Universidade do Minho
Escola de Engenharia

2017

Bioinformatics Open Days (BOD)

Book Conference

22-02-2017 to 24-02-2017

CONTENT

ORAL PRESENTATIONS

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 6th 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 22nd, 23rd and 24th February.

ORGANIZATION

General chairman

- Miguel Rocha

Organizing committee chairmen

- Andreia Campos (University of Minho)
- António Dias (University of Minho)
- Bruna Daniela Azevedo (University of Minho)
- Bruna Mendes (University of Minho)
- Catarina Santos (University of Minho)
- Daniel Moreira (University of Minho)
- Fernando Cruz (University of Minho)
- Gonçalo Figueiro (University of Lisbon)
- Hugo Magalhães (University of Minho)
- Isabela Mott (University of Lisbon)
- Joana Martins (University of Minho)
- João Sequeira (University of Minho)
- José Dias (University of Minho)
- Pedro Raposo (University of Minho)
- Raquel Simões (University of Minho)
- Sara Cardoso (University of Minho)
- Telma Afonso (University of Minho)
- Tiago Barbosa (University of Minho)
- Vera Manageiro (University of Lisbon)

PROGRAM

22 February, Wednesday – Workshops Day

HOURS	WORKSOPS	
14:30 – 17:00	Biomedical Text Mining	Bionode

23 February, Thursday

HOURS	
09:00 – 09:30	CHECK IN
09:30 – 10:00	Opening Session
10:00 – 10:45	Lecture *"NGS analysis of intratumoral heterogeneity"* David Posada
10:45 – 11:15	COFEE BREAK
11:15 – 12:30	Oral Presentations – Session 1
12:30 – 14:00	LUNCH
14:00 – 14:45	Lecture *"Genomic Epidemiology: How High Throughput Sequencing changed our view on bacterial strain similarity"* João Carriço
14:45 – 16:00	Oral Presentations – Session 2
16:00 – 16:30	COFEE BREAK
16:30 – 17:45	Oral Presentations – Session 3
18:00 – 20:00	Speed Meeting
20:30 – 00:00	Social Dinner

24 February, Friday

HOURS		
10:00 – 10:45	Lecture *“Human genome editing - using NGS to characterize the NHEJ-mediated Gene Editing events, a particular case of Hematopoietic Gene Therapy in Fanconi Anemia”* StabVida (Paulo Almeida, Francisco José Román Rodríguez)	
10:45 – 11:15	COFEE BREAK POSTER SESSION	
11:15 – 12:00	Companies Session	
12:00 – 13:00	Round Table	
	Heart Genetics	Ana Teresa Freitas
	IGC-Ophiomics	José Pereira Leal
	Astrazeneca	Marisa Cunha
	Coimbra Genomics	Bruno Soares
13:00 – 14:30	LUNCH	
14:30 – 16:00	Oral Presentation – Session 4	
16:15 – 16:30	CLOSING SESSION	

SPEAKERS BACKGROUND

David Posada

David Posada González is a professor at the University of Vigo, Spain, in the area of genetics, and an investigator in the areas of genetics and bioinformatics, being involved in the development of tools such as ModelTest, which carries out statistical selection of best-fit models of nucleotide substitution, CodABC, that coestimates recombination, substitution and molecular adaptation rates, and SimPhy. After graduating in biology, he took a PhD in zoology in Brigham Young University, in Utah, USA, and later worked as a scientist in the Oxford University, UK, Variagenics, Boston, USA, MIT, USA, among other places.

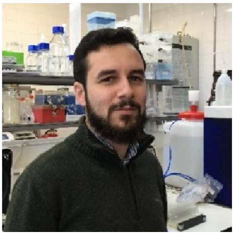
João Carriço

João André Carriço, Auxiliary Researcher at the Microbiology Institute of the Faculty of Medicine, University of Lisbon, graduated in Applied Chemistry – Biotechnology in 2000 and received a PhD in Biology in 2006 from the New University of Lisbon. He is also an Invited Assistant Professor in IST/UL in the Computational Biology Course. His research is centered on the bioinformatics aspects of the data management and analysis of microbial typing data. Currently he is focusing in the development of new methodologies for the analysis off High Throughput sequencing data and in the development of novel phylogenetic algorithms to analyse, interpret and visualize population structure using gene-by-gene methodologies. Further information can be found at his homepage: http://www.joaocarrico.info.

Francisco José Rodríguez

Francisco José Román Rodríguez, a 27-year-old PhD student from Úbeda, Spain, studied a Biotechnology degree in Universidad Pablo de Olavide, Seville (Spain). After graduating in 2012, he studied a Master in Biomedical Research in the University of Seville. During this time, he developed his master thesis in the Division of Human Genetics and Foetal Medicine at Hospital Virgen del Rocío, Seville. The project was focused in the identification of new genes and mutations involved in Hirschsprung disease. Later on, he was awarded with a grant from the Spanish Economics Ministry to develop a PhD project in the Division of Hematopoietic Innovative Therapies at Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT) in Madrid, where he has been working since January 2014. Here, they work in Fanconi Anemia, a low prevalence genetic disease characterised by the bone marrow failure at a very early age. As the only curative therapy available nowadays is hematopoietic stem cells transplantation, new alternatives are being investigated. Due to its monogenic behaviour, Fanconi Anemia is a good candidate to be treated by gene therapy. So the main efforts are focused in the development of new gene therapy approaches to directly correct the disease in hematopoietic stem cells from patients.

Paulo Almeida



Paulo Jorge Couto de Almeida obtained a degree in Genetics and Biotechnology awarded by Universidade de Trás-os-Montes e Alto Douro in 2011, and completed his master's degree in Biochemistry, awarded by Faculdade de Ciências e Tecnologia da Universidade de Coimbra in

2013. Part of his course was taken at Biogem s.c.a r.l. Institute in Ariano Irpino, where he deepened his interest in Bioinformatics. Since then, he has been working for STAB VIDA in the NGS Bioinformatics field. In terms of main functions his time is spread between NGS Project Design, NGS data analysis and development & adaptation of NGS analysis pipelines.

WORKSHOPS

Biomedical Text Mining

Francisco Couto, André Lamúrias e Luís Campos

Text mining is the process of obtaining information from a free text. In a biomedical context, this process is used in biological entities (genes, proteins, …) identification in free text, as well as in obtained experimental results validation with recourse to the corresponding literature. In this workshop, text mining solutions will be approached in order to facilitate the biomedical context concepts' search and annotation. This workshop will have three researchers from University of Lisbon as trainers.

BIONODE

Bruno Vieira

Bionode is an open source community focused on leveraging the Node.js ecosystem for the field of Bioinformatics. It is ambitioned to build highly reusable code and respective tools, and scale-up in order to process high volumes of data on any computational architecture, integrated with modern web technologies. Participants on the workshop will learn how to use Node.js to query and fetch data from the web, how to process it in real-time and in a scalable way using Node.js Streams (similar to UNIX pipes). The focus will be on genomic data repositories.

LECTURES

“NGS analysis of intratumoral heterogeneity”

David Posada

A tumor consists of an expanding population of clonal cells, that differentiate to a bigger or lesser extent and disperse to nearby or distant tissues. Different somatic evolutionary processes like mutation, selection, genetic drift, and/or migration result in ample intratumoral genomic heterogeneity, which has important clinical implications.

In this talk it will be present the current progress in understanding the process of tumor evolution analyzing NGS data from multiple regions in colorectal tumors and chronic lymphocytic leukemia.

“Genomic Epidemiology: How High Throughput Sequencing changed our view on bacterial strain similarity”

João Carriço

The development of High Throughput Sequencing (HTS) was a major revolution for Biological Sciences. The current ability to have a draft microbial genome in a matter of days for an average cost of 100 to 150 euros is a game-changer for Microbiology. In Clinical Microbiology where epidemiological studies relied in a plethora of typing methods to discriminate bacterial isolates at subspecies level, this methodology presently provides much more information and discriminatory power for a total fraction of the cost and time. HTS also created several new opportunities and challenges which are driving the novel field of Genomic Epidemiology. The ability to interrogate hundreds to thousands of draft genomes for a given microbial species, and the increasing availability of open data on public repositories, has its own computational challenges in data management and analysis, but it is, nevertheless, currently changing our views about species definitions and, within a given species, how to measure similarity between microbial strains.

In this talk, it will be described what are still the current challenges in the field and discuss the impact that Genomic Epidemiology has the surveillance and control of infectious diseases on a global scale.

“Human genome editing - using NGS to characterize the NHEJ-mediated Gene Editing events, a particular case of Hematopoietic Gene Therapy in Fanconi Anemia”

Paulo Almeida and Franscisco José Rodríguez

Gene editing with programmable nucleases has emerged as a new approach for the treatment of patients with inherited diseases. In this context CRISPR/Cas9 system is evidencing itself as an exceptional technology for genome engineering. However, such a powerful system requires an equivalent platform to properly analyze the editing generated. Due to its high resolution and detection capacity of low frequency mutations, next generation sequencing (NGS) has become an indispensable tool in genetic engineering because, when used properly and followed by a meticulous analysis, it constitutes one of the most accurate and sensitive methods for detecting and quantifying the frequencies of the repair events induced by the nucleases.

For these reasons NGS was the approach of choice to analyze the variations generated when targeting the most common mutation described in Fanconi Anemia Spanish patients. It consists on a premature stop codon in the fourth exon of FANCA that leads to a nonfunctional truncated protein. On particular, the goal was to remove the stop codon generated by the mutation as a result of the INDELS generated by NHEJ using CRISPR/Cas9 nucleases. NGS revealed that some of the repair events removed the pathogenic mutation while preserving FANCA open reading frame. These results agreed with the marked in vitro proliferative advantage observed in the edited cells, together with the characteristic MMC sensitivity reversion and the restoration of the FANCD2 foci formation.

