JOBIM 2021

06 > 09 JUIL

INSTITUT PASTEUR | PARIS

Proceedings

> Posters

HTTPS://JOBIM2021.SCIENCESCONF.ORG @JOBIM_2021

ORGANISÉ PAR

sfbi société française de bioinformatique

INSTITUT PASTEUR

cea

GENOSCOPE

PARTENAIRES

FRANCE GENOMIQUE

FRANCE MÉDECINE GÉNOMIQUE 2025 aviesan

GDR GÉNOMIQUE ENVIRONNEMENTALE

cnrs GDR Groupement de recherche BIM Bioinformatique moléculaire

GENOPOLE

Google Research

ifb INSTITUT FRANÇAIS DE BIOINFORMATIQUE

illumina

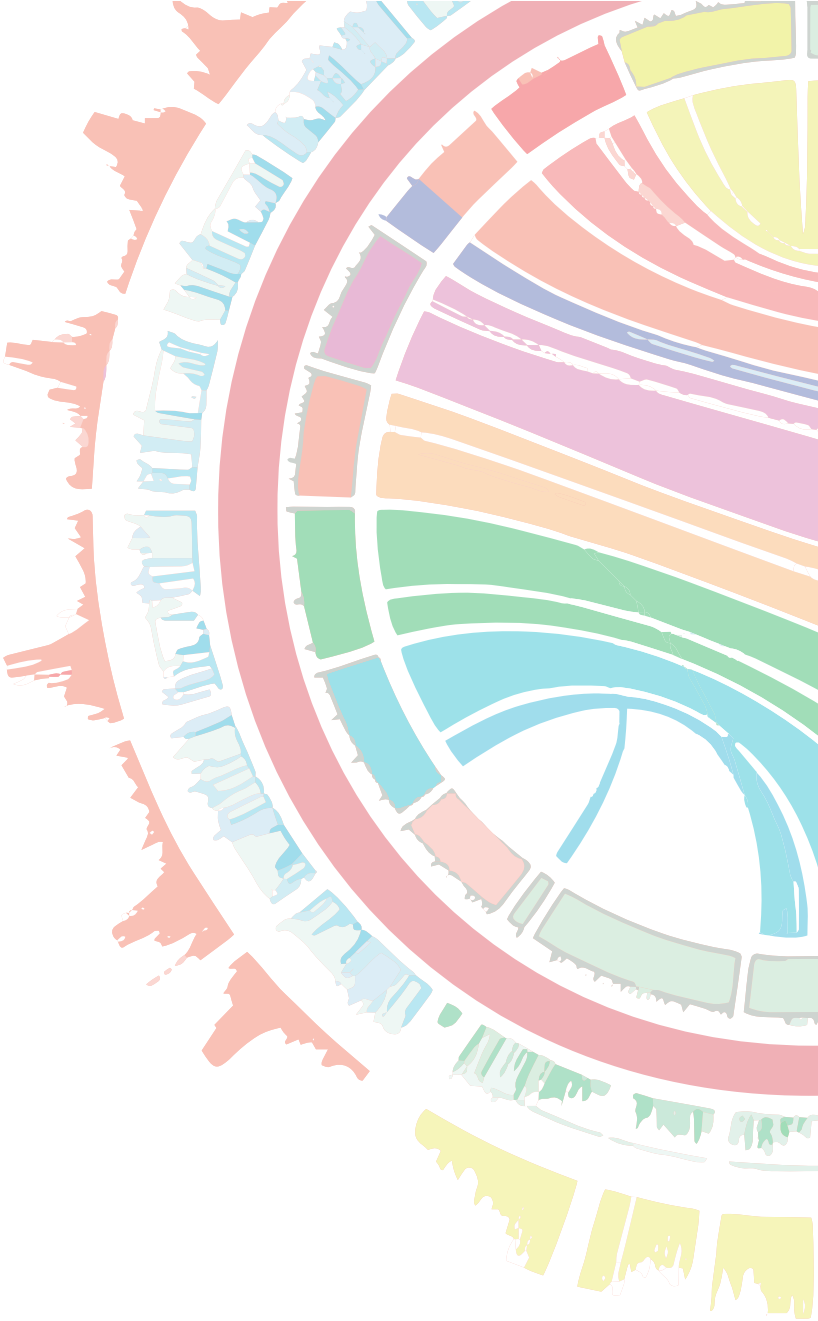
INCEPTION

INRAE

INVESTIS L'AVENIR

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

PSL UNIVERSITÉ PARIS | life



> Research

Towards a gold standard for enhancer-gene (E-G) interactions

Tristan HOELLINGER[1,2], Sarah DJEBALI[1]

[1]IRSD, Université de Toulouse, INSERM, INRAe, ENVT, UPS, Toulouse, France
[2]INSA de Toulouse, INP-ENSEEIHT, France

Corresponding Author: sarah.djebali@inserm.fr

Although cell type specific gene expression regulatory relationships such as enhancer-gene (E-G) are recognized to be tremendously important to consider for a better understanding of complex human genetic diseases, there is still no consensus about the best way to identify them genome-wide. There are nonetheless four broad types of approaches in this field: (1) cell type specific gene expression QTL (eQTL) analysis combined with a set of cell type specific enhancers, (2) computational prediction from several types of high-throughput functional genomic data 1D (RNA-seq, ATAC-seq, histone marks, etc), (3) computational prediction from a single type of chromosome conformation 3D data (promoter capture HiC, polII ChIA-PET data, etc) combined with a set of cell type specific enhancers, (4) cell type specific genetic screening data analysis.

Because (1) and (3) are costly, (4) is currently not genome-wide and many international projects have now generated plenty of functional genomic 1D data, (2) appear as more promising. For this reason we have tried to evaluate the few most recent and promising methods of the field, the ABC model [1] and the average rank method [2], together with the very popular distance method, based on training and/or evaluating on the two most recent reference sets of the field: the CRISPRi-FlowFISH set based on approach (4) in K562 cell line for 30 genes covering 5Mb [1] and the BENGI set made of data from (1), (3) and (4) [2].

Our results show a very different ranking of the evaluated methods based on the reference sets used for the evaluation, and the distance method to be the most robust in terms of equality of performance across reference sets. However this method is also known not to be satisfying on many real-life examples [3]. On the other hand the seven reference sets used here have been generated in many different ways and may be complementary : while 3D reference sets are genome-wide they have not been specifically designed to identify E-G interactions and may contain many false positives and false negatives; on the other hand CRISPRi-FlowFISH data have specifically been designed to identify E-G interactions in a reliable way (FDR < 0.05) [4], however they are not genome-wide. In fact 3D reference sets lack experimental confirmation and therefore cannot *stricto sensu* be considered reference sets of E-G interactions. Therefore, before designing a new method based on 1D data to identify E-G interactions genome-wide, we recommend the definition of a reliable and genome-wide set of E-G interactions in a cell line for which many types of 1D data are available.

References

[1] Fulco CP et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nature genetics. 2019 Dec;51(12):1664-9.

[2] Moore JE et al. A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods. Genome biology. 2020 Dec 1;21(1):17.

[3] Mumbach MR et al. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nature genetics. 2017 Nov;49(11):1602.

[4] Fulco CP et al. Systematic mapping of functional enhancer–promoter connections with CRISPR interference. Science. 2016 Nov 11;354(6313):769-73.

Resting and stimulated human PBMC single-cell RNA-seq data integration

Marie-Ange PALOMARES[1], Céline DERBOIS[1], Jean-François DELEUZE[1], Eric CABANNES[1] and Eric BONNET[1]

[1] Centre National de Recherche en Génomique Humaine (CNRGH), Institut de Biologie François Jacob, CEA, Université Paris-Saclay, Evry, France.

Corresponding Author: eric.bonnet@cnrgh.fr

DNA sequencing technology has scaled up very rapidly in throughput but also, through rapid advances in sample preparation, scaled down in terms of the amount of DNA that is required for analysis, to the point that it is now commonplace to analyze the DNA and RNA content of individual cells. This technology has triggered previously impossible applications in basic research and clinical science. Some examples are: transcriptome analysis of rare circulating tumor cells; characterization of early differentiation events in human embryogenesis; investigation of transcriptional noise and stochastic fate choice and creation of large-scale cell atlases such as the human cell atlas[1].

However, integrated analysis of different scRNA-seq data sets, consisting of multiple transcriptomic subpopulations or to integrate measurements produced by different technologies, remains challenging. It is especially difficult to distinguish between the composition of cell types in a sample and expression changes within a given cell type [2].

In this study, we used single-cell RNA-seq (scRNA-seq) to assess the expression of ~ 8,000 peripheral blood mononuclear cells (PBMC). PBMC were isolated from a healthy anonymous donor's whole blood specimen provided by EFS (Etablissement Français du Sang), using a Ficoll gradient. PBMC were then splitted into 2 samples : non-stimulated and LPS-stimulated (incubation for 4 hours with 1µg/mL Lipopolysaccharide).

Lipopolysaccharide (LPS) is the most abundant component within the cell wall of Gram-negative bacteria. It can stimulate the release of interleukin 8 and other inflammatory cytokines in various cell types, leading to an acute inflammatory response towards pathogens. Bacterial LPS has been extensively used in models studying inflammation as it mimics many inflammatory effects of cytokines [3].

The cells were then processed using the Chromium Next GEM Single Cell 3' GEM v3.1 kit following the manufacturer recommandations (10X genomics) to target 4000 cells/sample. Dual indexed libraries were sequenced on a NextSeq500 sequencer (Illumina). Sample demultipexing, barcode processing and unique molecular identifiers (UMI) counting were performed by using the 10X genomics pipeline Cellranger v4.0.0 with default parameters and target cells value of 4000 for each sample.

Quality control and analysis were performed using the R package Seurat v4.0.0. After removal of bad quality cells in each sample, we merged the samples using the Seurat function "merge", defining a global dataset of 6,232 cells. After normalization, scaling and dimension reduction with the UMAP algorithm, we could clearly see distinct groups of resting and stimulated cell types, such as T cells. A number of marker genes were visibly associated with either stimulated or non-stimulated groups. Then we performed data integration on the two data sets using the Seurat functions "FindIntegrationAnchors" and "IntegrateData". Data integration in the Seurat package is based on techniques used in computer vision for the alignment and integration of image data. Basically, these techniques identify shared correlations structures across data sets with canonical correlation analysis (CCA) and use basis vector to create aligned data sets. After integration, normalization, scaling and dimension reduction, we could identify 12 different cluster groups for which we could assign a cellular identity based on their typical markers, such as B cells, CD8 T cells, NK cells, etc. Such a dataset exemplify the utility of single-cell heterogeneous data set integration techniques.

References

1. Single-cell sequencing-based technologies will revolutionize whole-organism science. Shapiro E, Biezuner T, Linnarsson S. Nat. Rev. Genet. 2013, 14, pp618-630.
2. Comprehensive integration of single-cell data. Stuart S et al. 2019, Cell 177, 1888-1902.
3. Endotoxin signal transduction in macrophages. Sweet MJ, Hume DA, J Leukoc Biol. 1996 Jul; 60(1):8-26.

Metagenomic and metatranscriptomic analysis for the study of clouds and aerosols

Raphaëlle PEGUILHAN[1], Florent ROSSI[1], François ENAULT[2], Laurent DEGUILLAUME[3,4] and Pierre AMATO[1]

[1] Université Clermont Auvergne, CNRS, SIGMA Clermont, ICCF, F-63000 CLERMONT-FERRAND, France

[2] Université Clermont Auvergne, CNRS, Laboratoire Microorganismes : Genome et Environnement, F-63000 CLERMONT-FERRAND, France

[3] Université Clermont Auvergne, CNRS, Observatoire de Physique du Globe de Clermont-Ferrand, UMS 833, F-63000 CLERMONT-FERRAND, France

[4] Université Clermont Auvergne, CNRS, Laboratoire de Météorologie Physique, UMR 6016, F-63000 CLERMONT-FERRAND, France

Corresponding Author: raphaelle.peguilhan@uca.fr

Microorganisms are present in the outdoor atmosphere up to high altitude and, consequently, are prone to integrate clouds. Cloud droplets offer liquid airborne microenvironments to airborne cells and could thus favour microbial activity. Microbial diversity in clouds and aerosols has been already well documented. However, functional aspects are still largely unexplored.

Here, we aim at better understanding the functioning of microbial assemblages in clouds in comparison to aerosols, using metagenomics coupled with metatranscriptomics. We developed analytic protocols allowing the recovery of metagenomes and their associated metatranscriptomes from clouds and aerosols, by direct HiSeq shotgun sequencing with no preliminary amplification step such as PCR or MDA. Around 25 to 300 ng of total DNA and 33 to 240 ng of total RNA were extracted from samples. Data from the first cloud sample are used for developing bioinformatic analyses, while the sequencing of additional aerosols and cloud water samples is currently being processed. In order to analyse the consequent and complex dataset generated (around 190 to 260 million paired-end reads), a workflow was set up using the Galaxy platform and the resources and facilities deployed by the AuBi (Auvergne BioInformatique) network and the regional calculation cluster Mesocentre Clermont Auvergne. The workflow is based on tools such as Trimmomatic for filtering reads and trimming, MEGAHIT for assembly, Bowtie2 for mapping, Kraken2 for taxonomic affiliation against “PlusPF” kraken database and Blastx for functional assignation against UniprotKB. R environment is used for further data treatment and statistical analyses.

Preliminary results based on RNA to DNA sequences abundance ratios point Bacteria as the most active community members, in particular Clostridiales (eq. Eubacteriales), Burkholderiales and rare groups as Oceanospirillales, Rickettsiales and Parachlamydiales. In Eukaryota, Saccharomycetales in fungi and Cryptomonadales in microalgae are the most active. The functional analysis is still in progress. Preliminary results indicate the expression of numerous stress responses to a stressful environment, such as DNA repair activity, responses to osmotic and oxidative stresses, responses to hydrogen peroxide, light stimulus and starvation as well as autophagy and sporulation activities.

In conclusion, metagenomics coupled with metatranscriptomics analyses will permit to go deeper in the exploration of microbial functional diversity in clouds compared to aerosols.

Model learning to identify systemic regulators of the peripheral circadian clock

Julien Martinelli[1,2] and Annabelle Ballesta[2]
[1] Institut Curie, 35 Rue Dailly, 92210, Saint Cloud, France
[2] Lifeware Group Inria Saclay, 1 Rue Honoré d'Estienne d'Orves, 91120, Palaiseau, France

Corresponding author: julien.martinelli@inria.fr

Personalized medicine aims at providing patient-tailored therapeutics based on multi-type data towards improved treatment outcomes. Chronotherapy that consists in adapting drug administration to the patient's circadian rhythms may be improved by such approach [1]. Recent clinical studies demonstrated large variability in patients' circadian coordination and optimal drug timing. Consequently, new eHealth platforms allow the monitoring of circadian biomarkers in individual patients through wearable technologies (rest-activity, body temperature), blood or salivary samples (melatonin, cortisol), and daily questionnaires (food intake, symptoms). A current clinical challenge involves designing a methodology predicting from circadian biomarkers the patient peripheral circadian clocks and associated optimal drug timing. The mammalian circadian timing system being largely conserved between mouse and humans yet with phase opposition, the study was developed using available mouse datasets.

We investigated at the molecular scale the influence of systemic regulators (e.g. temperature, hormones) on peripheral clocks, through a model learning approach involving systems biology models based on ordinary differential equations. Using as prior knowledge our existing circadian clock model [2], we derived an approximation for the action of systemic regulators on the expression of three core-clock genes: *Bmal1*, *Per2* and *Rev-Erbα*. These time profiles were then fitted with a population of models, based on linear regression. Best models involved a modulation of either *Bmal1* or *Per2* transcription most likely by temperature or nutrient exposure cycles. This agreed with biological knowledge on temperature-dependent control of *Per2* transcription. The strengths of systemic regulations were found to be significantly different according to mouse sex and genetic background [3].

References

[1] Annabelle Ballesta, Pasquale F. Innominato, Robert Dallmann, David A. Rand, and Francis A. Lévi. Systems chronotherapeutics. *Pharmacological Reviews*, 69(2):161–199, 2017.

[2] Janina Hesse, Julien Martinelli, Ouda Aboumanify, Annabelle Ballesta, and Angela Relógio. A mathematical model of the circadian clock and drug pharmacodynamics to optimize irinotecan administration timing in colorectal cancer, 2021. Under review.

[3] Julien Martinelli, Sandrine Dulong, Xiao-Mei Li, Michèle Teboul, Sylvain Soliman, Francis Lévi, François Fages, and Annabelle Ballesta. Model learning to identify systemic regulators of the peripheral circadian clock, 2021. To appear in *Bioinformatics*.

Exploring the conformational rearrangement of the human Insulin Degrading Enzyme through Molecular Dynamics Simulations

M. Ghoula[1], G. Postic[1], A-C Camproux[1], G. Moroy[1]

[1] Université de Paris, BFA, UMR 8251, CNRS, ERL U1133, Inserm, F-75013 Paris, France

Corresponding Author : mariem.ghoula@inserm.fr

Abstract

Insulin Degrading Enzyme (IDE) is a metallopeptidase that degrades a large panel of amyloidogenic peptides and is thought to be a potential therapeutic target for type-2 diabetes and neurogenerative diseases like Alzheimer's disease [1]. Interestingly, IDE is a cryptidase. Its catalytic chamber, known as a crypt, is formed so that peptides can be enclosed and degraded [2]. However, the molecular mechanism of IDE function and peptide recognition remains elusive. It has been shown that, IDE undergoes several conformational changes and switches between closed and open states in order to regulate peptide degradation and cleavage [3]. Thereby, it is essential to unfold IDE mechanism and provide more information on how conformational dynamics can modulate the catalytic cycle of IDE.

In this aim, a free-substrate IDE crystallographic structure (PDB ID : 2JG4) was used to model a complete structure of IDE with the MODELLER software [4]. IDE stability and flexibility were studied through Molecular Dynamics simulations with the GROMACS software [5] and the CHARMM36m force field [6]. In total, we ran 7 simulations of 1μs each to cover a wide range of the IDE conformational space. The crypt volume as well as the Solvent Accessible Surface Area (SASA) were calculated to witness IDE conformational dynamics switching from a closed to an open state. The Gibbs free energy landscapes were also investigated to indicate the different conformational states accessible to the protein during the simulations. The Molecular Mechanics / Poisson-Boltzmann Surface Area (MM/PBSA) method was used to identify key residues involved in IDE rearrangement.

References

1. Tang WJ. Targeting Insulin-Degrading Enzyme to Treat Type 2 Diabetes Mellitus. *Trends Endocrinol Metab*. 2016;27(1):24-34.
2. Shen Y, Joachimiak A, Rosner MR, Tang WJ. Structures of human insulin-degrading enzyme reveal a new substrate recognition mechanism. *Nature*. 2006;443(7113):870-874.
3. B. Webb, A. Sali. Comparative Protein Structure Modeling Using Modeller. Current Protocols in Bioinformatics 54, John Wiley & Sons, Inc., 5.6.1-5.6.37, 2016.
4. M.J. Abraham, D. van der Spoel, E. Lindahl, B. Hess, and the GROMACS development team, GROMACS User Manual version 2019
5. Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M, de Groot, B.L., Grubmuller, H., and MacKerell, A.D., Jr., "CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins," *Nature Methods*, 14:71-73, 2016.

SURFMAP : a software for mapping in two dimensions any feature of a protein surface

Hugo SCHWEKE [1,2], Simon GOSSET[2], Anne LOPES[3] and Marie-Hélène MUCCHIELLI-GIORGI[2]

[1] Weizmann Institute of Science, Kimmelman Building, 76100, Rehovot, Israel

[2] Institut des sciences des plantes de Paris-Saclay,Batiment 630 rue Noetzlin, 91190, Gif sur Yvette, France

[3] Institut de Biologie Intégrative de la Cellule, 1 avenue de la terrasse, 91190, Gif sur Yvette, France

Corresponding Author: anne.lopes@i2bc.paris-saclay.fr

A lot of methods have been developed in order to analyze and compare protein surface features. Some of them focus on specific regions of the surface such as protein binding sites, active sites, binding pockets and ignore the rest of the surface which plays an important role in protein interactions by constantly competing with the interaction sites. Other methods focus on "molecular cartography » either by projecting the 3D structure of a protein into two dimensions (2D) and then projecting features of interest on the resulting 2D map [1], or by representing the protein surface based on a spherical approximation [2]. But no tool is available to easily visualize any physico-chemical characteristics of a protein surface: they are either limited to a restricted number of surface descriptors, and/or do not provide a standalone tool to compute the corresponding 2D maps.

We therefore developed SURFMAP, a free software designed to allow the two-dimensional projection of either predefined features of protein surface (electrostatic potential, hydrophobicity, stickiness and surface relief) or any descriptor encoded in the temperature factor column of a PDB file. SURFMAP uses a pseudo-cylindrical sinusoidal "equal-area" projection that has the advantage of preserving the area measures at the cost of distorting shapes locally. Indeed, we aim to preserve the size of the regions of interest rather than providing a precise representation of their shapes.

To map protein surface properties: (1) the values of the feature of interest are calculated for each protein residue or atom; (2) a set of particles localized at 3Å from the surface of the protein of interest is generated to approximate the protein shape. We will refer to this ensemble as the "protein shell"; (3) the value of the feature of interest of the closest protein residue (or atom) is assigned to each particle of the protein shell; (4) coordinates of the particles are projected onto a 2-dimensional plane by a sinusoidal projection, and each projected pair of spherical coordinates (φ, θ) is assigned to the corresponding particle feature value and represented on the 2-D plan at the point of coordinates ($\varphi \sin\theta, 90-\theta$); (5) the resulting map is then divided into a 72 x 36 cells and values of the points contained in the same cell are averaged; (6) the map can then be smoothed by averaging the score of each cell with the scores of the adjacent cells.

These 2D maps ease greatly the visualization and analysis of the distribution of a given feature compared to the 3D protein surface which is always difficult to handle despite the efforts made on protein 3D structure representation tools. They permit the visual comparison of surface features of homologs and are well suited for large-scale comparison since (i) it only involves the calculation of a map similarity through a straightforward numerical measure and (ii) the dimension reduction is robust against local irregularities of protein surfaces.

References

1. DW Fanning, JA Smith and GD Rose. Molecular cartography of globular proteins with application to antigenic sites. *Biopolymers*. 25:863–83, 1986.
2. JM Sasin, A. Godzik and JM Bujnicki. Protein classification by surface comparisons. *J Biosci.* 32:97–100, 2007

LRez: C++ API and toolkit for analyzing and managing Linked-Reads data

Pierre Morisse[1], Claire Lemaitre[1] and Fabrice Legeai[1,2]
[1] Univ Rennes, Inria, CNRS, IRISA, 35000, Rennes, France
[2] IGEPP, INRAE, Institut Agro, Univ Rennes, 35000, Rennes, France

Corresponding author: pierre.morisse@inria.fr

Abstract

Linked-Reads technologies, such as 10x Genomics, Haplotagging, stLFR and TELL-Seq, partition and tag high-molecular-weight DNA molecules with a barcode using a microfluidic device prior to classical short-read sequencing. This way, Linked-Reads manage to combine the high-quality of the short reads and a long-range information which can be inferred by identifying distant reads belonging to the same DNA molecule with the help of the barcodes. This technology can thus efficiently be employed in various applications, such as structural variant calling, but also genome assembly, phasing and scaffoling. To benefit from Linked-Reads data, most methods first map the reads against a reference genome, and then rely on the analysis of the barcode contents of genomic regions, often requiring to fetch all reads or alignments with a given barcode.

However, despite the fact that various tools and libraries are available for processing BAM files, to the best of our knowledge, no such tool currently exists for managing Linked-Reads barcodes, and allowing features such as indexing, querying, and comparisons of barcode contents. LRez aims to address this issue, by providing a complete and easy to use API and suite of tools which are directly compatible with various Linked-Reads sequencing technologies.

LRez provides various functionalities such as extracting, indexing and querying Linked-Reads barcodes, in BAM, FASTQ, and gzipped FASTQ files (Table 1). The API is compiled as a shared library, helping its integration to external projects. Moreover, all functionalities are implemented in a thread-safe fashion.

Our experiments show that, on a 70 GB Haplotagging BAM file from *Heliconius erato* [1], index construction took an hour, and resulted in an index occupying 11 GB of RAM. Using this index, querying time per barcode reached an average of 11 ms. In comparison, using a naive approach without a barcode-based index, querying time per barcode reached an hour.

LRez is available on GitHub at https://github.com/morispi/LRez and as a bioconda module. Additionally, its features are already used in the SV calling tool LEVIATHAN (https://github.com/morispi/LEVIATHAN) and in the gap-filling pipeline MTG-Link (https://github.com/anne-gcd/MTG-Link).

Command	Description
compare	Compute the number of common barcodes between pairs of regions or between pairs of contig ends
extract	Extract the barcodes from a given region of a BAM file
index bam	Index the BAM offsets or genomic positions of the barcodes contained in a BAM file
index fastq	Index by barcode the offsets of the sequences contained in a FASTQ or gzipped FASTQ file
query bam	Query the index to retrieve alignments in a BAM file given a barcode or list of barcodes
query fastq	Query the index to retrieve sequences in a FASTQ / gzipped FASTQ file given a barcode or list of barcodes

Tab. 1. LRez features.

Acknowledgements

This project has received funding from the French ANR ANR-18-CE02-0019 Supergene grant.

References

[1] Joana I. Meier et al. Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. *bioRxiv*, pages 1–27, 2020.

Development of integrative pipelines for transcriptional and epigenetic analyses: application to *Dot1l* and H3K79me2 during spermatogenesis

Manon Coulée[1], Clara Gobé[1], Mélina Blanco[1], Alban Lermine[2], Romain Daveau[2], Julie Cocquet[1*], Laila El Khattabi[1,3*]

[1] Institut Cochin, INSERM U1016, CNRS UMR8104, Université de Paris, Paris, France
[2] MOABI (Bioinformatic platform of APHP), Paris, France
[3] APHP centre-université de Paris, service de génétique, Paris, France

Corresponding Authors: julie.cocquet@inserm.fr & laila.el-khattabi@inserm.fr

Spermatogenesis is the biological process during which male germ cells develop into the highly specialized cells that are spermatozoa. It can be divided in 3 parts: proliferation of spermatogonial stem cells, meiosis, and extreme differentiation of round spermatids into spermatozoa. It is one of the most dynamic differentiation processes in terms of gene expression and chromatin remodelling[1]. Deregulation can lead to male infertility and/or could affect embryo development and offspring health[2].

Previous data from our group and others indicate that *Dot1l* (*Disruptor of telomeric silencing 1-like*) is essential for normal spermatogenesis[3]. *Dot1l* encodes the only methyltransferase able to methylate H3K79, a histone post translational mark highly enriched at the end of spermatogenesis[4]. To investigate the underlying molecular mechanism, we produced a mouse model with a knock-out (KO) of *Dot1l* gene. We have realized RNAseq analyses on KO and control male germ cells at different stages (primary and secondary spermatocytes and round spermatids) and H3K79me2 ChIPseq analyses on wild-type male germ cells. We have developed an RNAseq pipeline using STAR alignment tool and different downstream analyses, such as the analysis of differentially expressed genes (using DEseq2), ontology and GSEA (Gene Set Enrichment Analysis) analyses, as well as time course analyses (TCseq). In parallel, we have developed a ChIPseq pipeline (bowtie2 followed by MACS2) and characterized H3K79me2 peak distribution, genomic annotation and compared it with different epigenetic marks using ChromHMM. These two pipelines were combined to integrate results obtained from our RNAseq analyses in our ChIPseq pipeline and enable us to correlate the transcriptional changes induced by *Dot1l* loss with chromatin dynamic during spermatogenesis.

References

1. Rathke, C., Baarends, W. M., Awe, S. & Renkawitz-Pohl, R. Chromatin dynamics during spermiogenesis. *Biochim. Biophys. Acta BBA - Gene Regul. Mech.* 1839, 155–168 (2014).
2. Blanco, M., Cocquet, J. Genetic factor affecting sperm chromatin structure. *Adv Exp Med Biol.* (2019)
3. Moretti, C. *et al.* SLY regulates genes involved in chromatin remodeling and interacts with TBL1XR1 during sperm differentiation. *Cell Death Differ.* 24, 1029–1044 (2017).
4. Dottermusch-Heidel, C. *et al.* H3K79 methylation directly precedes the histone-to-protamine transition in mammalian spermatids and is sensitive to bacterial infections. *Andrology* 2, 655–665 (2014).

Talkmine, a workflow for the prediction of the interactions between secretome and surfaceome in the dialogue between cellular types

Manon CONNAULT[1], Jérémy TOURNAYRE[2], Céline BOBY[2], Muriel BONNET[2] and Nadia GOUÉ[1]

[1] AuBi plateform, Mesocenter, Clermont-Auvergne University, Turing Building, 7 Avenue Blaise Pascal, 63170 AUBIERE, France

[2] INRAE, Clermont-Auvergne University, Vetagro Sup, UMRH, 63122 SAINT-GENES-CHAMPANELLE, France

Corresponding Authors: nadia.goue@uca.fr, muriel.bonnet@inrae.fr

1 Context

The prediction of the protein-protein interactions is studied in different domains, for example for their roles in the interplay between cellular types or tissues[1]. In particular, the interactions between surfaceome and secretome may strongly improved the understanding of cellular crosstalk. This project aims to develop a Snakemake-based workflow[2] named Talkmine for the identification of molecular dialogue between two biologic tissues resulting from protein-protein interactions.

2 Materials and Methods

First objective was to identify and test some opensource tools known to predict proteins that belong to secretome or surfaceome. Publically available tools were classified according to three classes, i.e. peptide signal, subcellular location or topology prediction, and have been tested with a dataset of 165 UniProt protein entries whose subcellular location are known.

Secondly, we developed a workflow to connect the selected tools. In brief, user sends a gene or protein identifiers list to a python tool, g:Convert[3] which converts identifiers to a same format, according to a defined database. Then, Entrez-Direct tool[4] generates a multi-fasta file with all protein sequences listed in the NCBI database, injected to secretome and surfaceome tools. Finally, proteins tagged to secretome and surfaceome classes are send to PSICQUIC tool[5], which determines the interactions between proteins from these two classes.

3 Results

The resulting protein-protein interactions prediction from Talkmine is available both in standard output and in flat file. The user has access to the list of protein-protein interactions between the two tissues but also to the intermediate results.

4 Conclusion and perspectives

The workflow Talkmine is developed to predict the interactions between the proteins from the secretome and the surfaceome. This work was initiated to be applied to *Bos taurus* to determine the interactions between muscle and fat tissue but it is possible to apply it to other tissues and to application fields such as biomedical or food research.

References

[1] Bonnet M., et al. Prediction of the Secretome and the Surfaceome: A Strategy to Decipher the Crosstalk between Adipose Tissue and Muscle during Fetal Growth. *Int J Mol Sci* 2020;21:4375. doi:10.3390/ijms21124375

[2] Mölder F., et al. Sustainable data analysis with Snakemake. *F1000Res* 2021;10:33. doi:10.12688/f1000research.29032.1

[3] Raudvere U., et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res* 2019;47:W191–W198. doi:10.1093/nar/gkz369

[4] Kans J. Entrez Direct: E-utilities on the Unix Command Line. In: *Entrez Programming Utilities Help [Internet]*. National Center for Biotechnology Information (US), 2013.

[5] Aranda B., et al. PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat Methods*, 2011;8:528–529. doi:10.1038/nmeth.1637

Towards an integrative multi-omics workflow

Florian JEANNERET[1] and Stéphane GAZUT[1]
[1]Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France.

Corresponding author: florian.jeanneret@cea.fr

Abstract *The advent of high-throughput techniques has greatly enhanced biological discovery. Last years, analysis of multi-omics data has taken the front seat to improve physiological understanding. Handling functional enrichment results from various biological data raises practical questions.*
We propose an integrative workflow, wrapped in a Bioconductor R package multiSight, to better interpret biological process insights in a multi-omics approach. In this work, we present this workflow applied to breast cancer data from The Cancer Genome Atlas (TCGA) related to Invasive Ductal Carcinoma (IDC) and Invasive Lobular Carcinoma (ILC). Pathway enrichment, based on Reactome database, by Over Representation Analysis (ORA) and Gene Set Enrichment Analysis (GSEA) has been conducted with both features information from differential expression analysis (DEA) or selected features from multi-block sPLS-DA methods. Then, comprehensive comparisons of enrichment results have been carried out by looking at classical enrichment analysis, probabilities pooling by Stouffer's Z scores method and pathways clustering in biological themes.

Our multi-omics analysis workflow is fed by expression data for several biological actor (e.g. RNAseq and RPPA) and for the same patients that begets a multi-omics context. Firstly, to select a subset of relevant features for each omic block, either Differential Expression Analysis (DEA) are carried out by DESeq2 and limma analysis or the features are selected by sPLS-DA. Secondly, to reveal altered pathways between patient phenotypes, we are aiming at enrichment analysis by ORA or GSEA with selected feature subsets. Then, enrichment results from omic blocks are merged by p-values using Stouffer's probabilities pooling method. All these enrichment results are visualised by enrichment map to interpret easily highlighted biological themes.
Classical enrichment results, using DEA to feed ORA and GSEA, have shown pathways related to breast cancer and only one or two biological themes really associated to ILC or IDC carcinomas. On the other hand, multi-omics tables based on Stouffer's probabilities pooling method has led to more targeted and confident biological interpretation considering both enrichment results from each kind of omic features.

To complete the multi-omics approach we propose an assessing of a multi-omics features selection method substituting the DEA.
sPLS-DA driven features selection enables a more phenotype-targeted functional enrichment. The biological themes associated to ILC and IDC carcinomas differences are highlighted in enrichment maps after pathways clustering. Moreover, Stouffer's probabilities pooling method shows great benefits to more easily interpret several enrichment results in a multi-omics context.

Our work shows that ORA enrichment with selected sPLS-DA feature and pathway probabilities pooling by Stouffer's method lead to enrichment maps highly associated to the physiological knowledge of IDC or ILC phenotypes, better than ORA and GSEA with differential expression driven features. Lastly, following to these results, an R package named multiSight have been developped to handle functional enrichment results in a multi-omics context.

Keywords multi-omics, pathway enrichment, p-values pooling, sPLS-DA, R package

Cosimu: Connectivity-based simulation of derivative fold-changes from reference fold-changes

Catalina GONZALEZ GOMEZ[1,2], Julien FOURET[2]

[1] INSA Lyon, 20 Avenue Albert Einstein, 69100, Lyon, France

[2] Signia Therapeutics, 60 Avenue Rockefeller, 69008, Lyon, France

Corresponding Author: catalina.gonzalez@hotmail.com

Drug repurposing by analyzing gene expression profiles is a field that has been developed during the past few years. One of the first databases dedicated to exploit these profiles was Connectivity Map (Cmap) [1]. The profile querying of Cmap is based on a "connectivity score" ranging from +1 to -1, which characterizes the magnitude and the direction of the relation between two expression profiles. In the field of *in silico* repurposing, such a score is used to prioritize candidates for repurposing with the assumption that if a molecule profile is negatively connected with a pathology profile then this molecule could counteract this pathology. Unfortunately, there is no existing tool that enables the researchers to simulate linked gene expression profiles with known connectivity scores in order to test the performance of their methods. This is the reason why we are developing *cosimu*, a package that simulates fold change vectors related by a connectivity parameter as input, each element of the vectors represents a gene or transcript $\log_2$ fold changes in expression levels. In order to simulate related fold change vectors, we determined two types: the reference vector and the derivative ones.

There are a few steps before defining the reference fold change vectors. First, we randomly determine the regulation modality (up-regulated, down-regulated or non-regulated) of each gene/transcript (later referred as ‘entity’), thanks to input proportions for each category. Moreover, in order to be more realistic, we included sub-modalities among each modality as two up-regulated entities won’t have the same amplitude of deregulation. The sub-modalities are modeled by a normal law and with parameters (mean, SD) and proportion as inputs. Once each entity has its modality and sub-modality, the reference fold change vector is generated by sampling from the normal distribution associated.

The derivative fold change vector is more complex to generate, as we have to introduce the connectivity among input parameters; This is a three-step process.

(i) The modality of each entity in the derivative vector will be determined by a simple stochastic process based on 3x3 matrix defining the probability of transition from a modality to another or itself. Those probabilities are determined by the input connectivity parameter. However, we found that it is necessary to introduce a new parameter representing the noise added by the transition from and to non-regulated entities.

(ii) We handle the transition between sub-modalities based on a binomial distribution in order to determine the rank of the derivative sub-modality of each transcript. Alternately to this dependent and stochastic transition we have also implemented two alternative methods : independent and stochastic or dependent and deterministic.

(iii) Similarly we defined a dependent and stochastic method for fold-change value transitions with analogous alternatives.

We are currently using this tool to benchmark different approaches aiming to estimate the connectivity between transcriptomic signatures. As a baseline, we evaluated Pearson and Spearman correlations as connectivity scores along with a simple connectivity score based on Cmap. We then characterized different biases for all methods depending on the simulation parameters. Thanks to a better understanding of biases behind connectivity estimators, we hope it will be possible to develop less biased estimators.

References

[1] Justin Lamb, *et al*. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science*,313:1929-1935,2006.

Panache and the Linearized Pangenomes: a Visualization Story

Éloi DURANT[1,2,3,4], François SABOT[2,4], Matthieu CONTE[3] and Mathieu ROUARD[3,4]

[1] DIADE, Univ Montpellier, CIRAD, IRD, 34830, Montpellier, France

[2] Syngenta Seeds SAS, 12 Chemin de l'Hobit, 31790, Saint-Sauveur, France

[3] Bioversity International, Parc Scientifique Agropolis II, 34397, Montpellier, France

[4] French Institute of Bioinformatics (IFB) — South Green Bioinformatics Platform, Bioversity, CIRAD, INRAE, IRD, 34398, Montpellier, France

Corresponding Authors: eloi.durant@ird.fr, m.rouard@cgiar.org

1. Introduction

Pangenomes are conceptual entities inventorying unique and shared genomic material between genomes of a given group. They can focus on a complete repertoire of genes, or include all genomic material including intergenic sequences [1, 2]. The latter has been receiving an increasing attention these past years, with work being done on graph pangenomes (*sequence graphs* and *variation graphs* among others [3]). However, these are not the only representations used for sequence pangenomes and they face a lack of tools for generating or manipulating them. This extends to the field of visualization, where existing viewers do not scale well from gene to sequence pangenomes or lack user-friendly representations and exploration. We therefore developed *Panache*, a web-based viewer for linearized pangenomes.

2. Panache

A major difficulty encountered with graph pangenomes is the overload of branches created by the numerous variations between genomes, resulting in hard-to-explore visual clutter. We chose to focus on a linear representation instead, easier to read, parse and navigate. Linearity has its drawbacks (loss of the sense of structural variation, lack of malleability…), but can be a good alternative to graphs for exploration and comparison tasks between genomes.

We therefore created Panache [4], a JavaScript application built with the framework Vue.js and the library D3.js. Panache, which stands for PANgenome Analyzer with CHromosomal Exploration, displays pangenomes in a linear fashion similar to that of genome browsers. Only one (pan)chromosome at a time is available on screen, and pangenomic blocks (listed genes or sequences from the pangenome) are laid out on a single string, following a linear coordinate system from the input data. A presence absence matrix describes for each genome the pangenomic blocks that they do own, and additional tracks of information display details like the categorization of these blocks between *core* and *variable* genome.

Panache is available under an MIT license at https://github.com/SouthGreenPlatform/panache and is still undergoing development with regular addition of new features.

Acknowledgements

The authors wish to thank Romain Basset, Mel Florance and Romain Bousquet for their help during the development stages.

References

1. Tranchant-Dubreuil, C., M. Rouard, and F. Sabot, *Plant Pangenome: Impacts on Phenotypes and Evolution.* Annual Plant Reviews online, 2019(2).
2. Golicz, A.A., et al., *Pangenomics Comes of Age: From Bacteria to Plant and Animal Applications.* Trends in Genetics, 2019.
3. Eizenga, J.M., et al., *Pangenome Graphs.* Annual Review of Genomics and Human Genetics, 2020. **21**(1): p. 139-162.
4. Durant, É., et al., *Panache: a Web Browser-Based Viewer for Linearized Pangenomes.* 2021: p. 2021.04.27.441597.

FILT3R: AN NGS-BASED ALGORITHM FOR ROBUST DETECTION AND QUANTIFICATION OF FLT3-ITD IN CLINICAL PRACTICE

Augustin Boudry[1], Nicolas Duployez[1,2], Martin Figeac[3], Sandrine Geffroy[1], Maxime Bucci[1], Karine Celli-lebras[4], Matthieu Duchmann[5], Romane Joudinaud[1], Laurène Fenwarth[1,2], Olivier Nibourel[1], Sasha Darmon[6], Laure Goursaud[7], Raphael Itzykson[4,5], Hervé Dombret[4], Mathilde Hunault[8], Claude Preudhomme[1,2], Mikaël Salson[6]

[1] Laboratory of Hematology, CHU LILLE, F-59037, Lille, France
[2] UMR-S 1277, INSERM, F-59045, Lille, France
[3] Plate-forme de génomique fonctionnelle et structurale Go@L-GFS, Université Lille, F-59000, Lille, France
[4] Department of Hematology, Saint Louis Hospital, Assistance Publique-Hôpitaux de Paris (AP-HP), F-63110, Paris, France
[5] INSERM/CNRS UMR 944/7212, Saint-Louis Research Institute, Paris Diderot University, F-63103, Paris, France
[6] Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
[7] Department of Hematology, CHU LILLE, F-59037, Lille, France
[8] Department of Hematology, CHU Angers, F-49933, Angers, France
Corresponding Author: augustin.boudry@chru-lille.fr

Background: *Fms-like tyrosine kinase 3* internal tandem duplications (*FLT3*-ITD) occur in 20-30% of acute myeloid leukemia (AML) cases. These mutations represent both strong prognostic biomarkers and therapeutic targets. Hence, *FLT3*-ITD are systematically sought and quantified using quantitative fragment analysis, as per European LeukemiaNet guidelines. Although robust, this technique has several limitations (high limit of quantification 1%, determination of the size of the ITD, limited application). High-throughput sequencing (HTS) may represent an alternative method to detect these mutations with high detection sensitivity and the possibility of sequencing several other genes simultaneously. However, the use of NGS technologies to detect *FLT3*-ITD mutations remains challenging because these mutations are heterogeneous in terms of size and/or insertion site, and the perfect or near perfect duplication of the wild-type sequence. These specificities raise new concerns about bio-informatic processes, and several algorithms have been created to detect this duplication (GetITD, km, ScanITD...). However, most of them fail to detect and annotate the broad variety of ITDs.

Aims: This work aims to develop an ultrafast kmer-based algorithm named FiLT3r to detect and quantify ITDs and to compare it with existing algorithms to the reference method.

Methods: FiLT3r first uses a Bloom filter to quickly identify reads matching the *FLT3* gene and then uses their k-mer occurrences to detect *FLT3*-ITD. In a read having a duplication the k-mer positions on FLT3 will not monotonously increase. Such a signature is used to identify the duplication. The different algorithms were tested and compared to the reference method on 500 patients aged 18 to 60 years with de novo AML at diagnosis. Peripheral blood or bone marrow samples from AML patients were screened with our standard NGS routine with a capture method targeting 81 genes (SSQXT Agilent®) and sequenced with Illumina® technology.

Results: A total of 147 ITDs from 114 *FLT3*-ITD positive patients (23%) and 71 randomized *FLT3*-ITD negative patients (14%), determined according to the fragment method, were sequenced (with an average coverage of 2000 paired-end reads on exon 14-15 *FLT3*). FiLT3r used a fraction of the time and memory used by other software and provided results mostly comparable to those obtained by the validated technique, with a sensitivity of 1 and no false positive above a threshold of 1% (limit of quantification of the reference method). The quantification calculated with our algorithm showed a correlation with the reference method estimated at 0.92 which appears as the highest correlation coefficient of all the algorithms tested (range: 0.42 – 0.92).

ORSON: a nextflow workflow for transcriptome and proteome annotation

Cyril Noël[1] , Pierre Cuzin[1], Laura Leroi[1], Alexandre Cormier[1] and Patrick Durand[1]

[1]IFREMER-IRSI-Service de Bioinformatique (SeBiMER), Centre Bretagne - ZI de la Pointe du Diable, CS 10070 - 29280 PLOUZANE, FRANCE

Corresponding Authors: cyril.noel@ifremer.fr

One of the key steps in transcriptomic and proteomic analyses is to link sequences to biology through annotation. Namely, it consist in adding relevant biological information to these sequences by inferring their putative function and other features. However, this process requires a complex combination of successive tools and reference databases, as well as significant computing resources due to the amount of data. Then, it is quite difficult to group together results in order to obtain a complete biological understanding of the sequences because of the many tools and data formats involved. The implementation of an automated, standardized and user-friendly tool to process transcriptomic and proteomic annotations is therefore essential.

We have developed ORSON to combine state-of-the-art tools for annotation processes within a Nextflow [1] pipeline. ORSON combines sequence similarity search, functional annotation retrieval and functional prediction. Sequence comparison can be done using PLAST [2], BLAST [3] or Diamond [4]. Functional annotation retrieval is done using BeeDeeM [5] by gathering feature tables from well annotated reference banks such as Uniprot SwissProt. Functional prediction is performed using InterProScan [6]. It optionally integrates transcriptome completeness analysis using BUSCO [7] as well as ortholog search via eggNOG-mapper [8]. Choose of comparison tool and reference banks is fully customizable using standard Nextflow configuration file. While ORSON results can be analyzed through the command-line, it also offers the possibility to be compatible with BlastViewer graphical tool [9]. ORSON, combined with BlastViewer, offers a real alternative to the complex use of bioinformatic annotation tools by providing the best of both worlds: a scalable workflow running on the command-line to fit any computing infrastructures, and a GUI tool to analyze results. ORSON source code, documentation and installation instructions are freely available at https://github.com/ifremer-bioinformatics/orson.

References

1. Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo P Barja, Emilio Palumbo and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature biotechnology*, 35(4):316-319, 2017.
2. Hoa Van Nguyen and Dominique Lavenier. PLAST: parallel local alignment search tool for database comparison. *BMC bioinformatics*, 10(1):1-13, 2009.
3. Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers and David J Lipman. Basic local alignment search tool. *Journal of molecular biology,* 215:403-410, 1990.
4. Benjamin Buchfink, Chao Xie and Daniel H Huson. Fast and sensitive protein alignment using DIAMOND. *Nature Method*, 12:59-60, 2015.
5. https://github.com/pgdurand/BeeDeeM
6. Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siem-Yit Yong, Rodrigo Lopez and Sarah Hunter. InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236-40., 2014.
7. Mathieu Seppey, Mosé Manni and Evgeny M Zdobnov. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods in Molecular Biology*, 1962, 2019.
8. Jaime Huerta-Cepas, Kristoffer Forslund, Luis P Coelho, Damian Szklarczyk, Lars J Jensen, Christian von Mering and Peer Bork. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol Biol Evol*, 35(8):2115-2122, 2017.
9. https://github.com/pgdurand/BlastViewer

PLMdetect : *de novo* mapping of functional *cis*-regulatory motifs in 5'- and 3'-proximal regions from Arabidopsis and maize

Julien Roziere[1,2,3], Véronique Brunaud[1,2], Sylvie Coursol[3] and Marie-Laure Martin-Magniette[1,2,4]

[1] Université Paris-Saclay, CNRS, INRAE, Univ Evry, Institute of Plant Sciences ParisSaclay (IPS2), 91405, Orsay, France

[2] Université de Paris, CNRS, INRAE, Institute of Plant Sciences Paris Saclay (IPS2), 91405, Orsay, France

[3] Institut Jean-Pierre Bourgin, INRAE, AgroParisTech, Université Paris-Saclay, 78000, Versailles, France

[4] UMR MIA-Paris, AgroParisTech, INRAE, Université Paris-Saclay, 75005, Paris, France

Corresponding Author: julien.roziere@inrae.fr

Identifying cis-regulatory motifs controlling gene expression is an arduous challenge that is actively explored to discover key genetic factors responsible for traits of agronomic interest. The Preferentially Located Motif detection (PLMdetect) method was developed to identify over-represented motifs (PLMs) in promoters at a preferred distance from the transcription start site in the model plant Arabidopsis [1]. Here, we expanded the PLMdetect method to comprehensively analyze de novo the promoters as well as the untranslated transcribed regions of Arabidopsis and the important crop maize. We sought to determine how their differences in genome content and architecture would be reflected in features of their PLMs in 5'- and 3'-proximal regions of each gene locus. We have currently identified three groups of PLMs for each species in each targeted region. An assessment of these PLMs using known plant transcription factor (TF) binding site (TFBS) data [2] revealed that a subset of these PLMs (9.4% and 7.3% in Arabidopsis and maize, respectively) are previously characterized TFBSs (tPLMs), while the others represent novel and uncharacterized motifs (uPLMs), not captured by the current collection of plant TFBSs. Positional analyses of the tPLMs revealed positional preferences of TFBSs from several TF families as previously reported in Arabidopsis [3]. Furthermore, GO term enrichment analyses showed that 15.3% of the uPLMs are able to infer functional predictions which are not provided by tPLMs. In the near future, we will add comparisons between the datasets obtained from each species. Additionally, the development of the interactive PLMviewer website will provide the plant community with a valuable resource of PLM datasets for exploitation to investigate user-specific sequences.

Acknowledgements

This research is supported by a grant from the Plant2Pro Carnot Institute. IJPB and IPS2 benefit from the support of Saclay Plant Science-SPS (ANR-17-EUR-0007).

References

1. Bernard V, Lecharny A, Brunaud V. Improved detection of motifs with preferential location in promoters. Genome. 2010 Sep;53(9):739-52. doi: 10.1139/g10-042. PMID: 20924423.
2. Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranašić D, Santana-Garcia W, Tan G, Chèneby J, Ballester B, Parcy F, Sandelin A, Lenhard B, Wasserman WW, Mathelier A. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2020 Jan 8;48(D1):D87-D92. doi: 10.1093/nar/gkz1001. PMID: 31701148; PMCID: PMC7145627.
3. Yu CP, Lin JJ, Li WH. Positional distribution of transcription factor binding sites in Arabidopsis thaliana. Sci Rep. 2016 Apr 27;6:25164. doi: 10.1038/srep25164. PMID: 27117388; PMCID: PMC4846880.

Integration of regulatory and metabolic networks

Sophie Le Bars[1], Carito Guziolowski[1] and Jérémie Bourdon[1]

[1] Université de Nantes, Centrale Nantes, LS2N UMR 6004, Nantes, France

Corresponding Author: sophie.le-bars@univ-nantes.fr

1. Introduction

Our long-term objective is to conceive an integrated modeling framework of GRN (gene regulatory network) and metabolism. Our case study is Alzheimer's disease (AD). We firstly focused on a GRN derived from AD patients, in order to quantify how the computational predictions of 15 GRN enzymes impact brain metabolism. In a previous work [1], we illustrate the interest in using Iggy, a discrete GRN (gene regulatory network) modeling framework, over a continuous one, for computing predictions for the 15 enzymes. These predictions had a better agreement with experimental results. Iggy's benefits are that it can be used on a large-scale GRN and allow multi-perturbation without requiring a lot of parameters. However, Iggy's limitation is that its discrete enzyme predictions are not easily integrable into a metabolic model mathematical framework (linear program optimization). In this work, our objective is to transform Iggy output to allow this integration.

