Universidade do Minho
Escola de Engenharia

2018

Bioinformatics Open Days (BOD)

Conference Book

14-03-2018 to 16-03-2018

Contents

Bioinformatics Open Days (BOD)

In Portugal, Bioinformatics has experienced an outstanding growth over the past few years. This is reflected academically, through the development of prestigious post-graduations, and in the economical/business sector with the establishment of new start-up companies with international connections.

Bioinformatics Open Days is a student-led initiative, first held at University of Minho, Braga, in 2012. It aims to promote the exchange of knowledge between students, teachers and researchers from the Bioinformatics and Computational Biology fields. This symposium's 7th edition is a joint collaboration with the Bioinformatics and Computational Biology Master students of University of Minho and Faculty of Sciences of the University of Lisbon. This event will occur at University of Minho (Campus Gualtar) on the 14th, 15th, and 16th of March.

Organization

General chairman

Miguel Rocha

Organizing committee chairmen

Ana Monteiro[1]
Ana Ramos[1]
André Fonseca[1]
Amaro Morais[1]
Beatriz Magalhães[1]
Catarina Sousa[1]
Célia Domingues[2]
Cláudio Nóvoa[1]
Daniel Martins[1]
Daniela Pereira[1]
Davide Lagoa[1]
Francisco Santos[2]
Francisco Paupério[2]
Gisela Alves[1]
João Correia[1]
João Fernandes[1]
João Mendes[1]
João Rebelo[1]
José Bastos[1]
Jorge Gomes[1]
Marta Moreno[1]
Marta Sampaio[1]
Pedro Queirós[1]
Renato Cruz[1]
Sara Pereira[1]
Tiago Isabelinho[1]
Tiago Reis[1]

[1] University of Minho

[2] University of Lisbon

Program

14th March, Wednesday — Workshops Day

HOURS	WORKSHOPS
14H00-17H00	PhenoMeNal e-intrastructure Pablo Moreno
14H00-17H00	Understanding microbial dynamics and interventions with mathematical models Erida Gjini

15th March, Thursday

HOURS	
9H30-10H00	Check in
10H00-10H30	Opening Session
10H30-11H15	Lecture Emanuel Gonçalves
11H15-11H45	Coffee Break
11H45-13H00	Oral Presentations — Session 1
13H00-14H30	Lunch
14H30-15H15	Lecture Dina Ruano
15H15-16H15	Highlights Posters
16H15-17H00	Poster Session | Coffee Break
17H00-18H30	Oral Presentations — Session 2
19H00	Social Meeting (Sponsored by Coimbra Genomics)
20H30	Social Dinner

16th March, Friday

HOURS	
10H30-11H15	Lecture Francesca Ciccarelli
11H15-13H00	Oral Presentations — Session 3
13H00-14H30	Lunch
14H30-15:15	Lecture José Leal
15H15-16H00	Oral Presentations — Session 4
16H00-16H45	Coimbra Genomics
16H45-17H15	Round Table
16H15-17H45	Closing Session

Invited Speakers

Emanuel Gonçalves

Emanuel is a computational biologist within the Cancer Genomics group. He is mostly focused on the analysis of high-throughput genomic and pharmacological screens to unravel vulnerabilities in cancer cells that can then be translated into potential novel therapies.

Dina Ruano

Dr. Dina Ruano is a computational biologist working in cancer research and diagnostics at the Leiden University Medical Center since 2009. She has ample experience in variation analysis of tumor tissue, using whole-exome, and target sequencing among others. Recently, Dina's research focussed on detection of neoantigens. Dina obtained her Ph.D. degree from the Universidade do Minho in 2007

Francesca Ciccarelli

Dr. Francesca Ciccarelli coordinates the quantitative Genomics, Epigenomics and Biobank Programme in the Division of Cancer Studies at King's College (London). Her research work aims to understand the role of gene alterations in the development of cancer. Her group develops and applies a combination of computational and experimental methods to study cancer genes in the context of gene and network evolution. They analyse "omic" data deriving directly from cancer patients to identify patient-specific cancer genes and vulnerabilities that can be exploited in therapy. The group works with data of various types of cancer, but has a particular focus on colorectal cancer and oesophageal cancer. They set up and maintain the Network of Cancer Genes, a public resource of manually curated cancer genes.

José Leal

José Pereira Leal holds a PhD in Biomedical Sciences and has worked internationally in the field of Bioinformatics for the past 20 years. As a PhD student and then as a post-doctoral fellow, he worked as a researcher at Imperial College School of Medicine (London, UK), the Southwestern Medical School in Dallas (Texas, USA), the EMBL European Bioinformatics Institute (Cambridge, UK) and at the MRCl Laboratory of Molecular Biology (Cambridge, UK). After 10 years abroad, he returned to Portugal where he founded, and presently coordinates, the Computational Genomics Laboratory at the Instituto Gulbenkian de Ciência, supervising a research program in bioinformatics, comparative and medical genomics, and data integration. He also coordinated the Bioinformatics Unit of the Instituto Gulbenkian de Ciência, where he helped to establish the European biological data infrastructure ELIXIR, and its national counterpart, the Portuguese biological data infrastructure BioData.pt.

Lectures

Structural rearrangements determine CRISPR/Cas9 efficacy in cancer

Emanuel Gonçalves

CRISPR/Cas9 is now a mainstream mechanism to selectively target genes in cancer cells, and recent applications have demonstrated its potential as a tool for precision medicine. To that end, it is important to understand the determinants of CRISPR/Cas9 targeting and efficacy, in particular when considering cancer cells which undergo extensive genetic alterations. Here, we take advantage of one of the largest CRISPR/Cas9 screens available to date, comprising ~200 cell lines, and use it to identify chromosomal alterations that might affect the efficacy of CRISPR/Cas9 in cancer cells. CRISPR/Cas9 has many promising therapeutic applications and here we describe the importance of previously unstudied determinants of CRISPR/Cas9 targeting. Furthermore, we show how this could be used to synthetically target tumours, especially those enriched for structural rearrangements.