Companies

Astrazeneca

Astrazeneca (www.astrazeneca.com) is a British–Swedish multinational pharmaceutical and biopharmaceutical company that manufactures and sells pharmaceutical and biotechnology products to treat disorders in areas such as gastrointestinal, cardiac and vascular, neurological, oncology, among others. It also has a major research and development presence in Sweden.

This company will be represented by Marisa Cunha.

Coimbra Genomics

Coimbra Genomics (www.coimbra-genomics.com) is a portuguese biotech company that develops clinical decision support systems based on the patient's genome. Their software platform, ELSIE, allows physicians to consult their patient's genome and obtain reports with information relevant for clinical decisions.

This company will be represented by Sónia Martins.

HeartGenetics

HeartGenetics, Genetics and Biotechnology SA (www.heartgenetics.com) is a Portuguese biotech company that developed solutions of genetic tests for several cardiac pathologies. The computational tools developed support clinical diagnosis and prevention of cardiovascular diseases, by providing highly accurate and reproducible analysis and the integration of both genetic and clinical data.

This company will be represented by Ana Teresa Freitas.

Ophiomics

Ophiomics (www.ophiomics.com/en/home-en) is a portuguese biotech company that offers molecular diagnostic tests, based on DNA sequencing and tumor biopsies, enabling a precision medicine, in order to define personalised strategies for early detection of disease and therapies. They focus on chronic diseases, including cancer.

This company will be represented by José Pereira Leal.

ORAL PRESENTATIONS

Session 1

RNAplonc: identification of plant Long Non-Coding RNAs

Tatianne Da Costa Negri, Pedro Henrique Bugatti, Priscila Tiemi Maeda Saito, Douglas Silva Domingues and Alexandre Rossi Paschoal (Brazil)

Long non-coding RNAs (LncRNAs - > 200 nucleotides) is a class of large noncoding RNAs (ncRNA). LncRNAs have emerging attention in the last years as a potential layer of gene expression in cells. However, lncRNAs mechanisms in plants are still poorly known. Moreover, there is a lack of specific computational approaches for lncRNA prediction in plants, considering that the biological mechanism of this ncRNA class is different from mammals, which there are several tools for prediction. Having this in mind, we present the RNAplonc, an approach for lncRNA identification in plants. To build this tool, we used publicly avaliable lncRNA and transcript (mRNA) sequences from six plant genomes: A. thaliana, C. sativus, G. max, O. sativa and P. trichocarpa (extracted from the databases PLNlncRbase, GREENC and Phytozome). We applied pattern recognition techniques in a total of 5468 features based on sequence and structure from lncRNAs and transcripts (e.g. GC content, ORF, k-mer 1-6) in order to select the best features for classification. Sequences were also processed using: (i)- CD-Hit-EST: to avoid sequence redundancy; (ii)- txCDSPredict: for ORF prediction; (iii)- in-house PERL scripts to calculate k-mer nucleotides frequency, GC context, normalization and generating an ARFF file. All feature selection and classification processes were done using Weka 3.8.0. We detected 16 best features for classification after feature the selection process. These features were used to compare six classification methods. The REPTree method obtained the best results with (ROC curve = 0.933). We also compare against method in literature and obtain similar or best results. These results point out a promising approach to lncRNA identification in plants.

Genomic approaches to the study of adaptive evolution in Drosophila subobscura populations of contrasting biogeographical history

Marta Antunes, Pedro Simões, Inês Fragata, Gonçalo S Faria, Marta A Santos, Sofia G Seabra and Margarida Matos (FCUL/IGC/UK)

Joining the power of Experimental Evolution and genome-wide sequencing (evolve-and-resequence), we are performing genome-wide analysis in two populations of Drosophila subobscura founded from wild collections in the extremes of the European cline (Adraga, Portugal and Groningen, Netherlands). These populations (each three-fold replicated by generation three) have previously revealed fast convergence during laboratory adaptation in phenotypic traits but not at chromosomal inversion frequencies, despite significant changes across generations in the latter. We have taken two different approaches: 1) genome resequencing of pools of 50 females from each population from four different generations since foundation in the laboratory (generation 1, 6, 25 and 50 after founding), in a total of 24 samples; 2) Restriction site associated DNA sequencing (RADseq) of a total of 475 individual larvae (F1 crosses with a homokaryotypic stock, chcu) with known karyotype from two different generations (6 and 25) of the same populations. With these approaches we are addressing the impact of history, chance and selection in genomic variation and evolution, as well as the genomic impact of chromosomal inversions.

With these goals in mind, a draft genome was assembled. Our analyses of Pool-seq data included: alignment of reads to the reference genome; SNP calling; estimation of diversity and differentiation; detection of candidate SNPs under selection. We are characterizing candidate SNP regions through functional annotation and analysing the effects of mutations in terms of aminoacid changes. Variation in both genome-wide SNPs and SNPs with signal of selection revealed that initially differentiated populations followed different genetic routes during adaptation, with no genomic convergence detected between them.

For RADseq analysis we aligned the reads to the reference genome and obtained the haplotypes for each individual. We developed a pipeline to remove chcu haplotypes and construct the SNP matrix. In order to localize SNPs in relation to chromosomal inversions, we are developing a database with all published sequences that are cytologically mapped in this species. With this approach we will analyse if the genomic content of a given inversion differs between populations as well as if it changes during adaptation.

Pan-genome comparison between Streptococcus dysgalactiae subsp. equisimilis isolates from human and animal sources

Catarina Inês Mendes, Marcos Pinho, José Melo-Cristino, Mário Ramirez and João André Carriço (IMM)

Streptococcus dysgalactiae subsp. equisimilis (SDSE) is being increasingly reported in human infections colonizing various animal species, although no genomic analysis has been done to clarify the genomic identity and taxonomy of the isolates from animal origin.

In order to assess the genomic differences between SDSE isolates from human and animal sources, a pan-genome comparison was performed with a collection of SDSE isolates from human (n=40) and animal (n=32) sources. The collection was de novo assembled and annotated with Prokka, with quality control performed after the assembly. Roary software was used to create the pan-genome. Three different clades were readily distinguishable in all analyses, one containing isolates recovered from human sources (n=37), other containing isolates recovered from horses (n=15), and the third clade containing isolates recovered from various hosts (n=20), including human, horse, pig, dog, chicken, fish, duck, iguana and cow. A core-genome MultiLocus

Sequence Typing (cgMLST) analysis was also performed by creating the pan-genome of the 72 SDSE using the chewBBACA pipeline, to study allelic variation in core loci. As an approach to study recombination within the three clades, allelic segregation studies were carried out based on the cgMLST and whole-genome MultiLocus Sequence Typing (wgMLST) profiles.

Gene association studies, using Scoary, were carried out in the pan-genome of the three SDSE clades. The clade containing isolates recovered from various hosts had no significantly associated genes but there was a set of exclusive accessory genes for the human and horse clades. The exclusive accessory genomes of each clade were evaluated for potential virulence factors that could explain the host specificity of the clades. Several genes with similarity to recognized virulence factors in Streptococcus pyogenes were found in the human SDSE clade, and a set of potential virulence factors with similarity to diverse species were found in the horse SDSE clade.

Acorns of code: The role of bioinformatics in determining the consequences of climate change for cork oak populations.

Francisco Pina Martins and Octávio S Paulo

Global climatic changes have been proven to cause alterations in organisms' traits. Understanding how species respond to this alteration in their environmental context is becoming an ever more important question due to the pace at which these changes are taking place.

Here we discuss the consequences of these changes for Quercus suber and how cork oak is expected to respond to these aforementioned shifts. Unlike the situation with other European oaks, genomic resources for cork oak are scarce at the time of writing, despite the large number of genetic and physiological studies targeting this tree.

In order to answer this question several NGS datasets were obtained and analysed. From 454 data to GBS (illumina) genomic fragments,

analysis efforts were split between technical details (AKA, 'Getting it right') and biological interpretation of the results.
Here we detail which bioinformatics skills were essential and which tools were required / developed / contributed to reach an acceptable overview on how Q. suber's standing genetic variation influences the species' response to global climatic changes. A special focus is given on some of the most relevant implementation details and the whole learning experience.
Finally, some of the most relevant results are presented, together with why they matter in a more biologically broad perspective.

Session 2

In silico Identification of Metabolite Analogues for Rational Strain Improvement

Joao Cardoso, Ahmad Zeidan, Markus Herrgard and Nikolaus Sonnenschein (DTU - Denmark)

Antimetabolites are chemical compounds that can inhibit the utilization of other metabolites that are part of normal metabolism. They are typically structural analogues of natural metabolites, as they share a high degree of similarity with them. When present in the cell, antimetabolites can affect metabolic functions as they compete with the native compounds for the same enzymes. Genome-scale metabolic models (GEMs) comprise all biochemical reactions in an organism and their relation to the proteome and genome. These models have found applications across many different fields: metabolic engineering, drug discovery, evolution and microbial ecology. Because these models comprehensively represent natural metabolism in the species of interest, they proved useful for predicting the effect of antimetabolites, e.g., for identifying antibiotics or anti-cancer drugs. In this work, we present a framework to identify Metabolite Analogues for Rational Strain Improvement (MARSI).

MARSI provides an alternative approach to strain design by searching for metabolite targets, instead of genes or reactions. It uses GEMs to identify metabolites that when “knocked-out” can result in a desired phenotype of the cells. The identified metabolites can then be replaced by their respective antimetabolites to obtain the desirable phenotype. This approach does not require genetic engineering and thus enables metabolic rewiring without the use of GMOs. Two main approaches are implemented: re-evaluating strain designs obtained through previously published algorithms such as OptGene or OptKnock, and searching directly for metabolites using heuristic optimization search algorithms. MARSI also provides the basic tools to query and compare the identified targets to known drugs, toxic compounds, and metabolites classified as analogs.

