. 674–690 (cit. on p. 31).

[10] Bastian, M., Heymann, S., and Jacomy, M. “Gephi: An Open Source Software for Exploring and Manipulating Networks”. In: (2009) (cit. on p. 78).

[11] Bateman, A. et al. “UniProt: a hub for protein information”. In: *Nucleic Acids Res.* 43.Database issue (Jan. 2015), pp. D204–212 (cit. on p. 101).

[12] Benson, D. A., Karsch-Mizrachi, I., Clark, K., Lipman, D. J., Ostell, J., and Sayers, E. W. “GenBank”. In: *Nucleic Acids Res.* 40.Database issue (Jan. 2012), pp. 48–53 (cit. on p. 101).

[13] Berger, B., Peng, J., and Singh, M. “Computational solutions for omics data”. In: *Nat. Rev. Genet.* 14.5 (2013), pp. 333–346 (cit. on p. 13).

[14] Binneck, E., Silva, J. F., Neumaier, N., Farias, J. R., and Nepomuceno, A. L. "VSQual: a visual system to assist DNA sequencing quality control". In: *Genet. Mol. Res.* 3.4 (2004), pp. 474–482 (cit. on p. 32).

[15] Blake, J. A. et al. "The Gene Ontology: enhancements for 2011". In: *Nucleic Acids Res.* 40.Database issue (Jan. 2012), pp. D559–564 (cit. on p. 101).

[16] Bootstrap Core Team. *Bootstrap: HTML, CSS, and JavaScript framework for developing responsive web applications*. [Online; accessed 07-January-2016] (cit. on p. 84).

[17] Bordbar, A., Mo, M. L., Nakayasu, E. S., Schrimpe-Rutledge, A. C., Kim, Y. M., Metz, T. O., Jones, M. B., Frank, B. C., Smith, R. D., Peterson, S. N., Hyduke, D. R., Adkins, J. N., and Palsson, B. O. "Model-driven multi-omic data analysis elucidates metabolic immunomodulators of macrophage activation". In: *Mol. Syst. Biol.* 8 (2012), p. 558 (cit. on p. 31).

[18] Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data". In: *Nat. Genet.* 29.4 (Dec. 2001), pp. 365–371 (cit. on pp. 27, 43).

[19] Burbeck, S. *Applications Programming in Smalltalk-80: How to Use Model-View- Controller (MVC).* Softsmarts, Incorporated, 1987 (cit. on p. 40).

[20] Bussey, K. J., Kane, D., Sunshine, M., Narasimhan, S., Nishizuka, S., Reinhold, W. C., Zeeberg, B., Ajay, W., and Weinstein, J. N. "Match-

Miner: a tool for batch navigation among gene and gene product identifiers". In: *Genome Biol.* 4.4 (2003), R27 (cit. on p. 101).

[21] Camerlengo, T., Ozer, H. G., Onti-Srinivasan, R., Yan, P., Huang, T., Parvin, J., and Huang, K. "From sequencer to supercomputer: an automatic pipeline for managing and processing next generation sequencing data". In: *AMIA Jt Summits Transl Sci Proc* 2012 (2012), pp. 1–10 (cit. on p. 33).

[22] Cassman, M. "Barriers to progress in systems biology". In: *Nature* 438.7071 (2005), p. 1079 (cit. on p. 27).

[23] Chou, C. H., Chang, N. W., Shrestha, S., Hsu, S. D., Lin, Y. L., Lee, W. H., Yang, C. D., Hong, H. C., Wei, T. Y., Tu, S. J., Tsai, T. R., Ho, S. Y., Jian, T. Y., Wu, H. Y., Chen, P. R., Lin, N. C., Huang, H. T., Yang, T. L., Pai, C. Y., Tai, C. S., Chen, W. L., Huang, C. Y., Liu, C. C., Weng, S. L., Liao, K. W., Hsu, W. L., and Huang, H. D. "miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database". In: *Nucleic Acids Res.* 44.D1 (2016), pp. D239–247 (cit. on p. 98).

[24] Clark, M. R., Mandal, M., Ochiai, K., and Singh, H. "Orchestrating B cell lymphopoiesis through interplay of IL-7 receptor and pre-B cell receptor signalling". In: *Nat. Rev. Immunol.* 14.2 (2014), pp. 69–80 (cit. on p. 140).

[25] Clarke, C. J. and Haselden, J. N. "Metabolic profiling as a tool for understanding mechanisms of toxicity". In: *Toxicol Pathol* 36.1 (2008), pp. 140–147 (cit. on p. 20).

[26] Conesa, A. and Hernández-de-Diego, R. "Omics Data Integration in Systems Biology: Methods and Applications". In: *Applications of Advanced Omics Technologies: From Genes to Metabolites, Volume 64 (Comprehensive Analytical Chemistry).* Ed. by García-Cañas, V., A., Cifuentes, and C., Simó. Elsevier, 2014 (cit. on p. 26).