2. Methods

Iggy uses a sign-consistency approach, expressed as a logic program in Answer Set Programming [3]. Iggy confronts interaction graph models with observations of (signed) changes between two measured states. It discovers inconsistencies between data and network and applies minimal repair (by adding positive or negative influence over some nodes in the graph). It can also predict unobserved nodes in the network. Iggy will give as output discrete values (or coloring models): plus (+), minus (-), no-change (0), that express increasing, decreasing, or stable gene or protein expression. To quantify them, we tested different approaches such as thresholds and logic programming properties. We also studied the enumeration of all consistent coloring models (i.e. the discrete signs assigned to each node in all possible configurations) and calculated their sign's distribution.

3. Results and Perspectives

By studying all consistent coloring models we discovered that the solution space in our case study changes considerably when fixing different input-values (+,-,0) for 4 system proteins: HIF1A, CREBBP, ARNT, EP300. We explored the 3^4 possible sign configurations and observed that for some configurations the problems appeared more constrained (135 coloring models and about 20 repairs), whereas, for others, the problem is less constrained (more than 10^10 coloring models and about 9 repairs). For the more constrained cases, we showed that the full enumeration of the consistent coloring models allowed us to quantify the sign of several nodes, giving a different frequency of occurrence. However, for the less constrained cases, we were unable to list the consistent patterns. Our focus now is on 3 configurations (from the 3^4), which are close to the experimental data. For 2 of these configurations, Iggy offers the same sign as a prediction for the 15 enzymes. In order to observe a difference, we will test the impact on the predictions of using a more constrained logic in the sign constraints, such as the AND operator for signs.

References

[1] (Le Bars et al.) Le Bars S., Bourdon J., Guziolowski C. (2020) Comparing Probabilistic and Logic Programming Approaches to Predict the Effects of Enzymes in a Neurodegenerative Disease Model. In: Abate A., Petrov T., Wolf V. (eds) Computational Methods in Systems Biology. CMSB 2020. Lecture Notes in Computer Science, vol 12314. Springer, Cham. https://doi.org/10.1007/978-3-030-60327-4_8.

[2] Sven Thiele, Luca Cerone, Julio Saez-Rodriguez, Anne Siegel, Carito Guziołowski, and Steffen Klamt: Extended notions of sign consistency to relate experimental data to signaling and regulatory network topologies. BMC Bioinformatics (2015) 16:345 DOI 10.1186/s12859-015-0733-7.

[3] What Is Answer Set Programming? Vladimir Lifschitz. Third AAAI Conference on Artificial Intelligence (2008)

Improving scRNA-seq analysis in poorly-annotated genomes with matching long-read transcriptome

Nathalie LEHMANN, Rosette GOÏAME, Médine BENCHOUAIA, Kamal BOUHALI, Baptiste MIDA, Denis THIEFFRY, Xavier MORIN*, Evelyne FISCHER*, Morgane THOMAS-CHOLLIER*

Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France

Corresponding Authors: xavier.morin@bio.ens.psl.eu, evelyne.fischer@bio.ens.psl.eu, mthomas@bio.ens.psl.eu

In recent years, single-cell RNA-seq (scRNA-seq) has fostered the understanding of complex biological processes (e.g. cell differentiation, tumorigenesis) and the underlying cell heterogeneity. A crucial step in the analysis of scRNA-seq data is the generation of a count matrix summarizing the signal detected for all the genes and all the cells. The content of the count matrix is directly dependent on the annotation of the genome, as only signals covering the annotated genes or transcripts are taken into account. scRNA-seq signal obtained with 10x Genomics technology is limited to the 3' region of the transcripts, which may lead to signal loss, particularly in poorly-annotated genomes. For example, the annotation of the chicken *Gallus gallus* is not yet as complete as for the most studied organisms, such as human or mouse [1]. In order to assess to which extent such incomplete annotation affects scRNA-seq data analysis, we propose a novel approach to improve scRNA-seq analyses using long-read bulk transcriptome sequencing in matching cell samples.

We produced scRNA-seq data (10x Genomics / Illumina) from chicken cervical spinal progenitors at 66 hours of embryonic development. After quality filtering and alignment to the reference genome assembly (galGal6), up to 40% of the reads were not included in the count matrix. Visualizing the aligned reads in a genome browser revealed that significant signals fell outside of several known genes, and were thus not considered in the count matrix (as in the case of Sox2, a key marker for this study). Yet, the signal was often located in the vicinity of annotated genes. We thus concluded that loss of scRNA-seq signal was due to incomplete gene delineation, in particular at their 3' extremities.
To address this issue, we generated bulk long-read RNA-seq (MinION Oxford Nanopore Technologies, ONT) from samples matching our scRNA-seq data, in order to delineate the transcripts specific to these cells. ONT was chosen as it enables a sequencing of cDNAs from the 3' end, as for 10x Genomics / Illumina data. We exploited the long-reads data to expand the reference annotations collected from NCBI and Ensembl. We have evaluated various tools enabling the generation of gene annotations from aligned reads, such as StringTie2 [2] or Scallop [3], and selected the most appropriate ones to build our project-specific annotation. The resulting annotation combines the long-read bulk data, the scRNA-seq reads, and the reference annotation. Using this novel annotation, we were able to assign up to 87% of the reads at the genome scale, compared to 60% using only the reference annotation. We are currently evaluating the impact of this hybrid approach on the results of downstream scRNA-seq analyses.

This approach could be used to improve scRNA-seq analyses of other poorly-annotated genomes, i.e. the majority of available eukaryotic genomes, at a reasonable cost.

References

1. Kuo RI, Tseng E, Eory L, Paton IR, Archibald AL, Burt DW. Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. BMC Genomics. 2017;18: 323.
2. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 2019;20: 278.
3. Shao M, Kingsford C. Accurate assembly of transcripts through phase-preserving graph decomposition. Nat Biotechnol. 2017;35: 1167–1169.

Multimodal Single-cell Resolution Investigation of the Transcriptomic Program of Purified NK Cells in Multiple Myeloma Patients

Rüçhan Ekren[1,2,3], Marie Tosolini[1], Virginie Baylot[1,3], Nadege Carrie-Constantin[1,3], Marie-Véronique Joubert[1,3], Alexis Coulomb[1], Vera Pancaldi[1] and Ludovic Martinet[1,2,3]

[1] Cancer Research Center of Toulouse (CRCT), Institut National de la Santé et de la Recherche Médicale (INSERM) UMR 1037, Centre National de la Recherche Scientifique (CNRS), Université Paul Sabatier (UPS), Toulouse, France
[2] The Toulouse Graduate School of Cancer Ageing and Rejuvenation (CARe), 2nd Avenue Hubert Curien, 31100, Toulouse, France
[3] Institut Universitaire du Cancer, CHU Toulouse, Toulouse, France

Corresponding Authors: ruchan.ekren@inserm.fr,ludovic.martinet@inserm.fr

Multiple myeloma (MM) is a yet incurable disease characterized by the expansion of tumoral plasma B cells [1]. Natural Killer (NK) cells can recognize and kill MM cells in vitro, and we previously showed that NK cells can limit MM growth in vivo in mouse models [2] These observations suggest that harnessing NK cell activity could be a valid therapeutic strategy to improve current anti-MM treatments, especially new MM-targeting anti-CD38 antibodies, whose efficacy may rely on NK cells. Yet, data on NK cell status in MM patients are scattered, several articles show conflicting results and multiparametric analyses of NK cell phenotype and functions during MM progression are missing [2]. To address these points, we investigated the multimodal single-cell expression profile of a cohort of 10 MM patients and 10 healthy donors, extracting bone marrow and blood paired samples.

In this study, the multimodal single-cell expression profile of the samples was obtained by Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITEseq) [3]. This protocol enables the measurement of genome-wide transcriptomes (many thousands of genes) and protein expression with a selected panel of antibodies in single-cells. We linked multiple modalities of the cohort data with the Seurat v4 workflow [4]. Trajectory inference of the scRNAseq data was performed with Monocle v3 [5]. Our results revealed major changes in the NK cell populations of MM patients compared to healthy donors. Notably, we observed the expansion of NK cell clusters characterized by low levels of cytotoxicity-related genes (Prf1, CD16). These results were confirmed by flow cytometry in a retrospective cohort of 180 MM patients showing increased frequencies of CD16-/CD226- NK cells associated with a poor clinical outcome. Thus, important NK cell modifications occur during MM development, with a potential association with MM resistance to current therapy.

References

1. Camille Guillerey, Lucas Ferrari de Andrade, Christopher Chan, Slavica Vuckovic, David S. Ritchie, Leif Bergsagel, Marco Colonna, Daniel M. Andrews, Geoff R. Hill, Mark J. Smyth and Ludovic Martinet. "Immunosurveillance and therapy of multiple myeloma are CD226 dependent". *Journal of clinical investigation*, 2015, https://doi.org/10.1172/jci77181
2. Kyohei Nakamura, Mark J Smyth, Ludovic Martinet, "Cancer immunoediting and immune dysregulation in multiple myeloma". *Blood*, 2020.https://doi.org/10.1182/blood.2020006540
3. Marlon Stoeckius, Christoph Hafemeister, William Stephenson et al. "Simultaneous epitope and transcriptome measurement in single cells" *Nat Methods* 14, 865–868, 2017. https://doi.org/10.1038/nmeth.4380.
4. Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, et al "Integrated analysis of multimodal single-cell data" *bioRxiv* 2020.10.12.335331; doi: https://doi.org/10.1101/2020.10.12.335331
5. Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby et al "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells" *Nature Biotechnology*, 2014. https://doi.org/10.1038/nbt.2859

Update of ProteINSIDE: an online tool for proteome data mining

Jérémy Tournayre[1], Céline Boby[1] Matthieu Reichstadt[1] and Muriel Bonnet[1]

[1] INRAE, Université Clermont Auvergne, Vetagro Sup, UMRH, 63122, Saint-Genès-Champanelle, France
Corresponding Authors: jeremy.tournayre@inrae.fr and muriel.bonnet@inrae.fr

1. Introduction

We previously presented the ProteINSIDE tool to the scientific community [1,2]. ProteINSIDE was developed to mine results from large lists of proteins or genes, and thus to extract meaningful biological knowledge from "omics" datasets. The first version of ProteINSIDE gave 4 types of analyses or results: (1) identifiers conversion plus an overview of the biological information stored in public databases (NCBI and UniProt), (2) Gene Ontology enrichment analysis, (3) proteins that are predicted as secreted by mammalian cells, (4) protein protein interactions. Since then, we have improved this tool on many points, including an increase in the number of organisms considered, new functional enrichments (in addition to the GO ones) and the search for quantitative trait loci.

This new version of ProteINSIDE is available at the following address:
https://umrh-bioinfo.clermont.inrae.fr/ProteINSIDE_2/

2. Improvements

The workflow now uses the g:profiler API (Application Programming Interface) [3] for the conversion module (g:convert) and enrichment module (g:GOSt). The second version of ProteINSIDE analyzed lists of identifiers from more than 600 organisms rather than those from the 6 species previously targeted by the first version. Functional enrichment analysis previously focused on GO is now complemented by functional enriched analysis to find over-representation of information from several databases: (Human Proteome Atlas, Human Phenotype Ontology, Kegg, miRTarBase, Transfac, Reactome and WikiPathways).

A new functionality for protein protein interactions was added: the comparison between multiple lists experiments. For this, a user sends several lists of proteins. Then, ProteINSIDE, with the help of Psicquic [4], searches for any interactions between the proteins in these lists. The results are downloadable and can also be displayed on a network.

A complete overhaul of the tool was carried out in order to have a simplest interface and best optimization, in particular by parallelizing the calculations. Now the modules are independent of each other to allow the user to choose which one to launch, in order to access to the results as soon as possible without waiting for all the modules to finish running. For the visualization of networks, cytoscape web Flash was no longer supported and has been replaced by its Javascript version [5].

Finally, we added a fifth module, to search for quantitative trait loci using AnimalQTLdb [6] for the bovine species only.

References

1. Kaspric N, Picard B, Reichstadt M, Tournayre J, Bonnet M. ProteINSIDE to Easily Investigate Proteomics Data from Ruminants: Application to Mine Proteome of Adipose and Muscle Tissues in Bovine Foetuses. PLOS ONE. 22 mai 2015;10(5):e0128086.
2. Kaspric N, Reichstadt M, Picard B, Tournayre J, Bonnet M. Protein function easily investigated by genomics data mining using the proteINSIDE online tool. Genomics Comput Biol. 2015;1(1):e16.
3. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2 juill 2019;47(W1):W191-8.
4. Aranda B, Blankenburg H, Kerrien S, Brinkman FSL, Ceol A, Chautard E, et al. PSICQUIC and PSISCORE: accessing and scoring molecular interactions. Nat Methods. juill 2011;8(7):528-9.
5. Franz M, Lopes CT, Huck G, Dong Y, Sumer O, Bader GD. Cytoscape.js: a graph theory library for visualisation and analysis. Bioinformatics. 15 janv 2016;32(2):309-11.
6. Hu Z-L, Park CA, Reecy JM. Building a livestock genetic and genomic information knowledgebase through integrative developments of Animal QTLdb and CorrDB. Nucleic Acids Res. 8 janv 2019;47(D1):D701-10.

Ectopy: a new Python tool for prognosis biomarker discovery in cancers from omics data

Alexandre Flin[1], Florent Chuffart[1], Saadi Khochbin[1], Sophie Rousseaux[1] and Ekaterina Bourova-Flin[1]
EpiMed, Institute for Advanced Biosciences, INSERM U1209, CNRS UMR5309, University Grenoble Alpes, France

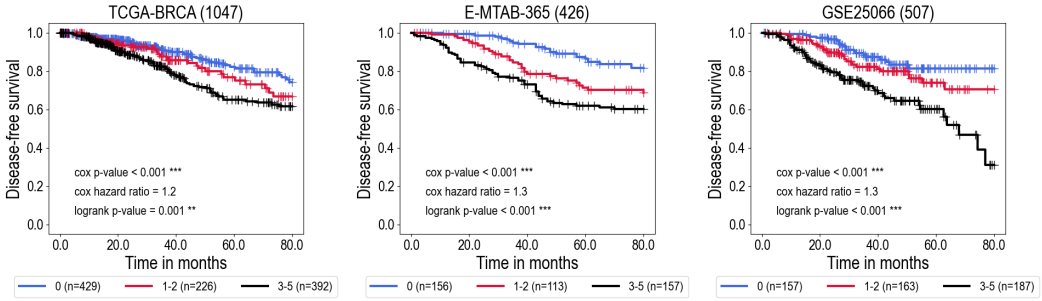
Corresponding author: ekaterina.flin@univ-grenoble-alpes.fr

1 Introduction

During the oncogenic process, the malignant transformation of cells is associated with systematic interconnected changes in the genome and the epigenome. The work of the EpiMed group at the Institute for Advanced Biosciences has demonstrated that any malignant tumor aberrantly expresses a number of genes which are normally silent in all adult non-germinal tissues and that these ectopic expressions represent a unique source of biomarkers and potential therapeutic targets. An aberrant activation of several of these genes in cancers is systematically associated with a shorter patient survival prognosis [1,2]. We present here a new open source Python tool "ectopy" [3] allowing to robustly identify candidate prognosis biomarkers based on ectopic expressions in a machine learning approach. It aims to identify the most aggressive forms of cancer.

2 Method and Results

Our biomarker discovery strategy can be applied to omics data of gene expression or gene abundance (transcriptome, proteome) to search a robust association with survival. In the first step of our method, we identify, among the tissue-specific genes, those which are aberrantly activated in the studied cancer in more than 10% of tumoral samples. In the second step, we define a robust threshold of activation for each gene by exploring a range of possible thresholds, from 10th to 90th percentile of expression levels. Each threshold is tested for its ability to discriminate between two groups of tumours, respectively of low and high expressions, corresponding to significantly different survival probabilities. The robustness of the thresholds is estimated using randomized k-fold cross-validations. We then select a number of candidate biomarkers robustly associated with survival probability to create a prognosis tool by combining them together. Finally, we stratify the tumours according to the number of gene activations among this subset of selected markers. An application of this method to define a prognosis tool in breast cancer is shown in the figure below. A panel of five prognosis biomarkers were identified using the "ectopy" software in the TCGA-BRCA training dataset and successfully validated in two independent test datasets E-MTAB-365 and GSE25066. The patients for whom none of the five genes were activated (blue line) have a significantly higher disease-free survival probability than the patients for whom one or more genes were simultaneously activated (red and black lines).



References

[1] Sophie Rousseaux, Jin Wang, and Saadi Khochbin. Cancer hallmarks sustained by ectopic activations of placenta/male germline genes. *Cell Cycle*, 12(15):2331–2332, 2013. PMID: 23856584.

[2] Jin Wang, Sophie Rousseaux, and Saadi Khochbin. Sustaining cancer through addictive ectopic gene activation. *Current Opinion in Oncology*, 26(1), 2014. PMID: 24275853.

[3] Alexandre Flin. Ectopy package, 2021. https://github.com/epimed/ectopy.

Redefining the archaea *Thermococcales* order with the biogeographic repartition and comparative genomic analysis of 104 isolates.

Violette DA CUNHA[1], Philippe OGER[2], Damien COURTINE[3], Loïs MAIGNIEN[3], Mohamed JEBBAR[3], Ghislaine MAGDELENAT[4], Valérie BARBE[4], Evelyne MARGUET[5], Patrick FORTERRE[1] and Jacques OBERTO[1*]

[1] Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France.

[2] UMR CNRS 5240 Microbiology, Adaptation & Pathogeny (MAP), Institut National des Sciences Appliquées - Bâtiment Pasteur - 11, Avenue Jean Capelle 69621 Villeurbanne, France.

[3] Université de Brest, CNRS, IFREMER, LIA1211, Laboratoire de Microbiologie des Environnements Extrêmes LM2E, IUEM, Rue Dumont d'Urville, F-29280 Plouzané, France.

[4] Génoscope, Laboratoire de Biologie Moléculaire pour l'Etude des Génomes C.E.A., Institut de Génomique - 2 rue Gaston Crémieux, EVRY, France

Corresponding Author: violette.da-cunha@iébc.paris-saclay.fr; jacques.oberto@i2bc.paris-saclay.fr

The Archaeal *Thermococcales* order comprises anaerobic hyperthermophilic organisms with some of them additionally barophilic. Based on the phylogeny of their 16S rRNA genes, *Thermococcales* have been ranked into three genera: *Pyrococcus*, *Thermococcus* and *Palaeococcus* [1]. These archaea all harbor small, fast evolving genomes and a number of integrated of self-replicating mobile genetic elements [2–4]. Due to the combined effort of different teams, several isolates are now genetically tractable and used as valid genetic models to investigate the extremophilic way of life.

With our complete genomic sequencing of 69 new *Thermococcales* genomes (55 *Thermococcus*, 12 *Pyrococcus* and 2 *Palaeococcus*), the total number of available *Thermococcales* genomes is now in excess of 100. The size, quality and completeness of this closed chromosomal sequences is making of these organisms one of the most represented archaeal order so far. The complete *Thermococcales* genomic dataset was analyzed using number of bioinformatics investigating techniques, core genome phylogeny, Average Nucleotide Identity (ANI) of the 16S or to the global genome (FastAni), the standardized microbial taxonomy approaches based on phylogeny developped for the Genome Taxonomy Database GTBD and all these analyses will be compared and linked to our biogeography analyses.

These analyses will be compared and linked to our biogeography analyses. The different approaches used in our deep comparative analysis concurred in unambiguously ranking all *Thermococcales* genomes into three genera differing significantly from the current classification based on 16S rRNA gene sequences. If the *Pyrococcus* genus remains unchanged, the extent of *Thermococcus* genus has been revisited with a number of genomes now belonging to the *Palaeococcus* genus.

The redefinition of the concepts of genus and species proposed here allowed the precise characterization of an entire archaeal order. The principles described in this work could be extended to rank accurately any groups of prokaryotic organisms for which a sufficient number of complete genomic sequences are available.

References

1. S.V. Albers, B. Siebers, The Prokaryotes: Other Major Lineages of Bacteria and The Archaea, Springer Heidelberg New York Dordrecht London, 2014.
2. Y. Zivanovic, J. Armengaud, A. Lagorce, C. Leplat, P. Guérin, M. Dutertre, V. Anthouard, P. Forterre, P. Wincker, F. Confalonieri, Genome analysis and genome-wide proteomics of Thermococcus gammatolerans, the most radioresistant organism known amongst the Archaea, Genome Biol. 10 (2009).
3. Zivanovic, P. Lopez, H. Philippe, P. Forterre, Pyrococcus genome comparison evidences chromosome shuffling-driven evolution, Nucleic Acids Res. 30 (2002)
4. M. Cossu, C. Badel, R. Catchpole, D. Gadelle, E. Marguet, V. Barbe, P. Forterre, J. Oberto, Flipping chromosomes in deep-sea archaea, PLOS Genet. 13 (2017) e1006847. https://doi.org/10.1371/journal.pgen.1006847.

Study of *de novo* gene birth in the genomes of plant-parasitic nematodes of the *Meloidogyne* genus

Mélanie MASSON, Dominique COLINET, Etienne G.J. DANCHIN

Université Côte d'Azur, INRAE, CNRS, ISA, Sophia Antipolis, France

Corresponding Author: melaniemasson@outlook.com

Genes lacking homology in other species are systematically found in newly-sequenced genomes. These so-called orphan genes include genes that diverged extensively from pre-existing ones but also genes that evolve *de novo* from DNA of non-genic origin. [1]

The objective of this study was to investigate *de novo* gene birth in the genomes of plant-parasitic nematodes of the *Meloidogyne* genus. The choice of this model is due to the availability of high quality genomes and preliminary results suggesting the presence of a large number of orphan genes. Seven species of this genus were studied: *M. arenaria*, *M. enterolobii*, *M. floridensis*, *M. incognita*, *M. graminicola*, *M. hapla*, and *M. javanica*.

From the 348,320 protein sequences predicted in the seven *Meloidogyne* genomes, 115,713 lacked homology in any other nematode and were thus specific to this genus, according to a Nematoda phylum-wide OrthoFinder analysis [2]. Using RSEM on RNA-seq data available for four of the *Meloidogyne* species; we confirmed that 43,774 proteins (or their homologous sequences) had corresponding genes supported by transcriptomic data. These 43,774 proteins were all considered as encoded by potential 'functional' orphan genes. A Diamond homology search revealed that 90% of these proteins had no further homology in the NCBI's nr database suggesting these 39,507 sequences are encoded by 'true' orphan genes. A maximum parsimony ancestral state reconstruction with Mesquite was then performed to determine the timing of appearance of each orphan gene in the *Meloidogyne* phylogeny. A large proportion of the identified orphan genes emerged in a common ancestor of the species *M. arenaria*, *M. enterolobii*, *M. floridensis*, *M. javanica* and *M. incognita*, all belonging to the *Meloidogyne* clade 1. We then searched for potential *de novo* genes specific to clade 1 using the closely-related *M. hapla* (clade 2) and *M. graminicola* (clade 3) as outgroup species. The protein sequences coded by the 32,273 clade 1 orphan genes were aligned on the genomes of *M. hapla* and *M. graminicola* using the Exonerate tool [3]. A total of 8,652 proteins aligned with at least 30% coverage and 30% identity (according to species specific thresholds previously determined) on the genomes of *M. hapla* and / or *M. graminicola*, while no gene was predicted at these loci. These clade 1-specific genes could thus be traced back to non-genic regions in these genomes and probably represent *de novo* gene birth events. Analyses were done to detect the events that allowed the transition from a non-genic sequence to a *de novo* gene (elimination of STOP codons, modification of splicing sites …).

Considering their number, *de novo* gene birth events have had a significant impact in the evolution of the *Meloidogyne* genus genomes. One might wonder about the biological impact of *de novo* gene birth. Insofar as they are specific to *Meloidogyne*, a part of these genes could be involved in plant parasitism mechanisms.

[1] van Oss et al., De novo gene birth. *PLoS Genetics*, 2019, *15*(5)

[2] Emms DM et al., OrthoFinder: Phylogenetic orthology inference for comparative genomics. Genome Biology, 2019, 20(1)

[3] Slater et al., Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005, 6 (31)

RSAT peak-motifs2: motif discovery in genome-wide regulatory genomics datasets

Walter SANTANA-GARCIA[1], Alejandra MEDINA-RIVERA[2], Jacques VAN HELDEN[3], Denis THIEFFRY[1] and Morgane THOMAS-CHOLLIER[1]

[1] Institute of Biology of ENS (IBENS), Department of biology, École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France

[2] International Laboratory for Human Genome Research – National Autonomous University of Mexico, Blvd. Juriquilla 3001, 76230, Querétaro, México

[3] Theories and approaches for Genomic Complexity (TAGC) – Aix-Marseille University, 163 Avenue de Luminity, 13288, Marseille, France

Corresponding Author: mthomas@bio.ens.psl.eu

Studies in regulatory genomics are mostly based on datasets produced by *in vivo* techniques at the genome-wide scale, such as ChIP-seq, ATAC-seq, CLIP-seq, etc. These epigenomic experiments yield thousands of genomic regions with sizes (typically >200 bp) larger than transcription factor (TF) binding sites (TFBSs, typically 5-30 bps). Thus, for such datasets that are far from the base-pair resolution, motif analysis is required to uncover the binding specificity of the profiled TF, to predict their potential cofactors, and to infer the precise locations of the TFBSs within these larger regions.

Nine years ago, we introduced *peak-motifs* [1,2], a time-efficient computational workflow designed for the analysis of ChIP-seq and similar epigenomic datasets, *e.g.* ATAC-seq, CLIP-seq, etc. This workflow is included in the user-friendly Regulatory Sequence Analysis Tools (RSAT, www.rsat.eu) [3] software suite. The RSAT *peak-motifs* performs *de novo* motif discovery from the provided sequences, compares the discovered motifs with user-provided or public motif databases, predicts TFBSs locations and creates custom tracks of the putative sites for UCSC browser visualization. Furthermore, *peak-motifs* can be used for differential motif analysis between two datasets to discover dataset-specific motifs.

Encouraged by the wide use of *peak-motifs*, we are now introducing *peak-motifs2*, a novel major release of RSAT *peak-motifs*. In *peak-motifs2,* the input web-form was re-designed to enhance user navigation and accessibility. In addition to peak sequences, a BED file can now be provided as an input, and users can choose to automatically retrieve the corresponding genomic sequences from an organism available at UCSC or locally-installed in RSAT. Low-complexity regions can be masked to prevent the generation of spurious motifs. Clustering of discovered motifs is performed to produce non-redundant motif collections. Positional distributions of predicted TFBSs in the sequences are computed to assist in ranking the discovered motifs. Finally, the HTML result report now takes the form of a dynamic dashboard to enhance biological interpretability of discovered motifs and to accommodate the new motif analysis results and features. The tool *peak-motifs2* is still in development, available at http://rsat-tagc.univ-mrs.fr/rsat/. The RSAT suite can be downloaded from https://github.com/rsa-tools/.

Acknowledgements

The development of *peak-motifs2* was supported by the *ITMO Cancer CID ModICeD project* and the Institut Universitaire de France.

References

1. Morgane Thomas-Chollier, et al. RSAT peak-motifs: motif-analysis in full-size ChIP-seq datasets. *Nucleic Acids Research*, (40):e31, 2012.
2. Morgane Thomas-Chollier, et al. A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. *Nature Protocols*, (7):1551-1568, 2012.
3. Nga Thi Thuy Nguyen, et al. RSAT 2018: regulatory sequence analysis tools 20th anniversary. *Nucleic Acids Research*, (46):W209–W214, 2018.

Tools for building and analyzing plant pangenomes

Ali Cuhadar[1,2], Marion Dupouy[2] and Clément Agret[3]
[1] Université de Bordeaux, 33076, Bordeaux, France
[2] ERA-BIO-IT, 31700, Mondonville, France
[3] CRIStAL, Centre de Recherche en Informatique Signal et Automatique de Lille, 59000, Lille, France

Corresponding author: ali.cuhadar@era-bio-it.com

1 Context

Nowadays, new sequencing technologies give access to a great and growing quantity of genomic data. New methods and tools have to be developed to process and analyze this much information effectively .Those new technologies also provoke a paradigm switch in our approach of analyzing genome. Instead of comparing data to an unique reference genome, we can now compare data to a pangenome. [1] Moreover, latest pangenomic studies showed the limit of the traditional reference genome approach used until now for agronomic purposes. The concept of pangenome has been introduce in 2005 by Tettelin et al [2] in a bacterial study, he defined the pangenome by categorizing three parts:
- Core genome for genes present in all strain,
- dispensable (or shell) -genome for genes absent in at least one strain
- and unique (or cloud) genome for genes present in only one strain.

The expanded definition now considers that a pangenome is a set of genomics sequences to be analyzed or served as a compliant reference and which integrates information on genomics structural variations between different varieties, groups of individuals or individuals of the same species. Studies using a pangenomic approach mainly use pipeline of tools[3] in order to build and analyze pangenome, Stand-alone tools are in active development during last years but the size and the complexity of plant genomes make it challenging.[4]

2 Purpose

Facing the diversity of pipeline approaches and stand-alone tools, often designed for microbial pangenomics, we are exploring multiple solutions through benchmarking. We expect to propose to plant molecular breeders and plant bioanalysts the optimal solution(s) to shift their analyzes from reference genomes to pangenomes in order to support their work on crop varietal improvement and evaluation.

Acknowledgements

ERA-Bio-IT is financed by Euralis, RAGT and Arvalis.

References

[1] Aamir W. Khan, Vanika Garg, Manish Roorkiwal, Agnieszka A. Golicz, David Edwards, and Rajeev K. Varshney. Super-pangenome by integrating the wild side of a species for accelerated crop improvement. *Trends in Plant Science*, 25(2):148–158, 2020.

[2] Tettelin and all. Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: Implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*, 102(39):13950–13955, 2005.

[3] Zhiqiang Hu, Chaochun Wei, and Zhikang Li. *Computational Strategies for Eukaryotic Pangenome Analyses*, pages 293–307. Springer International Publishing, Cham, 2020.

[4] The Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics*, 19(1):118–135, 10 2016.

Title: Cont-ID: detection of cross-contamination in metagenomic data

Johan Rollin[1,2] / Wei Rong[1] / Sébastien Massart[1]

1.Plant Pathology Laboratory, TERRA, Gembloux Agro-Bio Tech, University of Liège, Gembloux, Belgium
2.DNAVision, Gosselies, Belgium

Keywords: virus, bioinformatic, HTS, cross-contamination

Plant viruses cause a large part of the emerging plant diseases and pose a great threat to agricultural crops worldwide. INEXTVIR Marie-Curie Training Network proposes the use of High-Throughput Sequencing (HTS) technologies for studying the virome of agricultural crops across Europe.
In the frame of that project, we explored the issue of cross-contamination in viruses in metagenomic data. With the improvement of bioinformatic methods to detect viruses in these data, we now have some cases where tools (kraken, kaiju …) can rightfully predict something that is obviously wrong from a biologic perspective because of cross-contamination. We used controlled biological sequences generated in our own laboratory to test the detection of viruses (originating from cross-contamination or not) with known indexing status. Then, based on this dataset, we selected alternative strategies to rightfully predict the status of virus detection in these samples: infection or contamination.

We present Cont-ID, a method designed to check for cross-contamination in viruses previously identified in metagenomic datasets. It relies on a simple principle, every sample in a sequencing batch should have been processed the same way with at least one alien control. Cont-ID is an open-source python script that will be available (ongoing publication). It uses a decision tree to classify every species prediction on every sample of the sequencing batch into (true) infection, (cross) contamination. The method used seems to work regardless of the host (fruit tree, grass, human, animal …) or the sequencing technology used (dsRNA, Total-RNA, Small-RNA ...)

Large structural variant detection on Pearl millet genome using Bionano optical mapping

Marine SALSON[1,3], Julie ORJUELA[1,2], Cédric MARIAC[1], Christine TRANCHANT-DUBREUIL[1], Yves VIGOUROUX[1], and Cécile BERTHOULY-SALAZAR[1]

[1]DIADE Unit, Univ Montpellier, CIRAD, IRD, 911 Avenue Agropolis, F-34394, Montpellier Cedex 5, France

[2]PHIM Plant Health Institute, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France

[3]Master Sciences and Digital Technology for Healthcare Specialty Bioinformatics, Knowledge, Data, University of Montpellier, France

Corresponding Author: marine.salson@etu.umontpellier.fr

Pearl millet is a cereal cultivated in sub-Saharan Africa, India and South Asia and is the staple food for more than 90 million farmers. The richness of its grains in protein, micronutrients and fibers makes Millet a particularly interesting agronomic crop. This plant species is also adapted to arid soils and high temperatures. Thus, research projects aiming to improve this crop's productivity and resilience may help society to adapt to climate changes and contribute to greater food security.

The Pearl millet reference genome is composed of 7 chromosomes for ~ 1.79 Gb, with a percentage of repetitive DNA above 80% [1]. Genomic association studies conducted previously in our laboratory led to the identification of a large structural variation (SV) concerning two thirds of chromosome 3 and spanning ~ 200 Mbases. Self fertilization studies revealed that recombination is suppressed in this region. Given that inversions prevent recombination, the hypothesis has been made that the SV may be a large chromosomal inversion. Moreover, while this SV appears to be lethal in homozygous state, plants carrying the hypothetical inversion on only one copy of chromosome 3 display advantageous and adaptive traits.

In order to characterize this large structural variant, optical maps from several genotypes were generated with the Saphy System from Bionano Genomics, some of them carrying the potential inversion on one copy of chromosome 3. Optical mapping is a technique which enables identification of short sequence motifs on long DNA molecules ranging from 0.1 to 2 Mbases. The average molecule length of optical maps (~ 225 Kbases) being greater than the read lengths from both short and long read sequencing, they can span genomic regions usually difficult to study with sequencing data. Assembly of these raw molecules leads to the generation of contiguous and up to hundred megabases scale representations of genomes. Thus, optical maps can be used to improve genome assemblies, and especially to assist the scaffolding process of sequence contigs. This technique can also complement sequencing technologies for detecting large and complex genomic SVs [2].

We used two alignment tools, Refaligner from Bionano and OMBlast [3], to anchor large optical maps to the Pearl millet reference genome. Our aim is then to detect a potential large chromosomal inversion comparing the optical maps from different genotypes aligned to chromosome 3. Aligning optical maps is however different from aligning genomic sequences, and very few tools are currently available for data quality control and analysis, as well as for detecting SVs. Making use of these promising data is therefore not an easy task, especially with the complexity of plant genomes.

References

1. R. K. Varshney et al, Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments. *Nature biotechnology, 35(10) :969–976, Oct 2017*
2. Yuxuan Yuan et al, Advances in optical mapping for genomic research, *Computational and Structural Biotechnology Journal, Volume 18, :2051–2062, 2020*
3. Alden King-Yung Leung et al, OMBlast : alignment tool for optical mapping using a seed-and-extend approach, . *Bioinformatics, 33(3) :311–319, Feb 2017*

Detection of horizontal gene transfers of non-metazoan origin in the whitefly *Bemisia tabaci*

Mathilde BOYER[1], Georgios KOUTSOVOULOS[1], Arthur PERE[1], Etienne G. J. DANCHIN[1] and Dominique COLINET[1]

[1] INRAE, Université Côte d'Azur, CNRS, Institut Sophia Agrobiotech, 400 route des Chappes, 06903, Sophia-Antipolis, France
Corresponding Author: mathilde.boyer2@etu.uca.fr

Abstract

Horizontal gene transfer (HGT) is the transmission of genetic material by other ways than vertical transfer which is the natural inheritance of genes from parents to their offspring.
Although the phenomenon has been known for long between bacteria as a major mechanism of evolution and adaptation (e.g. antibiotic resistance), the occurrence of HGT in eukaryotes is still a matter of debate.
In the last few years, several publications have described genes of bacterial or fungal origin [1] but also of plant origin [2,3] acquired by horizontal transfer in *Bemisia tabaci*, one of the most globally significant crop pests. These were the first HGT events described from plants to animals [2,3].
The goal of our research was first to validate these results using a different approach. In a first step, the Alienness tool [4] was used to identify putative HGT events from the *B. tabaci* MEAM1 strain proteome available on the Whitefly Genome Database [1]. Alienness calculates an Alien Index score (AI) based on the E-value gap between the best non-metazoan and the best metazoan blast hits. We identified 752 genes with an AI score above 0, among which 157 were possible HGT (0<AI<14), 405 likely HGT (AI>14 and identity percent under 70), and 190 likely contaminations (AI>0 and identity percent above 70).
Then, we used AvP (Alienness vs Predictor) to study the phylogeny of each case and select the most likely HGT events. Among the putative HGT predicted by Alienness, 495 were classified as 'Strong HGT support' by AvP and, among them, 152 were of Viridiplantae origin. These included the genes described in the literature [2,3], among which BtPMaT1 encoding a phenolic glucoside malonyl transferase playing an important role in xenobiotic compounds detoxication. We also identified a pectin methylesterase coding gene of plant origin which constitutes a huge advantage for a phytophagous specie like *B. tabaci* because it is involved in the pecto-cellulosic plant cell wall degradation.
To strengthen HGT hypotheses and rule out the possibility of contamination, we checked the presence of the genes of interest in the genomes of two other *B. tabaci* strains using OrthoFinder[5]; which allowed classifying the query genes in orthogroups. As an outgroup species, we chose the pea aphid *Acyrthocyphon pisum*. All our putative HGT events were shared by at least two *B. tabaci* strains, discarding the contamination hypothesis.
Finally, we used GoFuncR [6] to identify significantly enriched functions in the set of proteins acquired by HGT according to AvP. Among the interesting enriched GO terms; we found the term 'transferase activity' (11 proteins including the BtPMaT1 gene) and 'oxydoreductase activity' both related to detoxication of xenobiotic compounds.
Overall, our analyses confirmed HGT events of plant origin in the genome of *B. tabaci* but also allowed identifying new HGT cases from different non-animal origins that will deserve further studies in the future.

References

[1] Wenbo Chen *et al.* The draft genome of whitefly *Bemisia tabaci* MEAM1, a global crop pest, provides novel insights into virus transmission, host adaptation, and insecticide resistance. *BMC Biol*, (14):110, 2016.
[2] Walter J. Lapadula *et al*. Whitefly genomes contain ribotoxin coding genes acquired from plants. *Sci Rep,* (10)**:**15503, 2020.
[3] Jixing Xia *et al*. Whitefly hijacks a plant detoxification gene that neutralizes plant toxins. *Cell*, (184):7, 2021.
[4] Corinne Rancurel *et al*. Alienness: Rapid Detection of Candidate Horizontal Gene Transfers across the Tree of Life. *Genes*, (8):248, 2017.
[5] David M. Emms et al. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*, (16):157, 2015.
[6] Steffi Grote. GOfuncR: Gene ontology enrichment using FUNC. *R package version*, 1(0), 2018.

DNA methylation repression of transcriptional activity and its reactivation in a demethylating context

F. Pittion, E. Bourova-Flin, S. Khochbin, S. Rousseaux and F. Chuffart
EpiMed group, Institute for Advanced Biosciences (INSERM U1209 - CNRS UMR5309 - UGA), Grenoble, France

Corresponding author: florent.chuffart@univ-grenoble-alpes.fr

DNA methylation is one of several epigenetic mechanisms that cells use to control gene expression. Since many cancers have low levels of DNA methylation compared to most mammalian somatic tissues [1,2], we want to investigate the implication of DNA methylation as a key regulator of gene expression contributing to cancer development [3,4]. The first objective of this study is to group genes according to the three following characteristics: i) CpG density of their promoter region, ii) level of methylation of the promoter region in normal tissues, iii) level of expression in a demethylating context. The aim is to define a set of genes whose promoter region is CpG-rich and widely methylated in most tissues and whose expression is up-regulated in a demethylating context. The second objective of this study is to investigate whether this subset of genes could be enriched in cancer aggressiveness biomarkers. To investigate the first question, we are developing `dmethr` (http://github.com/fchuffar/dmethr/), a three-step dedicated pipeline, and apply it on publicly available datasets.

1^{st} step: genomic regions surrounding the TSS of RefSeq genes (TSS +/- 2.5kb) are divided into 100bp-length bins. We compute the CpG density of the defined bins and obtain a matrix of 26067 genes and 25 bins. From this matrix, a PCA on genes separates **15928 CpG-rich and a 10139 CpG-poor associated genes**.

2^{nd} Second step: CpG-rich associated genes are clustered according to their methylation status in a publicly available dataset of normal tissues (GSE56515, GSE48472, GSE64096, GSE50192, GSE31848, GSE73375). We obtain a matrix of 166894 methylation probes and 370 samples. The 166894 methylation probes are mapped to the genomic regions surrounding the TSS of 14901 CpG-rich genes. The 370 samples are also mapped to the 30 healthy tissues. The initial methylation signal is then reduced by successively computing the average methylation by tissue and by gene. A matrix of average methylation values is obtained for 14901 genes and 30 tissues. A hierarchical clustering on this matrix reveals a set of **859 CpG-rich genes widely methylated in healthy tissues**.

3^{rd} step: A differential analysis of gene expression in control *vs.* demethylating context is performed. Two public expression datasets are selected for this analysis: i) RNA-seq from wild type (control) versus DNMT double KO HCT116 cells (colon cancer cells, GSE45332), ii) transcriptomic microarray from control and 5-azacytidine treated lung cancer cell lines (GSE5816). A Gene Set Enrichment Analysis [5] of these transcriptomic signatures confirms that they are significantly enriched in CpG-rich genes widely methylated in healthy tissues. Using a cut-off of adjusted p-value > 0.05 and foldchange > 2 we obtain **189 CpG-rich genes widely methylated in healthy tissues and upregulated in a demethylating context**. We now plan to look for the association between the expression of these genes and prognosis in The Cancer Genome Atlas dataset [6].

References

[1] Akpéli V. Nordor, Djamel Nehar-Belaid, Sophie Richon, et al. The early pregnancy placenta foreshadows DNA methylation alterations of solid tumors. *Epigenetics*, 12(9):793–803, 2017. PMID: 28678605.
[2] B. Zhang, M. Y. Kim, G. Elliot, et al. Human placental cytotrophoblast epigenome dynamics over gestation and alterations in placental disease. *Dev Cell*, 56(9):1238–1252, May 2021.
[3] Peter A. Jones. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484–492, Jul 2012.
[4] Mai Shi, Stephen Kwok-Wing Tsui, Hao Wu, and Yingying Wei. Pan-cancer analysis of differential DNA methylation patterns. *BMC Medical Genomics*, 13(10):154, Oct 2020.
[5] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.
[6] https://portal.gdc.cancer.gov.

BiSePS : Bisuflite Sequencing Processing Software

Skander Hatira[1], Jean-Marc Celton[1], Sandrine Balzergue[1], Claudine Landès[1] and François Laurens[1]

[1] IRHS-UMR1345, Université d'Angers, INRAE, Institut Agro, SFR 4207 QuaSaV, 49070, Beaucouzé, France

Corresponding Author: skander.hatira@inrae.fr

1 Introduction

DNA methylation differences among individuals can be identified through Whole Genome Bisulfite Sequencing. To analyze the substantial amount of data generated by this technique, several tools have been developed, yet most of them operate in a command line fashion, require some expertise and/or cover only a part of the whole process. The purpose of this work is to provide a user-friendly tool that can analyze data in bulk and provide reproducible and easily accessible results.

2 Methods

The snakemake [1] workflow relies on Trimmomatic [2] for reads quality control and adapter trimming. It prepares the genome by indexing it with Bismark [3] which also takes care of alignment and generates methylation reports (CX-reports) and other useful files that can later on be viewed in a self hosted Jbrowse2 instance [4]. We also convert methylation calls to CGmaps as per MethGET's [5] specifications for later use. We use DMRcaller [6] to identify Differentially Methylated Regions and some in-house scripts to correlate DMRs with close genes. Lastly, overall quality statistics and reports are generated by FastqQC [7] and Bismark and then merged into a MultiQC [8] report. A local MongoDB is deployed upon installation and keeps track of all analyses, which lets users directly access their data, and view configuration profiles and other important metrics.

3 Results

We present BiSePS, a Snakemake [1] Pipeline wrapped inside an Electron Desktop Application, compatible with Linux, Windows and Mac-OS that tackles the DMR identification task from A to Z.

Acknowledgments

We would like to thank teams VALEMA and BIDefI of IRHS for their support during the development of this tool. We are also grateful for Genouest Bioinformatics for granting us access to their cluster for validation. This research has been co-funded by the Community Plant Variety Office based on the Grant Agreement No 7513139 whom we also thank.

References

[1] Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., Köster, J., 2021. Sustainable data analysis with Snakemake. F1000Res 10, 33.

[2] Bolger, A. M., Lohse, M., & Usadel, B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, 2014.

[3] Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011.

[4] Buels, Robert, et al. JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biology 2016.

[5] Teng, CS., Wu, BH., Yen, MR. et al. MethGET: web-based bioinformatics software for correlating genome-wide DNA methylation and gene expression. BMC Genomics, 2020.

[6] Catoni, Marco, Tsang, MF J, Greco, P A, Zabet, Radu N. DMRcaller: a versatile R/Bioconductor package for detection and visualization of differentially methylated regions in CpG and non-CpG contexts. Nucleic Acids Research , 2018.

[7] Andrews S. FastQC: a quality control tool for high throughput sequence data, 2010.

[8] Philip Ewels, Måns Magnusson, Sverker Lundin, Max Käller, MultiQC: summarize analysis results for multiple tools and samples in a single report, Pages 3047–3048, Bioinformatics, 2016.

Data-Mining of GTEx Data for identification of potential regulators of Alternative Splicing

Tiffany Yen Kway [1], Yann Audic [1]

[1] Institute of Genetic and Development CNRS, IGDR - UMR 6290, University of Rennes, 35043 Rennes, France.

Corresponding Author: yann.audic@univ-rennes1.fr

In Eukaryotes, RNA splicing is a process by which introns are excised from pre-mRNAs and exons are joined together to form mature mRNAs. Through the selection of distinct 3' and 5' splice sites, alternative splicing can produce different mRNAs originating from the same gene. This process dramatically expand the proteome and it is now estimated that about 95% of human multiexon genes undergo alternative splicing [1].

By generating mRNAs with different coding sequences, alternative splicing complexify the proteome and is central in the development of tissue-specific gene expression programs. Muscle, brain, epithelial development for example, dramatically rely on controlled programs of alternative splicing to produce proteins that are almost unique to these tissues. Conversely, misregulation of alternative splicing may lead to many diverse pathological consequences [2].

The regulation of alternative splicing is highly dependent upon RNA binding proteins (RBPs) that through specific association to regulatory elements located in the pre-messenger RNA sequence dictate the splicing outcome [3]. The identification of these splicing regulators and their RNA targets is therefore central for the understanding of the mechanisms of gene expression in normal and pathological context. Hundreds of RBPs are expressed from the human genome and despite the existence of powerful biochemical methods it is a daunting task to experimentally identify specific RNA/RBP interactors,

From RNASeq data, alternative splicing is most precisely defined by the existence of reads that overlap the exon/exon junctions created during the splicing process. Quantification of such junction reads may serve as a base to quantify each splicing event in the cell or tissue. The GTEX portal [4] offers thousands of human RNAseq data originating from 55 tissues with such junction reads information available. The portal also offer gene expression data. By combining the analysis of junction usage and RBP expression among thousands of GTEX samples from diverse tissue origin, we aim to identify RBPs that are the most correlated to specific alternative splicing events of biological importance. As a proof of concept, we focused on the analysis of alternative splicing events with experimentally validated RBPs regulators to assess the results of the correlation analysis.

We envisage that this correlation analysis will allow use to prioritize the RNA binding proteins potentially involved in the regulation of specific alternative splicing event before we experimentally validate the causal relation between RBPs and specific alternative splicing events.

References

[1] Pan, Qun, Ofer Shai, Leo J. Lee, Brendan J. Frey, et Benjamin J. Blencowe. « Deep Surveying of Alternative Splicing Complexity in the Human Transcriptome by High-Throughput Sequencing ». *Nature Genetics* 40, nº 12 (décembre 2008): 1413-15.

[2] Jiang, Wei, et Liang Chen. « Alternative Splicing: Human Disease and Quantitative Analysis from High-Throughput Sequencing ». *Computational and Structural Biotechnology Journal* 19 (1 janvier 2021): 183-95.

[3] Chen, Mo, et James L. Manley. « Mechanisms of Alternative Splicing Regulation: Insights from Molecular and Genomics Approaches ». *Nature Reviews. Molecular Cell Biology* 10, nº 11 (novembre 2009): 741-54.

A comparative study of ancient DNA kinship estimation methods using pedigree simulations.

Maël Lefeuvre[1, 2 *], Céline Bon[1 *], Marie-Claude Marsolier-Kergoat[1, 3] and Aline Thomas[1]

[1] UMR7206 Éco-anthropologie – Muséum National d'Histoire Naturelle, 17 Place du Trocadéro, 75016, Paris, France

[2] Plateforme Bioinformatique – Institut de biologie de l'Ecole Normale Supérieure, 46 Rue d'Ulm, 75005, Paris, France

[3] Service de Biologie Intégrative et Génétique Moléculaire – I2BC/UMR9198 – CEA/DRF, 1 Avenue de la Terrasse, 91190, Gif-sur-Yvette, France

Corresponding Authors: mael.lefeuvre@bio.ens.psl.eu, celine.bon@mnhn.fr

Advances in next-generation sequencing techniques have in the last decade revolutionized the field of archaeogenetics by allowing the retrieval of whole genome sequencing data from ancient remains in a somewhat reliable manner. Hence, increasing attempts to unveil genetically related individuals among ancient burial grounds, charnels and necropolis can now be found scattered around scientific literature. Conjointly, a handful of recently published statistical methods now claim their ability to accurately estimate genetic relatedness between past individuals, while making do with the characteristic patterns of post-mortem damage and the extremely low sequencing depths typically recovered from ancient DNA. However, while gaining a firm grasp on the limits of these novel methods is a paramount precautionary step to safely interpret their results, a formal comparison regarding their respective predictive power and biases has to our knowledge never been published in literature.

In light of this, we implemented a scalable and reproducible Snakemake pipeline that is aimed at conducting a standardized comparative analysis between three published ancient DNA genetic relatedness estimation methods: READ, TKRelated and Grups [1,2,3]. Using Ped-sim [4], our pipeline carries out whole-genome pedigree simulations starting from randomly selected present-day individuals of the CEU population [5]. The resulting data is then artificially decayed to mimic ancient DNA sequencing data using Gargammel [6]. Following the current best practices, our pipeline finally performs data pre-processing and attempts to reconstruct genetic relatedness within each simulated pedigree, using our candidate methods.

To unveil the predictive biases that these methods may have and delineate the minimal required conditions for their proper usage, these simulations were performed under five scenarios of average sequencing depths, ranging from 0.01 to 1X (n=50), thus enabling us to obtain comparable estimates of the sensitivity, specificity and general classification performance of each method. Our study not only unveils a lack of specificity from TKRelated and Grups at such low sequencing depths, but equally demonstrates that estimating genetic relatedness between ancient individual remains a feasible prospect when using READ, with as few as 1700 overlapping SNPs and with a negligible risk (<1%) of obtaining false positive results.

References

1 Kuhn, J. M. M., Jakobsson, M. & Günther, T. Estimating genetic kin relationships in prehistoric populations. *PLOS ONE* **13**, e0195491 (2018).

2 Fernandes, D. *et al.* The Identification of a 1916 Irish Rebel: New Approach for Estimating Relatedness From Low Coverage Homozygous Genomes. *Sci. Rep.* **7**, 1–10 (2017).

3 Martin, M. D., Jay, F., Castellano, S. & Slatkin, M. Determination of genetic relatedness from low-coverage human genome sequences using pedigree simulations. *Mol. Ecol.* **26**, 4145–4157 (2017).