Systems biology to rebuild cancer evolution and identify new cancer drivers

Francesca Ciccarelli

Large-scale cancer genome projects in principle provide an extraordinary mine of molecular information on a vast range of cancer types and samples and offer the exciting potential of understanding the molecular mechanisms of cancer. Much knowledge is however still hidden in the data because of the current limitations in correctly analysing and modelling the data. This significantly reduces the effective contribution of cancer molecular profiling to the personalised medicine agenda. In my talk, I will review some of the technical, analytical and scientific challenges in cancer genomic data analysis. I will focus specifically on the identification of rare and sample-specific driver genes, which currently remain mostly undiscovered despite they are likely to contribute for a substantial fraction of cancer genes. I will illustrate

how we can use this information to identify cancer associated vulnerabilities and model cancer evolution.

Improving neoantigen detection in cancer

Dina Ruano

Tumor-specific mutations that result in aberrant peptides can lead to an effective anti-tumor immunity in humans. These peptides, known as neoantigens, are optimal targets for the anti-tumor immune response because they are not present in the normal tissue.

Vaccination with patient-specific neoantigens has been successfully applied in the treatment of cancers like melanoma and lung cancer, cancers know to have many different tumor-specific mutations. The success of such a strategy in cancers with fewer mutations have been hampered by a suboptimal identification of candidate neoantigens.

In this talk, we will cover challenges faced in the detection of neoantigens and discuss possible solutions.

Workshop Lecturers

Pablo Moreno

Pablo is the Senior Bioinformatician for the PhenoMeNal project, within the Cheminformatics & Metabolism group at EMBL-EBI. After his PhD and Post-doc in Bioinformatics with Christoph Steinbeck at EMBL-EBI, which dealt with the inference of complete Metabolomes based on multiple sources of data and various aspects related to metabolism representation, Pablo was appointed to lead the Bioinformatics Core facility at the Cambridge Institute for Medical Research (CIMR), dependent of the University of Cambridge. Two years later, Pablo came back to EMBL-EBI to work on PhenoMeNal.

Erida Gjini

Erida Gjini works at the interface between mathematics and biological sciences. She uses mathematics to answer biological questions, and biological questions as inspiration for new mathematical tools. Her primary interest lies in uncovering the mechanisms of pathogen population dynamics within and between hosts, with special focus on host and pathogen diversity. She develops mathematical and computational frameworks to understand intervention effects in multi-type pathogen communities, and to quantify the conditions for stability of coexistence in microbial ecosystems.

Workshops

PhenoMeNal e-infrastructure

Pablo Moreno

PhenoMeNal is a cloud deployable e-infrastucture for metabolomics data analysis. Designed to be completely cloud native, PhenoMeNal relies heavily on containers and container orchestration to provide the user the ability to do metabolomics data analysis through user friendly workflow environment systems. In this short workshop, we will give an overview of the PhenoMeNal project and architecture, explaining all the layers from bare bones cloud provider virtual hardware to an actual working deployment; show the PhenoMeNal portal, which easily allows users to deploy our cloud research environment to Amazon AWS, Google cloud or OpenStack; demonstrate how we package tools for PhenoMeNal and follow a short tutorial on assembling a workflow in the PhenoMeNal Galaxy deployment.

Understanding microbial dynamics and interventions with mathematical models

Erida Gjini

The workshop will focus on mathematical modeling as a powerful tool to describe and understand microbial system dynamics and interventions. There will presented two studies from a research in infectious diseases at the within- and between-host level. The first study regards antibiotic resistance and treatment optimization for bacterial infections, quantifying the crucial role of host immune defenses. The second study presents a multi-strain epidemiological model applied to pneumococcus data before and after vaccination. With a deeper quantitative understanding of infection processes across biological scales, it enables to propose better ways of designing and implementing interventions.

Companies

Coimbra Genomics

Bringing genomic knowledge into everyday clinical practice

Coimbra Genomics (www.coimbra-genomics.com) has developed and is commercializing ELSIE, a first-in-class digital platform that allows physicians of any speciality to use information on their patient's genome in an easy, fast and secure way, to make individualized decisions about diagnosis or prognosis during a regular medical appointment. By combining the capacity to provide fast, accurate, on-the-spot answers based on each patient's genome with an efficient articulation with multiple providers of specific genetic tests, ELSIE is the first simple, highly scalable and completely secure way to perform personalized medicine.

ELSIE taps into the multi-billion and fast-growing markets of both genetic testing and personalized diagnostics and therapeutics. An enormous opportunity lies in delivering genetic/genomic information to physicians in a simple, convenient and secure way, through a single channel designed in a friendly format, allowing them to make more informed, precise and personalized decisions. This will produce an enormous impact on patients' quality of life while also sharply cutting healthcare costs.

ELSIE is that channel. ELSIE can be used by any physician, regardless of previous knowledge in genetics, without stopping patient flow or requiring extensive learning on the spot. It articulates with third-party providers of specific genetic tests and gives physicians a simpler, unified ordering workflow. ELSIE's innovative and proprietary bioinformatics pipeline allows answers to be produced very quickly, with minimum direct human intervention, while the data security and privacy policies adopted ensure full data anonymity but still allow personalized answers.

ELSIE brings together patients, physicians, geneticists and genetic labs to create a symbiotic network composed of the key players necessary to make genomic based medicine truly part of everyday medical practice.

Oral presentations

Session 1

Annotator: a novel custom tool for genomic variants annotation and classification

D. Lemos, P. Oliveira, C. São José, J. Carvalho, C. Oliveira

Genome sequencing produces large amounts of data that enable the discovery of genomic variants. Using appropriate bioinformatics tools and public databases which aggregate clinical information, it is possible to understand the functional impact of such variants on phenotypic traits. Several bioinformatics tools already exist, however with limitations such as, the license costs and the limited choice of public databases consulted. Therefore, our aim was to implement a free tool that aggregates information collected from user-defined public databases, allowing the custom annotation of genomic variants and their classification in terms of pathogenicity. We implemented a tool, *Annotator*, which annotates and classifies genomic variants using data from 5 renowned public databases (Uniprot, OMIM, ClinVar, dbSNP, Pubmed). *Annotator* can also integrate keywords, defined according to sample characteristics, for advanced data mining, in order to refine collected data and variant classification. As validation, *Annotator* was used to analyse and classify a set of 151 genomic variants, detected in 52 probands with Familial Intestinal Gastric Cancer. These genomic variants had already been analysed and classified using a commercial software. The comparison of the results obtained with both analyses showed that *Annotator* was able to add relevant information for 5/42 somatic variants and 4/24 germline variants of unknown significance for the commercial software. In conclusion, *Annotator* is a valid tool for an accurate annotation and efficient classification of genomic variants derived from sequencing experiments.