Semantic annotation of electronic health records in a multilingual environment

Luís Campos, Vasco Pedro and Francisco Couto (FCUL)

Beyond the useful uses they already have, Radiology reports still have the potential of being a useful source of information for the improvement of health-related practices. They usually are better structured that other types of Electronic Health Records, making them more prone to be analysed by Natural Language Processing (NLP) tools. But most of these tools assume that the reports were written in English, which is not always true. One obvious solution is to translate the text in the native language to English, before applying the NLP techniques. But what kind of translation should be used? Machine Translation (MT) is one option, being way cheaper than the other obvious alternative, Human Translation (HT), but has the low of having worse quality. One possible profitable trade-off is to use Machine Translation with Post-Editing (MT+PE) by humans.

This work aims at studying how MT and MT+PE compares with HT on the simple task of Named-entity recognition of RadLex terms on biomedical texts related to Radiology, using a dictionary-based approach. The selected corpus for this work consists of research papers related to Radiology that are available both in Portuguese and English. The Portuguese versions of the papers were machine translated to English with Google (MT), Yandex (MT), and Unbabel (MT+PE) translation services and compared with the original (HT) English translation. The focus is to learn which kind of translation is more similar to HT in the task at hand. The results of this study can then serve as basis for decisions by researchers who want to integrate translation in existing or new tools. (e.g. Information Retrieval systems).

The Virtual Metabolic Human and ReconMap: resources for genome scale metabolic reconstructions and visualization

Alberto Noronha, Ronan Fleming and Ines Thiele (Luxemburgo)

The Virtual Metabolic Human (VMH) is a resource that can explicitly link the human metabolism, gut microbiome metabolism, nutrition, and disease. VMH integrates the global reconstruction of human metabolism (Recon 2), AGORA a comprehensive resource of typical gut microbes reconstructions, data for more than 200 metabolic diseases, a gene-phenotype diagnostic tool for mitochondrial disease, as well as the relations between all these entities. The nutrition resource contains 11 diets, each specifying a one day meal plan. The nutritional information contained in each food item can be integrated in a modelling fashion.

VMH also hosts ReconMap 2.0, a manually drawn comprehensive map, consistent with the content of Recon 2. ReconMap is available through a web interface that allows content query, visualization of custom datasets and submission of feedback to manual curators.

VMH provides users with information on human metabolism, disease and gut microbes, and ReconMap enhances the capability to tackle the complexity of the human metabolic network with a comprehensive and interactive visualization, making this combination a valuable tool for any researcher interested in metabolism. VMH and its resources are publicly available at http://vmh.life.

Simulating individual in space: a computer program to simulate complex demographic histories

Rita Rasteiro, Tiago Maié and Lounès Chikhi (UCL/IGC)

SINS (Simulating INdividuals in Space) is a computer program designed to simulate complex demographic histories using a spatial framework, similar to the SPLATCHE2 software but with a forward approach instead. SINS is thus slower but provides more realistic scenarios. With SINS, space is divided into layers, which are themselves subdivided into demes that harbour male and female individuals. Each deme is characterized by values which define the maximum population size (K) and the difficulty to move into that deme (F). SINS allows the user to simulate (1) vary K and F maps across time and space, (2) population expansions from multiple sources, (3) habitat contractions, expansions and habitat fragmentation, (4) sex-biased admixture and competition between populations from two or more layers corresponding to the same geographical space, (5) short and long distance sex-biased migration and (6) variance in reproductive success in males and females. The program uses an individual-based approach to simulate forward in time several types of molecular markers (sequences, SNPs and microsatellites) and genetic objects (X and Y chromosomes, autosomes and mitochondrial DNA). Being able to simulate population structure and at the same time sample the whole population allows SINS to be applied to many species and evolutionary questions.

Session 3

chewBBACA – an efficient framework for large-scale prokaryote whole genome/core genome MultiLocus Sequence Typing analyses

Mickael Silva, Mirko Rossi, Mário Ramirez and João André Carriço (IMM)

With the advent of high-throughput sequencing technologies, the field of microbial genomics experienced a paradigm shift from analyzing single or a few genomes to large-scale comparisons of thousands of genomes. These studies are computationally challenging since the current software and methodologies do not scale well to high number of isolates, taking days to weeks to compute even in large high-performance clusters.

Gene-by- gene methods have shown great promise in terms of scalability, since they reduce the complexity of genomic information to allelic profiles, facilitating the comparison of multiple genomes while at the same time providing a natural nomenclature that is crucial for outbreak detection and surveillance of the epidemiology of infections diseases. These methods allow pan genomic analysis, which include core (common to all strains from a single species) and accessory genome components. This genome-wide gene-by-gene comparison has been termed whole-genome MLST (wgMLST) and core genome MLST (cgMLST).
However, current solutions are either non-scalable, commercial, or only available through dedicated webservers. In order to overcome such caveats, we developed chewBBACA (Blast Score Ratio Based Allele Calling), a comprehensive and highly efficient open-source pipeline consisting of three modules: 1) Creation and validation of wg/cgMLST schemas; 2) Allele calling algorithm based on Blast Score Ratio (BSR); 3) a suite of functions to visualize and validate allelic variation in the loci for a given cg/wgMLST schema. chewBBACA was developed in Python language and runs on high-end single/multi

processor laptops or using SLURM in High Performance Clusters with the allele calling step taking on average 20-40 secs for each typical bacterial genome to run. The main innovation points over similar tools are the requirement for in coding sequences alleles and the use of Blast Score Ratio for sequence similarity inference. The allele calling engine of chewBBACA uses fasta files with draft assemblies and, based on a defined schema, returns as final output an allelic profile matrix and a set of fasta files containing all allelic forms for each gene. The software is freely available at https://github.com/mickaelsilva/chewBBACA.

ELIXIR Portugal: Platforms, Use Cases and Services

José Borbinha, Adelino Canário, João Cardoso, Inês Chaves, Isabel Sá-Correia, Bruno Costa, Cymon Cox, Daniel Faria, Pedro L. Fernandes, Ana Teresa Freitas, Célia Miguel, Pedro Monteiro, Gianluca De Moro, Arlindo Oliveira, Daniel Sobral, Miguel Cacho Teixeira and Mário J. Silva

We provide an update on the implementation of the ELIXIR infrastructure and in particular in its Portuguese Node (http://elixir-portugal.org). ELIXIR is organized into five platforms that support the bioinformatics community across Europe (http://elixir-europe.org/platforms): Tools, Data, Compute, Interoperability and Training. Additionally, four use cases are currently supported by those platforms (http://www.elixir-europe.org/use-cases): Human data, Rare diseases, Marine Metagenomics and Plant Sciences.

The Portuguese Node contributes actively to three of ELIXIR's platforms (Tools, Interoperability and Training) and two of its use cases (Plant Sciences and Marine Metagenomics). The Portuguese Node service delivery plan for 2017, includes six initial services (http://elixir-portugal.org/nodeservices): Computing Services, CorkOakDB, YEASTRACT, Plant sRNA Portal, Plant Experimental Assay Ontology (PEAO) and the Gulbenkian Training Programme in BioInformatics. The service delivery plan is to be updated periodically to include those services that are deemed relevant to the BioInformatics community.

The Portuguese Node of ELIXIR is deployed on a cloud computing infrastructure provided by the INCD (Infraestrutura Nacional de Computação Distribuída), as well as commercial cloud services providers.

One of the active developments is implementing a Breeding API (BrAPI, https://github.com/plantbreeding/API) endpoint. The BrAPI is a RESTful API that provides a standard programmatic access to data from a plant phenotyping and genotyping databases in a language agnostic way. BrAPI is an open source project that can be joined by anyone allowing data providers to improve calls that will enhance their data supply. Our endpoint implementation is intended to support access to the Portuguese community's curated datasets in the woody plants domain.

In addition to describing the current development status, we will discuss strategies to incorporate new services and extend the services offered to the Portuguese Bioinformatics community.

Tuning the engines: using Support Vector Machines to fix docking-based scoring functions

Carlos Simões, Cândida Silva and Rui Brito (UC)

Molecular docking is a computational chemistry method widely used in the structure-based drug design field. Modern docking tools have shown reasonable success at pose fidelity, i.e. the ability to reproduce X-ray ligand poses within the binding site of a target protein, but docking scoring functions (SF) are largely unable to correctly predict ligand binding affinities or simply discriminate active molecules from inactive (or decoy) molecules.

Here, we demonstrate how the use of support vector machines (SVM) can improve a SF's predictive power across a set of challenging pharmaceutical targets. AutoDock Vina was used as a docking engine to generate binding poses for all active and decoy molecules within the targets' binding site. Re-scoring of all docked complexes was carried out using RF-score. The energy parameters of Vina's scoring function and more than 30 RF-score terms depicting protein-ligand interactions were used to train classification models with SVM-light.

The results show that Vina provides acceptable pose prediction accuracy for most targets. As expected, however, Vina's scoring function performs poorly at discriminating actives from decoys. The higher overall performance of our SVM-based classification models confirms the potential of the use of machine learning methods to overcome the limitations of docking scoring functions by capturing the non-additive relationship between the SF's energy terms that describe ligand binding. The inclusion of additional terms produced by RF-score appears to be beneficial to improve scoring in the most challenging examples.

In conclusion, our results highlight the potential of our SVM-based protocols for fast, receptor-based virtual screening using freely-available docking and scoring software.