4 Caballero, M. *et al.* Crossover interference and sex-specific genetic maps shape identical by descent sharing in close relatives. *PLoS Genet.* 15, e1007979 (2019).

5 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* 526, 68–74 (2015).

6 Renaud, G., Hanghøj, K., Willerslev, E. & Orlando, L. gargammel: a sequence simulator for ancient DNA. *Bioinforma. Oxf. Engl.* 33, 577–579 (2017).

Portability and option extensions for Phos2Net, a tool for network reconstruction and pathway extraction based on phosphoproteomic data

Pauline Marie[1,2], Marion Buffard[1,3] and Ovidiu Radulescu[1]
[1] LPHI, Université de Montpellier, CNRS, F-34095 Montpellier, France;
[2] Master of Sciences and Digital Technologies for Healthcare Specialty Bioinformatics, Knowledge, Data, Montpellier University, 34090, Montpellier, France
[3] IRCM, Université de Montpellier, ICM, INSERM, F-34298 Montpellier, France

Corresponding author: pauline.marie@etu.umontpellier.fr

Protein phosphorylation acts as an efficient switch controlling deregulated key signalling pathways in cancer. In computational biology, prior knowledge networks can be extracted from pathway oriented interaction databases by using pathway enrichment methods. But this may promote well-known proteins and disadvantage less-studied proteins. Furthermore, prior networks contain redundant interactions. To overcome these issues, our team developed Phos2Net, a bioinformatics tool for identifying paths that connect signalling proteins to their targets in specific contexts [1,2]. Phos2Net suggests potential effective pathways that could explain cancer cell phenotypes such as proliferation or motility. The computational pipeline is publicly available (http://doi.org/10.5281/zenodo.3333687).

This tool starts by building the prior knowledge network, containing all the pathways enriched in differentially phosphorylated targets, using KEGG and Pathways Commons databases. Secondly, it relies on two graph theory algorithms (random walk with restart method and Dijkstra's shortest path algorithm) to select biologically relevant paths in this network. The potency of the tool was tested on various phosphoproteomic data (involving well-known tyrosine kinase SYK or PIK3CA kinase and unfamiliar SRMS kinase) [1,2].

As the tool was only available on Linux operating system, the first aim of this work is to extend the portability of the tool by the creation of an executable program working on Windows. This requires the translation into Python language of parts of the original code written in others languages. More specifically, we exploite SciPy and NumPy libraries. We hope this development will offer a better visibility to this tool. The second aim is to integrate a new option based on Cancer Cell Line Encyclopedia [3] in order to take in account the differences between various cancer cell lines in terms of presence or absence of network proteins. We modify the interface to allow users to select the cell line they are interested in. This option indicates the proteins present in the network which are not been detected as expressed in the selected cancer cell line from [3].

References

[1] Aurélien Naldi, Romain M. Larive, Urszula Czerwinska, Serge Urbach, Philippe Montcourrier, Christian Roy, Jérôme Solassol, Gilles Freiss, Peter J. Coopman, and Ovidiu Radulescu. Reconstruction and signal propagation analysis of the Syk signaling network in breast cancer cells. *PLoS computational biology*, 13(3):e1005432, 2017.

[2] Marion Buffard, Aurélien Naldi, Ovidiu Radulescu, Peter J. Coopman, Romain M. Larive, and Gilles Freiss. Network Reconstruction and Significant Pathway Extraction Using Phosphoproteomic Data from Cancer Cells. *Proteomics*, 19(21-22):e1800450, 2019.

[3] David P. Nusinow, John Szpyt, Mahmoud Ghandi, Christopher M. Rose, E. Robert McDonald, Marian Kalocsay, Judit Jané-Valbuena, Ellen Gelfand, Devin K. Schweppe, Mark Jedrychowski, Javad Golji, Dale A. Porter, Tomas Rejtar, Y. Karen Wang, Gregory V. Kryukov, Frank Stegmeier, Brian K. Erickson, Levi A. Garraway, William R. Sellers, and Steven P. Gygi. Quantitative Proteomics of the Cancer Cell Line Encyclopedia. *Cell*, 180(2):387–402.e16, January 2020.

Recruitment and functional analysis of the genus *Zobellia* in marine metagenomes

Mael GARNIER[1], Tangy GENTHON[2], Lison REBOUL[2], Francois THOMAS[3] and Erwan CORRE[1].

1 CNRS - Sorbonne Universite´ - Plateforme ABIMS - Station Biologique de Roscoff, Place Georges Teissier, 29680, Roscoff, France

2 Sorbonne Universite´ - Licence Biologie, Modélisation et Analyses de données - Station Biologique de Roscoff, Place Georges Teissier, 29680, Roscoff, France

3 CNRS - Sorbonne Universite´ - UMR8227 - Station Biologique de Roscoff, Place Georges Teissier, 29680, Roscoff, France

Corresponding Author: erwan.corre@sb-roscoff.fr

The recycling of macroalgal biomass influences the functioning of coastal ecosystems. It relies heavily on pioneer bacteria capable of attacking intact algal tissues and releasing degradation products into the water column. In this project, we explore the presence of some of these pioneer bacteria of the genus *Zobellia* in marine, coastal or alga-associated metagenomes.

Metagenomic libraries found in public databases MAGnify or IMG/G with taxonomic or functional match to *Zobellia* were initially selected. Reads associated to the genus *Zobellia* were detected using Kraken2 software, and the relative abundance of *Zobellia* in each metagenome was estimated with Bracken. Reads were then assembled (metaSPAdes) and annotated (Prokka) to verify the presence of this bacterial genus and study its functions. This bioinformatic methodology has proven its effectiveness by revealing the prevalence of bacteria on the surface of macroalgae, highlighting the specialization of the flavobacterium for algal degradation.

Analyses were conducted by several Bachelor students during their tutored or internships projects in close collaboration with the ABiMS IFB platform and local research team.

Characterization of the genomic diversity and gene content of a lactobacilli collection

Romane Junker[1], Victoria Chuat[2], Florence Valence[2], Michel-Yves Mistou[1] and Hélène Chiapello[1]

[1] Université Paris-Saclay, INRAE, MaIAGE, Domaine de Vilvert, 78350, Jouy-en-Josas, France
[2] INRAE, Agrocampus Ouest, STLO, 65 rue de Saint-Brieuc, 35042, Rennes, France

Corresponding author: romane.junker@inrae.fr

The *Lactobacillus* genus comprises 261 species displaying a great diversity of genotypes, phenotypes and habitats, some of them exhibiting key importance in food, biotechnology and therapeutic applications. The initial taxonomy of lactobacilli, mostly based on phenotyping traits and chemotaxonomic criteria such as DNA-DNA hybridization, was recently revised using a comparative genomics approach, leading to the creation of 23 novel genera [1]. In this context, the International Center for Microbial Resources dedicated to food associated bacteria at INRAE (CIRM-BIA) has recently decided to explore and characterize the genomic and functional diversity of a collection of 250 food associated strains from 21 species reflecting the three major lifestyle categories (free, commensal, nomadic) known for this group.

In order to analyze this dataset, we first designed a Snakemake bioinformatics workflow [2] allowing the fine characterization of the quality of genomic data, the assembly and annotation of genomes, and the analysis of genomic diversity of the 250 *Lactobacillus* strains, both at inter and intra-species levels. We computed different phylogenetic trees to evaluate and represent the dataset diversity at different scales, but also to enable metadata integration and evaluation of the relevance of the new taxonomy of lactobacilli on our dataset. We are currently working on the comparison of the gene content of the 250 assemblies to analyze the metabolic potential of this bacterial group through the construction of the pan-genome of the different species.

The work is still in progress and the poster will present the first results we obtained on phylogenetic trees constructed either using genomic distances such as MASH [3] or from phylogenomics approaches based on the core-genome alignment, and represented using the ggtree R package [4].

Acknowledgements

We are grateful to the ENS Paris-Saclay for the funding of this internship and to INRAE MIGALE bioinformatics facility (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi: 10.15454/1.5572390655343293E12) for providing help and computing and storage resources.

References

[1] Jinshui Zheng, Stijn Wittouck, Elisa Salvetti, Charles MAP Franz, Hugh MB Harris, Paola Mattarelli, Paul W O'Toole, Bruno Pot, Peter Vandamme, Jens Walter, et al. A taxonomic note on the genus lactobacillus: Description of 23 novel genera, emended description of the genus lactobacillus beijerinck 1901, and union of lactobacillaceae and leuconostocaceae. *International journal of systematic and evolutionary microbiology*, 70(4):2782–2858, 2020.

[2] Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

[3] Brian D Ondov, Todd J Treangen, Páll Melsted, Adam B Mallonee, Nicholas H Bergman, Sergey Koren, and Adam M Phillippy. Mash: fast genome and metagenome distance estimation using minhash. *Genome biology*, 17(1):1–14, 2016.

[4] Guangchuang Yu, David K Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 8(1):28–36, 2017.

Genome-wide analysis of DNA methylation robustness to environmental perturbations

Fabien Kon-Sun-Tack[1], Florent Chuffart[2] and Magali Richard[1]
[1] BCM team, TIMC-IMAG (CNRS UMR5525 - UGA)
[2] EpiMed group, Institute for Advanced Biosciences (INSERM U1209 - CNRS UMR5309 - UGA)
[1,2] Grenoble, France

Corresponding author: magali.richard@univ-grenoble-alpes.fr

Recent cancer-related epigenome studies have highlighted the main role of epigenetic modifications during tumorigenesis. Notably, variations of DNA methylation in cancer are closely associated with abnormal gene expression and oncogenic processes. Although universal systematic epigenetic deregulation patterns have been revealed in cancer, pan-cancer methylation studies of large-scale methylation data also indicates that differential methylation patterns vary significantly among cancers. Despite their crucial biological and clinical relevance, the constraints underlying DNA methylation variation remain poorly understood. In this work, we aim to investigate the robustness and sensitivity of DNA methylation patterns to environmental variations and oncogenic processes.

We addressed this question by applying a systematic statistical analysis on DNA methylation (Illumina Infinium methylation EPIC BeadChip) issued from multiple replicates of HCT 116 human colorectal carcinoma cell line under normal condition or under serum deprivation stress. First, we built an epigenome-wide association study (EWAS) and combined the resulting spatially correlated P-values to identify conserved and differentially methylated regions. We then assessed the variance of DNA methylation probes in those regions using biostatistical models accounting for uneven methylation data structure across the genome.

This should allow the identification of typical patterns of noise in DNA methylation signal, corresponding to epigenomic region less constrained and more prone to change under environmental or physiological perturbations, such as oncogenic processes. We then intend to evaluate the impact of these epigenetic hotspots on gene expression and to propose models that interpret or predict functionally meaningful cancer epigenetic "hot" domains. The results obtained will be correlated with clinical data (prognosis, cancer predisposition, treatment response, etc).

EuphausiiDB : A Transcriptomic reference Database for Krill Species

Fatoumata Binta BARRY[1,2], Mark HOEBEKE[1], Jean-Yves TOULLEC[2] and Erwan CORRE[1]

[1] CNRS - Sorbonne Universite´ - Plateforme ABIMS - FR2424 - Station Biologique de Roscoff, Place Georges Teissier, 29680, Roscoff, France
[2] CNRS - Sorbonne Universite´ - DYDIV Team- UMR 7144 - Station Biologique de Roscoff, Place Georges Teissier, 29680, Roscoff, France

Corresponding Author: erwan.corre@sb-roscoff.fr

Euphausiids, commonly called "krill", are the main prey of marine predators such as seabirds, fish or marine mammals, and they represent both a direct link between the extreme levels of the food chain and its pillar. They are as well a good organism model in terms of geographical diversity and latitudinal distribution within the oceans to study the diversity of physiological adaptations to temperature.

Indeed the geographic distribution of ectothermic species is determined in part by temperature, as seems to be evidenced by the distribution of species along the latitudinal thermal gradient and the bathymetric sinking of boreal marine species along this gradient [1, 2]. In addition to the species-specific evolutionary history, understanding the physiological mechanisms underlying this temperature-dependent distribution should help explain the current distribution of species but also predict how global warming is likely to impact species where they occur and how they will be able to cope [3, 4].

With 85 species, the *Euphausiidae* is the largest family in the order *Euphausiacea* (86 species).

We present in this study the constitution of the first genomic resource of annotated transcriptomes of Euphausiidae covering a wide spectrum of species (17 species).

All datasets were assembled and analyzed using the same workflows dedicated to *de novo* assembly and functional annotation.

The assembly workflow includes evaluation, filtering and trimming of raw data as well as de novo assembly and evaluation of assembled transcripts. Indeed, assemblies were first performed jointly with Trinity and rnaSpades, then selected for EuphausiiDB based on remapping rates (Salmon), BUSCO completeness rates (against eukaryotadb10 and arthropodadb10 databases) and average contig length. The annotation workflow was therefore mostly performed on rnaSpades transcriptomes. We predicted proteins with transdecoder and functional annotation is done with Interproscan, Diamond (vs. Uniprot/Swissprot, Uniref90) tools. We also performed an rRNA search with the barnap tool and a signal peptide search with SignalP.

The **EuphausiiDB** portal offers the possibility to users to explore the database by using a “simple or advanced” search function for a specific taxonomic level, a specific geographic location or a project origin and soon specific annotation. Statistical interactive charts, readsets location map and table and resulting datasets list are associated to the search functions. For each selected dataset, the user can access to both readset and assembly short summary page with cross-references to external databases (EBI SRA, NCBI taxID, WORMS) which allows better traceability and homogeneity across databases and the possibility of downloading all resulting files.

The idea of creating this database dedicated to euphausiids gradually developed due to the increasing amounts of data collected during different campaigns. The aim is not to build a new database on the emblematic species of the taxon, *Euphausia superba*, but to offer the scientific community a broader and evolutive transcriptomic resource that will support the development of new studies on less well known species or studies employing a comparative approach.

References

1. Bedulina DS et al. Expression patterns and organization of the hsp70 genes correlate with thermotolerance in two congener endemic amphipod species (*Eulimnogammarus cyaneu*s and *E. verrucosus*) from Lake Baikal. Mol Ecol doi:10.1111/mec.12136, 2013.
2. Morley SA et al. South Georgia: a key location for linking physiological capacity to distributional changes in response to climate change. Antarct Sci 22:774-781, 2010.
3. Somero GN. The physiology of climate change: how potentials for acclimatization and genetic adaptation will determine 'winners' and 'losers'. Journal of Experimental Biology 213:912-920, 2010.
4. Tomanek L. Variation in the heat shock response and its implication for predicting the effect of global climate change on species' biogeographical distribution ranges and metabolic costs. The Journal of experimental biology 213:971-979, 2010.

Modeling zero-inflated microbiome data: a sufficient dimension reduction approach

Eric Koplin[1], Diego Tomassi[3], Liliana Forzani[1,2] and Ruth Pfeiffer[4]
[1] National Scientific and Technical Research Council (CONICET), Argentine
[2] Universidad Nacional del Litoral, Santa Fe, Argentine.
[3] Biofortis - Mérieux NutriSciences, Saint-Herblain, France.
[4] National Cancer Institute, Division of Cancer Epidemiology and Genetics, Maryland, USA.

Corresponding author: diego.tomassi@mxns.com

1 Introduction

The microbiome is the genetic material of all bacteria, fungi, protozoa and viruses that live on and inside the human body. Features of the microbiome have been shown to impact human phenotypes and play a role in the activity of some therapeutic drugs. Given the recent availability of high throughput microbiome data much effort has been devoted to develop approaches to analyze and visualize microbiome data. We present a multivariate model-based approach that seeks to provide a unifying framework for analysis and visualization. We propose and study lower dimensional reductions of the data, that preserve all relevant information for the phenotype. A special feature of the data is the excess number of zeros, which we explicitly accommodate in our approach. The methodology builds upon our own previous results [1] and it is also related to conditional extensions of Hurdle models proposed to deal with single-cell RNAseq data [2].

2 Method

We introduce zero inflated conditional graphical models for the association between the microbiome data and the outcome or phenotype of interest. To model the excess of zeros we augment the microbiome compositions $\mathbf{X}$ with a zero-pattern vector $\nu = \nu(\mathbf{X})$, where $\nu(X_i) = I(X_i > 0)$. We propose several pairwise graphical models and derive score-based estimators [3] for the inverse regression model $P(\mathbf{X}, \nu \mid Y) = P^{+}(\mathbf{X}|\nu, Y)P(\nu|Y)$, where Y denotes the outcome of interest or phenotype. The first probability in the product models abundance relationships conditioned on the outcome (differential abundance when Y is binary), whereas the second models conditional presence-absence relationships. The two components in the joint-model are estimated separately. Once these models have been estimated, low-dimensional visualization and relevant biomarkers are derived from the model parameters. Importantly, interactions between different microbiome components are obtained as a by-product.

3 Results

Results obtained for synthetic data with a realistic proportion of zeros [4] show that the proposed models perform very well and indeed better than models that do not take into account the excess of zeros. Results with real data from the American Gut project with body mass index as the outcome also showed that the presence of zeros estimated separately in the reduction greatly improved the prediction of the outcome.

References

[1] D. Tomassi, L. Forzani, S. Duarte, and R. M Pfeiffer. Sufficient dimension reduction for compositional data. *Biostatistics*, 12 2019. kxz060.

[2] A. McDavid, R. Gottardo, N. Simon, and M. Drton. Graphical models for zero-inflated single cell gene expression. *The Annals of Applied Statistics*, 13(2):848 – 873, 2019.

[3] A. P. Dawid, S. Lauritzen, and M. Parry. Proper local scoring rules on discrete sample spaces. *The Annals of Statistics*, 40(1):593 – 608, 2012.

[4] R. Rong, S. Jiang, L. Xu, and et al. MB-GAN: Microbiome Simulation via Generative Adversarial Network. *GigaScience*, 10(2), 02 2021. giab005.

Improving the quality of public metagenomic data to study soil microorganisms

CAROLE BELLIARDO[1,2], Mathilde Clement[2], Justine Lipuma[2], Georgios D Koutsovoulos[1], Corinne Rancurel[1], Marc Bailly-Bechet[1] and Etienne GJ Danchin[1]

[1] Sophia Agrobiotech Institute, 400 route des Chappes, 06903, Sophia-Antipolis, France
[2] Mycophyto SAS, 400 route des Chappes, 06903, Sophia-Antipolis, France

Corresponding Author: Carole.Belliardo@Inrae.fr

Metagenome-Assembled Genomes (MAG) are a powerful resource to uncover the genetic diversity of uncultured microorganisms in different complex environments [1]. Specialized databases provide storage of worldwide samples and automated raw data processing to obtain MAG [2]. This type of open data represents unparalleled opportunities in the study of biotic interactions, including plant-associated organisms.

The rhizosphere is the area where multiple soil-dwelling microorganisms interact with plants. Among those, some pathogenic nematode species can invade roots and hijack plant resources. Although horizontal gene transfers (HGT) from bacteria and fungi have contributed to the evolution of phytophagy in nematodes; the donors have never been accurately identified [3]. In parallel, some other members of the rhizosphere, called arbuscular mycorrhizal fungi (AMF), provide benefits to plants, improving their growth and health. However, which combination of microorganisms benefits more to plant health across different environments remains poorly known [4]. We investigated whether mining soil MAGs could improve our knowledge on these two main questions related to plant health.

We collected 6,800 soil metagenomic datasets from the Joint Genome Institute's IMG/M server, the most extensive metagenomic resource [2]. The challenge was to filter and make this massive dataset more accurate and meaningful. First, we filtered the data based on assembly's quality by keeping only proteins from contigs of at least 1,000 bp or containing at least 3 genes. An important issue in MAGs is the underrepresentation of eukaryotes and their annotation with prokaryotic tools [5]. Thus, we filtered eukaryotic contigs and re-predicted proteins using Augustus, a eukaryotic dedicated gene predictor. Moreover, due to the bulk sequencing of all organisms and the combining of multiple samples; redundancy at the protein sequence level had to be eliminated. Finally, because the taxonomic assignment is solely based on the best blast hit against NR regardless of the percent identity; this information is unreliable. We thus, improved taxonomic assignment using a last common ancestor algorithm. After all these steps, we obtained an improved and non-redundant database (950 Million proteins) more representative of the soil natural biodiversity.

We are currently using this resource to study plant-associated organisms: (i) we are obtaining a more complete detection of HGT in nematodes with better identification of putative donors (ii) we are mapping the distribution of AMFs to deduce the correlation between the observed species and environmental characteristics. The results will improve our understanding of the interaction between plants and other organisms to consider new strategies for crop cultivation and protection.

Acknowledgements

We would like to thank Mycophyto SAS and the Plant Health and Environment Department of INRAE for supporting this project. We also thank Samuel Mondy and Laura Eme for their expertise and suggestions.

References

[1] Parks, D.H., Rinke, C., Chuvochina, M. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* **2,** 1533–1542, 2017.

[2] I-Min A. Chen, Victor M. Markowitz, Ken Chu *et al.* IMG/M: integrated genome and metagenome comparative data analysis system, *Nucleic Acids Research*, Volume 45, Issue D1, January, Pages D507–D516, 2017.

[3] Etienne G. J. Danchin, Marie-Noëlle Rosso, Paulo Vieira, Janice de Almeida-Engler, Pedro M. Coutinho, Bernard Henrissat, Pierre Abad. Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes. *Proceedings of the National Academy of Sciences*, 2010.

[4] Grace A Hoysted, Jill Kowal, Alison Jacob, William R Rimington, Jeffrey G Duckett, Silvia Pressel, Suzanne Orchard, Megan H Ryan, Katie J Field, Martin I Bidartondo. A mycorrhizal revolution. *Current Opinion in Plant Biology*, Volume 44, Pages 1-6, 2018.

[5] Patrick T. West, Alexander J. Probst, Igor V. Grigoriev, Brian C. Thomas and Jillian F. Banfield. Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Research*. 2018.

Analysis of cell identity loss in cancer using artificial intelligence

Alexis Pellerin[1], Florent Chuffart[1], Saadi Khochbin[1], Sophie Rousseaux[1] and Ekaterina Bourova-Flin[1]
EpiMed, Institute for Advanced Biosciences, INSERM U1209, CNRS UMR5309, University Grenoble Alpes, France

Corresponding author: pellerin_alexis@hotmail.com

Malignant transformation is associated with major abnormalities in the genome and its "markup" system, the epigenome, leading to deregulations of the gene expression program and subsequent modification, or even total loss, of cellular identity. Following these deregulations, cells can acquire new properties, which do not exist in the original normal cells, some of which are characteristic of cancer cells, such as the ability to spread to locations far from their place of origin. One of the consequences of these deregulations is the ectopic activation of genes which should normally be silent in healthy somatic adult tissues. For instance, as we observed from our previous work, tissue-specific genes, normally expressed only in one particular tissue, especially germline and placental genes, become aberrantly activated in tumour cells [1,2,3].

Today, considerable technological advances in the fields of genomics and post-genomics produce a large volume of high-quality data of genome sequencing. RNA sequencing (RNA-seq) data, measuring the gene expression levels in large datasets of normal and tumour human samples, are currently available in public repositories (GTEX, NCBI GEO and TCGA) and represent a great opportunity to address complex biological questions using artificial intelligence approaches. In this project, we aim to create a machine learning model which can accurately recognize the cellular identity of normal human tissues from large datasets of publicly available transcriptomic data. The classifier is built with more than 3600 samples and 30 tissue types and can be applied at different levels of tissue representation. The main steps of the pipeline are shown in Fig.1. This classifier, once trained on normal samples, represents a powerful tool to measure cell identity deregulations in different tumour samples. Using this tool, we are able to explore tumour heterogeneity in several cancers as well as in histological subtypes, and to integrate this information with molecular, biological and clinical data.

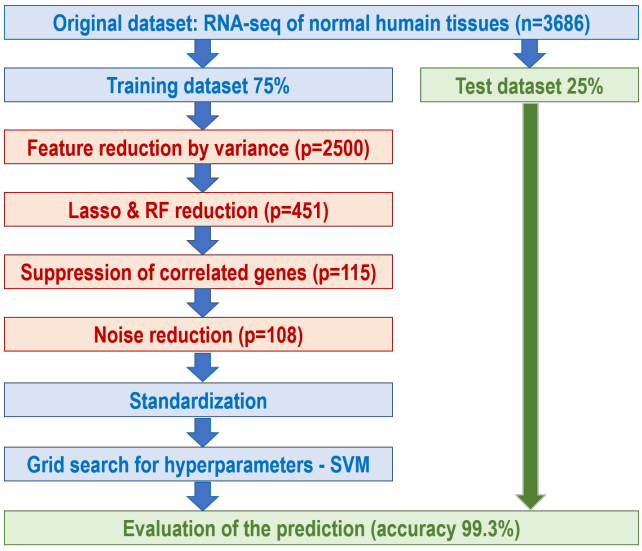


Fig. 1. The main steps of the pipeline to create a classifier for cell identity recognition. First, the original cohort is split into training and test datasets using stratified k-fold cross-validations. Then, several steps of feature reduction are progressively applied to the training dataset (red boxes). Finally, a data standardization and a grid search for hyperparameters are performed before the training of SVM model. The classifier predicts the cell identity in test datasets with the mean accuracy of 99.3%.

References

[1] Sophie Rousseaux, Jin Wang, and Saadi Khochbin. Cancer hallmarks sustained by ectopic activations of placenta/male germline genes. *Cell Cycle*, 12(15):2331–2332, 2013. PMID: 23856584.

[2] Jin Wang, Sophie Rousseaux, and Saadi Khochbin. Sustaining cancer through addictive ectopic gene activation. *Current Opinion in Oncology*, 26(1), 2014. PMID: 24275853.

[3] Sophie Rousseaux, Ekaterina Bourova-Flin, Mengqing Gao, Jin Wang, Jian-Qing Mi, and Saadi Khochbin. Unprogrammed gene activation: A critical evaluation of cancer testis genes. In *Encyclopedia of Cancer*, pages 523–530. Academic Press, Oxford, 2019.

AgroLD: A Knowledge Graph Database for plant functional genomics

Pierre Larmande [1,2], Ndomassi Tando [1,2], Bertrand Pitollat [2,3], Valentin Guignon [2,4], Mathieu Rouard [2,4], Gaetan Droc [2,3], Manuel Ruiz [2,3]

1 - DIADE, IRD, Univ. Montpellier, 911 av Agropolis, 34398 Montpellier, France
2 - French Institute of Bioinformatics (IFB)—South Green Bioinformatics Platform, Bioversity, CIRAD, INRAE, IRD, F-34398 Montpellier, France
3 - AGAP, CIRAD, INRAE, Univ. Montpellier, av Agropolis, 34398 Montpellier, France
4 - Bioversity International, Parc Scientifique Agropolis II, 34397 Montpellier, France

Corresponding Author: pierre.larmande@ird.fr

Exploring the links between genetic and phenotypic traits is an important area of research in agronomy. One of the main objectives of this is to accelerate the development of important traits that can positively impact the agricultural economy. However, due to the existence of complex molecular interactions, to gain complete understanding will warrant data analyses performed at different molecular and environmental levels for a given (plant) subject. For instance, to understand how rice genes involved in metabolism or signaling of growth regulators control the rice panicle architecture. While high-throughput technologies have played a key role in accelerating and generating the much-needed data, these can only partially capture the dynamics in genotype-phenotype relations. Consequently, our knowledge of the complex relationships between the different molecular actors responsible for the expression of the phenome in various plant systems remains fragmented. Hence, there is an urgent need to effectively integrate and assimilate complementary information to understand the biological system in its entirety.

We have developed AgroLD [1] (www.agrold.org), a knowledge graph system that exploits the Semantic Web technology and FAIR principles [2], to integrate information to integrate data about plant species of high interest for the plant science community e.g., rice, wheat, Arabidopsis and in this way facilitating the formulation of new scientific hypotheses. We present some integration results of the project, which currently focused on genomics, proteomics and phenomics. AgroLD is now an RDF knowledge base of 900M triples created by annotating and integrating more than 100 datasets coming from 15 data sources –such as Ensembl plants [3], Gramene.org [4] and TropGeneDB [5]– with 15 ontologies –such as the Gene Ontology [6] and Plant Ontology [7]. Our objective is to offer a domain specific knowledge platform to solve complex biological and agronomical questions related to the implication of genes in, for instances, plant disease resistance or high yield traits. We expect the resolution of these questions to facilitate the formulation of new scientific hypotheses to be validated with a knowledge-oriented approach.

Acknowledgements

Authors thank the technical staff of the South Green Bioinformatics platform for their support.

References

1. Venkatesan A, Tagny Ngompe G, Hassouni NE, Chentli I, Guignon V, Jonquet C, et al. Agronomic Linked Data (AgroLD): a Knowledge-based System to Enable Integrative Biology in Agronomy. PLoS ONE. 2018;:13:17.
2. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3.
3. Bolser D, Staines DM, Pritchard E, Kersey P. Ensembl Plants: Integrating Tools for Visualizing, Mining, and Analyzing Plant Genomics Data. Methods Mol Biol Clifton NJ. 2016;1374:115–40.
4. Tello-Ruiz MK, Naithani S, Stein JC, Gupta P, Campbell M, Olson A, et al. Gramene 2018: Unifying comparative genomics and pathway resources for plant research. Nucleic Acids Res. 2018.
5. Hamelin C, Sempere G, Jouffe V, Ruiz M. TropGeneDB, the multi-tropical crop information system updated and extended. Nucleic Acids Res. 2013;41.
6. Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res. 2015;43:D1049–56.
7. Cooper L, Walls RL, Elser J, Gandolfo MA, Stevenson DW, Smith B, et al. The plant ontology as a tool for comparative plant anatomy and genomic analyses. Plant Cell Physiol. 2013;54:e1.

ANNEXA: Analysis of Nanopore transcriptomic data with Nextflow for Extended Annotation

Matthias LORTHIOIS[1], Édouard CADIEU[1], Armel HOUEL[1], Catherine ANDRÉ[1], Christophe HITTE[1], Benoit HÉDAN[1], Thomas DERRIEN[1]

[1]Univ Rennes 1, CNRS, IGDR – UMR6290, F-35000 Rennes, France.

Corresponding Author: tderrien@univ-rennes1.fr

The development of long-read transcriptome sequencing (LR-RNAseq) promises to facilitate the process of genome annotation. By providing reads spanning repeats and direct exon/exon connectivity, LR-RNAseq represents an unfragmented vision of the transcriptome which thus allows the refinement of gene/transcript reconstruction and quantification. Yet, it requires the development of novel bioinformatic solutions specifically adapted to these novel technologies.
We first benchmarked four transcriptome reconstruction tools (Stringtie2, TALON, FLAIR and bambu) with respect to known reference annotation (Ensembl) and showed that bambu[1] provided the best performance in terms of sensibility and specificity for building exon and spliced transcript models. Then, based on bambu, we developed ANNEXA, an all-in-one reproductible pipeline, written in the Nextflow workflow manager, which allows users to analyze LR-RNAseq sequences from Oxford Nanopore Technologies (ONT), and to reconstruct and quantify known and novel genes and isoforms. More specifically, ANNEXA works by using only three parameter files (a reference genome, a reference annotation and mapping files) and provides users with an extended annotation distinguishing between novel protein-coding (mRNA) versus long non-coding RNAs (lncRNA) genes[2]. All known and novel gene/transcript models are further characterized through multiple features (length, number of spliced transcripts, normalized expression levels...) available as graphical outputs, including automatic clustering of samples if multiple conditions are provided (*e.g.* tumor/control).
To demonstrate the usability of the program in the context of comparative oncology studies, we sequenced 2 human and 7 dog cancer cell lines from mucosal melanomas and histiocytic sarcomas using ONT direct cDNA sequencing, representing ~60M nanopore reads (mean=6.5M reads/sample). We then applied ANNEXA on these two species-specific read sets and were able to reconstruct and quantify 1,842 and 8,262 new multi-exonic human and canine genes, respectively, all supported by at least 5 reads in both species. When including mono-exonic genes, most of new loci are classified as lncRNAs (59% and 70% in human and dog, respectively) which is expected given the higher number of lncRNAs with only one exon and their higher level of tissue/cell type-specificity compared to mRNAs.
Overall, our work presents a new bioinformatic pipeline to automatically reconstruct and characterize mRNAs and lncRNAs from ONT transcriptome data and is freely available at: https://github.com/mlorthiois/ANNEXA.

Acknowledgements

Authors would like to warmly thank the Bioinformatics Genouest platform (https://www.genouest.org) for providing the required infrastructure for this work.

References

1. Chen, Y. *et al.* A systematic benchmark of Nanopore long read RNA sequencing for transcript level analysis in human cell lines. *bioRxiv*2021.04.21.440736 (2021) doi:10.1101/2021.04.21.440736.
2. Wucher, V. *et al.* FEELnc: A tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Research* **45**, 1–12 (2017).

LINE-1 evolution in Mammals

Quentin Bouvier[1,2], Anthony Boureux[1], Nicolas Gilbert[1]
[1] IRMB-INSERM U1183, 80 av. Augustin Fliche, 34295, Montpellier, France
[2] Master Bioinformatique, Connaissaces et Données, University of Montpellier, Montpellier, France

Corresponding author: Anthony.Boureux@inserm.fr

Abstract

Secrets of evolution are contained in DNA sequences of all living creatures. Genetic drift is the main cause of biological diversity, of the multiplicity of biological functions and the divergence of species. For all theses reasons, multiples studies aim to understand the influence of DNA components on the appearance of different traits. Since the turn of the millennium, and with the development of NGS technologies numerous genomes have been sequenced and assembled. An important conclusion is that all higher eukaryotic genomes are mostly composed of transposable elements (TE). Moreover, multiple studies on TE revealed their implication on different biologic process like phenotypic modification, gene regulation or even evolution. A better knowledge of TE families in mammalian genomes would therefore contribute to better understand the implication of TE on mammalian evolution.

Since LINE-1 (L1) is the most abundant repeat element present in all mammalian genomes, we decided to study its structure and evolution in several assembled genomes. To identify L1 families, we chose the TE *de novo* characterization approach since we don't know the extent of the divergence between L1 elements in the different species. Many different detection programs exist but none can detect all the TEs. Therefore, in a pre-study, we used two well-known tools that both combine several and complementary programs dedicated to identify repeat sequences: REPET [1](using RECON [2], GROUPER [3] and PILER [4]) and RepeatModeler2 [5] (using RECON and RepeatScout [6]). The goal being here to determine which pipeline gives the best results in term of time, needs and TE identification. We analysed few mammalian genomes from different orders with both tools and the same computer settings. We compiled all TEs families obtained and classified them by class and subclass. Specially, we redefine consensus L1 sequence obtained in genomes for which no L1 sequences were available in public data bases. The newly define LINE-1 consensuses will be implemented in a L1 specific database, and in Public database. The propose of this pre-study is also to add this specific database to TE *de novo* pipeline to extend these analyzes to more species and ultimately to facilitate the automatic identification of L1 families in all the mammalian genomes available today.

Acknowledgements

I want to express my thanks to Anna-Sophie Fiston-Lavier for her support and to the master BCD for the knowledge transmission.

References

[1] Timothée Flutre, Elodie Duprat, Catherine Feuillet, and Hadi Quesneville. Considering transposable element diversification in de novo annotation approaches. 6(1):e16526, 2011.

[2] Z. Bao. Automated de novo identification of repeat sequence families in sequenced genomes. 12(8):1269–1276, 2002.

[3] Hadi Quesneville, Danielle Nouaud, and Dominique Anxolabehere. Detection of New Transposable Element Families in Drosophila melanogaster and Anopheles gambiae Genomes. *Journal of Molecular Evolution*, 57(0):S50–S59, August 2003.

[4] R. C. Edgar and E. W. Myers. PILER: identification and classification of genomic repeats. *Bioinformatics*, 21(Suppl 1):i152–i158, June 2005.

[5] Jullien M. Flynn, Robert Hubley, Clément Goubert, Jeb Rosen, Andrew G. Clark, Cédric Feschotte, and Arian F. Smit. RepeatModeler2 for automated genomic discovery of transposable element families. 117(17):9451–9457, 2020.

[6] A. L. Price, N. C. Jones, and P. A. Pevzner. De novo identification of repeat families in large genomes. 21:i351–i358, 2005.

JOBIM 2021 Pilot Project – Gender Speaking Differences in Academia

Junhanlu ZHANG, Rachel TORCHET and Hanna JULIENNE

Department of Computational Biology, Institut Pasteur 28 Rue du Docteur Roux, 75015, Paris, France

Corresponding Author: jobim-project@pasteur.fr

1. Objective of the study

We are launching a pilot project at the JOBIM 2021 conference to investigate gender speaking differences in academia. This mixed-method study intends to answer the following question: how to create the conditions for gender-equal expression in scientific conferences? Not only does gender bias establish barriers to women's academic achievement in externally constrained areas like the publication process [1,2,3], but it also shows significance in more subtle phenomena such as the number of questions asked in large scientific conferences [4,5]. Encouraged by the collective efforts made in the scientific community to mitigate the gender-gap in question-asking behaviour, we propose to move the discussion forward through this observation study during the JOBIM conference. In addition to surveys, semi-structured interviews will be conducted as complementary data collection methods. The research team consists of specialists in social/cultural anthropology, statistics and UX design. This observational and in-depth study will offer us insights on how to create a more inclusive scientific environment for all and to nurture a scientific community that would not overlook any worthy contributions.

2. Methodology and project implementation

This original study will be at the interface between social science and quantitative science. Data collection through surveys will be performed during the registration process and after the conference. During the conference, a participatory and observational study will be conducted under detailed guidelines. The post-survey questions will cover basic demographic information and focus on question-asking behaviours during the JOBIM conference. Finally, we will invite a number of conference attendees for an in-depth interview. All data will be analyzed in a scientific manner with adoption of qualitative research approaches. As an end result, a scientific paper will be produced to present the research findings of this study. In addition, guidelines for future conferences will be provided and shared in the form of a report. This poster session is an opportunity for JOBIM participants to ask questions about the project and discuss its methodology.

3. Acknowledgements

This study is funded and supported by the Institute Pasteur HUB of Bioinformatics and Biostatics. For further information about our data protection policy, please visit the research project page following the link: https://research.pasteur.fr/en/project/jobim-2021-pilot-project-gender-speaking-differences-in-academia/

References

1. Besselaar Peter van den and Sandström Ulf. Vicious circles of gender bias, lower positions, and lower performance: Gender differences in scholarly productivity and impact. *Plos one*, *12*(8), 2017.
2. Broderick Nichole A and Casadevall Arturo. Gender inequalities among authors who contributed equally. *ELife*, *8*, e36399, 2019.
3. Larivière Vincent, Ni Chaoqun., Gingras Yves, Cronin, Blaise and Sugimoto Cassidy R. Bibliometrics: Global gender disparities in science. *Nature News*, *504*(7479), 211, 2013.
4. Telis Natalie, Glassberg Emily C., Pritchard Jonathan K. and Gunter Chris. Public Discussion Affects Question Asking at Academic Conferences. *The American Journal of Human Genetics*, *105*(1), 189–197, 2019.
5. Carter Alecia J., Croft Alyysa, Lukas Dieter and Sandstrom Gillian M. Women's visibility in academic seminars: Women ask fewer questions than men. *Plos one*, *13*(9), e0202743, 2018.

Visualization of FAANG data with VizFaDa

Laura Morel[1], Peter Harrison[2] and Guillaume Devailly[1]

[1] GenPhySE, 24 chemin de Borde-Rouge, 31326, Castanet-Tolosan, France
[2] EMBL-EBI, Hinxton, United Kingdom

Corresponding Author: laura.morel@inrae.fr

The FAANG (*Functional Annotation of Animal Genomes*) international consortium aims to produce high-quality functional annotation of the genomes of domesticated animals [1]. Members of the community can submit their epigenomics, transcriptomics or genomics data to the FAANG Data Portal (https://data.faang.org) coordinated by a Data Coordination Centre at the EMBL-EBI [2]. FAANG data conforms to principles of findability, accessibility, interoperability and reusability (FAIR). The FAANG Data Portal allows users to find, select and download datasets relevant to their research using extensive sample and experimental metadata standards.

VizFaDa aims to produce interactive data visualization through web applications intended to be integrated to the FAANG Data Portal. In order to generate those visualizations, the raw data from the portal has to be processed. During this step, quality control reports are created, providing valuable and previously unavailable insight into the quality of the data. VizFaDa focuses on RNA-seq, ChIP-seq and DNA methylation data.

Interactive clustered correlation heatmap are generated, allowing the user to compare experiments from a certain assay type within a species. Experiments with similar results areclustered together. The user can use FAANG metadata to annotate the heatmap or to filter experiments from the database for a more focused visualization. Stacked epigenetic profiles are created from gene expression and epigenetic data obtained either from the same sample or from two comparable samples, notatbly at transcription start sites. This allow the investigation of relationship between epigenetic marks and transcription levels.

In future versions, additional functions will be added. Users will be able to upload their own processed data and use clustered heatmaps to find experiments with similar results. Data submitted to the portal will be automatically processed and added to VizFaDa, ensuring the long-term relevance and accuracy of the project.

Acknowledgements

The authors would like to thank Alexey Sokolov from EMBL-EBI, Sylvain Foissac and the GenEpi team at GenPhysSE for their support.

References

[1] The FAANG Consortium; Andersson, L.; Archibald, A. L.; Bottema, C. D.; Brauning, R.; Burgess, S. C.; Burt, D. W.; Casas, E.; Cheng, H. H.; Clarke, L.; et al. Coordinated International Action to Accelerate Genome-to-Phenome with FAANG, the Functional Annotation of Animal Genomes Project. *Genome Biology*, (16):57, 2015

[2] Harrison, P. W.; Fan, J.; Richardson, D.; Clarke, L.; Zerbino, D.; Cochrane, G.; Archibald, A. L.; Schmidt, C. J.; Flicek, P. FAANG, Establishing Metadata Standards, Validation and Best Practices for the Farmed and Companion Animal Community. *Animal Genetics*, (49):520–526, 2018

Model selection in Phylogeny : Performances and limitations

Anaïs Prud'homme[1,2], Anne-Muriel Arigon[2] and Vincent Lefort[2,3]
[1] Master Sciences et Numérique pour la Santé, parcours Bioinformatique, Connaissances, Données, Univ Montpellier, France
[2] LIRMM, Univ Montpellier, CNRS, Montpellier, France
[3] Institut Français de Bioinformatique, CNRS UMS 3601, France

Corresponding author: anais.prudhomme@etu.umontpellier.fr

Context In phylogeny, many evolutionary models have been proposed to describe sequences evolution. The most accurate phylogeny inference methods use probalistic models to estimate the equilibrium frequencies of the residues (nucleotides or amino-acids), the substitution rates of these residues, the variability of the substitution rates and the proportion of invariable sites. The selection of the model that best fits data is an important step in phylogenetic analysis. Thus, several model selection methods have been developed. Among these methods, those based on the information theory criteria (AIC [1]; AICc [2]; BIC [3]; and DT method [4]) are widely used. Some of them perform better depending on data characteristics.

Approach This study aims to analyse the performances of these model selection methods. To do so, we first study the validity conditions of each method by varying the ratio number of sites/number of taxa (which we call υ) and the relative evolution rate between taxa. Then we compare the quality of each method by studying the method error rates and under which condition a method performs better than the others. This study comprises three steps and is implemented in Snakemake [5].

Implementation The first step is to simulate data sets derived from 36 real trees selected on Orthomam [6] and corresponding to 3 ranges of relative evolution rate ([0;1], [1;2] and [2;5]) and 4 taxa numbers (50, 70, 90, 110). We use 16 evolutionary models (HKY85, GTR, LG, WAG with the combinations of +I, +Γ, +I+Γ) and 6 υ ratios (1, 10, 25, 50, 75, 100) in order to generate 34 560 alignments: 17 280 DNA alignments and 17 280 protein alignments. We use INDELible [7] to simulate the data.

The second step is to run the four selection model methods. We use JModeltest2 [8] and ProtTest3 [9] to estimate each alignment likelihood which is used to compute the information theory criteria values.

The third step is divided into two stages: 1) the analysis of the validity conditions for each method, 2) the comparative analysis of method results.

References

[1] Akaike Htrotugu. Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, 60(2):255–265, 1973.

[2] Clifford M Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.

[3] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[4] Vladimir Minin, Zaid Abdo, Paul Joyce, and Jack Sullivan. Performance-based selection of likelihood models for phylogeny estimation. *Systematic Biology*, 52(5):674–683, 2003.

[5] Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

[6] Vincent Ranwez, Frédéric Delsuc, Sylvie Ranwez, Khalid Belkhir, Marie-Ka Tilak, and Emmanuel JP Douzery. Orthomam: a database of orthologous genomic markers for placental mammal phylogenetics. *BMC evolutionary biology*, 7(1):1–12, 2007.

[7] William Fletcher and Ziheng Yang. Indelible: a flexible simulator of biological sequence evolution. *Molecular biology and evolution*, 26(8):1879–1888, 2009.

[8] David Posada. jModelTest: Phylogenetic Model Averaging. *Molecular Biology and Evolution*, 25(7):1253–1256, 04 2008.

[9] Diego Darriba, Guillermo L. Taboada, Ramón Doallo, and David Posada. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, 27(8):1164–1165, 02 2011.

Deep Learning model for canine gene expression predictions

Camille KERGAL[1], DoGA consortium[2], Catherine ANDRE[1], Marie-Dominique GALIBERT[1,3], Thomas DERRIEN[1] and Christophe HITTE[1]

[1] Univ Rennes 1, CNRS, IGDR – UMR 6290, F-35000 Rennes, France
[2] www.doggenomeannotation.org
[3] Somatic Cancer Genetics Department, Pontchaillou University Hospital, F-35000 Rennes, France

Corresponding Authors: camille.kergal@univ-rennes1.fr, hitte@univ-rennes1.fr, tderrien@univ-rennes1.fr

Deep learning algorithms have recently attracted a lot of attention in genomics and transcriptomics since they promise to extract biological knowledge from large dataset generated by high throughput sequencing technologies in a data-driven manner. For instance, it has been shown that deep learning strategies can outperform traditional machine learning approaches for complex tasks such as gene expression predictions [1]. As part of the team's work in comparative oncology between human and dog, we used and adapted the deep learning tool Basenji [2] based on a Convolutional Neural Network (CNN), developed for human dataset, to build a gene expression prediction model dedicated to the dog species. However, it is not clear whether neural network used to train models in one species can be easily generalized to other species or whether species-specific neural networks, specifically tuned with hyperparameters (HP) optimization would provide better predictive power [3]. The ultimate goal of this strategy is to propose the most powerful model to predict the impact of non-coding genome variations on gene expression and thus, prioritize regulatory variants associated to diseases and phenotypical traits in the dog species [2, 4].

For this purpose, we collected 134 canine expression data with 125 canine CAGE (Cap Analysis of Gene Expression) data as part of a partnership with the Dog Genome Annotation Project (DoGA) consortium [5], that represent 49 canine tissues and 9 public CAGE from the FANTOM consortium [6] corresponding to primary cell types. We were thus able to train a deep learning model to predict dog gene expression from DNA sequence based on the available Basenji architecture developed for human composed of 10 convolutional layers. Then, we evaluated its performance by calculating Pearson correlation coefficients between predicted gene expression and those measured experimentally in a test set of DNA sequences for each CAGE dataset. The median of those coefficients was 0.54, which is substantially lower than the median of coefficients from the human gene expression prediction model of Basenji (0.69) [2].

In order to better model canine expression data, we defined a 2-step approach to build a new deep CNN by including additional layers to the convolution and the dilated convolution steps and by HPs optimization. First, our results showed that deeper networks led to improve predictions of gene expressions (median r=0.54 and 0.60 for 10 and 19 layers, respectively). To optimize HP for the 19 layers network, it is necessary to establish an efficient strategy to avoid the grid search algorithm, aiming to test all combinations of potential HP. For instance, grid search techniques for testing only 6 values of 10 HP imply to train and assess up to 10^6 models. Hence, we are currently conducting a Bayesian optimization as implemented in the Skopt package Python library [3, 7]. It consists in finding the optimum set of HP, based on the generation of several models that at each stage seek the best HP from the previous stages.

Overall, our work highlights the development of deep learning frameworks based on high-throughput transcriptomic data for modeling species-specific gene expression levels.

Acknowledgements

PhD thesis of C. Kergal is funded by a grant from ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation. Authors thank the Bioinformatics Genouest platform for providing the GPU infrastructure, necessary to train prediction models.

References

1. Zou et al. A primer on deep learning in genomics. *Nature Genetics*. 2019.
2. Kelley et al. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*. 2018.
3. Cho et al. Basic enhancement strategies when using Bayesian optimization for hyperparameter tuning of deep neural networks. *IEEE Access.* 2020
4. Atak et al. Interpretation of allele-specific chromatin accessibility using cell state-aware deep learning. *Genome Research.* 2021.
5. Dog Genome Annotation Project. https://www.doggenomeannotation.org/ 2017.
6. Lizio M, et al. Data descriptor: Monitoring transcription initiation activities in rat and dog. *Scientific Data.* 2017.
7. Scikit-optimize contributors. https://scikit-optimize.github.io/ 2017-2020

Logical and incremental formalization of cell cycle checkpoints

Déborah Boyenval[1,2], Gilles Bernot[1], Jean-Paul Comet[1] and Franck Delaunay[2]
[1] I3S laboratory, Euclide B 2000 Route des Lucioles, 06900, Sophia Antipolis, France
[2] Institut de biologie de Valrose, Parc Valrose, 06108, Nice, France

Corresponding author: boyenval@i3s.unice.fr

Checkpoints ensure the integrity of DNA during the cell cycle, which is a succession of molecular and cellular events leading to the division of a mother cell into two genetically identical daughter cells. The DNA is first duplicated (S-phase), then equally distributed between the opposing poles of the cell, which finally divides into two independent daughter cells (M-phase). Two additional phases depict the process of cell preparation before S and M phases: respectively the G1 and G2 phases. Many modeling studies investigating checkpoints are based on ODE systems while the checkpoint concept itself is fundamentally discrete, described as follows: *a checkpoint prevents any event that initiates a phase from taking place before the end of all the events of the previous phase.* So far to our knowledge, very few qualitative models have yet attempted to formally define this discrete concept, as the notion of discrete cell cycle phase is still fuzzy, from a formal point of view.
Therefore, we propose a qualitative modeling study of cell cycle regulation dedicated to the logical specification of the G1, S, G2 and M phases, and the G1/S, S/G2, G2/M and mitosis exit checkpoints. This study has been made possible by using two types of formal methods: the "genetically modified" Hoare logic [1] and the model-checking for CTL [2]. The `TotemBioNet` tool efficiently combines these two methods to exhaustively identify the parameterizations (which govern the dynamics of the regulatory graph) compatible with all formalized biological knowledge [3].
Starting from a qualitative model of cell cycle progression regulation (Behaegel et al. [4]), the cell cycle was defined by a Hoare triple of the form $\{precondition\}\ path\ \{postcondition\}$, where the precondition is a single initial state, the path is a sequence of discrete events, and the postcondition is the single final state. The path was then divided into four canonical phases in order to identify non-permutable key events, which will constitute the main rule of the generic predicate $checkpoint(phase_i, phase_{i+1})$. Since the order of events within a phase is not necessarily known, the predicate (implemented in `Prolog`) includes rules which call `TotemBioNet` to extract all the orders of events of a phase compatible with all other static and dynamic biological knowledge of the system.
The results highlight that three "strong definitions" of checkpoint are validated, but no parameterization satisfies the mitosis exit checkpoint, that requires a *less restrictive* definition. Indeed the single event that can initiate G1 is permutable with the 3 events that can end the M phase. However, no abstract knowledge in our models is challenged since the first event of G1-phase (activation of $CycE/Cdk2$) has also been experimentally observed in M-phase [5]. This new highlighted knowledge has proved that our generic definition of the mitosis exit checkpoint is not biologically consistent, hence has been specifically revised. Finally, the *generic* nature of the checkpoint predicate opens a perspective to study any cyclic phenomenon genetically regulated by checkpoints.