Computational approaches in personalized medicine – ImmunoInformatics in gastric cancer

Bruno Cavadas, Nuno Fonseca, Rui Camacho, Luísa Pereira

Prediction of B- and T-cell epitopes is essential in the design of prophylactic and therapeutic vaccines that are currently being developed for personalized cancer

immunotherapies. Reduction in next generation cost have been responsible for the enrichment of immunological databases in such way that rendered the use of machine learning algorithms in epitope prediction more accurate. The efficiency of these algorithms depends highly on reliable pipelines for non-synonymous somatic variant discovery and on confirmation that the variant is translated in the tissue.

In this work, we developed and implemented a pipeline that: (1) couples four calling algorithms and infers consensus somatic mutations; (2) confirms that the set of non-synonymous somatic mutations are present in the expressed RNA-seq raw data; (3) characterizes the human leukocyte antigen (HLA) subtype of the individuals; (4) and predicts the personalized immunogenicity of variant somatic peptides given the HLA individual profile. We tested this pipeline in 441 gastric cancer cases from The Cancer Genome Atlas (TCGA) by whole-exome and RNASeq sequencing.

We confirmed that the consensus rule of mutation detection eliminates a high amount (approximately 47% of a total of 617000 mutations) of false variants. Given the high variance of the protein expression between tissues, in the specific case of stomach analyzed, step (2) decreased in 84% the amount of variants to be further investigated. For the HLA characterization, we confirmed the profile of the individuals was as expected given the geographic origin of the individuals analyzed. And, the main result, an amount of 5% somatic predicted T-cell epitopes (2586) could be used in personalized immunotherapy in 260 gastric cancer patients of the total 441 cohort.

Pathway expression optimization using the Ribosome Binding Site (RBS) Calculator tool

Joana L. Rodrigues, João Rainha, Lígia R. Rodrigues

Hydroxycinnamic acids and curcumin are plant metabolites with great therapeutic potential, including anti-inflammatory and anticancer activities. In this study, p-coumaric acid, caffeic acid and curcumin were produced in Escherichia coli using an artificial biosynthetic pathway. Their production was induced by heat using the dnaK and ibpA heat shock promoters. The ribosome binding sites (RBSs) used were tested and further optimized for each gene to assure an efficient translation. To optimize the RBSs we used the bioinformatic design tool RBS Calculator (v1.1) developed by Salis

Lab (Penn State University). This tool predicts the translation initiation rate (TIR) of mRNAs and designs synthetic RBS with specific TIRs. This allows to improve the translation efficiency and to reach a desired response and therefore obtain the expected production using novel genes or biosynthetic pathways.

Tyrosine ammonia lyase from Rhodotorula glutinis was used to produce p-coumaric acid from tyrosine. p-Coumaric acid was converted to caffeic acid using 4-coumarate 3- hydroxylase from Saccharothrix espanaensis or cytochrome P450 CYP199A2 from Rhodopseudomonas palustris. Curcumin was produced from ferulic acid using 4-coumarate-CoA ligase from Arabidopsis thaliana, diketide-CoA synthase and curcumin synthase from Curcuma longa. The optimization of the RBSs lead to an increase in the production of p-coumaric acid, caffeic acid and curcumin up to 97.8, 11.7 and 14.4 times, respectively. The highest p-coumaric acid, caffeic acid and curcumin production obtained were 2.5 mM, 370 μM and 17 μM, respectively. These results demonstrate that it is of utmost importance to consider the strength of the RBS when designing a biosynthetic pathway and user-friendly bioinformatic tools such as RBS Calculator can be very useful for that purpose.

Unveiling miRNA role on *Vitis* flowering, through transcriptomics

Miguel J. N. Ramos, João L. Coito, Maria M. R. Costa, Margarida Rocheta

RNA interference is a relevant mechanism in plants, animals and fungi. The RNA interference main functions are to defend cells against virus (small interference RNA) and to control, along with other mechanism, the gene expression, by promoting post-transcriptional gene silencing (micro-RNA; miRNA). miRNAs are originated from the transcription of MIRNA genes and subsequent processing of the respective non-coding RNA molecules produced. The presence of miRNAs in inflorescences of *Vitis vinifera* was already demonstrated in previous studies. Successful fruit production depends on flower production and fertilization and is a crucial step for the wine industry. However, the mechanisms that control *V. vinifera* flowering development are still obscure. One of the most intriguing mechanisms that are being studied refers to the sexuality differences between the commercial (*V. v. vinifera*) and wild-type (*V. v. sylvestris*) grapevines: the commercial varieties are hermaphrodite and self-fruitful whereas the

wild type is a dioecious subspecies. During floral development, the mechanisms controlling gene expression are determinant, which may explain the miRNAs identified previously.

In this ongoing study, sequences of transcripts that are under-represented on different flower transcriptomes (four developmental stages of male, female and hermaphrodite flower types), along with de novo identified transcripts, were submitted to miRNA databases (miRBASE and PMRD). The results suggest that 93 different sequences may be under the control of 72 different miRNAs (from 38 families). Also, there are 8 miRNA whose families were identified in *Vitis* for the first time by this study. Further studies under this project are currently being performed in order to validate the results obtained and to assess the role that these miRNAs may have on flower development.

Session 2

PLASMID ATLAS: the hitchhiker's guide to the plasmid galaxy

Jesus TF, Ribeiro-Gonçalves B, Silva DN, Valeria B, Ramirez M, Carriço JA

The identification of plasmids from bacterial high throughput sequencing (HTS) short read data is not trivial, mainly due to their modular and chimeric nature which interferes with the reconstruction of plasmid sequences. Often, HTS analysis of individual strains or even of metagenomic samples are performed with short read length technologies, which further complicates this issue.