[27] Cymerman, M. *Secure a Web application, Java-style*. [Online; accessed 29-October-2015]. 2000 (cit. on p. 42).

[28] Demirel, Y. *Energy: Production, Conversion, Storage, Conservation, and Coupling*. Springer, 2012. ISBN: 9781447123712 (cit. on p. 8).

[29] Desiderio, S., Lin, W. C., and Li, Z. "The cell cycle and V(D)J recombination". In: *Curr. Top. Microbiol. Immunol.* 217 (1996), pp. 45–59 (cit. on p. 140).

[30] DivShot Inc. *A field guide to STATIC APPS: Authentication and Authorization*. [Online; accessed 29-October-2015] (cit. on p. 41).

[31] Dweep, H. and Gretz, N. "miRWalk2.0: a comprehensive atlas of microRNA-target interactions". In: *Nat. Methods* 12.8 (2015), p. 697 (cit. on p. 98).

[32] ENCODE Project Consortium. "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414 (2012), pp. 57–74 (cit. on pp. 1, 26, 32).

[33] Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., Matthews, L., May, B., Milacic, M., Rothfels, K., Shamovsky, V., Webber, M., Weiser, J., Williams, M., Wu, G., Stein, L., Hermjakob, H., and D'Eustachio, P. "The Reactome pathway Knowledgebase". In: *Nucleic Acids Res.* 44.D1 (2016), pp. D481–487 (cit. on p. 157).

[34] Fan, J., Han, F., and Liu, H. "Challenges of Big Data Analysis". In: *Natl Sci Rev* 1.2 (2014), pp. 293–314 (cit. on p. 22).

[35] Ferreiros-Vidal, I., Carroll, T., Taylor, B., Terry, A., Liang, Z., Bruno, L., Dharmalingam, G., Khadayate, S., Cobb, B. S., Smale, S. T., Spivakov, M., Srivastava, P., Petretto, E., Fisher, A. G., and Merkenschlager, M. "Genome-wide identification of Ikaros targets elucidates its contribution

to mouse B-cell lineage specification and pre-B-cell differentiation". In: *Blood* 121.10 (2013), pp. 1769–1782 (cit. on pp. 54–56, 131).

[36] Filipowicz, W., Bhattacharyya, S. N., and Sonenberg, N. "Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight?" In: *Nat. Rev. Genet.* 9.2 (2008), pp. 102–114 (cit. on p. 13).

[37] Fowler, M. *Patterns of Enterprise Application Architecture*. Addison-Wesley, 2003. ISBN: 0321127420 (cit. on p. 38).

[38] Furio-Tari, P., Conesa, A., and Tarazona, S. "RGmatch: matching genomic regions to proximal genes in omics data integration". In: *BMC Bioinformatics* 17.Suppl 15 (2016), p. 427 (cit. on p. 95).

[39] G., Barchet (cit. on p. 20).

[40] Garcia-Alcalde, F., Garcia-Lopez, F., Dopazo, J., and Conesa, A. "Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data". In: *Bioinformatics* 27.1 (Jan. 2011), pp. 137–139 (cit. on pp. 29, 30, 79).

[41] Goble, C. and Stevens, R. "State of the nation in data integration for bioinformatics". In: *J Biomed Inform* 41.5 (2008), pp. 687–693 (cit. on p. 27).

[42] Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A., and Tegner, J. "Data integration in the era of omics: current and future challenges". In: *BMC Syst Biol* 8 Suppl 2 (2014), p. I1 (cit. on pp. 23, 26, 53).

[43] Goodacre, R., Broadhurst, D., Smilde, AK., Kristal, BS., Baker, JD., Beger, R., Bessant, C., Connor, S., Capuani, G., Craig, A., Ebbels, T., Kell, DB., Manetti, C., Newton, J., Paternostro, G., Somorjai, R., Sjöström, M., Trygg, J., and Wulfert, F. "Proposed minimum reporting

standards for data analysis in Metabolomics." In: *Metabolomics* 3.3 (2007), pp. 231–241 (cit. on p. 43).

[44] Goodenough J.and McGuire, B. *Biology of Humans: Concepts, Applications, and Issues (4th Edition)*. Pearson, 2011. ISBN: 0321707028 (cit. on p. 8).

[45] Google Inc. *AngularJS ? Superheroic JavaScript MVW Framework*. [Online; accessed 3-April-2017] (cit. on p. 84).

[46] Graves, P.R. and Haystead, T. A. "Molecular Biologist's Guide to Proteomics". In: *Microbiol Mol Biol Rev.* 66.1 (2002), 39?63 (cit. on pp. 19, 20).

[47] Hansen, K. D., Irizarry, R. A., and Wu, Z. "Removing technical variability in RNA-seq data using conditional quantile normalization". In: *Biostatistics* 13.2 (2012), pp. 204–216 (cit. on pp. 66, 72).