References

[1] Gilles Bernot, Jean-Paul Comet, Zohra Khalis, Adrien Richard, and Olivier Roux. A genetically modified hoare logic. *Theoretical Computer Science*, 765, 06 2015.

[2] Gilles Bernot, Jean-Paul Comet, Adrien Richard, and Janine Guespin. Application of formal methods to biological regulatory networks: Extending thomas' asynchronous logical approach with temporal logic. *Journal of theoretical biology*, 229:339–47, 09 2004.

[3] Déborah Boyenval, Gilles Bernot, Hélène Collavizza, and Jean-Paul Comet. What is a cell cycle checkpoint? the totembionet answer. In *18th International Conference on Computational Methods in Systems Biology (CMSB 2020)*, 2020.

[4] J. Behaegel, J.-P. Comet, G. Bernot, E. Cornillon, and F. Delaunay. A hybrid model of cell cycle in mammals. In *6th International Conference on Computational Systems-Biology and Bioinformatics (CSBio'2015)*, Bangkok (Thailand), November 22-25 2015.

[5] Sabrina Spencer, Steven Cappell, Feng-Chiao Tsai, K Overton, Clifford Wang, and Tobias Meyer. The proliferation-quiescence decision is controlled by a bifurcation in cdk2 activity at mitotic exit. *Cell*, 155, 09 2013.

Evolutionary history of SARS-CoV-2 interactome in bats and primates identifies key virus-host interfaces and conflicts

Marie CARIOU[1], Laurent GUEGUEN[2], Léa PICARD[1,2], Andrea CIMARELLI[1], Oliver FREGOSO[3,4], Dominique GUYOT[5,] Stéphanie JACQUET[1], Antoine MOLARO[6], Vincent NAVRATIL[5], and Lucie ETIENNE[1]

[1] CIRI – Centre International de Recherche en Infectiologie, Inserm U1111, Université Claude Bernard Lyon 1, CNRS UMR5308, Ecole Normale Supérieure de Lyon, Univ Lyon, F-69007, Lyon, France

[2] LBBE – Laboratoire de Biologie et Biometrie Evolutive, CNRS UMR 5558, Université Claude Bernard Lyon 1, Villeurbanne,France

3 Molecular Biology Institute, University of California, Los Angeles, California, USA

[4] Depertment of Microbiology,Immunology and Molecular Genetics, University of California, Los Angeles, California, USA

[5] PRABI, Rhône Alpes Bioinformatics Center, UCBL, Lyon1, Université de Lyon, Lyon, France.

[6] Genetics, Reproduction and Development (GReD) Institute, Université Clermont Auvergne, Clermont-Ferrand, France

Corresponding Authors: marie.cariou@ens-lyon.fr , lucie.etienne@ens-lyon.fr

The current COVID-19 pandemic is caused by a novel coronavirus strain, SARS-CoV-2. It originated from the cross-species transmission of a coronavirus from the bat reservoir, directly or through an intermediate host to humans. This catastrophic spillover underlines the necessity to better understand how viruses and hosts have shaped one another over evolutionary time.

Pathogenic viruses put a selective pressure on the host-viral interacting proteins. Identifying which host genes bear signatures of such evolutionary conflict (e.g. positive selection) can lead to the identification of the proteins that have been the most relevant in the response to a virus family. Here, we have used this evolutionary framework to decipher which interactions between the SARS-CoV-2-like viruses and our cells have been important *in vivo*. In addition, identifying traces of positive selection in different hosts phylogenetic lineages also sheds lights on ancient epidemics and how virus-host determinants may be species specific. This may help to understand differences in susceptibility and pathogenicity to SARS-CoV-like viruses between hosts.

To achieve this, we characterized the evolutionary history of the SARS-CoV-2 interactome identified in *in vitro* studies: 332 host proteins identified by mass-spectrometry by Gordon and collaborators [1], as well as two essential SARS-CoV-2 entry factors, the angiotensin converting enzyme 2 (ACE2) and the transmembrane serine protease 2 (TMPRSS2) genes. We characterized their evolution in primates (tracing the human history) and in bats (the natural viral reservoir). To do so, we used DGINN [2], a novel computational pipeline to Detect Genetic INNovations in protein-coding genes, which embeds gold-standard methods to perform phylogenetic and positive selection analyses in a high-throughput manner.

We found 88 and 38 proteins of the SARS-CoV-2 interactome under strong positive selection in primates and bats, respectively, with enrichment in cell cycle control, centrosome behavior, and DNA replication biological pathways. The ACE2 receptor, as well as seventeen other proteins, showed signatures of adaptation in both bats and primates, which (i) may be indicative of ancient epidemics by pathogenic SARS-like coronaviruses, and (ii) may define a core SARS-CoV interactome in primates and bats. Furthermore, we found other proteins with evidence of past adaptive events only in bats, or primates, highlighting species-specific adaptation. Lastly, positive selection signatures at specific sites in a handful of key genes identify putative molecular interfaces important in SARS-CoV replication and pathogenesis. Overall, these results highlight how the analysis of the “evolutionarily-relevant” interactome might point to primary drug targets.

References

[1] David E Gordon, Gwendolyn M Jang, Mehdi Bouhaddou. et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 583, 459–468, 2020.

[2] Léa Picard, Quentin Ganivet, Omran Allatif, Andrea Cimarelli, Laurent Gueguen, Lucie Etienne, DGINN, an automated and highly-flexible pipeline for the detection of genetic innovations on protein-coding genes, *Nucleic Acids Research*, 48(18) page e103, 2020.

How to explain QTL imbalance between ohnologous chromosomes?

Tanguy Lallemand[1], Sébastien Aubourg[1], Gilles Hunault[1], Jean-Marc Celton[1] and Claudine Landès[1]
Université d'Angers, Institut Agro, INRAE, IRHS, SFR QUASAV, F-49000 Angers, France.

Corresponding author: tanguy.lallemand@inrae.fr

1 Introduction

Polyploidy is a driver of genetic innovation in eukaryotic organisms, especially in plants [1]. A high-quality genome sequence was obtained recently for domesticated apple (*Malus* × *domestica Borkh*) [2]. This genome sequence confirmed a Whole Genome Duplication (WGD) event that occured 50 million years ago [3]. This WGD allows the study of ohnologous gene and makes the apple tree a model of choice to study gene evolution after a recent WGD.

Analysis of duplicated chromosome fragments allows the identification of syntenic blocks in which genes are homologous to each other and arranged in a conserved order. In apple, a Quantitative Trait Loci (QTL) imbalance has been identified. Indeed, for 7 chromosome pairs among the 14 main pairs of ohnologous chromosomes we find significantly more QTL on one chromosome compared to its homologous.

To understand this imbalance, we studied the rate of gene sequence evolution between syntenic blocks. We then investigated whether ohnologous transcript expression differences could cause the observed imbalance.

2 Method

A turnkey snakemake pipeline was written to compute Ka/Ks between triplets composed by Malus orthologous of each pair and a common reference (*Prunus persica*) as it was the closest related species without a WGD. Construction of triplets was made *via* a bi-directionnal best BLAST hit with a pair of ohnologous genes in apple and one gene in peach genome. To begin, by inputing homologous genes, a multiple alignment of the protein sequences is performed using MUSCLE. This proteic alignement is converted into a nucleic alignment using PAL2NAL. Finally, Ka/Ks rates are calculated using YN00 method from PAML. A paired t-test was set for a significant difference in Ka/Ks rate among triplets.

To investigate a potential transcription imbalance a mapping pipeline is used on all publicly available RNA-Seq runs that meet our quality criteria. Pseudo mapping are computed with Salmon and differential analysis with DESEQ2. Rather than testing two conditions against each other we have tested differential expression between pairs of ohnologous genes for each experimental condition.

3 Results

Ka/Ks pipeline was used on the apple gene sequences in order to determine whether the observed QTL imbalance could be explained by a different evolution rate of gene sequences. We conclude that the QTL imbalance cannot be explained by a sequence divergence between ohnologous genes.

We gathered 589 hight quality RNA-seq runs derived from 122 experiences. We tested transcriptionnal imbalance by testing if more than 50% of genes were significantly over-expressed in one chromosome pair. We show that several chromosome pairs including 1-7, 8-15, 4-12 6-14 and 2-15 are transcriptionally unbalanced, which is mostly consistent with the observed QTL imbalance. In addition, we identified 814 genes which expression is systematically higher in one of the two orthologues. In the future we plan to compare the evolution of transposable elements content located within syntenic blocks for which both QTL and transcriptional imbalance was identified.

References

[1] Pamela S Soltis and Douglas E Soltis. *Current Opinion in Plant Biology*, 30:159–165, 2016.

[2] Nicolas Daccord, Jean-Marc Celton, et al. *Nat Genet*, 49(7):1099–1106, 2017.

[3] Riccardo Velasco, Andrey Zharkikh, et al. *Nat. Genet.*, 42(10):833–839.

Network approaches and multi-omics integration applied to major depressive disorder.

Margot Derouin[1], Amazigh Mokhtari[1], El Chérif Ibrahim[2], Raoul Belzeaux [2], Bruno Etain[3], Cynthia Marie-Claire[3], Pierre-Eric Lutz[4,5], Andrée Delahaye-Duriez[1,6].

[1]NeuroDiderot - Inserm UMR 1141, Hôpital Robert Debré, Paris, France.
[2]Université Aix-Marseille, CNRS, Institut de Neurosciences de la Timone, Marseille, France.
[3]Université Paris Diderot, Sorbonne Paris Cité, Inserm, UMR-S1144, Paris, France.
[4]Institut des Neurosciences Cellulaires et Intégratives UPR 3212, CNRS, Université de Strasbourg, Fédération de Médecine Translationnelle de Strasbourg, Strasbourg, France.
[5]Douglas Mental Health University Institute, McGill University, Montréal, Canada.
[6]UFR Santé Médecine Biologie Humaine, Université Sorbonne Paris Nord, Bobigny, France.

margot.derouin@inserm.fr

Major Depressive Disorder (MDD) is a common and severe psychiatric disease that can be devastating, resulting in a higher risk of suicide and shorter life expectancy [1]. MDD is the most prevailing psychiatric syndrome in the general population, reaching a 12-month prevalence of 10.4% and a lifetime prevalence of up to 20.6% [2].

Despite the growing appeal of research towards molecular psychiatry, the etiology of MDD remains unclear with no identified blood biomarkers, mainly due to the complexity and heterogeneity of the disease and environmental interactions that play a significant role in triggering the symptoms.

Here, we took advantage of a new mRNA and miRNA genome-wide dataset, generated using RNAseq and miRNAseq, and peripheral blood samples from a cohort of N=80 MDD patients and N=89 healthy controls.

First, at the gene level, we characterized differentially expressed mRNAs between MDD and healthy control subjects, and obtained results significantly overlapped with a recently published meta-analysis [4]. Second, at the systems level, we performed a network analysis using the consensus Weighted Gene Correlation Network Analysis (WGCNA) method, in order to identify gene modules containing highly correlated and co-expressed genes. Functional annotations of differentially expressed gene and MDD-associated modules revealed enrichment for immune response processes, consistent with previous studies. Finally, we started implementing a MultiOmic integration approach, Similarity Network Fusion (SNF), [3] that combines mRNA and miRNA data in a non-linear fashion. This approach should allow for the identification of molecularly defined sub-groups of patients, potentially leading, in the long-term, to a more comprehensive understanding of pathophysiological pathways, a better patient stratification, and a more accurate prognosis of the illness.

References

1. AR Mathew, JW Pettit, PM Lewinsohn, JR Seeley, RE Roberts. Co-morbidity between major depressive disorder and anxiety disorders: shared etiology or direct causation?. Psychol Med. 2011;41(10).
2. DS Hasin, AL Sarvet, JL Meyers, TD Saha, WJ Ruan, M Stohl, BF Grant , (2018). Epidemiology of Adult DSM-5 Major Depressive Disorder and Its Specifiers in the United States. JAMA Psychiatry.
3. B Wang, AM Mezlini, F Demir, M Fiume, Z Tu, M Brudno, B Haibe-Kains & A Goldenberg , (2014). Similarity network fusion for aggregating data types on a genomic scale. Nature Methods 11, 333–337.
4. GM Wittenberg J Greene PE Vértescd WC Drevetse ET Bullmore, (2020).Major Depressive Disorder Is Associated With Differential Expression of Innate Immune and Neutrophil-Related Gene Networks in Peripheral Blood: A Quantitative Review of Whole-Genome Transcriptional Data From Case-Control Studies. Biological Psychiatry Volume 88, Issue 8, Pages 625-637.

Keywords: Major Depressive Disorder, multi-omics, transcriptomics, mirnomics, consensus WGCNA, SNF.

SVJedi-graph: Structural Variant genotyping with long-reads using a variation graph

Sandra Romain[1] and Claire Lemaitre[1]
Univ Rennes, Inria, CNRS, IRISA, 35000, Rennes, France

Corresponding author: claire.lemaitre@inria.fr

Abstract

Structural variants (SVs) are genomic segments of more than 50 bp that have been rearranged in the genome. The advent of third generation sequencing technologies has increased and enhanced their study, and a great number of SVs has already been discovered in the human genome. Complementary to their discovery, the genotyping of known SVs in newly sequenced individuals is of particular interest for several applications such as trait association and clinical diagnosis. Most of the SV genotypers currently available are designed for second generation sequencing data, although third generation sequencing data is more suited to study SVs due to their large range of sizes (up to few mega bases). As such, our team previously released SVJedi, the first SV genotyper dedicated to long read data[1]. The method is based on linear representations of the allelic sequences of each SV and each SV is represented and genotyped independently of the other ones. While this is very efficient for distant SVs, the method fails to genotype some closely located or overlapping SVs due to redundancy in representative allelic sequences.

To overcome this limitation, we present a novel approach, SVJedi-graph, which uses sequence graphs instead of linear sequences to represent the SVs. The use of sequence graphs to represent SVs for genotyping is fairly recent [2,3,4] and only designed for short-reads as for now. Here, we chose to represent only the SV sequences and that of the SV flanking regions in our graph, in order to reduce the long-read mapping time. This results in a variation graph composed of multiple connected components, each representing the possible alleles for a region of one, or several SVs in case of close SVs (less than 10 kb apart). In SVJedi-graph, the variation graph is built using VG toolkit[5] after a pre-processing step performed on the data. The long reads are then mapped on the graph using GraphAligner[6], and the mapping results are filtered to keep only the informative alignments. Finally, the genotype for each SV of the dataset is predicted using the estimation method implemented in SVJedi[1].

Tests on simulated long-reads on the human chromosome 1, with 1,000 deletions from the dbVar database, show a similar precision compared to SVJedi (98.1 %, against 97.8 %). Importantly, when additional deletions are added progressively closer to the original 1,000 in the dataset, SVJedi-graph maintains a 100 % genotyping rate with a high precision, when SVJedi is not able to assign a genotype to 21 % of the deletions when they are too close to each other (0-50 bp apart). SVJedi-graph also supports other SV types such as insertions and inversions, for which similar performances were obtained. We are planning to apply SVJedi-graph and to compare it to other approaches on real human re-sequencing data from the Genome In a Bottle consortium.

References

[1] L Lecompte et al. SVJedi: genotyping structural variations with long reads. *Bioinformatics*, 36(17):4568–4575, 2020.

[2] S Chen et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biology*, 20(1):291, 2019.

[3] H. P. Eggertsson et al. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature Communications*, 10(1):5402, 2019.

[4] G Hickey et al. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biology*, 21(1):35, 2020.

[5] E Garrison et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, 36(9):875–879, 2018.

[6] Mikko Rautiainen and Tobias Marschall. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biology*, 21(1):253, 2020.

Exploring the order and disorder content of *de novo* evolved proteins using structural signatures

Apolline Bruley, **Isabelle Callebaut and Elodie Duprat**

Sorbonne Université, Museum National d'Histoire Naturelle, UMR CNRS 7590, Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, IMPMC, Paris, France.

Corresponding Author: apolline.bruley@upmc.fr

De novo gene emergence, *i.e.* gene birth from previously non-coding regions, has been evidenced as an universal evolutionary process contributing to the adaptation of organisms, even viruses [1–8]. The structural properties of the corresponding proteins remain largely unexplored. Current debates concern the relationship between gene novelty and protein disorder [3–6]. *De novo* emergence of membrane proteins biased by nucleotide sequence composition was also discussed [5]. We wanted to give new insights by analyzing the structural signatures of a set of *de novo* proteins previously validated in the literature. We used an in-house approach allowing us to characterize, from the only information of protein sequences, their structural diversity from disorder to order in soluble and membrane environments.

We built a sequence dataset of reliable *de novo* proteins from different taxa (Viruses [2], Yeast [8] and *Oryza* [7]), as well as 4 reference non-redundant sequence datasets of: (i) soluble domains (from SCOPe [9] a, b, c, d classes), (ii) membrane domains (from SCOPe f class), (iii) disordered regions (from DisProt [10]) and (iv) a large dataset of non-categorized full-length proteins (UniProt [11]). We first delineated in these sequences the foldable segments based on the hydrophobic cluster density (correlated to the content in regular secondary structures) using the SEG-HCA tool [12]. Then, we refined this initial order segmentation using a sliding window estimating the content in strong hydrophobic amino acids, as well as hydrophobic cluster properties. This allowed us to distinguish soluble and membrane domains, as well as more disordered regions, within the foldable segments. Comparing the features of the *de novo* vs. reference datasets allows us to evaluate not only if *de novo* proteins are enriched in order or disorder, but also if they harbor taxonomic-specific or sequence-specific features which may rely on original structures.

The classification scheme of *de novo* proteins proposed by this work will open novel perspectives for high-throughput scanning of whole (meta)genomes, in order to explore molecular innovations without having to rely solely on homology searches in databases of known proteins.

References

1. Oss SBV, Carvunis A-R (2019) De novo gene birth. *PLOS Genet* 15: e1008160.
2. Rancurel C, Khosravi M, Dunker AK, et al. (2009) Overlapping Genes Produce Proteins with Unusual Sequence Properties and Offer Insight into De Novo Protein Creation. *J Virol* 83: 10719–10736.
3. Wilson BA, Foy SG, Neme R, et al. (2017) Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nat Ecol Evol* 1: 1–6.
4. Carvunis A-R, Rolland T, Wapinski I, et al. (2012) Proto-genes and de novo gene birth. *Nature* 487: 370–374.
5. Vakirlis N, Acar O, Hsu B, et al. (2020) De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nat Commun* 11: 1–18.
6. Bornberg-Bauer E, Hlouchova K, Lange A (2021) Structure and function of naturally evolved de novo proteins. *Curr Opin Struct Biol* 68: 175–183.
7. Zhang L, Ren Y, Yang T, et al. (2019) Rapid evolution of protein diversity by de novo origination in Oryza. *Nat Ecol Evol*.
8. Vakirlis N, Hebert AS, Opulente DA, et al. (2017) Yeast de novo genes preferentially emerge from divergently transcribed, GC-rich intergenic regions. *bioRxiv* 119768.
9. Fox NK, Brenner SE, Chandonia J-M (2014) SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res* 42: D304–D309.
10. Hatos A, Hajdu-Soltész B, Monzon AM, et al. (2020) DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res* 48: D269–D276.
11. The UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 49: D480–D489.
12. Faure G, Callebaut I (2013) Comprehensive repertoire of foldable regions within whole genomes. *PLoS Comput Biol* 9: e1003280.

Similarity estimation of PDX model multiOmic profiles with corresponding patient tumors.

Robin Droit[1], Jenny Costa[1], Antonin Marchais[1] and Nathalie Gaspar[1]

[1] Gustave Roussy institute, 114 Rue Edouard Vaillant, 94805, Villejuif, France

Summary

Osteosarcoma is the most frequent bone cancer in the adolescent and young adult population. [1] The relapse is the principal cause of mortality. Patient-derived xenograft (PDX) models are used to test treatments on this disease. The purpose of this study is to characterize molecularly the PDX models using a multiOmic approach. We investigate if the genomic and transcriptomic profiles are conserved from relapse to xenograft to identify PDX with promising characteristics as preclinical model.

In this study, we profiled tumors from 8 different patients sampled at diagnosis, relapse and in PDX models (orthotopic xenograft and subcutaneous xenograft). For each sample, Whole Exome Sequencing (WES) and RNA-seq have been produced, which results in 32 samples.

We analyzed the Somatic mutations, Copy number variations, Fusions and Expression profiles to dress the multiOmics landscape of every sample. Likewise, we estimated molecular similarities and discrepancies through time for each patient but also between all samples.

PDX samples are composed of a mixture of cells from two species and are usually produced in immune-depressed mice. Two features which introduce strong bias in comparison analysis with patient tumors. To minimize those bias, we developed a strategy based on Xenome [2] to identify the host or graft origin of the reads in RNA and WES data. For the RNA-seq, we, then, characterize in an unsupervised manner the genes not expressed in the mice due to immunodepression. Finally, we evaluated the genetic and transcriptomic proximity between the PDX and the human samples using unsupervised classification algorithms .

For the RNA-seq, the output of this method shows that the PDX samples are clusterizing with the corresponding relapse sample. However, the diagnostics samples are, in most cases, not clusterizing with the relapse and PDX samples, which suggest to the high difference between diagnosis and relapse Osteosarcomas.

For the WES, the similarity of genomic alteration is confirmed by the study. Due to the conservation of driver mutations the diagnostic samples are here clusterizing with the relapse and PDX samples. However, we observe a large amount of CNV and somatic mutations appearing in the relapse, which are conserved in the PDX models, supporting the clonal evolution model recently proposed by several group. [3]

In conclusion, most of the PDX models conserve the alterations driving the osteosarcoma and the expression profiles are similar when we remove the genes implied in the immunity. Which means that the PDX models have a high similarity with the human tumor and might be good preclinical model to test treatment response.

Acknowledgements

We thank the Preclinical Evaluation Platform for providing immunocompromised mice and animal care, the Imaging and Cytometry Platform for help on CTscan imaging.

References

[1] Trama, A. *et al.* Survival of European adolescents and young adults diagnosed with cancer in 2000–07: population-based data from EUROCARE-5. *Lancet Oncol.* (2016)

[2] Conway T, Wazny J, Bromage A, Tymms M, Sooraj D, Williams ED, Beresford-Smith B. Xenome--a tool for classifying reads from xenograft samples. Bioinformatics. (2012)

[3] Gambera, S., Abarrategi, A., González-Camacho, F. *et al.* Clonal dynamics in osteosarcoma defined by RGB marking. *Nat Commun* **9,** 3994 (2018)

Drawing interactive profiles of Percent Identical Positions with PIPprofileR – A Covid-19 use case

Thomas DENECKER[1,2], Hélène CHIAPELLO[1,2], Jacques VAN HELDEN[1,3]

1. CNRS, Institut Français de Bioinformatique, IFB-core, UMS 3601, Évry, France

2. Université Paris-Saclay, INRAE, MaIAGE, 78350 Jouy-en-Josas, France.

3. Aix-Marseille Univ, Inserm, laboratoire Theory and approaches of genome complexity (TAGC), Marseille, France.

Corresponding Author: thomas.denecker@france-bioinformatique.fr

Overview

Profiles of Percent Identical Positions (PIP) are widely used by virologists to analyse sequence similarities along viral genomes and detect recombinant regions. Noticeably, PIP profiles have been widely used in recent studies focusing on the SARS-CoV-2 origin [1,2]. However, existing tools propose restrictions for their installation (the most popular are Windows-specific) and user-friendliness. We will present PIPprofileR, a new web tool that provides a user-friendly and interactive interface to easily generate profiles of Percent Identical Positions (PIP) from a multi-sequence fasta file containing either nucleic or peptidic sequences. The results can be enriched with an annotation file to facilitate the exploration of regions of interest (i.e. annotated genes) and enhance the interpretation of the profiles.

Methods

PIPprofileR is built upon open-source technologies, written in R using the Shiny framework, and available on the collaborative development platform GitHub (https://github.com/IFB-ElixirFr/PIPprofileR/). It is developed in the open and licensed under the BSD 3-clause license. To ease the development and the deployment of the PIPprofileR, Docker is used to bundle the application as well as its dependencies. PIPprofileR is also available as an R package allowing the use of Shiny Proxy for deployment on a server.

In the demo

We will illustrate the use of PIPprofileR by showing how it can be used to compare several genomic and peptidic sequences of different coronaviruses with SARS-CoV-2 (considered here as the reference genome), to assess the degree of closeness in the different regions, and to identify abrupt changes in similarity likely to reflect recombinations between coronaviruses.

Keywords

Comparative genomics; PIP profiles; SARS-CoV-2; R; Shiny; Docker

References

1. E. Sallard, J. Halloy, D. Casane, J. van Helden, É. Decroly, Retrouver les origines du SARS-CoV-2 dans les phylogénies de coronavirus, Med Sci (Paris). 36 (2020) 783–796. https://doi.org/10.1051/medsci/2020123.

2. E. Sallard, J. Halloy, D. Casane, E. Decroly, J. van Helden, Tracing the origins of SARS-COV-2 in coronavirus phylogenies: a review, Environ Chem Lett. (2021). https://doi.org/10.1007/s10311-020-01151-1.

TF activity estimation sheds light on the role of TNF in response to immunotherapy in a mouse melanoma model

Matthieu Genais, Ludovic Martinet, Vera Pancaldi, Anne Montfort and Bruno Segui
CRCT, INSERM, 2 Avenue Hubert Curien, 31100 Toulouse, FRANCE

Corresponding author: matthieu.genais@inserm.com

1 Introduction

Immune checkpoint inhibitors (ICI) such as anti-PD-1 act on T cells to restore their ability to kill cancer cells. Cutaneous melanoma is a bad-prognosis skin cancer that can be treated by ICI. Despite major advances in the field of immunotherapy, melanoma kills more than half of all patients within 5 years of treatment induction due to primary or acquired resistance. Over the last 10 years, our team has identified a mechanism of resistance to immunotherapy that depends on the production of TNF, a major inflammatory cytokine that acts as a brake on the immune response against tumors [1][2].

2 materials and method/Results

In order to identify the impact TNF has on resistance to anti-PD-1 therapy, RNA-seq experiments were performed on mouse melanoma tumors. Tumors were collected 10 days post B16K1 melanoma cell injection in wild-type or TNF KO mice, injected with vehicle or anti-PD-1 at day 7. Using current methods such as immune cell deconvolution from this bulk data [3], as well as the study of the transcription factor activities modelled from RNAseq data [4], we were able to characterize the 4 groups. First, we determined that the absence of TNF could lead to a potentiation of the immune response, as illustrated by the activation of pathways such as "positive regulation of immune system process". Second, we were also able to confirm this potentiation with the anti-PD-1 treatment compared to the wild type, suggesting a potential synergy between inactivation of TNF and anti-PD-1. Finally, analysis of the transcription factors revealed higher activities of the trio STAT1, STAT2 and IRF9 in TNF KO groups compared to wild type groups. These factors are components of the ISGF3 complex (interferon-stimulated genes), which plays an essential role in various immune responses, including the activation of immune cells, immune cell propagation, and inflammatory cytokine production.

3 Conclusion

In conclusion, these preliminary results provide interesting leads for the study of biological samples from patients enrolled in clinical trials launched recently in our team, notably producing single cell transcriptomics data in patients treated with anti-PD-1 and anti-CTLA-4 in combination or not with TNF blockers with known response.

References

[1] Florie Bertrand, Anne Montfort, Elie Marcheteau, Caroline Imbert, Julia Gilhodes, Thomas Filleron, Philippe Rochaix, Nathalie Andrieu-Abadie, Thierry Levade, Nicolas Meyer, Céline Colacios, and Bruno Ségui. TNF blockade overcomes resistance to anti-PD-1 in experimental melanoma. 8(1):2256. Number: 1 Publisher: Nature Publishing Group.

[2] Florie Bertrand, Julia Rochotte, Céline Colacios, Anne Montfort, Anne-Françoise Tilkin-Mariamé, Christian Touriol, Philippe Rochaix, Isabelle Lajoie-Mazenc, Nathalie Andrieu-Abadie, Thierry Levade, Hervé Benoist, and Bruno Ségui. Blocking tumor necrosis factor enhances CD8 t-cell-dependent immunity in experimental melanoma. 75(13):2619–2628.

[3] Francisco Avila Cobos, José Alquicira-Hernandez, Joseph E. Powell, Pieter Mestdagh, and Katleen De Preter. Benchmarking of cell type deconvolution pipelines for transcriptomics data. 11(1):5650. Number: 1 Publisher: Nature Publishing Group.

[4] Mariano J. Alvarez, Yao Shen, Federico M. Giorgi, Alexander Lachmann, B. Belinda Ding, B. Hilda Ye, and Andrea Califano. Network-based inference of protein activity helps functionalize the genetic landscape of cancer. 48(8):838–847.

Greening R. Thomas' Framework: a Divide and Conquer Approach

Laetitia Gibart[1], Hélène Collavizza[1]and Jean-Paul Comet[1]
Université côte d'azur, I3S laboratory, UMR CNRS 7271, CS 40121, 06900 Sophia Antipolis Cedex, France

Corresponding author: Laetitia.gibart@univ-cotedazur.fr

When we model a complex biological system, we often try to understand the causality chains that explain the different behaviours observed. Qualitative models based on discrete mathematics tend to be powerfull to give this type of answer: A reason why the extended R.Thomas formalism completed by formal methods, has become a classic for regulatory networks [1,2].

However, observed behaviours often depend on environmental conditions (like nutrient availability in cell culture experimentation). Therefore, the construction of a right modelisation depends on our ability to take into account all this environemental information in a single modelling framework.

Moreover, even in a given modelisation framework, several modelling choices are possible. This is due to different instantiations of dynamical parameters piloting the behaviour of the model, that can lead to traces consistent with all observations. If the modeller chooses a particular setting, when new biological information is known about the system, the parameter identification step must be restarted from the beginning. The systematic approach would then consist in characterizing, at each step, all of the parameter settings consistent with current knowledge: when a new observation becomes available, the modeller just refines the previous set of consistent parameter settings by selecting only those that are also consistent with this new information.

The use of artefacts enables the simulations of successive environmental situations in a unique global network with the R. Thomas modelling framework. However, we recommend another option based on a "divide and conquer" approach: the green extension of R. Thomas' framework with the notion of environments. This approach has several steps. First, a specific (and thus smaller) regulatory network is built for each environment. Then, for each regulatory network, a consistent set of parameter settings compatible with the associated biological properties is searched. Finally, all consistent sets are intersected to obtained the settings which satisfy the properties for all environments[3].

This poster will show you that the addition of a new environmental context (calcium addition) in a running example of the *Pseudomonas aeruginosas* virulence regulation model. Pseudomonas is an opportunistic bacteria which can cause serious infections in the lung of cystic fibrosis patients with the production of thick mucus [4].

Acknowledgements

We are fully indebted to Gilles Bernot for his help to find how to use artefact with R. Thomas modelling framework in the first place.

References

[1] R. Thomas. Boolean formalization of genetic control circuits. *J.Theor.Biol*, 42(3):563–585, 1973.

[2] Gilles Bernot, Jean-Paul Comet, Adrien Richard, and Janine Guespin. Application of formal methods to biological regulatory networks: extending Thomas' asynchronous logical approach with temporal logic. *J.Theor.Biol*, 229(3):339 – 347, 2004.

[3] L. Gibart, H. Collavizza, and J.-P. Comet. Greening R. Thomas' framework with environment variables: a divide and conquer approach. *submitted*, 2021.

[4] Sankalp Malhotra, Don Hayes, and Daniel J. Wozniak. Cystic Fibrosis and Pseudomonas aeruginosa: the Host-Microbe Interface. *Clinical Microbiology Reviews*, 32(3), June 2019.

Integrated multi-omics of human early breast milk constituents

Camilo Broc[1], Karine Adel-Patient[2], François Fenaille[3], Mikaïl Berdi[2], Eric Venot[2], Blanche Guillon[2], Florence Castelli[3], Benoit Colsch[3] and Etienne Thevenot[3]

[1] CEA-LIST, Gif-sur-Yvette, France
[2] CEA-LIAA, Gif-sur-Yvette, France
[3] CEA-LEMM Gif-sur-Yvette, France

Corresponding author: camilo.broc@gmail.com

The impact of breast feeding on child health and development is clearly established in the literature [1]. However, the knowledge of all of its constituents and their post-partum evolution remains poorly understood. During the last decade, the number of molecules detected in human tissues and fluids is ever growing, thanks to constant instrumental developments such as those related to mass spectrometry. To give new insights about human EBM composition, we aim at combining the data obtained from four different types of molecular families assayed in EBM and analyzed though suitable multi-omics statistical tools.

Milk samples (n=257) were collected from days 2 to 6 within the EDEN mother-child cohort [2,3]. Untargeted analyses of Human Milk Oligosaccharides (HMOs), lipids and metabolites were performed using liquid chromatography coupled to high-resolution mass spectrometry, while targeted analysis of a large panel of cytokines, growth factors and antibodies was achieved thanks to multiplexed ELISA assays. A reusable workflow based on multi-omic R packages is proposed in order to detect groups of correlated features that are related to a day by day temporal evolution. In the data pre-processing steps, a strong effect of the collection place has been handled and corrected. Single-omic data analysis was first performed to assess the evolution of each type of molecules family independently (univariate hypothesis testing, PLS-DA modeling). We then used multi-block approaches including dimension reduction methods (multi-block PLS-DA) and clustering (Weighted Graph Correlation Network Analysis) to highlight potential associations between blocks of variables.

We evidenced that HMOs, lipids and metabolites have stronger temporal variations than cytokines. Especially, a large number of HMOs and lipid concentrations (around 50%) increase over time. Interestingly, multi-block methods infer associations between families of molecules, notably between cytokines and specific metabolites.

This study expands our knowledge about EBM composition. Combination of various omics approaches provides an unprecedented wide view of the biochemical composition of BM and new associations have been assessed, especially for metabolites. The further association of global milk composition with mother exposure or with infant health outcomes could lead to establishing relevant biomarkers.

References

[1] AnnA Petherick. Development: mother's milk: a rich opportunity. *Nature*, 468(7327):S5–S7, 2010.

[2] Barbara Heude, Anne Forhan, Rémy Slama, Lorraine Douhaud, Sophie Bedel, Marie-Josèphe Saurel-Cubizolles, Régis Hankard, Olivier Thiebaugeorges, Maria De Agostini, Isabella Annesi-Maesano, et al. Cohort profile: The eden mother-child cohort on the prenatal and early postnatal determinants of child health and development. *International Journal of Epidemiology*, 45(2):353–363, 2016.

[3] Mikaïl Berdi, Blandine de Lauzon-Guillain, Anne Forhan, Florence Anne Castelli, François Fenaille, Marie-Aline Charles, Barbara Heude, Christophe Junot, Karine Adel-Patient, and EDEN Mother-Child Cohort Study Group. Immune components of early breastmilk: Association with maternal factors and with reported food allergy in childhood. *Pediatric Allergy and Immunology*, 30(1):107–116, 2019.

TrEMOLODyn: how to monitor transposable element dynamics over generations?

Marion VAROQUI[1,2], Mourdas MOHAMED[3], Séverine CHAMBEYRON[3] and Anna-Sophie FISTON-LAVIER[2]

[1] Master Sciences et Numérique pour la Santé, Parcours Bioinformatique, Connaissances, Données, Université Montpellier, Montpellier, France

[2] ISEM, Université Montpellier, CNRS, UM, IRD, CIRAD, EPHE, Montpellier, France

[3] IGH, UM, Montpellier, France

Corresponding author: marion.varoqui@etu.umontpellier.fr

1 Abstract

Transposable Elements (TEs) are parasitic elements that are able to multiply within their host genome. Despite the fact that they are mostly deleterious, TEs and their underlying variations are drivers of adaptation [1]. Studying their dynamics can help to better understand their impacts on population dynamics and emergence of resistances [2]. The main objective of our study is to analyse data obtained from experimental evolution on *Drosophila melanogaster* to prospect the correlation between the dynamics of insertion and deletion of TEs and stress. Here we use a thermic shift of 5°C that inhibits the PiWI system, a system that block TE transposition [3]. Due to their repetitive nature, the use of the third generation sequencing technologies that allow to generate long-reads, is more appropriate than short-reads (Next-generation sequencing). To carry out an exhaustive inventory of insertions and deletions of TEs over 100 generations, bulks of 100 *Drosophila melanogaster* were pooled and sequenced. We sequenced with long-reads four different generations with the Oxford Nanopore technology (G30, G60, G80, G100). The pooled data were analysed with TrEMOLO, a bioinformatic tool that detects the presence and the absence of TEs with long-reads [4](*cf.* TrEMOLO poster #182). TrEMOLODyn is an add-on based on TrEMOLO output to deal with our problematic, the monitoring of TEs over the generations. The TrEMOLODyn is a Snakemake pipeline, a suite of Python scripts (available on https://github.com/MaVaroqui/TrEMOLODyn). This tool allows to report the transposition, insertion and the deletion rate in fine scale.

Over generations, as expected some known TE families show the strongest TE activity (*i.e.* ZAM and Gtwin). We confirm new TE insertions by detecting target site duplication that are hallmarks of new transposition events. Based on the RNA-seq analysis, we were able to show the impact of ZAM insertions on some gene expression. Preliminary results also highlight a slight decrease of the transposition rate after 80 generations suggesting the existence of another defense mechanism on top of the PiWI system that takes over.

With the emergence of new sequencing technologies, the tools that analyse their data and the interest on TEs, we can expect that a tool like TrEMOLODyn should be very helpful to the community.

References

[1] Rech G. E., Bogaerts-Márquez M., Maite G. Barrón, M. Merenciano, J. L. Villanueva-Cañas, Horváth V., Fiston-Lavier A.-S., Luyten I., Venkataram S., Quesneville H., Petrov D.A., and González J. *Stress response, behavior, and development are shaped by transposable element-induced mutations in Drosophila*. 2019.

[2] Benoît Chénais, Aurore Caruso, Sophie Hiard, and Nathalie Casse. The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. *Gene*, 509(1):7–15, November 2012.

[3] Bridlin Barckmann, Marianne El-Barouk, Alain Pélisson, Bruno Mugat, Blaise Li, Céline Franckhauser, Anna-Sophie Fiston Lavier, Marie Mirouze, Marie Fablet, and Séverine Chambeyron. The somatic piRNA pathway controls germline transposition over generations. *Nucleic Acids Research*, 46(18):9524–9536, October 2018.

[4] Mourdas Mohamed, Nguyet Thi-Minh Dang, Yuki Ogyama, Nelly Burlet, Bruno Mugat, Matthieu Boulesteix, Vincent Mérel, Philippe Veber, Judit Salces-Ortiz, Dany Severac, Alain Pélisson, Cristina Vieira, François Sabot, Marie Fablet, and Séverine Chambeyron. A Transposon Story: From TE Content to TE Dynamic Invasion of Drosophila Genomes Using the Single-Molecule Sequencing Technology from Oxford Nanopore. *Cells*, 9(8):1776, July 2020.

Study of variations associated with transposable elements in *Anopheles gambiae* natural populations

Clothilde Chenal[*1,2,3], Corentin Marco[*1,4], The Anopheles gambiae 1000 genomes consortium[5], Frédéric SIMARD[3], François Sabot[2,6], Michael C. Fontaine[5,7,8] and Anna-Sophie Fiston-Lavier[1]

[1] ISEM, Univ Montpellier, CNRS, UM, IRD, CIRAD, EPHE, Montpellier, France
[2] DIADE, Univ Montpellier, CIRAD, IRD, Montpellier, France
[3] MIVEGEC, Univ Montpellier, CNRS, IRD, Montpellier, France
[4] Master SNS, parcours Bioinformatique, Connaissances, Données, Univ de Montpellier, France
[5] https://www.malariagen.net/projects/ag1000g
[6] IFB, Southgreen Bioversity, CIRAD, INRAE, IRD, Montpellier, France
[7] Groningen Institute for Evolutionary Life Sciences (GELIFES), Univ Groningen, Groningen, Netherlands
[8] Centre de Recherche en Ecologie et Evolution de la Santé (CREES), Montpellier, France
(*) co-first authors

Corresponding authors: clothilde.chenal@ird.fr & corentin.marco@etu.umontpellier.fr

Malaria affects more than 228 millions of people around the world per year. In Africa, *Anopheles gambiae sensu stricto* is the flagship species of the *Anopheles gambiae* species complex, which includes the principal vector species of the *Plasmodium sp.* parasite, responsible for the disease. Understanding how genetic variation is shaped by environmental changes will allow improving our knowledge about how *An. gambiae* populations are spatially structured, how they have evolved, and the genetic processes involved in their dramatic adaptive abilities to local environmental variation. Taking advantage of the sequencing data from the MalariaGEN's Ag1000G consortium project, we aim to study population genetic variations that may be linked with local adaptations to environmental variation and stress (*e.g.* insecticides, pesticides, other environmental stress...). Previous studies focused mainly on single nucleotide polymorphisms (SNPs), however they represent only one type of genetic variation. Here, we focus on transposable elements (TEs, around 20% of the genome *An. gambiae* species complex; [1]), since they are source of genetic novelty and diversity [2]. To study the association between TE variations, population dynamics and local adaptations, we analyze new TE insertions, the more active ones, in *An. gambiae* mosquitoes in Cameroon populations. More than 300 mosquitoes were sampled in three different ecological conditions (forest, savanna and transition zone) and sequenced using Illumina (100bp paired-end reads). However, due the repetitive nature of TEs, computational tools implemented to detect new TE insertions with short-read data fail to limit the false-positive rate [3]. One solution to overcome this issue is to combine the calls from tools based on different approaches. After the selection of three computational tools dedicated for the detection of new TEs insertions with high accurate and complementary results: TEFLoN [4], PoPoolationTE2 [5] and Jitterbug [6], we implement a Snakemake Python pipeline to automatically combine the results from several computational tools by sample, and then for all the samples. The aim of the pipeline is to characterize accurately the shared and not shared TE insertions between our three natural populations. The results will be then validated by comparing the TE calls against the dispensable variations obtained using a pangenomic approach and previous studies. Thus, we hope to provide a catalog of the more active TE insertions associated with the environmental cline.

References

[1] Elverson Soares de Melo and Gabriel Luz Wallau. Mosquito genomes are frequently invaded by transposable elements through horizontal transfer. *PLOS Genetics*, 16(11):e1008946, 2020.
[2] Rech G. E., Bogaerts-Márquez M., Maite G. Barrón, M. Merenciano, J. L. Villanueva-Cañas, Horváth V., Fiston-Lavier A.-S., Luyten I., Venkataram S., Quesneville H., Petrov D.A., and González J. *Stress response, behavior, and development are shaped by transposable element-induced mutations in Drosophila.* 2019.
[3] Vendrell-Mir P., Barteri F., Merenciano M., González J., J. M. Casacuberta, and Castanera R. A benchmark of transposon insertion detection tools using real data. *Mobile DNA*, 10(1):1–19, 2019.
[4] Adrion J.R., Song M.J., Schrider D.R., M.W. Hahn, and Schaack S. Genome-wide estimates of transposable element insertion and deletion rates in drosophila melanogaster. *Genome Biology and Evolution*, 9(5):1329–1340, 2017.
[5] Kofler R., Gómez-Sánchez D, and Schlötterer C. PoPoolationTE2: Comparative Population Genomics of Transposable Elements Using Pool-Seq. *Molecular Biology and Evolution*, (10):2759–2764, 2016.
[6] Hénaff E., Zapata L., J.M. Casacuberta, and Ossowski S. Jitterbug: Somatic and germline transposon insertion detection at single-nucleotide resolution. *BMC Genomics*, 16(1):1–16, 2015.

New strategy for optimizing knowledge-based docking parameters: application to ssRNA-protein docking

Anna KRAVCHENKO[1], Malika SMAIL-TABBONE[1], Isaure CHAUVOT DE BEAUCHENE[1] and Sjoerd Jacob DE VRIES[2]

[1] LORIA,campus scientific, 54000, Nancy, France.
[2] RPBS, 35 rue Hélène Brion, 75013, Paris, France

Corresponding Author: anna.kravchenko@loria.fr

Computational prediction of a 3D structure of a molecular complex, also known as *docking*, is essential in modern biological research. It can complement MD, provide working directions to experimentalists, etc. We are invested in fragment-based docking, specifically for the single-stranded RNA-protein complexes. Why ssRNA specifically? Generally speaking, molecular flexibility is a scourge of docking: it increases its complexity and decreases results' reliability. High flexibility leads to a near-infinite number of docking models, the processing of which is too expensive computationally. Hence, highly flexible ssRNA is a challenging target to work on.

A fragment-based docking approach was developed to tackle this high flexibility issue [1]. Its core idea is to split the ligand into overlapping fragments, and dock them onto the rigid receptor separately, assembling the fragments back into the whole ligand afterwards. For the full procedure to succeed, each fragment must return at least one correct pose (so-called near-native): this is the *sampling* problem. The poses are obtained by minimisation using a differentiable *energy function*. Then, before assembling, docked fragments must be filtered, keeping a high percentage of near-natives. Otherwise, the assembly task once again becomes too expensive computationally: this is the *scoring* problem. The filtration is done using a *scoring function*. We are working with the ATTRACT docking engine, where the same function is used both for sampling and scoring. It has the shape of a Lennard-Jones potential, and 2 parameters per atom type pair (1054 in total). The current parameters were obtained in 2010 by extraction of the statistical potentials from RNA-protein crystal structures and were optimized by a random Monte Carlo-like strategy [2].

These parameters were not initially tailored to ssRNA and their performance is not flawless. Our goal is to optimize docking parameters, improving both sampling and scoring performance. To achieve it, we created an up-to-date dataset of ssRNA-protein complexes and set up a novel histogram-based approach. For each pair of interacting atom types, we (a) convert the current energy function into a log-odds histogram of the expected occurrences of atom-atom distances (discretized into bins) in native/non-native poses, using the Boltzmann equation; (b) obtain the corresponding histogram on a benchmark-wide docking test, which corresponds to the residual error of the energy function; (c) sum the predicted and real histograms; (d) analytically fit the energy parameters to the resulting histogram. Repeat until convergence - until the residual histogram is flat.

Our newly created dataset of ssRNA-protein complexes is expected to be sufficient for optimization. The proposed approach for the ssRNA-protein fragment-based docking parameter optimization is expected to be more robust in terms of the fitness of the initial parameters compared to the previous approach. It has potential to benefit both the sampling and the scoring problems.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 813239.

References

[1] Isaure Chauvot de Beauchene, Sjoerd Jacob de Vries, Martin Zacharias. Fragment-based modelling of single stranded RNA bound to RNA recognition motif containing proteins. Nucleic Acids Res., 44(10):4565-4580, 2016.

[2] Piotr Setny, Martin Zacharias. A coarse-grained force field for Protein-RNA docking. Nucleic Acids Res., Vol. 39, No. 21, 9118–9129, 2011.

deCIPHER: A pipeline for amplicon-based nanopore sequencing for tracking SARS-CoV-2 variants

Fatou Seck Thiam[1,2], Marine Combe[2], Georgina Rivera-Ingraham[3], Fabienne Justy[2], Damien Breugnot[2], Marie-ka Tilak[2] , Mohammad Salma[4,5], Jean-Claude Doudou[6] , Rodolphe Elie Gozlan[2], Emira Cherif [2]

[1]Master Bioinformatique, Connaissances, Données (BCD), Faculté des Sciences de Montpellier, Montpellier, France

[2]ISEM , IRD, CNRS, IPHE, Université de Montpellier, Montpellier, France

[3]ISEM , IRD, CNRS, IPHE, Université de Montpellier, Centre IRD de Cayenne, Guyane Française

[4]Institut de Génétique Moléculaire de Montpellier, Université de Montpellier, CNRS, Montpellier, France.

[5]Université de Paris, Laboratory of Excellence GR-Ex, Paris, France.

[6]IRD, Centre IRD de Cayenne, Guyane Française

Assessment of the dynamics of SARS-CoV-2 infection in human populations is limited to epidemiological data collected in an emergency setting that imperfectly reflects the actual dynamics of the COVID-19 disease and the viral strain diversity. SARS-CoV-2 is released in human feces and then discharged into sewage water or even in the environment in areas under poor sanitary conditions. Thus, wastewater-based epidemiology (WBE) is a valuable population-level approach to monitoring the viral variants' emerging and turnovers. Yet, the major efforts (epidemiology, genomics, bioinformatics. etc.) are oriented towards clinical data and approaches for environmental data are still lacking. Here, we present deCIPHER, a pipeline for amplicon-based Oxford Nanopore Technology (ONT) sequencing to analyze and assess SARS-CoV-2 genetic diversity in wastewater. The pipeline integrates 11 bioinformatics tools, including Seqkit, ARTIC bioinformatics tool, MultiQC, Minimap2, Medaka, Nanopolish, Pangolin (with the model database pangoLEARN), Deeptools (PlotCoverage, BamCoverage), MAFFT, RaxML and snpEff. deCIPHER is a standalone pipeline compatible with Ubuntu distributions, implemented in Python, including an easy-to-use configuration file. With a single command line and the raw sequencing data as input, the user can preprocess the data, obtain the statistics on sequencing quality, depth and coverage. Then, reconstruct the consensus genome sequences, identify the variants and their potential associated effects for each viral isolate, and, finally, perform the multi-sequence alignments and phylogenetic analyses. deCIPHER was used to analyze wastewater data sampled in more than ten localities in French Guiana from October 2020 to May 2021 and generated by adapting the ARTIC amplicon method and the ONT sequencing. Clinical data obtained by the same approaches were also used for the comparative analysis. Preliminary results showed considerably high variant turnovers over time and space. Among them, in the 2020 fall, the B.1.219 lineage from Suriname was the major lineage in wastewater and clinical samples. Later in winter and early spring 2021, the Guyanese lineage B1.160.25 became the dominant variant in most localities associated with a Brazilian variant P2. deCIPHER is intended to provide an additional tool for the WBE approaches to monitor at a large scale the diversity dynamics of the SARS-CoV2. It should be thus considered as a powerful tool for large disease surveillance of a multitude of water-linked emerging pathogens.

Keywords: SARS-CoV-2, variants detection, viral genomics analysis tool, wastewater-based epidemiology

Improving genome annotations with RNA-seq data: the TAGADA pipeline to combine transcript reconstruction and expression assessment

Cyril Kurylo[1], Cervin Guyomar[1], Sarah Djebali[2*], Sylvain Foissac[1*]

[1]GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France

[2]IRSD, Université de Toulouse, INSERM, INRAE, ENVT, UPS, Toulouse, France

Corresponding Authors: sylvain.foissac@inrae.fr, sarah.djebali@inserm.fr

In large scale genome annotation projects, multiple RNA-seq datasets are typically generated and/or analyzed to characterize the transcriptome of various tissues or cell types. While many RNA-seq pipelines are currently available to build de novo gene models or quantify gene expression levels using a provided gene annotation, few allow both transcript reconstruction and expression assessment from RNA-seq data in a reproducible way. To fill in this gap in the context of the EU H2020 GENE-SWitCH project, we have developed the TAGADA RNA-seq pipeline, for Transcripts And Genes Assembly, Deconvolution and Analysis.

TAGADA combines several reference RNA-seq bioinformatics tools into a containerized pipeline. It maps reads with STAR, reconstructs and quantifies genes and transcripts with StringTie and detects and characterizes long non-coding RNAs (lncRNAs) with FEELnc.