We developed an online tool (pATLAS) for the visualization and exploration of the metadata associated with all plasmids available in NCBI's refseq database. Additionally, each plasmid in pATLAS was annotated for antibiotic resistance and plasmid families using CARD, ResFinder and PlasmidFinder databases. A matrix of genetic distances between plasmid sequences was created using MASH software, and all the significant similarities between plasmids were used to construct a network. The visualization of this network is done through a force-directed graph, where each plasmid is a node and their relationships the connecting links.

With the goal of offering users a visual analytics tool to explore the existing plasmid database, pATLAS allows searches by plasmid name, taxa, resistance genes, plasmid families and sequence length.

In order to identify plasmids from users short read data, a set of scripts is provided for mapping the reads against the whole pATLAS database using Bowtie2 or to perform the search using a computational efficient algorithm named MASH. The scripts produce the results in a JSON format that can then be visualized in pATLAS website. The users can also add new plasmid sequences in Fasta format into the plasmid network. The correct and fast identification of plasmids in HTS data within a user-friendly environment, allowing the user to explore related plasmids, is of particular importance given the role of plasmids in horizontal gene transfer and antibiotic resistance.

pATLAS is freely available at www.patlas.site and its source code hosted at https://github.com/tiagofilipe12/pATLAS.

A mapping methodology for the characterization of modular genes

Catarina I. Mendes, Miguel P. Machado, José Melo-Cristino, Mário Ramirez, João A. Carriço

The presence of modularity allows for relatively small genomes, like bacteria, to obtain significant genetic variation, impacting phenotypic adaptation and genotypic diversity. The identification of the genes from high throughput sequencing (HTS) data underpinning this genetic variation is no easy task due to their modular conformation, often accompanied by the presence of large repeated areas within the gene, and that different alleles might be expressed in a single axonic culture. Therefore, the standard techniques of *de novo* assembly or mapping cannot be applied successfully.

We've developed a methodology for the characterization of modular genes and applied it to the the *hsdS* gene in *Strepcococcus pneumoniae*'s *ivr* locus, a six-phase variation type I restriction-modification systems composed of three polypeptides: R (restriction, encoded by *hsdR*), M (modification, encoded by *hsdM*), and S (specificity, encoded by *hsdS*). The *hsdS* contains two variable regions encoding the target recognition domains (TRD), and is usually accompanied by inverted silent units that, through recombination,

allow for phase variation switching events. We based our approach on the mapping of short-read paired-end HTS data to a conserved region in the expressed *hsdS*, using it as target and retrieving the read mates of the sequences that align to this region. By mapping these mates to the optional TRD, the first variable region can be determined, and is then used as the new target sequencing. Retrieving the mates of the sequences that align to this region and mapping them to the second optional TRD allows the determination of the other variable *hsdS* region. Steps of validation to assess the reliability of the results were implemented.

The *ivr* locus has been associated with epigenetic events that lead to different opacity phenotypes. Using ivrTyper and seroBA, we determined the dominantly expressed *hsdS* and serotype in 37,333 publicly available pneumococcal genomic data sets. Most of the isolates dominantly expressed alleles associated with the opaque phenotype shown to be more virulent in animal models, however, some serotypes, including the highly virulent serotypes 1 and 3, were associated with alleles responsible for a transparent phenotype.

ivrTyper is available at https://github.com/B-UMMI/ivrTyper.

Session 3

Inferring positive selection: ADOPS and B+ database

Cristina P. Vieira, Noé Vázquez, Miguel Reboiro-Jato, Hugo López-Fernández, Florentino Fdez-Riverola, Jorge Vieira

Amino acid changes in protein sequences can be adaptive, and when changes at few amino acid sites are the target of selection they can be detected using maximum likelihood methods based on models of codon substitution. This approach has been applied to infer positively selected amino acid sites at numerous proteins. The Automatic Detection of Positively Selected Sites (ADOPS) computer application allows the automated execution of all the steps needed to infer positively selected amino acid sites, starting from a FASTA file with non-aligned coding sequences. The batch option allows to run thousands of sequence files, and can now be made available at the B+ public database (http://bpositive.i3s.up.pt/), that has been designed to store and show

by a convenient interface the information contained in ADOPS project files. The availability of such a database ensures results repeatability, promotes data reuse with significant savings on the time needed for preparing datasets, and effortlessly allows further exploration of the data contained in ADOPS projects. Both large and small ADOPS datasets can be submitted to B+ (as compressed tar.gz files) along with a description containing the details about how the project was performed.

Lessons learned developing GUI in bioinformatics software: from end-user applications to resources for programmers

H. López-Fernández, M. Reboiro-Jato, Daniel Glez-Peña, Rosalía Laza, Reyes Pavón, Florentino Fdez-Riverola

Introduction: Software applications are an essential part of nowadays research in bioinformatics and computational biology. From the end-user point of view, these applications can have a Command Line Interface (CLI) and/or a Graphical User Interface (GUI), and both types of interfaces share the spectrum of publicly available bioinformatics applications. Nevertheless, GUI-based software are important since many life-scientists do not have advanced informatics skills to use the command line to conduct their research efficiently. When developing such GUI-based software, programmers can take advantage of data visualization libraries to show different representations to the users. However, there is a lack of libraries providing generic GUI components to help programmers developing such software. In this context, to ease the development of GUI for bioinformatics software, some years ago we developed the AIBench framework (http://www.aibench.org) and we are currently developing the GC4S library (http://www.sing-group.org/gc4s).

Results: Our group has been developing bioinformatics software for different areas for more than ten years. Our first experiences lead to the development of the AIBench framework in order to improve the productivity of software programmers by providing common functionalities present in scientific applications. This Java framework has been successfully used to develop end-user bioinformatics GUI applications such as Mass-Up, @Note or S2P, among others. A key factor to develop these GUI applications successfully was obtaining early feedback from actual users, thus increasing their

usability from the beginning. To complement the aforementioned framework, we are currently developing GC4S (GUI Components for Swing), a Java library providing a collection of extensible and reusable GUI components. These include different types of dialogs (including a Wizard assistant), components to retrieve user input (e.g. text fields, parameter configuration panels, etc.) and data visualization components (e.g. heat map, color legends, etc.).