[48] Haquin, S., Oeuillet, E., Pajon, A., Harris, M., Jones, A. T., van Tilbeurgh, H., Markley, J. L., Zolnai, Z., and Poupon, A. "Data management in structural genomics: an overview". In: *Methods Mol. Biol.* 426 (2008), pp. 49–79 (cit. on p. 32).

[49] Hawkins, R. D., Hon, G. C., and Ren, B. "Next-generation genomics: an integrative approach". In: *Nat. Rev. Genet.* 11.7 (2010), pp. 476–486 (cit. on p. 26).

[50] Hernández-de-Diego, R., de Villiers, E. P., Klingström, T., Gourlé, H., Conesa, A., and Bongcam-Rudloff, E. "The eBioKit, a stand-alone educational platform for bioinformatics". In: *PLoS Comput. Biol.* (forthcoming) (cit. on p. 153).

[51] Herzog, S., Reth, M., and Jumaa, H. "Regulation of B-cell proliferation and differentiation by pre-B-cell receptor signalling". In: *Nat. Rev. Immunol.* 9.3 (2009), pp. 195–205 (cit. on p. 140).

[52] Highcharts. *Interactive Javascript charts for your webpage*. [Online; accessed 11-December-2015]. 2015 (cit. on p. 85).

[53] Holzinger, E. R. and Ritchie, M. D. “Integrating heterogeneous high-throughput data for meta-dimensional pharmacogenomics and disease-related studies”. In: *Pharmacogenomics* 13.2 (2012), pp. 213–222 (cit. on p. 24).

[54] Hu, Z., Chang, Y. C., Wang, Y., Huang, C. L., Liu, Y., Tian, F., Granger, B., and Delisi, C. “VisANT 4.0: Integrative network platform to connect genes, drugs, diseases and therapies”. In: *Nucleic Acids Res.* 41.Web Server issue (2013), W225–231 (cit. on p. 78).

[55] Huang, d. a. W., Sherman, B. T., Stephens, R., Baseler, M. W., Lane, H. C., and Lempicki, R. A. “DAVID gene ID conversion tool”. In: *Bioinformation* 2.10 (2008), pp. 428–430 (cit. on p. 101).

[56] Hunter, J. *Java servlet programming*. Beijing Sebastopol, CA: O’Reilly, 2001. ISBN: 978-0-596-00040-0 (cit. on p. 36).

[57] Ideker, T., Galitski, T., and Hood, L. “A new approach to decoding life: systems biology”. In: *Annu Rev Genomics Hum Genet* 2 (2001), pp. 343–372 (cit. on p. 22).

[58] Iersel, M. P. van, Pico, A. R., Kelder, T., Gao, J., Ho, I., Hanspers, K., Conklin, B. R., and Evelo, C. T. “The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services”. In: *BMC Bioinformatics* 11 (2010), p. 5 (cit. on p. 101).

[59] Institute, The Broad. *The Broad?s glossary: Transcription Factor*. [Online; accessed 29-October-2016]. 2016 (cit. on p. 12).

[60] Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. “ForceAtlas2, a continuous graph layout algorithm for handy network visualization

designed for the Gephi software". In: *PLoS ONE* 9.6 (2014), e98679 (cit. on p. 110).

[61] Jha, U. C., Bhat, J. S., Patil, B. S., Hossain, F., and Barh, D. "Functional Genomics: Applications in Plant Science". In: *PlantOmics: The Omics of Plant Science*. Ed. by Barh, D., Khan, M. S., and Davies, E. Springer India, 2015, pp. 65–111. ISBN: 978-81-322-2172-2 (cit. on p. 14).

[62] Johnson, W. E. and Nazaire, M. D. *ComBat (v3): Combatting batch effects when combining batches of microarray data*. [Online; accessed 29-February-2016] (cit. on p. 66).

[63] Jones, E., Oliphant, T.and Peterson P., et al. *SciPy: Open source scientific tools for Python*. [Online; accessed 14-January-2016]. 2001– (cit. on p. 81).

[64] Kanehisa, M. and Goto, S. "KEGG: Kyoto Encyclopaedia of Genes and Genomes". In: *Nucl. Acids Res.* 28.1 (2000), pp. 27–30 (cit. on p. 86).

[65] Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. "Data, information, knowledge and principle: back to metabolism in KEGG". In: *Nucleic Acids Research* 42.D1 (2014), pp. D199–D205 (cit. on p. 86).

[66] Kanehisa Laboratories. *The KEGG Markup Language*. [Online; accessed 04-December-2016] (cit. on p. 90).

[67] Karp, P. D., Paley, S. M., Krummenacker, M., Latendresse, M., Dale, J. M., Lee, T. J., Kaipa, P., Gilham, F., Spaulding, A., Popescu, L., Altman, T., Paulsen, I., Keseler, I. M., and Caspi, R. "Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology". In: *Brief. Bioinformatics* 11.1 (2010), pp. 40–79 (cit. on p. 27).