TAGADA uses the Nextflow framework in line with the nf-core specifications. It provides a containerized environment that makes it compatible with a variety of high-performance computing platforms and workload orchestrators. It is designed to be easy to use and flexible. As such it requires a minimal set of inputs: a set of RNA-seq read files, a reference genome and its gene annotation. Optionally, a simple tabulated metadata file can also be provided to describe the experimental design and seamlessly merge samples according to specified factors.

The pipeline automatically generates a large variety of quality controls in the form of interactive charts and tables with statistics and metrics for various steps of the workflow. Expression tables are also provided with read counts for annotated and predicted genes and transcripts allowing further comparative expression analyses. We believe that the TAGADA pipeline offers a useful, powerful and easy-to-use way to process RNA-seq data, nicely complementing the existing nf-core catalogue of bioinformatics tools and contributing to the FAANG global action.

“This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the grant agreement n° 817998”.

A user-friendly web application to visualise genomic variation effects at gene and protein level

Laura DIEZ[1], Christophe DUNAND[1] and Catherine MATHÉ[1]

[1] Laboratoire de Recherche en Sciences Végétales, Université de Toulouse, CNRS, UPS, 24 chemin de Borde Rouge, 31320 Auzeville-Tolosane, France

Corresponding Author: laura.diez@lrsv.ups-tlse.fr

1. Introduction

Genetic variations are permanent changes in the DNA sequence that determine how diverse individuals or populations are from each other. The 1001 Genomes Project started in 2008 to discover detailed whole-genome sequence variation in at least 1001 geographically diverse *Arabidopsis thaliana* populations or accessions. In 2016 the detailed analysis of 1135 genomes have been published [1].

Although there exist several tools to visualise mutations on the genome such as ProteinPaint [2] or SnpHub [3], they do not suit the *A. thaliana* data that are available or do not allow the interactivity that is desired. Therefore, in order to visualize the available *A. thaliana* SNPs and InDels data from 1001 Genomes Project [4], a new tool has been created.

2. Tool description

An interactive web app, with three main functionalities, has been created using the R package Shiny. This app is inspired by already existing tools such as SnpHub and it uses functions of different R packages such as Trackviewer, GenomicRanges, Leaflet and RMyQSL.

2.1 Gene variants visualisation

The display takes a lollipop plot shape to represent the mutations either on genes or on proteins. The user can choose which gene or protein he wants to visualise and if he wants to visualise the whole structure or zoom in into a specific region. Together with the variants, the exon-intron gene structure and protein domains (PFAM or Prosite) can also be displayed. At the protein level, the user can choose various outputs such as which variant effects (for example, high impact as frameshifts or stop codon gain or loss), or amino acid changes type (synonymous codon, and amino acid similarity).

2.2 Geographic location

It displays, according to the user choices, all the populations or only those that present a certain mutation (specifying its chromosome and position) on an interactive map. Soon, geographic locations should be associated with climate information (temperature, rain) for further interpretation of variations.

2.3 Browse/download data

This serves to easily analyse and/or download the data from the database either as a table of variations or generated sequences (at DNA or protein level).

References

1. Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KM, Cao J, Chae E, Dezwaan TM, Ding W, *et al* (2016) 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* 166: 481–491
2. Zhou X, Edmonson MN, Wilkinson MR, Patel A, Wu G, Liu Y, Li Y, Zhang Z, Rusch MC, Parker M, *et al* (2016) Exploring genomic alteration in pediatric cancer using ProteinPaint. *Nat Genet* 48: 4–6
3. Wang W, Wang Z, Li X, Ni Z, Hu Z, Xin M, Peng H, Yao Y, Sun Q & Guo W (2020) SnpHub: an easy-to-set-up web server framework for exploring large-scale genomic variation data in the post-genomic era with applications in wheat. *GigaScience* 9: giaa060
4. Savelli B, Li Q, Webber M, Jemmat AM, Robitaille A, Zamocky M, Mathé C & Dunand C (2019) RedoxiBase: A database for ROS homeostasis regulated proteins. *Redox Biology* 26: 101247

A bioinformatic pipeline to elucidate the links between viruses and their hosts in microbial communities, applied to viruses in anaerobic digestion processes

Vuong Quoc Hoang NGO[1], Cédric MIDOUX[1,2,3], Mahendra MARIADASSOU[2,3], Valentin LOUX[2,3], François ENAULT[4], Ariane BIZE[1]

[1] Université Paris-Saclay, INRAE, PRocédés biOtechnologiques au Service de l'Environnement, 1 rue Pierre-Gilles de Gennes, CS10030, 92761, Antony, France.

[2] Université Paris-Saclay, INRAE, MaIAGE, Domaine de Vilvert, 78350, Jouy-en-Josas, France.

[3] Université Paris-Saclay, INRAE, BioinfOmics, MIGALE bioinformatics facility, Domaine de Vilvert, 78350, Jouy-en-Josas, France.

[4] UMR CNRS 6023 Microorganismes : Génome et Environnement, Bât. A, 24 avenue des Landais, 63177, Aubière, France.

Corresponding Author: ariane.bize@inrae.fr

Viruses are key-players in microbial ecosystems. However, predicting hosts from viruses is still a major challenge in microbial ecology. A few *in silico* methods for metagenomics data have proven useful for this purpose (e.g. in [1]), and they are highly invaluable when studying environmental samples. We developed a bioinformatic pipeline including the detection of CRISPR protospacers in viral contigs, a method previously used to predict hosts from marine viruses [1]. We applied our pipeline to anaerobic digestion (AD) ecosystems, in the context of organic waste treatment and valorisation. We focused on the diversity of viruses infecting methanogens, the latter being the key actors of methane production during AD. Viral diversity is only starting to be explored in AD processes [2], hence the great potential of new virus discovery in our study.

After enrichment of methanogenic archaea in AD microcosms by growth on formate as the sole carbon source, 2 DNA metaviromes and 5 cellular metagenomes were sequenced using Illumina NextSeq. Our pipeline was applied to all the obtained data. It was executed on the cluster of the INRAE MIGALE bioinformatics platform.

The most generic steps of our pipeline were scripted as a *snakemake* workflow, to favour reproducible and scalable data analysis (https://forgemia.inra.fr/cedric.midoux/workflow_metagenomics). After a pre-processing step, reads were assembled with *metaSPADES*. Coding regions were predicted with *Prodigal*. Taxonomic assignation of the contigs and of their predicted genes was obtained with *kaiju* against *NCBI nr* database. Functional annotation of the predicted genes was performed with *Diamond* against *Phrogs* (https://phrogs.lmge.uca.fr/), a database dedicated to prokaryotic viruses, and with *ghostKoala* against *KEGG* database (https://www.kegg.jp/ghostkoala/). For each dataset, reads were mapped to the assembled contigs.

Several steps specifically dedicated to the prediction of hosts from viral contigs were performed using *bash* and *python* scripts. For the cellular metagenomes, spacers were detected in contigs with *CRISPRdetect* and *CRISPRCasFinder*. A non-redundant spacer database was built from the obtained spacer sequences. The viral contigs were subsequently aligned with *blastn* against this specific database, enabling host prediction. In addition, metagenome-assembled genomes (MAGs) were constructed from cellular metagenomic data with *Metabat2*. Their quality was improved with *RefineM* and controlled with *CheckM*.

Thanks to this spacer-based approach, we were able to identify 77 viral contigs possibly originating from methanogenic archaea. We are currently further analysing them to confirm their nature and to study their gene content. The MAG reconstruction yielded 15 methanogenic archaea genomes. Thanks to these latter, we will search for archaeal proviruses with *Phaster* and *VirSorter* and we will also use a k-mer based method to identify additional putative archaeal viruses, using the tool *WiSH*.

References

1. Felipe H Coutinho, Cynthia B. Silveira, Gustavo B. Gregoracci, Cristiane C. Thompson, Robert A. Edwards, Corina PD Brussaard, Bas E. Dutilh, and Fabiano L. Thompson. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nature communications* 8, no. 1: 1-12, 2017.
2. Magdalena Calusinska, Martyna Marynowska, Xavier Goux, Esther Lentzen, and Philippe Delfosse. Analysis of ds DNA and RNA viromes in methanogenic digesters reveals novel viral genetic diversity. *Environmental microbiology* 18, no. 4: 1162-1175, 2016.

Contig error correction and scaffolding based on linked-read sequencing data

Andreea Dréau[1], Clément Birbès[1], Christophe Klopp[1] and Matthias Zytnicki[1]
INRAe, Unité de Mathématiques et Informatique Appliquées de Toulouse, Castanet-Tolosan, France

Corresponding author: andreea.dreau@inrae.fr

One of the main steps in genome assembly is contig assembly, which consists in reconstructing long and contiguous chromosomal parts based on the overlaps between the reads. The latest sequencing advances allow the construction of longer and more accurate contigs, but misassemblies are still present due to repeat sequences, heterozygosity and read errors. A technique that can be used for identifying these misassemblies is linked read sequencing since it provides long-range and low-error information. This type of sequencing is already used for correcting contigs by Tigmint [1], a tool that splits the contigs in loci with low molecule coverage. However, in case of contigs built from long reads and with the latest assemblers, the coverage drop is no longer sufficient for detecting misassemblies.

In this study we introduce a new correction pipeline adapted to more accurate contigs. In the first step the connecting errors are detected by analyzing linked reads molecules profiles for each contig. We start by aligning the linked reads to the contigs and identifying the molecules by grouping reads with the same barcode and aligned in the same region. Then our method computes several metrics, such as the molecule coverage, the mean read density per molecule and the mean molecule length, per 10kb window. For each metric we identify the outlier values and we split the contig if an interval is considered as outlier for at least two metrics. We tested the method by scaffolding several bovine assemblies with 3d-dna [2] and different Hi-C libraries. 3d-dna was able to connect more contigs into scaffolds with up to 80% increase in scaffold N50 and even obtain complete chromosomes when applied on contigs split with our method.

The main drawback of splitting the contigs before scaffolding them with Hi-C based methods is the increase in the number of very short contigs. These contigs will not have enough Hi-C contact points to be correctly oriented in a scaffold or even to be connected to another contig. Thus the number of inversions or of un-scaffolded contigs can increase. To solve these problems we propose a second step in our pipeline that will increase the assembled sequences length by scaffolding the split contigs based on the shared barcodes. Several scaffolding methods using linked read information were already proposed such as ARCS [3] or Scaff10x [4]. While these methods can make a difference on a highly fragmented assembly, the number of contigs connected into scaffolds are very low when tested on contigs constructed from long reads. This is due to higher quality of long read based assemblies where most of long contig extremities and short contigs are complex repetitive sequences that share barcodes with numerous other contigs. We propose a new method that constructs a scaffold graph based on multiple barcode sharing constraints such as a minimum number of shared barcodes, a minimum percentage of common barcodes between two contig extremities and correct molecules lengths if a connection between two contigs would be made. In this way we maximize the number of contigs situated in unbranched paths and we avoid local connecting decisions which are more error-prone.

Acknowledgements

This study is part of the SeqOccIn project (https://get.genotoul.fr/seqoccin/) which is conducted by Get and Bioinfo Platforms of Genotoul.

References

[1] Shaun D Jackman, Lauren Coombe, Justin Chu, Rene L Warren, et al. Tigmint: correcting assembly errors using linked reads from large molecules. *BMC Bioinformatics*, 19(1):1–10, 2018.

[2] Olga Dudchenko, Sanjit S Batra, Arina D Omer, et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, 356(6333):92–95, 2017.

[3] Sarah Yeo, Lauren Coombe, René L Warren, Justin Chu, and Inanç Birol. ARCS: scaffolding genome drafts with linked reads. *Bioinformatics*, 34(5):725–731, 2018.

[4] Zemin Ning and Ed Harry. Scaff10x Software. https://github.com/wtsi-hpag/Scaff10X, 2018. [Online; accessed 18-May-2021].

Leveraging multi-omics integration through co-expression networks to improve the diagnostic power of mitochondrial diseases

Justine LABORY[1,2], Samira AIT-EL-MKADEM SAADI[2] Sylvie BANNWARTH[2], Véronique PAQUIS-FLUCKINGER[2] and Silvia BOTTINI[1]

[1] Université Côte d'Azur, Center of Modeling, Simulation and Interactions, Nice, France
[2] Université Côte d'Azur, Inserm U1081, CNRS UMR 7284, Institute for Research on Cancer and Aging, Nice (IRCAN), Centre hospitalier universitaire (CHU) de Nice, Nice, France

Corresponding Author: silvia.bottini@univ-cotedazur.fr; silvia.bottini@unice.fr

Mitochondrial diseases (MD) are rare disorders caused by deficiency of the mitochondrial respiratory chain, which provides energy in each cell through oxidative phosphorylation. These diseases are extremely heterogeneous both clinically and genetically, with a broad range of age onset and very different symptoms, that makes their diagnosis very challenging. The estimated incidence is 1/5000 births, about 200 new cases per year in France. Although mitochondria possess their own genome, they need nuclear genes because both genomes encode proteins responsible of the mitochondrial biogenesis. Hence, MD are caused by pathogenic variants affecting either mitochondrial DNA or nuclear genes involved in mitochondrial functions.

The advent of whole-exome sequencing (WES) and whole-genome sequencing has accelerated the identification of variants on previously unknown disease genes (1). Although these technologies are mainstays in Mendelian disease diagnosis, their success rate for detecting causal variants is far from complete, ranging from 25 to 50%. Several variants remain as variants of unknown significance (VUS) or they are missed due to the inability to prioritize them such as intronic or non-coding variants. Recently, the employ of RNA sequencing (RNA-seq) has been proposed. This technology provides a direct probing of RNA abundance, including allele-specific expression and splice isoforms. Despite the promising premises of RNA-seq to detect new variants, the pioneering works employing this technique improved the diagnostic power of only 10% (2). The development of novel integrative computational approaches are essential to resolve diagnostic deadlock and improve our knowledge of mitochondrial disorders (3).

Here, we dispose of a cohort of 17 patients with clinical evidences of MD in diagnostic stalemate. WES was not informative since it did not allow to identify the responsible gene. Then, RNA-seq has been performed. We developed a novel approach, called MitoBook, that performs a co-expression network analysis based on the approach Multiscale Embedded Gene Co-expression Network Analysis (MEGENA) to identify functional co-expressed gene modules associated with MD (4). First, we tested this approach on a control cohort from GTEx database. Then we applied MitoBook on our patient cohort. Functional enrichment of biological pathways from several ontologies including GO ontology, metabolic pathways from KEGG and REACTOME, was performed on genes belonging to each module. To identify potential regulatory genes, we used a novel defined measure called patient-specificity that allowed to characterize the specificity of the modules. Finally, we integrated variants found by WES analysis to co-expression modules. This approach allowed to identify 115 variants for 81 potential candidate pathogenic genes for 5 patients. The output of MitoBook is a web interactive application that will summarize findings by patient, by pathway and by gene. In conclusion, MitoBook should increase the diagnostic power up and provides an end-to-end solution for identifying potential pathogenic genes and is suitable for use by mitochondrial diseases diagnostic platforms.

References

1. Plutino M, Chaussenot A, Rouzier C, Ait-El-Mkadem S, Fragaki K, Paquis-Flucklinger V, et al. Targeted next generation sequencing with an extended gene panel does not impact variant detection in mitochondrial diseases. 2018.
2. Kremer L, Bader D, Mertes C, Kopajtich R, Pichler G, Iuso A, et al. Genetic diagnosis of Mendelian disorders via RNA sequencing. 2017.
3. Labory J, Fierville M, Ait-El-Mkadem S, Bannwarth S, Paquis-Flucklinger V, Bottini S. Multi-omics approaches to improve mitochondrial disease diagnosis: challenges, advances and perspectives. Frontiers. 2020.
4. Song W-M, Zhang B. Multiscale Embedded Gene Co-expression Network Analysis. PLoS Comput Biol. 2015.

Structural prediction of protein-protein interactions using evolutionary information

Chloé QUIGNOT[1], Hélène BRET[1], Raphael GUEROIS[1] and Jessica ANDREANI[1]

[1] Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France

Corresponding Author: jessica.andreani@cea.fr

1. Introduction

Most biological processes rely on protein-protein interactions. Protein-protein docking aims to predict the most likely structural binding modes of interacting partners. Using information about the coevolution of protein partners improves this structural prediction of interfaces. Our team has contributed to the improvement of protein docking success rates by incorporating evolutionary information into docking strategies. Here, we present our most recent achievements in this respect.

2. Results and perspectives

In previous work, we developed a protein-protein docking pipeline that integrates evolutionary information in the docking process [1]. Recently, we designed a novel strategy to integrate evolutionary information into atomic-level scoring functions and found that it greatly improved their capacity to discriminate correct from incorrect interface models [2]. This strategy uses relatively shallow coupled multiple sequence alignments (coMSAs) containing 10 to 40 complex homologs. We benchmarked this strategy on a large dataset of docking targets based on unbound homology models [3]. We also integrated it into the *InterEvDock3* docking server [4], along with two other major improvements: the capacity to use information from covariation-based contact maps during the docking process and the ability to combine free docking with template-based assembly modeling. The server is available at https://bioserv.rpbs.univ-paris-diderot.fr/services/InterEvDock3/

We successfully applied the InterEvDock pipeline to recent targets of the international CAPRI assembly prediction challenge [5]. Future perspectives include a better combination of our atomic-level modeling of complex homologs with covariation-based approaches and the use of machine learning techniques to further enhance the extraction of (co)evolutionary signal from coMSAs.

Acknowledgements

This work was supported by the French National Research Agency under grants ANR-15-CE11-0008 to R.G. and ANR-18-CE45-0005 to J.A., by CEA doctoral funding to H.B. and by IDEX Paris-Saclay doctoral funding to C.Q. It was granted access to the HPC resources of CCRT under allocations 2018-7078 and 2019-7078 by GENCI (Grand Equipement National de Calcul Intensif).

References

1. Chloé Quignot, Julien Rey, Jinchao Yu, Pierre Tufféry, Raphael Guerois and Jessica Andreani. InterEvDock2: an expanded server for protein docking using evolutionary and biological information from homology models and multimeric inputs. *Nucleic Acids Res*, 46(W1):W408-W416, 2018.
2. Chloé Quignot, Pierre Granger, Pablo Chacón, Raphael Guerois and Jessica Andreani. Atomic-level evolutionary information improves protein-protein interface scoring. *Bioinformatics*, btab254 (in press), 2021.
3. Jinchao Yu and Raphael Guerois. PPI4DOCK: large scale assessment of the use of homology models in free docking over more than 1000 realistic targets. *Bioinformatics*, 32(24):3760-3767, 2016.
4. Chloé Quignot, Guillaume Postic, Hélène Bret, Julien Rey, Pierre Granger, Samuel Murail, Pablo Chacón, Jessica Andreani, Pierre Tufféry and Raphaël Guerois. InterEvDock3: a combined template-based and free docking server with increased performance through explicit modeling of complex homologs and integration of covariation-based contact maps. *Nucleic Acids Research*, gkab358 (in press), 202.
5. Aravindan Arun Nadaradjane, Chloé Quignot, Seydou Traoré, Jessica Andreani and Raphael Guerois. Docking proteins and peptides under evolutionary constraints in Critical Assessment of PRediction of Interactions rounds 38 to 45. *Proteins*, 88(8):986-998, 2020.

Statistical method to detect PPR binding sites at Arabidopsis Thaliana

Mathilde SAUTREUIL[1,2,3], Etienne DELANNOY[1,2] and Guillem RIGAIL[1,2,3]

[1] Université Paris-Saclay, CNRS, INRAE, Univ Evry, Institute of Plant Sciences Paris-Saclay (IPS2), Orsay, 91405, France

[2] Université de Paris, CNRS, INRAE, Institute of Plant Sciences Paris-Saclay (IPS2), Orsay, 91405, France

[3] Université Paris-Saclay, CNRS, Univ Evry, Laboratoire de Mathématiques et Modélisation d'Evry, Evry, 91037, France

Corresponding author: mathilde.sautreuil@u-psud.fr

Pentatricopeptide repeat (PPR) [1] are a large family of RNA-binding proteins which regulate several aspects of gene expression, primarily in the mitochondria and chloroplast but also in the nucleus. In particular, PPRs are involved in the maturation, splicing, editing, and translation of RNAs. About 500 PPRs exist in Arabidopsis Thaliana, but their binding to RNA stays partially known. The objective of our work was to predict PPR binding sites from RNA-seq data of small RNAs.

Classically, PPR binding sites are detected using differential RNA-footprinting: that is by comparing the number of small RNAs of a wild and mutant plant for one given PPR along the chloroplastic or mitochondrial genome. The disadvantage of this approach is the need to have a mutant PPR plant, restricting the use of this approach. Here we propose to predict PPR binding sites using only the small RNAs of wild plants.

We trained a statistical learning method (a Generalized Linear Model) to predict known PPR binding sites from small RNA datasets. We tested our approach on two public small RNA datasets of the chloroplastic and mitochondrial genome of Arabidopsis Thaliana. We trained our model on a first dataset from [2] containing many binding sites to predict. Using the cross-validation technique, we recovered a specificity of 0.999 and sensitivity of 0.756, showing that it is indeed possible to predict these binding sites. Furthermore, we visually validated some newly predicted sites. We used the second dataset [3] as a validation set. It contains very few known binding sites, but we were able to recover two of the known sites. Finally, we realized a differential analysis on the new detected PPR binding sites to identify those specifics at each condition.

References

[1] Sam Manna. An overview of pentatricopeptide repeat proteins and their applications. *Biochimie*, 113:93–99, June 2015.

[2] Hannes Ruwe, Gongwei Wang, Sandra Gusewski, and Christian Schmitz-Linneweber. Systematic analysis of plant mitochondrial and chloroplast small RNAs suggests organelle-specific mRNA stabilization mechanisms. *Nucleic Acids Research*, page gkw466, May 2016.

[3] Wenjuan Wu, Sheng Liu, Hannes Ruwe, Delin Zhang, Joanna Melonek, Yajuan Zhu, Xupeng Hu, Sandra Gusewski, Ping Yin, Ian D. Small, Katharine A. Howell, and Jirong Huang. SOT1, a pentatricopeptide repeat protein with a small MutS-related domain, is required for correct processing of plastid 23S-4.5S rRNA precursors in *Arabidopsis thaliana*. *The Plant Journal*, 85(5):607–621, March 2016.

Accelerating the diagnosis: Automated prioritization of CNVs with ACMG/ClinGen standards

Jiri Ruzicka[1,2], Sacha Beaumeunier[1], Nicolas Philippe[1] and Denis Bertrand[1]
[1] SeqOne Genomics S.A.S., 22 rue Durand, 34000 Montpellier, France
[2] Bioinformatics and Modeling, Biosciences Department, INSA Lyon, 20 avenue Albert Einstein, 691000 Villeurbanne, France

Corresponding author: jiri.ruzicka@seqone.com

Abstract

CNVs (Copy-number variations) are structural variants of the genome, consisting in deleting or repeating segments of DNA. These changes can have important consequences leading to cancer or rare diseases. With the great expansion of sequencing technologies, the automated prioritization of CNVs simplifies the clinical evaluation and rapidly gets the very needed information to the patient.

In 2020, ACMG and ClinGen[1] published new guidelines for CNV classification. These highly detailed and structured standards allow better and faster prioritization of CNVs with increased consistency between different laboratories.

Here, we present a novel implementation of the ACMG/ClinGen framework using the latest bioinformatics references. The highly structured pipeline uses the data from the latest available resources, including the Dosage Sensitivity Map from ClinGen[2] or the databases of structural variants from gnomAD[3] and DGV[4]. The classification is adapted to both small and large CNVs with the implementation of rare cases such as intragenic CNV in a haploinsufficient gene. Additional information is available to the user, like the frequency in common population of overlapping structural variants from gnomAD and DGV.

The predicted classification was evaluated with an original ACMG/ClinGen dataset consisting of 114 CNVs. The results were compared with the classification of two independent evaluators. The new classification showed a very high specificity in the detection of both pathogenic and benign variants. In 84.2% of CNVs, the prediction was the same as the prediction of at least one evaluator. For the 15.8% of predictions in disagreement, no variants classified as benign were predicted pathogenic and vice-versa, highlighting the good performance of the automated annotation method.

The new SeqOne implementation of ACMG/ClinGen standards provides a confident and fast classification of CNVs which could simplify the clinical interpretation of structural variants. The next step in getting an improved classification is the linkage between phenotypes and genes included in the CNV.

References

[1] Andersen E.F. Cherry A.M. et al. Riggs, E.R. Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the american college of medical genetics and genomics (acmg) and the clinical genome resource (clingen). *Genet Med*, (22):245–257, 2020.

[2] Brooks LD Bustamante CD Evans JP Landrum MJ Ledbetter DH Maglott DR Martin CL Nussbaum RL Plon SE Ramos EM Sherry ST Watson MS Rehm HL, Berg JS. Clingen. clingen–the clinical genome resource. *N Engl J Med*, (Jun 4;372(23)):2235–42, 2015.

[3] Brand H. Karczewski K.J. et al. Collins, R.L. A structural variation reference for medical and population genetics. *Nature*, (581):444–451, 2020.

[4] Yuen RK Feuk L Scherer SW MacDonald JR, Ziman R. The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.*, (Oct 29):42, 2013.

Multimodal analysis of small size data to understand the effects of low doses of ionizing radiation on atherosclerosis

ABDELOUAHAB, REY, EBRAHIMIAN, GLOAGUEN, KERESELIDZE, DEMARQUAY, KLOKOV, GARALI1 and CHIUSA EBRAHIMIAN
Institut de radioprotection et de sûreté nucléaire, laboratoire de radiotoxicologie et de radiobiologie expérimentale, 92260, Fontenay-aux Roses-France
macine-bachir.abdelouahab@irsn.fr

Background: Atherosclerosis is a hardening and narrowing of arteries. It can put blood flow at risk as arteries become blocked. Mechanistic understanding of the effects of low-dose ionizing radiation (LDIR) on atherosclerosis remains incomplete. The experimental studies have shown a protective effect of LDIR on atherosclerosis in rodent models [1][2]. How early responses in different cell types that are known to be involved in atherosclerosis contribute to this effect is not clear. Challenges in understanding biological mechanisms of LDIR include small size of experimental animal groups, low level changes and data heterogeneity and multimodality. Advanced methods of multimodal data integration, including machine learning algorithms, have proved useful in various biomedical applications but have not been used much in the studies of health effects of LDIR. In this study, we report results of applying a statistical and classical machine learning methods to reveal causal mechanistic links between biological responses of different modalities observed in ApoE-/- mice exposed to low to moderate doses of gamma-radiation. Revealing complex correlations and causal links related to health conditions, such as atherosclerosis, can help advance the concept adverse outcome pathway (AOP).

Analysis: Fourteen to 16 weeks old ApoE-/- mice were exposed to a single dose of 0, 0.05, 0.5, or 1 Gy from 137Cs at a dose rate of 10.35 mGy.min-1. Samples for biological assay were collected at 24 hours, 10 days and 100 days post-irradiation. Bone marrow cells from mice sacrificed at 24 hours or 10 days post-exposure were used to grow bone marrow derived macrophages for polarization into M1 or M2 phenotypes. Cells were treated with Il-4 or INFγ for M2 or M1 polarization, respectively. Polarization was validated by measuring the mRNA levels of polarization markers using RT-qPCR. Also, cell culture supernatants of polarized macrophages were harvested and cytokine expression were determined using ELISA. In this study, we focused on a single time point of 24 after irradiation with three modalities of data. Data from flow cytometry, gene expression and ELISA assay originating from the same animal were gathered. To carry out an integral analysis this complex and heterogeneous data obtained at different levels of biological organization, we employed a two-step analysis. In the first step, we used univariate and descriptive statistics analysis (ANOVA, PCA, PLS, RGCCA) and compared it with four classical machine learning methods (SVM, random forest, tree decision and Gradient boosting). In the second step, an integration of the obtained correlations into a biologically relevant mechanistic model that can be used as a basis for an AOP to atherosclerosis after exposure to LDIR was carried out.

Results:

Our preliminary results confirm the difficulty of visualizing the effects of LDIR. The results of univariate ANOVA analysis, although could test important comparisons, were limited in its ability to reveal a comprehensive integral insight when multiple variables were used. Using PCA allowed to reveal certain hidden trends in the assembly of variables, which was an improvement when compared to ANOVA. PLS and RGCCA were both more efficient in identifying those variables that are responsive to different treatment options and mediate biological phenotypic changes. Despite the low classification rate, machine learning methods allowed to identify those features/variables that contribute to differences between groups. These different analyses helped to reconstruct in more detail the mechanistic and/or causal links between various molecular endpoints measured under various treatment (radiation dose) and phenotypic (M0, M1 or M2 macrophages) conditions. Overall, our results highlight the need for comprehensive data analyses of LDIR studies and provide an example of such analyses using a multitude of statistical methods to understand mechanisms of LDIR effects on atherosclerosis.Pages must NOT be numbered. Final pagination will be set by the editors of the proceedings.

The list of references is headed *References*, it should be placed at the end of your contribution. It should be in *Times New Roman* 10-point font. Please do not insert a page break before the list of references. For citations in the text, please use square brackets [1] and consecutive ordered numbers [2,3] in list of references. Please

find below examples on how to format references corresponding to articles [1], books [2], book chapters and proceedings [3].

1. Ebrahimian TG et al. “Chronic Exposure to External Low-Dose Gamma Radiation Induces an Increase in Anti-inflammatory and Anti-oxidative Parameters Resulting in Atherosclerotic Plaque Size Reduction in ApoE-/- Mice.” Radiat Res. 2018
2. Rey et al. “Exposure to low to moderate doses of ionizing radiation induces a reduction of pro-inflammatory Ly6Chigh monocytes and a U-curved response of T cells in ApoE -/- mice.” Dose-Response.2021

PADLE - Profiling Analysis of DifferentiaL Expression: a tool and environment designed to analyse complex RNA-seq gene expression data

Arthur PÉRÉ[1], Frédérique HILLIOU[1], Etienne G.J. DANCHIN[1], Martine DA ROCHA[1] and Corinne RANCUREL[1]

[1] INRAE, Université Côte d'Azur, CNRS, 400 route des Chappes BP 167 06903 Sophia Antipolis Cedex, France

Corresponding Author: arthur.pere@inrae.fr

Abstract

PADLE analyses RNA-seq data in a user-friendly way for biologists with a graphical web interface. The tool allows for differential gene expression analyses, which can be combined with functional analyses and extracting expression patterns.

Although similar tools already exist, they do not provide all the features implemented in PADLE. For example, Shiny-Seq [1] does RNA-seq analysis, but doesn't offer several R packages to study differential expression. Moreover, it does not allow complex analysis (with several factors), and functional enrichment analyses (i.e. GO and KEGG) are limited to only two species.

For RNA-seq analyses, three different R packages are implemented: DESeq2 [2], EBSeq [3] and edgeR [4]. Thus, the user will select the R package most adapted to his experimental design (complex experimental designs with several factors, consideration of isoforms).

PADLE also proposes GO (Gene Ontology) [5] and/or KEGG pathway (Kyoto Encyclopedia of Genes and Genomes) [6] enrichment on differentially expressed genes, if a functional annotation table is provided. In addition to the stats for KEGG enrichment, the metabolic pathways are compiled and the enzymes that are activated or deactivated are highlighted in different colours on the metabolic pathways.

This tool makes it possible to group together genes with the same expression profile but also to highlight genes with an expression profile that is very different from the rest of the genes. For this purpose, the following unsupervised machine learning methods are used: ABOD, Isolation Forest and DBSCAN.

Finally, PADLE allows visualising experiments, taking all the biological comparisons as dimensions and reduce those to 2 dimensions using SOM, PCA and tSNE methods. The graph generated permits to see the clustering made and compare all the biological conditions at once.

PADLE is a web tool that will be released for public use soon. It will be possible to access it through this web page https://padle.sophia.inrae.fr/intro/. It allows analysis of complex RNA-seq data under a user-friendly interface for biologists.

References

[1] Z. Sundararajan, R. Knoll, P. Hombach, M. Becker, J. L. Schultze, et T. Ulas, « Shiny-Seq: advanced guided transcriptome analysis », *BMC Res. Notes*, vol. 12, nº 1, p. 432, juill. 2019, doi: 10.1186/s13104-019-4471-1.

[2] M. I. Love, W. Huber, et S. Anders, « Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 », *Genome Biol.*, vol. 15, nº 12, p. 550, déc. 2014, doi: 10.1186/s13059-014-0550-8.

[3] N. Leng *et al.*, « EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments », *Bioinformatics*, vol. 29, nº 8, p. 1035-1043, avr. 2013, doi: 10.1093/bioinformatics/btt087.

[4] M. D. Robinson, D. J. McCarthy, et G. K. Smyth, « edgeR: a Bioconductor package for differential expression analysis of digital gene expression data », *Bioinformatics*, vol. 26, nº 1, p. 139-140, janv. 2010, doi: 10.1093/bioinformatics/btp616.

[5] The Gene Ontology Consortium *et al.*, « The Gene Ontology resource: enriching a GOld mine », *Nucleic Acids Res.*, vol. 49, nº D1, p. D325-D334, janv. 2021, doi: 10.1093/nar/gkaa1113.

[6] M. Kanehisa, M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, et M. Tanabe, « KEGG: integrating viruses and cellular organisms », *Nucleic Acids Res.*, vol. 49, nº D1, p. D545-D551, janv. 2021, doi: 10.1093/nar/gkaa970.

Use of deep learning approaches for integrating multiple data types in the prediction of response to immunotherapy in melanoma

Haythem SRIHI[1,2], Matthieu GENAIS[1], Anne MONTFORT[1], Nicolas MEYER[1,3], Bruno SEGUI[1] and Vera PANCALDI[1]

1 CRCT,2 Avenue Hubert Curien, 31100 Toulouse, FRANCE
2 Faculty of Sciences of Montpellier,E. Bataillon Place, 34095, Montpellier, France

3 Dermatology department, Institut Universitaire du Cancer, 31100 Toulouse, France

Corresponding author: haythem.srihi@etu.umontpellier.fr

1 Introduction

Melanoma is the rarest but also the most serious form of skin cancer. The latest figures published by Public Health in 2019 show 15,513 new cases of melanoma of the skin identified in France. If it can be diagnosed in its early stages, with the choice of the appropriate treatment, the survival rate is generally good. Recent research studies show that in France the 5-year relative survival rate of a person living with melanoma is 87% (83% for men, 89% for women) when melanoma is diagnosed at an early stage. However, the 5-year overall survival rate is 52% for patients affected with advanced melanoma.

During the development of a tumor, cancer cells are surrounded by many other cells, including healthy cells and also immune cells, such as CD+ and CD8+ T lymphocytes, Natural Killer (NK), as well as myeloid cells. Some lymphocytes potentiate the defence against cancer cells such as CD8+, T lymphocytes and NK which have the potential of killing cancer cells by their cytotoxic action. Others, on the contrary, can act as immunosuppressive cells, such as regulatory T cells and myeloid-derived suppressor cells which compromise CD8 T cell-dependent immune responses, leading to tumor and metastases.

2 Methods

Here, we aim to predict patients' response to immunotherapy based on integration of multiple datasets extracted from their peripheral blood including phenotyping of circulating immune cells (by flow cytometry approaches that quantify cell-surface markers on single cells) and cytokine quantification in plasmas (Mesoscale). These datasets were used to create a feature matrix and various machine learning approaches were applied to predict patients' responses within the context of the TICIMEL clinical trials [1]. Finally, radiomics images available for each patient will be integrated through a Convolutional Neural Network to test whether therapy response can be predicted using this type of data at diagnosis and 6 weeks after immunotherapy induction.

3 Conclusion

Integrative machine learning approaches applied to the comprehensive datasets generated as part of this clinical trial will provide a better understanding of advanced melanoma resistance to immunotherapy.

4 References

[1] Anne Montfort, Thomas Filleron, Mathieu Virazels, Carine Dufau, Jean Milh_es, C_ecile Pag_es, Pascale Olivier, Maha Ayyoub, Muriel Mounier, Amélie Lusque, Stéphanie Brayer, Jean-Pierre Delord, Nathalie Andrieu-Abadie, Thierry Levade, Céline Colacios, Bruno Ségui, and Nicolas Meyer. Combining nivolumab and ipilimumab with iniximab or certolizumab in patients with advanced melanoma: First results of a phase ib clinical trial. 27(4):1037-1047.

Rattus norvegicus reference genome selection for RNA-seq data analysis

Christophe Le Priol[1], Anne-Elodie Receveur[1] and Andrée Delahaye-Duriez[1,2]
[1] Inserm UMR 1141 NeuroDiderot, Université de Paris, Paris, France
[2] UFR Santé Médecine Biologie Humaine, Université Sorbonne Paris Nord, Bobigny, France

Corresponding author: christophe.le-priol@inserm.fr

The sequencing of the transcriptome has enabled to assess genome-wide changes in gene expression in a variety of biological contexts. One of the first steps in a usual RNA-seq data analysis workflow is to estimate gene expression by counting the reads mapped to annotated genomic regions. The reference genome, which consist of both sequences and annotations, is mandatory to perform this step and may have a huge impact on the subsequent analyses, *e.g.* identification of differentially expressed genes. While tools have been developped to compare genome annotations for many years [1], the effect of the choice of a reference genome on RNA-seq data analysis remains rarely discussed. Yet, it has been demonstrated that the expression quantification and the differential expression assessment are indeed dramatically affected by the choice of reference genome for large sets of genes [2]. Besides, this effect is tissue-dependent [2]. As recommendations for classical RNA-seq data analysis, it is preferred to use a less complex annotation, *e.g.* NCBI RefSeq, for reproducibility and a more complex annotation, *e.g.* Ensembl, for exploratory research [3].

The Normay rat, *Rattus norvegicus* species, is a widely used experimental model in medical and biological research. The few studies quantifying the impact of the choice of reference genome on RNA-seq data analysis mostly deal with the human genome. No particular attention has already been paid to the *Rattus norvegicus* species on this topic. Here, we propose to evaluate the effect of common genome annotations, *i.e.* Ensembl and NCBI RefSeq, of the most widely used *Rattus norvegicus* genome sequence, Rnor_6.0 (RefSeq accession: GCF_000001895.5), on a classical differential expression workflow based on RNA-seq data. On 2020/11/10, a new genome assembly, mRatBN7.2 (RefSeq accession: GCF_015227675.2), was published, followed by NCBI RefSeq annotations a few months later. Taking benefit from significant improvements in sequencing since the release of the previous reference genome more than six years before, this new assembly exhibits much better assembly quality metrics. This lets hope significant refinements in RNA-seq data analysis. Since this new assembly and this new annotation were published very recently, no study has already used them. We propose to integrate them in the evaluation of *Rattus norvegicus* reference genomes and characterize the improvements brought by this new genome. We re-analyzed published studies using RNA-seq data from different hippocampal regions [4] and highlighted the impact of the choice of reference genome in the context of epileptogenesis. We emit recommendations in this particular context by focusing on some key genes using RT-qPCR data [5] as a gold standard for validation.

References

[1] Joel E Richardson. fjoin: simple and efficient computation of feature overlaps. *Journal of Computational Biology*, 13(8):1457–1464, 2006.

[2] Shanrong Zhao and Baohong Zhang. A comprehensive evaluation of ensembl, refseq, and ucsc annotations in the context of rna-seq read mapping and gene quantification. *BMC Genomics*, 16:97, 2015.

[3] Po-Yen Wu, John H Phan, and May D Wang. Assessing the impact of human genome annotation choice on rna-seq expression estimates. *BMC Bioinformatics*, 14 Suppl 11:S8, 2013.

[4] Lara Ianov, Matt De Both, Monica K Chawla, Asha Rani, Andrew J Kennedy, Ignazio Piras, Jeremy J Day, Ashley Siniard, Ashok Kumar, J David Sweatt, Carol A Barnes, Matthew J Huentelman, and Thomas C Foster. Hippocampal transcriptomic profiles: Subfield vulnerability to age and cognitive impairment. *Frontiers in Aging Neuroscience*, 9:383, 2017.

[5] Olga E Zubareva, Anna A Kovalenko, Sergey V Kalemenev, Alexander P Schwarz, Vladimir B Karyakin, and Aleksey V Zaitsev. Alterations in mrna expression of glutamate receptor subunits and excitatory amino acid transporters following pilocarpine-induced seizures in rats. *Neuroscience Letters*, 686:94–100, 2018.

Automated identification of a cancer patient treatment: from sequencing to treatment prioritization

Nicolas Soirat[1,2], Denis Bertrand[1], Sacha Beaumeunier[1], Raphaël Lanos[1], Nicolas Philippe[1], Dominique Vaur[2], Laurent Castera[2], Sophie Krieger[2] and Anne-Laure Bougé[1]

[1] SeqOne, 22 Rue Durand, 34000, Montpellier, France

[2] Laboratoire de Biologie et de Génétique du Cancer (Inserm U1245), 3 avenue général Harris, 14000, Caen, France

Corresponding author: nicolas.soirat@seqone.com

Abstract

The emergence of High Throughput Sequencing (HTS) allowed the scientific community to gather a tremendous amount of cancer genomic data. Those studies highlighted the genomic diversity of tumors and identified biomarkers responsible for tumorigenesis that might indicate potential treatments for which a tumor is sensitive. Those markers might predispose a person to cancer (germline mutation) or appear during cancer development (somatic mutation). So for a cancer patient, it is capital to be able to efficiently identify his tumor genomic landscape, the potential targetable biomarkers and his personalized treatment.

Biomarkers are encompassing different genomic events such that Single Nucleotide Variant (SNV), Copy Number Variation (CNV), Gene Fusion, abnormal mRNA isoforms and mutational signatures (Tumor Mutation Burden, Microsatellite Instability, Homologous Recombination Deficiency...). The use of short-read sequencing to identify cancer patient biomarkers and treatment is becoming a more common practice in hospitals and requires the development of automated analysis to help clinicians to efficiently identify the patient treatment.

For our study, we used a novel DNA/RNA panel of more than 500 genes, that was designed to detect actionable mutations in the tumor and can be used to identify most of the known tumor biomarkers. We develop a complete pipeline able to identify precisely all those different genomic events and apply the required filtering step, to remove FFPE biases, sequencing artefact and benign germline variants. A selection process allows us to only focus our tertiary analysis on mutations that are very likely to be associated with cancer[1]. Finally, we design a method able to identify and prioritize treatments and clinical trials. First, we identify all variant-treatment associations from a clinical drug/trial database. Second, as many patients do not harbor the known mutation, we developed a decisional tree to identify the most promising treatment by prioritizing all associations returned by the database. Our tests, on a selected set of patients, show very promising results.

References

[1] Zbyslaw Sondka, Sally Bamford, Charlotte G. Cole, Sari A. Ward, Ian Dunham, and Simon A. Forbes. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers, 2018.

MAGNETO: a complete workflow for the recovery of genomes from metagenomes using complementary strategies

Benjamin Churcheward[1], Maxime Millet[1], Audrey Bihouée[2], Guillaume Fertin[1] and Samuel Chaffron[1]

[1] Laboratoire des Sciences du Numériques de Nantes, UMR CNRS 6004, Université de Nantes, 2 Chemin de la Houssinière, BP 92208, 44322 Nantes Cedex, France

[2] IRS UN, Institut du Thorax - UMR INSERM 1087, 8 quai Moncousu, BP 70721, 44007 Nantes Cedex 1, France

Corresponding author: benjamin.churcheward@univ-nantes.fr, samuel.chaffron@univ-nantes.fr

Genomes are valuable resources for characterizing microbial organisms taxonomy, their functional and metabolic potential, as well as for evaluating and understanding their ecology and evolution. However, the reconstruction of bacterial genomes has been historically limited by the requirement to sequence isolates from pure cultures. Due to the fact that most bacterial species are difficult to cultivate it precluded genomic insights for the vast majority of microbial life. Thanks to the development of shotgun sequencing of microbial communities (metagenomics), a direct access to the diversity of uncultivated microorganisms is now possible *in situ*, bypassing the cultivation bottleneck.

During the last decade, various computational methods have been developed to reconstruct individual genomes from metagenomes, referred to as Metagenome-Assembled Genomes (MAGs), which have significantly expanded our knowledge about microbial genomic diversity by delineating thousands of previously unknown genomes inhabiting very diverse environments, such as the human gut, soil, and oceans. The reconstruction of MAGs from metagenomes is usually performed through two main steps: (i) the assembly of reads into contigs, followed by (ii) the binning of contigs into (draft) genome bins. The metagenomic assembly process can be performed using reads from a single sample, or from multiple samples (i.e., co-assembly). Recent studies have reconstructed MAGs using either single-assembly [7] or co-assembly [3] strategies, which both have their own benefits. While single-sample assemblies are well suited for large numbers of relatively low-diversity samples, co-assemblies may be beneficial for a moderate number of relatively high-diversity samples. While a recent work reported a wide range of advantages and limitations in current MAGs reconstruction methods [2], it still remains unclear which assembly strategies, in combination with which genome binning approach, offer the best performance for MAGs recovery. Besides, although a few metagenomic workflows have been recently developed (e.g., [4], [6]), they do not implement an automated co-assembly process, in combination with distinct genome binning strategies, for maximizing MAGs recovery.

In order to assess the impact of different reconstruction strategies on MAGs diversity and quality, we defined and compared four reconstruction strategies, by combining (i) single-sample assembly or co-assembly, with (ii) single-sample or multi-samples genome binning. Reads were assembled using Megahit [5]. We evaluated MAGs quality based on the presence/absence of single-copy core genes (SCG), and performed the dereplication of genomes independently recovered by each strategy for evaluation and comparison purposes. In addition, we also designed and implemented an automated co-assembly approach based on metagenomic distance, using Simka [1], to identify optimal sets of metagenomes to co-assemble. We implemented these complementary strategies in a Snakemake workflow called *MAGNETO* (shortly publicly available). Our tool also implements the reconstruction and annotation of metagenomic genes catalog, as well as the taxonomic and functional annotations of all recovered MAGs.

References

[1] Gaëtan Benoit, Pierre Peterlongo, Mahendra Mariadassou, Erwan Drezen, Sophie Schbath, Dominique Lavenier, and Claire Lemaitre. Multiple comparative metagenomics using multiset k -mer counting. *PeerJ Computer Science*, 2:e94, November 2016. doi:10.7717/peerj-cs.94.

[2] Lin-Xing Chen, Karthik Anantharaman, Alon Shaiber, A Murat Eren, and Jillian F Banfield. Accurate and complete genomes from metagenomes. *Genome research*, 30(3):315–333, 2020.

[3] Tom O. Delmont, Christopher Quince, Alon Shaiber, Özcan C. Esen, Sonny TM Lee, Michael S. Rappé, Sandra L. McLellan, Sebastian Lücker, and A. Murat Eren. Nitrogen-fixing populations of Planctomycetes

and Proteobacteria are abundant in surface ocean metagenomes. *Nature Microbiology*, 3(7):804–813, July 2018. doi:10.1038/s41564-018-0176-9.

[4] Silas Kieser, Joseph Brown, Evgeny M Zdobnov, Mirko Trajkovski, and Lee Ann McCue. Atlas: a snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data. *BMC bioinformatics*, 21(1):1–8, 2020. doi:10.1186/s12859-020-03585-4.

[5] Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*, 31(10):1674–1676, 2015.

[6] Sabrina Krakau; Daniel Straub; Hadrien Gourlé; Maxime Garcia; Gisela Gabernet; nf-core bot; Maxime Borry; Alexander Peltzer; Phil Ewels; Johannes Alneberg; Michael L Heuer. nf-core/mag. 2020. URL: https://zenodo.org/record/4529420#.YKWAgeuxV7g, doi:10.5281/zenodo.4529420.

[7] Edoardo Pasolli, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, Paolo Manghi, Adrian Tett, Paolo Ghensi, et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*, 176(3):649–662, 2019. doi:10.1016/j.cell.2019.01.001.

Automatic differential analysis of transcription variants in the chloroplast with changepoint detection

Arnaud Liehrmann[1,2], Benoît Castandet[2] and Guillem Rigaill[1,2]
[1] Laboratoire de Mathématiques et Modélisation d'Évry (LAMME), CNRS, 91037, Évry, France
[2] Institut des Sciences des Plantes de Paris-Saclay (IPS2), CNRS, INRAE, 91405, Orsay, France

Corresponding author: arnaud.liehrmann@universite-paris-saclay.fr

1 Background

Chloroplast gene expression is essential for photosynthesis and plant viability. Much of the expression is controlled post-transcriptionally. Gaining a global picture of chloroplast gene regulation requires detailed knowledge of the transcriptome along with the enzymatic and RNA-binding activities that shape it [1]. Several RNA-Seq based strategies have recently been developed to decipher its complexity. Most of the tools developed, however, only count the abundance of sequencing reads along annotated patterns (typically genes) and therefore neglect non-coding regions and regulatory events within genes that are pervasive in the chloroplast transcriptome. In the context of differential expression analysis, these events result in local changes in the log-ratio of coverage along the genome between compared conditions.

2 DiffSegR

Our method, DiffSegR, allows systematic identification of differential maturation events without relying on pre-existing annotations in a two-step design: (1) *Summary of the transcriptionnal landscape.* We assume that observed log-ratio of coverage, indexed on genomic positions, is a piecewise constant signal affected by K local changes, also called changepoints, and an added i.i.d noise. We can infer these changepoints using an efficient changepoint detection algorithm that optimizes a penalized likelihood criteria [2]. These changepoints define the limits of $K + 1$ segments / bins within which overlapping reads are then summarized. The first step ends by building a count matrix. (2) *Differential expression analysis.* Each segment, through its associated row in the count matrix, is statistically assessed for quantitative changes in expression levels between compared conditions using the negative binomial model of edgeR or DESeq2.

3 Empirical results

DiffSegR has been applied to two RNA-Seq datasets that were previously used in combination with traditional molecular biology techniques to decipher the role of the chloroplast ribonucleases PNPase [3] and MiniIII [4]. On these two sets of maturation events, DiffSegR returns results close to the expert annotation of the chloroplast RNA-Seq signals performed by biologists, while reducing the analysis time from several hours to a few minutes. We could rightfully predict the role of MiniIII in rRNA maturation and identify the pervasive role of PNPase in 3'-degradation of rRNA, mRNA and tRNA precursor. We believe DiffSegR will not only benefit the biologists working on organellar gene expression but the whole community working on transcriptomics as it allows access to information from a portion of the transcriptome that is not addressed by the classical differential expression analysis pipelines widely used today.

References

[1] David B. Stern, Michel Goldschmidt-Clermont, and Maureen R. Hanson. Chloroplast rna metabolism. *Annual Review of Plant Biology*, 61(1):125–155, 2010.

[2] Robert Maidstone, Toby Hocking, Guillem Rigaill, and Paul Fearnhead. On optimal multiple changepoint algorithms for large data. *Statistics and Computing*, 27:1573–1375, 2017.

[3] Benoît Castandet, Amber M. Hotto, Zhangjun Fei, and David B. Stern. Strand-specific rna sequencing uncovers chloroplast ribonuclease functions. *FEBS Letters*, 587(18):3096–3101, 2013.

[4] Amber M. Hotto, Benoît Castandet, Laetitia Gilet, Andrea Higdon, Ciarán Condon, and David B. Stern. Arabidopsis Chloroplast Mini-Ribonuclease III Participates in rRNA Maturation and Intron Recycling. *The Plant Cell*, 27(3):724–740, 2015.

