Interuniversity PhD program in Bioinformatics

Annual Internal Workshop

Faculty of Science

Universitat de Girona

Girona, February 2nd 2024

Book of abstracts

UAB Universitat Autònoma de Barcelona	Universitat de Barcelona	Universitat de Girona	Universitat de Lleida
UVIC UNIVERSITAT DE VIC UNIVERSITAT CENTRAL DE CATALUNYA	UOC Universitat Oberta de Catalunya	UPC UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH	UNIVERSITAT ROVIRA i VIRGILI

In collaboration with:

BIOINFORMATICS BARCELONA

Preface

This volume contains papers that had to be presented at the One-day Workshop of the PhD Interuniversity Doctorate Programme in Bioinformatics 2024 (PhDBioinformatics2024) celebrated in Girona, 2 February 2024. The Interuniversity Doctorate Programme in Bioinformatics is an official program jointly organised by the Universitat Autònoma de Barcelona (UAB), the Universitat Politècnica de Catalunya (UPC), the Universitat de Girona (UdG), the Universitat de Lleida (UdL), the Universitat Oberta de Catalunya (UOC), the Universitat de Vic – Universitat Central de Catalunya (UVic-UCC), the Universitat de Barcelona (UB), and the Universitat Rovira i Virgili (URV) with the participation of the Bioinformatics Barcelona Association (BIB). Each year, the main activity for the PhD training consists in a workshop to present the current status of the different PhD projects to develop networking among the participants and their research competences.

This workshop followed PhDBioinformatics2023 (UVic-UCC, Vic), PhDBioinformatics2021 (UdL, virtual), PhDBioinformatics2020 (UAB, Cerdanyola), and PhDBioinformatics2019 (UOC, Barcelona).

Beatriz López, Jose Luis Garcia Marin, Arnau Oliver, Sílvia Osuna
February 2024

Organised by

The organising committee is composed by the members of the academic commission of the PhD program:

- Margarida Julià, coordinator of the PhD program, UAB
- Xavier Daura, former coordinator, UAB, BIB
- Rui Alves, UdL
- Sergio Gómez, URV
- Beatriz López, UdG
- Marta Casanellas, UPC
- Ferran Prados, UOC
- Alexandre Sànchez, UB
- Jordi Villà-Freixa, UVic-UCC

The secretariat is composed by:

- Beatriz López, UdG
- Jose Luis Garcia Marin, UdG
- Arnau Oliver, UdG
- Sílvia Osuna, UdG
- Jonah Fernández, UdG
- Guillem Hernández, UdG
- Núria Pérez, UdG
- Oscar Raya, UdG
- Janet Sanchez, UdG

The event is supported by

Universitat de Girona
Escola de Doctorat

with the collaboration of the UdG research groups:

eXiT
Control Engineering an
Intelligent Systems

VICOROB

CompBioLab

Table of Contents

Program

8:30-9:15	**Registration**
09:15	**Opening** Dr. Gerardo Boto (Director Director of the PhD School, UdG) Dra. Margarida Julià (Coordinator of the Interuniversity PhD Program, UAB) Dra. Beatriz López (Organizing Committee, UdG)
9:30-11:15	**Session 1. Chair: Xavier Daura**
09:30	**Invited Talk.** Computational enzyme design: Towards the development of fast yet accurate approaches. **Silvia Osuna** (ICREA, Institut de Química Computacional i Catàlisi and Departament de Química, UdG)
10:15	Assessing Cell-Penetrating-Peptide Potential using Computational Electrophysiology **Èric Catalina Hernández** (Biophysics Unit, Biochemistry and Molecular Biology Department, Institute of Neuroscience, UAB)
10:35	Empirical Valence Bond Simulations of Glutathione Peroxidase **Nayanika Das Sekha**r (Research Group on Bioinformatics and Bioimaging (BI2); UVic -UCC)
10:55 - 11:50	**Coffe break and poster session S1**
10:50 - 13:10	**Session 2. Chair: Ferran Prados**
11:50	The landscape of gastrointestinal stromal tumour (GIST) progression uncovers a mutually exclusive chromosomal instability (CIN)-dependent and CIN-independent tumour evolution **David Gómez Peregrina** (Sarcoma Translational Research Laboratory, Vall d'Hebron Institute of Oncology (VHIO), Barcelona)
12:10	A synthetic data generation system for myalgic encephalomyelitis / chronic fatigue syndrome questionnaires **Marcos Lacasa** (e-Health Center, UOC)
12:30	Single-cell and spatial transcriptomic characterization of treatment resistance in high-grade serous ovarian cancer **Kathleen Imbach** (Josep Carreras Leukemia Research Institute, UAB)
12:50	Annotation of known molecules from MS2 spectra using a deep learning model based on Mol2vec and a Convolutional Neural Network (CNN) **Muhammad Faizan Khan** (Departament D'enginyeria electrònica Elèctrica i Automàtica, URV)
13:10-14:30	**Lunch time**
14:30-17:15	**Session 3. Chair: Sergio Gómez**
14:30	Modeling the effects of strigolactone levels on maize root system architecture **Abel Lucido Garbulo** (Systems Biology Group, Department Ciències Mèdiques Bàsiques, Faculty of Medicine, UdL, IRBLleida)
14:50	A BERT base model for the analysis of Electronic Health Records from diabetic patients **Enrico Manzini** (B2SLab, Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, UPC, Networking Biomedical Research Centre in the subject area of Bioengineering, Biomaterials and Nanomedicine, Madrid; Institut de Recerca Sant Joan de Deu, Barcelona)
15:10	Deciphering the role of spatio-temporal genome architecture in B cell differentiation **Laureano Tomás Daza** (Josep Carreras Leukaemia Research Institute, Badalona, Barcelona Supercomputer Center)
15:30-16:30	**Coffee break and poster session S2**
16:30-17:30	**Session 4. Chair: Rui Alves**
16:30	**Invited talk:** Enhancing Spatial Transcriptomics Resolution with Machine Learning. **Albert Pla Planas** (Computational Science Director - D4NT, Digital R&D, Sanofi)
17:15	Best poster announcement and closing. **Presenter: Xavier Daura**

Posters and other abstracts

Acedo Terrades	Ariadna	Targeting the WNT signaling pathway: a novel predictive signature for neoadjuvant chemotherapy response in muscle-invasive bladder cancer
Aviñó	Laura	Disentangling dynamic gene expression patterns from tissue movements: a computational approach
Bagué	Jaume	Personalized Medicine in Melanoma: biomarkers of prognosis and response to immunotherapy, and its relation to dietary habits and physical exercise
Balagué Dobón	Laura	Common genetic variants associated with urinary phthalate levels in children: a genome-wide study
Bárcenas López	Oriol	Structural determinants of α-synuclein binding to an inhibitory peptide studied by molecular dynamics simulations
Basallo Clariana	Oriol	Modeling the effects of circadian rhythm on the two alternative pathways for terpenoid precursor biosynthesis in plants
Butjosa Espín	Maria	Differences in silico in drug response between primary and metastatic cancer
Cabrera Gumbau	Jordi Manu	Machine learning for early prediction of vibrio vulnificus infections in the US
Canal Noguer	Pol	Methylation biomarkers for colorectal cancer early detection and survival prognostics impact gene expression and link to cancer-related biological pathways
Casals Franch	Roger	Gene expression prediction under novel conditions using ATAC-seq-informed regulons
Casanova Suárez	Carla	Diet adaptations in anatomically modern humans
Colomer i Vilaplana	Aina	Evaluating allele frequency trajectory and selection coefficient estimates from genealogies including ancient DNA
Diaz Hurtado	Marcos	Longitudinal segmentation of multiple sclerosis lesions
Diaz Ros	Maria	Conservation and evolution of human segmental duplications in mammal genomes
Engler	Camila	protAGOnist: an innovative NLS/NES prediction tool
Espin Garcia	Roderic	Pan-cancer vulnerability prompted by TGF⍰-hypoxia-mediated suppression of alternative end-joining
Fariña Morillas	Maria José	Epigenetic relationships to improve Synthetic Lethality prediction model for cancer treatment
García Illarramendi	Juan Manu	In-silico simulation and efficacy evaluation of anti-PD1 treatment on 4 triple-negative breast cancer molecular subtypes
Genius Serra	Patricia	Lipid mechanisms drive cerebrovascular disease in cognitively unimpaired individuals at low risk for late-life dementia
Giné Bertomeu	Roger	Datoma: A cloud computing platform for high-performance metabolomics data analysis
González Barriada	Rubén	Deep Learning Based Methods for Fundus Image Quality Evaluation
Lisa Molina	Julia	Bioactive peptides in Mediterranean plants: biological properties and pharmacological implications
Llano Viles	Joshua	Identification of epigenetic biomarkers for molecular subgrouping of ependymoma
Loreto Velázquez	Antonio	Evolution of morphological complexity under development-based genotypephenotype maps
Marín Montes	Raúl	Evolution of morphological complexity under development-based genotypephenotype maps
Martinez	Emilce Sole	Multimodal data integration to model, predict, and understand changes in plant biodiversity
Mateo Navarro	Ramon	dsFDL: DataSHIELD Federated Deep Learning for Secure and Collaborative AI in Healthcare
Mitjavila Ventura	Adrià	Transposons in the evolution of piRNA cluster expression in mice
Moreno Farre	Javier	Molecular Dynamics study of the kynureninase enzyme: an approach for the design of new therapeutic enzymes in cancer
Muntañola Valero	Cristina	Prognosis of patient groups with COVID-19, chronic diseases and polypharmacy. Mixed patient-centered approach
Munzón Gil	Daniel	Evaluation of the msGBS methodology for the taxonomic identification and quantification of diatoms
Muñoz López	Francisco N	Identification of clinical features associated with Sars-Cov-2 reinfections
Nieto	Andrea	Characterising the regulation B cell differentiation at a single cell resolution

Pelegrí	Maria Dolo	BigDataStatMeth: An R package to implement statistical methods for Big Data
Pérez-López	Carlos	Aquasearch: a new software for fast proteomic characterization and classification of wastewater samples analyzed using MALDI-TOF
Pintado Grima	Carlos	aSynPEP-DB: a database of biogenic peptides for inhibiting α-synuclein aggregation
Pose	Iria	Machine learning approaches for the characterization of COPD
Ratarac	Aleksa	Fly wing development in silico: A computational investigation of morphological plasticity in Drosophila wings
Ruiz	Gabriel	Extrapolation of pathogenicity between homologous variants
Sanchez Hernandez	Janet	Computational design of ganciclovir-dependent kinases for suicide cancer gene therapy
Sánchez Herrero	Sergio	Integrating Artificial Intelligence Methods in Pharmacokinetics & Pharmacodynamics Processes
Sarrat González	David	CADSETshield: Developing a secure and efficient platform for integrated medical imaging and genomic studies of COPD using DataSHIELD and OMOP CDM
Serrano Gómez	Gerard	Multi-omics microbiome dynamics in IBD
Tejada Gutiérrez	Eva Luz	Development and application of tools for automated integration and analysis of big data in forestry management
Tejero Laguna	Eudald	Reanalysis of next generation sequencing data from patients with cardiac diseases
Temprano Sagrera	Gerard	Identification of differential expressed genes between abdominal aortic aneurysm cases and controls in aortic tissue
Tarradas-Alemany	Maria	Discovering and tracking potential zoonotic species from metagenomic samples with a capture-based oriented pipeline
Tondar	Abtin	Identification of new BCL-2 inhibiting small molecules using machine learning, molecular docking, and MD simulation
Yakymenko	Illya	Characterization of strategies for structural variant imputation
Yang	Jing	A quantitative view of the heterogeneity-diversity axis in biological systems

Keynote 1.
Computational enzyme design: Towards the development of fast yet accurate approaches

Silvia Osuna[1,2]

1. Institut de Química Computacional i Catàlisi and Departament de Química, Universitat de Girona, Spain

2. ICREA, Barcelona, Spain2Affiliation

Abstract

Enzymes are essential for supporting life by accelerating chemical reactions in a biologically compatible timescale. These remarkable catalysts possess unique features like high specificity and selectivity, and they function under mild biological conditions. These extraordinary characteristics make the design of enzymes for industrially relevant targets highly appealing.

Enzymes exist as an ensemble of conformational states, and the populations of these states can be altered through substrate binding, allosteric interactions, and even by introducing mutations into their sequence. These conformational states can be altered through mutations, which facilitates the evolution of enzymes towards acquiring novel activities.[1] Interestingly, many laboratory-evolved enzymes exhibit a common pattern—a significant impact on the catalytic activity is often observed due to remote mutations located distal from the catalytic center.[2] Similar to allosterically regulated enzymes, distal mutations play a role in regulating enzyme activity by stabilizing pre-existing conformational states that are crucial for catalysis.

In this talk, the rational approaches we have developed for enzyme design along the years will be discussed. These approaches rely on inter-residue correlations derived from microsecond time-scale Molecular Dynamics (MD) simulations, enhanced sampling techniques, and more recently, the incorporation of AlphaFold2 predictions [1-4]. Over the years, our research on various enzyme systems has provided compelling evidence that the current challenge of predicting distal active sites to enhance functionality in computational enzyme design can ultimately be addressed [3].

References

[1] Maria-Solano, M. A.; Serrano-Hervás, E.; Romero-Rivera, A.; Iglesias-Fernández, J.; Osuna, S. Role of conformational dynamics for the evolution of novel enzyme function, *Chem. Commun.* **2018**, 54, 6622-6634.

[2] Osuna, S. The challenge of predicting distal active site mutations in computational enzyme design, *WIREs Comput Mol Sci*. **2020**, e1502.

[3] Casadevall, G.; Duran, C.; Osuna, S. AlphaFold2 and Deep Learning for Elucidating Enzyme Conformational Flexibility and Its Application for Design, *JACS Au* **2023**, DOI: 10.1021/jacsau.3c00188.

[4] Casadevall, G.; Duran, C.; Estévez-Gay, M.; Osuna, S. Estimating conformational heterogeneity of tryptophan synthase with a template-based AlphaFold2 approach, *Prot. Sci.* ***2022**, 31, e4426.*

Keynote 2.
Enhancing spatial transcriptomics resolution with machine learning

Albert Pla Planas, PhD

Computational Science Director – D4NT, DIGITAL R&D, Sanofi

Abstract

Spatial transcriptomics is emerging as a pivotal tool in bioinformatics, providing the ability to analyze the cellular composition of a tissue within its spatial context. These techniques, which bridge microscopy imaging and omics, is being instrumental in advancing our understanding of cellular interactions and functions in complex biological environments. Despite its transformative potential, spatial transcriptomics techniques like 10x Visium face significant challenges, both in terms of resolution and genome coverage. In this context, Machine learning methods offer great opportunities to overcome such limitations.

Artificial intelligence helps improving the interpretation of high-resolution microscopy image and analyzing the Artificial intelligence helps improving the interpretation of high-resolution microscopy image and analyzing the transcriptomics data associated to it. In this talk we will present how combining recent cell deconvolution algorithms and deep learning-based cell segmentation models we can estimate the cell composition of a tissue. By reaching pseudo-single cell resolution, we can significantly improve the interpretability of spatial transcriptomic data. This enables the identification of cellular interactions and colocation patterns in complex diseases like cancer or immune diseases

Oral presentations

Assessing Cell-Penetrating-Peptide Potential using Computational Electrophysiology

Eric Catalina-Hernandez[1], Mario Lopez-Martin[1], Alex Peralvarez-Marin[1]

1Biophysics Unit, Biochemistry and Molecular Biology Department, Institute of Neuroscience, Universitat Autonoma de Barcelona (Cerdanyola del Valles, Barcelona, Spain)

Abstract

The cell membrane is a highly selective and dynamic barrier that encloses the contents of all living cells, while regulating the flux of species between intra- and extra cellular compartments. Composed predominantly of phospholipids, it gains its selectively permeable nature. Cell penetrating peptides (CPPs) are small, positively charged peptides capable of traversing the cell membrane without inducing cellular toxicity (Deshayes *et al.*, 2004). CPPs demonstrate great potential in the delivery of various cargo such as proteins, nucleus acids, or nano particles, providing CPPs with substantial potential across various fields. The penetration mechanisms described are passive diffusion, pore formation, translocation, and endocytosis.

To assess whether a peptide possesses CPP-like capabilities, *in vivo* experiments can be performed. Nonetheless, this method does not allow the description of the translocation mechanism, which can be achieved through molecular dynamics (MD) simulations. Regrettably, to the best or our knowledge, there are no existing tools to conduct such MD experiments.

In this study, we introduce a novel approach involving a double membrane composition to perform Computational Electrophysiology (CompEL) for identifying peptides with these abilities (Kutzner *et al.*, 2011). Moreover, we have modelled nine-mers based on Arg9, a described CPP, and Leu9, a totally hydrophobic peptide. We have established a peptide continuum benchmark between these two ends, allowing us to determine the stage at which step Arg9 loses its CPP-like properties. The proposed benchmark opens the window for rational design of peptides with CPP and/or membrane disruptive potential.

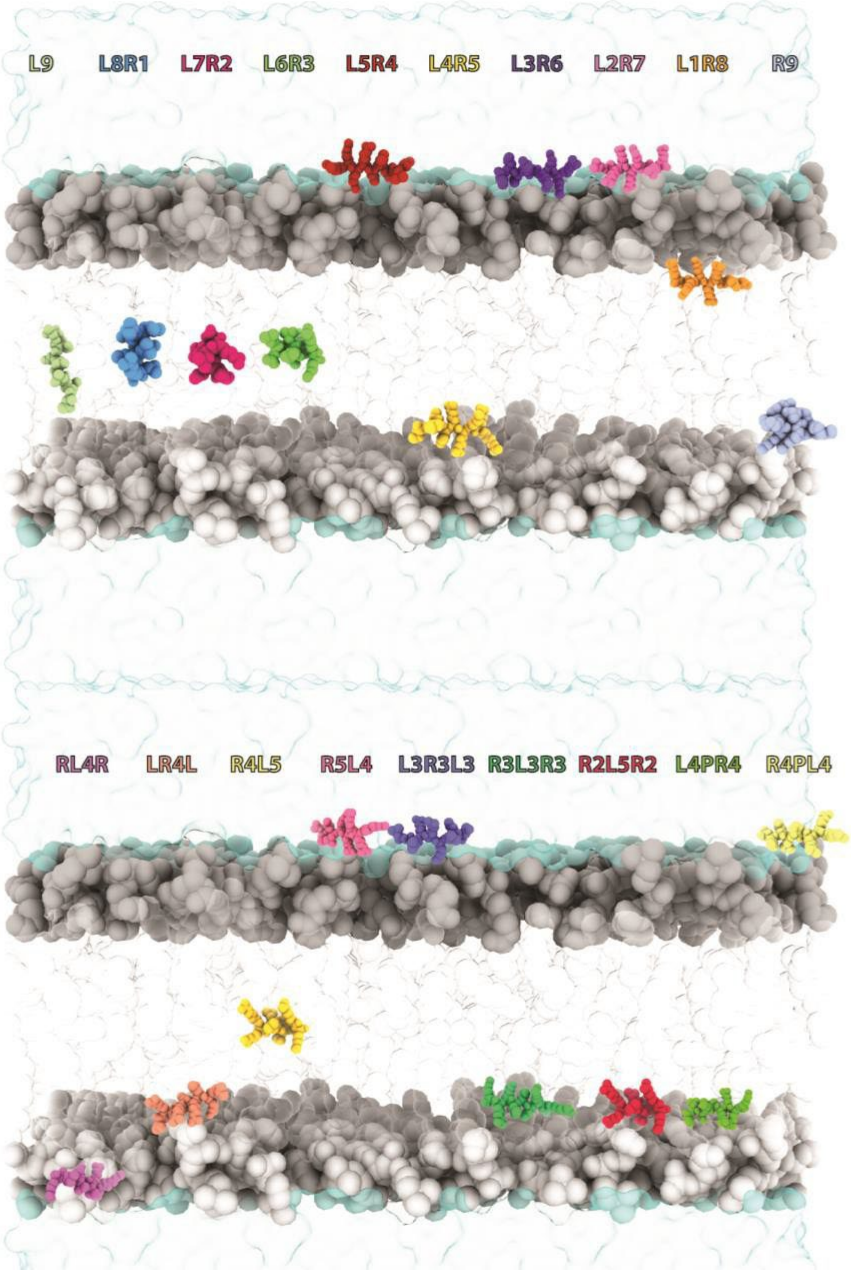


Figure 1. Illustrative representation of the peptides after the Computational Electrophysiology representation. 3 behaviors can be described: perturbation, insertion, and translocation.

References

Deshayes,S. *et al.* (2004) Insight into the Mechanism of Internalization of the Cell-Penetrating Carrier Peptide Pep-1 through Conformational Analysis. *Biochemistry*, **43**, 1449–1457.

Kutzner,C. *et al.* (2011) Computational Electrophysiology: The Molecular Dynamics of Ion Channel Permeation and Selectivity in Atomistic Detail. *Biophys J*, **101**, 809.

Empirical Valence Bond Simulations of Glutathione Peroxidase

Nayanika Das[1] , Jordi Villà-Freixa[1,2], Vijay Baladhye[3]

1 Research Group on Bioinformatics and Bioimaging (BI2); Facultat de Ciències, Tecnologia i Enginyeries; Universitat de Vic -Universitat Central de Catalunya, Vic, Barcelona

2 Institut de Recerca i Innovació en Ciències de la Vida i de la Salut a la Catalunya Central

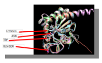
3 Savitribai Phule Pune University, Pune, India

Abstract

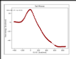
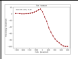
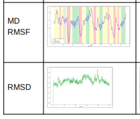
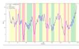
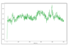
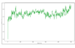
Selenoproteins are broadly divided into three families such as Glutathione peroxidases (GPXs), Thioredoxin reductases (TRs) and Iodothyronine deiodinases (DIOs) [1] [2]. GPX6 is a selenoprotein in humans and cysteine in rodents.[3] Preliminary experimental and computational results show that catalytic activity of several reconstructed ancestral structures of GPX6 recover their peroxidase activity when the active site is mutated from Cys to Sec keeping the binding of glutathione in all cases. [4]. Our goal is to study the epistasis linked to the accumulation of amino acid variants that may explain the presence and absence of peroxidase activity. Such effort can only be achieved by running molecular dynamics on simplified yet informative QM/MM and EVB (Empirical Valence Bond Simulations) based potential energy surfaces that allow for the exploration of the reaction free energy landscapes. We have first obtained molecular dynamics simulations in the ground state. Further, the concerted mechanism of the reaction has been used as the basis for the EVB simulations to obtain free energy profiles of the Cys/Sec GPX6 protein in mammals and mouse. Here we show the first results of our attempt to characterize the evolutionary landscape of reaction free energies.

All the data that we will use will be public data repositories (in particular, at the Repositori de Dades de Recerca, CSUC) and all the code used will be uploaded to a github repository. (https://github.com/ND7996)

Structural homology in GPX isoforms



MSA of mouse and human sequence

	Mouse	Mammals
EVB plots		
MD RMSF		
RMSD		

References

1. Hubert N, Walczak R, Sturchler C, Myslinski E, Schuster C, Westhof E, Carbon P, Krol A. RNAs mediating cotranslational insertion of selenocysteine in eukaryotic selenoproteins. Biochimie. 1996;78(7):590-6. doi: 10.1016/s0300-9084(96)80005-8. PMID: 8955902

2. Chen, Y. C., Sandeep Prabhu, K., & Mastro, A. M. (2013). Is selenium a potential treatment for cancer metastasis? In Nutrients (Vol. 5, Issue 4, pp. 1149–1168). MDPI AG. https://doi.org/10.3390/nu5041149

3. Scheerer, P., Borchert, A., Krauss, N., Wessner, H., Gerth, C., Höhne, W., & Kuhn, H. (2007). Structural basis for catalytic activity and enzyme polymerization of phospholipid hydroperoxide glutathione peroxidase-4 (GPX4).https://doi.org/10.1021/bi700840d

4. Ancient loss of catalytic selenocysteine spurred convergent adaptation in a mammalian oxidoreductase. Rees J Sarangi G Cheng Q Floor M Andrés A MiguelB Villà-Freixa JSj Arnér E Castellano S. Doi: 10.1101/2023.01.03.522577

The landscape of gastrointestinal stromal tumour (GIST) progression uncovers a mutually exclusive chromosomal instability (CIN)-dependent and CIN-independent tumour evolution

David Gómez-Peregrina[1], Alfonso García-Valverde[1], Sarah E McClelland[2], Nischalan Pillay[3], Christopher D Steele[4], Camille Stephan-Otto Attolini[5], Sebastian Bauer[6], Chiara Colombo[7], Piotr Rutkowski[8], Armelle Dufresne[9], Patrick Schöffski[10], Kjetil Boye[11], Pablo Marín-García[12], Marta Gut[13,14], Anna Esteve[13,14], Genís Parra[13,14], Claudia Valverde[15], Iván Olivares-Rivas[1], Jordi Rosell[1], Gemma Mur[1,16], César Serrano[1,15]

1. Sarcoma Translational Research Laboratory, Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain.
2. Barts Cancer Institute, Queen Mary University of London, Charterhouse Square, London, EC1M6BQ, UK.
3. Research Department of Pathology, Cancer Institute, University College London, London, WC1E 6BT, UK.
4. Department of Cellular and Molecular Medicine, UC San Diego, La Jolla, CA, 92093, USA.
5. Institute for Research in Biomedicine (IRB Barcelona), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain.
6. Department of Medical Oncology, Sarcoma Center, West German Cancer Center, University Duisburg-Essen, Medical School, Essen, Germany.
7. Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy.
8. Department of Soft Tissue-Bone Sarcoma and Melanoma, Maria Sklodowska-Curie National Research Institute of Oncology, Warsaw, Poland.
9. Medical Oncology Department, Centre Léon Bérard, Lyon, France.
10. Department of General Medicine Oncology, Leuven Cancer Institute, University Hospitals Leuven, Leuven, Belgium.
11. Department of Oncology, Oslo University Hospital, The Norwegian Radium Hospital, Oslo, Norway.
12. Zetta Genomics, Cambridge, UK.
13. CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona, Spain.
14. Universitat Pompeu Fabra (UPF), Barcelona, Spain.
15. Department of Medical Oncology, Vall d'Hebron University Hospital, Barcelona, Spain.
16. Clinical trials office, Vall d'Hebron Institute Oncology (VHIO), Barcelona, Spain.

Keywords: Sarcoma, Chromosome instability, TKI resistance.

Abstract

Gastrointestinal stromal tumours (GIST) are malignant mesenchymal neoplasms arising from the interstitial cells of Cajal and are classically referred to as "karyotypically simple" sarcomas (Heinrich *et al.*, 2002). The most common initiating event is the oncogenic activation of KIT (80%) or PDGFRA (10%) tyrosine kinase receptors by gain-of-function mutations, remaining present throughout the course of the disease (Hirota, 1998; Corless *et al.*, 2011). The exquisite oncogenic addiction to KIT/PDGFRA signalling explains the exceptional benefits of Tyrosine Kinase Inhibitors (TKIs) in the clinical setting (Demetri *et al.*, 2002, 2006, 2013; Blay *et al.*, 2020). Although first-line TKI imatinib induces major responses in metastatic GIST (ORR ~70%, mPFS ~30mo) (Demetri *et al.*, 2002), the selective pressure exerted by targeted agents triggers the polyclonal expansion of drug-resistant KIT secondary mutations (~90%) and KIT- downstream driver alterations (<10%) (Serrano and George, 2020; Liegl *et al.*, 2008; Serrano *et al.*, 2023). These also constitute the main mechanisms of therapy aversion during the subsequent TKI treatment lines available for GIST treatment. In parallel with initial KIT/PDGFRA activating mutations, there is a well-established multi-step cytogenetic progression involving particular chromosomal regions that affects specific genes not yet fully understood: 14q deletion (MAX) -> 22q deletion (DEPDC5) -> 1p deletion -> 9p deletion (CDKN2A) -> Xp deletion (DMD) (Serrano and George, 2020). This successive loss of specific regions mostly promotes cell cycle dysregulation, increased proliferation, aggressiveness and metastatic spread (Romeo *et al.*, 2009). Although TKI resistance has been mostly attributed to KIT secondary mutations, it is not yet known whether this cytogenetic progression synergises with the reduced efficacy of subsequent TKI lines after imatinib failure (ORR <10%, mPFS 4-6mo) (Serrano and George, 2020), suggesting the

emergence of new molecular mechanisms leading an attenuation of KIT/PDGFRA drivers' dependency.

To unravel the biological processes underlying GIST evolution, we have collected a unique cohort of GIST tumours (N=68) sequenced by WES, RNAseq and clinical information. This series recapitulates the clinical-biological evolution history of GIST, from treatment-naïve localized tumours to multi-TKI refractory metastases. From this collection, together with parallel studies in additional GIST patient cohorts, it has surprisingly become apparent that GISTs are much more genomically complex sarcomas than previously realised as a result of chromosomal instability (CIN) (*Figure 1*). GISTs are mostly affected by relative genomic losses and to a lesser extent by amplifications throughout their genomes, resulting in fractions of the genome altered (FGA) ranging from 3% to 82% that correlate with aneuploid alteration scores. In addition to the already known frequently aneuploid chromosome arms, we have detected new recurrent losses (9p, 10p/q, 13q, 17p, 18p/q and 19p/q) and gains (5p/q, 8p/q). Additionally, ~70% of GIST tumours have undergone at least one whole-genome doubling (WGD), which can appear in both early and late stages of the disease. WGD can promote increased CIN and is associated with higher aneuploidy rates (mostly arm losses), fractions of the genome altered and enrichment of CNA signatures with focal CNA oscillating regions indicative of chromothripsis (Cortés-Ciriano *et al.*, 2020; Voronina *et al.*, 2020; Steele *et al.*, 2019, 2022). The presence of highly enriched chromosomal regions with loss of heterozygosity (LOH) with and without WGD events also hints at a potential role of LOH as a driver of late WGD in GIST (López *et al.*, 2020). However, the presence of some whole-genome doubled samples with low proportions of LOH and FGA may indicate that WGD is also tolerated in the early stages of CIN (Vittoria *et al.*, 2023). Interestingly, genomic features associated with CIN are enriched in high-risk localized and metastatic disease (also reported in Gorunova *et al.*, 2022; Namløs *et al.*), but no clear associations were found with TKI treatments. Moreover, we have indetified alterations in cell cycle processes and checkpoints, increased proliferative properties, disrupted p53-network activity (with nearly no mutations in p53 or related genes) and impaired DNA damage sensing and response, ultimately providing a CIN-permissive context. Finally, by integrating multi-omics data from CNA profiles and signatures, mutational data and transcriptomics profiles, we are investigating the ongoing cytogenetic evolution in GIST and describing potential novel driver candidates of tumour progression and CIN to develop new stratification and therapy strategies.