Coarse-grain molecular dynamics simulations: shedding qualitative and quantitative light on biomolecular processes

Manuel N. Melo, Helgi I. Ingólfsson, Floris van Eerden, Joost Holthuis and Siewert-Jan Marrink

Molecular Dynamics (MD) simulations hold the promise of unparalleled structural and dynamic detail. In practice, however, these techniques are limited by computational power, and their applicability often falls short of the size and time scales of biological interest. I will show how the use of coarse-grained (CG) MD methods — namely, the Martini model — allows the simulation of systems into scales that cannot be reached by conventional atomistic MD. Through the use of CG MD, a range of new applications becomes accessible to simulation, from which both qualitative and quantitative mechanism information can be extracted.

Session 4

Selection of Novel Peptides Homing Alternative Targets in Triple Negative Breast Cancer

Débora Ferreira, Vera Silva, Franklin Nobrega, Ivone Martins, Leon Kluskens and Lígia Rodrigues (CEB-Uminho)

Breast cancer is the most frequent cancer amongst women, representing 25% of all cancer cases, with an estimated 1.67 million new cases in 2012 1. Phenotypically characterized by the lack of known receptors, the triple negative breast cancer (TNBC) subtype is responsible for 10 to 20% of all diagnosed breast cancers 2. Due to its unique profile, aggressive behavior and different patterns of metastasis, the search for effective diagnosis and treatment tools has intensified 3. However, the lack of specific cell targeting remains the main barrier for sensitive diagnostic tools. Therefore, peptide ligands that specifically recognize cell-surface receptors have been extensively used in cancer research. Evolutionary screening techniques as phage display 4 emerged as a powerful tool to recognize specific peptides and has been proved useful for the discovery of new biomarkers 5. To accomplish this purpose, we report the selection of two novel peptides by phage display, using two different libraries, homing the mammary adenocarcinoma murine 4T1 cell line (4T1pep1 –CPTASNTSC and 4T1pep2—EVQSSKFPAHVS) 6. This cell line has been shown to be an accurate model system as it closely resembles human TNBC. The high-affinity identified peptides were screened on the MimoDB database 7 and scanned with SAROTUP webserver 8 to detect homology with previously described cancer-specific peptides and to eliminate the existence of target unrelated peptides and false-positives. Therefore, the peptide sequences were further analyzed by the BLAST algorithm for homology to proteins with known or putative breast cancer correlations against the Homo sapiens and Mus musculus non-redundant protein databases. Bioinformatics analysis suggested that both peptides target human Mucin-16, a well characterized biomarker and its deregulation has been previously implicated in different types of cancer, showing overexpression in breast, prostate, lung and pancreas cancer 9. Docking experiments using CABS-dock webserver 10 were also performed to prove the role of Mucin-16 as a targeting receptor using these novel peptides. Our results strongly support the need of alternative targeting systems for TNBC and the peptides herein selected by phage display are very promising towards breast cancer therapy.

Next Generation Sequencing as a tool to detect antibiotic resistance mechanisms

Vera Manageiro and Manuela Caniça (RicardoJorge)

Antibiotic resistance is an emerging problem, becoming a serious threat to global public health. The causes of its spread are complex, as are the strategies to combat this threat. Following recent improvements in sequencing technologies, whole-genome sequencing (WGS) provides a comprehensive alternative in the evaluation and detection of antibiotic resistance mechanisms. In the scope of the analysis of nonsusceptibility of Gram-negative isolates recovered from human, veterinary and environment samples, we identified the presence of a high diversity of resistance mechanisms, with emphasis in the recently described plasmid-mediated mcr gene, conferring resistance to colistin. Therefore, the objective of this study was to characterize the phenotype and genotype of those isolates using conventional microbiological methods and WGS. Whenever appropriate, genetic relatedness of isolates was also investigated by pulsed-field gel electrophoresis (PFGE). To understand the genetic background of those resistance mechanisms, which included oxyimino-β-lactam, fluoroquinolone and colistin resistance-encoding genes, we performed whole genome and plasmid sequencing using a 454 (Roche) and/or MiSeq (Illumina) sequencing strategy. A set of bioinformatic web tools were used to estimate the presence of pathogenicity determinants, antibiotic resistance (AR) genes, and clinically relevant mobile genetic elements. Indeed, the efficient gene capture and spread of resistance determinants by mobile genetic elements are factors to be taken into account, due to their contribution for the co-selection of multidrug resistant strains in the different settings and environment. Furthermore, WGS might be used with great benefit in combination with phenotypic methods for surveillance purposes.

psichomics: Alternative Splicing Quantification, Analysis and Visualisation in Cancer Data

Nuno Saraiva-Agostinho and Nuno Barbosa-Morais (IMM)

Pre-mRNA alternative splicing (AS) allows proteins with distinct functions to be generated from the same gene, being involved in the control of many common cellular processes. Its deregulation may therefore foster the progression of a wide range of diseases. For instance, associations between most of the hallmarks of cancer and AS alterations have been reported [1].

The advent of next-generation sequencing has allowed transcriptome profiling beyond gene expression, enabling global studies of AS [2]. However, the currently available tools for the analysis of AS from RNA-Seq data are not user-friendly and primarily focus on AS quantification, having limited downstream analysis features (for example, [3,4]).

To overcome these limitations, we have developed psichomics, an R application with an easy-to-use graphical interface for the integrated analysis of AS from large transcriptomic datasets, namely from The Cancer Genome Atlas (TCGA) project. The tool interactively performs survival, principal component and median- and variance- based differentially splicing analyses. Amongst its innovative aspects are the analysis of variance (which our research shows to be important in the detection of otherwise unnoticed putative targets) and the direct incorporation of clinical features (such as tumour stage or survival) associated with TCGA samples. Interactive visual access to genomic mapping and functional annotation of selected AS events is also incorporated. We have successfully used the application in the revelation of cancer-specific AS signatures and associated novel putative prognostic factors.

The application's architecture is modular and extensible, aiming to stimulate contributions from its users, as well as to gradually expand its support to other data sources and file formats and the scope of its analysis and visualisation tools without modifying its core functionalities. The tool is freely available as open-source in Bioconductor (http://bioconductor.org/packages/psichomics) and in GitHub (http://github.com/nuno-agostinho/psichomics).

[1]: Oltean,S. and Bates,D.O. (2014) Hallmarks of alternative splicing in cancer. Oncogene, 33, 5311–5318.

[2]: Tsai,Y.S. et al. (2015) Transcriptome-wide identification and study of cancer-specific splicing events across multiple tumors. Oncotarget, 6, 6825–6839.

[3]: Katz,Y. et al. (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature methods, 7, 1009–1015.

[4]: Alamancos,G.P. et al. (2015) Leveraging transcript quantification for fast computation of alternative splicing profiles. RNA, 21, 1521–1531.

Mechanisms underlying human hemogenic reprogramming

Andreia Gomes, Carlos-Filipe Pereira, Dmitri Papatsenko, Kateri Moore and Ihor Lemischka (Ucoimbra/USA)

The direct conversion of somatic cells into hematopoietic stem cells (HSCs) by ectopic expression of transcription factors (TFs) provides promising approaches for regenerative medicine. Recent work from the lab has demonstrated the conversion of mouse fibroblasts into hemogenic cells that undergo endothelial-to-hematopoietic transition and acquire hematopoietic progenitor activity. Here we demonstrate that human adult dermal fibroblasts can be reprogrammed into hemogenic cells. These cells show multi-lineage engraftment capacity, express HSCs-related cell surface markers and have a similar transcriptional profile to hematopoietic stem cells from umbilical cord blood. It is however unclear how this combination of three TFs impose hematopoietic identity. To address this question we have used Chromatin ImmunoPrecipitation sequencing (ChIP-seq) to determine the genomic binding sites and initial engagement of the transcription factors in the fibroblast genome. Using this system we have also investigated the protein interactions between factors and characterized the induction at the single cell level. The genome-wide gene expression at the population and single cell level and genome location data were integrated to further define direct genetic networks mediated by hemogenic TFs. The uncovered mechanistic insights will broadly impact our understanding of blood stem cell specification and provide means to generate reprogrammed stem cells at high efficiency for transplantation.

Hidden treasures in omics datasets – inference of the microbiome

Bruno Cavadas, Nuno Fonseca, Rui Camacho and Luísa Pereira (IPATIMUP)

Advances in high-throughput sequencing technologies are conducting to the emergence of large-scale omics databases, focused on particular human diseases (such as The Cancer Genome Atlas, TCGA), worldwide human populations (1000 Genomes) or gene expression patterns in the diverse human tissues (GTEx). These human-centred genomic and transcriptomic datasets have a huge potential in providing information about microorganisms that colonize or infect humans, such as bacteria and viruses, the so-called human microbiome. In fact, when performing whole genome/exome or RNA sequencing (RNA-seq), a proportion of the reads will not align in the human reference, aligning instead in the microorganisms that are present in the human sample. The study of the human microbiome is of particular importance due to the microbial influence in disease, a good example being the carcinogenic Helicobacter pylori in the development of gastric cancer.

Since currently there is no sound pipeline for the indirect analysis of the microbiome, we have tested several alignment algorithms in order to establish a reliable pipeline to be applied to RNA-seq data. We tested this pipeline in human gastric tumor and normal samples (a total of 399 samples from TCGA, equivalent to 6.3TB of RNA-seq data). The obtained results are biologically sound, as the identified bacteria are known to colonize the human gut: Proteobacteria (Pseudomonas, Enterobacter, Escherichia, Klebsiella and Helicobacter) and Firmicutes (Bacillus, Clostridium and Lactobacillus) are abundant, with traces of Actinobacteria.