Precise detection of antibiotic resistance genes in chicken microbiota

Anne-Carmen SANCHEZ[1], Guillaume KON KAM KING[1], Sylvie BAUCHERON[2], Fanny CALENGE[3], Hélène CHIAPELLO[1], Pierre NICOLAS[1], Sébastien LECLERCQ[2] and Anne-Laure ABRAHAM[1]

[1] Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France

[2] Unité Infectiologie et Santé publique, INRAE Val de Loire, 37380 Nouzilly, France

[3] Université Paris-Saclay, INRAE, AgroParisTech, GABI, Jouy-en-Josas, France

Corresponding Author: anne-carmen.sanchez@inrae.fr

Antibiotic resistance genes (ARGs) presence is largely documented in gut microbiomes of farm animals and represent a major risk for animal and human health but the routes and mechanisms of their dissemination are yet only partially understood. At the epidemiological level, it would be interesting to better understand the respective roles of contamination from the environment and transmission between generations of animals. These transmission events, at the genetic level, may be mediated by horizontal transfer of ARGs and clonal propagation of ARG harbouring bacterial lineages. To study these questions, an experiment was conducted aiming at following ARGs by shotgun metagenomics across 3 generations of broiler chicken in two seperate buildings.

The study of ARG in metagenomic samples is challenging, because of sequencing errors, short read length, and variable abundance of genes, which suggest that specific methods may be needed to properly identify ARG flows. Moreover, preliminary results have shown that several haplotypes can co-exist for each gene, and that these haplotypes may be shared or not between samples. Analyzing this micro-evolutionary diversity is the key idea of our work. The identification of haplotypes in metagenomic data is a challenge that has been recently addressed with methods to identify strains relying on SNP frequencies in selected core genome marker genes (such as ConStrains[1], or DESMAN[2]), or with methods using reads as a phasing information to resolve haplotypes on individual genes (Hansel & Gretel [3]).

In order to accurately detect ARG flows between samples, we are working on a new workflow based on aligning metagenomic reads on a non redundant database of reference ARG sequences (Resfinder4.1 [4] clustered at 95% of identity) with BWA-MEM, and stringent quality filters are applied on reads and alignments using samtools. Nucleotide frequencies at each position of these reference sequences are then computed and statistics (number of polymorphic positions, number of alleles…) are recorded in order to quantify the diversity in ARGs within samples and the distance between samples. Our first results indicate a huge variability between genes in terms of number of polymorphic positions and number of haplotypes. Future steps will consist in identifying shared ARG haplotypes between samples by using approaches derived from those used for strain resolution to resolve the different haplotypes and determine their abundances in each sample, or by trying to infer ARG fluxes by comparing samples based on diversity and distance statistics only, without haplotype resolution.

References

[1] Chengwei Luo, Rob Knight, Heli Siljander, Mikael Knip, Ramnik J Xavier, Dirk Gevers, ConStrains identifies microbial strains in metagenomic datasets. *Nat Biotechnology*, 33:1045–52, 2015.

[2] Christopher Quince, Tom O. Delmont, Sébastien Raguideau, Johannes Alneberg, Aaron E. Darling, Gavin Collins & A. Murat Eren, DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biology*, 18, 181, 2017.

[3] Samuel M. Nicholls, Wayne Aubrey, Arwyn Edwards, Kurt de Grave, Sharon Huws, Leander Schietgat, André Soares, Christopher J. Creevey, Amanda Clare, Recovery of gene haplotypes from a metagenome. *BioRxiv*, 223404, 2019.

[4] Valeria Bortolaia, Rolf S. Kaas, Etienne Ruppe, Marilyn C. Roberts, Stefan Schwarz, Vincent Cattoir, Alain Philippon, Rosa L. Allesoe, Ana Rita Rebelo, Alfred Ferrer Florensa, Linda Fagelhauer, Trinad Chakraborty, Bernd Neumann, Guido Werner, Jennifer K. Bender, Kerstin Stingl, Minh Nguyen, Jasmine Coppens, Basil Britto Xavier, Surbhi Malhotra-Kumar, Henrik Westh, Mette Pinholt, Muna F. Anjum, et al, ResFinder 4.0 for predictions of phenotypes from genotypes. *Journal of Antimicrobial Chemotherapy*, 75(12), 2020.

Integration of publicly available high-throughput transcriptomic and proteomic studies of human tissues

Mitra Barzine[1], Nuno Fonseca[1], James Wright[2,3], Jyoti Choudhary[2,3] and Alvis Brazma[1]
[1] EMBL-EBI, Wellcome Genome Campus, CB10 1SD, Cambridge, UK
[2] Wellcome Sanger Institute, Wellcome Genome Campus, CB10 1SA, Cambridge, UK
[3] Institute of Cancer Research, 123 Old Brompton Road, SW7 3RP, London, UK

Corresponding author: mitra.barzine@gmail.com

The sampling of undiseased tissues is often challenging for obvious ethical reasons. Thus, the community usually refers to a few selected expression studies to support or validate new findings. Most of the time, external data is retrieved (e.g., from an atlas) and is integrated as-is. Examining the soundness of this approach has become increasingly imperative. Besides, while these reference studies have many overlapping tissues, few have attempted to integrate the data in any way. The study [1] investigates gene expression consistency between high-throughput transcriptomics and proteomics across tissues and studies.

Raw data from bulk RNA sequencing [2,3,4,5] and bottom-up label-free tandem-mass spectrometry [6,7] studies have been processed with consistent processing pipelines and gene annotation. The data have been compared and integrated for each biological layer before their joint analyses. Note that mitochondrial protein-coding genes (n=13) proved to be a significant biasing source and were removed from most analyses.

Transcriptomic samples with identical histological origin have higher levels of correlation than samples collected for the same study. Globally, mRNAs (including ubiquitous and more tissue-specific ones) show similar expression profiles across studies for a given set of tissues. On the other hand, the proteomic data present high discrepancies between studies due to the high detection variability of the mass spectrometry.

Nonetheless, the joint study of mRNA and protein expressions highlights that, for most tissues, we can observe quite good (and significant) correlation levels across biological layers, even when the samples have a different genetic background (due to the independent sampling). Many genes present a similar expression pattern for their mRNA and protein, e.g., genes for which the proteins have been detected in a single tissue are more likely to have an mRNA that is tissue-specific. Additionally, gene groups have presented interesting functional enrichments of biological processes: highly correlated protein and mRNA pairs are enriched in catabolic processes. In contrast, the most anticorrelated ones are enriched for ribosomes and ncRNAs regulation. Finally, highly tissue-specific genes are enriched for signalling processes.

Integrating proteomics and transcriptomics, even from independent studies, can help to improve our general knowledge and current biological models. This work has established an extensive list of genes presenting a robust transcriptomic and proteomic expression across tissues and studies. Besides, the highlighting of gene groups with distinct functional enrichment profiles has laid a framework for further research.

References

[1] M. Barzine. *Investigating normal human gene expression in tissues with high-throughput transcriptomic and proteomic data.* PhD thesis, University of Cambridge, July 2020.

[2] J.C. Castle and al. Digital genome-wide ncrna expression, including snornas, across 11 human tissues using polya-neutral amplification. *PLOS ONE*, 5:1–9, 07 2010.

[3] D. Brawand and al. The evolution of gene expression levels in mammalian organs. *Nature*, 478:343–348, 10 2011.

[4] M. Uhlén and al. Tissue-based map of the human proteome. *Science*, 347, 2015.

[5] M. Melé and al. The human transcriptome across tissues and individuals. *Science*, 348:660–665, 05 2015.

[6] M.-S. Kim and al. A draft map of the human proteome. *Nature*, 509:575–581, 5 2014.

[7] M. Wilhelm and al. Mass-spectrometry-based draft of the human proteome. *Nature*, 509:582–587, 05 2014.

Spatial omics network analysis with tysserand and mosna

Alexis Coullomb[1], Vera Pancaldi[1]

[1] Centre de Recherches en Cancérologie de Toulouse, 2 Avenue Hubert Curien, 31100, Toulouse, France

Corresponding Author: alexis.coullomb@inserm.fr

The last decade has seen the emergence of technologies to produce maps of protein or RNA content of single cells in intact tissue slices. These spatial and single cell - resolved omics methods offer great opportunities to further understand developmental processes and diseases such as cancer. Yet, data generated by these technologies require adapted methods to extract as much information as possible - such as the spatial relations between variables (RNA or protein counts) - and to cope with the size and high dimensionality of these datasets.

Here we will show how representing solid tissue samples as networks is fruitful to analyse cell-cell interactions and find local communities. In these networks nodes are cells and edges represent contact between them. We present *tysserand* [1], a Python library to reconstruct spatial networks from bioimages with high speed and accuracy. *tysserand* can reconstruct networks from segmentation images or coordinate arrays, it implements several reconstruction methods and utilities to choose the most appropriate parameters, and can output networks in formats compatible with several libraries dedicated to network and single-cell analysis.

We demonstrate how these networks can be analyzed with *mosna*, the Multi-Omics Spatial Network Analysis Python library. *mosna* can compute z-scored mixing matrices and assortativity [2], a measure of how cell types tend to interact preferentially with each other in whole samples. To study cell interactions at a more local scale , *mosna* also features the *Neighbors Aggregation Statistics* (NAS) method, which is designed to find spatial areas or local communities of specific cell types or states. In this method, variables of each cell and their first order neighbors are aggregated and summarized by their central tendency (mean or median) and variability (standard deviation, interdecile range, …). These "aggregation statistics" are clustered with a noise-aware algorithm that can define arbitrary-shaped clusters. The clusters define spatially coherent areas, some of them contain several cell types, and a given cell type is not necessarily limited to one specific cluster. Furthermore, we can subtract for each cell the contribution of its phenotype to RNA or protein counts before the aggregation analysis, in order to highlight variables that are modulated by spatial factors. These spatial areas can then be compared between each other or across patients to assess how spatial patterns are associated with specific biological processes such as development or tumor progression.

The *tysserand* and *mosna* libraries can thus be used on various multi-omics spatial datasets to discover spatial clusters and patterns, uncovering local communities and interactions between different cell types and states.

References

1. Coullomb, Alexis, and Vera Pancaldi. "Tysserand-Fast reconstruction of spatial networks from bioimages." *bioRxiv* (2020).
2. Newman, Mark EJ. "Assortative mixing in networks." Physical review letters 89.20 (2002): 208701.

Small non-coding RNA profiling in Germ Cell Tumours

Sean LAIDLAW[1,2], Luz ALONSO-CRISOSTOMO[3], Matthew J. MURRAY[3], Nicholas COLEMAN[3], Raheleh RAHBARI[2]

[1] Master Bioinformatique, Université de Montpellier
[2] Wellcome Sanger Institute, Hinxton, Cambridge
[3] Department of Pathology, University of Cambridge

Corresponding Author: sean.laidlaw@etu.umontpellier.fr

Germ Cell Tumours (GCT) are thought to be derived from aberrant primordial germ cells[1]. Although when they emerge and how they diversify remains unclear. They are a heterogenous group of tumours, presenting diverse histological subtypes, and are the most common malignancy in young males[2].

Regulatory RNA plays a key role in modulating gene expression, and small non-coding RNA has previously been profiled in various types of cancer[3,4]. MicroRNAs (miRNA) are the most frequently studied class of small noncoding RNA. They integrate with argonaute proteins to create a sequence specific silencing complex (miRISC) to hybridize to specific mRNA and silence target-gene translation[5]. Another regulatory class, piRNAs, are especially present in germline cells and are thought to interact with the PIWI family of proteins to silence transposable elements[6]. Both of these have been extensively studied in model organisms but study of these small non-coding RNA in subtypes of human GCT has so far been limited.

Here we present findings from a transcriptome sequence study of small RNA isolated from 8 gonadal controls and 10 testicular and ovarian GCTs, further classified by histological subtype. We demonstrate that there is a general underexpression of piRNA in GCT compared to controls. Furthermore, the signature of miRNA overexpression allows differentiation between the histological subtypes of testicular GCT, and demonstrates similar profiles between metastases and primary GCT for both classes of small RNA. We observed mir-371-373 to be universally upregulated in GCT regardless of histological subtype, as well as significant downregulation of the tumour suppressing let-7 class of miRNA. Our data suggests global piRNA downregulation across all testicular and ovarian derived GCTs.

References

1. Palmer, et al. Malignant Germ Cell Tumors Display Common MicroRNA Profiles Resulting in Global Changes in Expression of Messenger RNA Targets. *Cancer Research* 70. 7(2010): 2911–2923.
2. Hayes-Lattin, et al. Testicular Cancer: A Prototypic Tumor of Young Adults. *Seminars in Oncology* 36. 5(2009): 432–438.
3. Bloomston, et al. MicroRNA Expression Patterns to Differentiate Pancreatic Adenocarcinoma From Normal Pancreas and Chronic Pancreatitis. *JAMA* 297. 17(2007): 1901.
4. Takamizawa, et al. Reduced Expression of the let-7 MicroRNAs in Human Lung Cancers in Association with Shortened Postoperative Survival. *Cancer Research* 64. 11(2004): 3753–3756.
5. Sana, et al. Novel classes of non-coding RNAs and cancer. *Journal of Translational Medicine* 10. 1(2012): 103.
6. Siomi, M., et al. "PIWI-interacting small RNAs: the vanguard of genome defence. *Nat Rev Mol Cell Biol* 12, (2011): 246–258 .

Comparison of two pipelines to study spatial transcriptomics

Anne-Elodie RECEVEUR[1], Christophe LE PRIOL[1], Andrée DELAHAYE-DURIEZ[1,2]

[1] Inserm UMR 1141 NeuroDiderot, Université de Paris, 75019, Paris, France
[2] UFR Santé Médecine Biologie Humaine, Université Sorbonne Paris Nord, Bobigny, France

Corresponding Author: anne-elodie.receveur@inserm.fr

Spatial transcriptomics is a new type of genome-wide expression profiling methods that allow to count and localize the transcripts on histological sample sections. This technique has been used to lead some researches about human diseases, like the characterisation of some specific cells inducing the Alzheimer's pathology [1]. For these methods, specific slides that capture poly(A) mRNA are needed. These slides contain spots with spatially barcoded mRNA-binding oligonucleotides in addition to UMIs (Unique Molecular Identifier). As the sequence of each barcode is known, it can be related to coordinates. The RNAs are fixed by the oligo by diffusion through the tissue section. Reverse transcription occurs while the tissue is still in place, generating a cDNA library that incorporates the spatial barcodes and preserves spatial information. After sequencing of the strands, their localization is retrieved. Tools and workflows to process and analyse these particular sequencing data have been developed. One of the first companies that market products and instruments to carry out spatial transcriptomics, 10X Genomics (under the name of Visium Spatial Gene expression Solution), provides a workflow with an open source software: Space Ranger (https://support.10xgenomics.com/spatial-gene-expression/software/pipelines/ latest/using/count). Here, we propose to compare Space Ranger ability to generate gene expression count matrices from sequencing data with another widely used open source pipeline: ST Pipeline [2].

ST Pipeline and Space Ranger are both of them an informatic pipeline dedicated to the spatial transcriptomics, allowing to study the data coming from the sequencer and generates count matrices and other data useful for downstream analysis. Here, we used the same dataset of mouse brain (https://support.10xgenomics.com/spatial-gene-expression/datasets/1.1.0/V1_Adult_Mouse_Brain) publicly available on the website of 10X Genomics to evaluate both pipelines.

Between the two pipelines, several steps are similar, *e.g.* the quality filtering, the mapping, the demultiplexing, the annotation and the read counting, but we can notice some differences with the parameters given during the call of the same software used, like STAR [3] for the read mapping step. The scripts called during the process were analysed and to take into consideration the parameter differences. Moreover, each tool has its own specificities, which implies that some analyses are not found in both of the pipelines but just only in one, *e.g.* contaminant discarding only performed by ST Pipeline. We compared the outputs of the two pipelines, and more particularly the count matrices that are retrieved.

We will provide recommendations and criteria to help future users to choose between these two tools based on this comparison. As an example, we will apply these recommendations to select the best tool for the analysis of newly generated spatial transcriptomics data from 16 hippocampus sections through 3 different phases of the epileptogenesis in a rat model for Mesial Temporal Lobe Epilepsy (MLTE).

References

1. Wei-Ting Chen *et al,* Spatial Transcriptomics and *In Situ* Sequencing to Study Alzheimer's Disease, *Cell*, (182):1-16, 2020
2. José Fernandez Navarro *et al*, ST Pipeline: an automated pipeline for spatial mapping of unique transcripts. *Bioinformatics*, (33/16):2591-2593, 2017
3. Alexander Dobin *et al*, STAR: ultrafast universal RNA-seq aligner, *Bioinformatics*, (29/1):15-21, 2013

Decoding the RNA regulatory network to identify microRNA-sponge mechanisms involving circular RNA

Rose-Marie FRABOULET [1], Sébastien CORRE [1], Marie-Dominique GALIBERT [1,2], Yuna BLUM [1]

1 IGDR (Institut de Génétique et Développement de Rennes)—UMR6290, CNRS, Univ Rennes, F-35000 Rennes, France.
2 Department of Molecular Genetics and Genomics, Hospital University of Rennes (CHU Rennes), F-35000 Rennes, France.

Presenting Author
Corresponding Authors: yuna.blum@univ-rennes1.fr

1. Background

Non-coding RNAs (ncRNAs) represent a large component of the human transcriptome and have been shown to play important roles in cancers. The ability of ncRNAs to control gene expression makes them critical components of the cell plasticity and attractive targets for drug development. In particular, circular RNA (circRNA), a class of ncRNAs that have been poorly studied until now, can act as microRNA (miRNA) sponges[1] and may thus indirectly regulate expression of genes involved in cell plasticity. In this project, we propose to infer the RNA regulatory network underlying treatment resistance in melanoma, by integrating multi-level transcriptomic data and accounting for co-expression and predicted physical bindings. We aim at automatically identifying at a large-scale, sponge mechanisms involving circRNAs associated to treatment resistance, that could be targeted by specific therapies.

2. Results

We generated a multi-level transcriptomic data for the different types of RNAs (mRNA, miRNA and circRNA) from melanoma cell lines, resistance or sensitive to treatment (BRAF inhibitor). From these data, we identified RNA signatures associated to treatment resistance. We then built an extensive miRNA-binding database using the TargetScan[2] program, considering both mRNA and circRNA as potential targets. For each predicted miRNA-binding sequence (MRE), the context score representing the targeting efficacy was calculated as proposed by TargetScan. This score is calculated using cumulative scores considering all the features of the site (seed-pairing stability, location, target site abundance…). We inferred a RNA-regulatory network, where nodes represent the different types of RNAs and edges represent the co-expression between two miRNA-targets or a physical binding between a miRNA and its putative target.
Sponge candidate were retrieved according to two main criteria: (1) shared miRNA(s) and (2) co-expression between the circRNA and mRNA(s). We proposed different metrics to automatically estimate the sponge efficiency and the extent of sponge impact for each circRNA, by taking into account the number of impacted genes and the previously calculated MRE context scores. Based on our metrics, a set of experiments will be carried out to validate the best circRNA sponge-candidates.

3. Conclusion

Our approach allows inferring a RNA regulatory network based on expression and binding interactions. We introduce different metrics to automatically identify circRNA sponges and prioritize the most likely and relevant sponge mechanisms for functional validations. A shiny web app is under development that will allow biologists to explore and visualize the constructed RNA network and focus on sponge mechanisms of interest.

References

1. Memczak, S., Jens, M., Elefsinioti, A. & Torti, F. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495**, 333–8 (2013).
2. Agarwal, V., Bell, G. W., Nam, J.-W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4**, e05005 (2015).

Investigating resistance to IDH inhibitors in acute myeloid leukemia

Alexis Hucteau[1], Nina Verstraete[1], Feng Wang[2], Courtney Dinardo[2], Koichi Takahashi[2], Vera Pancaldi[1] and Jean-Emmanuel Sarry[1]

[1] INSERM, Centre de Recherches en Cancérologie de Toulouse (CRCT, Toulouse, France)
2 The University of Texas MD Anderson Cancer Center (Houston, TX, United States)

Corresponding Author: alexis.hucteau@inserm.fr

1. Introduction

The mutation in the gene isocitrate dehydrogenase 1 (IDH1) is implicated in Acute Myeloid Leukemia (AML), as cells with the alteration abnormally produce an oncometabolite 2-hydroxyglutarate (2-HG). 2-HG was found to cause widespread changes in DNA methylation [3]. IDH inhibitors have shown good clinical response in AML patients. However, primary or acquired resistance to IDH inhibitor therapies represent a major problem limiting their efficacy. The mechanisms that mediate resistance to IDH inhibition are poorly understood. We present an analysis of gene expression, inferred Transcription Factor (TF) activity, and clinical outcome in AML patients to uncover pathways related to IDH inhibitor resistance.

2. Materials and Methods

We studied 64 patients with relapsed or refractory AML who received IDH inhibitor therapy [1]. Gene expression was profiled by RNAseq, at diagnosis and relapse. We used the R package VIPER [2] and the dorothea network [3] to infer TF activities.

3. Results

The analysis of TF activities in relation to the response to IDHi showed that the activity of RUNX1 and CEBPa are linked to the clinical outcome. A Vitamin D metabolic process is found to be enriched in Non-Responders and is currently studied for its link to CEBPa in AML IDHm cell lines [4]. Statistical analyses showed correlation between the activity of RUNX1 and the presence of IDHm in a cohort from an independent dataset [5] where RUNX1 is more active when IDH is mutated. Some other TFs are significantly differentially inactive in Complete remission samples compared to other cases, including SMAD3, FLI1, NFE2L2, whereas the combination of the two TFs FOXA1--GATA3 is found to have higher activity. CEBPA, RUNX1, FOXA1, GATA3 and SMAD3 are associated with differentiation, confirming that leukemia stemness is an important pathway in primary resistance [1]. NFE2L2 is linked to the coordinated up-regulation of genes in response to oxidative stress and seems to also be linked to the Vitamin D pathway in AML [6].

4. Discussion

Co-occurring mutations in RUNX1 and RAS pathways genes, for example, are frequent in IDHm AML cases and studies have already made a link between these genes and response to the treatment [1] but the regulatory network behind that is still unclear. These preliminary results have to be confirmed experimentally and epigenetic analyses using chromatin conformation are in progress to link the regulation of these TF to epigenetic mechanisms that are impacted by IDH mutation. Mitochondrial metabolism is an important pathway of resistance in AML and these results confirmed its close relationship to stemness.

References

1. Feng Wang, Courtney Dinardo, Koichi Takahashi & al, Leukemia stemness and co-occurring mutations drive resistance to IDH inhibitors in acute myeloid leukemia, Nature Communication, 2021.
2. Alvarez MJ, Shen Y, Giorgi FM, Lachmann A, Ding BB, Ye BH & Califano, A. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. Nature Genetics, 2016.
3. Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J. “Benchmark and integration of resources for the estimation of human transcription factor activities.” Genome Research. 2019
4. Marie Sabatier, Jean-Emmanuel Sarry et al. Activation of Vitamin D Receptor Pathway Induces Differentiation in Acute Myeloid Leukemia with Isocitrate Dehydrogenase Mutations, 2021.
5. Maria E Figueroa & al. DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia, Cancer Cell, 2010.
6. Matan Nachlielya, Michael Danilenkoa, Dimethyl fumarate and vitamin D derivatives cooperatively enhance VDR and Nrf2 signaling in differentiating AML cells in vitro and inhibit leukemia progression in a xenograft mouse model, J Steroid Biochem Mol Biol, 2018

Identification of LpxC enzyme as a novel drug target against a resistant strain of *Acinetobacter baumannii* by using subtractive genomic approach

Manel ZOGHLAMI[1] ,Najla SADFI-ZOUAOUI[1] and Abdelmonem MESSAOUDI *[1,2]

[1] Laboratory of Mycology, Pathologies, and Biomarkers, Faculty of Science of Tunis, University of Tunis El Manar 2092, Tunis, Tunisia;

[2] Higher Institute of Biotechnology of Beja, Jendouba University, Habib Bourguiba Street, 9000 Beja, Tunisia

Corresponding Author: messaoudiabdelmonemster@gmail.com

Abstract:

The World Health Organization declared in its report in 2017 a list of microorganisms more critical and priority for research and development of new antibiotics, including *Acinetobacter baumannii*, it is a multi-resistant opportunistic bacteria responsible for nosocomial infections that can be sometimes serious in frail people; it can cause septicemia, pneumonia, and bacteriemia [1]. A Bioinformatics study was conducted on *A.baumannii* to identify a new therapeutic target not yet exploited in antibiotic therapy by using a subtractive genomics approach and to seek new therapeutic molecules by applying the computational methods, i.e. homology modeling, docking, and virtual screening.

The Data of Essential Genes database [2] allowed us to identify 452 proteins essential to the survival of the bacterium. The second step consists of eliminating the redundant paralogous sequences with a percentage of identity >60 by using the CD-HIT server [3], two protein sequences were eliminated and the final number remaining is 450 proteins. The next step is to eliminate proteins homologous to the human proteome, to do this, a comparative analysis of protein sequences was performed by BlastP [4] of the 450 proteins of *A.baumannii* against the human proteome (*Taxid: 9606*), proteins with a percentage of identity ≤38% are considered non-homologous [5]. 220 proteins were retained. The latter decreased to 164 proteins after eliminating hypothetical proteins (proteins predicted by gene identification tools, but for which there is no experimental evidence [6] and those with an amino acid number <100 since they are not suitable to represent essential proteins [7]. This work was followed by a study based on a manual comparison of metabolic pathways between pathogen and host proteins by the Kyoto Encyclopedia of Genes and Genomes server [8] to identify proteins unique to the pathogen involved only in metabolic pathways specific to it. The final number identified was 17 proteins. Determination of cellular localization by PSORT-B [9] and calculation of molecular weight by the Protein Molecular weight tool, are also important criteria for the identification of therapeutic targets since it has been shown that cytoplasmic proteins having molecular weight <100 are likely to be potential therapeutic targets[10]. The 17 proteins previously identified underwent a protein druggability analysis via Drug Bank [11]. The objective is to identify the percentage of similarity that these proteins share with known target proteins that can interact with existing drugs (high affinity) without modulating their activity [12]. The result provided by this analysis allowed us to identify 8 proteins considered druggable, included the lpxC enzyme. Among these, we have to choose the most appropriate one to establish the virtual targeting. Referring to the literature, and based on several criteria, we could choose the enzyme LpxC, since it represents a potential therapeutic target in other microorganisms [13] and The importance of the vital metabolic pathway in which it is involved "biosynthesis of lipopolysaccharides" [14]. In addition, LpxC presents a cytoplasmic cellular localization and a molecular mass of 30.35 KDa. Regarding the availability of the three-dimensional structure, a search was made using the PDB database [15] to distinguish proteins with a 3D structure already determined experimentally from those with a 3D model requiring the adoption of the concept of homology modeling to build a 3D structure of the chosen target. The result confirmed that the enzyme LpxC presents the most adequate model (PDB ID: 5N8C) with the highest percentage of identity (59%). This leads us to choose the enzyme LpxC, then, building a 3D model using the SWISS-MODEL server [16] and validated by the SAVES server.

The second step of this study, is to perform the molecular docking, to do this , a search in the Binding Database [17] was performed, we could collect 424 chemical compounds that can act on the enzyme LpxC and after performing the screening by the software Autodock (v.1.5.6) , we chose the 5 best ligands with the lowest binding energies (Ranging from -7.91 kDa/Mol and -8.54 kDa/Mol) that can be proposed as potential inhibitors including the compound PubChem 70701903 (Belongs to the class of benzamides, antibiotics capable of acting on the enzyme lpxC of *Escherichia coli* and *Pseudomonas aeruginosa* [18]). It has a maximum energy of -8.54 kDa/Mol and it is stabilized by 4 hydrogen bonds near the active site of the enzyme (his240). Predicting the ADME-T properties to obtain more active molecules (leads) and performing a complimentary in vitro study are an asset to validate the obtained result.

References:

[1] : Kerry Montefour. An emerging multidrug resistant pathogen in critical care. *Critical care nurse,* (28/1): 15-25,2008.

[2] : Ren Zhang. DEG: a database of essential genes. *Nucleic Acids Research*,D271-D272,2004.

[3] : Weizhong Li. Cd hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 1658-1659,2006.

[4] :SF Altschul. Basic Local Alignment Search Tool. *J Mol Biol*, 403-10, 1990.

[5] : MEENU Goyal. *In silico* identification of novel drug targets in *Acinetobacter baumannii* by subtractive genomics approach. *Asian J pharm clin Res*, (11/3) : 230-236,2018.

[6] : Gert Lubec. Searching for hypothetical proteins: Theory and practice based upon original data and literature. *Prog Neurobiol*, (1/2): 90-127, 2005.

[7] : Gupta Sunil Kumar. Definition of potential targets in *Mycoplasma pneumonia* through subtractive genome analysis. *Journal of antivirals and antiretrovirals*, (2): 038-041,2010.

[8] : Hiroyuki Ogata. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, (27/1): 29-34, 1999.

[9] : Nancy Y.Yu. Psortb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes.*Bioinformatics*,(26/13): 1608-1615,2010.

[10] : Shakhinur Islam Mondal. Identification of potential drug targets by subtractive of genome analysis of *Escherichia coli* O57:H7: an *in silico* approach. *Adv Appl Bioinform Chem*, (8): 49-63,2015.

[11] : David S.Wishart. DrugBank: A knowledgebase for drugs , drug actions and drug targets. *Nucleic Acids Res*, (36): D901-D906,2008.

[12] : T Liu. Identifying druggable targets by protein microenvironments matching: Application to transcription factors. *CPT Pharmacometrics syst pharmacol*, (3/1): e39,2014.

[13] : Chin-sheng yu. Predicting subcellular localization of proteins for gram –negative bacteria by support vector machines based on n-peptide composition. *Protein sci,(*13/5): 1402-1406,2004.

[14] : Szalo I.M. Le lipopolysaccharide d'*Escherichia coli* : structure, biosynthese et roles. *Ann. Med. Vet*, (150): 108-124,2006.

[15] : Helen M. Berman. The protein Data Bank. *Acta cryst*, D58: 899-907, 2002.

[16] : Torsten Schwede. Swiss-Model: an automated protein homology-modeling server. *Nucleic Acids research*, (31/13): 3381-5,2003.

[17] : Tiqing Liu. Binding DB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids research*, (35): D198-201,2007.

[18] : Adam W Barb. Mechanism and inhibition of lpxC: an essential Zinc-dependent deacetylase of bacterial lipid A synthesis. *Curr pharm Biotechnol*, (9/1): 9-15,2008.

Identifying explicit and tacit knowledge in a life science knowledge base

Johanna Saoud[1], Alain Gutierrez[1], Marianne Huchard[1], Pierre Silvie[2,3] and Pierre Martin[2]
[1] LIRMM, Univ Montpellier, CNRS, Montpellier, France
[2] CIRAD, UPR AIDA, F-34398 Montpellier, France
AIDA, Univ Montpellier, CIRAD, Montpellier, France
[3] PHIM Plant Health Institute, Montpellier University,
IRD, CIRAD, INRAE, Institut Agro, Montpellier, France

Corresponding author: johanna.saoud@etu.umontpellier.fr

An alternative to the use of synthetic pesticides and antibiotics in agriculture is to spray local plants extracts, in aqueous or essential oil form. To this end, the Knomana knowledge base [1] compiles various knowledge sets on plant use such as the 42000 descriptions of pesticidal plant uses for plant, animal, and public health presented in the literature. As the One Health approach dictates to be aware of the additional uses of these pesticidal plants to prevent their unintended effects on the animal, the human, and their environment, the challenge for the domain experts (e.g. entomologist, pathologist) is thus to identify the pesticidal plants in Knomana considering the One Health approach.

With the aim to present knowledge to the expert using a compact and comprehensive formalism, in [2], we computed the Duquenne-Guigues basis (DGB) of implications on an excerpt of Knomana, in which each plant is described using its taxonomy (i.e. species, genus, and family), to be consumed as food, and to be used in medical care. The DGB method is based on Formal Concept Analysis (FCA) and provides a cardinality-minimal set of non-redundant implications. By considering a reduced knowledge set, this work identified 3 types of knowledge elements in the implications: knowledge on plant use at diverse taxonomy levels (e.g. *Plants from Meliaceae family are not consumed as food*), plant taxonomy (e.g. *A plant from Salvia genus is from Lamiaceae family*), and side effect of the knowledge set, e.g. *a plant from the Piperaceae family is from the genus Piper*. This latter illustration is not in accordance with taxonomic referential and thus informs on the extend of knowledge inserted in Knomana. Moreover, as plant taxonomy is known by the experts, removing it from the implications eases their reading but makes it tacit knowledge.

Implementing this method to select pesticidal plants requires to consider Knomana as a multidimensional (ternary) dataset, and thus to use the extension of FCA devoted to this kind of knowledge discovery, i.e. Relational Concept Analysis (RCA). Therefore, computing the DGB of implications based on RCA provides linked set of implications which includes the existential quantifier. Converting this formulation as practical expression is a need for the domain experts.

This poster describes the product line that formulates Knomana knowledge on pesticidal plants as implications, from which the implicit knowledge elements were removed and the side effects are highlighted to alert the expert. This product line was developed using the library fca4j from Cogui software (http://www.lirmm.fr/cogui/), that provides the RCA based DGB of implications, and using a post-process which differentiates the 3 types of knowledge elements within the implications. As an illustration, this poster presents the implications on *Spodoptera frugiperda*, a highly polyphagous insect that is close to invade South of Europe. The perspective of this work is to identify pesticidal European plants species that share chemical components similarities with plants used to control this pest in its native area.

References

[1] Pierre J. Silvie, Pierre Martin, Marianne Huchard, Priscilla Keip, Alain Gutierrez, and Samira Sarter. Prototyping a knowledge-based system to identify botanical extracts for plant health in sub-saharan africa. *Plants*, 10(5), 2021.

[2] Johanna Saoud, Alain Gutierrez, Marianne Huchard, Pascal Marnotte, Martin Silvie, and Pierre Martin. Explicit versus Tacit Knowledge in Duquenne-Guigues Basis of Implications: Preliminary Results. Montpellier, France, May 2021. Submitted to Analyzing Real Data with Formal Concept Analysis, RealDataFCA'2021.

Exploring the transcriptional landscape of gonadotroph tumor microenvironment with single cell RNA-seq

Benoît ALIAGA[1,2], Julie RIPOLL[1], Marie CHANAL[3], Hector HERNANDEZ-VARGAZ[3], Emmanuel JOUANNEAU[3,5,7], Alexandre VASILJEVIC[3,4,7], Gerald RAVEROT[3,6,7], Eric RIVALS[1] and Philippe BERTOLINO[3]

[1] Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, UMR 5506, Univ. Montpellier et CNRS, 161 rue Ada, 34095, Montpellier, France

[2] Master Sciences et Numérique pour la santé, Faculté des Sciences, Univ. Montpellier, Campus Triolet Place Eugène Bataillon, 34095, Montpellier, France

[3] Centre de Recherche en Cancérologie de Lyon, INSERM U1052, CNRS UMR5286, Université Claude Bernard Lyon1, 28 rue Laennec, 69008, Lyon, France

[4] Centre de Pathologie Est, Groupement Hospitalier Est, Hospices Civils de Lyon, 69677, Bron, France

[5] Université de Lyon1, Service de Neurochirurgie, Hôpital Neurologique, Hospices Civils de Lyon, 69677, Bron, France

[6] Fédération d'Endocrinologie, Groupement Hospitalier Est, Hospices Civils de Lyon, 69677, Bron, France

[7] Faculté de Médecine Lyon Est, Université de Lyon1, 69677, Lyon, France

Corresponding Authors: philippe.bertolino@lyon.unicancer.fr, rivals@lirmm.fr

Pituitary is a small endocrine gland located at the base of brain. The anterior part of the gland produces and secretes several hormones which play important roles in the growth, the development and the functions of numerous organs. Different subtypes of endocrine pituitary tumors exist. While somatic mutations, copy number alterations, cell-cycle dysregulations and epigenetics changes have been identified in most pituitary tumor-subtypes, the transformation and tumor-driving mechanisms in the gonadotroph subtype remain unknown [1]. Here, we want to explore whether the implication of the tumor microenvironment plays a role in gonadotroph adenoma based on its emerging roles as candidate actor and therapeutic target in pituitary tumors [2]. To that extent, our main objective is to characterize the composition and transcriptional landscape of gonadotroph tumors microenvironment through the use of single cell genomics.

To address our objective, we performed single cell RNA sequencing (scRNAseq) on surgically resected gonadotroph tumors. Single cells dissociated from tumors were encapsulated using a chromium controller (10X Genomics) prior to library generation and sequencing. During my internship, I develop a single cell RNA-seq (scRNA-seq) pipeline with conda and snakemake [3] to ensure an automatized and reproductible pipeline. The latter comprises three steps: (i) data acquisition (reads quality check, read trimming, demultiplexing and mapping with StarSolo), data cleaning (cells quality check, batch effect correction, and normalization) and cell subpopulation identification both with Seurat.

Here, we aim (i) at optimizing a scRNA-seq pipeline to explore the cellular composition of gonadotroph tumors, (ii) at determining if the microenvironment influences the tumorigenesis or if tumor cells have a genetic origin, and (iii) at pointing out new tumorigenesis factors/biomarkers in gonadotroph adenoma.

Acknowledgements

The participation at JOBIM 2021 is granted by the Faculté des Sciences of the University of Montpellier (master Sciences et Numérique pour la Santé, BA). The work is supported by La Région-Rhône Alpes-Auvergne, La Ligue de la Loire et du Rhône, la Fondation ARC pour la Recherche sur le Cancer et les Hospices Civils de Lyon. ER & JR are supported by project FluoRib (INCa grant n°2018-131). I thank Philippe Bertolino, Julie Ripoll and Eric Rivals for agreeing to supervise me on this work.

References

1. Melmed, S. Pituitary-Tumor Endocrinopathies. *N. Engl. J. Med.* **382**, 937–950 (2020).
2. Ilie, M. D., Vasiljevic, A., Raverot, G. & Bertolino, P. The microenvironment of pituitary tumors-biological and therapeutic implications. *Cancers (Basel).* **11**, 1–22 (2019).
3. Koster, J. & Rahmann, S. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).

Use of deconvolution for oncology: a methods review

Jacobo Solórzano, Corinne Bousquet, Yvan Martineau and Vera Pancaldi

Centre de Recherches en Cancérologie de Toulouse (CRCT) 2 Avenue Hubert Curien, 31037, Toulouse, France

Corresponding Author: jacobo.solorzano@inserm.fr

In tumours, cancer cells are surrounded by a diverse collection of immune and normal stromal cells, which have an impact on tumour progression and response to therapy. Quantification of cell type abundance from bulk datasets coming out of tumour biopsies (deconvolution) is one of the most relevant computational approaches for immuno-oncology, as, ideally, it aspires to be a cheaper alternative to single cell experiments, applicable to any kind of omics data. Deconvolution can be classified in (1) supervised, where bulk data as well as prior knowledge are needed to infer each cell type's abundance; and (2) unsupervised, where some specific expression profiles (components) that define the bulk data are inferred and quantified across the samples[1,2]. Unsupervised deconvolution could potentially quantify the presence of novel cell types/phenotypes, whose expression signatures can be projected onto other datasets. Nevertheless, as components and sources are inferred mathematically, these methods are sensitive to noise, hindering the identification of the source behind the detected components, and thus compromising their usefulness for cell type deconvolution [2].

Using unsupervised deconvolution when accounting for the presence of cell types is a double-edged sword, as biological sources with similar expression signatures as technical noise could be confounded into similar components if the number of components is low. For instance, we replicated Puleo *et al.* 2018 [3] Independent Component analysis (ICA) and ran several supervised deconvolution methods, showing that the component initially associated with the age of the paraffin block is also consistently correlated with the abundance of Neutrophils as well as NK cells. Here we present a project to study the usefulness of unsupervised deconvolution approaches for detecting the presence of specific cell types over technical and biological noise, by comparing supervised and unsupervised deconvolution methods, and integrating the available metadata over pancreatic cancer samples. In particular using GEM-DeCan [4] for the supervised deconvolution and different unsupervised strategies involving the JADE ICA algorithm [5] on the well known datasets of Puleo et al. 2018 [3], TCGA [6] and ICGC [7]. Altogether we aim to provide guidelines on how to apply existing deconvolution techniques to characterize immune infiltrates which could be potentially useful for novel target discovery, exploring drug resistance mechanisms and performing patient stratification.

1. Stein-O'Brien, G. L., Arora, R., Culhane, A. C., Favorov, A. V., Garmire, L. X., Greene, C. S., ... & Fertig, E. J. (2018). Enter the matrix: factorization uncovers knowledge from omics. *Trends in Genetics*, *34*(10), 790-805.
2. Jin, H., & Liu, Z. (2021). A benchmark for RNA-seq deconvolution analysis under dynamic testing environments. *Genome biology*, *22*(1), 1-23.
3. Puleo, F., Nicolle, R., Blum, Y., Cros, J., Marisa, L., Demetter, P., ... & Maréchal, R. (2018). Stratification of pancreatic ductal adenocarcinomas based on tumor and microenvironment features. *Gastroenterology*, *155*(6), 1999-2013.
4. Xie, T., Pernet, J., Verstraete, N., Madrid-Mencía, M., Kuo, M. S., Hucteau, A., ... & Pancaldi, V. (2021). GEM-DeCan: Improving tumor immune microenvironment profiling by the integration of novel gene expression and DNA methylation deconvolution signatures. bioRxiv.
5. Miettinen, J., Nordhausen, K., & Taskinen, S. (2017). Blind Source Separation Based on Joint Diagonalization in R: The Packages JADE and BSSasymp. Journal of Statistical Software, 76(2), 1 - 31. doi:http://dx.doi.org/10.18637/jss.v076.i02
6. Samur, M. K. (2014). RTCGAToolbox: a new tool for exporting TCGA Firehose data. PloS one, 9(9), e106397.
7. Bailey, P., Chang, D. K., Nones, K., Johns, A. L., Patch, A. M., Gingras, M. C., ... & Grimmond, S. M. (2016). Genomic analyses identify molecular subtypes of pancreatic cancer. Nature, 531(7592), 47-52.

Mining novel biosynthetic gene clusters from metagenomics

Léonard Dubois[1] and Nicola Segata[1]

[1] Laboratory of Computational Metagenomics, CIBIO, Via Sommarive 9, 38123, Trento, Italy

Corresponding Author: nicola.segata@unitn.it

By sequencing the whole genetic material present in a sample, metagenomics provides a culture-independent approach to characterize microbial communities. Over the last decade, metagenomic profiling of the human gut microbiome led to the identification of species associated with conditions including several diseases and different lifestyles. As the resolution of metagenomic profiling increases, it is becoming clear that genetic variability within microbial species should be carefully considered as it can explain both microbial phenotypic variability and variable association with host conditions [1,2,3]. While improved computational methods are trying to profile the tremendous conspecific variability of microbial taxa within metagenomes [4], current methods are lacking the ability to identify the most functionally relevant portion of such variability which is frequently due to biosynthetic gene clusters (BGCs) of potential mobile origin [5].

In order to identify BGCs that can functionally explain intra-species variability of members of the gut microbiome and their potential association with host conditions, we developed a new method that *de novo* identifies them using a combination of intra species strain-level comparative genomics, gene synteny information from reference and metagenome-assembled genomes, and sequence composition features. We applied the new method to profile BGCs of 23 bacterial species across ~9,000 human gut samples. This highlighted the extensive presence of BGCs in the accessory genome of the analysed species. Indeed, we found more than 3,000 BGCs consisting of 6 to 16 genes (10th and 90th percentiles) that are strictly co-present or co-absent and adjacent on the genome. These BGCs are organized in a non-redundant resource that can be used to identify such mobile elements into metagenomic samples. The number of clusters ranges from a few tens to a few hundred per species (up to ~370 for *Bacteroides vulgatus*) and can represent from 8 to 25% of the gene content of the pangenome of a species.

The functional analysis of the identified BGCs reinforced the mobile nature of most of them as they are enriched for genes involved in mobile genetics elements such as transposases or relaxosome proteins. Moreover, functional analysis also revealed that in some species up to 40% of the clusters contain carbohydrate active enzymes useful for complex sugar degradation (e.g. *Ruminococcus gnavus* 34.6%, and *Bifidobacterium longum* 39.8%). Furthermore, the prevalence of specific BGCs varies a lot across conditions and particularly across levels of Westernization. For example, 81% of *Prevotella copri* vs 4% for *R. gnavus*' BGCs are differentially prevalent in Westernized versus non-Westernized populations (Fisher's exact test corrected p_value < 0.05). Altogether, our preliminary study confirms the ubiquity of BGCs, the possibility of identifying them *de novo*, and the potential relevance of profiling BGCs in under-characterized intestinal microbial species. We also provide a resource summarizing common BGCs of relevance for downstream analysis and for generation of functional host-microbiome interaction hypotheses.

References

[1] Karcher, N., Pasolli, E., Asnicar, F. et al. Analysis of 1321 Eubacterium rectale genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations. Genome Biol 21, 138 (2020).

[2] Zeevi, D., Korem, T., Godneva, A. et al. Structural variation in the gut microbiome associates with host health. Nature 568, 43–48 (2019).

[3] Fehlner-Peach, Hannah, et al. Distinct polysaccharide utilization profiles of human intestinal Prevotella copri isolates. Cell host & microbe 26.5 (2019).

[4] Van Rossum, T., Ferretti, P., Maistrenko, O.M. et al. Diversity within species: interpreting strains in microbiomes. Nat Rev Microbiol 18, 491–506 (2020).

[5] Cimermancic P., et al. "Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters." Cell 158.2 (2014).

Differential analysis with a single sample in the disease group

Omran Allatif[1]and Antoine Corbin[1]
CIRI, Inserm U1111, CNRS UMR5308, Université Claude Bernard Lyon 1, École Normale Supérieure de Lyon, Univ Lyon, F-69007, Lyon, France

Corresponding author: omran.allatif@ens-lyon.fr

1 Introduction

Differential features analysis such as in transcriptomics and proteomics is based on the biological variability between independent samples in each of the two compared groups. Doing differential analysis when one group consists of a single sample, while the second group has several biological replicates, reduces the analyzes to a descriptive level or leads to speculative assumptions based on previous experiences with similar data [1] & [2] .

2 Context

This situation is particularly frequent in clinical studies, for example in rare diseases, where sometimes it is not possible to have more than one sample. It also concerns clinical groups formed from heterogeneous diagnoses. Indeed, the criteria used to assign an individual to a disease group could depend on the practitioner's judgment and experience, which could be somewhat subjective. Even when the diagnosis is based on objective criteria, assigning individuals in groups is not always in line with the biological reality. Thus, same genetic disorder could have two different clinical manifestations, and the same clinical presentation could result from various genetic sources. While heterogeneity of profiles can hardly induce a Type I statistical error, it remains strongly associated with a Type II error; thus homogeneity is essential to reliably conclude. Comparing each patient *separately* to the control group, determining its profile in terms of differential features, then assigning him to a group, ensure the consistency between patients appearing clinically together. This step could be routinely applied as a preliminary step in differential analysis, where samples are taken from real patients or living ecological entities.

3 Method

We set up a non-parametric approach to infer a deregulation in gene/protein expression when the disease group consists of a single sample while the control group has several. This method is therefore very useful for correctly grouping the patients who have similar clinical manifestations and genetic profile by performing set operations for which we had previously developed and deployed a Shiny application. Statistically, our method is inspired and adapted from existing approaches based on the median of the distribution and the inter-quartile range. A feature is qualified as differentially relevant when it exceeds a relative threshold, which remains parametric taking into account the average variability per matrix expression row.

4 Conclusion

We created our method when working on proteomic data obtained by protein arrays as part of the study of the detection of auto-antibodies in Lupus patients (SLE). We could transpose this method for analyzing RNA-Seq data, one Vs. four samples having highly homogeneous sequencing depths. Comparing the results with the same data being analysed, four Vs. four samples, with the robust approach of limma-trend, common transcripts qualified as differential in both approaches is encouraging. Although still needs some improvement, especially to correct for multiple testing, our approach shows already good results.

References

[1] McCarthy DJ Robinson MD and Smyth GK. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, (26):139–140, 2010.

[2] Chen Y McCarthy DJ and Smyth GK. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic Acids Research*, (40):4288–4297, 2012.

A method combining two complementary approaches to trace the evolution of alpha-solenoid protein repertoires in Archaeplastida.

Céline Cattelin[1], Yves Choquet[1], Chantal Prevost[2], Olivier Vallon[1], Francis-André Wollman[1], Charles H. Robert[2] and Ingrid Lafontaine[1]

[1] UMR7141 CNRS/Sorbonne Université, IBPC Fondation Edmond de Rothschild, 13 rue Pierre et Marie Curie, 75005 PARIS, France

[2] UMR9080 CNRS/Université de Paris/PSL Research University, IBPC Fondation Edmond de Rothschild, 13 rue Pierre et Marie Curie, 75005 PARIS, France

Corresponding Author: ingrid.lafontaine@ibpc.fr

Abstract

In eukaryotes, the nucleus controls organelle gene expression at every post-transcriptional step, relying almost exclusively on nuclear-encoded proteins that are imported into the organelle and bind to their RNA target in a sequence-specific manner. Evolution of these RNA-binding proteins, together with the network of their RNA targets, is instrumental in the response of eukaryotic organisms to environmental changes.

In photosynthetic eukaryotes of the green lineage, the RNA-binding proteins involved in the nuclear control of chloroplast gene expression belong mostly to the PPR (pentatrico-peptide repeat) and OPR (octatrico-peptide repeat) families of alpha-solenoid proteins [1], which harbor 35 residue- (PPR) or 38 residue- (OPR) tandem repeats, each consisting of a pair of antiparallel alpha-helices. As expected from their evolutionary role, the PPR and OPR repertoires are remarkably diverse between organisms, with land plants containing more than 400 PPRs [2] and a few OPRs, whereas the microalga *Chlamydomonas reinhardtii* encodes only 14 PPRs [3] but as many as 127 OPRs [4], some of them possibly acting as endoribonucleases [5]. The evolution of these repertoires in terms of gene gain and loss likely reflects their genetic adaptation to different lifestyles or ecological niches.

Here we present two approaches for annotating alpha-solenoid proteins targeted to the chloroplast and the mitochondria using a semi-automated pipeline. The first approach retrieves candidate proteins with known OPR and PPR motifs using a profile-based similarity search, while the second approach selects candidate proteins likely to adopt an alpha-solenoid fold with key characteristics of OPR and PPR. This work will facilitate further investigations of the evolutionary history of alpha-solenoid proteins that are potentially involved in the regulation of organelle gene expression. Observed sequence variations will also allow us to better understand the determinants of their RNA affinity and specificity through physical-chemical and modelling studies.

Applied to Archaeplastida our method efficiently retrieves known OPR and PPR proteins, as well as TPR proteins and identifies new candidates. Results of the method applied to Diatoms will also be presented.

References

[1] A. Barkan, I. Small. Pentatricopeptide Repeat Proteins in Plants. *Annual Review of Plant Biology*, (65):415–442, 2014.

[2] C. Lurin et al.. Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *The Plant Cell*, (16):2089–2103, 2004.

[3] N. J. Tourasse et al.. PPR proteins of green algae. *RNA Biology*, (10):1526–1542, 2013.

[4] S. Eberhard et al.. Dual functions of the nucleus-encoded factor TDA1 in trapping and translation activation of atpA transcripts in Chlamydomonas reinhardtii chloroplasts. *The Plant Journal*, (67):1055–1066, 2011.

[5] Boulouis et al.. Spontaneous dominant mutations in chlamydomonas highlight ongoing evolution by gene diversification.A. *Plant Cell*, (27):984–1001, 2015.