Conclusions: Over the years, the experience developing different GUI-based bioinformatics software allowed us to identify key points for creating user-friendly, effective interfaces with success. In addition, this experience also lead to the development of two valuable resources for programmers: AIBench and GC4S.

Identification of Mutations in an Adaptive Laboratory Evolution Experiment for Butanol Producing Bacteria Strains of Clostridium Saccharoperbutylacetonicum

Hüseyin Demirci, Rafael F. Alves, Miguel Rocha, Oscar Dias, Isabel Rocha

Clostridium Saccharoperbutylacetonicum is known to be a butanol producing bacteria which makes it an attractive subject for biofuel studies. In this work, we have worked on the analysis of mutants of the strain N1-4 (HMT), which were derived from an adaptive laboratory evolution process. From this strain one wild type and 4 mutants were derived and sequenced where the mutants have the ability to grow and produce solvents under inhibitors which present in hemicellulosic hydrolysate obtained from sugarcane bagasse. We will briefly explain the NGS pipeline from raw data to variants that consists of the following steps: Quality Control, NGS quality and adaptor filtering, mapping sequencing reads to the reference genome, variant calling, visualization and analysis of the variants. We will point out the results obtained by the analysis of the mutants. For each mutation type, a few significant mutations have been observed where the mutation clones share some mutations. We will also mention the software and platforms used for the analysis. We will focus on the lessons learned from this sequencing experiment and analysis steps. This work can be a model for the design of adaptive laboratory evolution experiments. In the later steps of this work, metabolic models of the mutants will be obtained in order to investigate the mechanisms underlie the survival of the mutations under harsh environmental conditions.

Whole-genome analysis of daptomycin-resistant Staphylococcus aureus isolates

Vera Manageiro, Vanessa Salgueiro, Catarina Silva, Luís Vieira, Manuela Caniça

Background: Daptomycin (DAP) is a cyclic lipopeptide with in vitro activity against a variety of Gram-positive pathogens, including multidrug-resistant organisms. Although DAP-resistance is uncommon in clinical practice, development of this phenomenon during therapy has been widely described in clinically important organisms such as *Staphylococcus aureus*. Here, we performed whole-genome sequencing (WGS) to identify potential determinants of DAP-resistance in 3 *S. aureus* clinical strains.

Methods: Antibiotic susceptibility were determined by both disk diffusion and the microdilution technique. Interpretation of results was done according to the EUCAST clinical breakpoints. Genomic DNA was extracted using the MagNA Pure96 System (Roche), and quantified using Qubit 1.0 Fluorometer (Invitrogen). The Nextera XT DNA Sample Preparation Kit (Illumina) was used to prepare sequencing libraries from 1ng of genomic DNA according to the manufacturer's instructions. WGS was performed using 150 bp paired-end reads on a MiSeq (Illumina). Sequence reads were trimmed and filtered according to quality criteria, and de novo assembled into contigs by means of CLC Genomics Workbench 10.0 (Qiagen). *In silico* phenotyping and molecular typing (*agr*- typing, *spa*-typing and MLST) was performed using web-based bioinformatics tools.

Results: One S. aureus was susceptible to all antibiotics tested, except to DAP, and two presented also resistance to ciprofloxacin and cefoxitin being methicillin-resistant *S. aureus* (MRSA). We identified heterogeneity of lineages, such as ST22 (t2357 and t3914) in MRSA isolates, and a new ST (single locus variant of ST20) with a *spa* typing t22 in Sa850. Regarding DAP-resistance, we detected SNPs and deletions in the multipeptide resistance factor gene (*mprF*) and the *yycFG* components. Both loci are involved in key cell membrane events, with *mprF* being responsible for the synthesis and outer cell membrane translocation of the positively charged phospholipid, lysyl-phosphotidylglycerol, while the *yyc* operon is involved in the generalized response to stressors such as antimicrobials.

Results: One *S. aureus* was susceptible to all antibiotics tested, except to DAP, and two presented also resistance to ciprofloxacin and cefoxitin being methicillin-resistant *S. aureus* (MRSA). We identified heterogeneity of lineages, such as ST22 (t2357 and t3914) in MRSA isolates, and a new ST (single locus variant of ST20) with a spa typing t22 in Sa850. Regarding DAP-resistance, we detected SNPs and deletions in the multipeptide resistance factor gene (*mprF*) and the *yycFG* components. Both loci are involved in key cell membrane events, with *mprF* being responsible for the synthesis and outer cell membrane translocation of the positively charged phospholipid, lysyl-phosphotidylglycerol, while the *yyc* operon is involved in the generalized response to stressors such as antimicrobials.

Conclusions: In summary, we confirm that point mutations in genes coding for membrane phospholipids are associated with the development of reduced susceptibility to DAP in *S. aureus*. Furthermore, we attested that WGS has the potential of providing a better tool for the detection of the mechanisms involved in DAP-resistance.

Session 4

Making Plant Data FAIR

Daniel Faria, Inês Chaves, Bruno Costa, Célia Miguel

The FAIR data principles of Findability, Accessibility, Interoperability and Reusability have been proposed as a means to tackle the challenge of managing and analyzing the vast amounts of data being produced in domains such as the life sciences. They have been embraced by the European life science network, ELIXIR, and its Portuguese representative, BioData.pt.

Plant sciences raise particular challenges to the implementation of FAIR due to their multidisciplinarity and the heterogeneity of their data, which includes high-throughput molecular and phenotypic data. As such, they have been recognized as a key use case by both ELIXIR and BioData.pt. Here, we relate the efforts undertaken by this use case towards FAIRness.

One such effort has been the extension and refinement of the Minimal Information About Plant Phenotyping Experiments (MIAPPE) metadata standard, which details the metadata attributes that are required or optional for interpreting plant phenotyping experiments, with detailed instructions on how to fill them and recommended ontologies for that purpose. Another effort has been the extension and implementation of the Plant Breeding API (BrAPI), which specifies a standard interface for searching and retrieving data from plant phenotype/genotype databases. This extension aims at ensuring that BrAPI covers all fields required by MIAPPE, and at mapping the two resources so that MIAPPE-compliant datasets can be directly submitted to BrAPI-implementing databases. We are in the process of implementing a national BrAPI end-point within BioData.pt, which will be part of an integrated network of ELIXIR end-points and accessible from an aggregating search engine.