By inferring the microbiome, "hidden" in human multi-omics platforms, we open new avenues in the research of the microbiota impact in cancer. The microbiota diversity can thus be easily tested against diverse human features provided in these databases, such as: somatic and inherited genomic diversity, gene expression profiles, methylome patterns, proteomic/epigenetic modifications, clinical characteristics and response to drug treatment.

Transcriptional Analysis as a tool to understand drug effects – The case of the use of Anthracyclines in Inflammation

André Barros, Ana N. Costa, Dora Pedroso and Luís F. Moita (IGC)

Sepsis is still a poorly understood medical condition, that is responsible for high mortality rates due to few specific treatment options. It has been reported that low dose anthracyclines, a group of clinical approved drugs used mainly for cancer treatment, confer robust protection against severe sepsis. However, the molecular mechanisms of anthracycline-induced protection remain incomplete. Global gene expression patterns and pathway analysis are important tools to generate novel hypothesis for experimental testing.

For this purpose, we performed RNA Sequencing of mRNA (mRNA-Seq) to analyze the transcriptome of bone marrow-derived mouse macrophages. After isolation, these cells were divided into two experimental groups: unstimulated or stimulated with lipopolysaccharide (LPS) to promote inflammation; afterwards, each group was subdivided accordingly to the drug treatment received: epirubicin, aclarubicin, doxyrubicin, etoposide and untreated.

The obtained results on gene read counts demonstrated clear differences in the overall genetic expression profiles between unstimulated and LPS-stimulated samples; the drug treatment lead to also very interesting patterns, particularly for aclarubicin. We then performed differential expression analysis for several combinations of the experimental groups; in turn, these datasets were further explored following two different approaches: first, based on a supervised method, by getting the differential expression of genes of interest in several combinations of comparisons; and the second, in a unsupervised way, by performing Functional Analysis and Enrichment Analysis.

Both approaches led to expected results but also to unexpected findings, mainly on pathways enriched in specific conditions. Overall, these observations allowed us to further uncover the gene expression profiles and pathway mechanics of LPS-caused inflammation and the anthracyclines effect on this condition; in addition, the conclusions will allow us to choose better genes and pathway targets, as well as to boost therapeutic efficacy with both established and novel drugs.

POSTER PRESENTATIONS

1. Automated identification of amino acid misincorporations and their related codons: creating a new R package for Mass-spectrometry data analysis

Andreia Reis, Joana Tavares, Mafalda Santos, Rui Vitorino, Manuel Santos and Gabriela Moura - Ibimed (UA) e IPATIMUP

Protein synthesis accuracy is central to life, and so a high number of mediators functions to maintain translational fidelity and efficiency. For instance, tRNAs can be charged with the wrong amino acid by aminoacyl tRNA synthetases, or ribosomes can incorrectly bind a near-cognate tRNA with one base mismatch relative to the mRNA codon being decoded. If a near-cognate tRNA coding for a different amino acid is used by the ribosome, the wrong amino acid will be incorporated, creating a “missense translational error”. The incorporation of the wrong amino acid in a polypeptide chain (misincorporation) accounts for most of the missense errors that occur during genetic information transfer. In our lab, we are interested in studying such random events that lead to the production of aberrant or even toxic proteins in cells. For this we have developed a method to detect random amino acid misincorporations using mass-spectrometry data from cells growing under different experimental conditions.The main goal of this work was to automatize a process that could retrieve faster and more accurate results on the amino acid that was being misincorporated at each position and also to be able to pinpoint their correspondent codon. In a first phase, scripts were created using R and Perl languages for building a dataset of putative peptide sequences, allowing for a maximum of 2 misincorporations using the Saccharomyces cerevisiae proteome. In a second phase, this dataset was compared to real mass-spectrometry data to determine what misincorporations could actually be found in real samples. Finally, the pipeline was adapted to allow for more complex species to be used, i.e. Mus musculus. The main advantage of this pipeline is to allow for fast and accurate identification of single amino acid misincorporations in a proteome-scale analysis and to be able to integrate DNA sequence information to identify the codons that were mistranslated in the process. This bioinformatics tool also builds a user-friendly output that can be easily managed by researchers with minimum programming skills. The next step will be to upgrade the pipeline so that it can be submitted as a new R package for proteomics analysis.

2. Co-selective patterns of antibiotic resistance and virulence in environmental and human gut microbiomes

Pedro Escudeiro, Joël Pothier, Francisco Dionisio and Teresa Nogueira (FCUL/IGC/Paris)

Pathogenic and non-pathogenic bacteria have gradually become resistant to antibiotics. We tested the hypothesis on which, under antibiotics selective pressure, resistance and virulence traits are co-selected amongst bacterial communities and that the dissemination of resistant phenotypes leads to the dissemination of virulence. Through metagenomic data analysis, we have studied 64 environmental metagenomes, and 110 human gut metagenomes issuing from different individuals belonging to distinctive human populations across the world, having contrastive levels of access to antibiotics. We have found that there is a great diversity of antibiotic resistance (ARd) and virulence factor (VFd) genetic traits amongst metagenomes in general, and different correlations between ARd and VFd. In the human gut there are less ARd and VFd than in more versatile environments, yet some ARd/VFd correlations are still strong. They can vary from very high in Malawi, a poverty-stricken African country, where there is high prevalence of infectious diseases and unattended antibiotic consumption, to inexistent in the uncontacted native populations of the Venezuelan Amazon, thus allowing us to link this effect to antibiotic exposure. These results drive us to conclude that resistance and virulence traits are being co-selected, and that amongst bacterial communities naturally occurring in the human gut microbiota, this correlation can be coupled to antibiotic exposure. Therefore, by selecting for resistant bacteria, we may be selecting for more virulent strains, as a side effect of antimicrobial therapy.

3. Selection of a molecule that specifically targets the cell-surface Human Epidermal growth factor Receptor 2: in silico docking simulation

Diana Sousa, Débora Ferreira, Andrea Silva, Leon Kluskens and Ligia Rodrigues (CEB-UMInho)

Breast cancer is the most frequently diagnosed cancer and the leading cause of cancer death among females, accounting for 23% of the total cancer cases and 14% of the cancer deaths.1 Human Epidermal growth factor Receptor 2 (HER2) is a protein that is overexpressed in 25-30% of breast cancers and is involved in cell growth regulation, survival and differentiation.2 Aptamers generated from Systematic Evolution of Ligands by EXponential Enrichment (SELEX) emerged as a potential new tool for the development of targeted cancer therapies due to their three-dimensional structures that specifically recognize cell surface receptors, such as HER2.3 In this study, HER2-aptamers were screened and identified using SELEX. To design an approach for computational analysis of the isolated aptamers, their structures were modelled by mfold4, a web-based methodology for DNA structure prediction and hybridization software. The HER2 protein structure was obtained from Protein Data Bank (PDB) and using ZDOCK server5, the aptamer-target interactions were predicted through a combination of shape complementarity and statistical potential terms for scoring. Finally, the interactions scores were compared with the experimental results. In silico interaction scores and the experimental outcomes suggest that the best docking results are obtained for the lower binding energy constants. The good agreement between in silico and experimental results highlight the reliability of this new approach. Also, the results provide valuable guidelines for the application of docking simulations for the prediction of aptamer-ligand structures, as well as for the design of novel features of ligand-aptamer complexes.

4. A Pipeline for the Prediction of Fungi Secretomes

João Baptista, Artur Alves, Ana Cristina Esteves and Octávio Paulo (FCUL)

The secretome can be defined as "the set of proteins secreted into the extracellular space by a cell, tissue, organ or organism". Secreted proteins play important roles in a wide variety of cellular functions. In fungi, secreted proteins are necessary for nutrition, but pathogenic fungi also use secreted enzymes for host tissue penetration and colonization.

This pipeline was designed to predict and annotate the secretome of fungi from their predicted proteome by our need to find secreted aspartic peptidases of Ascomycota fungi for a molecular evolution study.

In silico prediction of fungi secretome was performed using a battery of computational tools on the predicted proteome for the prediction of signal peptides, trans-membrane regions, subcellular localization, endoplasmic reticulum targeting motif and GPI-anchored proteins.

To predict signal peptides and trans-membrane regions we used two different methods and only the proteins that were predicted by both methods were selected for further analysis. The first method uses SignalP 4.1 and TMHMM 2.0 to predict signal peptides and trans-membrane regions respectively. The second method uses Phobius 1.01 which predicts both signal peptides and trans-membrane regions. By using two methods that rely in different approaches for their predictions we were able to reduce the rate of false positives. The subcellular localization prediction was made using two predictors, WolfPsort 0.2 and ProtComp 9.0 and their output was combined. Ps-scan 1.86 with the Prosite motif entry: PS00014 was used to predict the endoplasmic reticulum proteins. In the final step of the pipeline we used PredGPI to discover which proteins are GPI-anchored.

After predicting the secretome, its annotation was performed using the tools HMMscan and Blast2Go.

This pipeline allows for an automated prediction of an annotated secretome with a low false positive rate, the possibility for the comparison of different secretomes and to find potential secreted proteins of interest.

Thanks are due, for the financial support to CESAM (UID/AMB/50017-POCI- 01-0145-FEDER-007638) and to the research project ALIEN (PTDC/AGR- PRO/2183/2014-POCI-01-0145-FEDER-016788), to FCT/MEC through national funds, within the PT2020 Partnership Agreement and Compete 2020. The authors acknowledge FCT financial support to A Alves (IF/00835/2013) and AC Esteves (BPD/102572/2014).