Extensive benchmark of machine learning methods for quantitative microbiome data

Sébastien Fromentin[1], Florian Plaza Oñate[1], Nicolas Maziers[1], Samar Berreira Ibraim[1], Guillaume Gautreau[1], Oscar Gitton-Quent[1], Manolo Laiola[1], Soufiane Maski[1], Raphaëlle Momal[1], Florence Thirion[1], Franck Gautier[1], Nicolas Pons[1] and Magali Berland[1]

[1] Université Paris-Saclay, INRAE, MGP, 78350, Jouy-en-Josas, France

Corresponding Author: magali.berland@inrae.fr

1. Introduction

Characterization of microbial communities with omics technologies shed to light powerful biomarkers for diagnosis and prognosis in human health [1]. In particular, shotgun metagenomics allows a highly precise microbiome profiling. Indeed, prediction of phenotypic features, such as clinical status or disease states can help to stratify patients which is the first step toward precision medicine. Many machine learning (ML) methods have been developed to tackle classification and regression problems yet statistical specificities of metagenomic data make difficult the learning task [2]. In the present work, we compare the commonly used ML methods on a quantitative metagenomics dataset.

2. Methods

We developed a workflow in R to browse and compare ML methods for classification or regression implemented in the caret package [3]. A table where microbial features (species, functions, metabolites) are quantified across a set labelled samples (e.g: control/disease) is taken as input. Then, the selected models are trained and evaluated with repeated 10 fold cross validation. Each model is trained 100 times with different random splits. The Activeon Proactive workflow engine was used to efficiently distribute the computing load on multiple servers. The code is available upon request.

3. Results

We applied our workflow on a dataset where the fecal microbiota of patients with cardiovascular diseases is compared to healthy controls using shotgun metagenomics. Each model was tested on regression or classification problems, expected to be easy or difficult to predict. We compared the models with several indicators including predictive performance (F-score, R^2), stability across iterations and computational resources consumption. We also explored the impact of common preprocessing steps to remove non informative variables (near zero variance features, linear combo, etc.). We observed that a wide range of common methods show similar predictive performance (svm, pls, glm, rf, etc.) although some can be very slow (spls). Finally, the choice of the best model may be guided by other criteria like the interpretability that can give insights in the underlying biological hypotheses in order to provide insightful medical decisions.

References

1. Laura Judith Marcos-Zambrano *et al*. Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. *Frontiers in microbiology*, vol. 12, p. 313, 2021.
2. Isabel Moreno-Indias *et al*. Statistical and machine learning techniques in human microbiome studies: contemporary challenges and solutions. *Frontiers in Microbiology*, vol. 12, p. 277, 2021.
3. Max Kuhn. Caret: classification and regression training. *Astrophysics Source Code Library*, ascl-1505, 2015. https://topepo.github.io/caret/

Keywords – microbiome, machine learning, disease prediction, shotgun metagenomics, biomarker identification, precision medicine

MethMotif 2022: An update of the transcription factor binding motifs database that integrates tissue-specific features and DNA methylation profiles

Matthew Dyer[1], Quy Xiao XuanLin[2], Aida Ghayour-Khiavi[1], Roberto Tirado-Magallanes[2], Morgane Thomas-Chollier[3], Denis Thieffry[3] and Touati Benoukraf[1,2]

[1] Discipline of Genetics, Faculty of Medicine, Memorial University of Newfoundland, A1B 3V6, St. John's, Canada

[2] Cancer Science Institute of Singapore, National University of Singapore, Singapore 117599, Singapore

[3] Computational Systems Biology Team, Institut de Biologie de l'École Normale Supérieure (IBENS), CNRS, INSERM, École Normale Supérieure, PSL Research University, 75005 Paris, France

Corresponding author: tbenoukraf@mun.ca

The first release of our MethMotif database [1], an integrative cell-specific database of transcription factor (TF) binding motifs coupled with DNA has clearly shown how transcription factor binding site motifs are cell-specific and dependant to DNA methylation profiles. To facilitate cistrome/methylome data analysis and integration, we developed TFregulomeR, an R-library that combined an up-to-date compendium of ChIPseq and whole-genome bisulfite sequencing datasets [2]. These resources provide a novel framework that opens a new avenue for large-scale and multi-dimensional integrative analyses which has been proven to be useful to determine context-specific TF partners and to brought to light TF's cell-specific functions. Using TFregulomeR, we expanded the range of information available in the new release of MethMotif by listing a breakdown of context-specific TFs' co-factors and ontology terms of gene targets. Using our toolbox, we have shown that TF's target ontologies can differ notably depending on their partners, making the 2022 release of Methmotif (methmotif.org) the first TFBS database that incorporates context-specific position weight matrices coupled to epigenetic information and context-specific TFs' function. In this new release, we introduced Forked-Position Weight Matrices and Forked-Sequence Logos to better portray TF dimers at the DNA motif and methylation levels. These new representations better depict TFBS of a TF of interest connected to its segregated list of partners and improves PWM models of dimerized TFs, to enhance TFBS prediction power. Ultimately, this toolbox aims to facilitate the analysis of the consequences of epigenetic aberrations (DNA methylation) seen in diseases such as cancers. Overall, this update turns MethMotif into an integrative TFBS database with a diverse set of regulatory element analysis tools accessible to a broad audience.

Acknowledgements

This research was enabled in part by support provided by ACENET (www.ace-net.ca) and Compute Canada (www.computecanada.ca), as well as the Center for Health Informatics and Analytics (CHIA) at Memorial University.This work has been supported in part, thanks to funding from the Canada Research Chairs program and by the National Research Foundation, the Singapore Ministry of Education under its Centres of Excellence initiative.

References

[1] Quy Xiao Xuan Lin, Stephanie Sian, Omer An, Denis Thieffry, Sudhakar Jha, and Touati Benoukraf. MethMotif: an integrative cell specific database of transcription factor binding motifs coupled with DNA methylation profiles. *Nucleic Acids Research*, 47(D1):D145–D154, 10 2018.

[2] Quy Xiao Xuan Lin, Denis Thieffry, Sudhakar Jha, and Touati Benoukraf. TFregulomeR reveals transcription factors' context-specific features and functions. *Nucleic Acids Research*, 48(2):e10–e10, 11 2019.

Search algorithms for dinucleotide Position Weight Matrices

Marie Mille[1,2], Julie Ripoll[1] and Eric Rivals[1]

[1] Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier - UMR 5506 CNRS, Univ. Montpellier, 860 rue de St Priest, 34095, Montpellier, France

[2] Master Sciences et Numérique pour la Santé, Faculté des Sciences, Univ. Montpellier, Campus Triolet Place Eugène Bataillon, 34095, Montpellier, France

Corresponding author: eric.rivals@lirmm.fr; marie.mille01@etu.umontpellier.fr

Transcription regulation is an important cellular process. Specialized proteins, called Transcription Factors (TF), bind on short, specific, DNA sequences to regulate the expression of nearby genes. The sequences recognized by a TF in the vicinity of different genes are not identical, but similar. One captures the similarity of those binding site in different representations, which are generally called *motifs*. The most widely used sort of motifs are Position Weight Matrices (PWMs) (also known as a position-specific weight matrix (PSWM) or position-specific scoring matrix (PSSM)). A PWM is built from a multiple alignment of observed binding sequences and capture the observed variation of nucleotides at the different positions. Several databases (JASPAR, TRANSFAC, etc.) collect PWMs for known TFs. Those PWMs are used to scan new DNA sequences to find putative binding sites and possibly to annotate them. In the case of complete genomes, the scanning procedure for many PWMs may last a long time [1].

PWMs assume that the distinct positions of the binding sequence are independent of each other. However, several studies have observed that a mutation at a given position influences the probability of mutation at neighboring positions. To overcome this limitation of PWMs, Kulakovskiy et al. have proposed a more complex sort of motifs, called di-PWMs, which model the frequency of occurrence of dinucleotides in the binding sites (instead of mononucleotides for PWMs) [2]. Their studies show that di-PWMs improve in sensitivity compared to PWMs, and thus produce less false positives when scanning a sequence.

Our aim is to design new search algorithms for di-PWMs, either online or offline, and to implement them. Our online scanning algorithm computes a partial score for some positions in the current window, and estimates the maximum achievable score for the whole window. If this score does not match the user defined threshold, the window can be discarded. Our offline approach works in two steps. As for read mapping, the genome is first preprocessed to produce an index data structure. Then, for any given motif and a threshold score, we enumerate potential matching words (i.e. words whose score lies above the threshold) and search their occurrences in the index in optimal time. The difficulty is to design an efficient enumeration algorithm. Such an approach was developed for PWMs in the MOTIF software [3]. If time suffices, we plan to compare experimentally our implementations to that of an existing search tool for di-PWMs, called SPRY-SARUS (github) [1].

Acknowledgements

We thank the GEM Flagship project funded from Labex NUMEV (ANR-10-LABX-0020) for the internship of M. Mille. JR is supported JR by project FluoRib (INCa grant n°2018-131).
The participation of M. Mille at JOBIM 2021 is granted by the Faculté des Sciences of the University of Montpellier, Master Sciences et Numérique pour la Santé, parcours Bioinformatique, connaissances et données.

References

[1] Cinzia Pizzi, Pasi Rastas, and Esko Ukkonen. Finding significant matches of position weight matrices in linear time. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(1):69–79, January 2011.

[2] Ivan Kulakovsky, Victor Levitsky, Dmitry Oshchepkov, Leonid Bryzgalov, Ilya Vorontsov, and Vsevolod Makeev. From binding motifs in ChIP-seq data to improved models of transcription factor binding sites. *Journal of Bioinformatics and Computational Biology*, 11(01):1340004, February 2013.

[3] David Martin, Vincent Maillol, and Eric Rivals. Fast and accurate genome-scale identification of dna-binding sites. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 201–205, 2018.

Inferring genetic regulation of tumor microenvironment composition using mutliomic personalised analysis in cancer

Fabien Quinquis[1], Clémentine Decamps[1], Daniel Jost[2] and Magali Richard[1]
[1] Laboratory TIMC-IMAG, UMR5525, Univ.GrenobleAlpes, CNRS,38000, Grenoble,France
[2] LBMC, UMR5239, ENS de Lyon, 46 Allée d'Italie, CNRS, 69007, Lyon, France

Corresponding author: magali.richard@univ-grenoble-alpes.fr

High throughput multi-omic cancer studies have led to well-defined molecular classifications accounting for inter-patient variations. Current understandings of cancer biology and corresponding therapeutic strategies are based on these classifications. Nevertheless, these findings only reflect the most abundant tumor subtype in the examined sample, thus neglecting the fact that cancers consist of cells with different identities and origins (cell heterogeneity).

Here we propose to take advantage of recent advances in multi-omic high throughput sequencing technologies to study how the gene expression of 'pure' tumor cells specifically contributes to the regulation of the immune microenvironment. We designed a novel method and a corresponding R package, named *RiTMIC* (RegulatIon of Tumor MICrovenvironment). First, we reconstitute a surrogate differential expression matrix specific to tumor cells, at the patient/individual level, using a reference free deconvolution algorithm (EDec [1]) and a personalized differential expression approach (PenDA [2]). Second, we statistically infer which genes are involved in the regulation of the microenvironment composition. Using a realistic benchmark of simulated pancreatic tumor, we demonstrated that *RiTMIC* achieved high specificity and sensitivity to detect tumor-specific genetic regulation of immune cell fractions. We are currently applying our pipeline on several independent cohorts of non-small cell lung cancer to validate the method in real pathological context.

Reconstitution of gene expression specific to pure unmixed tumor cells, in each individual sample, should allow us to unravel currently hidden genetic regulators of tumor composition. These results will contribute to interrogate and revisit current tumorigenesis understandings, in the light of genetic models accounting for tumor heterogeneity.

References

[1] Vitor Onuchic, Ryan J. Hartmaier, David N. Boone, Michael L. Samuels, Ronak Y Patel, Wendy M. White, Vesna D. Garovic, Steffi Oesterreich, Matt E. Roth, Adrian V. Lee, and Aleksandar Milosavljevic. Epigenomic deconvolution of breast tumors reveals metabolic coupling between constituent cell types. *Cell reports*, 17(8):2075–2086, November 2016.

[2] Magali Richard, Clémentine Decamps, Florent Chuffart, Elisabeth Brambilla, Sophie Rousseaux, Saadi Khochbin, and Daniel Jost. PenDA, a rank-based method for personalized differential analysis: Application to lung cancer. *PLoS Computational Biology*, 16(5), May 2020.

BioGraph: A Julia package to extract the longest representative path from a pangenome graph

Nguyet Dang[1,2], Tuan Do[3] and Francois Sabot[1,2]
[1] DIADE, Univ Montpellier, IRD, Montpellier, France
[2] South Green Bioinformatics Platform, Bioversity-CIAT Alliance, CIRAD, INRA, IRD, Montpellier, France
[3] N2TP Technology Solutions JSC., Hanoi, Vietnam

Corresponding author: thi-minh-nguyet.dang@ird.fr

To compare multiple genomes, a linear reference genome was often used as a coordination system to describe genes, variations and other functional annotations across individuals. However, this single reference was shown not to be sufficient to grasp every existing genomic variation such as copy number variations (CNV), presence/absence variations (PAV) or more general structural variations (SV) [1]. To overcome this limitation, the concept pangenome composing a core-genome and a dispensable genome was applied to investigate a group of genomes[2]. Graph-based data model generated by incrementally incorporating genome-to-graph alignment information was one of the novel approaches to represent pangenome information [3]. Here, we propose a Julia package to extract the longest representative path from a pangenome graph that are usable for available tools working with linear reference genome while conserving the additional information provided by a graph.

Considering the pangenome graph as a simple directed graph $G = (V, E)$ with set of vertices V (non-repetitive DNA sequences among individuals) and set of edges E (directed linkage between two sequences). The longest representative path is defined as either the path containing the greatest number of base pairs or the path containing the most common vertices among individuals. In both cases, we have to find the longest path from one vertex in set $\mathcal{S}_1$ - the set of all source vertices to one vertex in set $\mathcal{S}_2$ - the set of all sink vertices knowing that all vertices in the graph having a weight value equal either the length of the DNA sequence or the number of individuals having this DNA sequence, correspondingly. We construct a Linear Programming model to select the path having maximal weight value satisfying these following constraints: (1) the edges must be existed in the edges of graph G, (2) only one edge is outgoing from the set of source vertices, (3) only one edge is incoming to the set of sink vertices, (4) there is at most one incoming edge for all vertex, (5) if there is an incoming edge, there would be an outgoing edge for all vertex. These ideas were implemented in the package BioGraph, which is able to access at https://github.com/nguyetdang/BioGraph.jl

We used minigraph to generate a pangenome graph of 12 near-gap-free reference genomes sequences 12 subpopulations of cultivated Asian rice [4] and the common reference *Oryza sativa Nipponbare* acting as the first genome. The longest representative path were extracted from the graph using BioGraph. In both cases, the size of the longest path is approximately 100 MBs more than the size of *Oryza sativa Nipponbare* reference. Based on the rank of the graph, this Julia package also separates sub-graphs having other cultivated Asian rice as the first genome which can be used later to access information that are not available in the reference. In this case, all of these computations were optimized to run parallel in less than 10 minutes.

References

[1] Xiaofei Yang, Wan-Ping Lee, Kai Ye, and Charles Lee. One reference genome is not enough. *Genome Biology*, 20(1):104, 2019.

[2] Christine Tranchant-Dubreuil, Mathieu Rouard, and Francois Sabot. Plant Pangenome: Impacts On Phenotypes And Evolution. *Annual Plant Reviews*, May 2019.

[3] Heng Li, Xiaowen Feng, and Chong Chu. The design and construction of reference pangenome graphs with minigraph. *Genome Biology*, 21(1):265, 2020.

[4] Yong Zhou, Dmytro Chebotarov, Dave Kudrna, Victor Llaca, Seunghee Lee, Shanmugam Rajasekar, Nahed Mohammed, Noor Al-Bader, Chandler Sobel-Sorenson, Praveena Parakkal, Lady Johanna Arbelaez, Natalia Franco, Nickolai Alexandrov, N. Ruaraidh Sackville Hamilton, Hei Leung, Ramil Mauleon, Mathias Lorieux, Andrea Zuccolo, Kenneth McNally, Jianwei Zhang, and Rod A. Wing. A platinum standard pan-genome resource that represents the population structure of asian rice. *Scientific Data*, 7(1):113, 2020.

Multi-modal omics data integration in Galaxy

Etienne CAMENEN[1] and Francois-Xavier LEJEUNE[2]

[1] Inserm U 1127, CNRS UMR 7225, Sorbonne Universités, UPMC Univ Paris 06 UMR S 1127, Institut du Cerveau, ICM, F-75013, Paris, France

Corresponding Author: etienne.camenen@icm-institute.org

Biomedical data are very heterogeneous (*i.e.*, different kinds of omics data) and high dimensional (*i.e.*, a large number of variables compared to the number of individuals). Data complexity dealt with different structure: for example, the same individuals but different groups of variables or modalities. This structure is called "multimodal". Its analysis is possible for clinicians thanks to RGCCA [1, 2, 3] (a generalization of canonical correlation analysis). The RGCCA statistical framework includes a set of multivariate analysis methods for a single (PCA; principal component analysis), two (*e.g.*, PLS; partial least squares analysis), or multiple data sets (*e.g.*, RGCCA). The algorithm will maximize the correlation between each modality to understand the relationship between individuals. Similar to PCA, RGCCA will reduce this set of matrices into a consensus space of a small number of synthetic variables also called "the components of the analysis".

Galaxy can be defined as a public platform, a database of omics data analysis tools that a user can connect to each other through custom pipelines [4]. We have developed a simplified, user-friendly interface for the RGCCA available on the Intergalactic Utilities Commission's European Galaxy server (https://usegalaxy.eu/). After uploading the quantitative data matrix, the analysis can then be directly launched and automatically visualized through several synthetic graphical outputs. For users more familiar with the RGCCA, an "advanced mode" allows them to adjust the analysis parameters and those specific to visualization. With this tool, the clinician can identify the variables in each modality that are most correlated to the first two components (or axes) of the analysis on a "corcircle" (*e.g.*, transcriptomic, metabolomic, etc.) [5]. These sets of potential biomarkers can potentially be associated with a clinical response or groups of patients in the same representation space. These main outputs of the software (*e.g.*, metabolomic or transcriptomic fingerprint) can then be reused by other tools in *ad hoc* workflows (*e.g.*, pathway or gene set enrichment).

Our team is currently working on the functionalities of the next R package: prediction/cross-validation, taking into account the temporal dimension, missing values, etc. A shiny version is being distributed to our partners (https://github.com/rgcca-factory/RGCCA/tree/3.0.0/inst/shiny). These interfaces will allow clinicians to explore their data with a certain number of methods that will be directly accessible to them.

References

[1] Arthur Tenenhaus and Michel Tenenhaus. Regularized generalized canonical correlation analysis. Psychometrika, 76:257-284, 2011.

[2] Michel Tenenhaus, Arthur Tenenhaus and Patrick JF Groenen. Regularized generalized canonical correlation analysis: A framework for sequential multiblock component methods. Psychometrika, 82(3):737–777, 2017

[3] Arthur Tenenhaus and Vincent Guillemot. RGCCA Package, 2017. http://cran.project.org/web/packages/RGCCA/index.html

[4] Afgan, Enis, Dannon Baker, Marius Van den Beek, Daniel Blankenberg, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, and Carl Eberhard. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic acids research, 44(W1), W3–W10, 2016.

[5] Imene Garali, Isaac M. Adanyeguh, Farid Ichou, Vincent Perlbarg, … and Arthur Tenenhaus, A. A strategy for multimodal data integration: application to biomarkers identification in spinocerebellar ataxia. Briefings in bioinformatics, 19(6), 1356-1369, 2017.

Comparison of high-dimension mediation methods to estimate impact of environmental effects on phenotype via DNA methylation

Basile Jumentier[1], Claire-Cécile Barrot[1], Johanna Lepeule[2] and Olivier François[1]

[1] TIMC-IMAG UMR5525, Université Grenoble-Alpes, 38000, Grenoble, France
[2] IAB-SLAMA U1209, Université Grenoble-Alpes, 38000, Grenoble, France

Corresponding Author: basile.jumentier@univ-grenoble-alpes.fr

Mediation analysis is a statistical tool for understanding the relationships between two variables, through the inclusion of mediation variables on the causality path. Since its development in the 1980s, it has been widely used in medical and biological research. The emergence of high throughput sequencing methods has led to the use of mediation models in epigenetic studies, for example, to estimate the causal role of DNA methylation in health outcomes linked to environmental exposures. Several high-dimension mediation methods have been developed to estimate indirect factors on given phenotypes; however, to date, there is no method that makes consensus. We searched to estimate the impact of environmental effect on a specific phenotype via DNA methylation.

To estimate the impact of environmental effect on a phenotype via DNA methylation (DNAm), we compared the speed and accuracy of five high-dimensional mediation methods found in literature: HIMA[1], Tobi[2], HDMT[3], SBMH[4] and ScreenMin[5], as well as one developed in our group, Max2. Mediation methods use regressions models, the most common are based on the methods for estimation cell composition: RefFreeEWAS[6] or Refactor[7]. We have compared their accuracy to the latent factor mixed model (LFMM)[8]. All comparisons were made by simulating CpG methylation data in a cohort of n = 500 patients, with 3 confounding factors and 6 cell types. For each method, we ran 12 simulations, varying 3 parameters: the environmental effect on the methylation level (0.2 or 0.4), the effect of the methylation level on the phenotype (0.2 or 0.4), and the number of mediators (8, 16, or 32).

From the six mediation methods tested, Max2 showed the shortest time of execution. HDMT and Max2 had the best F1-scores regardless of the strengths of environmental effects on DNAm levels and of the effects of DNAm levels on phenotype. Tobi's method also reached a good F1-score when the environmental effect on methylation is low and the effect of methylation on phenotype is strong. In addition, the methods with the higher F1-scores increased their F-score with more mediators. Regarding regression models, LFMM was the best performing method in terms of F1-scores and allows to better estimate confusion factors; the Refactor method obtained the lowest F1-score. In conclusion, our comparisons of 6 high-dimensional mediation methods and 3 regression models showed that LFMM regression coupled with the Max2 for mediation obtained the best results in terms of accuracy and speed.

References

[1] Zhang H et al. Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*. 32(20):3150-3154, 2016.
[2] Tobi EW. DNA methylation as a mediator of the association between prenatal adversity and risk factors for metabolic disease in adulthood. *Science Advances*, 2018.
[3] Dai JY et al. A Multiple-Testing Procedure for High-Dimensional Mediation Hypotheses. *Journal of the American Statistical Association*, 2020.
[4] Sampson JN et al. FWER and FDR control when testing multiple mediators. *Bioinformatics*, 2018.
[5] Djordjilović V et al. Global test for high-dimensional mediation: Testing groups of potential mediators. *Statistics in medicine*, 2019.
[6] Houseman EA et al. Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC Bioinformatics*, 2016.
[7] Rahmani E et al. Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nature Methods*, 2016.
[8] Caye et al. LFMM 2: Fast and Accurate Inference of Gene-Environment Associations in Genome-Wide Studies. *Molecular biology and evolution*, 2019.

Solving conflicts when gene and genome evolution disagree in paleopolyploid plants

Lada Isakova[1], Alexandra Louis[1], Elise Parey[1] and Hugues Roest Crollius[1]

[1] Institut de Biologie de l'École Normale Supérieure (IBENS), École Normale Supérieure, CNRS, INSERM, Université PSL, 46 Rue d'Ulm, 75005, Paris, France

Corresponding Author: lada.isakova@ens.psl.eu

1. Introduction

Whole-genome duplications (WGDs) are evolutionary events that result in the doubling of a genome. Gene deletions then gradually return the gene content to a diploid state for most of the genome, leading to complex gene families where orthologs and paralogs are difficult to identify reliably. The SCORPiOs pipeline was developed to correct phylogenetic gene trees at duplication nodes corresponding to a WGD event using local syntenic context [1]. However, this approach was only tested on vertebrates to correct nodes corresponding to the Teleost fish WGD. While WGDs are rare in vertebrates, they are especially common in angiosperms, among which many lineages exhibit multiple rounds of polyploidization [2]. Here we apply the SCORPiOs approach to correct for a specific WGD event in plant genome evolution which took place ~55 Ma years ago in Papilionoideae, the legume subfamily associated with nitrogen-fixing bacteria.

2. Methods

Data preparation. Gene phylogenies, alignments, and coordinates for 96 plant species were downloaded from Ensembl Plants v.49. Genome continuity metrics such as scaffold number, N50, and L70 statistics were calculated for all genomes and used for the selection of the most appropriate WGD. The Papilionoideae WGD event was chosen for analysis. 7 Papilionoideae species constituted the duplicated group, and the *Prunus persica* and *Cannabis sativa* genomes were used as outgroups. Gene trees were filtered to select for duplicated and outgroup species genes and reconciled with TreeBeST to label the duplication nodes. This set represents a total of 24 524 gene trees, 10 879 of which can be corrected (>2 genes, at least 1 outgroup gene).

Correction of gene trees. SCORPiOs was run on the dataset with default parameters. Sliding window size of 15, 20, 25 neighboring genes for determination of synteny similarity was tested as well as different outgroup(s): *Prunus persica* only, *Prunus persica* + *Cannabis sativa*.

Evolutionary categorization of duplicated families. Papilionoideae gene families (defined as subtree(s) in a given gene tree with the root node in Papilionoideae) were classified into three categories with respect to their fate after the duplication: genes retained on two branches after the WGD (in two or more copies) across all descendant species ("systematic ohnologs"), genes found on a single branch in all species ("singletons"), and genes retained on two branches in at least one species but not in all ("facultative ohnologs").

3. Results

Trees correction by SCORPiOs. Utilization of the sliding window of 20 genes and 2 outgroup species proved to correct the largest number of gene trees (1130 trees). With these parameters among 14 048 families tested, SCORPiOs identified 3371 (24%) synteny-inconsistent families and was able to correct 1322 (39%).

We then analyzed the set of 10 879 gene trees (corrected and uncorrected) and assigned evolutionary categories to 14 798 families. We classified 66% of families as singletons, 23% as facultative, and 11% as systematic ohnologs. After correction we found a considerable change in the counts of different categories - the number of families classified as systematic and facultative ohnologs increased by 32% and 89% respectively and the number of singleton families decreased by 32% compared to the counts in original trees.

Conclusion. The application of SCORPiOs on the Papilionoideae WGD resulted in the correction of 1130 gene trees and an increase in ohnologs counts, consistent with the results reported for vertebrates [1].

References

1. Parey E, Louis A, Cabau C, Guiguen Y, Roest Crollius H, Berthelot C. Synteny-guided resolution of gene trees clarifies the functional impact of whole genome duplications. Mol. Biol. Evol., (37/11):3324–3337, 2020.
2. Ren R., Wang H., Guo C., Zhang N., Zeng L., Chen Y., Ma H., and Qi J. Widespread Whole Genome Duplications Contribute to Genome Complexity and Species Diversity in Angiosperms. Mol. Plant. (11):414–428, 2018.

Towards a novel framework for large scale RNAseq data analysis in human health

Chloé Bessiere [1], Benoit Guibert [1], Camille Bernadas [1], Sébastien Riquier [1], Florence Ruffle [1], Anthony Boureux[1], Daniel Gautheret [2] and Thérèse Commes [1]

[1] IRMB, University of Montpellier, INSERM U1183, 80 rue Augustin Fliche, 34295 Montpellier, France;
[2] Institute for Integrative Biology of the Cell, CEA, CNRS, Université Paris Saclay, Gif sur Yvette, France.

Corresponding Author: chloe.bessiere@inserm.fr

With its ability to reveal both altered gene expression levels and the production of aberrant transcripts, RNAseq is popular in the field of precision medicine. An increasing number of clinical trials uses this technology in order to discover functionally relevant alterations. Driven by myriads of projects, public RNAseq databases are exploding, to date, there is over 164,000 RNA-seq on SRA for human. This huge body of publicly available RNAseq libraries is a precious resource to identify specific transcriptional events. However, the challenges lie in the complexity of RNA biological content and the exponential increase in data volume. We want to make RNAseq data easily accessible, providing a capture of the whole transcriptome complexity, in the context of biological and human health applications. Therefore, we developed a new framework based on a k-mer approach, constructed with several modules: 1/ a new RNAseq indexing structure that will serve as an efficient platform to request any transcribed information, 2/ a complete module to generate unique k-mers as signature of transcripts, 3/ a supporting web site to facilitate the queries for the biologists.
The indexing step uses Reindeer, a new k-mer based indexation structure. To our knowledge, it's the first method capable of performing fast mapping-free quantification of variant transcripts in thousands of RNAseq libraries [1]. The methodology is already efficiently implemented for several biological applications based on public datasets (from ten to thousand of RNAseq corresponding to 100Go to 10To of raw data). The k-mer designing module uses Kmerator, a tool developed to extract specific k-mers (*https://github.com/Transipedia/kmerator*) [2]. Finally, the web application is already available to facilitate large RNAseq datasets queries by the biologists with their sequences of interest as input (fasta format). Concerning biological and medical applications, we already requested and identified in selected public datasets, genes co-expressions, tissue specific biomarkers, as well as tumor specific signatures comparing normal and tumoral samples. In perspectives, advanced Machine Learning approaches could be tested and combined with our k-mer based framework, in order to select the best signatures and to improve diagnosis and prognosis models in human health.

Acknowledgements

This work was supported by the Agence Nationale de la recherche for the project ”Transipedia” [ANR-10-INBS-09]; the Canceropole Grand-Sud-Ouest ”Trans-kmer” project [2017-EM24]; and the Region Occitanie for the project ”SuriCare” [R19073FF].

References

[1] Marchet C et al. REINDEER: efficient indexing of k-mer presence and abundance in sequencing datasets. Bioinformatics (2020)
[2] Riquier S*, Bessiere C*, Guibert B, Bouge AL, Boureux A, Ruffle F, Audoux J, Gilbert N, Xue H, Gautheret D, Commes T. Kmerator Suite: design of specific k-mer signatures and automatic metadata discovery in large RNA-Seq datasets. Nucleic Acids Research (2021) (in revision).

A workflow based on k-mers to select specific biomarker candidates in a tumor microenvironnement context

Cédric RIEDEL[1,2,3], Raíssa SILVA[1,3], Benoit GUIBERT[1], Anthony BOUREUX[1], Florence RUFFLE[1], Chloé BESSIERE[1] and Thérèse COMMES[1]

[1] IRMB, University of Montpellier, INSERM U1183, 80 rue Augustin Fliche, 34295 Montpellier, France
[2] Faculty of Medicine, University of Montpellier, Montpellier, France
3. Equal authors

Corresponding Author: therese.commes@inserm.fr

Publicly available human RNA-sequencing (RNAseq) datasets are precious resources for biomedical research. Indeed, RNA-seq is widely used to identify actively transcribed genes or any transcribed alterations, and quantify gene or transcript expression. RNA-seq analysis also substantially contributes to our understanding of the processes involved in human disease. Then, the need for tools enabling fast and specific quantification of candidate sequences in large RNA-seq datasets is more and more required. Lately, approaches relying on k-mers from raw sequencing files have emerged and are used for the query of transcriptomic data. We therefore developed tools based on a k-mer approach (Riquier et al, 2021; *https://github.com/Transipedia/kmerator*) that propose a new way to explore RNAseq data and can be used for fast and in-depth exploration of transcriptomes.

In this context, we focus on tumor microenvironment measures in cancer patient cohorts. Indeed, many studies characterizing the tumor microenvironment have been published as well as prognosis associations of tumor immune and stromal response measures. However, the proposed gene signatures are generally very large and shared by different cancer types. Collection of signatures adapted to a specific tumor are often lacking, particularly in hematopoietic malignancies. Moreover, the quantification of relevant microenvironnement markers can give a score of the tumor contamination that is a crucial parameter for further analysis on primary cancer biopsies.

We then propose a workflow to select relevant candidate signatures specific for one cancer type. In a first step, the tumor microenvironment known genes/transcripts were submitted to Kmerator tool [1] to design their specific k-mers, and the RNA-seq cancer patient cohorts are indexed with the Reindeer software [2] which enables an ultra fast k-mer counting. Then, the workflow we propose identifies biomarker candidates in the dataset with: 1) a feature selection step where we both remove the non-expressed genes, with several possible selection criteria, and the overlaping genes present in the tumor itself by using appropriate cancerous cell lines, to retain only the non-tumor candidates of the microenvironnement, and 2) a more refined genes selection, by showing the effect of several chosen selection approaches, on the patients clustering. Several published microenvironment gene signatures could be easily quantified and compared. Finally, our objective is to support the framework with Machine Learning process in order to propose biomarker candidates for specific cancer predictions and to extend the workflow to other applications such as tissue specific biomarkers selection.

References

[1] Riquier S*, Bessiere C*, Guibert B, Bouge AL, Boureux A, Ruffle F, Audoux J, Gilbert N, Xue H, Gautheret D, Commes T. Kmerator Suite: design of specific k-mer signatures and automatic metadata discovery in large RNA-Seq datasets. Nucleic Acids Research (2021) (in revision).

[2] Marchet C et al. REINDEER: efficient indexing of k-mer presence and abundance in sequencing datasets. Bioinformatics (2020)

RNA-Seq analysis of temporal artery biopsies to identify novel pathogenic pathways based on inflammatory infiltration patterns in Giant Cell Arteritis

Michal ZULCINSKI[1,2], Ann MORGAN[1] and Mark ILES[2]

[1] Molecular and Personalised Medicine Group, University of Leeds, LS2 9JT, Leeds, UK
[2] Leeds Institute for Data Analytics, University of Leeds, LS2 9JT, Leeds, UK

Corresponding Author: M.Zulcinski@leeds.ac.uk

Introduction. Giant cell arteritis (GCA) is the most common form of vasculitis in people over 50 years old and can lead to serious complications, such as permanent visual loss, if undiagnosed in a timely manner [1]. Appropriate therapy can make GCA controllable and prevent complications. However, current treatment options are very limited and comprise mostly the use of glucocorticoids. Better understanding of the molecular and genetic mechanisms in GCA is still needed in order to discover novel pathogenic pathways that could be therapeutically targeted [2]. This study aims at using RNA-Seq dataset generated from GCA positive biopsies to identify candidate pathways, by exploring transcripts associations with inflammatory infiltration patterns previously described and validated in GCA patients [3, 4], and by performing pathway analysis for lists of statistically significant transcripts.

Materials and Methods. The cohort used in the study comprises forty-one patients (25 women & 16 men) selected from the UK GCA Consortium. All the subjects underwent temporal artery biopsy (TAB) resulting in positive GCA diagnosis, and were subsequently admitted for treatment. Various clinical variables, such as, symptoms, comorbidities, drugs taken during treatment, etc., were collected for each patient. A serial section of the biopsies was conducted, followed by scoring for twenty-one features with a research interest in GCA. The remaining biopsy slides were used for RNA extraction, and then sequencing using the NextSeq500 Illumina System. All statistical testing was performed using the non-parametric Mann-Whitney-Wilcoxon test and False Discovery Rate was used for multiple testing correction.

Results. A series of downstream analyses was performed using the gene expression dataset along with clinical and histological variables. These analyses aimed in particular at assessing the influence of confounding factors, such as sex, age and the duration of steroid treatment, on gene expressions profiles and examining the associations of transcripts levels with histological and clinical phenotypes. No clear confounding influence of patients' age or steroid treatment duration was found, however, from the clinical perspective, such effects were felt to be very likely to occur, therefore this aspect will be further investigated in the extended dataset. Some confounding effects of gender were observed and they were found to be secondary to inclusion of the sex chromosomes. Statistical testing for associations with histological features revealed statistically significant lists of transcripts (i.e. after multiple testing correction) for the following variables: "Giant cells presence", "Inflammation in intima", "Inflammation in adventitia" and "Inflammation in media". Preliminarily functional enrichment analysis, using the lists of statistically significant results for different features, was performed and these results will to be soon evaluated and further explored.

Future work. This work is an ongoing study and makes one part of the PhD project of the corresponding author. The current cohort of 41 samples will be extended to include 96 supplementary samples which will significantly increase statistical power of the study. Additional methods for pathways analysis and evaluation of its results will also be implemented. Moreover, within the other part of his PhD project, genome-wide association study is being currently undertaken to explore the associations in GCA from genetic perspective.

References

1. Borchers, A. T., & Gershwin, M. E. (2012). Giant cell arteritis: a review of classification, pathophysiology, geoepidemiology and treatment. Autoimmunity reviews, 11(6-7), A544–A554.
2. Samson, M., Corbera-Bellalta, M., Audia, S., Planas-Rigol, E., Martin, L., Cid, M. C., & Bonnotte, B. (2017). Recent advances in our understanding of giant cell arteritis pathogenesis. Autoimmunity reviews, 16(8), 833–844.
3. Hernández-Rodríguez, José et al. "Description and Validation of Histological Patterns and Proposal of a Dynamic Model of Inflammatory Infiltration in Giant-cell Arteritis." Medicine vol. 95,8 (2016): e2368.
4. Nekane Terrades-Garcia, Maria C Cid, Pathogenesis of giant-cell arteritis: how targeted therapies are influencing our understanding of the mechanisms involved, Rheumatology, Volume 57, Issue suppl_2, February 2018.

Accurate transposable element detection and allele frequency estimate using long-read sequencing data combining assembly and mapping-based approaches

Anna-Sophie Fiston-Lavier*[1], François Sabot*[2,3] Mourdas Mohamed[4], Marion Varoqui[1,5], Bruno Mugat[4], Alain Pelission[4] and Séverine Chambeyron[4]
[1] ISEM, Université Montpellier, CNRS, IRD, CIRAD, EPHE, Montpellier, France
[2] DIADE, UM, CIRAD, IRD, Montpellier, France
[3] IFB-Southgreen Biodiversity, CIRAD, INRAE, IRD, Montpellier, France
[4] IGH, Université Montpellier, Montpellier, France
[5] Master Sciences et Numérique pour la Santé, Parcours Bioinformatique, Connaissances, Données, Montpellier, France
(*) co-first authors

Corresponding authors:
anna-sophie.fiston-lavier@umontpellier.fr;
francois.sabot@ird.fr;
severine.chambeyron@igh.cnrs.fr

Abstract

Transposable Elements (TEs) are genetic elements that are able to multiply within their host genome. They are ubiquitous and can represent over 90% of a genome [1]. Due to their repeated nature and length (from 100bp to more than 10kb), the use of short-reads does not provide efficient TE calls to study the TE insertion genomic localization and then the TE dynamics [2]. Long-read sequencing technologies offer the unprecedented opportunity to detect with precision the position and estimate with accuracy the allele frequency of newly integrated TEs in long-read data. Few computational tools have been developed to detect TE insertions in long-read sequencing data. They all use mapping-based approaches. After the mapping of the long reads on reference genome sequences, they launch variant calling tools such as Sniffles. However, none of these approaches return accurate TE calls with low quality reference genome sequences. As the assembly process is still a challenge, we present here *Transposable Element MOnitoring with LOngReads (TrEMOLO)*, a computational tool that uses long-read sequencing to detect recent TE insertions combining assembly and mapping-based approaches. TreMOLO could detected most of the TE insertions using high or low quality reference genome and for TE insertions at high or low frequency. We assessed the veracity of TreMOLO by performing a comparative study with the other computational tools available. The TE insertions detected and the frequency estimates were experimentally validated. We also analyzed the TreMOLO results by testing several datasets with variable quality and from diverse taxons (drosophila, plants and human). Taking together, TreMOLO appears as a good computational tool to study with accuracy the dynamics of TEs (*cf.* poster TreMOLOdyn #128).

References

[1] Benoît Chénais, Aurore Caruso, Sophie Hiard, and Nathalie Casse. The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. *Gene*, 509(1):7–15, November 2012.

[2] Mourdas Mohamed, Nguyet Thi-Minh Dang, Yuki Ogyama, Nelly Burlet, Bruno Mugat, Matthieu Boulesteix, Vincent Mérel, Philippe Veber, Judit Salces-Ortiz, Dany Severac, Alain Pélisson, Cristina Vieira, François Sabot, Marie Fablet, and Séverine Chambeyron. A Transposon Story: From TE Content to TE Dynamic Invasion of Drosophila Genomes Using the Single-Molecule Sequencing Technology from Oxford Nanopore. *Cells*, 9(8):1776, July 2020.

Phylogeny of *Meloidogyne incognita* populations associated with *Malpighia emarginata* in the northeastern region of Brazil

Francisco J. C. SOUZA JUNIOR[1], Mayara C. ASSUNÇÃO[1] and Jaime C. SANTOS NETO[1]

[1] Nematology Laboratory, Federal Rural University of Pernambuco, Department of Agronomy, Dois Irmãos, 52.171-900, Recife, Brazil

Corresponding Author: jorgesouza@alu.ufc.br

The acerola tree is one of the main fruit trees in Brazil, with a cultivated area of approximately 10.000 hectares in this country. Among the main phytosanitary problems of the crop, the attack of *Meloidogyne* nematodes stands out as one of the most important, due to the damage caused to the roots, impairing the absorption of water and nutrient and causing losses in productivity, in addition to most varieties are susceptible. The objective of this work was to identify *Meloidogyne* species associated with acerola tree. The collections were carried out in aceroleira orchard (8º1'4''S 34º56'44''O, elevation 20 m) in the municipality of Recife, located in the Metropolitan mesoregion of the State of Pernambuco, between September 2019 to March 2020. The procedure for extracting nematodes from soil and root samples. For molecular characterization, three fragments of ribosomal DNA (rDNA) were sequenced (D2-D3 region of 28S rRNA, ITS and 18S rRNA) and two regions of mtDNA (coxI and coxII-16S). Bayesian inference (BI) was used for phylogenetic reconstruction. In total 72 isolates of *Meloidogyne* spp. Were obtained, in preliminary analysis they were separated into 2 different haplotypes (H1-H2), 60 were represented by the haplotype (H1) and the sequences were grouped in *Meloidogyne* sp. 1 CN0007. Twelve isolates represented by the haplotype (H2) were separated for *Meloidogyne* sp. 2 CN0008. A total of 2 isolates representative of the haplotypes were chosen for sequencing the remaining locus and subsequent analyzes. *Meloidogyne* isolates of *M. emarginata* were identified in a single species, according to the GCPSR criterion. The two isolates CN0007 and CN0008 were grouped with the *M. incognita* clade with maximum support in multilocus BI analysis. The *Meloidogyne incognita* species was identified associated with the aceroleira culture in an orchard located in the city of Recife in the state of Pernambuco, Brazil. The presence of this species of nematode in aceroleira production areas confirms the damage caused by the parasitic action of this etiological agent, requiring the adoption of management measures to keep *Meloidogyne* populations below the level of economic damage.

References

1. Magno B. Silva, Jairton F. Araujo, Elaine R. Galvão, and Fabiana P. R. Batista. Produção e qualidade de acerola com biofertilizantes líquidos sob cultivo biodinâmico. *Revista Ouricuri*, (2/2):125-137, 2019.
2. Maria C. L. Silva, Carmem D. G. Santos, and Gilson S. Silva. Espécies de *Meloidogyne* associadas a vegetais em microrregiões do estado do Ceará. *Revista Ciência Agronômica*, (47/4):710-719, 2016.

A phylogenetic approach for functional module characterisation of the ADAMTS / ADAMTS-like protein family

Olivier Dennler[1,2], Samuel Blanquart[2], François Coste[2], Catherine Belleannée[2] and Nathalie Théret[1,2]

[1] Univ Rennes, Inserm, EHESP, Irset - UMR_S1085, F-35043 Rennes
[2] Univ Rennes, Inria, CNRS, IRISA, UMR 6074, Rennes, France

Corresponding author: olivier.dennler@inria.fr, nathalie.theret@univ-rennes1.fr

1 Introduction

ADAMTS and ADAMTS-like proteins are involved in microenvironment remodelling and are now considered as new potential therapeutic targets in numerous diseases. However the characterisation of these proteins is still under progress and it is necessary to develop new approaches for their functional annotation [1]. The high number of genes, the multi-domain structures of the proteins and the lack of experimental data challenge the conventional approaches used for protein functional characterization. We propose here a new method adapted to multi-paralog and multi-domain protein sequences based on module decomposition and phylogenetic inferences, in order to identify functional protein regions. Our purpose is to identify conserved modules (*i.e.* a set of ungapped conserved sequence regions, shared by at least 2 sequences) and to characterise the function (*i.e.* protein-protein interaction) associated to these conserved modules, while dealing with phylogenetic uncertainty and sparse knowledge as best as possible.

2 Methods

Our method is based on two of the most widely used methods for functional prediction, namely the sequence conservation and the phylogenomics inference. The principle of our method is to segment without *a priori* protein sequences in modules, then to infer protein-protein interactions (PPI) in parallel of conserved module phylogeny, and to integrate both on a common gene phylogeny. This method combines and integrates multiple phylogenetic inference strategies ; 1) gene phylogeny inference, 2) ancestral modules composition inference using Domain-Gene-Species (DGS) reconciliation [2] and 3) inference of function (using ancestral scenario reconstruction [3]). The final pipeline allows us to identify changes in module composition occurring at the same time than protein-protein interactions. We hypothesize that, during evolution, the modules gained at the same time than a protein-protein interaction might be implicated in this interaction.

3 Results and Perspectives

This new strategy is implemented as a pipeline allowing to correlate the conserved sequence modules and the functions evolution in order to identify the co-appearance of conserved modules and functions. Applying this method to ADAMTS-TSL proteins permits us to retrieve a region of the protein previously known to be involved in COMP-ADAMTS7 [4] protein-protein interaction (PPI) but also new modules that co-occur with the gain of this PPI but were not previously known to be involved in this function.

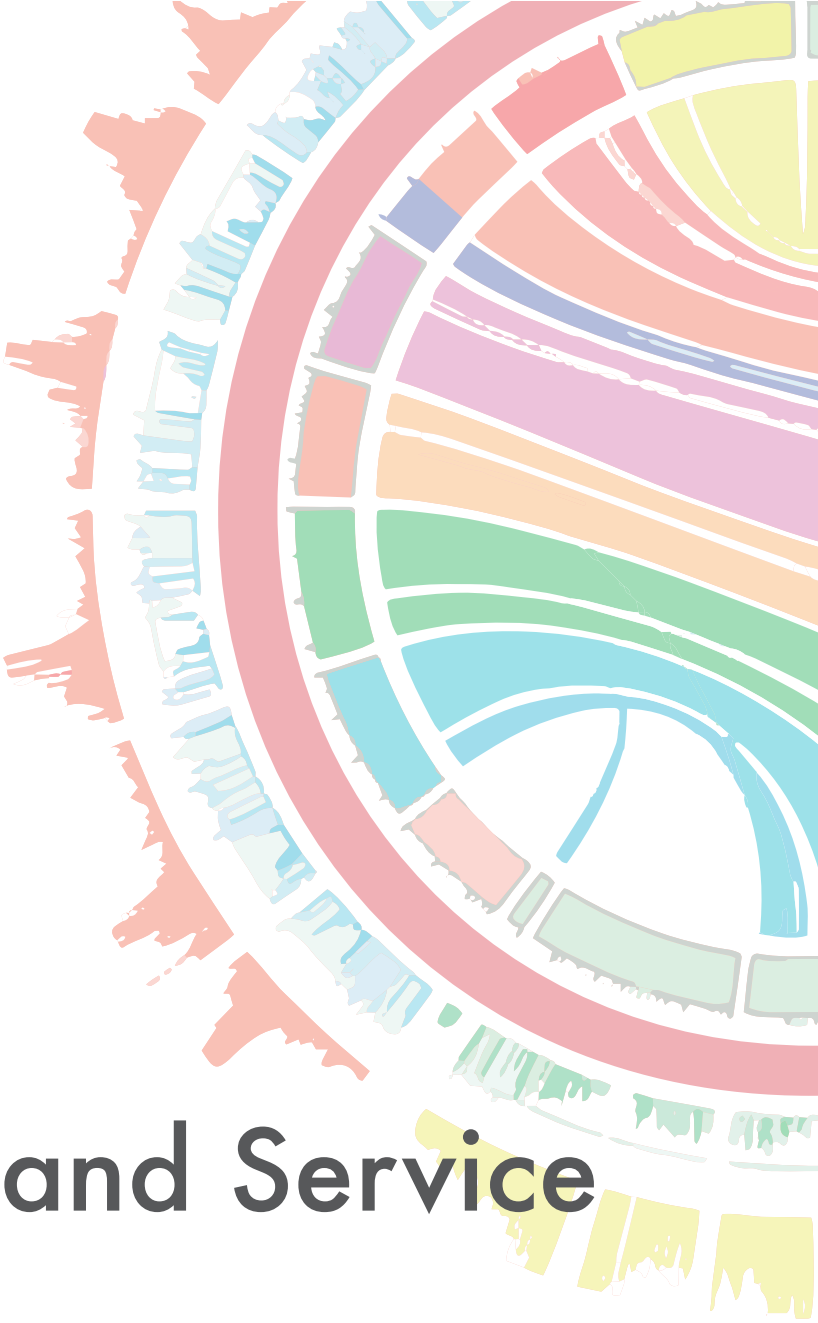
References

[1] Suneel S. Apte. ADAMTS Proteins: Concepts, Challenges, and Prospects. *Methods in Molecular Biology (Clifton, N.J.)*, 2043:1–12, 2020.

[2] Lei Li and Mukul Bansal. Simultaneous Multi-Domain-Multi-Gene Reconciliation Under the Domain-Gene-Species Reconciliation Model. pages 73–86. January 2019.

[3] Sohta A Ishikawa, Anna Zhukova, Wataru Iwasaki, and Olivier Gascuel. A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios. *Molecular Biology and Evolution*, 36(9):2069–2085, September 2019.

[4] Chuan-Ju Liu, Wei Kong, Kiril Ilalov, Shuang Yu, Ke Xu, Lisa Prazak, Marc Fajardo, Bantoo Sehgal, and Paul E. Di Cesare. ADAMTS-7: a metalloproteinase that directly binds to and degrades cartilage oligomeric matrix protein. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology*, 20(7):988–990, May 2006.



> Plateform and Service Activities

Multi-omics interactive tool set for differential expression, enrichment analysis and visualization

Miriam Riquelme-Perez[1,2], Fernando Perez-Sanz[3], Jean-François Deleuze[2], Carole Escartin[1], Eric Bonnet[2] and Solene Brohard[2]

[1] Université Paris-Saclay, CEA, CNRS, MIRCen, Laboratoire des Maladies Neurodégénératives, 92265, Fontenay-aux-Roses, France.
[2] Centre National de Recherche en Génomique Humaine (CNRGH), Institut de Biologie François Jacob, CEA, Université Paris-Saclay, Evry, France.
[3] Biomedical Informatics & Bioinformatics Service, Institute for Biomedical Research of Murcia (IMIB), 30120 Murcia, Spain

Corresponding author: miriam.riquelme-perez@cea.fr

1 Introduction

We are at a time of considerable growth in the use and development of *in silico* researches. A typical example of this could be differential expression analysis, that generates gene lists from large datasets and complex designs which require increasingly higher computing capacities and expertise. However, trained or qualified personnel is not always available in experimental laboratories, to carry out this processing from raw data to meta-analysis. In this way, biologists involved in the experiments may lose versatility and control of their results because they are unable to explore them to their fullest extent easily on their own.

Despite the undeniable development of software applications over the years, there is still room for improvement; in particular, to make these tools more accessible to non-experts in the field of bioinformatics. This implies also a commitment of appropriate maintenance and extensive customization as well as clear export options of the results, still lacking in this area.