The last effort has been the development of ontologies to bridge gaps in the coverage of the present plant science ontology landscape. We have participated in the development of the Woody Plant Ontology, covering traits of woody plants, and the Plant Phenotyping Experiment Ontology, an ontological representation of MIAPPE. We are also developing the Plant Experimental Assay Ontology, which aims at modelling plant experimental pipelines.

A National Bioinformatics Training Programme for BioData.pt

Joana Marques, Pedro L. Fernandes

Biological research is increasingly data intensive, often involving large and complex data sets. This raises huge challenges associated with data storage, interoperability and usability, requiring skilled professionals, permanent access to training, and strategies to keep the community up-to-date. The aims of BioData.pt include a plan and implementation of training at the national level as part of the infrastructure. Feasibility and sustainability are major concerns, as there is a successful model that we will try to scale-up into a national programme.

The programme will primarily focus on short-term face-to-face training events, providing practical, hands-on training in both general and specific skills to academic

and industrial users of data, in the domains covered by the BioData.pt infrastructure. It will include training courses in: general bioinformatics and data management; domain-specific bioinformatics and data analysis; continuing education activities for training Bioinformatics instructors; lifelong, on-the-job education activities for data curators; and closer collaborations with international infrastructures. The training programme will aim at continuously updating the community in the most current standards, tools and data sources.

The success of a national program is highly dependent on its sustainability. The long-term objective is to decentralise the physical installations and provide themes tailored for specific institutions and/or regions. In addition to face-to-face training, other training methods will be explored, such as distance and e-learning, when feasible. The teaching methods will be based on ELIXIR's training best practices, highlighting the value of active learning, and promoting learner engagement and participation.

To achieve a fully functional, sustainable and reproducible training programme, we are using the GTPB programme as a model, capitalising on more than 18 years of experience in bioinformatics training, and over 5000 course participants so far. It is reproducible and has gradually reached full sustainability. This model has been carefully documented and that is the basis of a complete set of recommendations that will be made available.

We are currently gathering the information needed to propose a plan for a National Bioinformatics Training Programme as part of the BioData.pt infrastructure, by the end of 2018.

Poster Presentations

Developing a standalone toolbox for processing FASTA files

Hugo López-Fernández, Miguel Reboiro-Jato, Noé Vázquez, Pedro Duque, Florentino Fdez-Riverola, Cristina P. Vieira, Jorge Vieira

Introduction: One of the most important types of data used in biological research is DNA or protein sequence data. They are usually stored in FASTA files, which can store one or more sequences. Public databases such as GenBank, NCBI or Ensembl provide huge collections of genomes, genome annotations, and so on, in FASTA format. Nevertheless, downloaded files usually must be preprocessed before subsequent analysis depending on each researcher needs. Despite the simplicity of these preprocessing operations (e.g. remove sequences without a minimum number of bases), processing of large batches of FASTA files is a complex task that usually requires advanced bioinformatics skills and the combination of different tools (including the bash command line) to achieve the desired result. In order to allow researchers to easily perform these operations we are developing the SEDA software application (http://www.sing-group.org/seda/).

Results: SEDA (SEquence DAtaset builder) is a Java desktop multiplatform application specifically created to perform processing of FASTA sequence files. Currently, SEDA allows researchers to filter sequences based on different criteria (including text patterns), translate nucleic acid sequences into amino acid sequences, execute Blast analyses, remove duplicated sequences, and sort, merge, split or reformat files, among others. Moreover, its plugin-based architecture makes it useful for programmers of bioinformatics software that want to make use of the SEDA core operations or extend it by creating new plugins.

Conclusions: SEDA is completely free, distributed under license GPLv3, and provide a friendly graphical user interface designed to allow researchers saving time in processing FASTA files.

Using public consortia omics databases to address clinical questions: evaluating the risk of carriers of hemochromatosis mutations to develop liver cancer

Joana Ferreira, Bruno Cavadas, Verónica Fernandes, Graça Porto, Luísa Pereira

The Cancer Genome Atlas (TCGA) is a large collaboration that has generated multi-omics data in 33 types of cancer. More than two petabytes of genomic data are publicly available, allowing us to address several clinical questions when using suitable bioinformatics tools. In this work, we evaluate the risk of carriers of hemochromatosis mutations to develop liver cancer. Hemochromatosis is a genetic disease of European origin that results in low levels of hepcidin protein synthesis and, eventually, iron overload in the body. Liver is one of the main affected organs in this disease. Two mutations, C282Y and H63D, in the HFE gene have been associated with iron overload in hemochromatosis.

The code of the SAMTOOLS was changed in-house in order to extract only HFE sequence from 10,477 exome sequences available for the 33 cancer types. Then HaplotypeCaller was used for variant calling in the two positions of interest. The frequency of C282Y mutation in the European cohort of liver cancer (0.067) was significantly (p-value=0.03) increased in relation, for instance, to the frequency in the European cohort of gastric cancer (0.035), corresponding to an odds ratio of 1.946. No statistical differences were observed for H63D mutation.

We also checked the relation between carrying C282Y mutation and level of viral infection in the liver, by establishing a pipeline using RNASeq data. The RNASeq unmapped-human reads were run against a reference database of human viruses, by using the SAMTOOLS, Bowtie and PRINTSEQ. Each virus was quantified as parts per million reads (ppm), and an infection threshold of 10 ppm was considered. An amount of 32% of the liver cancer cases were infected by hepatitis B virus. Nevertheless, the C282Y-carriers had no infection, showing that mutation and viral infection are independent liver cancer risk factors.