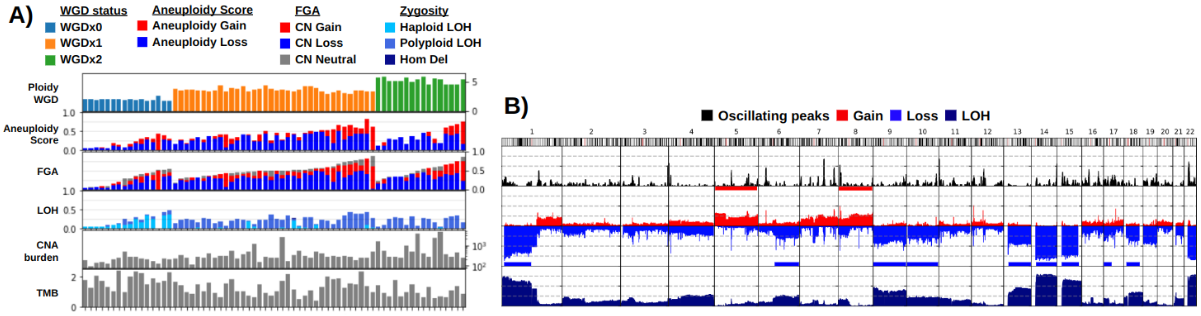


Figure 1: CIN landscape in GIST patients. A) Overview of main CIN readouts inferred from allele-specific CNA profiles of 68 GIST patients. The number of WGD events in each sample were calculated using a LOH-adjusted ploidy model (Steele *et al.*, 2019, 2022). Gains, losses and copy neutral (LOH regions without changes in copy number) segments were used to calculate the fraction of the genome altered (FGA) and aneuploidy scores (fraction of chromosome arms with >50% with losses/gains). Allele-specific CNA calls allowed to assess haploid and polyploid LOH regions throughout genomes. CNA burden was calculated by the total count of CNA segments per sample. Tumor mutation burden was calculated by the number of somatic mut/Mb adjusted by DNA content (tumor ploidy). **B)** Summary of recurrent regions with oscillating segments potentially associated with chromothripsis, gains/losses (with statistically significant enriched chromosome arm alterations in coloured lines) and LOH. Dashed horizontal lines represent percentiles 25%, 50%, 75%.

References

1. Blay,J.-Y. *et al.* (2020) Ripretinib in patients with advanced gastrointestinal stromal tumours (INVICTUS): a double-blind, randomised, placebo-controlled, phase 3 trial. *Lancet Oncol*, 21, 923–934.

2. Corless,C.L. *et al.* (2011) Gastrointestinal stromal tumours: origin and molecular oncology. *Nat Rev Cancer*, 11, 865–878.

3. Cortés-Ciriano,I. *et al.* (2020) Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat Genet*, 52, 331–341.

4. Demetri,G.D. *et al.* (2002) Efficacy and Safety of Imatinib Mesylate in Advanced Gastrointestinal Stromal Tumors. *New England Journal of Medicine*, 347, 472–480.

5. Demetri,G.D. *et al.* (2013) Efficacy and safety of regorafenib for advanced gastrointestinal stromal tumours after failure of imatinib and sunitinib (GRID): an international, multicentre, randomised, placebo-controlled, phase 3 trial. *The Lancet*, 381, 295–302.

6. Demetri,G.D. *et al.* (2006) Efficacy and safety of sunitinib in patients with advanced gastrointestinal stromal tumour after failure of imatinib: a randomised controlled trial. *Lancet (London, England)*, 368, 1329–38.

7. Gorunova,L. *et al.* (2022) Cytogenetic and molecular analyses of 291 gastrointestinal stromal tumors: site-specific cytogenetic evolution as evidence of pathogenetic heterogeneity. *Oncotarget*, 13, 508–517.

8. Heinrich,M.C. *et al.* (2002) Biology and genetic aspects of gastrointestinal stromal tumors: KIT activation and cytogenetic alterations. *Human Pathology*, 33, 484–495.

9. Hirota,S. (1998) Gain-of-Function Mutations of c-kit in Human Gastrointestinal Stromal Tumors. *Science*, 279, 577–580.

10. Liegl,B. *et al.* (2008) Heterogeneity of kinase inhibitor resistance mechanisms in GIST. *The Journal of Pathology*, 216, 64–74.

11. López,S. *et al.* (2020) Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. *Nat Genet*, 52, 283–293.

12. Namløs,H.M. *et al.* Chromosomal instability and a deregulated cell cycle are intrinsic features of high-risk gastrointestinal stromal tumours with a metastatic potential. *Molecular Oncology*, n/a.

13. Romeo,S. *et al.* (2009) Cell cycle/apoptosis molecule expression correlates with imatinib response in patients with advanced gastrointestinal stromal tumors. *Clin Cancer Res*, 15, 4191–4198.

14. Serrano,C. *et al.* (2023) Circulating tumor DNA analysis of the phase III VOYAGER trial: KIT mutational landscape and outcomes in patients with advanced gastrointestinal stromal tumor treated with avapritinib or regorafenib☆. *Annals of Oncology*, 0.

15. Serrano,C. and George,S. (2020) Gastrointestinal Stromal Tumor: Challenges and Opportunities for a New Decade. *Clinical Cancer Research*, 26, 5078–5085.

16. Steele,C.D. *et al.* (2022) Signatures of copy number alterations in human cancer. *Nature*, 1–8.

17. Steele,C.D. *et al.* (2019) Undifferentiated Sarcomas Develop through Distinct Evolutionary Pathways. *Cancer Cell*, 35, 441-456.e8.

18. Vittoria,M.A. *et al.* (2023) Whole-genome doubling in tissues and tumors. *Trends in Genetics*, S0168952523001877.

19. Voronina,N. *et al.* (2020) The landscape of chromothripsis across adult cancer types. *Nat Commun*, 11, 2320.

A synthetic data generation system for myalgic encephalomyelitis / chronic fatigue syndrome questionnaires

Marcos Lacasa[1], Ferran Prados[1,2,3,4], José Alegre[5], Jordi Casas-Roma[1]

1 - e-Health Center, Universitat Oberta de Catalunya, Barcelona, Spain

2 - Center for Medical Image Computing, University College London, London, United Kingdom

3 - National Institute for Health Research Biomedical Research Centre at UCL and UCLH, London, United Kingdom

4 - Queen Square MS Center, Department of Neuroinflammation, UCL Institute of Neurology, Faculty of Brain Sciences, University College London, London, United Kingdom

5- ME/CFS Unit, Division of Rheumatology, Vall d'Hebron Hospital Research Institute Universitat Autònoma de Barcelona, Barcelona, Spain

Citation (published 31 August 2023):

Lacasa, M., Prados, F., Alegre, J. *et al.* A synthetic data generation system for myalgic encephalomyelitis/chronic fatigue syndrome questionnaires. *Sci Rep* **13**, 14256 (2023). https://doi.org/10.1038/s41598-023-40364-6

Abstract

Background

Artificial intelligence or machine-learning-based models have proven useful for better understanding various diseases in all areas of health science. Myalgic Encephalomyelitis or chronic fatigue syndrome (ME/CFS) lacks objective diagnostic tests. Some validated questionnaires are used for diagnosis and assessment of disease progression. The availability of a sufficiently large database of these questionnaires facilitates research into new models that can predict profiles that help to understand the etiology of the disease. A synthetic data generator provides the scientific community with databases that preserve the statistical properties of the original, free of legal restrictions, for use in research and education.

Methods and Results

The initial databases came from the Vall Hebron Hospital Specialized Unit in Barcelona, Spain. 2522 patients diagnosed with ME/CFS were analyzed. Their answers to questionnaires related to the symptoms of this complex disease were used as training datasets. They have been fed for deep learning algorithms that provide models with high accuracy [0.69-0.81]. The final model requires SF-36 responses and returns responses from HAD, SCL-90R, FIS8, FIS40, and PSQI questionnaires.

Conclusions

A highly reliable and easy-to-use synthetic data generator is offered for research and educational use in this disease, for which there is currently no approved treatment.

Single-cell and spatial transcriptomic characterization of treatment resistance in high-grade serous ovarian cancer

Kathleen Jane Imbach[1,2], Vikram Naganathan[1], Arola Fortian[3], Daniela Grases[1], Adrià Bernat-Peguera[3], Sara Bystrup[3], Mustafa Sibai[1,2], Lorena Valdivieso Almeida[3], Margarita Romeo[4], Eduard Porta Pardo[1], Jordi Barretina[3]

1 Josep Carreras Leukemia Research Institute

2 Universitat Autonoma de Barcelona

3 Germans Trias i Pujol Research Institute

4 Institut Català d'Oncologia

Abstract

High-grade serous ovarian carcinoma (HGSOC) is the most common epithelial ovarian cancer, characterized by genetic alterations imposing DNA repair deficiency and poor prognosis. Elucidating the molecular mechanisms of cancer development, treatment resistance and immune response in HGSOC is essential to improve clinical targeting and patient outcomes. In this pilot study, we leverage the capabilities of single-cell (sc) and spatial transcriptomics (ST) (10X Genomics platform) to characterize tumor samples from 9 HGSOC patients with variable chemotherapy (CT) and PARP-inhibitor (PARPi) response. We analyze gene expression from the tumor landscape of 5 patient samples taken at diagnosis, 3 therapy-unresponsive patient samples taken after CT, and 3 patient samples taken after PARPi. Combining our sc data with existing data permits robust classification of cells for further characterization (Vazquez *et al.*, 2022). Transcription-based inference of clonal copy number variants in the cancer compartment supports commonly altered genes implicated in HGSOC (Patel *et al*., 2014; Smith *et al*., 2023), and clones exhibit distinct spatial distributions. Our findings indicate strong expression differences in both the malignant and tumor microenvironment compartments of patient tumors according to treatment. Overall, we show that this proof-of-concept undertaking can be expanded upon to further distinguish treatment response patterns in HGSOC.

References

1. Vázquez-García,I. *et al*. (2022) Ovarian cancer mutational processes drive site-specific immune evasion. *Nature*, **612**, 778-786.

2. Patel,A.P. *et al*. (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, **344**, 1396–1401.

3. Smith,P. *et al.* (2023) The copy number and mutational landscape of recurrent ovarian high-grade serous carcinoma. *Nature Communications*, **14**, 4387.

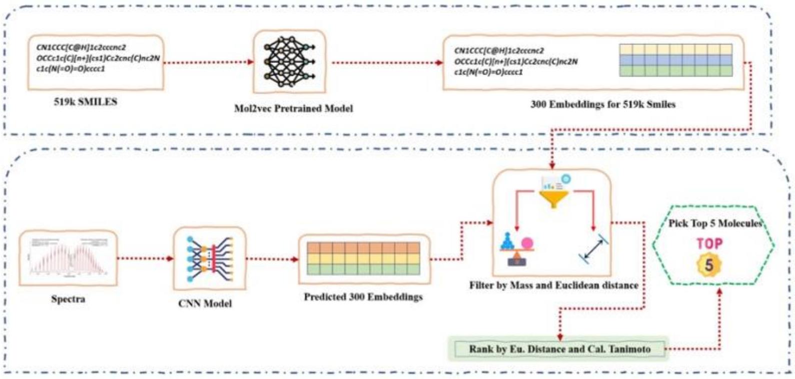
Annotation of known molecules from MS2 spectra using a deep learning model based on Mol2vec and a Convolutional Neural Network (CNN)

Muhammad Faizan Khan, Óscar Yanes, Roger Guimerà and Marta Sales-Pardo

1. Departament D'enginyeria electrònica Elèctrica i Automàtica, Universitat Rovira i Virgili

Abstract

Predicting the structure of a small molecule based on its tandem MS/MS spectrum is a challenging and unresolved task in metabolomics. Here, we have developed a deep learning model to predict molecules from MS/MS spectra. Our strategy consists of two blocks: first, we used the deep learning model Mol2vec (1) to obtain 300-features vector embeddings, capturing the chemical properties of a reference database comprising 519k molecules (SMILES). Next, we created a convolutional neural network (CNN) (2) model from the MS/MS spectra of 38k (positive mode) and 14k (negative model) unique compounds present in common mass spectral databases (NIST20, Agilent METLIN Metabolomics database, GNPS, and MSDial) as input data to predict Mol2vec vectors. The dataset was divided into training (80%), validation (10%) and test data (10%). The 300-features embeddings predicted for the *test data* were compared using the cosine similarity and Euclidian distance against our reference of 519k mol2vec embeddings. Finally, each pair of SMILES were ranked and the Tanimoto score was determined focusing on the top-1 and top-5 ranked molecules.



Using this method, we demonstrate that spectral information is critical: merging fragment ions from multiple MS/MS collision energies of the same small molecule improved the prediction of the Mol2vec vector by the CNN, as reflected in smaller Euclidian distances and higher cosine similarities with respect to the reference vector embeddings. Consequently, we went from a performance of 25.05% (top 1 hit) and 55.68% (top 5 hits) using individual spectra, to 40% (top 1 hit) and 73% (top 5 hits) using merged spectra. This can be further improved to 41% and 76% by adding neutral losses, defined as the mass difference between precursor and fragment ions.

Finally, our new method was applied to two real experimental datasets of unknown or non-annotated metabolites: the CASMI contests 2016 and 2022, and the Annotated Recurrent Unidentified Spectra (ARUS) from the NIST Mass Spectrometry Data Center.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 945413 and from the Universitat Rovira i Virgili (URV).

References

1. Jaeger, S., Fulle, S., & Turk, S. (2018). Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling*, *58*(1), 27-35.

2. Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., ... & Wilson, K. (2017, March). CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 131-135). IEEE.

Modeling the effects of strigolactone levels on maize root system architecture

Abel Lucido[1, 2], Johan-Fabian Andrade[1, 2], Oriol Basallo[1, 2], Abderrhamane Eleiwa[1, 2], Alberto Marin-Sanguino[1, 2], Ester Vilaprinyo[1, 2], Albert Sorribas[1, 2], Rui Alves[1, 2]

[1]Systems Biology Group, Department Ciències Mèdiques Bàsiques, Faculty of Medicine, Universitat de Lleida, Lleida, Spain

[2]IRBLleida, Lleida, Spain

Abstract

Maize is the most in-demand staple crop globally. Its production relies strongly on the use of fertilizers for the supply of nitrogen, phosphorus, and potassium, which the plant absorbs through its roots, together with water. The architecture of maize roots is determinant in modulating how the plant interacts with the microbiome and extracts nutrients and water from the soil. As such, attempts to use synthetic biology and modulate that architecture to make the plant more resilient to drought and parasitic plants are underway. These attempts often try to modulate the biosynthesis of hormones that determine root architecture and growth. Experiments are laborious and time-consuming, creating the need for simulation platforms that can integrate metabolic models and 3D root growth models and predict the effects of synthetic biology interventions on both, hormone levels and root system architectures. Here, we present an example of such a platform that is built using Mathematica. First, we develop a root model, and use it to simulate the growth of many unique 3D maize root system architectures (RSAs). Then, we couple this model to a metabolic model that simulates the biosynthesis of strigolactones, hormones that modulate root growth and development. The coupling allows us to simulate the effect of changing strigolactone levels on the architecture of the roots. We then integrate the two models in a simulation platform, where we also add the functionality to analyze the effect of strigolactone levels on root phenotype. Finally, using *in silico* experiments, we show that our models can reproduce both the phenotype of wild type maize, and the effect that varying strigolactone levels have on changing the architecture of maize roots.

References

1. Arite,T. *et al.* (2012) Strigolactone Positively Controls Crown Root Elongation in Rice. *J Plant Growth Regul*, **31**, 165–172.
2. Pagès,L. *et al.* (1989) A simulation model of the three-dimensional architecture of the maize root system. *Plant Soil*, **119**, 147–154.
3. Ruyter-Spira,C. *et al.* (2011) Physiological Effects of the Synthetic Strigolactone Analog GR24 on Root System Architecture in Arabidopsis: Another Belowground Role for Strigolactones? *Plant Physiology*, **155**, 721–734.

A BERT base model for the analysis of Electronic Health Records from diabetic patients

Enrico Manzini[1,2,3], Alexandre Perera Lluna[1,2,3]

1 B2SLab, Departament d'Enginyeria de Sistemes, Auto àtica i Informàtica Industrial, Universitat Politècnica de Catalunya, Barcelona, Spain.

2 Networking Biomedical Research Centre in the subject area of Bioengineering, Biomaterials and Nanomedicine, Madrid, Spain.

3 Institut de Recerca Sant Joan de Deu, Barcelona, Spain

Abstract

In recent years the digitization of health data and the increasing availability of Electronic Health Records (EHRs) is opening new opportunities -as well as new challenges- for improving the health care process through precision medicine (Noura 2019), i.e. that process of creating better diagnostic and treatment response models tailored on patients data. Deep learning (DL), a subfield of Machine Learning, is having a big impact in the way EHRs are analyzed and used to create personalized prediction models. In particular, DL has been shown to get better results compared to traditional approaches in different tasks: from disease detection to sequential prediction of clinical events; from data augmentation to concept embedding (Xiao 2018). One of the biggest limitations in training DL models for specific tasks is the availability of labeled data for training and validation. For this reason the concept of transfer learning, i.e. training a model on a generic domain to transfer this knowledge on a different, more specific, domain (Weiss 2016), is gaining more and more strength in DL research. This is especially true in the field of Natural Language Processing (NLP), where the BERT model achieved state of the art performances in different tasks. In this work we proposed a BERT based model designed to work with diagnosis and medicament codes and different continuous variables. Moreover we introduced: a state vector describing static information about the patient that helps the model to better learn the sequence of EHRs; and a mechanism of relative time attention based on the Relative Position Representation (RPR) (Shaw 2018), in order to better learn the irregularity of the data. Results of this model outperformed classical supervised learning techniques such as recurrent neural networks and random forest algorithms.

References

1. Noura S. Abul-Husn and Eimear E. Kenny (2019), Personalized medicine and the power of electronic health records, Cell, 177, 58–69

2. Cao Xiao, et al (2018), Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review, Journal of the American Medical Informatics Association, 25, 1419–1428

3. Karl Weiss, et al. (2016)A survey of transfer learning. Journal of Big Data, 3:9, 12

4. Peter Shaw, et al (2018) Self-attention with relative position representations. NAACL HLT 2018-Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 2, 464–468

Deciphering the role of spatio-temporal genome architecture in B cell differentiation

Laureano Tomás-Daza[1,2], Alfonso Valencia[2] and Biola M. Javierre[1]

1 Josep Carreras Leukaemia Research Institute, Badalona, Barcelona, Spain
2 Barcelona Supercomputer Center, Barcelona, Barcelona, Spain

Abstract

B cells harbour a vast diversity of antibodies that efficiently recognize specific pathogens, facilitating their neutralization and destruction. This diversification is the result of mutations and translocations in their immunoglobulin loci during the differentiation process. However, aberrations in this stage may lead to cancer.

Despite the clinical relevance of these processes, we do not have a complete understanding of the molecular mechanisms that regulate them. Our preliminary data suggest that the 3D chromatin organization plays a key role in the differentiation, but this still remains unexplored.

In this project we have addressed this gap of knowledge from the 3D chromatin modelling perspective(Di Stefano and Cavalli, 2022). We have implemented 3D chromatin modelling using structural data (top-down modelling) from our new method low input capture Hi-C (liCHi-C(Tomás-Daza *et al.*, 2023; Rovirosa *et al.*, 2023)) and we have also adapted the pipeline of 3D chromatin modelling using epigenomic data (bottom-up modelling) to gain resolution in the epigenomic composition of the monomers in the models.

Collectively, we have implemented a framework to work separately or together with both top-down and bottom-up modelling, to properly understand the structural and epigenomic contribution of the 3D genome in B cell differentiation.

References

Rovirosa,L. *et al.* (2023) An Integrated Workflow to Study the Promoter-Centric Spatio-Temporal Genome Architecture in Scarce Cell Populations. *J. Vis. Exp.*, **2023**.

Di Stefano,M. and Cavalli,G. (2022) Integrative studies of 3D genome organization and chromatin structure. *Curr. Opin. Struct. Biol.*, **77**.

Tomás-Daza,L. *et al.* (2023) Low input capture Hi-C (liCHi-C) identifies promoter-enhancer interactions at high-resolution. *Nat. Commun.*, **14**, 268.

Posters and other abstracts

TARGETING THE WNT SIGNALING PATHWAY: A NOVEL PREDICTIVE SIGNATURE FOR NEOADJUVANT CHEMOTHERAPY RESPONSE IN MUSCLE-INVASIVE BLADDER CANCER

Ariadna Acedo-Terrades[1], Júlia Perera-Bel[1], Marta Bódalo-Torruella[1], Maria Gabarós[1], Nuria Juanpere[1], Marta Lorenzo[1], Alejo Rodriguez-Vida[1], Oscar Buisan[2], Eulàlia Puigdecanet[3], Eduardo Eyras[1–4], Tamara Sanhueza[2], Lara Nonell[5], Joaquim Bellmunt[1–6].

[1]Hospital del Mar Medical Research Institute; Barcelona, Spain, [2]Germans Trias i Pujol Research Institute (IGTP); Badalona, Spain, [3]UVic-UCC; Barcelona, Spain, [4]EMBL Australia Partner Laboratory Network at the Australian National University; Canberra, Australia, [5]Vall d'Hebron Institute of Oncology (VHIO); Barcelona, Spain, [6]Division of Hematology and Oncology, Beth Israel Deaconess Medical Center; Boston, USA

In muscle-invasive bladder cancer (MIBC), neoadjuvant cisplatin-based chemotherapy (NAC) has become a standard of care prior to cystectomy for eligible patients based on the improved disease-specific and overall survival. Downstaging to non-MIBC at cystectomy leads to an enhanced outcome with 5-year overall survival of 80–90%. High-throughput DNA and RNA profiling technologies might help to overcome the inability to predict responders. Since most MIBC patients undergo NAC followed by cystectomy, pre-treatment tumor biopsy and post-chemotherapy cystectomy specimens are clinically available, creating an ideal setting to study the genomic and transcriptomic effects of NAC.

Here we present RNA sequencing of a cohort of 113 MIBC patients treated with NAC from different hospitals. For each patient, FFPE pre (n=71) and post-treatment (n=29) samples were obtained from biopsy and cystectomy respectively. Response (n=58) was defined as downstaging to non-MIBC (<pT2) at cystectomy. Differential expression analysis, GSEA, deconvolution and weighted correlation network analysis (WGCNA) was performed to assess differences between responders (R) and non responders (NR) in pre-treatment samples.

We found several differentially expressed (DE) genes (p.val < 0.05) upregulated in NR before treatment, associated with cancer growth and worse prognosis. On the other hand, R showed upregulated pathways related to the cell cycle. Interestingly, no differences were observed in immune cell proportions between the two groups. However, in the WGCNA, we identified a gene group negatively correlated with response, linked to crucial signaling pathways such as Wnt signaling and cell proliferation. WNT signature was obtained through performing the intersection between DE genes and genes related to several WNT pathways. This group of genes shows a significant correlation between low expression of those genes and overall survival, as well as response to NAC, in MIBC patients.

Disentangling dynamic gene expression patterns from tissue movements: a computational approach

Laura Aviñó[1], James Sharpe[1,2]

1 - European Molecular Biology Laboratory (EMBL) Barcelona, European Molecular Biology Laboratory, Barcelona, Spain.

2 - Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain.

Abstract

Organogenesis involves an interplay between growing/moving tissues, and dynamic gene expression patterns whose domains of expression sometimes move "through" the tissue. For model species which cannot be monitored by time-lapse imaging (such as the mouse) (Dalmasso et al., 2022) we must create an integrated framework which can capture, and distinguish between changes in expression pattern that are due to tissue movements, versus those which are due to active gene regulation. Here we present such a framework for early limb development, based on the integration of hundreds of images of gene expression patterns and previously computed tissue movements (Marcon et al., 2011). This has allowed us to create the first detailed reconstruction over time and space, for a handful of genes that are critical to limb development including the Sox9, the main skeletal marker.

Fig. 1

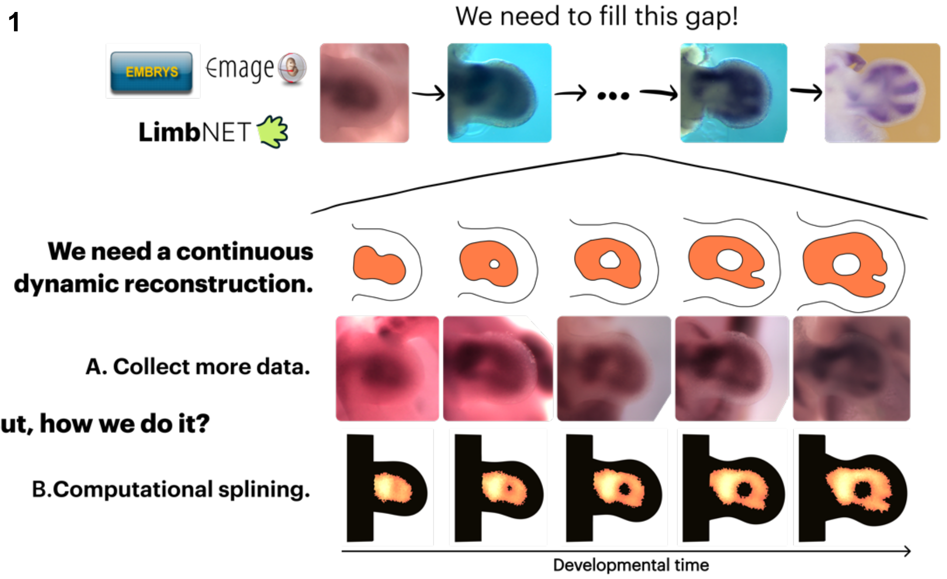

We need to fill this gap!
EMBRYS
Emage
LimbNET
We need a continuous dynamic reconstruction.
A. Collect more data.
But, how we do it?
B.Computational splining.
Developmental time

Fig. 1. Graphical representation of the main research question. Even though over the years the "limb development" community has collected hundreds of images of the gene expression patters ((Yokoyama et al., 2009), (Richardson et al., 2013)), these only show discrete snapshots. Hence, we don't have the complete trajectory of the gene expression patterns that is needed for its detailed computational modeling.

References

1. Dalmasso,G. *et al.* (2022) 4D reconstruction of murine developmental trajectories using spherical harmonics. *Developmental Cell*, **57**.

2. Marcon,L. et al. (2011) A computational clonal analysis of the developing mouse limb bud. *PLoS Computational Biology*, **7**.

3. Yokoyama,S. *et al.* (2009) A systems approach reveals that the myogenesis genome network is regulated by the transcriptional repressor RP58. *Developmental Cell*, **17**, 836–848.

4. Richardson,L. *et al.* (2013) Emage Mouse embryo spatial gene expression database: 2014 update. *Nucleic Acids Research*, **42**.

Personalized Medicine in Melanoma: biomarkers of Prognosis and Response to Immunotherapy, and its Relation to dietary habits and physical exercise

Jaume Bagué[1,2], Teresa Torres[1,2,3], Ferran Reverter[4,] Susana Puig[1,2,3],

[1]Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain

[2]Dermatology Department, Melanoma Unit, Hospital Clínic de Barcelona, Universitat de Barcelona, Barcelona, Spain

[3]CIBER of Rare Diseases (CIBERER), Instituto de Salud Carlos III, Barcelona, Spain

[4]Genetic, Microbiology and Statistic Department, Faculty of Biology Universitat de Barcelona, Barcelona, Spain

Abstract

This thesis aims to identify and validate prognostic and response classifiers for immunotherapy in patients with melanoma.

It has been observed that among patients in the initial stages of melanoma, there is a group that relapses after several years of surgical removal of the melanoma (on average between 3 and 4 years)[1]. There is currently no standardized clinical follow-up strategy for these patients.[2] Many monitoring strategies, such as using periodic imaging techniques, have not proven to be sufficiently cost-effective[3]. For this reason, distinguishing patients who are at greater risk of relapse after removal is key. On the other hand, some patients in more advanced stages, but who are already disease-free, are administered immunotherapy to prevent relapse (adjuvant therapy). This therapy has associated toxicities, and yet, about a 50% of these patients relapse despite following the treatment. Therefore, knowing the probabilities of a successful response to these therapies is key to prescribing the most appropriate therapy for each patient.

Methodologically, the first phase of this thesis will consist of creating classifiers based on protein biomarkers extracted from blood plasma and responses to a questionnaire on eating habits and physical exercise. Two classifiers will be implemented—one for prognostic purposes and the other for assessing responses to adjuvant immunotherapy. These classifiers will be conducted using mass spectrometry (LS-MS) followed by a directed proteomics technique (PRM-MS). The combination of proteomics and questionnaire results is expected to facilitate personalized treatment through precise classifiers[4]. The second phase will consist of transferring these classifiers to the clinic. This will be done by validating the biomarkers using ELISA (Enzyme-Linked ImmunoSorbent Assay) technology, through a prospective cohort of patients who will also complete the dietary habits and physical exercise questionnaire, to fully validate the classifiers in the new cohort.

The expected results of this study will have direct clinical application. The classifiers generated will be useful to optimize periodic reviews and invasive practices in patients free of disease in early stages and to help in making decisions about the immunotherapy regimen, assessing not only the side effects and the stage of the disease but also the probability of response. In addition, the project will lead to greater knowledge of the relationship between diet and physical exercise with the response to melanoma and the response to immunotherapy. This will help to more accurately advise patients with this disease about their lifestyle habits.

References

1. Tas, F. & Erturk, K. Early and late relapses of cutaneous melanoma patients. *Postgrad Med* **131**, 207–211 (2019).

2. Marciano, N. J., Merlin, T. L., Bessen, T. & Street, J. M. To what extent are current guidelines for cutaneous melanoma follow up based on scientific evidence? *Int J Clin Pract* **68**, 761–70 (2014).

3. Ribero, S. *et al.* Ultrasound-based follow-up does not increase survival in early-stage melanoma patients: A comparative cohort study. *Eur J Cancer* **85**, 59–66 (2017).

4. Picard, M., Scott-Boyer, M.-P., Bodein, A., Périn, O. & Droit, A. Integration strategies of multi-omics data for machine learning analysis. *Comput Struct Biotechnol J* **19**, 3735–3746 (2021).