[68] Kawashima, S., Katayama, T., Sato, Y., and Kanehisa, M. "KEGG API: A Web Service Using SOAP/WSDL to Access the KEGG System". In: *Genome Informatics* 14 (2003), pp. 673–674 (cit. on p. 115).

[69] Kazic, T. "Ten Simple Rules for Experiments' Provenance". In: *PLoS Comput. Biol.* 11.10 (2015), e1004384 (cit. on p. 76).

[70] Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. "The human genome browser at UCSC". In: *Genome Res.* 12.6 (2002), pp. 996–1006 (cit. on pp. 28, 29).

[71] Khatri, P., Sirota, M., and Butte, A. J. "Ten years of pathway analysis: current approaches and outstanding challenges." In: *PLoS Comput. Biol.* 8.2 (2012), e1002375. DOI: 10.1371/journal.pcbi.1002375 (cit. on p. 104).

[72] Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions". In: *Genome Biol.* 14.4 (2013), R36 (cit. on p. 66).

[73] Klingström, T., Mendy, M., Meunier, D., Berger, A., Reichel, J., Christoffels, A., Bendou, H., Swanepoe, C., Smit, L., Mckellar-Basset, C., Bongcam-Rudloff, E., Söderberg, J., Merino-Martinez, R., Amatya, S., Kihara, A., Kemp, S., Reihs, R., and Müller, H. "Supporting the development of biobanks in low and medium income countries". In: *2016 IST-Africa Week Conference*. 2016, pp. 1–10 (cit. on p. 153).

[74] Kozomara, A. and Griffiths-Jones, S. "miRBase: annotating high confidence microRNAs using deep sequencing data". In: *Nucleic Acids Res.* 42.Database issue (2014), pp. 68–73 (cit. on p. 98).

[75] Kuo, T. C., Tian, T. F., and Tseng, Y. J. "3Omics: a web-based systems biology tool for analysis, integration and visualization of human

transcriptomic, proteomic and metabolomic data". In: *BMC Syst Biol* 7 (2013), p. 64 (cit. on pp. 29, 78).

[76] Laibe, C. and Le Novere, N. "MIRIAM Resources: tools to generate and resolve robust cross-references in Systems Biology". In: *BMC Syst Biol* 1 (2007), p. 58 (cit. on p. 101).

[77] Laird, P. W. "Principles and challenges of genomewide DNA methylation analysis". In: *Nat. Rev. Genet.* 11.3 (2010), pp. 191–203 (cit. on p. 19).

[78] Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J., and Pavlidis, P. "Coexpression analysis of human genes across many microarray data sets". In: *Genome Res.* 14.6 (2004), pp. 1085–1094 (cit. on p. 25).

[79] Lin, K., Kools, H., Groot, P. J. de, Gavai, A. K., Basnet, R. K., Cheng, F., Wu, J., Wang, X., Lommen, A., Hooiveld, G. J., Bonnema, G., Visser, R. G., Muller, M. R., and Leunissen, J. A. "MADMAX - Management and analysis database for multiple omics experiments". In: *J Integr Bioinform* 8.2 (2011), p. 160 (cit. on p. 33).

[80] Linkurious. *Linkurious, visualize graph data easily*. [Online; accessed 11-December-2015]. 2015 (cit. on p. 85).

[81] Lodish, H., Berk, A., Zipursky, L., Matsudaira, P., Baltimore, D., and Darnell, J. *Molecular Cell Biology*. W. H. Freeman, 1999. ISBN: 071673706X (cit. on pp. 10, 12).

[82] Luo, W. and Brouwer, C. "Pathview: an R/Bioconductor package for pathway-based data integration and visualization". In: *Bioinformatics* 29.14 (2013), pp. 1830–1831 (cit. on p. 79).

[83] Lyne, R., Smith, R., Rutherford, K., Wakeling, M., Varley, A., Guillier, F., Janssens, H., Ji, W., Mclaren, P., North, P., Rana, D., Riley, T., Sullivan, J., Watkins, X., Woodbridge, M., Lilley, K., Russell, S., Ash-

burner, M., Mizuguchi, K., and Micklem, G. “FlyMine: an integrated database for Drosophila and Anopheles genomics”. In: *Genome Biol.* 8.7 (2007), R129 (cit. on p. 27).

[84] Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. “Entrez Gene: gene-centered information at NCBI”. In: *Nucleic Acids Res.* 39.Database issue (2011), pp. D52–57 (cit. on p. 101).

[85] Maglott, D. R., Katz, K. S., Sicotte, H., and Pruitt, K. D. “NCBI’s LocusLink and RefSeq”. In: *Nucleic Acids Res.* 28.1 (2000), pp. 126–128 (cit. on p. 101).

[86] Marcel Hellkamp. *Bottle: Python Web Framework*. [Online; accessed 14-January-2016] (cit. on p. 81).

[87] Mariette, J., Escudié, F., Allias, N., Salin, G., Noirot, C., Thomas, S., and Klopp, C. “NG6: Integrated next generation sequencing storage and processing environment”. In: *BMC Genomics* 13 (2012), p. 462 (cit. on pp. 33, 50).