sim16S: A new bioinformatics tool for benchmarking of metagenomics software

Luís Vieira, Octávio S. Paulo

Sequencing of metagenomes is a widely used approach within ecological and human clinical studies. DNA sequences can be derived from all fragments in the sample (shotgun sequencing) or from amplicons of marker genes, such as the highly conserved 16S rRNA gene. In the latter case, accurate taxonomic classification of sequencing reads continues to be a challenging task mainly due to platform-specific sequencing errors. It is therefore important to have a tool that can simulate a metagenomic dataset composed of reads from different prokaryotic 16S rRNA sequences and which contain errors mimicking real world sequencing data. In this work we present sim16S, a new program developed in Matlab that allows the user to generate a customized 16S rRNA amplicon dataset by introducing its own primer sequences and deciding the proportion of reads containing 1 or more substitution errors per read. sim16S produces several statistics, including the proportion of primers on target, sizes of amplicons, global substitution rate and the number of reads belonging to each taxon from phylum to species level. To test the usefulness and potential of sim16S, several different metagenomic datasets were produced and classified with QIIME, one of the mostly used taxonomic classification programs. As expected, the accuracy of classification decreased between 99.94% and 79.78%, and between 99.99% and 82.91%, from phylum to genus, for error-free sequencing reads derived from V3 and V4 regions of the 16S rRNA gene, respectively. The accuracy of classification was kept at similar levels in the presence of ~1% and ~10% of reads with 1 sequencing error per read but decreased 6.53% at the genus level in the presence of 100% of reads with 1 error per read. In contrast, the presence of 4 errors per read in ~10% of reads resulted in only a minor reduction of classification accuracy ranging between 0.10% (phylum) and 0.64% (genus), indicating that different error profiles have distinct impacts in taxonomy classification. We conclude that sim16S is a useful bioinformatics tool to benchmark taxonomic classification programs and that it can easily be adapted to other prokaryotic or eukaryotic genes for which sequence databases are already available.

In silico analysis of transcriptional nucleotide modifications in Vitis vinifera sylvestris

David F. Silva, Miguel J. N. Ramos, João L. Coito, Wanda Viegas, Margarida Rocheta

Vitis vinifera L., includes the cultivated *Vitis vinifera* subsp. *vinifera* and the wild form *Vitis vinifera* subsp. *sylvestris*. The most discriminating characteristic between them is the sexual system (Arnold et al., 2007). Flowers of cultivated varieties are mainly hermaphroditic, whereas wild plants are dioiceous with female and male separated individuals that follow a hermaphroditic pattern during early stages of floral development.

In this study, we sequenced the genomes from male and female plants of *Vitis vinifera sylvestris* and performed RNA-Seq from three flower developmental stages in both flower types resulting in six transcriptome. The availability of big data, obtained in this kind of projects, provides unprecedented study opportunities, but also raises new challenges for data analysis that can only be processed with a robust bioinformatics analysis.

The data analysis was performed using a RStudio cluster running R, programmed using parallel computing instead of a serial approach. The bigger advantage of parallel was the reduced time used to process data when compared with a serial approach.

The comparison between genomes and the six transcriptomes allowed the identification of transcriptional nucleotide modifications otherwise impossible e nova assemblagem. Comparisons between male and female flowers clearly demonstrate that the majority of the modifications occur on female plants which also present a marked increase on modifications frequencies from the initial to the intermediary flower developmental stage.

SPINET (Syndemic Protein Interaction NETwork): a web server for HIV, M. tuberculosis and host protein-protein interaction networks

Ana Santos-Pereira, João Correia, Miguel Rocha, Nuno S. Osório

Tuberculosis (TB) and HIV-1/AIDS are major public health global problems. In 2015, 1.8 million people died from TB, with 22% of these deaths occurring in HIV-1 coinfected individuals. HIV-1 greatly increases the risk of active TB and coinfection

leads to an acceleration of both diseases. The synergy between the two diseases is very complex and more information on how these pathogens interact with humans is necessary to devise better therapeutic strategies. Although numerous interactions between proteins of HIV-1, M. tuberculosis (Mtb) and human proteins have been reported in databases or in the scientific literature there is currently no single database where this information can be efficiently accessed and visualized. Thus, the goal of this project is to provide researchers a concise but informatic network of all known interactions of HIV-1/Mtb and human proteins. The webserver will be developed in node.js and we will resort to Cytoscape.js for the network visualization. It will allow the user to visualize the complete network of interactions or to introduce one or more proteins of interest (from Mtb, HIV-1 or human) and display graphical representations of protein-protein interaction networks involving these proteins. This will promote easy access to this valuable information that is not available in an integrated way, making it time-consuming to study by individual researchers. This webserver based tool has the potential to provide helpful insights to the HIV/TB-research communities.

Implementing High-Throughput Sequencing in bacterial foodborne pathogen surveillance: The INNUENDO Platform

Bruno Ribeiro-Gonçalves, Diogo Silva, Miguel P. Machado, Mickael Silva, Jani Halkilahti, Anniina Jaakkonen, Federica Palma, Mario Ramirez, Mirko Rossi, João André Carriço

Background: Outbreak investigations and pathogen surveillance are crucial tasks to control transmission of foodborne transmitted diseases. The decreasing costs of High-Throughput Sequencing (HTS) are boosting application of HTS for molecular typing in routine surveillance and outbreak investigation, maximizing discriminatory power. However, lack of standardized bioinformatics infrastructures for data processing and integration, together with limited bioinformatics skills, continue to be major hurdles of HTS routine implementation. To overcome these limitations, we developed the INNUENDO platform, an infrastructure that provides a user-friendly interface and the required framework for data analysis, from raw data quality assurance to integration of epidemiological data and visualization of the final analyses, providing the tools for the use of HTS techniques in everyday surveillance and outbreak investigation.

Method: The INNUENDO platform includes the INNUca pipeline for automatic QC from reads to draft genome assemblies, which ultimately aims at producing consistently high-quality and comparable genomic data. The curated genome assemblies are then analysed following a gene-by-gene typing based approach. The chewBBACA software is used to perform the allele calling for whole genome MLST (wgMLST) profile definition. The wgMLST profiles generated for each isolate of interest can then be compared with profiles already stored in the platform's database. The wgMLST profiles of the isolates of interest, together with a selection of the closest ones in the database, are then filtered to produce a core genome MLST (cgMLST) and the data sent to PHYLOViZ Online for the construction of a minimum spanning tree annotated with metadata, allowing the exploration of possible epidemiological scenarios.

Results and conclusion: INNUENDO platform was developed with a modular design allowing the incorporation of different bioinformatic tools for the characterization of specific pathogens. It also aims to facilitate data sharing and communication between different institutions, promoting cooperation in surveillance and outbreak investigation. The use of open source tools and standardized protocols will allow a future accreditation of the INNUENDO platform.