Common genetic variants associated with urinary phthalate levels in children: a genome-wide study

Mariona Bustamante [a,b,c,*], Laura Balagué [a *], Zsanett Buko[d], Amrit Kaur Sakhi[e], Maribel Casas[a,b,c], Lea Maitre[a,b,c], Sandra Andrusaityte[f], Regina Grazuleviciene[f], Kristine B Gützkow[e], Anne Lise Brantsæter[e], Barbara Heude[g], Claire Philippat[h], Leda Chatzi[i], Marina Vafeiadi [j], Tiffany C Yang[k], John Wright[k], Amy Hough[k], Carlos Ruiz-Arenas[m], Ramil Nurtdinov[n], Geògia Escaramís[l,c], Juan Ramon Gonzalez[a,b,c], Cathrine Thomsen[e], Martine Vrijheid[a,b,c]

[a]Environment and Health over the Lifecourse, ISGlobal, Barcelona, Spain
[b]Universitat Pompeu Fabra (UPF), Barcelona, Spain
[c]CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain
[d]Department of Oncological Science, Huntsman Cancer Institute, Salt Lake City, United States of America
[e]Division of Climate and Environmental Health, Norwegian Institute of Public Health, Oslo, Norway
[f]Department of Environmental Science, Vytautas Magnus University, Kaunas, Lithuania
[g]Université Paris Cité and Université Sorbonne Paris Nord, Inserm, INRAE, Center for Research in Epidemiology and StatisticS (CRESS), F-75004 Paris, France
[h]University Grenoble Alpes, Inserm U-1209, CNRS-UMR-5309, Environmental Epidemiology Applied to Reproduction and Respiratory Health Team, Institute for Advanced Biosciences, 38000, Grenoble, France
[i]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA
[j]Department of Social Medicine, Faculty of Medicine, University of Crete, Heraklion, Greece [k]Bradford Institute for Health Research, Bradford Teaching Hospitals NHS Foundation Trust, Bradford, UK
[l]Departament de Biomedicina, Institut de Neurociències, Universitat de Barcelona (UB), Barcelona, Spain
[m]Computational Biology Program, CIMA University of Navarra, Pamplona, 31008, Spain
[n]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona 08003, Catalonia, Spain

[*]Equal contribution

Abstract

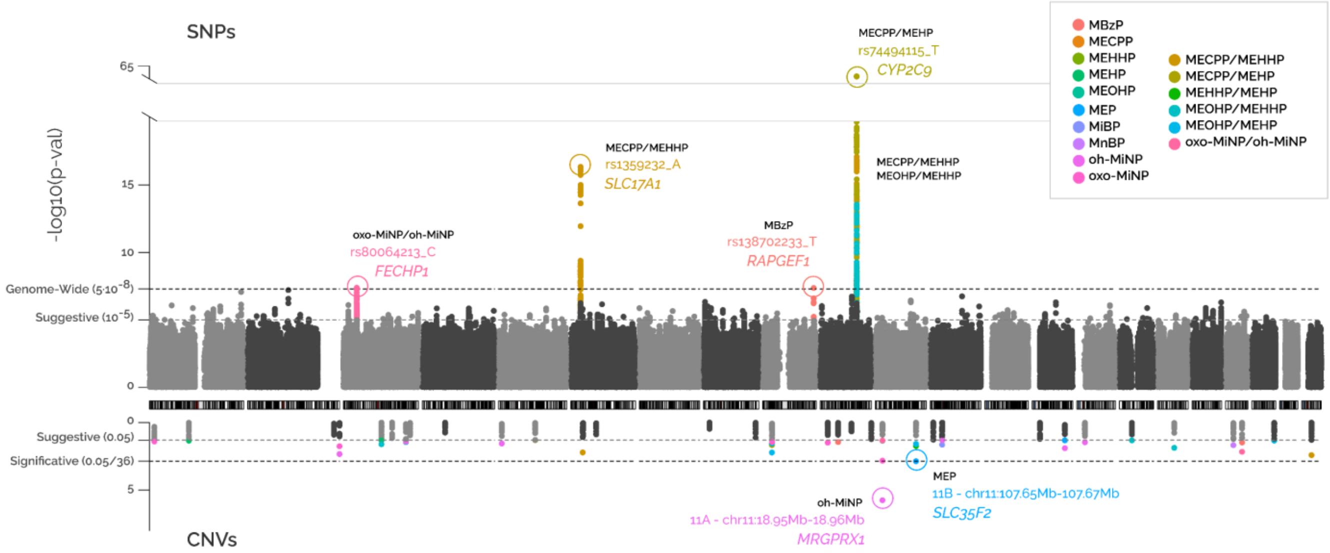
Phthalates, or dieters of phthalic acid, are a ubiquitous type of plasticizer used in a variety of common consumer and industrial products. They act as endocrine disruptors and are associated with increased risk for several diseases. (Praveena *et al.*, 2018; Wang *et al.*, 2019) Once in the body, phthalates are metabolized through partially known mechanisms, involving phase I and phase II enzymes (Domínguez-Romero and Scheringer, 2019). In this study we aimed to identify common single nucleotide polymorphisms (SNPs) and copy number variants (CNVs) associated with the metabolism of phthalate compounds in children through genome-wide association studies (GWAS).

The study used data from 1,044 children with European ancestry from the Human Early Life Exposome (HELIX) cohort. Ten phthalate metabolites were assessed in a two-void urine pool collected at the mean age of 8 years. Six ratios between secondary and primary phthalate

metabolites were calculated. Genome-wide genotyping was done with the Infinium Global Screening Array (GSA) and imputation with the Haplotype Reference Consortium (HRC) panel. PennCNV (Wang *et al.*, 2007) was used to estimate copy number variants (CNVs) and CNVRanger (Da Silva *et al.*, 2020) to identify consensus regions. GWAS of SNPs and CNVs were conducted using PLINK (Purcell *et al.*, 2007) and SNPassoc (González *et al.*, 2007), respectively. Subsequently, functional annotation of suggestive SNPs (p-value <1E-05) was done with the FUMA web-tool (Watanabe *et al.*, 2017).

We identified four genome-wide significant (p-value <5E-08) loci at chromosome (chr) 3 (FECHP1 for oxo-MiNP_oh-MiNP ratio), chr6 (SLC17A1 for MECPP_MEHPP ratio), chr9 (RAPGEF1 for MBzP), and chr10 (CYP2C9 for MECPP_MEHPP ratio). Moreover, 113 additional loci were found at suggestive significance (p-value <1E-05). Two CNVs located at chr11 (MRGPRX1 for oh-MiNP and SLC35F2 for MEP) were also identified. Functional annotation pointed to genes involved in phase I and phase II detoxification, molecular transfer across membranes, and renal excretion.

Through genome-wide screenings we identified known and novel loci implicated in phthalate metabolism in children. Genes annotated to these loci participate in detoxification and renal excretion.



Miami plot of the association between common SNPs (top panel) and CNVs (bottom panel) vs phthalate levels or ratios

Each dot represents the association of a SNP or CNV. The x-axis indicates the position of the SNP or CNV in the genome. The y-axis shows the statistical significance, the –log10(p-value). Colours indicate the phthalate compound or ratio the SNPs or CNVs are associated with. The SNPs/CNVs passing multiple-testing correction (5E-08 for SNPs and 1.39E-03 for CNVs) are annotated to the closest gene.

References

Domínguez-Romero,E. and Scheringer,M. (2019) A review of phthalate pharmacokinetics in human and rat: what factors drive phthalate distribution and partitioning? *Drug Metab. Rev.*, **51**, 314–329.

González,J.R. *et al.* (2007) SNPassoc: an R package to perform whole genome association studies. *Bioinformatics*, **23**, 654–655.

Praveena,S.M. *et al.* (2018) Recent updates on phthalate exposure and human health: a special focus on liver toxicity and stem cell regeneration. *Environ. Sci. Pollut. Res.*, **25**, 11333–11342.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, **81**, 559–575.

Da Silva,V. *et al.* (2020) CNVRanger: association analysis of CNVs with gene expression and quantitative phenotypes. *Bioinformatics*, **36**, 972–973.

Wang,K. *et al.* (2007) PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.*, **17**, 1665–1674.

Wang,Y. *et al.* (2019) A Review of Biomonitoring of Phthalate Exposures. *Toxics*, **7**.

Watanabe,K. *et al.* (2017) Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.*, **8**, 1826.

Structural determinants of α-synuclein binding to an inhibitory peptide studied by molecular dynamics simulations

Oriol Bárcenas1,2, Salvador Ventura1, Ramon Crehuet2

1 Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Bellaterra, Barcelona 08193, Spain

2 CSIC-Institute for Advanced Chemistry of Catalonia (IQAC), E-08034 Barcelona, Spain

Abstract

Parkinson's disease is the world's second most prevalent neurodegenerative disease and is characterized by the loss of dopaminergic neurons in the substantia nigra pars compacta (Pringsheim *et al.*, 2014; Tysnes and Storstein, 2017). α-Synuclein aggregation into amyloid fibers is the main pathological hallmark of Parkinson's disease. Thus, it is the focus of many studies that aim to understand and find treatment for this debilitating disorder (Vázquez-Vélez and Zoghbi, 2021).

α-Synuclein is an Intrinsically Disordered Protein: it does not have a defined structure in solution (Lashuel *et al.*, 2013). Thus, it cannot be studied using classical methods that rely on a single structure, such as X-ray crystallography or Cryogenic Electron Microscopy (CryoEM). On the other hand, methods that measure the protein's structure in solution, such as nuclear magnetic resonance (NMR) or small-angle X-ray scattering (SAXS), are more adequate but only produce ensemble-averaged data with reduced information content. These methods provide only marginal information on α-Synuclein conformation, and alternative strategies are required, such as computational models and molecular simulations.

Given the importance of Parkinson's disease, many efforts have focused on finding therapeutics targeting the formation of aggregates. In one such study, PSMa-3, an antimicrobial peptide synthesized by *S. aureus*, was identified as a highly potent inhibitor of the aggregation of α-Synuclein (Santos *et al.*, 2021). It exerts this activity by binding preferentially to an α-Synuclein region comprising residues 24-58, which is highly relevant for Parkinson's disease, as almost all inherited mutations of familial Parkinson's disease are located in this region. For this reason, the peptide of the 24-60 amino acid region is simulated.

This work focuses on simulating and studying the interaction of α-Synuclein with PSMa-3 in solution and elucidate whether its interaction significantly impacts their structure. Each peptide will also be analyzed separately to compare the obtained results. To this end, the Molecular Dynamics open-source software GROMACS (Abraham *et al.*, 2015) and supercomputing resources are used. Enhanced sampling using "Parallel Tempering, Well-Tempered Ensemble" (PT-WTE) is employed to visit a more expansive conformation space of both peptides and their complex (Deighan *et al.*, 2012). With this method, rare conformations are visited more often than would otherwise be feasible with a classical simulation method.

The simulation results agree with the experimentally observed behavior of the two peptides. Two distinct interaction surfaces between PSMa3 and α-Synuclein are described, with strand secondary structure governing these contacts. Moreover, these contacts can induce secondary structures (strand and helix) in neighboring residues. Finally, the key role of the hydrophobic effect in promoting these contacts is described.

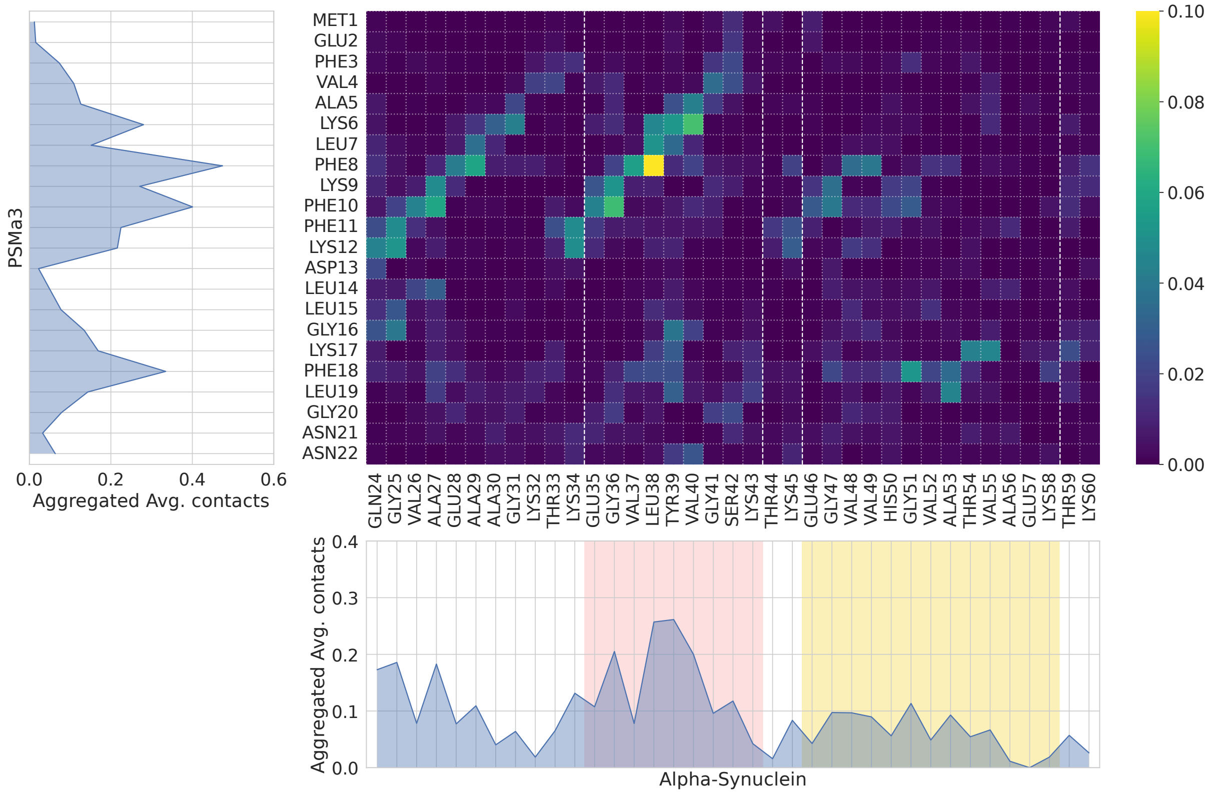


Figure 1. Heatmap of the contacts between residues of PSMa3 (y-axis) and aSyn (x-axis). These contacts have been averaged over the complete trajectory. In each axis, the total amount of contacts has been aggregated for each residue. For aSyn, the P1 and P2 regions, which concentrate most familiar mutations, have been highlighted in red and yellow, respectively.

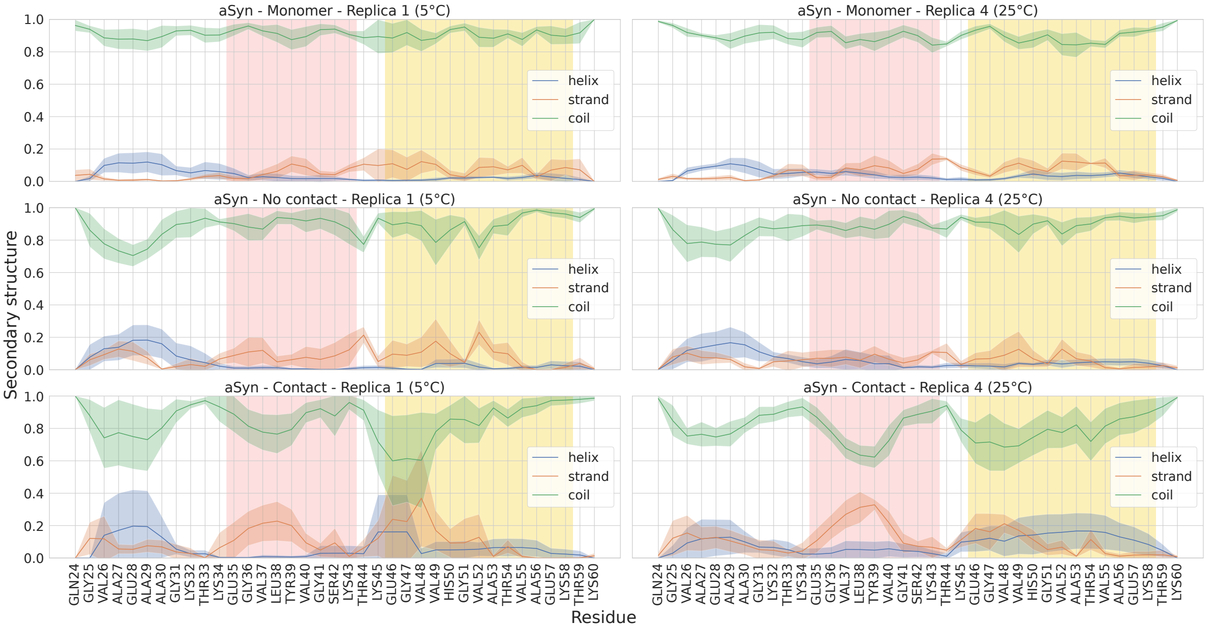


Figure 2. Average secondary structure and error of the mean of the aSyn peptide for each residue. For each replica (columns), three analyses are provided (rows): the system where aSyn is simulated by itself (”Monomer”), the frames where the complex system had no contacts (”No contact”), and the frames where the complex system had contacts (”Contact”).

References

Abraham,M.J. *et al.* (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, **1–2**, 19–25.

Deighan,M. *et al.* (2012) Efficient Simulation of Explicitly Solvated Proteins in the Well-Tempered Ensemble. *J. Chem. Theory Comput.*, **8**, 2189–2192.

Lashuel,H.A. *et al.* (2013) The many faces of α-synuclein: from structure and toxicity to therapeutic target. *Nat Rev Neurosci*, **14**, 38–48.

Pringsheim,T. *et al.* (2014) The prevalence of Parkinson’s disease: A systematic review and meta-analysis. *Movement Disorders*, **29**, 1583–1590.

Santos,J. *et al.* (2021) α-Helical peptidic scaffolds to target α-synuclein toxic species with nanomolar affinity. *Nat Commun*, **12**, 3752.

Tysnes,O.-B. and Storstein,A. (2017) Epidemiology of Parkinson’s disease. *J Neural Transm*, **124**, 901–905.

Vázquez-Vélez,G.E. and Zoghbi,H.Y. (2021) Parkinson’s Disease Genetics and Pathophysiology. *Annu. Rev. Neurosci.*, **44**, 87–108.

Modeling the effects of circadian rhythm on the two alternative pathways for terpenoid precursor biosynthesis in plants

Oriol Basallo[1,2], Abel Lucido[1,2], Ester Vilaprinyo[1,2], Rui Alves[1,2]

[1]Faculty of Medicine, Universitat de Lleida. [2]IRBLleida.

Abstract

Many highly valued chemicals in the pharmaceutical, biotechnological, cosmetic, and biomedical industries belong to the terpenoid family. Biosynthesis of these chemicals relies on polymerization of the terpenoid precursors Isopentenyl di-phosphate (IPP) and/or dimethylallyl diphosphate (DMAPP) monomers, which plants synthesize using two alternative pathways: a cytosolic mevalonic acid (MVA) pathway and a plastidic methyleritritol-4-phosphate (MEP) pathway. We were interested in understanding the effects of circadian rhythm and yearly seasons on the regulation of these pathways.

To study those effects, we created a mathematical model that describes the dynamic behavior of both pathways and adapted it to receive input signals from the daily light cycle. We implemented circadian regulation of the MVA and MEP pathways at every level we found in literature: availability of carbohydrates and organic acids due to photosynthesis, and regulation of gene expression of enzymes, both upstream and downstream of IPP/DMAPP.

Steady state, stability and sensitivity analysis of our basal model (first described in Basallo *et al.* (2023)) show robustness to enzyme mutations and compatibility with biological homeostasis conditions. Adding the signal of a circadian clock to the model shows that the kinetics prevent precursors IPP and DMAPP from depleting at the same rate as other intermediate metabolites. We also test the effect of light cycles and seasons on the regulation of alternative modules of the model by analyzing how the dynamic behavior changes when regulation patterns change.

References

Basallo,O. *et al.* (2023) Changing biosynthesis of terpenoid percursors in rice through synthetic biology. *Front. Plant Sci.*, **14**, 1133299.

Differences *in silico* in drug response between primary and metastatic cancer

Maria Butjosa Espín[1,2] , Jose Antonio Seoane[1]

1. Vall d'Hebron Institute of Oncology (VHIO), Barcelona
2. Universitat Autònoma de Barcelona (UAB), Bellaterra

Abstract

Metastasis causes 90% of cancer-related deaths, urging innovative therapy research due to ineffective treatments and therapy resistance (Ganesh and Massagué, 2021). New treatments, often first tested in metastatic patients, lack success in primary tumors. To tackle this challenge, our aim is to identify drugs with a better response in metastasis, taking advantage of distinct genetic patterns in metastases (Poturnajova, M. *et al.*, 2021; Paul, M. R. *et al*., 2020) and using databases containing drug response of thousands of drug - cell line pairs (Smirnov, P *et al*. 2018).

The first part of the project has involved the use of two drug response database, PRISM and GDSC2, to identify and compare differential drug response between metastatic and primary cell lines by employing a logistic regression model. Moreover, a 'drug set enrichment analysis', an analysis analogous to GSEA (Gene set enrichment analysis) was conducted to identify enriched mechanisms of actions.

The analysis revealed drugs and drug families with differential effects in metastatic versus primary cell lines, particularly at a pan-cancer level. Specifically, the Akt inhibitors group was enriched in metastatic cell lines in both the PRISM (FDR < 0.01) and GDSC2 (FDR < 0.25) databases. However, when stratifying by cancer types, no significant results were found in GDSC2, limited by its lower statistical power. Further exploration in PRISM did unveil the enrichment of EGF receptor inhibitors in colon adenocarcinoma (COAD) metastatic cell lines (FDR < 0.05), and the significance of epigenetic regulation in lung adenocarcinoma (LUAD).

Our approach offers a promising strategy for identifying optimal drug candidates for specific cancer types in metastatic versus primary diseases. This is for now limited by the number of available drug -cell line pairs in the public databases used. To overcome this limitation, the next step of the project will involve using deep learning architectures (Manica, M*. et al,* 2019) for prediction of drug response for unavailable drug – cell line pairs, allowing application of our methodology in cancer subtypes cell line models which could improve patient prognosis in a targeted way.

References

1. Ganesh, K. & Massagué, J. (2021) Targeting metastatic cancer. *Nat. Med.* **27**, 34–44.
2. Poturnajova, M. *et al.* (2021) Molecular features and gene expression signature of metastatic colorectal cancer (Review). *Oncol. Rep.* **45**, 1–18.
3. Paul, M. R. *et al.* (2020) Genomic landscape of metastatic breast cancer identifies preferentially dysregulated pathways and targets. *J. Clin. Invest.* **140**, 4252–4265.
4. Smirnov, P *et al.* (2018). PharmacoDB: an integrative database for mining in vitro anticancer drug screening studies. *Nucleic Acids Research,* **46**(D1), D994–D1002.
5. Manica, M. *et al.*(2019) Toward Explainable Anticancer Compound Sensitivity Prediction via Multimodal Attention-Based Convolutional Encoders. ***Mol. Pharm.*** 4797–4806.

MACHINE LEARNING FOR EARLY PREDICTION OF VIBRIO VULNIFICUS INFECTIONS IN THE US

J.M. Cabrera-Gumbau [1], J. Triñanes-Fernandez [2], J. Martinez-Urtaza [1].

[1]*Universitat Autonoma de Barcelona - Barcelona (Spain)*

[2]*Universidad de Santiago de Compostela - Santiago De Compostela (Spain)*

Abstract

Background

Pathogenic members of the *Vibrio* genus present in tropical marine waters have recently emerged at higher latitudes due to climate change.[1] *Vibrio vulnificus* (Vv) is an opportunistic pathogen that causes vibriosis with 25% of fatality on healthy people in the US, and 56% on Immunosuppressed or people with hepatic diseases. Cases in the US have mainly been confined on the Caribbean area, however reports have shown a constant northward expansion, generating a public health concern.[2]

Ecological conditions on which the bacterium can bloom are complex, being sea water temperature and salinity, the key variables driving abundance in the environment and infections.[3] Nevertheless, efforts to use those conditions as proxy to identify the risk of infections proved to be inefficient. Machine learning offers a new alternative to analyses massive amount of data within complex contexts and has been used to model and forecast other climate-sensitive infectious diseases (*West Nile Virus* or cholera infections).[4,5]

Methods

Here we use machine learning algorithms such as Random Forest, XGBoost[6] and Logistic Regression Model, to predict cases in the coasts of the US with county and daily resolution, aiming to build a generalized model able to predict cases all around the globe. We have combined clinical confirmed cases of Vv from the COVIS dataset and environmental and oceanic satellite data, to train our models using the cross-validation approach.

Results

We compared the best performing models and selected the best one, which is a XGBoost model with a performance with unseen data of 70% accuracy and AUC_ROC of 0.751. However, 22% of the results generated were false positives that cannot be taken into consideration because of the non-reporting situation of Vv, where only 1 out of 143 cases are reported. Meaning that some of the false predicted cases could be cases that actually happened but were not reported, making this model even more robust, reporting what was not reported.

Conclusions

Improvements on the model have already been planned for future works, working mainly on epidemiological data acquisition, spatial resolution and human behaviour implementation.

References

1. Baker-Austin, C. *et al.* Emerging Vibrio risk at high latitudes in response to ocean warming. *Nat. Clim. Change* **3**, 73–77 (2013).
2. Haftel, A. & Sharman, T. Vibrio Vulnificus. in *StatPearls* (StatPearls Publishing, 2023).
3. Amaro, C. & Carmona-Salido, H. Vibrio vulnificus, an Underestimated Zoonotic Pathogen. *Adv. Exp. Med. Biol.* **1404**, 175–194 (2023).
4. Farooq, Z. *et al.* European projections of West Nile virus transmission under climate change scenarios. *One Health* **16**, 100509 (2023).
5. Campbell, A. M., Racault, M.-F., Goult, S. & Laurenson, A. Cholera Risk: A Machine Learning Approach Applied to Essential Climate Variables. *Int. J. Environ. Res. Public. Health* **17**, 9378 (2020).
6. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016). doi:10.1145/2939672.2939785.

Methylation biomarkers for colorectal cancer early detection and survival prognostics impact gene expression and link to cancer-related biological pathways

Pol Canal Noguer[1,3,4,5]; Alejandro Requena Bermejo[1]; Francesco Mattia Mancuso[1]; Juan Carlos Higareda[1]; Marina Manrique López[1]; Marko Chersicola[2]; Pablo Pérez Martínez[2]; Pablo Antonio Camino[1]; Primoz Knap[2]; Vivian Erklavec Zajec[2]; Kristi Kruusmaa[2]

1Universal Diagnostics S.L., Sevilla, Spain

2Universal Diagnostics d.o.o., Ljubljana, Slovenia

3B2SLab, Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, Barcelona, Spain

4Networking Biomedical Research Centre in the subject area of Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Madrid, Spain

5Institut de Recerca Pediàtrica Hospital Sant Joan de Déu, Esplugues de Llobregat, Barcelona, Spain

Abstract

Background

DNA methylation has been previously shown to have diagnostic and predictive potential for colorectal cancer (CRC). Aim of this study was to evaluate putative methylation markers in the context of early cancer development and diagnostics as well as further investigate the biological significance of these regions.

Methods

Biomarker discovery was done by whole genome bisulfite sequencing (WGBS) of 88 CRC, 48 advanced adenoma (AA) and corresponding adjacent normal tissue (NAT) samples. Short-list of significantly hypermethylated regions (DMRs) was correlated to transcriptomics data from 512 CRC patients in The Cancer Genome Atlas (TCGA) cohort. Pathway enrichment for biological pathway analysis of the DMRs was done by using Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database. Survival analysis was performed using Kaplan–Meier method on sub-groups of patients divided by the methylation status of individual markers. Finally, individual marker significance of selected regions was evaluated by analyzing 26 plasma samples from early stage (stage I-IIA) CRC samples and 42 colonoscopy verified controls (CNT) with targeted methylation sequencing assay.

Results

4167 putative marker regions were identified from biomarker discovery with WGBS. Differential signal could be observed both between AA and NAT and CRC and NAT, while several of these regions were differentially methylated also between AA and CRC samples, indicating biological signal change with adenoma progression to cancer. 84 hypermethylated DMRs from several verification studies were further evaluated against transcriptome data from TCGA, where overlap for 69 genes was found. 19 of these genes showed a significant down- regulation ($p < 0.05$), indicating a link between hypermethylation and gene expression. 2 genes showed significant up-regulation ($p < 0.05$), which could indicate other epigenic processes to be in place. KEGG pathway analysis revealed that the top pathways involved were axonal guidance, ephrin receptor signaling, epithelial-mesenchymal transition and FGF signaling, which all play significant role in the context of cancer development and progression. Kaplan-Meier analysis showed significant correlation to patients 5- year survival prediction linked to 3 genes: FGF14 ($p=0.025$, HR = 1.75) , DPY19L2P1 ($p=0.012$, HR = 1.86), PTPRO ($p=0.046$, HR = 1.63). Targeted sequencing analysis on plasma samples of patients with early stage (I-IIA) colorectal cancer and age and gender matching colonoscopy-verified controls, showed high individual marker accuracy with AUC= 0.78 for FGF14, AUC= 0.81 for DPY19L2P1 and AUC= 0.73 for PTPRO.

Conclusions

Methylation markers have distinct signals in early development of CRC, with high individual accuracy for separating early-stage cancers from matching controls. These regions have impact on gene expression and can be linked to relevant biological pathways. Extending early detection potential of the markers to further prognostics and stratification, could lead to better outcomes and improved survival of the patients.

Gene expression prediction under novel conditions using ATAC-seq-informed regulons

Roger Casals Franch[1,2,3]

Supervisors: Adrián García López de Lomana[1,2], Jordi Villà Freixa[1,2]

1. Institut de Recerca i Innovació en Ciències de la Vida i de la Salut a la Catalunya Central (IRIS-CC)
2. Research Group on Bioinformatics and Bioimaging (BI2); Facultat de Ciències, Tecnologia i Enginyeries; Universitat de Vic −- Universitat Central de Catalunya (UVic-UCC), Vic, Barcelona
3. Ph.D. Bioinformatics Programme UVic-UCC

Abstract

The prediction of cell state changes due to genetic and environmental perturbations is a paramount long-sought research question in biology [1]. Understanding how cells transition from one state to another is crucial for unravelling the mechanisms underlying various biological processes, including development, disease progression, and response to therapy.

Recently, the application of deep learning models such as variational autoencoders to single-cell omics profiles has enabled the accurate prediction of cell state transitions in response to a broad variety of perturbations, not only within a given biological scenario but across dissimilar systems like different studies or even different species [2]. This tremendous capability has facilitated the recent emergence of a new field into the deep learning bioinformatics [3,4]. Yet, a mechanistic understanding of the molecular regulators behind these transitions is currently lacking.

Having a single-cell perturbation dataset [5] and different state-of-the-art single-cell omics integration methods [6] our goal is building Gene Regulatory Networks (GRNs), which encode cell identity by the regulation of target gene expression through interactions between transcription factors (TFs) and sets of cis-regulatory elements (CREs), those TF are detected based on the TF-binding sites (TFBS), frequently cell-specific. Such candidate enhancers can be better predicted using open chromatin data. Our approach uses SCENIC+ to integrate single-cell ATAC-sequencing data and single-cell RNA-sequencing data to infer a gene regulatory network to find regulons to enhance the accuracy and resolution of cell state predictions [7].

Our primary hypothesis states that predicting perturbations on those TF will offer a more mechanistic understanding of cell state transitions. To achieve this, we

will focus on transcription factor to gene target interactions, exploring the potential of regulons in capturing cellular plasticity.

Appropriate metrics, including accuracy, precision, and recall, as well as biologically known truths to validate the outputs, will be employed to assess the results.

The proposed research has the potential to advance our understanding of cell state transitions, providing insights into molecular mechanisms crucial for developmental biology, cancer research, and regenerative medicine.

Through collaborations with clinical researchers and adaptation to project-specific requirements, the developed computational framework will contribute to the identification of novel therapeutic targets and further the field of predictive biology.

Ethical considerations will be paramount, adhering to clinical standards when working with relevant information.

All the data that we will use will be public data repositories (in particular, at the Repositori de Dades de Recerca, CSUC) and the code and the results uploaded to my Github account.

References

[1] Moris, N., Pina, C., & Arias, A. M. (2016). Transition states and cell fate decisions in epigenetic landscapes. *Nature Reviews Genetics*, *17*(11), 693-703.

[2] Lotfollahi, M., Wolf, F. A., & Theis, F. J. (2019). scGen predicts single-cell perturbation responses. *Nature methods*, *16*(8), 715-721.

[3] Ji, Y., Lotfollahi, M., Wolf, F. A., & Theis, F. J. (2021). Machine learning for perturbational single-cell omics. *Cell Systems*, *12*(6), 522-537.

[4] Stein-O'Brien, G. L., Ainslie, M. C., & Fertig, E. J. (2021). Forecasting cellular states: from descriptive to predictive biology via single-cell multiomics. *Current opinion in systems biology*, *26*, 24-32.

[5] Daniel Burkhardt, Andrew Benz, Richard Lieberman, Scott Gigante, Ashley Chow, Ryan Holbrook, Robrecht Cannoodt, Malte Luecken. (2023). Open Problems – Single-Cell Perturbations. Kaggle. https://kaggle.com/competitions/open-problems-single-cell-perturbations

[6] Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., ... & Satija, R. (2019). Comprehensive integration of single-cell data. *Cell*, *177*(7), 1888-1902.

[7] Bravo González-Blas, C., De Winter, S., Hulselmans, G., Hecker, N., Matetovici, I., Christiaens, V., ... & Aerts, S. (2023). SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nature Methods*, *20*(9), 1355-1367.