[88] Medina, I., Salavert, F., Sanchez, R., Maria, A. de, Alonso, R., Escobar, P., Bleda, M., and Dopazo, J. “Genome Maps, a new generation genome browser”. In: *Nucleic Acids Res.* 41.Web Server issue (2013), W41–46 (cit. on p. 28).

[89] Microsoft MSDN Library. *Data Transfer Object*. [Online; accessed 21-October-2015]. 2010 (cit. on p. 38).

[90] Miko, I. and LeJeune, L. *Essentials of Cell Biology*. [Online; accessed 29-October-2016]. 2009 (cit. on p. 10).

[91] Mohammad, F., Flight, R. M., Harrison, B. J., Petruska, J. C., and Rouchka, E. C. “AbsIDconvert: an absolute approach for converting genetic identifiers at different granularities”. In: *BMC Bioinformatics* 13 (2012), p. 229 (cit. on p. 101).

[92] MongoDB, Inc. *MongoDB for GIANT Ideas*. [Online; accessed 11-December-2015] (cit. on p. 80).

[93] MongoDB, Inc. *PyMongo, interacting with MongoDB database from Python*. [Online; accessed 11-December-2015] (cit. on p. 81).

[94] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. "Mapping and quantifying mammalian transcriptomes by RNA-Seq". In: *Nat. Methods* 5.7 (2008), pp. 621–628 (cit. on p. 16).

[95] Nardini, C., Dent, J., and Tieri, P. "Editorial: Multi-omic Data Integration". In: *Frontiers in Cell and Developmental Biology* 3.46 (2015) (cit. on p. 54).

[96] Ng, A., Bursteinas, B., Gao, Q., Mollison, E., and Zvelebil, M. "Resources for integrative systems biology: from data through databases to networks and dynamic system models". In: *Brief. Bioinformatics* 7.4 (2006), pp. 318–330 (cit. on p. 27).

[97] Nicol, J. W., Helt, G. A., Blanchard, S. G., Raja, A., and Loraine, A. E. "The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets". In: *Bioinformatics* 25.20 (2009), pp. 2730–2731 (cit. on p. 79).

[98] Nueda, M. J., Sebastian, P., Tarazona, S., Garcia-Garcia, F., Dopazo, J., Ferrer, A., and Conesa, A. "Functional assessment of time course microarray data". In: *BMC Bioinformatics* 10 Suppl 6 (2009), S9 (cit. on p. 110).

[99] O'Connor, C. M. and Adams, J. U. *Essentials of Cell Biology, Unit 1: What Is a Cell? What Are the Essential Characteristics of Cells?* [Online; accessed 29-October-2016]. 2010 (cit. on p. 8).

[100] O'Neill, L. A., Kishton, R. J., and Rathmell, J. "A guide to immunometabolism for immunologists". In: *Nat. Rev. Immunol.* 16.9 (Sept. 2016), pp. 553–565 (cit. on p. 141).

[101] Oracle Corporation. *Oracle Technology Network: Java Servlet Technology Overview*. [Online; accessed 21-October-2015] (cit. on p. 36).

[102] Osmani, A. *Learning JavaScript Design Patterns*. O'Reilly Media, 2012. ISBN: 1449331815 (cit. on p. 40).

[103] Park, P. J. "ChIP-seq: advantages and challenges of a maturing technology". In: *Nat. Rev. Genet.* 10.10 (2009), pp. 669–680 (cit. on p. 18).

[104] Pavlopoulos, G. A., Wegener, A. L., and Schneider, R. "A survey of visualization tools for biological network analysis". In: *BioData Min* 1 (2008), p. 12 (cit. on pp. 77, 78).

[105] Perry, B. *Java servlet and JSP cookbook*. Sebastopol, Calif: O'Reilly, 2004. ISBN: 978-0-596-00572-6 (cit. on p. 36).

[106] Phillips, T. "Regulation of transcription and gene expression in eukaryotes". In: *Nature Education* 1.1 (2008), p. 199 (cit. on p. 13).

[107] Python Software Fundation. *Python Language Reference, version 2.7.* [Online; accessed 14-January-2016]. Python Software Fundation (cit. on p. 80).

[108] R Core Team. *R: A Language and Environment for Statistical Computing*. [Online; accessed 14-January-2016]. R Foundation for Statistical Computing (cit. on p. 80).

[109] Rajasundaram, D. and Selbig, J. "More effort - more results: recent advances in integrative 'omics' data analysis". In: *Curr. Opin. Plant Biol.* 30 (2016), pp. 57–61 (cit. on pp. 23, 24).

[110] Reference, Genetics Home. *Help Me Understand Genetics*. [Online; accessed 29-October-2016]. 2016 (cit. on pp. 10, 12).