5. Genomic analysis of Acinetobacter baumannii prophages reveals remarkable diversity and suggests significant impact on bacterial virulence and fitness

Rodrigo Monteiro, Ana Rita Costa and Joana Azeredo (CEB-UMINHO)

Bacterial genomics has revealed substantial amounts of prophage DNA in bacterial genomes. This integrated viral DNA has been shown to play important roles in the evolution of bacterial pathogenicity. Acinetobacter baumannii has shown a fast progression as a nosocomial multi-resistant pathogen in recent years, and is now considered one of the most dangerous microorganisms in hospital environments. The role of prophages in the evolution of A. baumannii pathogenicity has not yet been explored. In this context, we aimed at evaluating the impact of prophages on A. baumannii genomic diversity and pathogenicity.

Approximately 959 strains were analyzed for the presence of intact and defective prophages using PHAST. A total of 6691 prophages were detected, with all strains having at least one prophage and 83.4% encoding intact prophages. A subset of 184 prophages (from 134 strains) were analyzed in more detail. Prophages were classified by comparing of specific structural proteins to those of previously classified phages using BLASTp. Among the prophages possible to classify, all belonged to the Caudovirales order, with higher prevalence of Siphoviridae (39.67%), followed by Myoviridae (19.57%) and Podoviridae (8.7%) families. The prophages sequences were aligned using MAFFT, and a distance matrix and a phylogenetic tree were constructed to evaluate similarities. A high diversity was found among the prophages, which may contribute to the diversity of A. baumannii, since some strains differ only on the integrated prophages. Furthermore, numerous potential virulence factors encoded by the prophages were detected, implicated in A. baumannii pathogenicity, namely antibiotic resistance, toxins, host interaction, survival and fitness.

Overall, our results demonstrate a high prevalence of prophages in A. baumannii. The amount and diversity of potential virulence factors encoded by the identified prophages point towards a significant contribution of these mobile elements for the dissemination and evolution of pathogenicity in this bacterial species.

6. Copy Number Variants (CNV) association patterns in Suicidal Behavior (SB) and in Major Depression Disorder (MDD)

Katarzyna Kwiatkowska and Lisete Sousa (FCUL)

Etiology of some mental disorders such as schizophrenia, bipolar disorder or autism, has been assigned to number of interacting factors, including genetic variations. This study attempts to examine the association between copy number variants (CNV) and two mood pathologies: Suicidal Behavior (SB) and Major Depression Disorder (MDD), applying a package of Bioconductor.

Raw microarray data was retrieved from EMBL-EBI ArrayExpress repository, generously provided by Turecki G. et al [1] as the result of the work dedicated to potential contribution of CNVs in SB phenotypes, and completed utilizing A-GEOD-8882 - Illumina HumanOmni1-Quad BeadChip platform. The R package selected to perform the association analysis was CNVassoc, since - as referred in literature - it results in increased efficiency and sensitivity. The workflow consisted of four principal steps: aberration calling, reduction of dimensions, CNV association and correction for multiple comparison.

The random subset of samples was prepared in order to reproduce proportions of phenotypic groups as originally observed in collected population. Thus, the study included mono-phenotypic individuals presenting solely SB or MDD (respectively 5 and 7 cases), combined phenotypes of SB and MDD (9 cases) and 28 healthy controls. The analysis was limited exclusively to 22 autosomes, since those are the most promising in terms of relevance of results and permit forming conclusions independent on the sex. Expression samples differed slightly in statistical dispersion, presenting median absolute deviation (MAD) between 0.12 – 0.19. Step of creating regions brought notable reduction of dimension, reaching approximately 99%. The null-hypothesis statement, formed as “CNV state is not associated with a phenotype”, was tested for both mood phenotypic traits. There were found 177 and 165 chromosomal regions among respectively SB and MDD individuals, indicating potential association (identified with individual, not corrected p-values < 0.05) between aberrations and mood disorders. However, after adjustment with Benjamini and Hochberg approach, it was concluded that obtained final results did not provide any evidence on significant (at the level of significance 0.05) association between CNVs and mood disorders.

Despite there exist some indications suggesting that chromosomal variations might be one of the risk factors for SD and MDD, it was not confirmed in study, nonetheless conclusions here formed give a new insight into the matter.

7. Identification of miRNAs related with the response to heat stress in two Triticum durum varieties

Brígida Meireles, Daniel Gaspar, Anabel Usié, Pedro Barbosa, Paula Scotti, José Semedo, Isabel Pais, Benvindo Maçãs, Ana Sofia Almeida, Rita Costa, José Coutinho, Nuno Pinheiro, José Matos, Fernanda Simões, Diogo Mendonça, Joana Guimarães and António Marcos Ramos (Beja)

The global agricultural production was affected by climate changes, which caused a dramatic decrease in the crop area and production. The production of food with nutritional value is fundamental to the population growth and sustainability, with wheat being one of the most significant crops used for human consumption. Due to the global warming effect, wheat yields were negatively affected by drought and heat in the last decade.

Recent studies demonstrated that small-RNAs, specifically miRNAs, regulate the expression of a wide range of genes in plants. They play a regulatory role in plant development and response to stress,being essential in the post-transcriptional gene regulation.

Hence, to discover the role of miRNAs in the response to heat stress in wheat development, two varieties of Triticum durum (Celta and TE-1330) were sequenced using small-RNA sequencing approach. For each variety, two libraries for leaf and stem tissues were constructed, including control and heat stress conditions, and sequenced using the Illumina Hiseq 4000 platform. A total of 188,580,550 raw single-end reads were obtained. After adapter removal, a total of 173,998,257 (92.3%) reads with length ranging from 19 to 27 nt were kept and used for further analyses. These reads were mapped with bowtie against the TGACv1 wheat draft reference genome (Triticum aestivum - common wheat), using very restrictive and optimized parameters due to (1) the short length of the reads and (2) the complexity and size of the reference genome, and a total of 162,534,610 (93.4%) mapped reads were obtained.

The mapped reads were subsequently used in miraligner, selecting Triticum aestivum, Aegilous tauschii, Citrus sinensis, Linum usitatissimum, Brachypodium distachyon and Hordeum vulgare species from MiRBase 21.0, in order to identify conserved miRNAs. A total of 211 conserved miRNAs were identified, including miR160, miR166 and miR167, which were previously associated with heat stress in wheat in other studies.

Further analyses are ongoing to identify novel miRNAs in durum wheat. Additionally, a quantification of expression of the conserved and novel miRNAs identified will be performed in order to identify those miRNAs associated to the response of heat stress in wheat development.

8. Comparative analysis of transcriptional response to heat stress in two durum wheat (Triticum durum) varieties

Daniel Gaspar, Brígida Meireles, Anabel Usié, Pedro Barbosa, Paula Scotti, José Semedo, Isabel Pais, Benvindo Maçãs, Ana Almeida, Rita Costa, José Coutinho, Nuno Pinheiro, José Matos, Fernanda Simões, Diogo Mendonça, Joana Guimarães and António Ramos

Presently, climate changes and global warming are undoubtedly one of the most pressing issues worldwide. The gradual heating of Earth's surface, which has led to extreme weather events, including heat waves and severe droughts, produces a negative impact in farming practices and crop development. Hence, these environmental phenomena have been responsible for high losses in the agricultural sector, being harmful to a large range of cultures such as wheat, which may have implications for world food supply over the next decades. Recent studies revealed a reduction of wheat global production as a consequence of the temperature rise. This decay may represent a big concern since wheat is one of the most significant crops for human consumption.

To further understand the potential impact caused by warming in wheat development, two varieties of Triticum durum (Celta and TE-1330) were selected for sequencing following a whole transcriptome profiling approach. Four RNA pools for each variety, with a total of eight individuals per pool, were created from different tissues and conditions: (1) leafs control, (2) leafs heat stress, (3) stems control and (4) stems heat stress. Thus, eight libraries were constructed and sequenced using the Illumina Hiseq 4000 platform, yielding a total of 683,317,184 raw reads. From this set, 649,703,587 (95.1%) reads were kept after preprocessing. These reads were mapped against the TGACv1 wheat draft genome reference, and a total of 543,288,155 (83.6%) unique mapped reads were obtained, which were used to perform a differential expression analysis conducted with EdgeR for each tissue per variety. While a total of 2,024 (1,105 up and 919 down regulated) and 989 (455 up and 534 down regulated) differentially expressed (DE) genes were obtained between the two conditions in Celta stems and leafs respectively, in TE-1330, a total of 3,162 (1,460 up and 1,702 down regulated) DE genes were identified in stems and 2,065 (753 up and 1,312 down regulated) in leafs tissues. These results were achieved after correcting for multiple testing with a false discovery rate value of 0.01.

These results represent a breakthrough towards the discovery of genes related with the transcriptional response of Triticum durum varieties to heat stress.

9. 16S as a key complementary analysis for study of Inflammation Therapeutics

André Barros, Henrique Colaço and Luís F. Moita (IGC)

The study of the gut microbiome has become a field of intense research that has important medical implications. The variation in composition and density of the bacterial community has been frequently implicated in disease symptoms and their treatment. The analysis of gut microbiota is, therefore, an important component for disease related studies, especially those on obesity and inflammation. The study of the gut microbiome may be of particular relevance for the understanding of severe sepsis of abdominal origin. It has been reported that low dose anthracyclines, a group of clinical approved drugs used mainly for cancer treatment, confer robust protection against severe sepsis. However, recent experiments in a different mouse facility have shown substantial variability in the robustness of the protective phenotype. To understand how the gut microbiome contributes to this variability, we performed 16S rRNA gene sequencing of the microbiota present in the gut of two different experimental groups which yield different survival rates; for one of them, we performed a time-course analysis as well.

In addition to a similar density of the same bacterial groups between experimental groups and across time, our results show a clear clustering for the experimental groups concerning community diversity. These results suggest that the presence or absence of less represented bacteria taxons, rather than their effective density, might be the key modulators of anthracyclines treatment of severe sepsis.