Diet adaptations in anatomically modern humans

Carla Casanova Suárez1, Oscar Lao Grueso1

1 Institute of Evolutionary Biology (UPF-CSIC)

Abstract

Anatomically modern humans (AMH) initially evolved in a particular environment within the African continent. However, humans have managed to conquer a wide range of diverse environments after the Out of Africa event in an extremely short period of time from an evolutionary point of view. This has been possible by cultural and biological adaptations (incorporating genetic variants conferring a fitness advantage from archaic populations and/or via mutation). Nowadays, the mismatch between our original environment and the current living conditions is being drastically increased by recent events, such as the industrial revolution and the digital era. This disparity has been suggested by the field of Evolutionary Medicine to be a key factor in the development of common complex diseases. Until recently there was no information of genetic variants associated with these complex phenotypes. Nevertheless, Genome Wide Association Studies (GWAS), even if they explain a very limited amount of the variance of these phenotypes (Manolio et al., 2009), provide a first hint of the role of genetics in common diseases. In addition, advances in the field of ancient DNA have provided a large number of publicly available datasets covering a huge range of evolutionary history in Europe after the Out of Africa event. Taking all these elements together, this opens the possibility to study the recent evolution of complex phenotypes (Sella and Barton, 2019; Marciniak and Perry, 2017). The "Diet" phenotype is a particularly interesting trait for understanding human evolution, since it is usually geographically restricted and conditioned by the environment. This implies that it should be easier to identify selective pressures among human populations by comparing them. Consequently, the current study will implement a bioinformatic pipeline for tracking the selection fingerprint of a series of genetic markers with diet susceptibility across European populations from different regions and periods by integrating data from ancient genomes. Nevertheless, this is conditioned to the development of new methods for enhancing the power of GWAS data by considering machine learning techniques, notably deep learning (DL), and genetic algorithms for integrating genetic data from diet-related phenotypes (Li *et al.*, 2019). These analyses can provide new insights into the repercussions of dietary changes on human genetic adaptation and health.

References

Li,Y. *et al.* (2019) Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. *Methods*, **166**, 4–21.

Manolio,T.A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.

Marciniak,S. and Perry,G.H. (2017) Harnessing ancient genomes to study the history of human adaptation. *Nat. Rev. Genet.*, **18**, 659–674.

Sella,G. and Barton,N.H. (2019) Thinking About the Evolution of Complex Traits in the Era of Genome-Wide Association Studies. *Annu. Rev. Genomics Hum. Genet.*, **20**, 461–493.

Evaluating allele frequency trajectory and selection coefficient estimates from genealogies including ancient DNA

Aina Colomer-Vilaplana[1,2], Sònia Casillas[1,2], Antonio Barbadilla[1,2], Leo Speidel[3,4]

1. Institute of Biotechnology and Biomedicine, Universitat Autònoma de Barcelona, Parc de Recerca, Mòdul B, 08193 Cerdanyola del Vallès, Catalonia, Spain
2. Department of Genetics and Microbiology, Facultat de Biociències, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Catalonia, Spain
3. Genetics Institute, University College London, 99-105 Gower St, London WC1E 6AA, United Kingdom
4. Francis Crick Institute, 1 Midland Road, London NW1 1AT, United Kingdom

Abstract

Humans have successfully adapted to different environments during their migration across the continents. Yet, a question still unsolved is the extent to which selection has played a role in shaping our genomes throughout these migrations. Recent advances in genomic methodologies, including the use of ancient DNA, have provided new opportunities to study our genetic past. The availability of large cohorts of ancient DNA samples from single populations has enabled the inference of allele frequency trajectories and the associated selection coefficients.

Recently, new methods for inferring genealogies -such as Relate [1] and tsinfer[2]- as well as new methods that feed from these genealogies -like Clues[3]- have made it possible to extrapolate allele frequency trajectories from sequencing data of modern-day samples. Here we evaluate the effectiveness of Relate and Clues benchmarking these methods against known and inferred genealogies from simulated data using SLiM[4]. Moreover, we develop our own strategy to infer a selection coefficient from a pre-estimated genealogy incorporating ancient DNA. We test this method under different selective regimes ranging from neutral to strong selection, showing that aDNA substantially improves selection estimates. With our proposed method we aim for a better understanding of the genomic marks left by selection over the past tens of thousands of years.

Applying this approach at the genome-wide level could provide new light on understanding the role of selection in shaping our genomes during human past migrations.

References

1. Speidel,L. *et al.* (2019) A method for genome-wide genealogy estimation for thousands of samples. *Nature genetics*, **51**(9), 1321-1329.
2. Kelleher,J. *et al.* (2019) Inferring whole-genome histories in large population datasets. *Nature genetics*, **51**(9), 1330-1338.
3. Stern,A. *et al.* (2019) An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. *PLoS Genetics*, **15**(9), e1008384.
4. Haller,B.C. *et al.* (2019) Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology Resources*, **19**(2), 552–566.

2nd February 2024 Workshop

LONGITUDINAL SEGMENTATION OF MULTIPLE SCLEROSIS LESIONS

Marcos Diaz Hurtado , Ferran Prados Carrasco , Jordi Casas Roma - UOC PhD Bioinformatics

INTRODUCTION

Multiple sclerosis (MS) is a leading cause of neurological disability in young individuals, often diagnosed and monitored using magnetic resonance imaging (MRI). Traditional lesion quantification in MRI is time-consuming and prone to errors. Various methods, including statistical classification, machine learning, and deep learning like U-NET, have been used for automatic lesion measurement (Diaz-Hurtado et al. 2022). The emerging transformer architecture, known for its impact in computer vision, is underexplored but promising in MS lesion segmentation.

OBJECTIVES

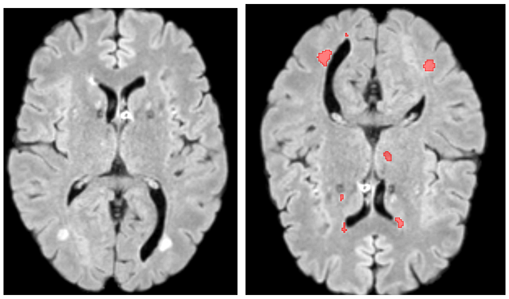
We propose a U-Net-based approach for segmenting MS lesions in MRI images, both cross-sectional and longitudinal, using baseline and follow-up images along with the baseline mask to predict longitudinal lesion mask changes.

MATERIAL AND METHODS

The ISBI dataset was employed alongside the MONAI API, implemented in Python and executed within a Docker MONAI image hosted on an AWS VPS with a Linux system boasting a 16 GB GPU.

RESULTS

The model is currently in the development phase, with preliminary results indicating segmentation accuracy of at least 50%.



DISCUSSION

The development of a U-Net-based approach for MS lesion segmentation presents a promising avenue for improving diagnosis and monitoring of MS patients. The integration of longitudinal data, once fully implemented, will enhance our ability to track lesion changes over time, aiding in treatment planning and evaluation. Further refinements and optimizations are ongoing to enhance the model's performance and generalizability.

CONCLUSION

The ongoing work on this U-Net-based MS lesion segmentation model, with a focus on longitudinal data integration, holds significant potential to improve clinical outcomes for MS patients by providing more accurate and timely information for medical decision-making.

BIBLIOGRAPHIC REFERENCES

Diaz-Hurtado, Marcos, Eloy Martínez-Heras, Elisabeth Solana, Jordi Casas-Roma, Sara Llufriu, Baris Kanber, and Ferran Prados. 2022. “Recent Advances in the Longitudinal Segmentation of Multiple Sclerosis Lesions on Magnetic Resonance Imaging: A Review.” *Neuroradiology* 64 (11): 2103–17.

CONSERVATION AND EVOLUTION OF HUMAN SEGMENTAL DUPLICATIONS IN MAMMAL GENOMES

Maria Diaz-Ros[1,2], Mario Cáceres[1,2,3]

1 Research Program on Biomedical Informatics (GRIB), Hospital del Mar Research Institute (IMIM), Barcelona, Spain

2 Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain

3 ICREA, Barcelona, Spain

Abstract

Segmental duplications (SDs) are defined as highly identical duplicated DNA fragments longer than 1 kb which promote non-allelic homologous recombination (NAHR), contributing to recurring rearrangements such as inversions, deletions and duplications[1-5]. In humans, SDs show considerable variation and they have long been recognized as a potential source of phenotypic diversity and rapid evolution of new genes, but also as a basis for disease[1,2,6,7]. However, the repetitive nature of the SD sequences makes them difficult to sequence and assemble reliably, and therefore, they are one of the least known regions of the genome[8]. Recently, with the development of newer sequencing techniques based on long-reads and the efforts of ongoing projects such as the Vertebrate Genomes Project[9] or the Zoonomia Project[10], whose goal is building complete high-quality reference genomes for a wide variety of species, for the first time it is possible to analyze SD conservation and evolution throughout the mammalian lineage[10]. For that, we have created two human sets of SDs to compare with each other, one including 480 regions with inverted orientation between repeats and another with 684 directly-oriented repeats. Then, we have assessed SD conservation for each of these regions in 41 species across the mammalian phylogeny with a wide range of divergence times. These has allowed us to estimate the origin of each pair of SDs and the rate of gains and losses suffered throughout the mammalian lineage. Results show an increased conservation of SDs in chromosome X compared to autosomes. They also show an increase in conservation of inverted SDs compared to direct SDs, especially in chromosome X, which could help to identify the most conserved regions that could have functional implications.

References

1. Marques-Bonet,T. *et al.* (2009) The origins and impact of primate segmental duplications. *Trends in Genetics*, **25**, 443–454.

2. Vollger,M.R. *et al.* (2022) Segmental duplications and their variation in a complete human genome. *Science*, **376**, 6965.

3. Sharp,A.J. *et al.* (2006) Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nature Genetics,* **38**, 1038–1042.

4. Puig,M. *et al.* (2015) Human inversions and their functional consequences. *Briefings in Functional Genomics,* **14**, 369–379.

5. Giner-Delgado,C. *et al.* (2019) Evolutionary and functional impact of common polymorphic inversions in the human genome. *Nature Communications*, **10**, 1–14.

6. Sudmant,P.H. *et al.* (2015) Global diversity, population stratification, and selection of human copy-number variation. *Science*, **349**, 6253.

7. Dennis,M.Y. *et al.* (2017) The evolution and population diversity of human-specific segmental duplications. *Nature Ecology & Evolution,* **1**, 1–10.

8. Chaisson,M.J.P. *et al.* (2015) Resolving the complexity of the human genome using single-molecule sequencing, *Nature*, **517**, 608–611.

9. Rhie,A. *et al.* (2021) Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746.

10. Zoonomia Consortium. (2020) A comparative genomics multitool for scientific discovery and conservation, *Nature* **587**, 240–245.

protAGOnist: an innovative NLS/NES prediction tool

Camila Engler1, Belen Moro1, Andrea Martin1, Antonela Lavatelli1, Luciano Abriata2, Nicolás Bologna1

1Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB, Bellaterra, Barcelona 08193, Spain 2 École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Abstract

Introduction RNA silencing is a mechanism that regulates gene expression by using small RNA molecules. These small RNAs play a crucial role in various biological functions, such as regulating the stress response, controlling transposon activity, and responding to viruses. In order to perform their function, small RNAs are loaded into ARGONAUTE (AGO) proteins that recognize and regulate their target genes. Depending on the organism and the small RNA pathway, AGOs have nuclear and/or cytoplasmic localization (Bologna et al., 2018). However, determining the subcellular localization of ARGONAUTE proteins is not always reliable due to the existing prediction tools, which limits our understanding of their function and movement within cells.

Design and architecture We have developed a computational tool called protAGOnist, which combines standard sequence-based predictions with the biophysical properties of amino acids, their evolutionary conservation, and molecular modelling. This tool considers the three-dimensional structure, exposure, flexibility, and region of each subcellular localization signal, and improves the scoring of true nuclear localization signals and nuclear export signals. By doing so, protAGOnist reduces the number of false-positive signals that are common in standard prediction programs.

The tool have been developed in a modular way (Fig1), mainly in Python, including subprocess to T_coffee (Notredame et al., 2000), Iupred3 (Erdős et al., 2021), NACESS (Hubbard, 1993), NESmapper (Kosugi et al. 2014), NLStradammus (Nguyen Ba et al., 2009) and AlphaFold2 (Jumper et al., 2021). The only mandatory input is a FASTA (or multifasta) file but the user may upload a PDB file with the structure and/or a csv file with signals for a more personalized analysis. An interface prototype has been developed using streamlite in order to make the tool more user-friendly.

Output The output of protAGOnist includes a CSV file with the analyzed data and the label true or false for each putative signal, a graph to visualize the position and signal before and after filtering, and a PDB file with the 3D structure of the protein with the exposed signals highlighted (Fig2).

Results Our tool has been applied to more than 52 AGO proteins in various eukaryotic organisms, such as Arabidopsis thaliana, Drosophila melanogaster, Caenorhabditis elegans, Mus musculus, and humans.

We were able to validate several signals obtained by protAGOnist in several AGO proteins from different eukaryotic organisms. Using protAGOnist, we were able to significantly improve the NLS/NES prediction by reducing the number of false positives by 78%.

Conclusion The results demonstrate that the developed computational method can reduce the number of putative signals, making it easier to conduct functional analysis on the studied proteins.

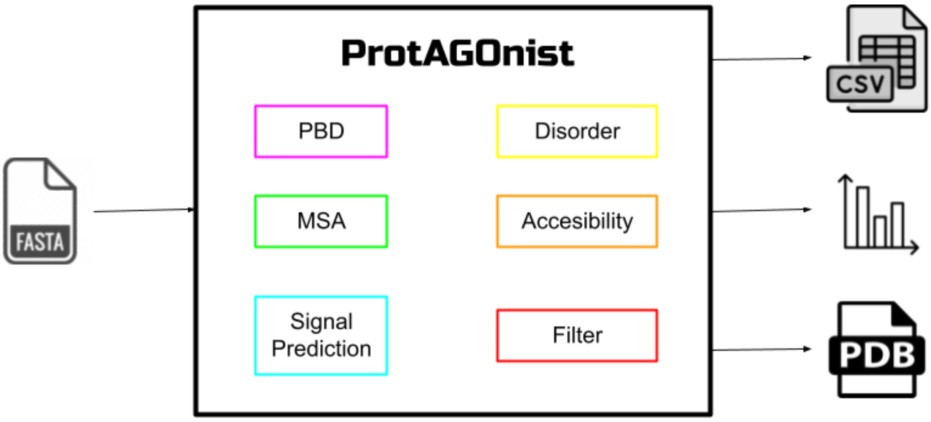


fig1: ProtAGOnist architecture

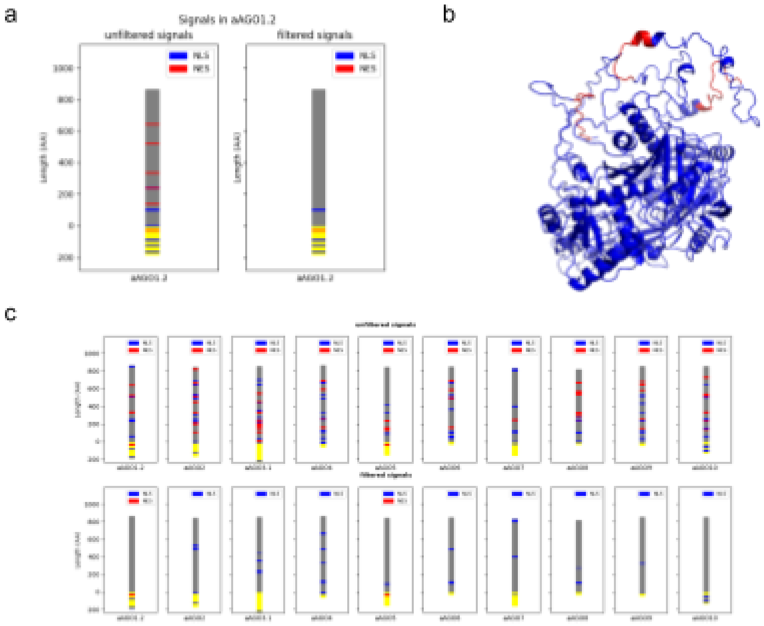


fig2: ProtAGOnist output

References

1. Erdős,G. et al. (2021) IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. Nucleic Acids Research, 49, W297–W303.

2. Hubbard,S.J. (1993) NACCESS-Computer program.

3. Jumper,J. et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature, 596, 583–589.

4. Kosugi,S. et al. (2014) NESmapper: Accurate Prediction of Leucine-Rich Nuclear Export Signals Using Activity-Based Profiles. PLoS Comput Biol, 10, e1003841.

5. Nguyen Ba,A.N. et al. (2009) NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction. BMC Bioinformatics, 10, 202.

6. Bologna,N.G. et al. (2018) Nucleo-cytosolic Shuttling of ARGONAUTE1 Prompts a Revised Model of the Plant MicroRNA Pathway. Mol Cell, 69, 709-719.e5.

7. Notredame,C. et al. (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment 1 1Edited by J. Thornton. Journal of Molecular Biology, 302, 205–217.

Pan-cancer vulnerability prompted by TGFβ-hypoxia-mediated suppression of alternative end-joining

Roderic Espín[1,2], Ferran Medina-Jover[1,2,3], Javier Sigüenza[1,2], Sònia Farran-Matas[1,2], Francesca Mateo[1,2], Agnes Figueras[1,2], Alexandra Baiges[1,2], Rosario T Sanz[4], Guillermo Pablo Vicent[4], Lídia Franco[1,2], Arzoo Shabbir[1,2], Adrián Martínez[1,2], Miguel Angel Pardo[1,2], María Martínez-Iniesta[1,2], Xieng Chen Wang[1,2], Elisabet Cuyàs[1,5], Javier Menéndez[1,5], Marta Lopez-Cerda[2], Purificacion Muñoz[2], Ivonne Richaud[6,7], Angel Raya[6-8], Rehna Krishnan[9], Razqallah Hakem[9,10], Isabel Fabregat[2,11], Oriol Casanovas[1,2], Jordi Bruna[2,12], Inés Guix[13], Josep Maria Piulats[2], Alberto Villanueva[1,2], Mary Helen Barcellos-Hoff[13], Francesc Viñals[1,2]*, Álvaro Aytes[1,2]* & Miquel Angel Pujana[1,2,14]*

1 ProCURE, Catalan Institute of Oncology, L'Hospitalet del Llobregat, Barcelona, Spain
2 Oncobell, Bellvitge Institute for Biomedical Research (IDIBELL), L'Hospitalet del Llobregat, Barcelona, Spain
3 Department of Physiological Sciences, University of Barcelona, L'Hospitalet del Llobregat, Barcelona, Spain
4 Molecular Biology Institute of Barcelona, Spanish National Research Council (IBMB-CSIC), Barcelona, Spain
5 Girona Biomedical Research Institute (IDIBGI), Girona, Spain
6 Regenerative Medicine Program and Program for Clinical Translation of Regenerative Medicine in Catalonia—P-CMR[C], Bellvitge Institute for Biomedical Research (IDIBELL), L'Hospitalet del Llobregat, Barcelona, Spain
7 Biomedical Research Network Centre in Bioengineering, Nanomaterials, and Nanomedicine (CIBER-BBN), Instituto de Salud Carlos III, Madrid, Spain
8 Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain
9 Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada
10 Department of Medical Biophysics, University of Toronto, Ontario, Canada
11 Biomedical Research Networking Centre in Hepatic and Digestive Diseases (CIBEREHD), Instituto de Salud Carlos III, Madrid, Spain
12 Neuro-Oncology Unit, University Hospital of Bellvitge, Catalan Institute of Oncology, Bellvitge Institute for Biomedical Research (IDIBELL), L'Hospitalet del Llobregat, Barcelona, Spain
13 Department of Radiation Oncology and Helen Diller Family Comprehensive Cancer Centre, University of California, San Francisco, San Francisco, CA, United States
14 Biomedical Research Network Centre in Respiratory Diseases (CIBERES), Instituto de Salud Carlos III, Madrid, Spain

Abstract

The alternative end-joining (alt-EJ) DNA repair pathway is an error-prone mechanism activated in homologous recombination-deficient (HRD) cancer cells(Gelot *et al.*, 2023; Brambati *et al.*, 2023). Targeting alt-EJ by inhibiting poly (ADP-ribose) polymerase (PARP) or DNA polymerase theta (POLθ) has shown clinical benefit in HRD cancers(Lord and Ashworth, 2013; Zhou *et al.*, 2021). However, it is not well understood how alt-EJ is regulated, which limits progress in cancer therapies based on the balance between HR and alt-EJ. Here we show that alt-EJ is suppressed by signalling mediated by TGFβ and hypoxia and that this regulation can be therapeutically exploited. Functional gene expression signatures of the TGFβ/hypoxia and alt-EJ pathways are found to be anticorrelated in stem cell-like states, including normal and

cancer settings. Machine-learning modelling identifies a cancer cell state that suppresses alt-EJ through TGFβ-hypoxia while, conversely, alt-EJ is promoted by MYC. Combined inhibition of the hypoxia-inducible factor 1α (HIF1a) using the drug *PX-478*, and of PARP or POLθ, shows synergistic activity in reducing the clonogenic capacity of cancer cells *in vitro*. Combined inhibition of HIF1a and PARP reduces ovarian tumour growth *in vivo.* The findings reveal opportunities for further tackling cancer cells by enforcing the use of alt-EJ.

Figure 1

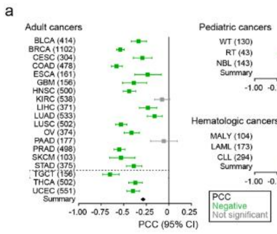


Figure 2

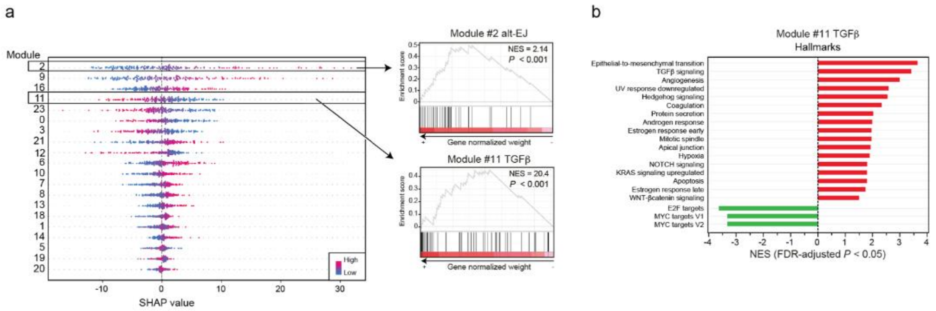


Figure 3

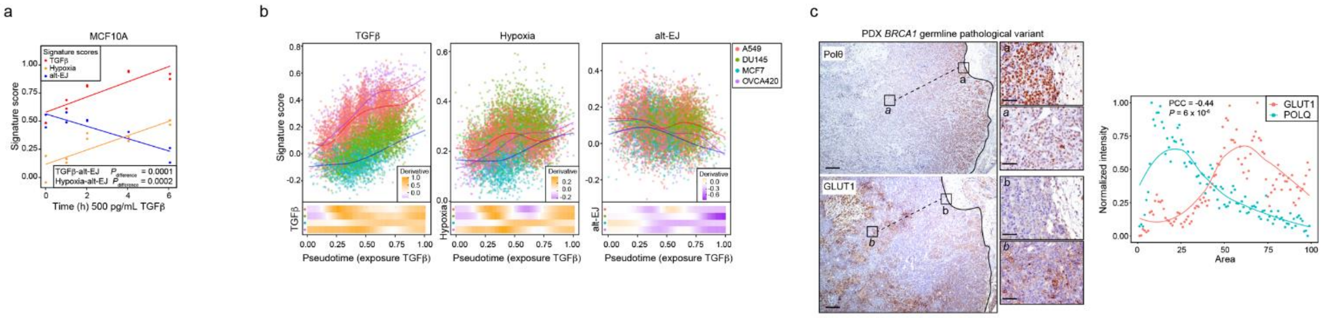


Figure 4

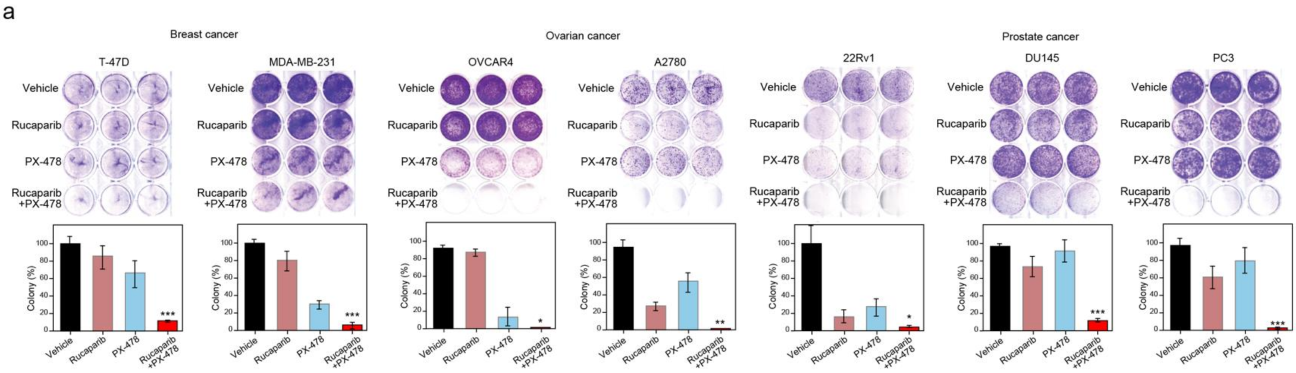
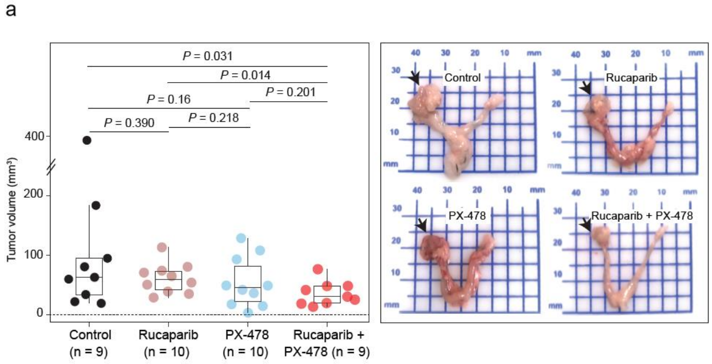


Figure 5



References

Brambati,A. *et al.* (2023) RHINO directs MMEJ to repair DNA breaks in mitosis. *Science*, **381**, 653–660.

Gelot,C. *et al.* (2023) Polθ is phosphorylated by PLK1 to repair double-strand breaks in mitosis. *Nature*, **621**, 415–422.

Lord,C.J. and Ashworth,A. (2013) Mechanisms of resistance to therapies targeting BRCA-mutant cancers. *Nat. Med.*, **19**, 1381–1388.

Zhou,J. *et al.* (2021) A first-in-class polymerase theta inhibitor selectively targets homologous-recombination-deficient tumors. *Nat. Cancer*, **2**, 598–610.

Epigenetic relationships to improve Synthetic Lethality prediction model for cancer treatment

Maria Farina-Morillas [1,2], Jose A. Seoane [1]

(1) Cancer Computational Biology Group, Vall d'Hebron Institute of Oncology (VHIO), Barcelona

(2) Universitat Autònoma de Barcelona (UAB), Barcelona

Abstract

Synthetic Lethality (SL) is a specific interaction between genes where the perturbation of either gene individually does not alter cell viability, although the perturbation of both genes simultaneously leads to loss of viability [1]. Several computational methods have been developed to predict synthetic lethality between genes integrating multi-omics data. Common analysis such as Differential Expression Analysis, pathway analysis or co-expression analysis can be used as input features to develop prediction algorithms [2], yet current methods do not take epigenetic alterations into account despite their role in transcriptional reprograming that could induce new SL relationships. Therefore, due to their role controlling chromatin accessibility and transcription, we have defined a set of Chromatin Regulatory Genes (CRGs) for which we aim to predict their candidate SL partners.

To perform this prediction we are training a Random Forest model with epigenetic SL as target using gene expression data and incorporating epigenetic relationships. We are first mimicking the DiscoverSL [2] approach while still focusing on our set of CRGs, by combining 4 input features in a Random Forest model, so we can benchmark the performance of our model compared to an already established approach. These input features are a Differential Expression Analysis to evaluate changes due to mutations in CRGs, a Mutually Exclusive analysis of mutations, amplifications and deletions between CRGs and protein-coding genes, a co-expression analysis between CRGs and protein-coding genes and a CRGs pathways analysis. In the next steps, we will develop new epigenetic-related features by including methylation events with the mutations and Copy Number Alterations, as well as ATAC-seq data to improve our model.

References

(1) O'Neil, N. J et al. (2017). Synthetic lethality and cancer. Nature Reviews Genetics 2017 18:10, 18(10), 613–623.

(2) Das, Shaoli et al. (2019) DiscoverSL: an R package for multi-omic data driven prediction of synthetic lethality in cancers. Bioinformatics 2019 35,4: 701-702.

In-silico simulation and efficacy evaluation of anti-PD1 treatment on 4 triple-negative breast cancer molecular subtypes

Juan Manuel García-Illarramendi[1,2], Pedro Matos-Filipe[1,3], Guillem Jorba[1,3], Chiara Sopegno[1,3] Xavier Daura[2], Judith Farrés[1]

[1]Anaxomics Biotech SL, Barcelona, Spain

[2]Institute of Biotechnology and Biomedicine, Universitat Autònoma de Barcelona, Barcelona, Spain

[3]Structural Bioinformatics Group, Research Programme on Biomedical Informatics, Department of Experimental and Health Science, Universitat Pompeu Fabra, Barcelona, Spain

Abstract

Triple-negative breast cancer (TNBC) accounts for up to 20% of breast cancer diagnoses and remains a major therapeutic challenge (Mandapati and Lukong). Currently, the combination of immunotherapy and chemotherapy is presented as one of the main regimes to treat TNBC patients (Cortes, et al.). Nevertheless, the association between immunotherapy efficacy and TNBC subtypes is still to be elucidated.

124 TNBC patients from the GEO series *GSE167213* (Hartung, et al.) were assigned to the 4 Lehmann's TNBC subtypes (Chen, et al., 2012; Lehmann, et al.; Lehmann, et al., 2016). Differential expression between the patients of each subtype and the remaining subtypes was done and the identified differentially expressed (DE) proteins were used to build *in-silico* initial state Therapeutic Performance Mapping System (TPMS) models (Gutierrez-Casares, et al.; Jorba, et al.) for each subtype. Anti-PD1 treatment was simulated on these models and *in-silico* efficacy of the treatment was determined for each subtype (see study steps in Figure 1).

Target-based simulation of anti-PD1 treatment on a knowledge-based TNBC protein set was done for each of the 4 subtypes based on their initial state TPMS models. Although no significant difference on the efficacy of anti-PD1 treatment was observed (ANOVA, p-value > 0.05) (see Figure 2), M subtype had the lowest efficacy among the 4 subtypes. Independent evaluation of 4 different signatures of the TNBC subtypes identified in an independent TNBC cohort (Akhouayri, et al.) also identified the M subtype as the lowest *in-silico* efficacy subtype (see Figure 3).

M subtype showed the lowest *in-silico* anti-PD1 efficacy in the TPMS models, which goes in line with previous observations made (Lehmann, et al.). A patient-level analysis using TPMS models is still to be made to further confirm these results. Mechanistic differences associated with efficacy differences will also be analyzed.