[111] Rohn, H., Junker, A., Hartmann, A., Grafahrend-Belau, E., Treutler, H., Klapperstuck, M., Czauderna, T., Klukas, C., and Schreiber, F. "VANTED v2: a framework for systems biology applications". In: *BMC Syst Biol* 6 (2012), p. 139 (cit. on p. 29).

[112] Romero, P., Wagg, J., Green, M. L., Kaiser, D., Krummenacker, M., and Karp, P. D. "Computational prediction of human metabolic pathways from the complete human genome". In: *Genome Biol.* 6.1 (2005), R2 (cit. on p. 27).

[113] Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T. R., Fujita, P. A., Guruvadoo, L., Haeussler, M., Harte, R. A., Heitner, S., Hickey, G., Hinrichs, A. S., Hubley, R., Karolchik, D., Learned, K., Lee, B. T., Li, C. H., Miga, K. H., Nguyen, N., Paten, B., Raney, B. J., Smit, A. F., Speir, M. L., Zweig, A. S., Haussler, D., Kuhn, R. M., and Kent, W. J. "The UCSC Genome Browser database: 2015 update". In: *Nucleic Acids Res.* 43.Database issue (2015), pp. D670–681 (cit. on p. 26).

[114] Rousseeuw, P. J. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53 –65 (cit. on p. 110).

[115] Schmid, N., Pessi, G., Deng, Y., Aguilar, C., Carlier, A. L., Grunau, A., Omasits, U., Zhang, L. H., Ahrens, C. H., and Eberl, L. "The AHL- and BDSF-dependent quorum sensing systems control specific and overlapping sets of genes in Burkholderia cenocepacia H111". In: *PLoS ONE* 7.11 (2012), e49966 (cit. on p. 31).

[116] Scholtalbers, J., Rößler, J., Sorn, P., Graaf, J. de, Boisguérin, V., Castle, J., and Sahin, U. "Galaxy LIMS for next-generation sequencing". In: *Bioinformatics* 29.9 (2013), pp. 1233–1234 (cit. on pp. 33, 50).

[117] Sencha Inc. *ExtJS, Client-side JavaScript Framework*. [Online; accessed 11-December-2015] (cit. on pp. 38, 84).

[118] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. "Cytoscape: a software environment for integrated models of biomolecular interaction networks". In: *Genome Res.* 13.11 (2003), pp. 2498–2504 (cit. on pp. 29, 78).

[119] SigmaJS. *SigmaJS, JavaScript library dedicated to graph drawing*. [Online; accessed 11-December-2015]. 2015 (cit. on p. 85).

[120] Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., and Kasprzyk, A. "BioMart–biological queries made easy". In: *BMC Genomics* 10 (2009), p. 22 (cit. on p. 27).

[121] Smith, R. N., Aleksic, J., Butano, D., Carr, A., Contrino, S., Hu, F., Lyne, M., Lyne, R., Kalderimis, A., Rutherford, K., Stepan, R., Sullivan, J., Wakeling, M., Watkins, X., and Micklem, G. "InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data". In: *Bioinformatics* 28.23 (2012), pp. 3163–3165 (cit. on p. 27).

[122] Song, C. X., Szulwach, K. E., Dai, Q., Fu, Y., Mao, S. Q., Lin, L., Street, C., Li, Y., Poidevin, M., Wu, H., Gao, J., Liu, P., Li, L., Xu, G. L., Jin, P., and He, C. "Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming". In: *Cell* 153.3 (2013), pp. 678–691 (cit. on p. 31).

[123] SQLAlchemy. *SQLAlchemy, The Database toolkit fot Python*. [Online; accessed 14-January-2016] (cit. on p. 81).

[124] Suhre, K. and Schmitt-Kopplin, P. "MassTRIX: mass translator into pathways". In: *Nucleic Acids Res.* 36.Web Server issue (2008), W481–484 (cit. on p. 30).

[125] Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W., Fiehn, O., Goodacre, R., Griffin, J. L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A. N., Lindon, J. C., Marriott, P., Nicholls, A. W., Reily, M. D., Thaden, J. J., and Viant, M. R. "Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI)". In: *Metabolomics* 3.3 (2007), pp. 211–221 (cit. on p. 43).

[126] Sun Microsystems Inc. *Core J2EE Patterns - Data Access Objects*. [Online; accessed 21-October-2015]. 2002 (cit. on p. 38).

[127] SVGJS. *SVG.JS, a lightweight library for manipulating and animating SVG.* [Online; accessed 11-December-2015]. 2015 (cit. on p. 85).

[128] Tarazona, S., Hernández-de-Diego, R., Silberberg, G., Ferreiros, I., van del Kloet, F., Ramirez, R., Schmidt, A., Marabita, F., Lagani, V., Papoutsoglou, G., Hankemeier, T., Westernhuis, J., Imhof, A., Ballestar, E., Meier, D., Lappe, M., Tsamardinos, I., Mortazavi, A., Merkenschlager, M., Tenger, J., Gomez-Cabrero, D., and Conesa, A. In: () (cit. on p. 131).