A distributed nation-wide bioinformatics service facility

Daniel Pinto-Neves, Jorge Vieira, Daniel Sobral

BioData.pt is a Portuguese distributed infrastructure for biological data that operates as the Portuguese node of ELIXIR (European Life Sciences Infrastructure for Biological Information). Its vision for the future is to provide seamless access to biological information and tools to researchers in the life sciences, both in academia and industry. As part of these efforts, we are implementing a distributed service facility, operating nation-wide, to provide bioinformatics support to researchers in the life sciences.

Currently operating in two nodes, one based in Instituto Gulbenkian de Ciência (IGC), Oeiras, and the other in Instituto de Biologia Molecular e Celular (IBMC), Porto, we provide consulting services in bioinformatics and computational biology, from the initial steps of study design to downstream analyses, as well as direct support on a fee-

for-service basis, in the analysis of a broad range of biological data, building upon the expertise currently present in the operating nodes, and resources available in the wider BioData.pt network and in other infrastructures such as INCD and GenomePortugal.

Our main objective is the empowerment of researchers by capacitating them with the necessary tools, infrastructure and knowledge necessary to autonomously perform their own analyses. For this, considerable effort is put into the development of user-friendly specialized software, as well as reusable data analysis pipelines and workflows. The later are provided to users as value-added cloud computational resources on a per need basis in the form of virtual machines pre-configured with analysis pipelines and tools such as Galaxy. To further incentivize user autonomy, we also provide hands-on tutorial sessions on a range of bioinformatics tools and interface with in-depth training resources such as what is provided by the Gulbenkian Training Program in Bioinformatics (GTPB).

A Proposal for a Allele Nomenclature Server for gene-by-gene methods

Mickael Silva, António Paulo, Ramirez M, Carriço JA

Due to the fast advances of DNA high throughput sequencing technology, microbial strain identification has become focused on sequence-based methodologies, due to the high portability of results and much higher discriminatory power when compared with traditional gel-based or phenotypic techniques. One of the most common approaches, Gene-by-gene methods, extend the concepts of Multilocus Sequence Typing (MLST), to the genes present in the core genome of a given species (cgMLST) or even the pangenome (wgMLST) defined by a set of strains. We have previously defined a framework, chewBBACA, that allows the creation of gene-by-gene schemas and performed allele calls based on that schemas from assembled draft genomes. Nevertheless allele calls obtained with chewBBACA need to be shareable and comparable within the community at a global level. For that purpose we developed a Nomenclature Server that is based on the the TypOn ontology, and integrates with the chewBBACA framework, to provide a centralized process of defining allele identifiers.

The implemented nomenclature server provides a public and centralised web service, allowing an easy way to users worldwide to use chewBBACA to download the

necessary data for the cg/wgMLST schemas, analyze locally their own samples against that data and query/submit the results to the database. An important novelty point in our approach is that, unlike in other publicly available webservices such as PubMLST or Enterobase, users can run their analysis in their local machines circumventing possible concerns due to data privacy issues, and uses the advantages of ontologies/triple store databases. An example of the implementation is available at http://137.205.69.51/app/v1/NS and the REST API is described in https://app.swaggerhub.com/apis/mickaelsilva/nomenclature_server/1.0.0 . The web service was built with a Python web development microframework, Flask, and uses the virtuoso triple store as main database.

Development of an automated pipeline for metagenomics and metatranscriptomics data analyses

João C. Sequeira, Miguel Rocha, M. Madalena Alves, Andreia F. Salvador

The dynamics of complex microbial communities within their natural environment can be investigated by using molecular biology based technologies, such as metagenomics (MG) and metatranscriptomics (MT). MG and MT target the study of metagenomes and metatranscriptomes, respectively, and generate large quantities of information that require bioinformatics analyses. Several tools are available to perform specific steps, including pre-processing of files obtained by Next Generation Sequencing, assembly of reads into contigs, annotation of reads and/or contigs and finally differential gene expression analysis. The available pipelines developed for MG or MT are either hard to handle or web based and lack several analysis steps. The goal of this work was to create a pipeline that integrates all the steps for a complete MG and MT analysis, with emphasis on automation and independence from web access. Several wrappers were created in Python 3 to run the tools through the command line interfaces. FastQ reads are the input for the pipeline, which are first submitted to quality check and preprocessing by using FastQC and Trimmomatic. The report obtained with FastQC is used to define the parameters for trimming, specifically the removal of low-quality regions and artificial sequences. Removal of rRNA is done with SortMeRNA. Two different assemblers, MetaSPAdes and Megahit, and two quality control tools for the

assembly step, MetaQUAST and Bowtie2, are available in the pipeline. The annotation steps begin with the gene calling, performed with FragGeneScan, and then by the alignment of sequences to a FASTA database (e.g., UniProt) with DIAMOND. The pipeline uses the UniProt's ID mapping service to automatically retrieve taxonomic and functional information on the annotated genes. The output consists on CSV files containing all taxonomic and functional information, and formatted for generating krona plots for data visualization. Differential expression analysis is performed with DeSEQ2, and results are represented in heatmaps. This pipeline is an alternative for combined MG and MT data analysis, since it is the only pipeline available which includes all major steps, that is fully automated and almost independent from web services.

Antibiotics and host immune responses in intracellular infections

Francisco Paupério, Erida Gjini

In this work, we study intracellular infection dynamics combining effects of antibiotic treatment and adaptive immune responses. The ODE models are based on parameters for acute and chronic bacteria, such as Listeria monocytogenes and Mycobacterium tuberculosis. We find that driving factors of infection outcome are the balance between intracellular and extracellular processes, modulated by the apoptosis rate of infected macrophages, death rate of extracellular microbes and burst size. We also study mutational processes leading to emergence of resistance. We present a formal sensitivity analysis of key model parameters, with and without treatment, and extract optimization principles for infection over a range of scenarios. Our results highlight the potential of new targeted treatments of intracellular infection, with low antibiotic doses and duration, combined with sufficient immune action. We call attention on the clearance of the intracellular infection compartment to effectively eliminate bacteria and minimize the risk of resistance.

Sponsors & Partners