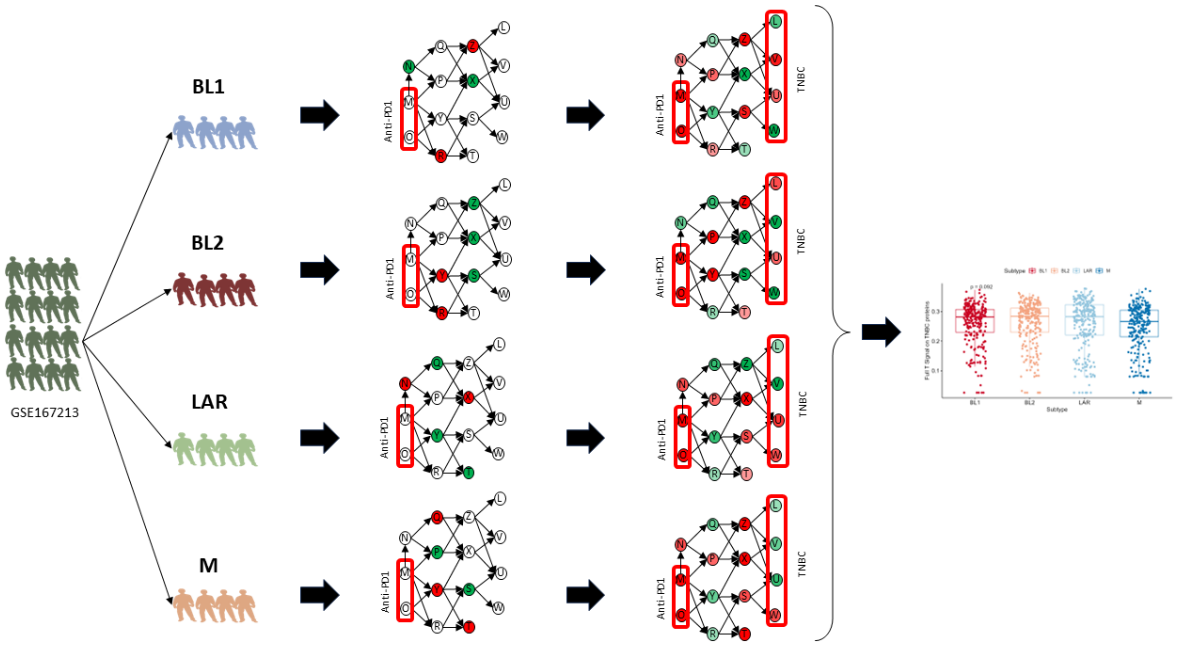


Figure 1: Diagram of the steps followed in the study.

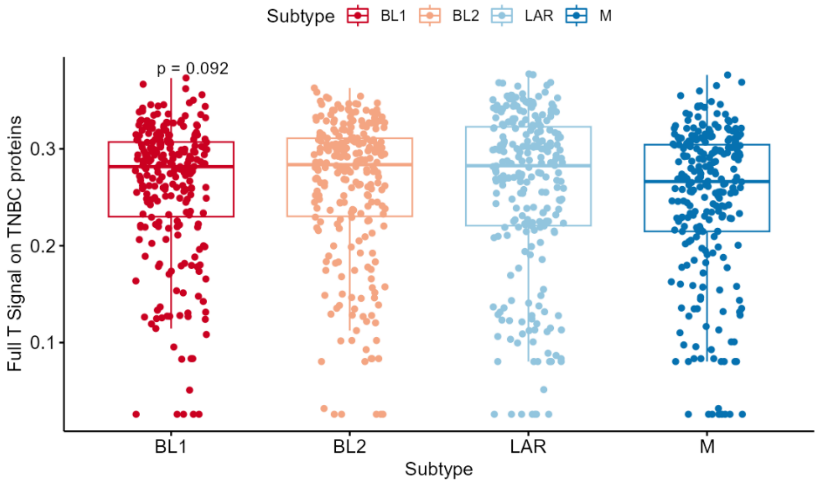


Figure 2: *In-silico* efficacy of anti-PD1 treatment of the 4 TNBC subtypes derived from *GSE167213* GEO serie (Hartung, et al.).

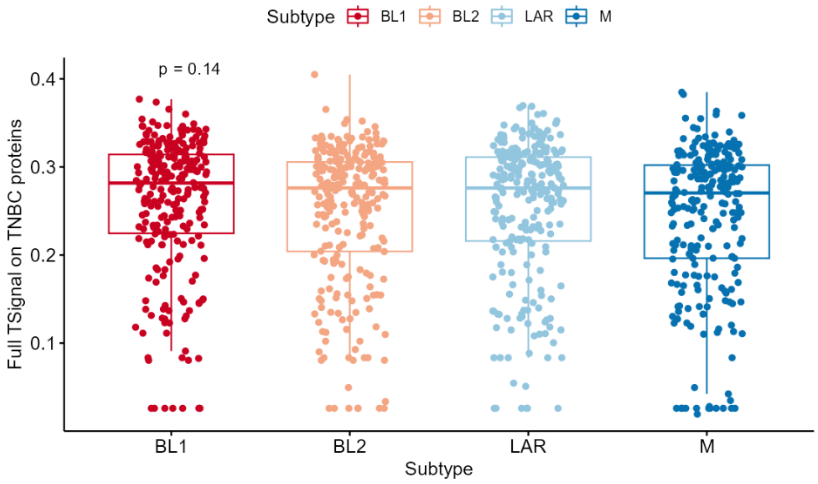


Figure 3: *In-silico* efficacy of anti-PD1 treatment of the 4 TNBC subtypes derived from the gene signature identified in an external TNBC population (Akhouayri, et al.).

References

Akhouayri, L.*, et al.* Identification of a minimum number of genes to predict triple-negative breast cancer subgroups from gene expression profiles. *Hum Genomics* 2022;16(1):70.

Chen, X.*, et al.* TNBCtype: A Subtyping Tool for Triple-Negative Breast Cancer. *Cancer Inform* 2012;11:147-156.

Cortes, J.*, et al.* Pembrolizumab plus chemotherapy versus placebo plus chemotherapy for previously untreated locally recurrent inoperable or metastatic triple-negative breast cancer (KEYNOTE-355): a randomised, placebo-controlled, double-blind, phase 3 clinical trial. *Lancet* 2020;396(10265):1817-1828.

Gutierrez-Casares, J.R.*, et al.* Methods to Develop an in silico Clinical Trial: Computational Head-to-Head Comparison of Lisdexamfetamine and Methylphenidate. *Front Psychiatry* 2021;12:741170.

Hartung, C.*, et al.* Identifying High-Risk Triple-Negative Breast Cancer Patients by Molecular Subtyping. *Breast Care (Basel)* 2021;16(6):637-647.

Jorba, G.*, et al.* In-silico simulated prototype-patients using TPMS technology to study a potential adverse effect of sacubitril and valsartan. *PLoS One* 2020;15:e0228926.

Lehmann, B.D.*, et al.* Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Invest* 2011;121(7):2750-2767.

Lehmann, B.D.*, et al.* Multi-omics analysis identifies therapeutic vulnerabilities in triple-negative breast cancer subtypes. *Nat Commun* 2021;12(1):6276.

Lehmann, B.D.*, et al.* Refinement of Triple-Negative Breast Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection. *PLoS One* 2016;11(6):e0157368.

Mandapati, A. and Lukong, K.E. Triple negative breast cancer: approved treatment options and their mechanisms of action. *J Cancer Res Clin Oncol* 2023;149(7):3701-3719.

Lipid mechanisms drive cerebrovascular disease in cognitively unimpaired individuals at low risk for late-life dementia

Patricia Genius[1], Blanca Rodríguez-Fernández, Carol Minguillón, Manel Esteller, Carole H. Sudre, Arcadi Navarro, Juan Domingo Gispert, Natalia Vilor-Tejedor, for the ALFA study

1. Barcelonaβeta Brain Research Center (BBRC), Pasqual Maragall Foundation, Barcelona, Spain.

Abstract

Cardiovascular risk factors (CVRF) increase the risk of cerebrovascular disease. However, asymptomatic middle-aged individuals with a low cardiovascular risk profile display cerebrovascular lesions, particularly white matter hyperintensities (WMH). WMH are a hallmark of cerebral small vessel disease (SVS) and have been linked to a higher risk of dementia. Understanding modifiable mechanisms leading to cerebrovascular disease is fundamental for implementing preventive strategies.

We aimed to elucidate the biological mechanisms underlying the presence of SVS in cognitively unimpaired (CU) middle-aged individuals at low risk for late-life dementia.

We included 1,139 CU participants from the ALFA study with magnetic resonance imaging data, genotyping, and Alzheimer's disease-related risk factors assessments. We assessed genetic predisposition to WMH (Persyn *et al.*, 2020) using polygenic scoring (PRS-WMH). Individuals were classified into risk groups for late-life dementia using the CAIDE score (Kivipelto *et al.*, 2006). Covariate-adjusted Spearman's rank correlation tests evaluated the association between the PRS-WMH and global WMH volumes, adjusting for age and sex. An enrichment analysis (Wu *et al.*, 2021) of the PRS-annotated genes unveiled the biological mechanisms leading to WMH burden. Group-specific effects were explored based on dementia-related CVRF.

Genetic predisposition to WMH was associated with larger WMH volumes in individuals at low cardiovascular risk for late-life dementia [Figure 1]. Lipid-related biological processes were driving WMH genetic risk [Figure 2]. Individuals genetically predisposed to display larger WMH volumes were either hypercholesterolemic, older than 55 or with lower educational attainment [Figure 3].

Lipid-related mechanisms contribute to SVS in individuals at lower cardiovascular risk for late-life dementia. These individuals should be considered for lipid-modifying therapies to prevent dementia later in life.

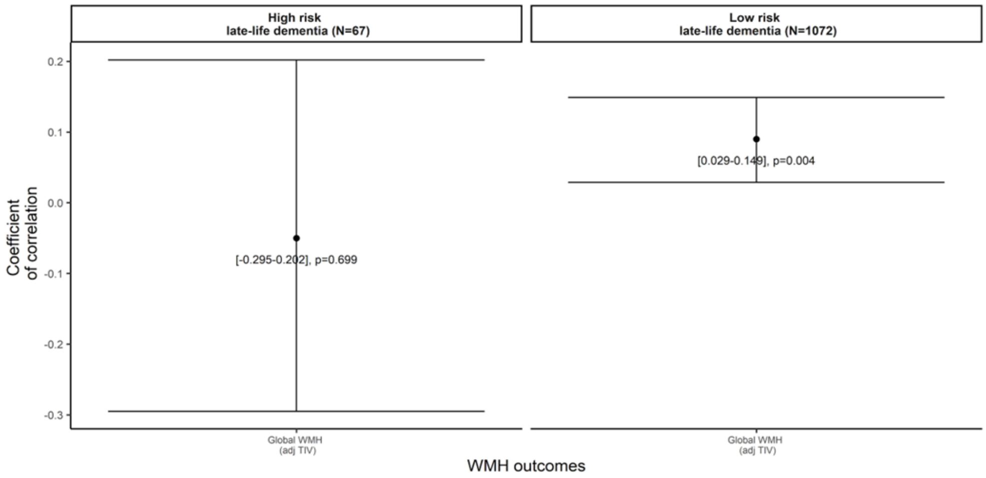


Figure 1. Covariate adjusted Spearman's rank correlation test assessing the association between WMH volumes with the genetic predisposition to WMH volumes. Models were stratified by the 20-years risk of dementia and adjusted for age and sex. Confidence intervals and p-values were reported. *Legend: WMH (White matter hyperintensities), TIV (Total Intracranial Volume).*

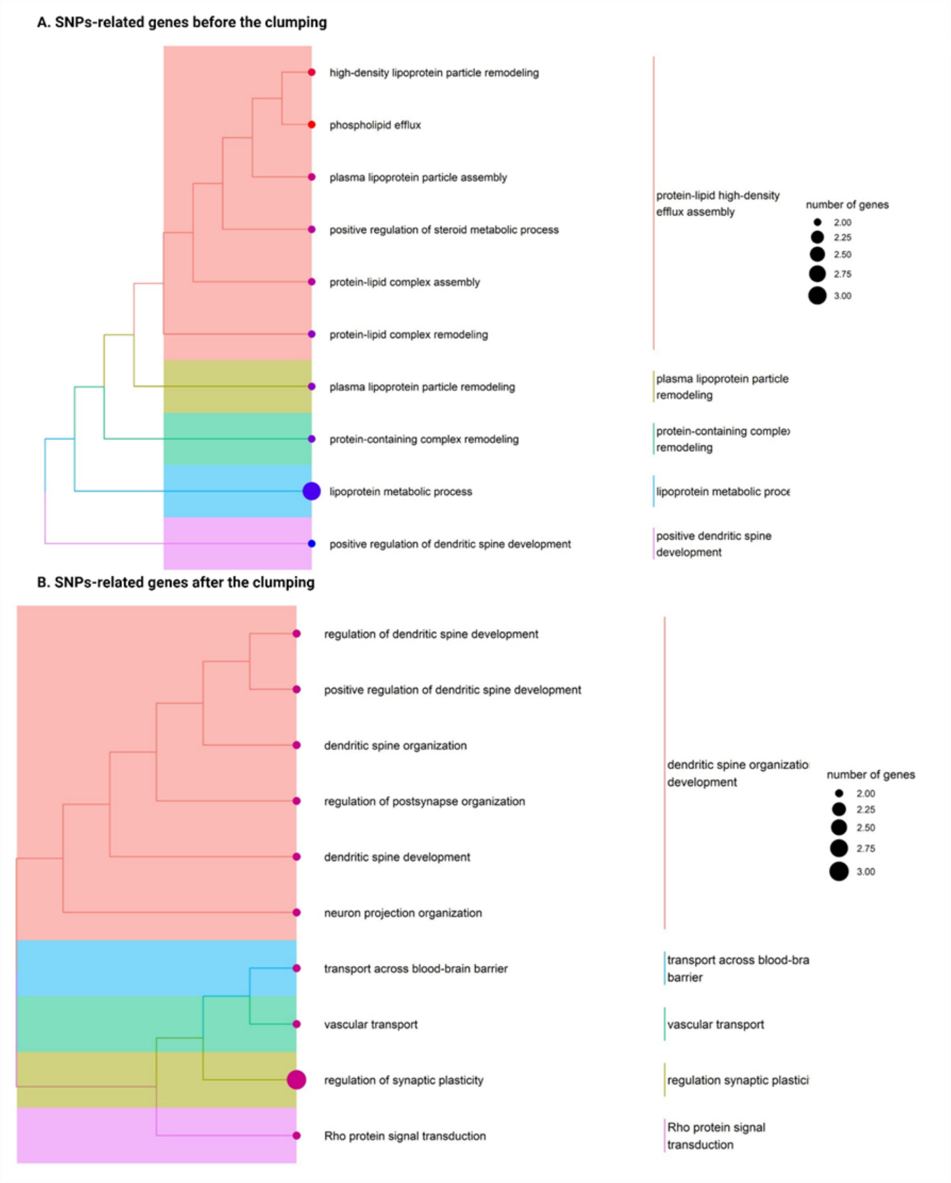


Figure 2. Enrichment analysis results display the main biological processes in which SNPs-annotated genes are involved. Biological mechanisms are grouped into main functions based on their similarity. Panel (2.A) displays the enrichment analysis based on the full spectrum of genetic variants associated with WMH. Panel (2.B), shows the enrichment analysis working with the specific SNPs that remained after the clumping. Significant results were reported at nominal p-value <0.05.

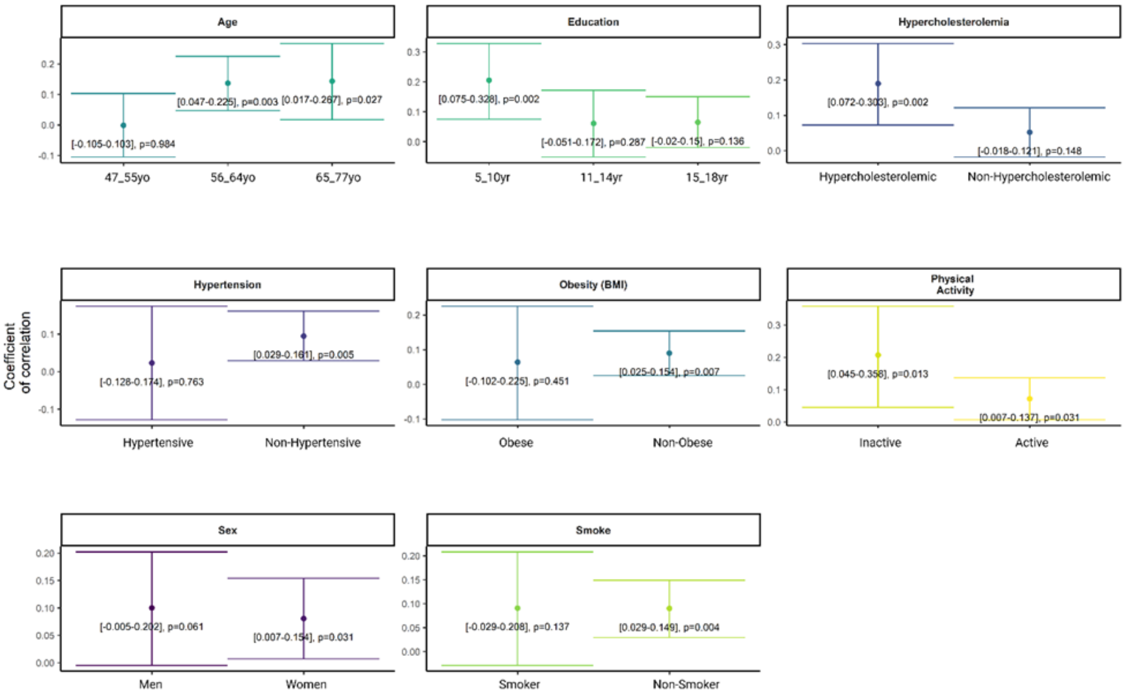


Figure 3. Covariate adjusted Spearman's rank correlation test assessing the association between global WMH volumes and genetic predisposition to white matter hyperintensities volumes. Models were stratified by CAIDE-I components. Age and sex were included as covariates. Confidence intervals and nominal p-values were reported.

References

1. Kivipelto,M. *et al.* (2006) Risk score for the prediction of dementia risk in 20 years among middle aged people: a longitudinal, population-based study. *Lancet Neurol.*, **5**, 735–741.
2. Persyn,E. *et al.* (2020) Genome-wide association study of MRI markers of cerebral small vessel disease in 42,310 participants. *Nat. Commun.*, **11**, 2175.
3. Wu,T. *et al.* (2021) clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)*, **2**, 100141.

Datoma: A cloud computing platform for high-performance metabolomics data analysis

Roger Giné Bertomeu[1], Enrique Martínez Martínez[2], Eduard Bel Ribes[1,3], Pedro García López[2], Òscar Yanes Torrado[1,3,4]

1 Universitat Rovira I Virgili. Department d'Enginyeria Elèctrica, Electrònica i Automàtica (DEEEiA). Metabolomics Interdisciplinary Laboratory (MILAB)

2 Universitat Rovira I Virgili. Department d'Enginyeria Informàtica (DEI). Cloud and Distributed Systems Lab (CLOUDLAB)

3 Institut d'Investigació Sanitària Pere Virgili (IISPV). Metabolomics Platform.

4 Centro de Investigación Biomédica en Red de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM)

Abstract

Cloud computing responds to the ever-expanding need for fast, scalable and reproducible data analysis workflows. Using cloud computing, research groups can process large data throughputs without maintaining their own in-house clusters. In the field of metabolomics, several web-based and cloud-based data analysis platforms have arisen in the past years: Worfkflow4Metabolomics (Giacomoni *et al.*, 2014), Metaboanalyst (Pang et al., 2022), Phenomenal (Peters et al., 2018), XCMSonline (Tautenhahn et al., 2012), METLIN and GNPS (Aron et al., 2020). However, to this date, all such platforms have been designed with a rather narrow use scope, offering a limited set of data analysis tools.

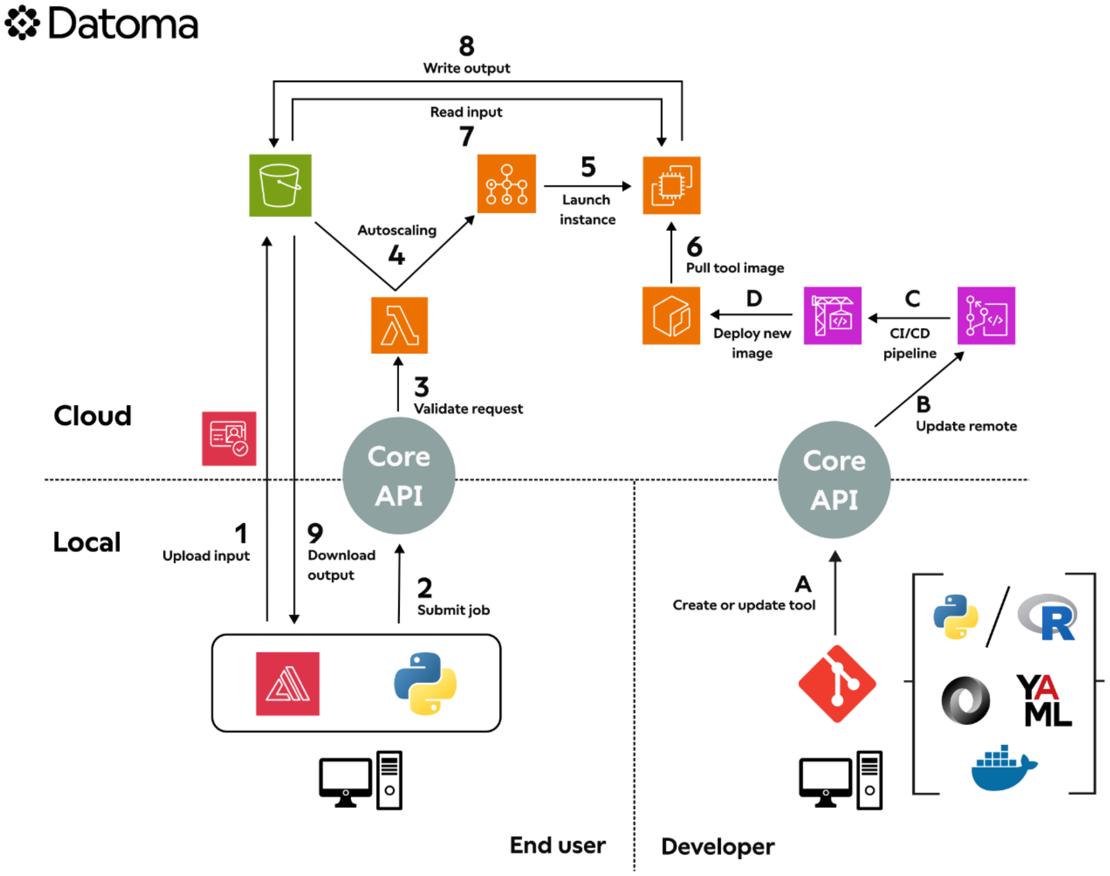


Figure 1: Simplified Datoma infrastructure for two use cases: Running a job (1-9) and updating a bioinformatics tool (A-D). Running a job: Users use either the Python package or the web application to upload their files to their

personal Datoma storage folder (1). Then users perform a SubmitJob request (2), which is validated by the core API (3) and, based on the uploaded files and the autoscaling configuration, a Batch job request is created with the needed resources. Then an appropriate instance is launched (5) and the corresponding tool image is pulled (6). The instance then reads the input files (7) and writes the outputs to the user folder, which can then be downloaded (9). Updating a tool: the bioinformatics developer creates a Git repository with the parametric task scripts in Python or R, configuration files and a Dockerfile. They then perform a UpdateTool request to the CoreAPI (A), the upstream Git remote is updated (B), and a CI/CD pipeline is run based on the code (C). If the pipeline is successful, the created Docker image is deployed and ready to be used by the end users.

We present Datoma, a cloud-native computing platform (Figure 1) that allows users to execute >20 curated metabolomics bioinformatics tools using an intuitive web-based interface or programmatically via a dedicated Python package and API. Datoma has been designed with modularity and interoperability in mind: any bioinformatics tools (R-based or Python-based) can be easily migrated to Datoma by creating a reproducible Docker image and a parametric task script. Complex workflows can then be easily defined by mapping tool output files to the inputs of other tools using regular expressions. Cost-efficiency has also been considered in Datoma: tasks dynamically scale the required compute resources needed based on the inputs.

We processed a large-scale (>100GB) imaging-MS dataset with the tool rMSI (Ràfols et al., 2017) and achieved a >10x speed-up compared to a regular workstation (from 8-10 hours, to <30 minutes). Datoma thus allows reproducible, scalable and faster data analysis while avoiding dependency management and using tool-specific APIs.

References

1. Giacomoni,F. et al. (2014) Workflow4Metabolomics: A collaborative research infrastructure for Computational Metabolomics. Bioinformatics, 31, 1493–1495.
2. Pang,Z. et al. (2022) Using Metaboanalyst 5.0 for LC–HRMS spectra processing, multi-omics integration and covariate adjustment of Global Metabolomics Data. Nature Protocols, 17, 1735–1761.
3. Peters,K. et al. (2018) Phenomenal: Processing and analysis of Metabolomics data in the cloud. GigaScience, 8.
4. Tautenhahn,R. et al. (2012) XCMS online: A web-based platform to process untargeted metabolomic data. Analytical Chemistry, 84, 5035–5039.
5. Aron,A.T. et al. (2020) Reproducible molecular networking of untargeted mass spectrometry data using GNPS. Nature Protocols, 15, 1954–1991.
6. Ràfols,P. et al. (2017) RMSI: An R package for MS Imaging Data Handling and Visualization. Bioinformatics, 33, 2427–2428.

Deep Learning Based Methods for Fundus Image Quality Evaluation

Rubén G.Barriada[1] and David Masip[1]

[1]AIWell Research Group, Faculty of Computer Science Multimedia and Telecommunications
Universitat Oberta de Catalunya, 08018 Barcelona, Spain

Abstract

Retinal imaging has become an important source of information in the evaluation of different eye disorders, such as diabetic retinopathy, age-related macular degeneration or glaucoma. Besides, in recent years, retinal imaging has gained a relevant importance in the field of oculomics, helping in the diagnosis of complex systemic diseases such as Alzheimer's disease, dementia, or cardiovascular diseases (ischaemic stroke, myocardial infarction and heart failure). Retinal images are obtained by different acquisition devices, by people with different levels of experience and with significant variance in the illumination and relative position of the retinal quadrants. Therefore, there is a large variation in the quality of the images used for automated diagnosis. Thus, objectively evaluating the quality of retinal images is essential for a reliable diagnosis and this is possible by applying deep learning methods. In this work, we implemented a Convolutional Neural Network (CNN) to evaluate the quality of a retinal image based on a three-level quality grading system (i.e., 'Good', 'Usable' and 'Reject') using the EyePACS dataset with 28,792 retinal images. Based on the work in *Fu, H. et al. (2019)*, we trained a multi-class classifier to evaluate the quality of the images with an accuracy of 90.50 ± 0.71. In addition, we visually evaluated the performance of the model on two external retinal datasets dedicated to the prediction of Coronary Artery Calcium and cardiovascular events. As future work, we intend to integrate this mechanism in deep learning applications in order to evaluate how a poor quality images filtering may impact in the learning process and thus on the predictive capability.

Keywords— Retinal image, Quality assessment, Deep learning

References

[1] Fu, H. et al. (2019). Evaluation of Retinal Image Quality Assessment Networks in Different Color-Spaces. In: Shen, D., et al. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science(), vol 11764. Springer, Cham.

[2] Zhou, Y., Chia, M.A., Wagner, S.K. et al. A foundation model for generalizable disease detection from retinal images. Nature 622, 156–163 (2023)

Bioactive peptides in Mediterranean plants: biological properties and pharmacological implications

Julia Lisa-Molina[1 2 3 *], María Mulet[1 2], Jose Antonio Sánchez-Milán[1 2], María Fernández Rhodes[1], Aida Serra[1+], Xavier Gallart-Palau[2+]

1. Department of Basic Medical Sciences, University of Lleida (UdL) – Biomedical Research Institute of Lleida (IRB Lleida) - +Pec Proteomics Research Group (+PPRG) - Neuroscience Area, 25198 Lleida, Spain.
2. Biomedical Research Institute of Lleida Dr. Pifarré Foundation (IRBLLEIDA) - +Pec Proteomics Research Group (+PPRG) - Neuroscience Area – University Hospital Arnau de Vilanova (HUAV) – Health Campus, Psychology Dept., University of Lleida, Spain.
3. Algèmica Barcelona S.L. Camí Sant Bartomeu 5, 08600 Berga (Barcelona), Spain.

+ Correspondence: Xavier Gallart-Palau, PhD. xgallart@irblleida.cat ; Aida Serra, PhD. aida.serra@udl.cat.
*Presenter: Julia Lisa-Molina jlm13@alumnes.ub.edu

Abstract

The Mediterranean basin is one of the richest plant biodiversity hotspots in the World, with a vast history of traditional uses of medicinal plants[1,2]. Nonetheless, countless bioactive principles of said plants (especially peptidic ones) remain to be uncovered[3]. Bioactive Peptides (BPs) are short sequences of amino acids (<10kDa) with high potential to modulate organismic functions beyond basic nutritional needs. A class of BPs, known as cysteine-rich peptides (CRPs), is distinguished by the abundance of cysteines in their sequences[4]. The enhanced presence of disulfide bonds in CRPs, thus, implies heightened thermal and proteolytic stability, making them highly valuable for nutraceutical and pharmaceutical applications[3,4]. The Plant Kingdom becomes a riveting source for the isolation of BPs[5], and even though a handful of plants have been found to produce CRPs[6,7], many species in this life kingdom remain to be explored and delved into.

This project aims to expand our characterization and understanding of endogenously synthesized CRPs by Mediterranean plants, thereby exploring the potential for new local medicinal products. To achieve this, we focused on 3921 species belonging to 27 different Mediterranean plant families that are potentially capable of synthesizing CRPs. A bioinformatics sequence-pattern-based approach was then employed to screen transcriptomic, genomic, and proteomic data available in public repositories and to predict potential CRP sequences associated with the mentioned plant species. To date, a total of 53 plants have been thoroughly screened, among which 31 have the potential to produce CRPs. Specifically, our bioinformatics analyses have identified a total of 118,392 potential CRPs.

The predicted CRP sequences in these plants are currently being screened using chromatographic techniques along with Mass Spectrometry-driven proteomics. Simultaneously, additional bioinformatics analyses are being conducted to explore potential functional applications associated with these sequences. Furthermore, isolated CRPs will undergo in vitro assays based on the aforementioned predictions to evaluate and characterize potential bioactivities. These findings will also eventually be validated through *in vivo* approaches. The data generated in this project, thus, holds promise to expand the current available libraries including potential nutraceutical and therapeutic compounds for thorough screening in several biotechnological and biomedical applications.

References

1. Gonzalez-Tejero, M.R*., et al.* Medicinal plants in the Mediterranean area: synthesis of the results of the project Rubia. *J Ethnopharmacol* 2008;116(2):341-357.

2. Ramos-Gutiérrez, I*., et al.* Atlas of the vascular flora of the Iberian Peninsula biodiversity hotspot (AFLIBER). *Global Ecol Biogeogr* 2021;30(10):1951-1957.

3. Meena, S*., et al.* Biologia futura: medicinal plants-derived bioactive peptides in functional perspective-a review. *Biol Futur* 2020;71(3):195-208.

4. Lavergne, V., Taft, R.J. and Alewood, P.F. Cysteine-rich mini-proteins in human biology. *Curr Top Med Chem* 2012;12(14):1514-1533.

5. Fan, H*., et al.* Review on plant-derived bioactive peptides: biological activities, mechanism of action and utilizations in food development. *Journal of Future Foods* 2022;2(2):143-159.