[129] Taylor, C. F., Paton, N. W., Lilley, K. S., Binz, P. A., Julian, R. K., Jones, A. R., Zhu, W., Apweiler, R., Aebersold, R., Deutsch, E. W., Dunn, M. J., Heck, A. J., Leitner, A., Macht, M., Mann, M., Martens, L., Neubert, T. A., Patterson, S. D., Ping, P., Seymour, S. L., Souda, P., Tsugita, A., Vandekerckhove, J., Vondriska, T. M., Whitelegge, J. P., Wilkins, M. R., Xenarios, I., Yates, J. R., and Hermjakob, H. "The minimum information about a proteomics experiment (MIAPE)". In: *Nat. Biotechnol.* 25.8 (2007), pp. 887–893 (cit. on pp. 27, 43, 72, 73, 75).

[130] The Django Software Foundation. *Django, The web framework for perfectionists with deadlines*. [Online; accessed 14-January-2016] (cit. on p. 81).

[131] The Functional Genomics Data Society. *MINSEQE: Minimum Information about a high throughput Nucleotide SeQuencing Experiment - a proposal for standards in functional genomic data reporting*. [Online; accessed 21-October-2015]. 2012 (cit. on pp. 27, 43).

[132] The jQuery Foundation. *jQuery: The Write Less, Do More, JavaScript Library*. [Online; accessed 11-December-2015] (cit. on pp. 38, 85).

[133] The National Human Genome Research Institute. *Fact Sheets on Science, Research, Ethics and the Institute: Epigenomics*. [Online; accessed 08-November-2016] (cit. on p. 18).

[134] The STATegra Project Consortium. *STATegra, User-driven Development of Statistical Methods for Experimental Planning, Data Gathering, and Integrative Analysis of Next Generation Sequencing, Proteomics and Metabolomics*. [Online; accessed 02-February-2015]. 2012 (cit. on pp. 53–57, 66, 121, 123, 140, 151, 153, 161).

[135] Thimm, O., Blasing, O., Gibon, Y., Nagel, A., Meyer, S., Kruger, P., Selbig, J., Muller, L. A., Rhee, S. Y., and Stitt, M. "MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes". In: *Plant J.* 37.6 (2004), pp. 914–939 (cit. on pp. 30, 78).

[136] Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration". In: *Brief. Bioinformatics* 14.2 (2013), pp. 178–192 (cit. on pp. 28, 79).

[137] Title=. In: () (cit. on p. 98).

[138] Tokimatsu, T., Sakurai, N., Suzuki, H., Ohta, H., Nishitani, K., Koyama, T., Umezawa, T., Misawa, N., Saito, K., and Shibata, D. "KaPPA-view: a web-based analysis tool for integration of transcript and metabolite

data on plant metabolic pathway maps". In: *Plant Physiol.* 138.3 (2005), pp. 1289–1300 (cit. on pp. 30, 78).

[139] Troshin, P. V., Postis, V. L., Ashworth, D., Baldwin, S. A., McPherson, M. J., and Barton, G. J. "PIMS sequencing extension: a laboratory information management system for DNA sequencing facilities". In: *BMC Res Notes* 4 (2011), p. 48 (cit. on p. 33).

[140] Tseng, G. C., Ghosh, D., and Feingold, E. "Comprehensive literature review and statistical considerations for microarray meta-analysis". In: *Nucleic Acids Res.* 40.9 (2012), pp. 3785–3799 (cit. on p. 25).

[141] Tuncbag, N., McCallum, S., Huang, S. S., and Fraenkel, E. "SteinerNet: a web server for integrating 'omic' data to discover hidden components of response pathways". In: *Nucleic Acids Res.* 40.Web Server issue (2012), W505–509 (cit. on p. 29).

[142] Tyanova, S., Temu, T., Carlson, A., Sinitcyn, P., Mann, M., and Cox, J. "Visualization of LC-MS/MS proteomics data in MaxQuant". In: *Proteomics* 15.8 (2015), pp. 1453–1456 (cit. on p. 72).

[143] Vallon-Christersson, J., Nordborg, N., Svensson, M., and Häkkinen, J. "BASE–2nd generation software for microarray data management and analysis". In: *BMC Bioinformatics* 10 (2009), p. 330 (cit. on p. 32).

[144] Van Regenmortel, M. H. "Reductionism and complexity in molecular biology. Scientists now have the tools to unravel biological and overcome the limitations of reductionism". In: *EMBO Rep.* 5.11 (2004), pp. 1016–1020 (cit. on p. 21).

[145] Van Rossum, T., Tripp, B., and Daley, D. "SLIMS–a user-friendly sample operations and inventory management system for genotyping labs". In: *Bioinformatics* 26.14 (2010), pp. 1808–1810 (cit. on p. 32).