Our work demonstrates the importance of using 16S rRNA sequencing of gut microbiota to understand variability of responses to drug treatment. In conjugation with RNA Sequencing of relevant cell types involved in this condition, it has the potential to expand our knowledge of the deeply complex and dynamic processes that operate in severe sepsis and will open new opportunities for novel therapeutic strategies.

10. Inference of viral infection in human cancer samples from TCGA database

Joana Ferreira, Bruno Cavadas, Pedro Soares and Luisa Pereira

It is known that viral infection is responsible for 15-20% of human malignancies worldwide. With the large amount of genomic and metagenomic information available on public international consortia, such as The Cancer Genome Archive (TCGA), it is nowadays possible to indirectly infer viral infections from human centred omics studies. Here we focused on cervical (CESC), hepatocellular (LIHC) and head and neck squamous cell (HNSC) carcinomas, which are known to show a high proportion of viral-positive cases. Our aims were to: establish a pipeline to infer viral infection from RNASeq data; estimate infection rates; fully characterize the viral strains; and infer gene expression of relevant viruses.

We downloaded RNASeq data for 1299 tumor samples (8,62TB). The unmapped-human reads were run against a reference database of human viruses, by using the tools Batch, SAMTOOLS, Bowtie and PRINTSEQ. Each virus was quantified as parts per million reads (ppm), and an infection threshold of 10 ppm was considered. We confirmed the following infection rates: 94% in CESC, mostly by HPV (human papillomavirus) and specifically by the HPV16 strain; 32% in LIHC, by HBV (hepatitis B virus); 17% in HNSC, most commonly by HPV16. Finally, the most expressed viral genes in HPV16 were identified by using the HTSeq tool, leading to the confirmation that E6 and E7 genes, said to be preferentially integrated in the host genome and responsible for initiating carcinogenesis, are the most overexpressed.

These omics-based viral infection rates are identical to the ones evaluated by standard methods, showing that public international consortia can indirectly provide interesting insights into the involvement of viral infection in tumorigenesis. The high number of samples per tumor, the wide geographic origin of the samples, and the high-throughput characterisation for different omics platforms allows multilayer comparisons and evaluations, in a scale not affordable before.

11. Alternative splicing detection across different tissues in cork oak

Pedro Barros, Pedro Barbosa, Anabel Usié, Cátia Pesquita and António Marcos Ramos (Beja)

Alternative splicing is a process used during gene expression to yield different transcript variants (or isoforms) and protein products derived from a single gene. This process significantly increases the overall protein diversity in a given organism, and is implicated in the genetic regulation of complex traits. In cork oak (Quercus suber), a forest species with high economic and social significance, no studies have ever been executed regarding the characterization of alternative splicing events.

With this work we aim to characterize the extent of alternative splicing events in cork oak, using a RNA-seq dataset generated for four different cork oak tissues, which include leaf, phellogen, xylem and inner bark. Two different bioinformatics strategies will be tested for read mapping and isoform reconstruction, using state of the art algorithms. Read mapping, against the cork oak draft genome sequence, is ongoing using HISAT2 [1] and STAR [2] and the most suitable output will be used for isoform reconstruction and quantification with Cufflinks suite and StringTie/BallGown [1]. Results will be compared and evaluated, providing a standard workflow to support further studies in this species. Characterization of alternative splicing within and between different tissues will be performed by quantification of isoform diversity and abundance. With the recent availability of the cork oak genome sequence, this study will provide important data to improve gene structure annotation.

[1] Pertea et al. (2016) Nat Protoc 11(9), 1650–1667;

[2] Dobin et al. (2013) Bioinformatics 29(1), 15–21;

[3] Trapnell et al. (2012) Nat Protoc 7(3), 562–578.

12. Dynamic cgMLST phylogenetic analysis: a tool for exploring large gene-by-gene datasets at different discrimination levels

Bruno Gonçalves, Mirko Rossi, Mário Ramirez and João Carriço (IMM)

Microbial typing is widely used in different fields of microbiological research such as epidemiological studies to determine source of infection, identify outbreaks, or to determine phylogenetically related clusters with relevant biological properties such as antibiotic resistance or relevant virulence factors. One of the most frequently typing methods is Multi Locus Sequence Typing (MLST), a profile based method that uses seven housekeeping genes for classifying bacteria.

With the rise of Next Generation Sequencing (NGS), researchers now have tools to query a larger number of loci across whole genomes in a method called whole genome MLST (wgMLST). This can be further refined to all shared loci in the set of bacterial isolates being analyzed without any missing data: the core genome MLST (cgMLST). These profiles can then be used to infer relationships between bacterial isolates by constructing phylogenetic trees using tools such as PHYLOViZ Online.

Nevertheless, the discriminatory power of this approach is dependent on the number of genes used to create the profile. However, due to either intrinsic properties of the species under analysis or due to sequencing, assembly or variant calling errors, the number of loci that can be found in all isolates is generally inversely proportional to the number of isolates under analysis. Since missing loci can lead to miss-identification of closely related strains, these are generally excluded from the analysis.

Here we present a new approach to dynamically increase discriminatory power in the comparison between profiles by using PHYLOViZ Online’s interactive visualization. From an uploaded wgMLST profile, the application constructs a cgMLST profile using the entire group of isolates. Then it allows visual interactive selection of isolate subsets that can be re-analyzed by automatically constructing a new tree from a new cgMLST profile that maximizes the shared loci.

This process can then be repeated for further discrimination of a novel subset. This approach reduces the impact of missing data when analyzing similar strains allowing the user to make the most of the available data. A demonstration video is available at goo.gl/t5q6HF.

13. 3D Identification of clusters: a cancer cell case-study

Gonçalo Figueiró, Francisco Couto and André Moitinho (FCUL)

Whether in astronomy or bioinformatics, data driven scientific research is characterized by methods to handle large data sets. Biomedical datasets are no exception and are becoming increasingly complex, particularly because they're derived from multiple sources and hence heterogeneous, unstructured and high-dimensional [1]. It is expected that in coming years the number of biomedical datasets based in 3D point clouds will largely increase. These can be natural point clouds from nuclear medicine and scanners or representative point clouds in which 2D images are represented as 3D points.

The interactive 3D visualization of point clouds allows the discovery of patterns and relationships within the data, however if the dataset is large enough it might be impossible to visualize all points at once due to constraints in memory and computational power. But even if possible it often isn't the right approach since using traditional visualization techniques with these datasets leads to confusing and even useless representations due to cluttering and overplotting problems. Thus there is a need to represent the data without overbearing the user and allow the retrieval of information from the dataset.

We propose creating a method to identify regions of interest in a Point Cloud that takes into account not only the 3D positional coordinates but also all other dimensions.

In order to implement and evaluate our method we created a synthetic dataset to model healthy cells and cancer growth. An adapted version of UPMASK [2] was used to compute membership assignment. UPMASK is an unsupervised clustering method for membership assignment, that was previouly used to assess stellar clusters using only photometry and positions. It is data-driven and doesn't rely on models. It is particularly useful for situations where cluster members are buried in field stars. These characteristics make it a good candidate to use with our problem and case-study since cancer cells are also buried in healthy cells.

An initial version of this method allowed the correct identification of cancer cells (Fig. 1) even in situations where healthy cells where very close or even within the structure of the "tumor".

14. Integrating semantic distances in ontology matching algorithms for biomedical ontologies

Isabela Mott, João Ferreira and Cátia Pesquita (FCUL)

To handle the challenges in managing and exploring life sciences data the biomedical research community has created ontologies to represent life-sciences related knowledge in a standard way. The proliferation of ontology development efforts lead to the creation of ontologies representing the same or similar domains (e.g. anatomy, clinical terms etc.), which in turn introduced the need to merge them into a unified source of knowledge. Ontology matching addresses this need by identifying mappings between equivalent or related entities from different ontologies employing specific matching algorithms to produce a set of mappings, also called an alignment.

An alignment is calculated, for example, by lexical or structural similarity. Lexical similarity is calculated by the likeness between the concept's labels and their synonyms; the structural similarity between two concepts depends on the existence of similar concepts in their neighbourhood, for instances siblings, parents, or children concepts. Calculating this neighbourhood is a challenge because in many biomedical ontologies, structural distance does not correspond to true semantic distance, e.g.: a parent concept can be more similar to one of its children that it is to the rest. However, these matching algorithms can improve the quality of an alignment by finding new matches that would otherwise be missed. This study presents preliminary results from the employment of four semantic similarity based approaches to neighbourhood similarity matching, where the matcher uses the semantic distance between neighbouring concepts to arrive at a final confidence score. An evaluation was run on three sets of biomedical ontologies, and results from this approach showed an improvement over existing alignments, with the detection of new mappings as simple as ‘cuboid’ and ‘cuboid bone’ or more complex like ‘Ligament of upper limb’ and ‘Structure of ligament of shoulder and upper extremity’. The most successful alignments improved the baseline algorithm up to 8.2% in F-measure.

The mentioned algorithms have been integrated into the AML system, a top performing ontology matching system, particularly geared for biomedical ontology matching, and will be further evaluated using other ontologies and tasks.