6. Zorin, E.A*., et al.* A variable gene family encoding nodule-specific cysteine-rich peptides in pea (Pisum sativum L.). *Front Plant Sci* 2022;13:884726.

7. Tam, J.P*., et al.* Ginsentides: Cysteine and Glycine-rich Peptides from the Ginseng Family with Unusual Disulfide Connectivity. *Sci Rep* 2018;8(1):16201.

Identification of epigenetic biomarkers for molecular subgrouping of ependymoma

Joshua Llano-Viles[1,2,3,4], Soledad Gómez-González[3,4], Cinzia Lavarino[3,4], Alexandre Perera-Lluna[1,2]

1 B2SLab, Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, Barcelona, Spain

2 Networking Biomedical Research Centre in the subject area of Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Madrid, Spain

3 Institut de Recerca Pediàtrica Hospital Sant Joan de Dèu, Esplugues de Llobregat, Barcelona, Spain

4 Pediatric Cancer Center Barcelona (PCCB), Hospital Sant Joan de Déu, Barcelona, Spain.

Abstract

Ependymomas represent rare neoplasms occurring within the central nervous system (CNS), manifesting in either the supratentorial region or the posterior fossa of the brain, as well as within the spinal cord. Their occurrence demonstrates a slightly lower frequency in children (3/1,000,000) in contrast to adult men and women (5/1,000,000 and 4/1,000,000 respectively). Notably, pediatric patients exhibit a particularly unfavorable prognosis (Saleh et al., 2022).

The prognosis of ependymomas varies based on clinical, histopathological, and molecular characteristics (Pajtler et al., 2015). In 2016, the World Health Organization (WHO) introduced a classification system for ependymoma subtypes, integrating histopathological and molecular parameters (Louis et al., 2016). However, numerous studies have indicated that risk stratification using molecular subgrouping derived from methylation profiling surpasses the conventional reliance on histological grading. Consequently, the 2021 revision of the WHO classification for CNS tumors marks a significant departure from the prior histomorphological classification of ependymal tumors. This updated classification delineates ten distinct types of ependymomas, characterized by both their anatomical location and molecular features (Louis et al., 2021; Saleh et al., 2022).

Nevertheless, employing array-based technology in a standard diagnostic setting poses challenges due to its time-consuming nature, substantial cost, and, in many cases, limited accessibility for numerous medical centers globally that manage patients with brain tumors. As a result, a considerable cohort of patients is unable to avail themselves of the clinical advancements linked to the methylation-based classification of ependymomas.

The primary objective of this study is to extract epigenetic biomarkers derived from the methylation profiles distinctive to each molecular subgroup of ependymoma. A key criterion for these biomarkers is their potential integration into routine clinical practice. To achieve this, we amassed DNA methylation microarray data from multiple sources, encompassing 8 distinct studies (n=1748), utilizing both the Illumina Infinium HumanMethylation 450

BeadChip (HM450K) and the Illumina methylation EPIC BeadChip array (EPIC) (Mack et al., 2014; Pajtler et al., 2018; Fukuoka et al., 2018; Brabetz et al., 2018; Rogers et al., 2018; Capper et al., 2018; Cavalli et al., 2018; Michealraj et al., 2020). First, we updated the subgroup assignment of the samples according to the WHO 2021 classification. Then, we conducted an unsupervised analysis of the samples. In this process, we implemented a filter to eliminate cytosines that were missing in any of the data matrices in order to merge them. Subsequently, we normalized the data to mitigate any potential bias arising from batch variations. Following this, we performed a principal component analysis using a set of cytosines selected based on their standard deviation. During this stage, we noticed a significant influence of the anatomical location of the samples on our cytosine measurements. To mitigate this effect, we used linear models to eliminate the location information. After this correction, we employed a random forest model to train the adjusted data and successfully predict the subgroups accurately. In the current phase of our work, our focus is on reducing the number of cytosines used by the random forest model to effectively adapt it for clinical settings. This reduction process will enable us to simplify the model and facilitate its practical application in clinical medicine. In addition, we will compare these cytosines from ependymoma samples with cytosines of other pediatric tumors and healthy tissues in order to identify and filter cytosines capable of exclusively discriminating ependymoma subtypes.

References

1. Saleh, A. H. et al (2022). The biology of ependymomas and emerging novel therapies. *Nature Reviews Cancer*, *22*(4), 208–222. https://doi.org/10.1038/s41568-021-00433-2

2. Pajtler, K. W. et al (2015). Molecular Classification of Ependymal Tumors across All CNS Compartments, Histopathological Grades, and Age Groups. *Cancer Cell*, *27*(5), 728–743. https://doi.org/10.1016/j.ccell.2015.04.002

3. Louis, D. N. et al (2016). The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathologica*, *131*(6), 803–820. https://doi.org/10.1007/s00401-016-1545-1

4. Louis, D. N. et al (2021). The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro-Oncology*, *23*(8), 1231–1251. https://doi.org/10.1093/neuonc/noab106

5. Mack, S. C. et al (2014). Epigenomic alterations define lethal CIMP-positive ependymomas of infancy. *Nature 2014 506:7489*, *506*(7489), 445–450. https://doi.org/10.1038/nature13108

6. Pajtler, K. W. et al (2018). Molecular heterogeneity and CXorf67 alterations in posterior fossa group A (PFA) ependymomas. *Acta Neuropathologica*, *136*(2), 211–226. https://doi.org/10.1007/S00401-018-1877-0/METRICS

7. Fukuoka, K. et al (2018). Significance of molecular classification of ependymomas: C11orf95-RELA fusion-negative supratentorial ependymomas are a heterogeneous group of tumors. *Acta Neuropathologica Communications*, *6*(1), 134. https://doi.org/10.1186/S40478-018-0630-1/FIGURES/5

8. Brabetz, S. et al (2018). A biobank of patient-derived pediatric brain tumor models. *Nature Medicine 2018 24:11*, *24*(11), 1752–1761. https://doi.org/10.1038/s41591-018-0207-3

9. Rogers, H. A. et al (2018). Limitations of current in vitro models for testing the clinical potential of epigenetic inhibitors for treatment of pediatric ependymoma. *Oncotarget*, *9*(92), 36530–36541. https://doi.org/10.18632/ONCOTARGET.26370

10. Capper, D. et al (2018). DNA methylation-based classification of central nervous system tumours. *Nature 2018 555:7697*, *555*(7697), 469–474. https://doi.org/10.1038/nature26000

11. Cavalli, F. M. G. et al (2018). Heterogeneity within the PF-EPN-B ependymoma subgroup. *Acta Neuropathologica*, *136*(2), 227–237. https://doi.org/10.1007/S00401-018-1888-X/METRICS

12. Michealraj, K. A. et al (2020). Metabolic Regulation of the Epigenome Drives Lethal Infantile Ependymoma. *Cell*, *181*(6), 1329-1345.e24. https://doi.org/10.1016/j.cell.2020.04.047

13. Gomez, S. et al (2018). A novel method for rapid molecular subgrouping of medulloblastoma. *Clinical Cancer Research*, *24*(6), 1355–1363. https://doi.org/10.1158/1078-0432.CCR-17-2243

Evolution of morphological complexity under development-based genotype-phenotype maps

Loreto Velázquez Antonio[1], Salazar-Ciudad Isaac[1,2]

1 Genomics, Bioinformatics and Evolution group, Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain.

2 Centre de Rercerca Matemàtica, Cerdanyola del Vallès, Spain.

Abstract

Morphological complexity in organisms arises through development; however, how does development produce complexity in the first place is an open question that can only be fully understood when considering how developmental mechanisms have evolved. This study aims to investigate how does morphological complexity arise in evolution and development. To achieve this, we will study evolution by combining EmbryoMaker, a realistic computational model of development (Miquel *et al.,* 2016) with a population genetics model that considers reproduction, mutation and uses morphological complexity as the selection criteria. In EmbryoMaker, a morphology, i.e., a specific distribution of cells and gene expression in 3D, is generated by specifying a gen regulatory network, how gene products influence cell signaling, animal cell behaviors (cell division, apoptosis, contraction, adhesion, etc.) and by stating the biophysics that govern cells and tissues. We expect to identify the kind of changes (e.g., gene network topology, type of developmental mechanism, etc.) that lead to complexity and, secondly, to understand the influence of morphological complexity/simplicity on adaptive dynamics in both the short and long-term.

References

1. Miquel Marin-Riera. *et al.* (2016) Computational modeling of development by epithelia, mesenchyme and their interactions: a unified model, *Bioinformatics*, Volume 32, Issue 2, Pages 219–225.

ClinBioNGS: an integrated clinical bioinformatics pipeline for the analysis of somatic NGS cancer panels

Marín R[1,2], Alay A[1,2], Hijazo-Pechero S[2], Moreno V[1,2], Nadal E[1,2], Solé X[3]

[1]Unit of Bioinformatics for Precision Oncology, Catalan Institute of Oncology (ICO), L'Hospitalet de Llobregat, Spain.

[2]Preclinical and Experimental Research in Thoracic Tumors (PrETT), Molecular Mechanisms and Experimental Therapy in Oncology Program (Oncobell), Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Spain.

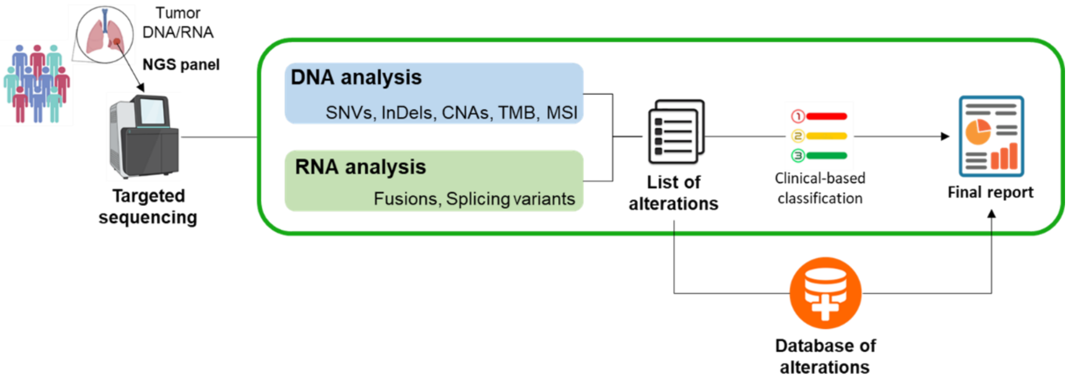
[3]Molecular Biology CORE, Center for Biomedical Diagnostics (CDB), Hospital Clínic de Barcelona, Spain.

Abstract

Tumor targeted sequencing panels currently in use cover both DNA and RNA alterations to improve the molecular clinical diagnostics process. However, in terms of bioinformatics analysis, commercial panels often provide proprietary and non-customizable solutions which cannot be tailored to the user preferences. Additionally, these tools offer very limited graphical reports, hindering the interpretability of the results. Here we present ClinBioNGS, an open-source and customizable clinical bioinformatics pipeline to identify both DNA and RNA alterations in targeted NGS panels. ClinBioNGS provides interpretable and visual results, and can also keep an up-to-date database of all identified alterations in the samples sequenced in the laboratory. ClinBioNGS currently works both for Illumina TruSight Oncology 500 (TSO500) and ThermoFisher Oncomine Precision (OPA) and Comprehensive (OCA) panels.

We have compared the results of our pipeline to the commercial one in ~700 samples profiled in-house with Illumina TSO500. ClinBioNGS detected more than 99% of all the relevant alterations reported by Illumina. Additionally, we reported 6% (~300) more SNVs/InDels and ~4X more CNAs (264 vs 1022). Both pipelines detect the same number of clinically relevant splice variants (16), but we also identified some extra cancer-related ones (100) that TSO500 failed to report. We also detected 3 extra cancer-related fusions (78 vs 81; *TPM3-NTRK1*, *KMT2A-MLLT3*, *MKRN2-PPARG*), with 1 of them being potentially actionable (*NTRK1* fusion). A similar analysis with ~300 samples profiled with ThermoFisher OPA is currently ongoing, and our pipeline not only detects the relevant alterations reported by ThermoFisher, but also some extra ones providing more genomic insight from the data.

All this makes ClinBioNGS a robust bioinformatics pipeline that allows for a better detection, visualization, and interpretation of tumor alterations in the context of somatic molecular diagnostics of cancer patients.



Multimodal data integration to model, predict, and understand changes in plant biodiversity.

Martinez, E. S. [1, 2], Tejada Gutierrez, E. L. [1], Defacio, R. A.[2], Vaqueiro, R. C. [1]

1 Department of Basic Medical Sciences, University of Lleida, Spain.

2 National Institute of Agricultural Technology (INTA), Argentina.

Abstract

Biodiversity conservation is crucial for the maintenance of ecosystem services, food security, and human health. Climate change affects plant biodiversity but knowledge about its status and the threats it faces is often incomplete (Framework for Action on Biodiversity for Food and Agriculture, 2022). Studies about climate change and biodiversity still focus on short-term effects, analyze only two variables at the same time, or use data sources with low temporal and spatial resolution. In this context, the ForestForward database was created by Tejada Gutierrez *et. al.* (2022). It contains over 3.000 datasets with information on plant abundance dating back to the last century. Still, ForestForward needs to incorporate information about climate change and other important factors that affect biodiversity. My thesis aims to develop strategies to integrate and analyze multimodal data to understand changes in plant biodiversity over time and geography. First, we will search for different sources to obtain data on climate, topography, and land use and align it, spatially and temporally, with data on plant abundance to create a time series dataset in the ForestForward. Second, we will analyze the time series dataset using Autoregressive Integrated Moving Average Models (ARIMA) and quantify the relationship between all variables. Third, we will develop tools for predictive modeling of the changes in biodiversity, employing machine learning algorithms. We will then use these models to test future alternative scenarios. Finally, we will use our analysis to identify critical areas for agrobiodiversity conservation, prioritizing species such as corn (*Zea mays* L.). We are adding multimodal data to the knowledge base of ForestForward to enhance its ability to analyze and understand the changes in plant biodiversity over time. The expected results will contribute to our understanding of the relationship between the changes in plant biodiversity and climate change over the past few decades.

References

Framework for Action on Biodiversity for Food and Agriculture. (2022). FAO. https://doi.org/10.4060/cb8338en

Tejada Gutierrez, E.L., et al. (2022). ForestForward: aplicación web para la visualización a nivel mundial de datos forestales del último siglo. 8° Congreso Forestal Español: La ciencia forestal y su contribución a los objetivos de desarrollo sostenible. 8CFE-103. Sociedad Española de Ciencias Forestales. ISBN 978-84-941695-6-4.

dsFDL: DataSHIELD Federated Deep Learning for Secure and Collaborative AI in Healthcare

Ramon Mateo Navarro[1,2] and Juan R. Gonzalez[1,3]

[1]Barcelona Institute for Global Health (ISGlobal)
[2]Institut de Bioenginyeria de Catalunya (IBEC)
[3]Universitat Autònoma de Barcelona (UAB)

Abstract

Managing medical data across hospital networks demands rigorous privacy preservation, especially crucial when handling data distributed through these networks. This approach is vital as it deals with extremely sensitive information subject to regulations such as GDPR. To address these needs, we introduce *dsFDL*, a software package tailored for medical applications in federated learning, image processing, and radiomic feature extraction. Moreover, this analysis tends to be highly computationally demanding, focusing on high-dimensional medical imaging. Our software offers support for both R and Python, the latter enabling GPU-accelerated performance. This acceleration is essential for conducting deep analysis of medical imaging within a reasonable time frame.*dsFDL* seamlessly integrates with DataSHIELD, leveraging robust data protection mechanisms of DataSHIELD and the computational prowess of GPU-accelerated federated learning. This approach not only meets the stringent privacy standards required in health data ecosystems but also facilitates advanced image analysis and the ethical application of AI in medical research.

References

[1] Yang Q, Liu Y, Cheng Y, Kang Y, Chen T, Yu H (2020). Horizontal Federated Learning. En: *Federated Learning*. Springer, Cham. DOI: 10.1007/978-3-031-01585-4_4.

[2] Kirillov A, Xiao T, et al. (2023). Segment Anything. arXiv preprint arXiv:2304.02643. Available at: http://arxiv.org/abs/2304.02643.3.

Transposons in the evolution of piRNA cluster expression in mice

Adrià Mitjavila Ventura[1,2,3], Tanya Vavouri[2]

1. Germans Trias i Pujol Research Institute (IGTP), Can Ruti Campus, 08916 Badalona, Spain.

2. Regulatory Genomics group, Josep Carreras Leukaemia Research Institute (IJC), Can Ruti Campus, 08916 Badalona, Spain.

3. PhD Program in Bioinformatics, Universitat Autònoma de Barcelona (UAB), 08193 Cerdanyola del Vallès, Spain.

Abstract

Piwi-interacting RNAs (piRNAs) are small non-coding RNAs expressed in the animal germline[1]. They are produced from long single-stranded transcripts that derive from discrete genomic loci called piRNA clusters[1]. piRNAs and piRNA clusters are highly diverged between species showing almost no evidence of selection constraint[2,3]. Considering their fast turnover, we wondered how the expression of piRNA clusters evolves in short evolutionary time scales. To address this, we focused on differences in postnatal piRNA expression in different inbred strains of mice and closely related murine species. We found significant differences in piRNA clusters within and across species. Comparing the expression of the piRNA clusters across mouse species, we found that piRNA expression level correlated with conservation of the piRNA clusters, while species-specific clusters showed fewer and more variable piRNA production. We found that clusters with polymorphic endogenous retroviruses were overrepresented among those with highly variable piRNA cluster expression, likely contributing to transcriptional activation and post-transcriptional processing of novel piRNA clusters. Taken together our results suggest young endogenous retroviruses as potent drivers of piRNA cluster gains and that piRNA abundance constrains piRNA evolution.

References

1. Ozata et al. (2019) PIWI-interacting RNAs: small RNAs with big functions. Nat Rev Genet. doi: 10.1038/s41576-018-0073-3.

2. Chirn et al. (2015) Conserved piRNA expression from a distinct set of piRNA cluster loci in eutherian mammals. PLoS Genet. doi: 10.1371/journal.pgen.1005652.

3. Ozata et al (2020). Evolutionary conserved pachytene piRNA loci are highly divergent among modern humans. Nat Ecol Evol. doi: 10.1038/s41559-019-1065-1.

Molecular Dynamics study of the kynureninase enzyme: an approach for the design of new therapeutic enzymes in cancer

Javier Moreno Farre[1]

1 Universitat de Girona

Abstract

Controlling tumour production of L-kynurenine (KYN) is the focus of many studies and has the potential to treat, among others, cancers such as breast cancer, colorectal cancer and lung adenocarcinoma [1,2,3].

The present project overcomes a major deficiency in the art by providing enzymes that comprise bacterial and mammalian polypeptide sequences capable of degrading KYN and 3-hydroxy-L-kynurenine, thus displaying favourable activity desired for cancer therapy. The study will use molecular dynamics (MD) to determine the key factors that can affect the conformational preference of kynureninase (KYNU) and correlate them with its catalytic activity, as well as, the evaluation of the effect of mutations on the conformational dynamics of the enzyme.

The study is aiming to have a deeper understanding on the KYNU structure and activity by performing an exhaustive computational MD study of the enzyme and demonstrate the two main factors that influence the tautomeric equilibrium of the PLP -Schiff base: (a) the protonation state of the pyridine ring, and (b) the substituent on the imino nitrogen of the Schiff base as reported in other pyridoxal 5′- phosphate dependent enzymes such as L-dopa decarboxylase and alanine racemase [4]. The study will also focus on (1) the role of the strictly conserved Asp-168 and Asp-250 among kynureninases in the PLP pyridine nitrogen hydrogen bonded with the side chains and (2) will explore the PLP carbinolamine formation and the role of Histidine 253 conformation change.

A number of studies to evaluate the KYNU conformation changes with different substrates such as kynurenine and OH-kynurenine and the effects of distal and site mutations have been reported. These studies are mainly focused on pre-steady-state and hydrogen−deuterium exchange mass spectrometry (HDX-MS) methodologies to study the conformation of KYNUases (Homo Sapines HsKYNase and Pseudomonas fluorescencens PfKYNase) [5]. A recent study has also been reported to study the effects of active site and distal mutations using HDX-MS experiments in which distal mutations are more important in the activity of KYNU towards different substrates [6]. This research will aim to replicate these studies using MD methodologies

References

1. Et al Liu Y, Feng X, Lai J, et al. A novel role of kynureninase in the growth control of breast cancer cells and its relationships with breast cancer. J Cell Mol Med. 2019

2. Et al Al-Mansoob M. KYNU, a novel potential target that underpins CD44-promoted breast tumor cell invasion. J Cell Mol Med. 2021;25:2309–2314

3. Et al Fahrmann, J.F.; M.; et al. Mutational Activation of the NRF2 Pathway Upregulates Kynureninase Resulting in Tumor Immunosuppression and Poor Outcome in Lung Adenocarcinoma. Cancers 2022, 14, 2543. https://doi.org/10.3390/cancers14102543

4. Molecular dynamics simulations of the intramolecular proton transfer and carbanion stabilization in the pyridoxal 5′- phosphate dependent enzymes L-dopa decarboxylase and alanine racemase Biochim Biophys Acta. 2011 November ; 1814(11): 1438–1446. doi:10.1016/j.bbapap.2011.05.002.

5. Conformational Dynamics Contribute to Substrate Selectivity and Catalysis in Human Kynureninase ACS Chem. Biol. 2020, 15, 3159−3166

6. Bypassing evolutionary dead ends and switching the rate-limiting step of a human immunotherapeutic enzyme https://www.nature.com/articles/s41929-022-00856-6.

IDENTIFICATION OF CLINICAL FEATURES ASSOCIATED WITH SARS-COV-2 REINFECTIONS

Francisco Muñoz-López[1,4], Antoni E. Bordoy[7,8], Ignacio Blanco[7,9], Elisa Martró[7,8,11], Francesc Català-Moll[1], Mariona Parera[1], Pere-Joan Cardona[4,7,10], Cristina Casañ[7], Ana Blanco-Suárez[7], Roger Paredes[1,2,3,5,6], Bonaventura Clotet[1,2,3,5,6], Lourdes Mateu[2,3,5,10], Marc Noguera-Julian[1,5,6], José Ramón Santos[2,3], Marta Massanella[1,5,6]

[1]IrsiCaixa-AIDS Research Institute, Badalona, Spain

[2]Infectious Diseases Department, Germans Trias i Pujol Hospital, Badalona, Spain

[3]Fight Against AIDS Foundation (FLS), Germans Trias i Pujol Hospital, Badalona, Spain

[4]Autonomous University of Barcelona (UAB), Barcelona, Spain

[5]University of Vic - Central University of Catalona (UVic-UCC), Vic, Spain

[6]Centro de Investigación Biomédica en Red de Enfermedades Infecciosas (CIBERINFEC), Madrid, Spain

[7]Northern Metropolitan Clinical Laboratory, Microbiology Department, Hospital Universitari Germans Trias i Pujol (HUGTiP), Badalona, Spain.

[8]Fundació Institut d’Investigació en Ciències de la Salut Germans Trias i Pujol (IGTP), Badalona, Spain.

[9]Clinical Genetics Department, Northern Metropolitan Clinical Laboratory, Hospital Universitari Germans Trias i Pujol (HUGTiP), Badalona, Spain.

[10]Respiratory Diseases Networking Biomedical Research Centre (CIBERES). Instituto de Salud Carlos III, Madrid, Spain

[11]Epidemiology and Public Health Networking Biomedical Research Centre (CIBERESP). Instituto de Salud Carlos III, Madrid, Spain

Over 700 million of COVID-19 cases have been reported. A remarkable fragment of these cases are reinfections, which are mostly explained by the genomic variability of the SARS-CoV-2 variants. However, little is known about other factors fostering these reinfections.

We recorded clinical and demographic data from subjects (N=3303, March 2020 - March 2022) with at least 2 PCR+ events separated by ≥90 days, analyzed by the Microbiology Department, Northern Metropolitan Clinical Laboratory from Germans Trias i Pujol Hospital (Spain). Data collected included: age, sex, comorbidities, adjusted morbidity group (AMG), hospitalization, symptomatology, NAAT (PCR, TMA) tests, antigen tests, serology, and vaccination. Temporal data was encoded using Python, and demographic characterization was performed under R.

We identified 2344 cases of confirmed reinfections, where the 2 PCR+ events were separated by ≥90 days and a negative test was obtained between episodes. 72.2% of reinfected subjects were females with a median age of 45 IQR [28-63] years. Age density analysis showed three peaks at 24, 45, and 85 years, probably mostly composed of young people, who usually are less cautious, healthcare workers, and people living in nursing homes, respectively, being all of them groups prone to be tested. Regarding health status, 86.2% of participants had at least one chronic condition, with 40.5% of patients having chronic conditions in ≥4 systems based on AMG assessment. Interestingly, 75.2% of reinfected subjects <26 years had at least one chronic condition. 121 (4.2%) participants were hospitalized during a COVID-19 episode, highlighting 8.3% (N=10) of them hospitalized during the reinfection (half of them vaccinated before hospitalization), and 5% (N=6) of them during both infections. The severity of the second infection may be caused by a diminished acquired immunity after the first infection. Time between reinfections density analysis provided three peaks at ~200, ~400, and ~600 days, corresponding with time between waves. A decrease of reinfections was observed between 40 and 100 days after vaccination, which would be the period of highest protection against reinfection.

SARS-CoV-2 reinfections are more prevalent among women. Importantly, people with an undermined health status, independently of age, are more sensitive to reinfections, but in most of the cases no hospitalization was required. Finally, vaccination seems to have a short protective effect on reinfection.

Prognosis of patient groups with COVID-19, chronic diseases and polypharmacy. Mixed patient-centered approach

Cristina Muntañola Valero[1, 2, 3], Beatriz López Ibáñez, Isabel del Cura González[1, 2, 4, 5, 6]

1. Network for Research on Chronicity, Primary Care, and Health Promotion (RICAPPS), Spain

2. Research Unit, Primary Care Management, Madrid Health Service, Madrid, Spain

3. Fundación para la Investigación e Innovación Biosanitaria de Atención Primaria (FIIBAP)

4. Department of Medical Specialities and Public Health. Rey Juan Carlos University, Alcorcón, Madrid, Spain

5. Instituto Investigación Sanitaria Gregorio Marañón IiSGM

6. Ageing Research Center, Karolinska Institutet and Stockholm University, Stockholm, Sweden.

Abstract

Coronavirus is an infectious disease whose patients can be grouped together (1). These groups differ in their symptoms and clinical characteristics, hospital stay and mortality (2), which can be used to predict the potential prognosis of each (3). The prognosis of these patients is greatly influenced, inter alia, by their chronic diseases and their story of multimorbidity (MM) (4, 5, 6). Furthermore, individuals with polypharmacy have a higher risk of coronavirus infection, which is even higher if they have MM (7, 8), and they too can be effectively grouped (9, 10). Even so, these patients who take certain drugs commonly used in chronic diseases differ between each each other in the development and evolution of COVID-19 infection (3, 7, 11, 12, 13). Therefore, a study that groups multimorbid patients and/or with polypharmacy is necessary to study their trajectories in the future based on mixed data. A methodology based on machine learning is an interesting approach (14, 15) and we know that its use in the study of patients with COVID was effective (3, 7, 16, 17), but the conventional statistics approach also has advantages (14). We believe that the development of tools of both classes, based on a patient-centered analysis with chronic drugs, is necessary for a good management and improvement of the prognosis of patients with multimorbidity who have or have had certain diseases, such as COVID-19 (7, 2). Our objective is to set and describe multimorbidity and polypharmacy groups from a cohort of COVID-19 patients, and evaluate their relationship with the severity/mortality of the infection.

References

1. van den Houdt SCM, Slurink IAL, Mertens G. Long COVID is not a uniform syndrome: Evidence from person-level symptom clusters using latent class analysis. J Infect Public Health. 2023 Dec 29;17(2):321-328. doi: 10.1016/j.jiph.2023.12.019. Epub ahead of print. PMID: 38183882.

2. Han, L., Shen, P., Yan, J., Huang, Y., Ba, X., Lin, W., ... & Tu, S. (2021). Exploring the clinical characteristics of COVID-19 clusters identified using factor analysis of mixed data-based cluster analysis. *Frontiers in medicine*, *8*, 644724.

3. San-Cristobal, R., Martín-Hernández, R., Ramos-Lopez, O., Martinez-Urbistondo, D., Micó, V., Colmenarejo, G., ... & Martínez, J. A. (2022). Longwise cluster analysis for the prediction of COVID-19 severity within 72 h of admission: COVID-DATA-SAVE-LIFES cohort. *Journal of Clinical Medicine*, *11*(12), 3327.

4. Lip, G. Y., Genaidy, A., Tran, G., Marroquin, P., Estes, C., & Sloop, S. (2022). Effects of multimorbidity on incident COVID-19 events and its interplay with COVID-19 event status on subsequent incident myocardial infarction (MI). *European Journal of Clinical Investigation*, *52*(5), e13760.

5. Wong, K. C. Y., Xiang, Y., Yin, L., & So, H. C. (2021). Uncovering clinical risk factors and predicting severe COVID-19 cases using UK Biobank data: machine learning approach. *JMIR public health and surveillance*, *7*(9), e29544.

6. Gimeno-Miguel A, Bliek-Bueno K, Poblador-Plou B, Carmona-Pírez J, Poncel- Falcó A, González-Rubio F, Ioakeim-Skoufa I, Pico-Soler V, Aza-Pascual-Salcedo M, Prados-Torres A, Gimeno-Feliu LA; PRECOVID Group. Chronic diseases associated with increased likelihood of hospitalization and mortality in 68,913 COVID-19 confirmed cases in Spain: A population-based cohort study. PLoS One. 2021 No12;16(11):e0259822. doi: 10.1371/journal.pone.0259822. PMID: 34767594; PMCID: PMC8589220.

7. Pezoulas, V. C., Mylona, E., Papaloukas, C., Liontos, A., Biros, D. I., Milionis, O. I., ... & Fotiadis, D. I. (2022, September). A hybrid approach based on dynamic trajectories to predict mortality in COVID-19 patients upon steroids administration. In *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)* (pp. 01-04). IEEE.

8. Hayhoe, B. W., Powell, R. A., Barber, S., & Nicholls, D. (2022). Impact of COVID-19 on individuals with multimorbidity in primary care. *British Journal of General Practice*, *72*(714), 38-39.

9. Calderón-Larrañaga, A., Gimeno-Feliu, L. A., González-Rubio, F., Poblador-Plou, B., Lairla-San José, M., Abad-Díez, J. M., ... & Prados-Torres, A. (2013). Polypharmacy patterns: unravelling systematic associations between prescribed medications. *PLoS One*, *8*(12), e84967.

10. Patterns of multimorbidity and polypharmacy in young and adult population: Systematic associations among chronic diseases and drugs using factor analysis.

11. Zangrillo, A., Landoni, G., Monti, G., & Yavorovskiy, A. G. (2021). Dexamethasone in COVID-19: does one drug fits all?. *Medicina intensiva*, S0210-5691.

12. Shams, E., Kamalumpundi, V., Cheng, L., Taiwo, A., Shibli-Rahhal, A., Dokun, A. O., & Correia, M. L. (2023). Associação entre o Antagonismo do Sistema Renina-Angiotensina-Aldosterona e a Mortalidade Relacionada à COVID-19 em Pacientes com Hipertensão Relacionada ao Sobrepeso/Obesidade: um Estudo Retrospectivo de Coorte. *Arquivos Brasileiros de Cardiologia*, *120*, e20220277.