[146] Venco, F., Vaskin, Y., Ceol, A., and Muller, H. "SMITH: a LIMS for handling next-generation sequencing workflows". In: *BMC Bioinformatics* 15 Suppl 14 (2014), S3 (cit. on p. 76).

[147] Villaveces, J. M., Koti, P., and Habermann, B. H. "Tools for visualization and analysis of molecular networks, pathways, and -omics data". In: *Adv Appl Bioinform Chem* 8 (2015), pp. 11–22 (cit. on p. 78).

[148] Waegele, B., Dunger-Kaltenbach, I., Fobo, G., Montrone, C., Mewes, H. W., and Ruepp, A. "CRONOS: the cross-reference navigation server". In: *Bioinformatics* 25.1 (2009), pp. 141–143 (cit. on p. 101).

[149] Wang, L., Wang, G., and Alexander, C. "Big Data and Visualization: Methods, Challenges and Technology Progress". In: *Digital Technologies* 1.1 (2015), pp. 33–38 (cit. on p. 147).

[150] Wang, R., Dillon, C. P., Shi, L. Z., Milasta, S., Carter, R., Finkelstein, D., McCormick, L. L., Fitzgerald, P., Chi, H., Munger, J., and Green, D. R. "The transcription factor Myc controls metabolic reprogramming upon T lymphocyte activation". In: *Immunity* 35.6 (2011), pp. 871–882 (cit. on p. 141).

[151] Wang, Z., Gerstein, M., and Snyder, M. "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nat. Rev. Genet.* 10.1 (2009), pp. 57–63 (cit. on p. 15).

[152] Wei, G., Abraham, B. J., Yagi, R., Jothi, R., Cui, K., Sharma, S., Narlikar, L., Northrup, D. L., Tang, Q., Paul, W. E., Zhu, J., and Zhao, K. "Genome-wide analyses of transcription factor GATA3-mediated gene regulation in distinct T cell types". In: *Immunity* 35.2 (2011), pp. 299–311 (cit. on p. 31).

[153] Weinstein, J. N. et al. "The Cancer Genome Atlas Pan-Cancer analysis project". In: *Nat. Genet.* 45.10 (2013), pp. 1113–1120 (cit. on p. 1).

[154] Wirapati, P., Sotiriou, C., Kunkel, S., Farmer, P., Pradervand, S., Haibe-Kains, B., Desmedt, C., Ignatiadis, M., Sengstag, T., Schutz, F., Goldstein, D. R., Piccart, M., and Delorenzi, M. "Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures". In: *Breast Cancer Res.* 10.4 (2008), R65 (cit. on p. 25).

[155] Witkos, T. M., Koscianska, E., and Krzyzosiak, W. J. "Practical Aspects of microRNA Target Prediction". In: *Curr. Mol. Med.* 11.2 (2011), pp. 93–109 (cit. on p. 98).

[156] Xia, J., Lyle, N. H., Mayer, M. L., Pena, O. M., and Hancock, R. E. "INVEX–a web-based tool for integrative visualization of expression data". In: *Bioinformatics* 29.24 (2013), pp. 3232–3234 (cit. on pp. 25, 77).

[157] Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., Giron, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Keenan, S., Lavidas, I., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Nuhn, M., Parker, A., Patricio, M., Pignatelli, M., Rahtz, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Birney, E., Harrow, J., Muffato, M., Perry, E., Ruffier, M., Spudich, G., Trevanion, S. J., Cunningham, F., Aken, B. L., Zerbino, D. R., and Flicek, P. "Ensembl 2016". In: *Nucleic Acids Res.* 44.D1 (2016), pp. D710–716 (cit. on pp. 26, 28, 126).

[158] Zhang, Z., Bajic, B., J., Yu, Cheung, K. H., and Townsend, J. P. "Data Integration in Bioinformatics: Current Efforts and Challenges". In: *Bioinformatics - Trends and Methodologies*. Ed. by Mahdavi, M. A. InTech, 2011 (cit. on p. 26).

[159] Zhu, Y., Sun, L., Garbarino, A., Schmidt, C., Fang, J., and Chen, J. "PathRings: a web-based tool for exploration of ortholog and expression

data in biological pathways". In: *BMC Bioinformatics* 16 (2015), p. 165 (cit. on p. 111).

Development of Bioinformatics Resources for the Integrative Analysis of Next Generation Omics Data

Rafael Hernández de Diego

In recent years Systems Biology has established itself as a multidisciplinary area of research which tries to model the dynamic behaviour of biological systems by holistically studying the interactions between the different types of molecules that are essential for life, including DNA, RNA, proteins, and metabolites. Systems Biology is an interdisciplinary area that requires biologists, mathematicians, biochemists, and other researchers to work closely together, and in which computer sciences plays a fundamental role because of the volume and complexity of the data handled.

This thesis addresses the problem of data management, integration, and analysis in multi-omics studies. More specifically, this research focused on two of the most characteristic computational challenges in Systems Biology: the development of integrated databases and the problem of integrative visualisation.

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

EDITORIAL