15. INNUca, a standardized pipeline for bacteria genome assembly and quality control

Miguel P. Machado, Mirko Rossi, Inês Mendes, Yucel Nalbantoglu, Mário Ramirez, Vítor Borges and João André Carriço (IMM)

Currently, surveillance by public health institutes/organizations rely on systematic collection of microbial samples that are processed in order to monitor pathogen trends and produce early warnings in cases of outbreaks or rise in frequency of virulent/antibiotic-resistant strains. Recent advances in the field lead to a shift from traditional microbiological characterization to whole genome sequencing (WGS) methods. Each WGS sequencing dataset needs to be properly handled to produce high quality data with which epidemiologic studies can confidently work. However, the potential routine use of WGS analysis in public health real-time surveillance is hampered by the absence of accessible and established bioinformatics frameworks, as well as by limited bioinformatics skills in handling these novel methodologies. Moreover, the different bioinformatics pipelines applied by different laboratories hinder the sharing of data that would greatly benefit public health surveillance worldwide.

The production of high quality and comparable draft genomes data were the motivation for the development of a standardised, fully automate, flexible, portable and pathogen-independent bioinformatics pipeline for bacterial genome assembly. The INNUca pipeline processes raw sequencing data to de novo assembly and species confirmation. To achieve high quality standards, INNUca makes use of several already available tools considered as golden standards in de novo bacteria genome assembly production. The de novo assembly is performed with SPAdes, and then subsequently corrected using Pilon in order to significantly improve the draft genome by correcting bases, fixing mis-assemblies and filling gaps. Additionally, it can better estimate the true bacterial chromosome coverage via read mapping against reference gene sequences distributed throughout in the genome, avoiding coverage estimate bias introduced by mobile genetic elements. This module can also detect contamination with different strain or species. All the INNUca steps are subject to quality control using clearly defined thresholds to ensure data quality, resulting in a simple "FAIL/PASS" flag for each module, which are compiled in an endpoint report.

INNUca was developed in the scope of “INNUENDO - a new cross-sectorial platform for genomics integration in surveillance of food-borne pathogens” (financed by a Thematic Grant of the European Food Safety Authority) (https://sites.google.com/site/innuendocon/), and can be found in https://github.com/B-UMMI/INNUca.

16. Biodiversity of Cultivable Microorganisms from Antarctica

Adriana Isabel Rego et al (FCUP/CIMAR)

Microorganisms represent an extremely rich reservoir of potentially valuable natural small molecules on the planet, such as polyketides, nonribosomal peptides, terpenes and alkaloids, between other classes of compounds. Among microorganisms, Bacteria are the most prolific producers, including Cyanobacteria, which have proved an extremely rich source of secondary metabolites together with Actinobacteria, Firmicutes and Myxobacteria. One of the strategies employed to uncover new secondary metabolites relies on the study of microorganisms inhabiting extreme environments, as is expected that a large part of them are still unknown and possess unique adaptations to their habitats, including the production of unique chemical entities with unprecedented biological activities.

Here, we present our results on the microbial biodiversity of environmental samples collected in a cold hyper-arid polar desert, the McMurdo Dry Valleys, Antarctica. The 16S rRNA gene was amplified and sequenced using Roche's 454 pyrosequencing technology. Also, the samples were inoculated in different selective media with the final purpose to search for secondary metabolites produced by the isolated microorganisms. Samples were obtained from a rock with endolithic colonization and from a soil transect with decreasing water availability. From the rock, two endolithic cyanobacterial strains with high similarity to Leptolyngbya antarctica were obtained. It was found that these strains possess the diterpenoid dehydroabietic acid, a secondary metabolite from the Terpene family, reported by the first time to be present in Cyanobacteria [1]. From the soil samples, a diversity of Firmicutes species with high similarity to Paenisporosarcina macmurdoensis, Paenisporosarcina indica and Sporosarcina antarctica were isolated as well as two Fungi species, with high similarity to Penicillium citrinum and Dicyma pulvinata. Further, four cyanobacterial strains are in isolation process.

Future work will include the isolation and identification of the remaining strains, followed by large-scale cultivation and screening of the bioactivity of the compounds produced.

[1] Costa MS, Rego A, Ramos V, et al. The conifer biomarkers dehydroabietic and abietic acids are widespread in Cyanobacteria. Sci Rep. 2016;6(23436):1-11. doi:10.1038/srep23436.

17. Genome-wide profiling of alternative splicing in aging across human tissues

Ana Carolina Leote and Nuno Barbosa-Morais (IMM)

One of the biggest modern challenges lies in keeping an increasingly aged population healthy [1]. In order to do so, a clear knowledge of the human aging mechanisms is necessary. Alternative splicing (AS), a tightly regulated process by which a single gene can give origin to multiple proteins, has recently been implicated in the aging of humans and mice. Age related changes in expression of genes involved in mRNA processing, as well as an increase in the number of alternatively spliced genes in tissues of aged mice, affecting RNA processing pathways, have been described [2,3], pointing to an important and still partially undisclosed role of AS in the aging process.

In this work, RNA sequencing data collected by the Genotype-Tissue Expression (GTEx) project from over five hundred donors were analyzed using the R language and AS signatures of the aging process were derived for human healthy tissues. Those signatures point to aberrant splicing of spliceosome components in most tissues with the exception of brain, where neuronal functions are the most affected. Tissue-specific models based on the AS signatures were used to estimate tissue biological age and compare aging trends between tissues, grouping brain regions, as well as circulatory and respiratory systems. Our analyses also suggest an increase of variability in alternative exon usage with age in most tissues, except for colon, kidney and salivary gland. Furthermore, samples from cells progressing into replicative senescence were analyzed to obtain an AS signature of cellular senescence which was used to quantify the contribution of this process to human aging.

These results highlight the importance of AS in the aging process across tissues and provide novel insights into the aging mechanisms.

[1] Harper, S. Economic and social implications of aging societies. Science 346, 587–591 (2014).

[2] Harries, L. W. et al. Human aging is characterized by focused changes in gene expression and deregulation of alternative splicing. Aging Cell 10, 868–878 (2011).

[3] Rodríguez, S. A. et al. Global genome splicing analysis reveals an increased number of alternatively spliced genes with aging. Aging Cell 15, 267–278 (2016).

18. Development of text mining tools for information retrieval and extraction from patents

Tiago Alves, Hugo Costa and Miguel Rocha

Biomedical literature is composed of a large and ever increasing number of publications, written in natural language. Patents are a relevant fraction of these publications, considered important sources of information due to all the curated information available in the documents, from the granting process. Although being real technological libraries, their unstructured data turns the search of information within these documents a challenging task. Biomedical text mining is a scientific field that explores this task, creating methodologies to search and structure the information in the biomedical literature.

Information retrieval (IR) is one of the biomedical text mining tasks, in which the relevant information is obtained from an extensive collection of documents using several text retrieval methodologies. Getting all the information available on a patent document requires the download of all the bibliographic information as well as the respective patent PDF document, that is then converted into a machine-readable text by technologies as Optical Character Recognition (OCR).

To achieve that goal, a “patent pipeline” was developed and integrated into @note2, an open-source computational framework for biomedical text mining. The patent pipeline can be disintegrated into four different tasks that were used to build different new processes on @note2: the patent search and the retrieval of patent metadata (relative to bibliographic data) were introduced as a new IR Search tool and the retrieval of patent PDF files as well as the subsequent extraction of all the information from them were introduced as a new IR Crawling and a new PDF to text conversion processes, respectively. A set of patents from the BioCreative V CHEMDNER task was used to test the developed pipeline, evaluating the framework performance and the real capacity to retrieve the requested patents and extract their unstructured information to machine readable text. The results were promising, and showed that is possible bringing the published patent information to the scientific community, allowing the posterior implementation of other biomedical text mining processes over these documents, automating the search of structured information on patents and taking advantage of all the great informative capacity applied on them.

19. Evidence for lineage-specific CD4+ T cell-driven diversifying selection in Mycobacterium tuberculosis pinpoints targets of co-evolution with humans

Carlos Magalhães, Iñaki Comas, Gil Castro, Jorge Pedrosa, Sebastien Gagneux, Margarida Saraiva and Nuno Osório

Tuberculosis (TB), a devastating disease caused by bacteria from the Mycobacterium tuberculosis complex (MTBC), remains vastly uncontrolled. There is increasing evidence that the long parallel evolution of MTBC and the human host is in fact co-evolution. The precise identification of the molecular bases underlying immune-related co-evolution is of major interest, as it may lay the foundations for novel and more effective TB vaccines. We investigated the existence of MTBC-varying CD4+ T cell epitopes by analyzing the diversity, evolution and immunogenicity of peptidic sequences encoded by 270 MTBC genomes. We found specific aminoacid residues that were predicted to be under diversifying selection and to induce large alterations in the number of encoded peptides with predicted high binding affinity to class II human leukocyte antigen (HLA) proteins. Overall, our findings highlight potential targets of co-evolution that may have been selected to modulate host CD4+ T cell-mediated immunity. Our study uncovers unprecedented paths to exploit evolving pathogen loci in favor of the host, namely for the development of more efficient vaccines and lineage-specific diagnostic tools.

Sponsors

COMBOIOS DE PORTUGAL

geneious
Inspirational software for biologists!

DELTA
CAFÉS

COPYSCAN
CENTRO DE CÓPIAS

ADEGA DE
FAVAIOS
1952 - 2012
60 ANOS

TRILHAMUNDOS

coimbra
genomics
genetically informed medicine

NOVO BANCO

BioDiamante Extra
Rua da Adega S° Pedro | 4750 - 842 Barcelos
Telef: 253 813 609 | Telem: 960 150 161

Padaria
Velha
Padaria Jardim
Velho Unipessoal, Lda

padarias pastelarias
Montalegrense

Maria Bolacha

Allkits
CRIATIVAMENTE DIFERENTE

SG
Santa Eugénia
Restaurante - Pastelaria
Padaria - Pizzaria

PARTNERS

Universidade do Minho
Escola de Engenharia

Ciências
ULisboa
Faculdade
de Ciências
da Universidade
de Lisboa

NEBIUM