13. Caro-Codón, J., Rey, J. R., Iniesta, A. M., Rosillo, S. O., Castrejon-Castrejon, S., Rodriguez-Sotelo, L., ... & CARD-COVID Investigators. (2022). Impact of the withdrawal of renin-angiotensin-aldosterone inhibitors on mortality in COVID-19 patients. *Revista Portuguesa de Cardiologia*, *41*(10), 823-830.

14. Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Lee, M. J., & Asadi, H. (2018). eDoctor: machine learning and the future of medicine. *Journal of internal medicine*, *284*(6), 603-619.

15. Allam, A., Feuerriegel, S., Rebhan, M., & Krauthammer, M. (2021). Analyzing patient trajectories with artificial intelligence. *Journal of medical internet research*, *23*(12), e29812.

16. Garcia-Gutiérrez, S., Esteban-Aizpiri, C., Lafuente, I., Barrio, I., Quiros, R., Quintana, J. M., & Uranga, A. (2022). Machine learning-based model for prediction of clinical deterioration in hospitalized patients by COVID 19. *Scientific reports*, *12*(1), 7097.

17. Chimbunde, E., Sigwadhi, L. N., Tamuzi, J. L., Okango, E. L., Daramola, O., Ngah, V. D., & Nyasulu, P. S. (2023). Machine learning algorithms for predicting determinants of COVID-19 mortality in South Africa. *Frontiers in Artificial Intelligence*, *6*.

Evaluation of the msGBS methodology for the taxonomic identification and quantification of diatoms

Daniel Munzon Gil [1], Josep Piñol [2] and Miquel Àngel Senar [1]

[1] Departament d'Arquitectura de Computadors i Sistemes Operatius (DACSO). Universitat Autònoma de Barcelona (UAB).

[2] Centro de Investigación Ecológica y Aplicaciones Forestales (CREAF). Universitat Autònoma de Barcelona (UAB).

Abstract

Biodiversity monitoring using DNA gives certain advantages over traditional methods (de Souza et al., 2016; Hunter et al., 2018; Supple & Shapiro, 2018). Despite their many advantages, these methods however entail some limitations, like amplification errors and lack of standard regions for some biological groups. For those reasons, there is a need to provide new technologies to overcome the limitations of the most used method, metabarcoding (Taberlet et al., 2012). Multispecies Genotyping by Sequencing or msGBS, is a new proposal that uses restriction enzymes to obtain thousands of genetic regions species-specific (Wagekamer et al., 2022). With those regions, it should be possible to obtain the list of species in a heterogeneous sample and quantify the relative abundance of the species in the mixture. We compared different enzymes to improve this method. We calculated the number of fragments produced per 10^6 bp, using one and two enzymes. Also, in the case of using two enzymes, we calculated the percentage contribution of each of the enzymes. Finally, we calculated the number of informative reads according to the number of enzymes used. Our results showed that enzymes with a recognition site of 6 bp were the best option, because the number of fragments produced is high enough to avoid storage and taxonomic identification problems, using two enzymes did not differ statistically from using one enzyme and, regardless of the number of enzymes used, most of the reads were informative.

References

de Souza, C. P., Guedes, T. D. A., & Fontanetti, C. S. (2016). Evaluation of herbicides action on plant bioindicators by genetic biomarkers: a review. Environmental Monitoring and Assessment, 188, 1-12.

Fuentes-Pardo, A. P., & Ruzzante, D. E. (2017). Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. Molecular Ecology, 26(20), 5369-5406.

Hunter, M. E., Hoban, S. M., Bruford, M. W., Segelbacher, G., & Bernatchez, L. (2018). Next-generation conservation genetics and biodiversity monitoring. Evolutionary Applications, 11(7), 1029-1034.

Supple, M. A., & Shapiro, B. (2018). Conservation of biodiversity in the genomics era. Genome Biology, 19, 1-12.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular Ecology, 21(8), 2045-2050.

Wagemaker, C. A., Mommer, L., Visser, E. J., Weigelt, A., van Gurp, T. P., Postuma, M., Smit-Tiekstra, A., & de Kroon, H. (2021). msGBS: a new high-throughput approach to quantify the relative species abundance in root samples of multispecies plant communities. Molecular Ecology Resources, 21(4), 1021-1036.

Characterising the regulation B cell differentiation at a single cell resolution

Andrea Nieto-Aliseda Sutton[1], Biola M. Javierre[2]

[1] Josep Carreras Leukaemia Research Institute, Badalona, Barcelona, Spain

[2] Josep Carreras Leukaemia Research Institute and Institute for Health Science Research Germans Trias i Pujol, Badalona, Barcelona, Spain

Effective development of B lymphocytes plays a pivotal role in the hematopoietic system, given that they provide a humoral immune response against external pathogens. The intricacies of human B cell lymphopoiesis involve a multifaceted developmental process marked by specific surface protein profiles, with hematopoietic stem cells (HSCs) emerging in the foetal liver as early as six weeks post-conception (Popescu *et al.*, 2019). Postnatally, B cell differentiation initiates in the bone marrow, where HSCs give rise to multipotent progenitors that progressively differentiate into various cell types essential for the humoral immune response, including plasma cells and memory B cells. The orchestration of concise transcriptional control at each cellular transition, encompassing gene activation and silencing, proves critical for the accurate development of B lymphocytes. Conversely, the dysregulated establishment of cell and lineage-specific gene transcriptional programs during different stages of B cell lymphopoiesis precipitates the onset of B cell malignancies, encompassing conditions such as leukaemia, lymphoma, multiple myeloma, immunodeficiencies, and autoimmunity (Alizadeh *et al.*, 2000; Klein *et al.*, 2001; Staudt and Dave, 2005). This project aims to understand the epigenetic regulation of gene expression normal B cell lymphopoiesis, to then use as a baseline to decipher the exact mechanisms by which aberrant differentiation occurs. Employing a multi-omics approach, including cutting-edge single-cell technologies such as the recently developed scCUT&TAG-pro (Zhang *et al.*, 2022), our objective is to, for the first time, delineate the gene regulatory network and its evolution throughout B cell differentiation at a single-cell resolution. Through this innovative methodology, we aspire to define the dynamicity of gene regulatory networks throughout B cell differentiation together with its inherent heterogeneity within cell populations and unravel the unique molecular mechanisms steering aberrant developmental pathways. In doing so, we aim to advance our understanding of B cell lymphopoiesis and, in turn, facilitate targeted therapeutic strategies in personalised medicine.

Alizadeh,A.A. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Klein,U. *et al.* (2001) Gene expression profiling of B cell chronic lymphocytic leukemia reveals a homogeneous phenotype related to memory B cells. *J. Exp. Med.*, **194**, 1625–1638.

Popescu,D.-M. *et al.* (2019) Decoding human fetal liver haematopoiesis. *Nature*, **574**, 365–371.

Staudt,L.M. and Dave,S. (2005) The biology of human lymphoid malignancies revealed by gene expression profiling. *Adv. Immunol.*, **87**, 163–208.

Zhang,B. *et al.* (2022) Characterizing cellular heterogeneity in chromatin state with scCUT&Tag-pro. *Nat. Biotechnol.*, **40**, 1220–1230.

BigDataStatMeth: An R package to implement statistical methods for Big Data

Dolors Pelegrí-Sisó1, Juan R González1,2

1. Barcelona Research Institute for Global Health (ISGlobal), Doctor Aiguader 88, 08003 Barcelona, Spain.

2. CIBER in Epidemiology (CIBERESP), Av. Monforte de Lemos, 3-5. Pabellón 11. Planta 0 28029 Madrid, Spain
2Affiliation

Abstract

In recent years there has been a large increase in the amount and complexity of data available for analysis. These include data from different fields such as biomedical data, geographical information systems, and clinical, among many others. High-dimensional datasets analyses require scalable and computationally efficient algorithms, starting with algorithms to handle matrix and vector operations and perform basic algebra operations. In addition, there is a pressing need for methods to integrate more than two big tables and to be able to perform multivariate analyses. Multivariate analyses are used in all stages of data analysis, from feature extraction to pre-processing. For instance, principal Component Analysis (PCA) through Single Value Decomposition (SVD) is routinely used in genomics research to extract underlying genetic characteristics of the populations (1)(2) or to genotype structural variation such as inversions (3). In different contexts, SVD can be used in data reduction to remove undesired technical signals, such as to estimate surrogate variables that relate to the persistent batch effects in transcriptomic and epigenomic studies (4). In addition, SVA can be used to detect biological signals that may confound posterior analysis (5) (6). Most of these statistical methods can be naturally implemented in R language. However, problems arise in terms of memory allocation and computation efficiency when they are applied to big data. Most R packages are not designed to deal with big datasets, and hence, they are not computationally efficient. R is very well-suited for the development of new methods and statistical techniques but does not seamlessly handle massive data. BigDataStatMeth is an R package designed to perform algebraic operations and multivariate analysis on big datasets. BigDataStatMeth works directly with data stored in HDF5 data files simultaneously loading several but small partial blocks of data to perform calculations and thus avoid memory overflows and increase computation speed. All functions are written in C++ and integrated into R using the Rcpp and RcppEigen packages. Future development of methods for big data in R can easily incorporate the efficient algebraic operations implemented in BigDataStatMeth. Methods can be developed using the R language with the application programming interface (API) for R implemented in the BigDataStatMeth package or in C++ by using Rcpp and the API for C++ also implemented in the package to take full advantage of the benefits of using C++ in R.

References

1. Price, A., et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38, 904–909 (2006). https://doi.org/10.1038/ng1847

2. Price, A., et al. New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11, 459–463 (2010). https://doi.org/10.1038/nrg2813

3. Alejandro Cáceres, et al. Following the footprints of polymorphic inversions on SNP data: from detection to association tests, Nucleic Acids Research, Volume 43, Issue 8, 30 April 2015, Page e53, https://doi.org/10.1093/nar/gkv073.

4. Leek JT, et al. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012 Mar 15;28(6):882-3. doi: 10.1093/bioinformatics/bts034. Epub 2012 Jan 17. PMID: 22257669; PMCID: PMC3307112.

5. Houseman EA, et al. Reference-free cell mixture adjustments in analysis of DNA methylation data. Bioinformatics. 2014 May 15;30(10):1431-9. doi: 10.1093/bioinformatics/btu029. Epub 2014 Jan 21. PMID: 24451622; PMCID: PMC4016702.

6. Alquicira-Hernandez, J., et al. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. Genome Biol 20, 264 (2019). https://doi.org/10.1186/s13059-019-1862-5

Aquasearch: a new software for fast proteomic characterization and classification of wastewater samples analyzed using MALDI-TOF.

Carlos Pérez-López[1], Montserrat Carrascal, Antoni Ginebreda[1], Damiá Barcelò[1] and Joaquín Abian

1 Institute of Environmental Assessment and Water Studies (IDAEA-CSIC), Department of Environmental Chemistry, Barcelona, Spain

2 Institute of Biomedical Research of Barcelona, Spanish National Research Council (IIBB-CSIC/IDIBAPS), Barcelona, Spain

Abstract

The study of wastewater is a valuable source of information about the environment, health and industrial activities of the inhabitants of an area. Although the study of wastewater has traditionally focused on small molecules such as pharmaceuticals or illegal drugs, recent studies have reported the valuable information that can be obtained from large molecules in wastewater, introducing proteomics as an emerging field in environmental monitoring (Carrascal *et al.,* 2020; Perez-Lopez *et al.,* 2021; Carrascal *et al.,* 2023).

Liquid Chromatography coupled with High-Resolution Mass Spectrometry (LC-HRMS) instrument was used to identify the proteins in wastewater in the studies with a shotgun approach. Although the entire process reports comprehensive and accurate results, it is expensive and time-consuming. Therefore, Matrix-Assisted Laser Desorption/Ionization coupled with Time of Flight (MALDI-TOF) is proposed as a high-throughput instrumental approach for faster and more cost-effective sample characterization. In this work, we present Aquasearch, a newly developed software in Python for the characterization and classification of samples in a multisampling analysis. Aquasearch primarily performs two tasks: 1) signal filtering from wastewater proteomics samples analyzed by MALDI-TOF and identification of peptides belonging to livestock and human biomarkers using an in-house database and 2) using the identification results to classify the samples based on their proteomic profile in a non-supervised analysis. To facilitate the use of Aquasearch, including the parameter selection and result visualization, the program can be run through a graphic user interface (GUI).

To test the program, 4 wastewater samples collected from 4 WWTPs in Catalonia, Spain (Besòs, Girona, Vic and Figueres), were analyzed by MALDI-TOF. The Aquasearch analysis of the corresponding protein profiles showed the dominance of human biomarkers in Besòs and Girona, while pig and chicken biomarkers were the major components in Vic and Figueres. Finally, these proteomic profiles clustered the samples in the non-supervised multisampling analysis based on their origin.

References

Carrascal,M. *et al.* (2020) Discovery of large molecules as new biomarkers in wastewater using environmental proteomics and suitable polymer probes. *Science of the Total Environment*, **747**.

Carrascal,M. *et al.* (2023) Sewage Protein Information Mining: Discovery of Large Biomolecules as Biomarkers of Population and Industrial Activities. *Environ Sci Technol*, **57**, 10929–10939.

Perez-Lopez,C. *et al.* (2021) Non-target protein analysis of samples from wastewater treatment plants using the regions of interest-multivariate curve resolution (ROIMCR) chemometrics method. *J Environ Chem Eng*, **9**.

aSynPEP-DB: a database of biogenic peptides for inhibiting α-synuclein aggregation

Carlos Pintado-Grima[1], Oriol Bárcenas[1], Valentín Iglesias[1], Jaime Santos[1,2], Zoe Manglano-Artuñedo[1], Irantzu Pallarès[1], Michał Burdukiewicz[1,3] and Salvador Ventura[1]

[1]Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Bellaterra, Barcelona 08193, Spain

[2]Center for Molecular Biology of Heidelberg University (ZMBH), Heidelberg 69120, Germany

[3]Clinical Research Centre, Medical University of Białystok, Kilińskiego 1, Białystok 15-369, Poland

Abstract

Parkinson's disease (PD) is the second most prevalent neurodegenerative disorder, yet effective treatments able to stop or delay disease progression remain elusive. The aggregation of a presynaptic protein, α-synuclein (aSyn), is the primary neurological hallmark of PD and, thus, a promising target for therapeutic intervention. However, the lack of consensus on the molecular properties required to specifically bind the toxic species formed during aSyn aggregation has hindered the development of therapeutic molecules. Recently, we defined and experimentally validated a peptide architecture that demonstrated high affinity and selectivity in binding to aSyn toxic oligomers and fibrils, effectively preventing aSyn pathogenic aggregation (1). Human peptides with such properties may have neuroprotective activities and hold a huge therapeutic interest (2). Driven by this idea, in this work we develop a discriminative algorithm for the screening of human endogenous neuropeptides, antimicrobial peptides and diet-derived bioactive peptides with the potential to inhibit aSyn aggregation. We identify over 100 unique biogenic peptide candidates and ensembled a comprehensive database (aSynPEP-DB: https://asynpepdb.ppmclab.com/) that collects their physicochemical features, source datasets and additional therapeutic-relevant information, including their sites of expression and associated pathways (3). Besides, we provide access to the discriminative algorithm to extend its application to the screening of artificial peptides or new peptide datasets. aSynPEP-DB is a unique repository of peptides with the potential to modulate aSyn aggregation, serving as a platform for the identification of previously unexplored therapeutic agents.

References

1. Santos J., Gracia P., Navarro S. et al. (2021) alpha-Helical peptidic scaffolds to target alpha-synuclein toxic species with nanomolar affinity. Nat. Commun., 12, 3752.

2. Santos J., Pallares I. and Ventura S. (2022) Is a cure for Parkinson's disease hiding inside us? Trends Biochem. Sci., 47, 641–644.

3. Pintado-Grima C., Bárcenas O., Iglesias V. at. (2023). aSynPEP-DB: a database of biogenic peptides for inhibiting α-synuclein aggregation. Database, 2023, baad084.

Machine Learning approaches for the Characterization of COPD

Iria Pose-Lagoa[#1], Beatriz Urda-García[#2], Jose Carbonell-Caballero[#3], Alfonso Valencia[#*4]

[#]Life Sciences, Barcelona Supercomputing Center, Plaça d'Eusebi Güell, 1-3, 08034, Barcelona, Spain

[*]ICREA, Barcelona, Spain

[1]iria.poselagoa@bsc.es, [2]beatriz.urda@bsc.es, [3]jon.sanchez@bsc.es,
[3]jose.carbonell@bsc.es, [3]alfonso.valencia@bsc.es

***Keywords*— Chronic Obstructive Pulmonary Disease, Machine Learning, feature selection, gene expression**

ABSTRACT

Chronic Obstructive Pulmonary Disease (COPD) is a complex and heterogeneous disease, comprising a wide range of nonidentical patient profiles [1]. Its diagnosis is not straightforward - it is underdiagnosed, especially in women - usually appearing with severe airflow obstruction profiles, leading to a need for improved strategies to identify individuals who are at greater risk of developing COPD or who have early-stage [2].

Understanding the diversity of the disease is important for diagnosing and treating COPD, enabling the implementation of more individualized therapies. Here, we aim to enhance the binary patient classification of COPD using gene expression data from the Lung Tissue Research Consortium. To achieve this, we employ various feature selection criteria to identify the most relevant genes. These filtering approaches include knowledge extracted from intrinsic data characteristics (data-driven), external information from DisGeNET of genes associated with COPD (curated COPD-related genes), and their respective biological expansions based on physical interaction partners (OmniPath) [3] and network-based prioritization algorithms (GUILDify) [4]. Subsequently, we evaluate the performance of different classifiers: Random Forest, Support Vector Machines - polynomial and radial kernel, k-Nearest Neighbors, Generalized Linear Models, and XGBoost.

Our results show that the data-driven and curated COPD-related expansion gene selection approaches yield the highest cross-validation and independent test data performances, respectively. Our techniques demonstrate their ability to accurately classify COPD patients, outperforming previous studies [5-7] with accuracies up to 84,8%, and the selected genes represent relevant biomarkers for disease prediction.

References

1. Roca Torrent, J., Vargas, C., Cano Franco, I., Selivanov, V., Barreiro, E., Maier, D., ... & Gomez Cabrero, D. (2014). Chronic Obstructive Pulmonary Disease heterogeneity: challenges for health risk assessment, stratification and management. Journal of Translational Medicine, 2014, vol. 12, num. Suppl 2, p. s3.
2. Choi, J. Y., & Rhee, C. K. (2020). Diagnosis and treatment of early chronic obstructive lung disease (COPD). Journal of Clinical Medicine, 9(11), 3426.
3. Türei, D., Korcsmáros, T., & Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nature methods, 13(12), 966-967.
4. Aguirre-Plans, J., Piñero, J., Sanz, F., Furlong, L. I., Fernandez-Fuentes, N., Oliva, B., & Guney, E. (2019). GUILDify v2. 0: A tool to identify molecular networks underlying human diseases, their comorbidities and their druggable targets. Journal of molecular biology, 431(13), 2477-2484.
5. Mostafaei, S., Kazemnejad, A., Azimzadeh Jamalkandi, S., Amirhashchi, S., Donnelly, S. C., Armstrong, M. E., & Doroudian, M. (2018). Identification of novel genes in human airway epithelial cells associated with chronic obstructive pulmonary disease (COPD) using machine-based learning algorithms. Scientific reports, 8(1), 15775.
6. Yao, Y., Gu, Y., Yang, M., Cao, D., & Wu, F. (2019). The gene expression biomarkers for chronic obstructive pulmonary disease and interstitial lung disease. Frontiers in genetics, 10, 1154.
7. Mahmudah, K. R., Purnama, B., Indriani, F., & Satou, K. (2021, February). Machine Learning Algorithms for Predicting Chronic Obstructive Pulmonary Disease from Gene Expression Data with Class Imbalance. In BIOINFORMATICS (pp. 148-153).

Fly wing development *in silico*:

A computational investigation of morphological plasticity in *Drosophila* wings

Aleksa Ratarac[1], Carlos Mora Martinez[2], Isaac Salazar Ciudad[1,3]

1 Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, Spain

2 Institute of Biotechnology, University of Helsinki, Finland

3 Centre de Recerca Matemàtica, Barcelona, Spain

***interested in giving an oral presentation**

Although the wing of the fruit fly (*Drosophila melanogaster*) is one of the best understood animal organs in terms of development, crucial aspects such as the mechanics and the succession of its phases of isotropic growth, elongation and asymmetrical contraction remain largely elusive. Examining morphological plasticity – the phenomenon of environmentally induced variation in shape and size – may be key for progress in our understanding of this developmental system, as well as of the development of animal organs in general.

Our group constructed a computational model simulating *Drosophila* wing development, building upon the framework of 2D apical vertex models (Farhadifar *et al.*, 2007) and resting on the assumption that all morphogenetic movements can be explained by a finite set of behaviors and properties of cells and extracellular structures. An early iteration of the model successfully reproduced the main properties of the asymmetrical contraction phase as well as some distinct mutant phenotypes solely by tuning the system- or tissue-wide parameters (Ray *et al.*, 2015). Based on previous advancements using other models (Salazar-Ciudad and Jernvall, 2010), we aim to take this approach further by expanding and improving the model in attempt to reproduce much finer variation (e. g. populational variation) and simulate all experimentally observable phases of the pupal stage. In addition, we ran phenotypic plasticity experiments in order to provide the morphological data on populational variation in wing morphology resulting from different temperatures and population densities.

Our primary objectives encompass generating a representative wild-type morphology, reproducing the direction and extent of experimentally observed variation through controlled parameter perturbations, and interpreting these results within the context of environment-phenotype and environment-genotype-phenotype interactions. The results are expected to provide insight and suggest hypotheses on broader principles governing environmental influences on morphogenesis, opening the way for formulating and testing new mechanistic explanations on how particular environmental changes lead to different morphologies.

References

Farhadifar,R. *et al.* (2007) The Influence of Cell Mechanics, Cell-Cell Interactions, and Proliferation on Epithelial Packing. *Current Biology*, **17**, 2095–2104.

Ray,R.P. *et al.* (2015) Patterned Anchorage to the Apical Extracellular Matrix Defines Tissue Shape in the Developing Appendages of Drosophila. *Developmental Cell*, **34**, 310–322.

Salazar-Ciudad,I. and Jernvall,J. (2010) A computational model of teeth and the developmental origins of morphological variation. *Nature*, **464**, 583–586.

Extrapolation of pathogenicity between homologous variants

Mireia Olivella[1], Arnau Cordomí[2], Sergi Soldevila[1], Gabriel Ruiz[1]

1Bioinformatics and Bioimaging, UVic-UCC

2Bioinformatics, ESCI-UPF

Abstract

When we compare an individual's genome with the reference, several mutations are encountered. Most of these mutations are neutral, but some others can lead to pathogenic consequences. Given the rapid increase in the amount of generated sequencing data, there is an urgent need to accurately determine whether genetic variants detected in patients are disease causing or not. While numerous computational predictive tools exist, their ability to make accurate predictions is still limited. In this study, we focus on missense variants, those that modify the coding amino acid, and our aim is to determine if the pathogenicity of these variants can be extrapolated to homologous variants, i.e., variants affecting the same position in homologous proteins and exhibiting the same or similar amino acid change. With this purpose, we extracted homologous variants in a dataset composed of all reported disease-causing (ClinVar) and neutral (gnomAD) human missense variants from proteins with autosomal dominant (AD) inheritance. We collected 21,734 pairs of homologous variants from which 19,731 were disease-causing, 1,081 were neutral and 922 of them disagreed in pathogenicity annotation, achieving an error rate of 4.24%. Thus, our data supports that pathogenicity can be extrapolated, with a high accuracy and reliability, between homologous variants. This approach expands the number of variants for which pathogenicity can be annotated with a high precision.

References

Auton, A. *et al.* (2015) A global reference for human genetic variation. Nature, 526, 68–74.

Eilbeck, K. *et al.* (2017) Setting the score: variant prioritization and Mendelian disease. Nat Rev Genet, 18:599–612.

EURORDIS (2005) Rare Diseases: understanding this Public Health Priority

Chong, JX. *et al.* (2015) The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am J Hum Genet, 97(2):199–215.

Flanagan, SE. *et al.* (2010) Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. Genet Test Mol Biomarkers;14(4):533–7.

Computational design of ganciclovir-dependent kinases for suicide cancer gene therapy

Janet Sánchez[1], Miquel Estévez-Gay [1] and Sílvia Osuna [1,2].

[1] Institut de Química Computacional i Catàlisi (IQCC) and Departament de Química, Carrer Maria Aurèlia Capmany 69, 17003 Girona.

[2] ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain

Abstract

Computational enzyme design gives a wholly comprehension between mutations and their impact on enzyme activity. In many cases, enzymes that play an important role in biological processes present complications with their application in the industry. This is the case of one of the most promising biotherapy against cancer, the Herpes Simplex Virus- Thymidine Kinase (HSV-TK) phosphorylating the prodrug ganciclovir (GCV)[1,2]. HSV-TK has a multifunctional activity in the pyrimidine salvage pathway catalysing the γ-phosphate transfer from ATP in the presence of Mg^{2+} to thymidine (THM), the natural substrate[3]. Previous studies revealed that the activity of HSV-TK is substantially decreased (100-fold higher KM and 5-fold lower kcat) when the non-natural GCV is used, which hampers the HSV-TK/GCV application in suicide cancer gene therapy[4].

With the final aim of designing new improved variants, we evaluate the conformational dynamics of Wild-Type (WT) HSV-TK with the natural and non-natural GCV substrate and other experimental variants(SR39[4],Ala68Hie[5]). From the molecular dynamics simulations the Free Energy Landscapes (FELs) were reconstructed, which elucidated mainly two important conformations for the natural substrate and also Shortest Path Map (SPM)[6] was perform to identify those residues that contributes most on enzyme conformal dynamics.

Despite SPM identifies promising mutation points, sometimes it can be difficult to choose which residues can be mutated and which repercussions would occur on enzyme f.e loss of activity. To palliate this situation, SPM has been combined with Multiple Sequence Alignment (MSA) identifying at the same time those important residues and their conservation states allowing further selection criteria on rational enzyme design.

References

1. Barese,C.N. et al. (2012) Thymidine Kinase Suicide Gene-mediated Ganciclovir Ablation of Autologous Gene-modified Rhesus Hematopoiesis. Mol Ther, 20, 1932–1943.

2. Zeng,Z.-J. et al. (2014) The cell death and DNA damages caused by the Tet-On regulating HSV-tk/GCV suicide gene system in MCF-7 cells. Biomed Pharmacother, 68, 887–892.

3. Wurth,C. et al. (2001) The effect of substrate binding on the conformation and structural stability of Herpes simplex virus type 1 thymidine kinase. Protein Sci, 10, 63–73.

4. Kokoris,M.S. and Black,M.E. (2002) Characterization of Herpes Simplex Virus type 1 thymidine kinase mutants engineered for improved ganciclovir or acyclovir activity. *Protein Sci*, 11, 2267–2272.

5. Balzarini,J. et al. (2006) Engineering of a Single Conserved Amino Acid Residue of Herpes Simplex Virus Type 1 Thymidine Kinase Allows a Predominant Shift from Pyrimidine to Purine Nucleoside Phosphorylation. J Biol Chem, 281, 19273–19279.

6. Romero-Rivera,A. et al. (2017) Role of Conformational Dynamics in the Evolution of Retro-Aldolase Activity. ACS Catal., 7, 8524–8532.

Integrating Artificial Intelligence Methods in Pharmacokinetics & Pharmacodynamics Processes

Sergio Sánchez-Herrero[1], Laura Calvet[2], Angel A Juan[3]

[1]Department of Computer Science, Multimedia and Telecommunication, Universitat Oberta de Catalunya, 08018 Barcelona, Spain
[2]Telecommunications and Systems Engineering Department, Universitat Autònoma de Barcelona, Carrer Emprius, 2, 08202 Sabadell, Spain
[3]Research Center on Production Management and Engineering, Universitat Politècnica de València, Plaza Ferrandiz-Salvador, 03801 Alcoy, Spain

Abstract

Traditionally, determining pharmacokinetic (PK) and pharmacodynamic (PD) parameters for therapeutic drug monitoring (TDM) in humans has relied on in vitro and in vivo methods. These parameters, essential for understanding the correlation between drug exposure and response, are commonly explored through pharmacokinetic/pharmacodynamic (PKPD) population studies, employing metrics such as the area under the concentration–time curve (AUC). They play a crucial role in the pharmacological regulatory drug approval processes [1].

A notable recent trend involves the integration of machine learning (ML) methods into pharmacokinetics methodology has provided a robust approach to managing complex relationships within extensive datasets and analyzing high-dimensional data in clinical practice. The infusion of artificial intelligence (AI) into ML has further accelerated its application in drug-dose predictions, demonstrating remarkable computational efficiency and substantial potential in the realm of drug development [2]. Consequently, the goal of these works is to apply ML methodologies to illustrate how machine-learning methods could enhance PK/PD predictions in the PK/PD workflow (see Figure 1).

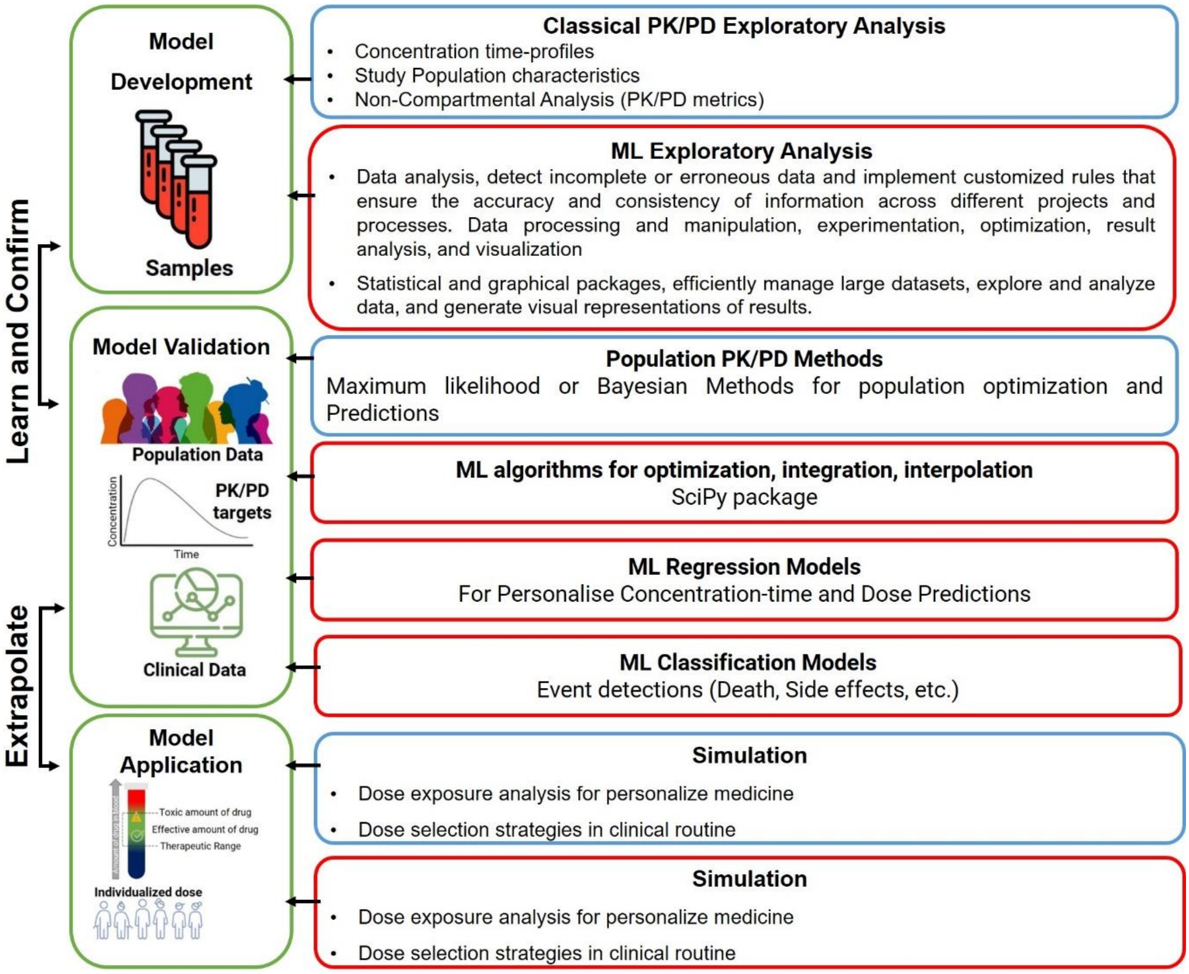
Firstly, exploratory analysis of PK/PD data is crucial, involving the examination of concentration-time profiles, study population characteristics, and Non-Compartmental Analysis. While various PK and PD software tools exist for generating diagnostic tables and plots, their built-in tools can often be inflexible and inefficient. External programming languages such as Python, Julia, R, or MATLAB offer a plethora of sophisticated and comprehensive packages for data analysis, scientific computing, application development, back-end web development, and machine learning. Our study underscores the potential of integrating open-source software, replete with an array of innovative packages, to elevate predictive capabilities and streamline analyses in PK methods. This integration ushers in new avenues for an advanced intelligent simulation modeling within the realm of PK, thus holding significant promise for the advancement of drug research and development [3].

Second, certain drugs, characterized by a narrow therapeutic index, significant toxicity, adverse effects, and interindividual variability, require frequent therapeutic drug monitoring and dose adjustments in renal transplant recipients. This study focuses on comparing machine learning (ML) models that utilize pharmacokinetic data to predict tacrolimus blood concentration. Various ML models were employed, and their performances were systematically compared. While all models demonstrated favorable fit outcomes, the ExtraTreesRegressor (ETR) stood out with superior performance. It achieved measures of -0.161 for MPE, 0.995 for AFE, 1.063 for AAFE, and 0.8 for R2, indicating accurate predictions that meet regulatory standards. These findings underscore the predictive potential of ML [4].

Third, the common PK/PD methods and functions employed for parameter estimation and final model validation often include the sequential quadratic programming (SQP) method and a genetic algorithm linked to First-order conditional estimation (FOCE-i) methods. The scipy.optimize.minimize function in Python is utilized for the analysis of optimization methods. SciPy optimize offers a range of functions for minimizing objective functions, encompassing solvers for nonlinear problems, linear programming, constrained and nonlinear least-squares, root finding, and curve fitting. These optimization methods were applied to estimate clearance (CL) and volume of distribution (Vc) in a one-compartment PK model using real patient data derived from plasma concentrations (in μg/mL) of Cefepime administered intravenously. Among the optimization techniques, COBYLA and Nelder-Mead exhibited superior results [5].

As both PK/PD and ML encounter various challenges, there is a rising interest within their respective communities to explore ways to integrate expertise from these two fields [6]. This highlights the significance of fostering collaboration between these disciplines, driven not only by time constraints but also by the necessity to collectively tackle PK/PD challenges.

For these reasons, open research lines will be focus on looking for new ML approaches in pharmacokinetics area combining PK models and ML methods, or applying meta-models for PK regression or classification problems or improving any PK analysis that could be helpful for regulatory drug approval processes.



Classical PK/PD Exploratory Analysis

- Concentration time-profiles
- Study Population characteristics
- Non-Compartmental Analysis (PK/PD metrics)

ML Exploratory Analysis

- Data analysis, detect incomplete or erroneous data and implement customized rules that ensure the accuracy and consistency of information across different projects and processes. Data processing and manipulation, experimentation, optimization, result analysis, and visualization
- Statistical and graphical packages, efficiently manage large datasets, explore and analyze data, and generate visual representations of results.

Population PK/PD Methods

Maximum likelihood or Bayesian Methods for population optimization and Predictions

ML algorithms for optimization, integration, interpolation

SciPy package

ML Regression Models

For Personalise Concentration-time and Dose Predictions

ML Classification Models

Event detections (Death, Side effects, etc.)

Simulation

- Dose exposure analysis for personalize medicine
- Dose selection strategies in clinical routine

Simulation

- Dose exposure analysis for personalize medicine
- Dose selection strategies in clinical routine

Figure 1. Pharmacokinetics and pharmacodynamics process workflow.

References

1. Rajman I. (2008) PK/PD modelling and simulations: utility in drug development. Drug Discov Today 2008; **13**(7-8):341–46.

2. Vora, L. K., *et* al., (2023). Artificial intelligence in pharmaceutical technology and drug delivery design. Pharmaceutics, **15**(7), 1916.

3. Sánchez-Herrero, S. (2023) Embedding R inside the PhysPK Bio-simulation Software for Pharmacokinetics Population Analysis. BIO Integration, **4**(3), 97-113.

4. Sánchez-Herrero, S., Calvet L., Juan A. A. (2023). Machine Learning Models for Predicting Personalized Tacrolimus Stable Dosages in Pediatric Renal Transplant Patients. BioMedInformatics, **3**(4), 926-947.

5. Sánchez-Herrero, S., *et al.,* (2023). Integrating Python optimization algorithms inside PhysPK® (PK/PD/PBPK) software for improving PK estimation methods. PAGE 31 Conference. Abstr 10534.

6. McComb M., Bies R., M Ramanathan M. (2021). Machine learning in pharmacometrics: Opportunities and challenges. Br. J. Clin. Pharmacol. **88**:1482–1499.

CADSETshield: Developing a Secure and Efficient Platform for Integrated Medical Imaging and Genomic Studies of COPD Using DataSHIELD and OMOP CDM

David Sarrat González[1], Juan R González[1,2]

1 Barcelona Institute for Global Health (ISGlobal)

2 Universitat Autònoma de Barcelona (UAB)

Abstract

Chronic Obstructive Pulmonary Disease (COPD) is a major health issue worldwide that requires advanced research approaches[1]. The integration of detailed medical imaging and genomic data, or radiogenomics, is essential for better understanding and managing COPD[2]. However, combining such vast and sensitive data sets raises significant challenges in data standardization, harmonization, and privacy, especially under strict regulations like the General Data Protection Regulation (GDPR)[3].

To address these issues, we have developed CADSETshield, a platform that effectively integrates DataSHIELD and the OMOP Common Data Model (OMOP CDM) for COPD research. While OMOP CDM helps in standardizing varied clinical data, DataSHIELD ensures the privacy and security of this data. The key to blending these systems is our newly developed tool, 'dsOMOP'.

dsOMOP is an R package that is specifically designed to simplify the interaction between OMOP CDM and DataSHIELD. This makes it easier for researchers to manage and analyze large amounts of harmonized data. The introduction of dsOMOP is a significant step forward in clinical research, allowing for large-scale federated data analysis.

By enabling researchers to perform in-depth analyses without compromising data privacy, we are opening new paths for discovery in COPD research. CADSETshield and dsOMOP together offer a practical solution for handling complex datasets in clinical studies, paving the way for significant advancements in the field.

References

1. Celli, B. R., & Agustí, A. (2018). 'COPD: Time to improve its taxonomy?', ERJ Open Research, 4(1). doi: 10.1183/23120541.00132-2017

2. Agustí, A., Celli, B. & Faner, R. (2017) 'What does endotyping mean for treatment in chronic obstructive pulmonary disease?', Lancet, 390(10098), pp. 980–987. doi: 10.1016/S0140-6736(17)32136-0

3. F, S. et al. (2018) To share or not to share? Expected pros and cons of data sharing in radiological research. Eur. Radiol., 28, 2328–2335.

Multi-omics microbiome dynamics in IBD

Gerard Serrano Gómez[1,3], Chaysavanh Manichanh[1,2]

1. Gut Microbiome Group, Vall d'Hebron Institut de Recerca (VHIR), Vall d'Hebron Hospital Universitari, Vall d'Hebron Barcelona Hospital Campus, Passeig Vall d'Hebron 119-129, 08035 Barcelona, Spain
2. Medicine Department, Autonomous University of Barcelona (UAB), 08193 Cerdanyola del Vallès, Spain
3. Institute of Biotechnology and Biomedicine, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, 08193, Spain.

Abstract

Inflammatory Bowel Disease (IBD) is the term used to describe two of the most common chronic inflammatory diseases of the gastrointestinal (GI) tract: Crohn's Disease (CD) and Ulcerative Colitis (UC). Both IBD subtypes are characterised by the alternation of periods of clinical remission and relapse.

Although the factors that trigger CD and UC development are still unknown, several studies demonstrated that gut microbiota alterations are associated with IBD, with an increase of Proteobacteria and depletion of Firmicutes in CD-affected individuals (Halfvarson *et al.*, 2017; Baumgart *et al.*, 2007; Manichanh *et al.*, 2006) and a decrease of butyrate-producing bacteria in UC (Kumari *et al.*, 2013; Machiels *et al.*, 2014). However, environmental factors such as diet, smoking habits, antibiotic usage, stress or sleeping schedule have also been linked to the development of IBD (Gomaa, 2020). It is believed that the increase of pathogenic bacteria in the GI tract alters gut permeability, causing a disbalance in the microbial community known as dysbiosis, and an alteration of the metabolite composition in the GI tract (Strugala *et al.*, 2008; Schmitz *et al.*, 1999; Nishida *et al.*, 2018), which ultimately leads to gut inflammation.

To address this knowledge gap, we analyzed a total of 421 unique measurements generated by metagenomics, metatranscriptomics or metabolomics techniques from 67 IBD patients and 67 healthy controls. Our results revealed novel microbial signature species in CD, as well as shifts on the expression of key pathways in the gut microbial ecosystem and the alteration of the concentration of several metabolites in the GI tract. Finally, integrative analysis gave insight into the interaction of these factors contributing to dysbiosis in CD.

References

Baumgart,M. *et al.* (2007) Culture independent analysis of ileal mucosa reveals a selective increase in invasive Escherichia coli of novel phylogeny relative to depletion of Clostridiales in Crohn's disease involving the ileum. *ISME J.*, **1**, 403–418.

Gomaa,E.Z. (2020) Human gut microbiota/microbiome in health and diseases: a review. *Antonie Van Leeuwenhoek*, **113**, 2019–2040.

Halfvarson,J. *et al.* (2017) Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat. Microbiol.*, **2**, 17004.

Kumari,R. *et al.* (2013) Fluctuations in butyrate-producing bacteria in ulcerative colitis patients of North India. *World J. Gastroenterol.*, **19**, 3404–3414.

Machiels,K. *et al.* (2014) A decrease of the butyrate-producing species Roseburia hominis and

Faecalibacterium prausnitzii defines dysbiosis in patients with ulcerative colitis. *Gut*, **63**, 1275–1283.

Manichanh,C. *et al.* (2006) Reduced diversity of faecal microbiota in Crohn’s disease revealed by a metagenomic approach. *Gut*, **55**, 205–211.

Nishida,A. *et al.* (2018) Gut microbiota in the pathogenesis of inflammatory bowel disease. *Clin. J. Gastroenterol.*, **11**, 1–10.

Schmitz,H. *et al.* (1999) Altered tight junction structure contributes to the impaired epithelial barrier function in ulcerative colitis. *Gastroenterology*, **116**, 301–309.

Strugala,V. *et al.* (2008) Thickness and continuity of the adherent colonic mucus barrier in active and quiescent ulcerative colitis and Crohn’s disease. *Int. J. Clin. Pract.*, **62**, 762–769.

Development and application of tools for automated integration and analysis of big data in forestry management

TEJADA-GUTIERREZ, E. L.[1], SOLSONA TEHAS, F.[2], MATEO FORNÉS, J.[2] , ALVES, R.[1]

1.Dept. Ciències Mèdiques Bàsiques, Universitat de Lleida

2.Dept. d’Enginyeria Informàtica, Universitat de Lleida.

Abstract

Recently there has been an increase in both the number of datasets and the amount of forest-related data from around the world. Because of the amount of data, its fragmentation and lack of uniformity, it is very hard to develop tools for finding patterns that can contribute to decision-making about land management, sustainability, and mitigate the effects of climate changes. As such, it is useful to have access to this data in an integrated manner for analysis, in order to develop tools for understanding and mitigating the impact of climate change on biodiversity.

This work integrates more than 3500 available forest data sets into a unique curated Mongo database of uniform format and quality, packaging it into a webtool we name ForestForward. Subsequently, the data have been preprocesed and grouped by regions whose quality and quantity allow different biodiversity indices to be calculated, as all datasets are integrated with uniform format and quality.

This open access web platform contains information about geography, species, number of individuals and year of the observations. By analyzing the data to calculate biodiversity over different geographic regions and year we can see how that biodiversity evolved over the time, and contribute to understand the impacts of climate change on ecosystems. This analysis shows that data for the European region is of highest quality for a longer period of time in all processed datasets. That data reveals that the richness and abundance of species changed differently in different parts of the continent. However, changes were always subtle.

References

1. Bellard, C., Bertelsmeier, C., Leadley, P., Thuiller, W., & Courchamp, F. (2012). Impacts of climate change on the future of biodiversity. In Ecology Letters (Vol. 15, Issue 4, pp. 365–377).

2. Kent, M. (2005). Biogeography and macroecology. Progress in Physical Geography, 29(2), 256–264.

3. Serra-Diaz, J. M., Enquist, B. J., Maitner, B., Merow, C., & Svenning, J. C. (2017). Big data of tree species distributions: how big and how good? Forest Ecosystems, 4(1).

Reanalysis of next generation sequencing data from patients with cardiac diseases

Eudald Tejero[1,2], Carles Moliner[5], Clara Serra-Juhé[3,4], Marta de Antonio[5], Ivon Cuscó[3,4], Marta Campreciós[5], Antonio Barros[5], José M. Guerra[5], Sònia Mirabet[5], Jordi Surrallés[1-4], Benjamín Rodríguez-Santiago[3,4]

1. Genome Instability and DNA Repair Syndromes Group. IR-Sant Pau - Sant Pau Institute of Biomedical Research (IIB)

2. Joint Unit in Genomic Medicine UAB-IR Sant Pau

3. Center for Biomedical Network Research on Rare Diseases (CIBERER)

4. Genetics Department, Hospital de la Santa Creu i Sant Pau, Barcelona

5. Cardiology Service, Hospital de la Santa Creu i Sant Pau, Barcelona

Abstract

The utilization of exome sequencing technology in genetics diagnosis of rare diseases patients is becoming routinary in the majority of laboratories (Vinkšel *et al.*, 2021). Depending on patients' clinical phenotypes the percentage of successful diagnosis is between 25 and 58% (Fung *et al.*, 2020). Recent literature recommends reanalysing negative cases after a period of time to reach a genetic diagnosis (Dai *et al.*, 2022), taking advantage of variant databases updates and improved pipelines. Here we conducted a Next Generation Sequencing (NGS) data reanalysis of 456 patients from Sant Pau Hospital (Barcelona) with different cardiac diseases, including dilated, hypertrophic and arrhythmogenic cardiopathies, and aortic pathology. Patients were initially negative for NGS genetic testing performed during 2018-2022 period years by the Genetics Department using the NGS pipeline and gene lists corresponding to that time and patient's indication. Two different kinds of reanalyses were carried out: 1) Deep intronic variants analysis using specific tools, and 2) SNVs, indels and CNVs identification in a comprehensive cardiologic disease gene panel from the existing clinical exome data. Regarding the first group, data in VCF format was scanned by the prediction tool SpliceAI (Jaganathan *et al.*, 2019) to obtain a score per variant, then the variants were filtered according to the program recommendations and by allelic frequency. Two pathogenic variants associated to phenotype were identified in two patients in canonical splice sites. No variant of interest was detected in deep intronic regions. Regarding the second analysis, a large number of variants were filtered taking into account whether they were classified as pathogenic by variant databases and/or their population frequency. We found 24 candidate variants that deserved a follow-up with cardiologists that is still ongoing.

References

Dai,P. *et al.* (2022) Recommendations for next generation sequencing data reanalysis of unsolved cases with suspected Mendelian disorders: A systematic review and meta-analysis. *Genetics in Medicine,* 24, 1618–1629.

Fung,J.L.F. *et al.* (2020) A three-year follow-up study evaluating clinical utility of exome sequencing and diagnostic potential of reanalysis. *Genom. Med.*, 5, 37.

Jaganathan,K. *et al.* (2019) Predicting Splicing from Primary Sequence with Deep Learning. *Cell*, 176, 535-548.e24.

Vinkšel,M. *et al.* (2021) Improving diagnostics of rare genetic diseases with NGS approaches. *J Community Genet*, 12, 247–256.

Identification of differential expressed genes between abdominal aortic aneurysm cases and controls in aortic tissue

Gerard Temprano-Sagrera MSc,[1] Ana Viñuela PhD#,[2] Mercedes Camacho PhD#,[1] Maria Sabater-Lleal PhD#[1,3]

1Unit of genomics of Complex Disease, Institut de Recerca Sant Pau (IR SANT PAU), Sant Quintí 77-79, 08041 Barcelona, Spain.
2Biosciences Institute, Faculty of Medical Sciences, Newcastle University, UK.
3Cardiovascular Medicine Unit, Department of Medicine, Karolinska Institutet, Stockholm, Sweden
#These authors equally contributed as senior authors.

Abstract

Abdominal aortic aneurysm (AAA) is a cardiovascular disease that is clinically significant due to its asymptomatic nature and potential for high mortality rates upon rupture, which can reach 85% (Golledge, 2019; Sakalihasan *et al.*, 2018). To uncover new underlying molecular mechanisms, we conducted comprehensive transcriptomic analyses comparing 96 AAA tissue samples with 44 aortic samples from deceased organ donors. The first analysis identified 7,454 differentially expressed genes (DEGs) between cases and controls (FDR < 0.05). We aimed to account for the effect of ischemic time (IT) on control samples, using GTEx data,(The GTEx Consortium atlas of genetic regulatory effects across human tissues, 2020), obtaining a list DEGs by IT, to refine the DEGs between cases and controls, while acknowledging that their role on AAA cannot be completely excluded. Cluster analysis, based on enriched gene ontology terms, revealed four distinct clusters strongly associated with AAA (see Figure 1). Moreover, exploring AAA of different diameters identified 32 DEGs, 8 of which overlapped with the AAA-associated DEGs, suggesting their involvement not only in disease onset but also in its progression (see Figure 2). Our study on alternative splicing identified 11 genes with differential patterns between cases and controls (FDR < 0.05), 7 of which were also AAA-associated DEGs. For example, *SPP1* (osteopontin), an important inflammation regulator, showed a higher frequency of exon skipping events in cases than in controls, which could have contributed to its increased expression in cases. We also studied allelic specific expression (ASE) in 12 genotyped AAA individuals and compared it to GTEx control individuals. We identified 90 genes that showed differential ASE in 5 or more of our samples (FDR < 0.05). For instance, *SNURF* gene, which also exhibited divergent ASE patterns between cases and controls, shedding light on potential mechanisms driving differential expression. Overall, our differential expression study identified 3,568 new DEGs between AAA cases and controls compared to the largest previous microarray study. (Lindquist Liljeqvist *et al.*, 2020) In addition, the clusters obtained considering the IT effect of calcium regulation and ATP synthesis had not been identified in this type of study, although they had been studied in relation to AAA. Finally, the study of the effect of splicing and ASE, allows us to deepen in the causes of the altered metabolic pathways in AAA.

References

Golledge,J. (2019) Abdominal aortic aneurysm: update on pathogenesis and medical treatments. *Nat Rev Cardiol*, **16**, 225–242.

Lindquist Liljeqvist,M. *et al.* (2020) Tunica-Specific Transcriptome of Abdominal Aortic Aneurysm and the Effect of Intraluminal Thrombus, Smoking, and Diameter Growth Rate. *Arteriosclerosis, Thrombosis, and Vascular Biology*, **40**, 2700–2713.

Sakalihasan,N. *et al.* (2018) Abdominal aortic aneurysms. *Nat Rev Dis Primers*, **4**, 1–22.

The GTEx Consortium atlas of genetic regulatory effects across human tissues (2020) *Science*, **369**, 1318–1330.

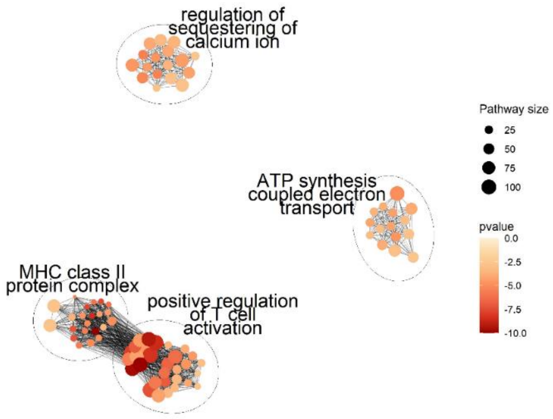


Figure 1: Hierarchical clustering analysis results after removing differentially expressed genes by ischemic time.

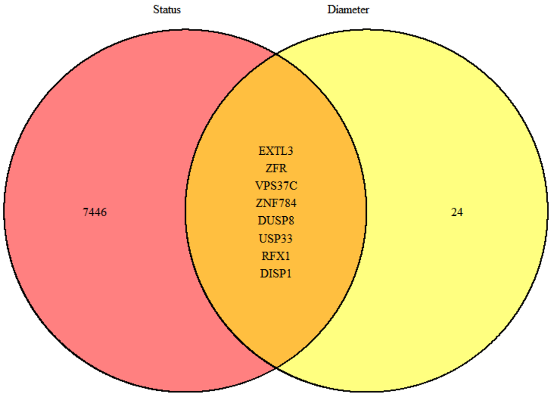


Figure 2: Venn diagram showing the overlap between differentially expressed genes in cases and controls and differentially expressed genes by AAA diameter.

Discovering and tracking potential zoonotic species from metagenomic samples with a capture-based oriented pipeline

Tarradas-Alemany, Maria[1,2]

PhD programe in bioinformatics
PhD Workshop
February 2, 2024

1. Computational Genomics Lab., Department of Genetics, Microbiology and Statistics, Universitat de Barcelona (UB); Institut de Biomedicina UB (IBUB).

2. Laboratory of Viruses Contaminants of Water and Food, Department of Genetics, Microbiology and Statistics, Universitat de Barcelona (UB).

From the dawn of Next Generation Sequencing(NGS) technologies, those strategies have become crucial in the study of microbial communities from environmental samples. However, there are still some challenges to overcome, either from biological and computational perspectives, to characterize their virome composition. Viral metagenomics has to deal with low quality sequences, possible sample biases (due to chemical inhibitors, degradation, etc), challenging data analysis, and more specifically the lack of standardized regions for classification, the arduous purification of enough biomass for sequencing, and the limited completeness of the available virus databases. In addition, most of the viral particles found in environmental samples correspond to bacteriophages, which further complicates the detection of specific viral families and species[3].

The proposed aproach to overcome some of those issues focuses on the use of capture probes specifically designed to hybridate a set of species of interest, with the aim to enrich the sample with their genomic sequences and similar ones[1]. For a specialized bioinformatic analysis of these datasets we introduce `CAPTVRED` (*Capture-based metagenomics Analysis Pipeline for tracking ViRal species from Environmental Datasets*), a NextFlow[2] automated pipeline purposely designed to provide comprehensive results of capture-based metagenomics datasets. The pipeline includes a pre-filtering stage to discard non-viral sequences, taking advantage of a curated viral database, which also excludes phage viral sequences, as reference. Unlike other available protocols, `CAPTVRED` offers the flexibility to adjust almost any parameter at each step, making it adaptable to the unique characteristics of viral metagenomic datasets.

The virome present in a set of samples retrieved from sewage and bat guano have been already analyzed with this pipeline; moreover, sequences obtained by whole-genome shotgun and probe-based viral capture approaches have been also considered, in order to assess the performance of the capture kit, as well as for the pipeline. The results show an increased number of assigned viral contigs in the capture approach (using RVDB database), which also recalls higher coverage and similarity with respect to reference sequences of potentially zoonotic viruses.

References

[1] Briese, T. *et al.* (2015). Virome capture sequencing enables sensitive viral diagnosis and comprehensive virome analysis. *MBio*, **6**(5), e01491–15.

[2] Di Tommaso, P. *et al.* (2017). `Nextflow` enables reproducible computational workflows. *Nature Biotechnology*, **35**(4), 316–319.

[3] Garner, E. *et al.* (2021). Next generation sequencing approaches to evaluate water and wastewater quality. *Water Research*, **194**, 116907.

Identification of New BCL-2 Inhibiting Small Molecules using Machine Learning, Molecular Docking, and MD Simulation.

Abtin Tondar[1], Laura Calvet Liñán[2], David Hervas Marin[3]

1-Department of Computer Science, Multimedia and Telecommunication, Universitat Oberta de Catalunya (UOC), Spain

2-Telecommunications and Systems Engineering Department, Universitat Autònoma de Barcelona (UAB), Spain

3-Department of Applied Statistics and Operational Research, and QualityAlcoy, Universitat Politècnica de València (UPV), Spain

This study presents an innovative approach to identifying small molecule therapeutics targeting B-cell lymphoma 2 (BCL-2), a critical protein in cancer pathogenesis. Leveraging the power of machine learning, molecular docking, and molecular dynamics (MD) simulations, we developed a practical framework for the virtual screening of compounds with potential BCL-2 inhibitory activity. Our methodology combines the predictive accuracy of deep neural networks (DNN) and the robustness of Random Forest (RF) algorithms, integrated with molecular docking techniques, to identify promising candidates from a vast chemical space. Through rigorous MD simulations, we validated the stability and binding affinity of top-performing molecules, highlighting their therapeutic potential in cancer. The study not only exemplifies the synergy of computational techniques in drug discovery but also marks a significant step forward in the search for effective BCL-2 inhibitors, offering promising avenues for cancer therapy development.

Keywords: Computational Biology, Machine Learning, Virtual Screening, High-Throughput Screening, Molecular Docking, MD Simulation, Cancer Drug Discovery, Process Optimization

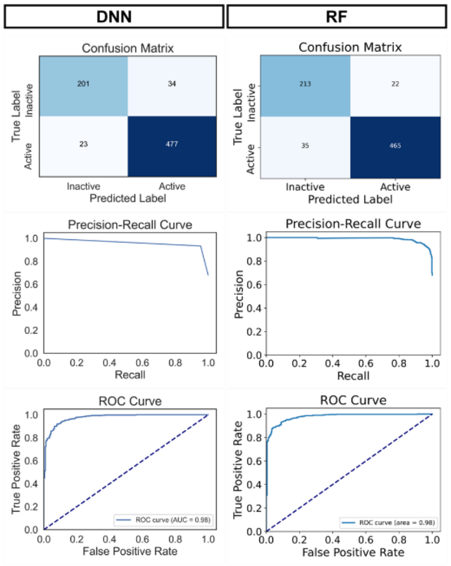


Fig1- Classification reports of the two machine learning models.

References:

Kawiak,A. and Kostecka,A. (2022) Regulation of bcl-2 family proteins in estrogen receptor-positive breast cancer and their implications in endocrine therapy. *Cancers*, **14**, 279.

Kim,S. *et al.* (2022) PubChem 2023 update. *Nucleic Acids Research*, **51**.

Characterization of strategies for structural variant imputation

Illya Yakymenko[1,2], Ruth Gómez-Graciani[2], Mario Cáceres[1,2,3]

1 Research Program on Biomedical Informatics (GRIB), Hospital del Mar Research Institute (IMIM), Barcelona, Spain

2 Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain

3 ICREA, Barcelona, Spain

Abstract

Structural variants (SVs) are genomic rearrangements of large segments of DNA. Despite being relatively common, they tend to be omitted or poorly characterized due to the complexity of variant calling. Focused on inversions, inverted duplications and inversion-associated deletions, we explored different strategies to optimize genotype prediction in genomic datasets. Inversions originated by non-homologous (NH) mechanisms and inverted duplications can be genotyped from short read data with approaches as BreakSeq [1]. Inversions originated by non allelic homologous recombination (NAHR) requires PCR-based techniques for genotyping. Using a well characterized reference panel, we tested BreakSeq genotyping in GEUVADIS [2] and imputation in GEUVADIS and GTEx [3] v8 datasets for 198 of the aforementioned SVs. BreakSeq genotyped 99.37% and 99.75% of the samples on average for African and European populations, respectively. Imputation was performed by inference with variants in perfect linkage disequilibrium (tagging variants), IMPUTE2 [4] and scoreInvHap [5]. In total, 143 and 104 of the SVs were resolved by BreakSeq and imputation with tagging variants in GEUVADIS and GTEx, respectively. On the other hand, a combination of imputation and genotype probability filtering has exhibited a high-quality imputation for 72.73% and 87.67% of the remaining SVs in GEUVADIS and GTEx, respectively. However, strategies for a reliable imputation of a significant fraction of the analyzed SVs with low linkage disequilibrium with neighboring variants are still missing.

References

1. Lam, H. *et al.* (2010) Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat Biotechnol* 28, 47–55.

2. Lappalainen, T. *et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506–511.

3. The GTEx Consortium, (2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369,1318-1330.

4. Howie, B. *et al.* (2011). Genotype imputation with thousands of genomes. G3 (Bethesda, Md.), 1(6), 457–470.

5. Ruiz-Arenas, C.*et al.* (2019). scoreInvHap: Inversion genotyping for genome-wide association studies. *PLoS genetics*, *15*(7), e1008203.

A Quantitative View of the Heterogeneity-Diversity Axis in Biological Systems

Jing Yang[1], Jordi Villà-Freixa[1,2], Adrián López García de Lomana[1,2]

1 Research Group on Bioinformatics and Bioimaging (BI^2); Facultat de Ciències, Tecnologia i Enginyeries; Universitat de Vic −- Universitat Central de Catalunya (UVic-UCC), Vic, Barcelona

2 Institut de Recerca i Innovació en Ciències de la Vida i de la Salut a la Catalunya Central (IRIS-CC)

Abstract

Glioblastoma (GBM) remains the most fatal type of brain tumor in adults, with overall survival among the worst in the spectrum of cancers. Its poor prognosis roots in the very nature of its high inter- and intratumoral heterogeneity driven by genetic and epigenetic alterations, which lead to therapeutic resistance and tumor relapse. The landscape of glioblastoma characterization has markedly advanced since the introduction of single-cell RNA sequencing, resulting in the classification of GBM cells into three identifiable subtypes. These subtypes span from stem-like to more differentiated states, offering insights that mirror features of neurodevelopmental hierarchies. However, there has been limited focus on quantifying intertumoral heterogeneity hindering a comprehensive understanding of compositional and functional differences among patients. In addressing this gap, we propose the introduction of entropy measures to capture the degree of disorder or randomness in the highly variable gene expression shared by each primary and recurrent GBM ecosystem. We delve into the importance and intricacy of inter- and intratumoral heterogeneity in the context of GBM drug-induced evolution.

The PhD thesis project has been recently modified to include research to be done in collaboration with the Aquatic Ecology group at the UVic-UCC, in which the exploration of the heterogeneity of biological systems will span pollutants influenced biodiversity heterogeneity in European pondscapes influenced by climate change gradients.

All the data that we will use will be public data repositories (in particular, at the Repositori de Dades de Recerca, CSUC) and all the code used will be uploaded to a github repository.

UVIC
UNIVERSITAT DE VIC
UNIVERSITAT CENTRAL
DE CATALUNYA

Bi^2
Bi-squared Research Group
Bioinformatics and Bioimaging
UVIC·UCC