Therefore, we have developed an intuitive and user-friendly web application in shiny R, which generates a complete set of figures and tables, accessible and personalizable for users who are not bioinformatics specialists. It is based on a wrapper of several databases temporally updated (such as Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Set Enrichment Analysis (GSEA) computational method gene sets) as well as various functions to visualize results obtained with distinct statistical approaches.

Multiple inputs may be used to run the application covering a range of data types. Depending on the complexity of the object entered, the app is able to extract different elements and produce several results. Thus, from a simple **'Gene List'** it can obtain the gene names to perform the enrichment analysis. If statistical values are added to the list as extra columns, the application may extend the analysis to many statistical plots that imply knowing the p-value or fold-change of each gene.

By adding a raw **'Expression Matrix'** file and a **'ColData'** table with the information corresponding to the samples (such as the group to which they belong, sex, age...), the application calculates the genes differentially expressed and the statistics with the design according to the column of choice and extract preliminary plots and tables related to these values. It will also perform enrichment analysis results. Finally, if a **'DESeq2 object'** is entered, which already contains the differential expression information, all figures and tables mentioned before are available, without any calculation from the app.

With this web application, we expect to offer in an intuitive and simple way for unskilled users, a complete set of figures, tables and results related to differential expression and enrichment analysis across different databases and functions. It aims to incorporate into a single software complex analyses that would require substantial processing in diverse tools to not experienced bioinformatics users.

The set of applications is accessible at http://shiny.imib.es/entryApp/ and the code is freely available on github.

Acknowledgements

MRP holds a PhD fellowhsip from the CEA.

BioInformatics and Genomics platform at Institut Sophia Agrobiotech

Martine DA ROCHA[1], Arthur PÉRÉ[1], Etienne DANCHIN[1] and Corinne RANCUREL[1]

[1] Institut Sophia Agrobiotech, INRAE, Université Côte d'Azur, CNRS, 400 route des Chappes BP 167, F-06903 Sophia Antipolis Cedex, France

Corresponding Author: martine.da-rocha@inrae.fr, corinne.rancurel@inrae.fr

The BioInformatics and Genomics (BIG) platform of Institut Sophia Agrobiotech (ISA: INRAE - CNRS - Univ. Côte d'Azur) offers expertise in bioinformatics and solutions for processing, integrating, analyzing and visualizing multi-omics data in the field of plant health and protection. The BIG platform is part of PlantBios (Biocontrol and Plant Biostimulation, Facilities and Expertise), labeled as a collective scientific infrastructure by INRAE. PlantBIOs offers equipment and expertise for studies ranging from gene level to the whole agroecosystem scale with analytical tools (Imagery and Microscopy, Biochemistry and Mass Spectrometry, Bioinformatics and Genomics), experimental tools, and collections of rare biological resources.

Since the end of 2020, BIG has been an IFB contributing platform. The core of the BIG platform is composed of three bioinformatics engineers: Martine Da Rocha (from INRAE), Arthur Péré (from INRAE) and as operational manager Corinne Rancurel (from CNRS). The core is complemented by a scientific advisor: Etienne Danchin (INRAE senior scientist).

BIG has a main expertise in comparative genomics, transcriptomics and molecular evolution. More recently, BIG has been involved in epigenomics, small RNA as well as metagenomics studies. The tools and resources produced by BIG are made available to the scientific community (website, forge and integrative portals) and can address similar problems encountered in other research areas. For instance, the Alienness tool, which allows rapid detection of candidate horizontal gene transfers in genomes has been used 1019 times by 132 different users and the corresponding paper [1] (Rancurel et al. 2017) has already been cited 22 times, since its launch in August 2017.

In addition to methodological developments, the platform offers support and training to biologists in the use of bioinformatics tools and pipelines, including the one developed by BIG itself.

The BIG platform is open for collaboration and can be contacted at the following e-mail address: spiboc.big@inra.fr

This web page summarizes the activities and organization of the BIG platform: https://www6.paca.inrae.fr/institut-sophia-agrobiotech/Infrastructure-PlantBIOs/Equipements-Ressources-biologiques-et-Expertises/Plateau-de-bioinformatique or http://tinyurl.com/y9qkho4v

References

1. Rancurel C, Legrand L, Danchin EGJ. Alienness: Rapid Detection of Candidate Horizontal Gene Transfers across the Tree of Life. Genes (Basel). 2017 Sep 29;8(10):248. doi: 10.3390/genes8100248. PMID: 28961181; PMCID: PMC5664098.

MicroScope: an Integrated Platform for the Annotation and Exploration of Microbial Gene Functions through Genomic, Pangenomic and Metabolic Comparative Analysis

Alexandra Calteau[1], Mathieu Dubois[1], Jérôme Arnoux[1], Adelme Bazin[1], Mylène Beuvin[1], Stéphanie Fouteau[1], Frédéric Kuhner[1], Aurélie Lajus[1], Félix Legrelle[1], David Roche[1], Zoé Rouy[1], Mark Stam[1], Claudine Médigue[1] and David Vallenet[1]

[1] LABGeM, Génomique Métabolique, CEA, Genoscope, Institut François Jacob, CNRS, Université d'Évry, Université Paris-Saclay, Evry, France

Corresponding Author: acalteau@genoscope.cns.fr

Large-scale genome sequencing and the increasingly massive use of high-throughput approaches produce a vast amount of new information that completely transforms our understanding of thousands of microbial species. However, despite the development of powerful bioinformatics approaches, full interpretation of the content of these genomes remains a difficult task. To address this challenge, we develop the MicroScope platform, which is an integrated Web platform for management, annotation, comparative analysis and visualization of microbial genomes (https://mage.genoscope.cns.fr/microscope) [1]. The platform enables collaborative work in a rich comparative genomic context and improves community-based curation efforts.

Launched in 2005, the platform has been under continuous development within the LABGeM team at Genoscope. MicroScope provides analyses for complete and ongoing genome projects together with metabolic network reconstruction and transcriptomic experiments allowing users to improve the understanding of gene functions. Besides automatic functional annotations, we integrated several tools to analyze a wide range of biological systems (antibiotic resistance, virulence, secondary metabolites, integrons, secretions systems, CRISPR-Cas clusters…). Particularly, tools from the PPanGGOLiN software suite (https://github.com/labgem/PPanGGOLiN) allows users to analyze pangenomes from several hundreds of genomes of the same species and to explore their content in regions of genomic plasticity [2,3]. The platform also has extensive functionality to explore and compare metabolic pathways.

MicroScope platform is widely used by microbiologists from academia and industry all around the world for collaborative studies and expert annotation. To date, MicroScope contains data for >14,500 microbial genomes, part of which are manually curated and maintained by microbiologists (>5,400 user accounts in March 2021). The platform is also a useful resource for academic training.

This poster gives an overview of the platform and its evolution and presents new methodologies and tools already integrated or currently being developed.

Acknowledgements

This work was supported in part by FRANCE GENOMIQUE [ANR-10-INBS-09–08] and INSTITUT FRANÇAIS DE BIOINFORMATIQUE [ANR-11-INBS-0013].

References

1. David Vallenet, Alexandra Calteau, Mathieu Dubois *et al*. MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. *Nucleic Acids Research*, Jan 8;48(D1):D579-D589, 2020.
2. Guillaume Gautreau, Adelme Bazin, Mathieu Gachet *et al.* PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. *PLoS Computational Biology*, Mar 19;16(3):e1007732, 2020.
3. Adelme Bazin, Guillaume Gautreau, Claudine Médigue *et al.* panRGP: a pangenome-based method to predict genomic islands and explore their diversity. *Bioinformatics*, Dec 30;36(Suppl_2):i651-i658, 2020.

Biogenosis: a start-up specialized in the creation and development of innovative bioinformatic tools supporting scientific research

Lysiane Hauguel[1] and Benjamin Bourgeois[1]
Biogenosis, 17A avenue Jacques Chastellain, 76100, Rouen, French

Corresponding author: lysiane.hauguel@biogenosis-bioinfo.com

1 Biogenosis

Biogenosis is a French start-up from the region of Rouen[1], founded in January 2021 by Lysiane Hauguel and Benjamin Bourgeois. This start-up was admitted, as of 12 may 2021, to Normandie Incubation's incubator.

In the world of science and research, researchers and scientists are confronted with large amounts of data and information that they just can't manage or exploit easily alone. That's where we, bioinformaticians, intervene, helping in the process of analyzing data.

Bioinformatics being a new area the low number of bioinformatic engineer doesn't meet the ever-growing demand.

Unfortunately overworked bioinformatic engineers leads to scientists analyzing data by themselves, often using methods and softwares that are not best suited for their data at the expense of quality, reproducibility, time, money and effort.

Our goal is to provide custom-built bioinformatics, user-friendly tools that will allow biologists to work efficiently without the need to ask for bioinformatician's support.

With their cleaned, simplified and user-friendly interface, our tools are state of the art, making the access and visualisation of omic data easier than ever. Our tools also allow biologist to conduct bioinfomatic analysis routine (alignment, variant calling, various samples comparisons etc ..)

2 Biogenosis Data Analyser

Biogenosis Data Analyser is the tool that we are developing in partnership with a research team (apicomplexes comparative biology team) from the Cochin Institute directed by the PU-PH Frédéric Ariey.

Our tool aims to facilitate and help research team in their daily work. We are developing this tool with the aim of offering an optimal response to the demand of researchers and biologists. Our tool include many features.

1. Storage, management and access to data stored in a secure database.
2. A Genome Browser to visualize genomes, genes and alignements.
3. Epidemiological monitoring to follow and study the evolution of pathogens over time.
4. Other bioinformatic analysis:
   — Alignement against a reference genome
   — Extraction, study and comparison of genes and mutations between samples
   — Copy number variation (CNV) visualization along the genome
   — Blast

This tools is under development, new features will be added soon (international databases API, epidemic propagation visualization, phylogeny, metagenomic species identification data, etc.).

References

[1] Biogenosis – a startup creating bioinformatic tools for biologists, https://biogenosis-bioinfo.com/.

SynTViewJS : a dynamical viewer for the microbial genome analysis

Rachel Bellone[1], Pierre Lê-Bury[2], Catherine Dauga[1], Javier Pizarro-Cerda[2] and Pierre Lechat[3]

[1] Institut Pasteur, Arboviruses and Insect Vectors, 75015, Paris, France

[2] Institut Pasteur, Yersinia Laboratory, 75015, Paris, France

[3] Institut Pasteur, Hub de Bioinformatique et Biostatistique, 75015, Paris, France

Corresponding Author: plechat@pasteur.fr

1. Introduction

SynTView[1] is a published interactive multi-view genome browser for next-generation comparative microorganism genomics. SynTViewJS is the rewrite of the software in javascript with the addition of new features. The software is characterised by the presentation of syntenic organisations of microbial genomes and the visualisation of polymorphism data obtained from next generation sequencing.

2. Microbial genome analysis with SynTViewJS

SynTViewJS is built as a generic genome browser including sub-maps that hold information about genomic objects. After selecting genomes of interest, the users can explore them visually by genomic location, or directly go to specific genes by name. Several genomic maps can be stacked ordered by a phylogenetic tree according to biological metadata on top of each other. The creation of a SynTView website is very helpful in the analysis of a large number of strains, bringing together phylogeny, polymorphisms, larger variants such as indels, coverage, as well as functional annotations and strains meta-data. SynTViewJS is designed to visualise information about polymorphism across a large number of bacterial strains. The SNP maps allow the user to navigate through polymorphism data sets. The non javascript tool has been used in many projects such as the study of Legionella[3] bacterial strains. I will show in the poster the study of the mutational dynamics of chikungunya virus as a function of temperature with visibility filters (mutation frequency, specificity ...) with the possibility of zooming to the sequence.

SynTView has been also integrated to the Listeriomics[2] web site, a platform for visualizing and analysing every heterogeneous Listeria "omics" dataset published to date and will be integrated soon in Yersiniomics (same platform dedicated to Yersinia dataset).

The tool can be uploaded to a website and the data made accessible on a server or directly added by drag and drop. Source code is available at https://gitlab.pasteur.fr/plechat/syntviewjs.

References

1. Lechat, P, Souche E & Moszer I SynTView an interactive multi-view genome browser for next-generation comparative microorganism genomics. BMC Bioinformatics 14, 277 (2013).
2. Bécavin C, Koutero M, Tchitchek N, Cerutti F, Lechat P, Maillet N, Hoede C, Chiapello H, Gaspin C, Cossart P. Listeriomics: an Interactive Web Platform for Systems Biology of Listeria. mSystems. 2017 Mar 14;2(2)
3. David S, Rusniok C, Mentasti M, Gomez-Valero L, Harris SR, Lechat P, Lees J, Ginevra C, Glaser P, Ma L, Bouchier C, Underwood A, Jarraud S, Harrison TG, Parkhill J, Buchrieser C. Multiple major disease-associated clones of Legionella pneumophila have emerged recently and independently. Genome Res. 2016 Nov;26(11):1555-1564

Ases: Alternative Splicing Evolution Server

Diego Javier Zea[1], Hugues Richard[2] and Elodie Laine[1]
[1] Laboratoire de Biologie Computationnelle et Quantitative, 4 Place Jussieu, 75005, Paris, France
[2] Bioinformatics Unit (MF1), Robert Koch Institute, Nordufer 20, 13353, Berlin, Germany

Corresponding author: diegozea@gmail.com

Ases is a novel *web server* that allows the user to assess *alternative splicing (AS)* potential impact on protein evolution. It relies on gene annotations from Ensembl [1] as its primary source of information and works with the translated amino acid sequences of the transcripts. Ases takes as input a gene and species name and, optionally, a list of species. By default, the server considers twelve species, ranging from human to nematode. Ases processes the input data by running *ThorAxe* [2] and *PhyloSofS* [3]. ThorAxe defines the *s-exons*, groups of putative orthologous exonic regions, and creates an *evolutionary splicing graph*. PhyloSofS takes ThorAxe input and produces the *phylogenetic reconstruction*. To allow the exploration of those outputs, Ases has:

— an interactive *evolutionary splicing graph* summarizing the transcript variability observed for the query gene across the selected species. Each transcript corresponds to a path in the graph. The transcripts are divided into minimal building blocks called *s-exons*, the nodes in the graph. The *AS events* are defined by pairs of canonical and alternative subpaths in the graph. The user can move the nodes and easily visualise their conservation levels, *multiple sequence alignments*, and the involved AS events.
— two interactive *tables* giving detailed information about the graph. For example, the user can select all transcripts annotated for a given species, get access to the list of species where a given AS event is observed, or list the s-exons for a given transcript, among other options.
— an interactive *phylogenetic forest* where each tree represents the phylogeny of a transcript. The tree's root indicates the appearance of a transcript in evolution, and dead ends indicate transcript losses. When the user selects an internal node (ancestral transcript) or a leaf (observed transcript), s/he can visualise the corresponding nodes in the evolutionary splicing graph.
— an interactive representation of the *gene structure using s-exons* as units.

Ases provides a way to rapidly and easily get an overview of the *sequence variations induced by AS* at the amino acid resolution. Thus, for example, a structural biologist interested in a particular region interacting with a partner can check whether AS impacts this region and extract conserved AS signatures. Likewise, a clinician having identified some transcript isoform whose overexpression is associated with a disease can rapidly check if this isoform is found in other species and date its appearance in evolution. Hence, Ases meets a need in the scientific community that is not met by any other openly accessible web server to the best of our knowledge.

You can access the Ases web server at *http://www.lcqb.upmc.fr/Ases*

References

[1] Andrew D Yates, Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, et al. Ensembl 2020. *Nucleic acids research*, 48(D1):D682–D688, 2020.

[2] Diego Javier Zea, Sofya Laskina, Hugues Richard, Elodie Laine, and Alexis Baudin. Assessing conservation of alternative splicing with evolutionary splicing graphs. *bioRxiv*, 2020.

[3] Adel Ait-Hamlat, Diego Javier Zea, Antoine Labeeuw, Lélia Polit, Hugues Richard, and Elodie Laine. Transcripts' evolutionary history and structural dynamics give mechanistic insights into the functional diversity of the jnk family. *Journal of molecular biology*, 432(7):2121–2140, 2020.

ERA-BIO-IT: A bioinformatics platform for applied research in Plant Breeding

Marion Dupouy[1], Daniel Cabero[2], Bruno Claustres[3], Boris Demenou[4], Mila Garcia[2], Delphine Hourcade[4], Michel Romestant[3], Jean-Pierre Cohan[4], Philippe Dufour[3] and Jean-Marc Ferullo[2]

[1] ERA-BIO-IT, 31700, Mondonville, France
[2] Euralis Semences, 31700, Mondonville, France
[3] RAGT 2n, 12510, Druelle, France
[4] Arvalis, 31450, Baziège, France

Corresponding author: marion.dupouy@era-bio-it.com

1 Presentation

ERA-Bio-IT was created in 2020 by three partners, two French seeds companies (Euralis Semences[1] and RAGT 2n[2]) and a technical Institute (Arvalis[3]) wishing to share bioinformatics resources for crops breeding and cultivars evaluation.

2 Objectives

Firstly, the aim of ERA-Bio-IT is to set up basics tools (such as JBrowse[4] or Galaxy[5]) for each partners bioanalysts, oriented toward the genetic and genomic study of major crops (maize, wheat, barley...). Once the platform fully functional, more advanced genomic analyses and bioinformatics developments are expected as support for each partner. The platform will be open for multi-partnership projects, including with new private and public partners.

Finally, the ERA-Bio-IT platform aims to support the omics technology and methodology watch for its partners in order to provide a cutting-edge expertise in these fast-evolving fields.

3 Infrastructure

The computing infrastructure is managed by Portalliance Engineering[6]. It is composed of a virtualization server, hosting various virtual machines (JBrowse, Galaxy...), an High-Performance Computing (HPC) server managed with SLURM and data storage Qumulo solutions. Those computing resources are scalable to answer quickly to each partner needs

Acknowledgements

ERA-Bio-IT is financed by Euralis, RAGT and Arvalis.

References

[1] Euralis Semences - Multi-species seed producer among the leaders in Europe - www.euralis-semences.fr.

[2] RAGT, RAGT-Semences: European seed producer in corn, sunflower, sorghum, fodder, cereals, rapeseed, lawns, field seeds - www.ragt-semences.fr.

[3] ARVALIS, the french arable crops R&D institute - www.arvalisinstitutduvegetal.fr.

[4] Robert Buels, Eric Yao, Colin M. Diesh, Richard D. Hayes, Monica Munoz-Torres, Gregg Helt, David M. Goodstein, Christine G. Elsik, Suzanna E. Lewis, Lincoln Stein, and Ian H. Holmes. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biology*, 17(1):66, December 2016.

[5] Enis Afgan, Dannon Baker, Bérénice Batut, Marius van den Beek, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Björn A Grüning, Aysam Guerler, Jennifer Hillman-Jackson, Saskia Hiltemann, Vahid Jalili, Helena Rasche, Nicola Soranzo, Jeremy Goecks, James Taylor, Anton Nekrutenko, and Daniel Blankenberg. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, 46(W1):W537–W544, July 2018.

[6] Portalliance Engineering - Design office - Calculation and modeling experts - Toulouse - www.portalliance.fr.

Automated Deployment of Genome Databases for Marine Algae Using Galaxy Genome Annotation (GGA)

Loraine Brillet-Guéguen[1,2], Arthur Le Bars[2], Anthony Bretaudeau[3], J. Mark Cock[1], Susana M. Coelho[1] and Erwan Corre[2]

[1] Sorbonne Université, CNRS, Integrative Biology of Marine Models (LBI2M), Station Biologique de Roscoff (SBR), 29680, Roscoff, France

[2] CNRS, Sorbonne Université, FR2424, ABiMS, Station Biologique, 29680, Roscoff, France

[3] INRA, UMR IGEPP, BIPAA/GenOuest, Campus Beaulieu, 35000, Rennes, France

Corresponding Author: loraine.gueguen@sb-roscoff.fr

As part of the French and European projects PhaeoExplorer, IDEALG and SEXSEA, 53 genomes and transcriptomes of marine brown algae have been sequenced. To scientifically exploit these genomic resources, the community needs effective tools to analyze and value the data.

Based on the GGA project (https://galaxy-genome-annotation.github.io) we have deployed an integrated environment dedicated to the management of genomic data (genomes and their annotations) for the community through user-friendly interfaces in an automated way. The GGA project uses GMOD tools (JBrowse, Apollo, Tripal, Chado, etc.) and Galaxy, as a data loading orchestrator for administrators, with Docker lightweight virtualization technologies and Python libraries.

To facilitate the deployment and the administration of the GGA services of a first release of 53 brown algae genomes, we have developed a set of Python tools allowing mass deployment of Docker containers and automated data loading through Galaxy with the Bioblend API: http://gitlab.sb-roscoff.fr/abims/e-infra/gga_load_data.

To provide the community with a collaborative hub for accessing, visualizing and analyzing the brown algal genome and transcriptome resources, we have developed a web portal (https://phaeoexplorer.sb-roscoff.fr) giving access to the datasets, the GGA environments, the SequenceServer BLAST interfaces and some external resources.

In the context of genome sequencing programs of a large and diverse number of species such as the Vertebrate Genomes Project or the European Reference Genome Atlas project, the automated deployment of such e-infrastructures could bring to scientific annotation consortiums a simplified solution for collaborative environments.

Of Quality Control and Taxonomy

Laure Lemee[1], Etienne Kornobis[1,2], Dimitri Desvillechabrol[1,2], Juliana Pipoli da Fonseca[1], Deborah Garnier[1], Laurence Ma[1], Thomas Cokelaer[1,2]

1 Institut Pasteur – Plateforme Technologique Biomics – Center For Technological Resources and Research (C2RT) – 75015 Paris, France
2 Institut Pasteur – Bioinformatics and Biostatistics Hub – Département de Biologie Computationnelle – 75015 Paris, France,

Corresponding Author: thomas.cokelaer@pasteur.fr

1. Summary

During the 2019-2021 period, the Biomics Platform (Institut Pasteur, Paris) has sequenced about 200 runs a year (if we consider short read technologies only) covering various type of sequencing (DNA-seq, RNA-seq, ChiP-seq, Capture-seq, etc). When required (e.g., NextSeq sequencers), we performed demultiplexing. For each run, we also systematically perform standard QCs and taxonomic analysis. We provide an overview of the pipelines used within the platform including QC, demultiplexing and taxonomic analysis, which have been applied on thousand of samples. The underlying tools are standard and established tools in the field of NGS such as FastQC [1], MultiQC [2] and Kraken [3]. They have been wrapped within Snakemake pipelines available within Sequana [4] that benefit from simple user interface and GUI available within the Sequana project. They may be of interest for our users who wish to use the same tools and therefore to students and researchers you wish to move quickly from raw data to sequence analysis.

Acknowledgements

This work has been supported by France Génomique (ANR-10-INBS-09-09) and IBISA.

References

1. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
2. Philip Ewels, Måns Magnusson, Sverker Lundin and Max Käller (2016) MultiQC: Summarize analysis results for multiple tools and samples in a single report, Bioinformatics
3. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology 2014,15:R46
4. Cokelaer T., Desvillechabrol D., Legendre R. and Cardon M. (2017). 'Sequana': a Set of Snakemake NGS pipelines. Journal of Open Source Software. doi:10.21105/joss.00352

LORA, a Sequana Pipeline for Bacterial Genome Assemblies

Dimitri Desvillechabrol[1,2], Etienne Kornobis[1,2], Juliana Pipoli da Fonseca[1] Laurence Ma[1], and Thomas Cokelaer[1,2]

[1] Institut Pasteur - Platforme Technologique Biomics - Center For Technological Resources and Research (C2RT) - Paris, France.

[2] Institut Pasteur - Bioinformatics and Biostatistics Hub - Département de Biologie Computationnelle - Paris, France

Corresponding author: dimitri.desvillechabrol@pasteur.fr

In the context of the Biomics Sequencing Platform (Institut Pasteur), the PacBio long read technology [1] is used to sequence and assemble tens of bacterial genomes every year. In this poster, we describe the standard pipeline used within the platform. It is called LORA for LOng Read Assemblies. It is based on the Snakemake workflow manager [2] and is part of the Sequana project [3]. It has been tested mostly on bacterial genomes; in such case, the output of the pipeline is an annotated genome ready for publications. Since the underlying assembler is based on Canu [4], assemblies should also be possible on eukaryotes and nanopore data.

The LORA pipeline provides global quality assessments that computes common metrics like N50 or percentage of mapped reads. BUSCO [5] gives a quantitative assessment of the completeness in terms of expected gene content. Moreover, LORA produces a contig level quality assessment with BLAST or coverage analysis with Sequana to detect misassembly. At the end, a HTML report is provides where all quality assessment results can be introspected.

The open-source pipeline will be released under the Sequana project and be accessible on github where suggestions to include new features can be added.

Acknowledgements

This work has been supported by France Génomique (ANR-10-INBS-09-09) and IBISA.

References

[1] Anthony Rhoads and Kin Fai Au. Pacbio sequencing and its applications. *Genomics, proteomics & bioinformatics*, 13(5):278–289, 2015.

[2] Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

[3] Thomas Cokelaer, Dimitri Desvillechabrol, Rachel Legendre, and Mélissa Cardon. 'sequana': a set of snakemake ngs pipelines. *Journal of Open Source Software*, 2(16):352, 2017.

[4] Sergey Koren, Brian P Walenz, Konstantin Berlin, Jason R Miller, Nicholas H Bergman, and Adam M Phillippy. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*, 27(5):722–736, 2017.

[5] Mathieu Seppey, Mosè Manni, and Evgeny M Zdobnov. Busco: assessing genome assembly and annotation completeness. In *Gene prediction*, pages 227–245. Springer, 2019.

Capturing SARS-CoV-2

Juliana Pipoli da Fonseca[1], Etienne Kornobis[1,2], Elodie Turc[1], Thomas Cokelaer[1,2], Marc Monot[1]

1 Institut Pasteur – Plateforme Technologique Biomics – Center For Technological Resources and Research (C2RT) – 75015 Paris, France
2 Institut Pasteur – Bioinformatics and Biostatistics Hub – Département de Biologie Computationnelle – 75015 Paris, France

Corresponding Author: juliana.pipoli-da-fonseca@pasteur.fr
etienne.kornobis@pasteur.fr

1. Summary

The very high sensitivity of capture panels gives the opportunity to detect viruses within samples that are considered negative in RT-qPCR but present positive clinical symptoms [1]. In order to evaluate the efficacy of capture panels to access the SARS-CoV-2 genomeon patient samples, we performed capture of patient samples positive for SARS-CoV-2 by qPCR. For this we used 4 different ready-to-use commercial designs of capture probes including two SARS-CoV-2 only panels (Twist Bioscience and Arbor Bioscience) and two multiviral panels (Illumina and Twist Bioscience). We present preliminary analysis of the sequenced data obtained with this benchmark.

We performed capture on 19 patient samples positive for SARS-CoV-2 by qPCR. Samples were pooled by 4, according to their CT to reduce bias related to target abundance within the samples. The four capture panels were applied separately on all samples. Bioinformatics analysis were performed to study the off an on target performances of the panels including denovo analysis [2] and taxonomic contents [3].

The results obtained by the two multiviral panels tested suggests that both panels are able to capture high abundance targets, but fails to capture low abundance targets, having a high off target percentage translated by the abundance of host sequences. In our study, we observe that Arbor SARS-CoV-2 panel shows the best percentage on target of all panels. The efficacy of the double capture is even more evident for samples with higher CT (less viral copies) showing once more the power of the double capture. Even though Arbor panel with its double capture generates more on-target reads, it does not affect the percentage of final breadth of coverage results [4].

Acknowledgements

This work has been supported by France Génomique (ANR-10-INBS-09-09) and IBISA.

References

1. Tao Ai, Zhenlu Yang, Hongyan Hou, Chenao Zhan, Chong Chen, Wenshi Lv, qian Tao, Zyong Sun, Liming Xia, (2020) Correlation of Chest CT and RT-PCR Testing for Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases, Radiology, v296, N2
2. Cokelaer T, Desvillechabrol D, Legendre R and Cardon M (2017). 'Sequana': a Set of Snakemake NGS pipelines. Journal of Open Source Software.
3. Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology.
4. Desvillechabrol D, Bouchier C, Kennedy S, Cokelaer T . Detection and characterization of low and high genome coverage regions using an efficient running median and a double threshold approach. Giga science. 2018

A Comparison of MGI and Illumina Sequencing Technologies

Zachary Allouche[1], Etienne Kornobis[1,2], Juliana Pipoli da Fonseca[1] Dimitri Desvillechabrol[1,2], Georges Haustant[1], Julian Garneau[1], Laure Lemee[1], Laurence Ma[1],Elodie Turc[1], Imène Najjar[1],Valérie Briolat[1], Marc Monot[1], Thomas Cokelaer[1,2]

1 Institut Pasteur – Plateforme Technologique Biomics – Center For Technological Resources and Research (C2RT) – 75015 Paris, France
2 Institut Pasteur – Bioinformatics and Biostatistics Hub – Département de Biologie Computationnelle – 75015 Paris, France

Corresponding Author: zachary.allouche@pasteur.fr
thomas.cokelaer@pasteur.fr

1. Summary

In 2018, the Chinese company MGI[1] subsidiary of BGI, launched several sequencers using an innovative technology called the DNA NanoBall Sequencing (DNB-Seq). This technology is announced to be as performant as the Illumina [2,3] sequencing by synthesis but at a lower cost. Recently, the Biomics platform (Institut Pasteur) acquired a DNB-Seq G400 sequencer. We describe the runs that were performed with the MGI technology. Then, we compare MGI's DNB-Seq data with Illumina's TruSeq data including RNA-seq and DNA-seq sequencing runs[4]. We have sequenced 5 libraries with both MGI and Illumina runs, including RNA-seq and DNA-seq. We then compared these runs. In terms of quality, we have seen that they are broadly the same. Concerning the RNA-seq analysis we may have seen that the results between the two technologies were really close (PCA almost stackable) and further analysis are required to explain the discrepancies in terms of differentially regulated gene lists. Concerning the variant calling analysis we have found that these two technologies give very similar results; specific variants may be caused by the difference of coverage and not related with the technologies themselves. Recently, we also experimented on the MGI conversion protocol that takes ready-to-load Illumina libraries and converts them to MGI librairies with success. Finally, although MGI sequencers produces standard FastQ files, they are barcoded and an a priori merging is required, which can be done with MergeGI, which was recently developed [5].

Acknowledgements

This work has been supported by France Génomique (ANR-10-INBS-09-09) and IBISA.

References

1. https://en.mgi-tech.com/
2. https://www.illumina.com
3. MGI internal communication (2020)
4. Cokelaer T, Desvillechabrol D, Legendre R and Cardon M (2017). 'Sequana': a Set of of Snakemake NGS pipelines. Journal of Open Source Software.
5. Merge MGI barcoded fastq files into final fastq files. Available on https://pypi.org/project/MergeGI/ Open source code on https://github.com/sequana/MergeGI

Swiss-PO: a new tool to analyze the impact of mutations on protein three-dimensional structures for precision oncology

Fanny S. Krebs[1], Vincent Zoete[1,2], Maxence Trottet[1], Timothée Pouchon[2], Christophe Bovigny[2], Olivier Michielin[2,3]

[1] Computer-aided molecular engineering - UNIL, route de la corniche 9A, 1066, Epalinges, Switzerland

[2] Molecular Modelling Group - SIB, quartier UNIL-Sorge, 1015, Lausanne, Switzerland

[3] Centre d'oncologie de précision - CHUV, rue du Bugnon 46, 1011, Lausanne, Switzerland

Corresponding Author: vincent.zoete@unil.ch, Olivier.michielin@chuv.ch

Swiss-PO is a new web tool to map gene mutations on the three-dimensional (3D) structure of corresponding proteins and to intuitively assess the structural implications of protein variants for precision oncology. Swiss-PO is constructed around a manually curated database of 3D structures, variant annotations, and sequence alignments, for a list of 50 genes taken from the Ion AmpliSeqTM Custom Cancer Hotspot Panel. The website was designed to guide users in the choice of the most appropriate structure to analyze regarding the mutated residue, the role of the protein domain it belongs to, or the drug that could be selected to treat the patient. The importance of the mutated residue for the structure and activity of the protein can be assessed based on the molecular interactions exchanged with neighbor residues in 3D within the same protein or between different biomacromolecules, its conservation in orthologs, or the known effect of reported mutations in its 3D or sequence-based vicinity. Swiss-PO is available free of charge or login at https://www.swiss-po.ch [1].

1. Krebs, F.S., Zoete, V., Trottet, M. et al. Swiss-PO: a new tool to analyze the impact of mutations on protein three-dimensional structures for precision oncology. npj Precis. Onc. 5, 19 (2021). https://doi.org/10.1038/s41698-021-00156-5

Galaxy Genome Annotation (GGA) environment in the cloud

Romain DALLET[1], Gildas LE CORGUILLÉ[1], Loraine BRILLET-GUÉGUEN[1,2] and Erwan CORRE[1]

[1] Sorbonne Université, CNRS, FR2424, ABiMS, Station Biologique de Roscoff, Place Georges Teissier, 29680, Roscoff, France
[2] Sorbonne Université, CNRS, Integrative Biology of Marine Models (LBI2M), Station Biologique de Roscoff, Place Georges Teissier, 29680, Roscoff, France

Corresponding Author: romain.dallet@sb-roscoff.fr

Abstract :

The EOSC-Life project aims to create an open collaborative digital space for life science in the European Open Science Cloud (EOSC). The ABiMS bioinformatics platform, member of the European Marine Biological Resource Centre (EMBRC) research infrastructure, is involved in the work package 2 (WP2) dedicated to make computational tools, workflows and registries findable, accessible, interoperable and reusable (FAIR).

The Galaxy Genome Annotation [1](GGA) (https://galaxy-genome-annotation.github.io) project consists of several projects and tool suites that are working closely together to deliver a comprehensive, scalable and easy to use Genome Annotation experience. The Galaxy Genome annotation environment not only offers a wide array of high-profile tools in Galaxy [2] for structural and functional annotation, but also a highly integrated set of "dockerized" GMOD tools [3], a collection of open-source applications for visualizing, annotating, and managing genomic data (JBrowse, Apollo, Tripal, Chado). Galaxy is used as a data loading orchestrator for administrators, with dedicated Galaxy tools and workflows to interact with GMOD tools, and Python libraries to make all tools work together.

As part of the EOSC-Life WP2, we planned to provide the GGA environment available in the cloud. So we are currently developing Ansible recipes (https://github.com/abims-sbr/GGA_Cloud) to deploy the GGA environment in an Openstack cloud infrastructure. These Ansible recipes allow the deployment of a Virtual Machine in a cloud via the Terraform software, the installation of the GGA environment and its dependencies, as well as loading data into the Galaxy library.

References

1. Bretaudeau A, Rasche H, Boudet M *et al.* Galaxy genome annotation: easier genome annotation using Galaxy and GMOD tools. *F1000Research* 2019, **8**:1026 (poster) (https://doi.org/10.7490/f1000research.1116992.1)
2. Enis Afgan, Dannon Baker, Bérénice Batut, Marius van den Beek, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Björn Grüning, Aysam Guerler, Jennifer Hillman-Jackson, Vahid Jalili, Helena Rasche, Nicola Soranzo, Jeremy Goecks, James Taylor, Anton Nekrutenko, and Daniel Blankenberg. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update, Nucleic Acids Research, Volume 46, Issue W1, 2 July 2018, Pages W537–W544, doi:10.1093/nar/gky379
3. Generic Model Organism Database (GMOD), http://gmod.org

ADLIN Science, an end-to-end digital platform for multi-omics data integration

Luciana De Oliveira[1], Marion Crespo[1], Mohamed Ouasti[1], Richard Smadja[1] and Paul Rinaudo1[1]
ADLIN Science - 4, Rue Pierre Fontaine, Pépinière "Genopole Entreprises" Campus 3, 91058, Évry-Courcouronnes Cedex, France

Corresponding author: paul@adlin-science.com

While multi-omics data are becoming more popular, it is not a surprise that one of the consequences is the increase in the generation of massive and complex biological data sets. Further, combining the heterogeneity of bio-data and the multi-factorial aspects of the biological systems results in rising the complexity of high-dimensional data [1,2]. On this basis, we face a challenge: understand the modular dynamic of each -omic layer and integrate it into a molecular network system. On the other hand, multiscale exploratory analysis is also critical for biologists, demanding biological knowledge and a high level of computational skills [3].

In order to help to solve those complex problems, ADLIN Science [4], a digital health-tech company, develops innovative digital solutions for researchers, clinicians, biotech and pharmaceutical companies in the molecular biology field, and more particularly in multi-omics sciences.

ADLIN Science aims to facilitate multi-omics data exploration by creating a tool adapted for biologists, bio-informaticians and mathematicians, and by enabling the constitution of the multi-disciplinary team since the beginning of the research protocol up to the publication with traceability and security. Our company helps laboratories and researchers manage all of their omics data and developing state-of-the-art approaches to produce publication-ready results. Our solution provides data and metadata structuring relating to the research protocols. Also, the multi-omics data integration, statistics analysis, IA (Intelligence Artificial) approach and data visualization are combined into a collaborative workspace where the research team can share and synchronize their information.

Finally, ADLIN can provide an innovative application for multi-omics analysis and catches scientific and economic values in research and health fields by converting data into therapeutic insights.

Acknowledgements

ADLIN Science want to say thank you to the support and collaboration of GENOPOLE, Université Paris-Saclay, La French Tech Paris Saclay, Université de Paris, Inserm, École Polytechnique - IP Paris, Institut Curie, Université Paris Diderot - Paris 7, Agoranov and Inserm.

References

[1] Ana Conesa and Stephan Beck. Making multi-omics data accessible to researchers. *Scientific data*, 6(1):1–4, 2019.

[2] Roman Schulte-Sasse, Stefan Budach, Denes Hnisz, and Annalisa Marsico. Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. *Nature Machine Intelligence*, pages 1–14, 2021.

[3] Indhupriya Subramanian, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. Multi-omics data integration, interpretation, and its application. *Bioinformatics and biology insights*, 14:1177932219899051, 2020.

[4] ADLIN Science. https://adlin-science.com/. Accessed: 2021-06-01.

A FAIR toolkit for fast visualization of omics data and metadata

Clémence Carcopino[1], Pierre Cuzin[1], Laura Leroi[1], Alexandre Cormier[1], Cyril Noël[1] and Patrick Durand[1]

[1]IFREMER-IRSI-Service de Bioinformatique (SeBiMER), Centre Bretagne - ZI de la Pointe du Diable, CS 10070 - 29280 PLOUZANE, FRANCE

Corresponding Authors: laura.leroi@ifremer.fr

When studying relations between marine habitats and model organisms, IFREMER research teams develop integrative approaches at various scales, from molecular mechanisms to evolutionary processes. For that purpose, scientists collect numerous omics data sets that should be analyzed as quickly as possible. Therefore, the challenge for a Bioinformatics Core Platform consists in providing tools to inventory, explore and interpret those data as quickly as possible, too.

To tackle this challenge, we first developed an interactive genome catalog based on the Keshif data visualization tool [1]. It provides the list of available model organisms, along with metadata such as lineage, genome version, submitter, assembly level, etc. The catalog enables view filtering through specific features defined according to the metadata. In addition, it gives access to a genome browser session implemented using JBrowse2 viewer [2]. Data preparation for JBrowse2 has been fully automated with a Nextflow [3] pipeline. It combines omics data preprocessing using dedicated tools (e.g. samtools [4], bcftools [4], genometools [5]) with JBrowse CLI. The pipeline also automatically handles files deployment to meet JBrowse2 requirements, as well as IT specificities (e.g. web server isolated from computing nodes and data storage).

In conclusion, we propose a simple FAIR toolkit for fast visualization of omics data and metadata, and an easy way to share omics information at various levels (private, restricted or public access). The toolkit is available at: https://github.com/ifremer-bioinformatics/omics-catalog

References

1. Mehmet Adil Yalçın, Niklas Elmqvist, Benjamin B. Bederson. Keshif: Out-of-the-Box Visual and Interactive Data Exploration Environment. *Proc. of IEEE VIS 2016 Workshop on Visualization in Practic*e, http://www.github.com/adilyalcin/keshif
2. Robert Buels, Eric Yao, Colin M. Diesh, Richard D. Hayes, Monica Munoz-Torres, Gregg Helt,David M. Goodstein, Christine G. Elsik, Suzanna E. Lewis, Lincoln Stein and Ian H. Holmes. JBrowse: a dynamic web platform forgenome visualization and analysis. *Genome Biology* **,** 17: 66, 2016.
3. Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature biotechnology*, 35(4):316-319, 2017
4. Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16): 2078–2079, 2009.
5. Gordon Gremme, Sascha Steinbiss, Stefan Kurtz. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans Comput BiolBioinform*, 10(3):645-56, 2013

A digital space for EMERGEN, the French plan for SARS-CoV-2 genomic surveillance and research

Thomas DENECKER[1,2], Adeline FERI[3], François GERBES[1], Julien SEILER[4], Gildas LE CORGUILLÉ[5], Nicole CHARRIÈRE[1], Bruno COIGNARD[3], Abdelkader AMZERT[6], Franck LETHIMONNIER[6], Claudine MÉDIGUE[1,7], David SALGADO[8], Christophe ANTONIEWSKI[9] & Jacques VAN HELDEN[1,10]

1. CNRS, Institut Français de Bioinformatique, IFB-core, UMS 3601, Évry, France
2. Université Paris-Saclay, INRAE, MaIAGE, 78350 Jouy-en-Josas, France.
3. Santé Publique France, 12, rue du Val d'Osne 94 415 Saint-Maurice Cedex
4. CNRS UMR7104, Inserm U1258, Université de Strasbourg, Institut de Génétique et de Biologie Moléculaire et Cellulaire, Illkirch, France
5. Sorbonne Université, CNRS, FR2424, ABiMS, Station Biologique, Place Georges Teissier, Roscoff 29680, France
6. Inserm, Institut national de la santé et de la recherche médicale, 101 rue de Tolbiac 75013 Paris.
7. UMR 8030, CNRS, Université Evry-Val-d'Essonne, CEA, Institut de Biologie François Jacob - Genoscope, Laboratoire d'Analyses Bioinformatiques pour la Génomique et le Métabolisme, Evry, France.
8. Aix Marseille Univ, INSERM, MMG, 13005, Marseille, France
9. Sorbonne Université, CNRS FR3631, Inserm US037, Institut de Biologie Paris Seine (IBPS), ARTbio Bioinformatics Analysis Facility, Paris, France.
10. Aix-Marseille Univ, Inserm, laboratoire Theory and approaches of genome complexity (TAGC), Marseille, France.

Corresponding Author: thomas.denecker@france-bioinformatique.fr

In January 2021, the French ministers of Health (MSS) and Research (MESRI) launched a national plan for SARS-CoV-2 genomic surveillance which aims at monitoring the evolution of the COVID-19 in France, at detecting new variants, and at supporting the integration of viral genomic data and health data for both surveillance and research. The project is co-lead by the national agency in charge of emerging infectious diseases (ANRS-MIE) and Santé Publique France (SPF), and relies on a consortium involving 4 high-throughput sequencing platforms, including the National Reference Centers for viruses, 43 teams of the ANRS-MIE network, the Inserm and the Institut Français de Bioinformatique (IFB).

The IFB, in collaboration with the Inserm, is in charge of developing the digital space and bioinformatics services. This work package is structured in 3 phases:

1. **SARS-CoV-2 variant monitoring.** The EMERGEN-DB database gathers the non-sensitive metadata (variant annotations of the virus) produced by the sequencing platforms of the consortium. It is equipped with both user-friendly and programmatic interfaces, enabling users to upload data and query the databases manually or in batch.
2. **Digital space for non-sensitive data.** We launched a workgroup of biologists and bioinformaticians from the sequencing teams to deploy a hardware and software environment and implement workflows for high-throughput analyses of SARS-CoV-2 sequencing data. The digital space will be powered to support the automatic processing of 10,000 full viral genomes per week, and will allow managing workflows via either a Galaxy server, or a Unix terminal. The server will be equipped with the standard tools for NGS analysis as well as specific tools for the analysis of viral variants.
3. **Pairing SARS-CoV-2 sequences and personal patient data.** We will deploy a secured digital space to support the pairing between health data of patients and the sequences of their infecting SARS-CoV-2 strains, thus making it possible to carry out epidemiological or clinical studies.

EMERGEN will also include a service of data brokering by supporting the automated submission of genomic sequences to two international repositories: GISAID to ensure a rapid sharing of the consensus genomes and variant annotations, and EBI-ENA to ensure an open access to the raw sequences and consensus genomes.

Keywords : SARS-CoV-2; COVID-19; genomic surveillance; health data; EMERGEN ; EMERGEN-DB ;

Meta-analysis of RNAseq runs performed on the IMRB genomics and bioinformatics platform: the dark side of QC

Sidwell RIGADE [1], Carole CONEJERO [2], Stéphane KERBRAT [2], Denis MESTIVIER [1]

[1] Bioinformatics core of Institut Mondor de Recherche Biomédicale, 8 rue du Général Sarrail, 94010, Créteil, France

[2] Genomics core of Institut Mondor de Recherche Biomédicale, 8 rue du Général Sarrail, 94010, Créteil, France

Corresponding Author: sidwell.rigade@u-pec.fr

RNAseq is a sequencing technique widely used in genomics. At the Institut Mondor de Recherche Biomedical (IMRB), for the past two years, every RNASeq performed by the Genomics Platform are quality checked (QC) by the bioinformatics platform. Starting from standard QC (mainly fastQC [1]), we encountered sequencing situations that leads us to add more QC to our QC-pipeline, especially regarding rRNA level and contamination. For example, the kits used in genomics for library preparation do not necessarily live up to the manufacturers' promises. In the case of a 'total RNA' kit, a depletion must be done to delete rRNA and this depletion can be effective depending on the sample, the treatment, and the experimental conditions. Contamination can also occur at different levels of the experimentation.

We expand our QC-pipeline with tools such as SortMeRNA [2] which detects rRNAs and Kraken2 [3] which searches reads against a large database Database built from the Refseq bacteria, archaea, viral libraries, the GRCh38 human and GRCm38 genomes..

Here, we performed a retrospective meta-analysis of all our RNAseq QCs performed on the mouse genome. This corresponds to 25 runs, 422 samples and 12,324,596,035 reads in total.

If we show that a sample have a high level of rRNA (> 10%), this is as many reads lost since the number of total reads per run does not change. Moreover, high level of rRNA will inevitably have an impact on the statistical analysis later. It is sometimes possible to detect poor quality samples at the time of analysis if they are too different from the others via a PCA but this is not always the case.

Our results shows that samples with high level of rRNA or contamination are not isolated problems. However, to the best of our knowledge, RNAseq literature report QC often limited to the fastQC tool (or similar). Our results shows that more statistics (%r RNA/contamination) should include.

References

1. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
2. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 2012
3. Wood, D.E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. Genome Biol 20, 257 (2019).
4. Eliah G. Overbey et al. NASA GeneLab RNA-seq consensus pipeline: standardized processing of short-read RNA-seq data. iScience. 2021 Apr 23

The Migale bioinformatics facility

Valentin LOUX[1,2], Mouhamadou BA[1,2], Damien BERRY[1,2], Hélène CHIAPELLO[1,2], Olivier INIZAN[1,2], Mahendra MARIADASSOU[1,2], Véronique MARTIN[1,2], Cédric MIDOUX[1,2,3], Olivier RUÉ[1,2], Valérie VIDAL[1,2] and Sophie SCHBATH[1,2]

[1] Université Paris-Saclay, INRAE, MaIAGE, Domaine de Vilvert, 78350, Jouy-en-Josas, France.

[2] Université Paris-Saclay, INRAE, BioinfOmics, MIGALE bioinformatics facility, Domaine de Vilvert, 78350, Jouy-en-Josas, France.

[3] Université Paris-Saclay, INRAE, PRocédés biOtechnologiques au Service de l'Environnement, 1 rue Pierre-Gilles de Gennes, CS10030, 92761, Antony, France.

Corresponding Author: valentin.loux@inrae.fr

The Migale bioinformatics facility is a team of INRAE's MaIAGE research unit (Applied Mathematics and Computer Science, from Genome to the Environment). It has been providing services to the life sciences community since 2003.

Migale is an open platform, that offers four types of services ;

- an open infrastructure dedicated to life sciences data processing,
- dissemination of expertise in bioinformatics,
- design and development of bioinformatics applications,
- genomic, metagenomic and metatranscriptomic analysis.

Migale is part of the French Institute of Bioinformatics (IFB) and France Génomique projects. It has an ISO9001 certification and is also one of the four platforms which compose BioinfOmics, the national Research Infrastructure in bioinformatics of INRAE.

The poster will illustrate the platform services with examples chosen from recent achievements : i) how to switch trainings from face-to-face to online , ii) traceability of analyses and projects reports as a way to train and empower users.

A complete description of Migale facility's service offer is available on its website : https://migale.inrae.fr

The IBENS Genomics core facility

Médine BENCHOUAIA[1], Corinne BLUGEON[1], Jade GUGLIERI[1],Laurent JOURDREN[1], Sophie LEMOINE[1], Catherine SENAMAUD-BEAUFORT[1], Stéphane LE CROM[1,2] and Morgane THOMAS-CHOLLIER[1]

[1] Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France

[2] Sorbonne Université, CNRS, Institut de Biologie Paris-Seine (IBPS), Laboratory of Computational and Quantitative Biology (LCQB), 75005 Paris, France

Corresponding Author: jourdren@bio.ens.psl.eu

The **genomics core facility of the Institut de Biologie de l'École normale supérieure** (IBENS) [1,2] was created in 1999. We focus on functional genomics in **eukaryotes** , including classical model organisms, as well as more exotic organisms (jellyfish, birds, butterflies…). **The facility has always been a well-balanced structure between wet-lab and bioinformatics**: half of the team is involved on the wet-lab part; the other half being involved on the data analysis part. Our goal is to assist laboratories during their **high-throughput sequencing projects** from the experimental design to data analysis for publication. We are part of the **France Génomique consortium** and have been following the **ISO 9001** quality international standard since March 2013.

Our genomics core facility offers many services to the community: **library preparation** (RNA-seq, scRNA-seq and ChIP-seq), **sequencing** (including "ready-to-load" libraries) for short (Illumina) and long reads (Oxford Nanopore); and **bioinformatics analysis** (RNA-seq and scRNA-seq).

All the staff working on the facility gets a balanced schedule between the core **production service** and **research and development projects** to propose **up-to-date and reliable experimental solutions** to our collaborators. To cope with the experimental constraints of our users among the research teams, we invest time in **testing library protocols** (very low quantities, ribosome depletions…). We are also deeply involved in **software development** to manage our project analyses (65% of projects are analysed on the facility). The tools we develop are distributed on an open source basis on **GitHub** [3] and we now provide most of them as **Docker** images [4] to **ease the distribution** of our work. We develop workflows to achieve **reproducible and transparent data analysis** of our high throughput experiments.

Since 2016, our facility has been offering two new technologies. The first one is devoted to **single cell RNA-seq** witha **Chromium** system from **10X Genomics** based on the Drop-seq protocol. The second one is dedicated to **long read** sequencing in RNA-seq. We use **Oxford Nanopore Technologies MinION** system in order to sequence full length transcripts for isoform abundance estimation.

This year we have released a rewritten and enhanced version of **ToulligQC** [5], our QC tool for Oxford Nanopore sequencers and we are currently testing scNaUmi-seq protocol [6] to improve our scRNA-seq service with long read sequencing.

All these developments allow us to be at the **state of the art in functional genomics** applications, so that we can provide to our users all the tools needed to succeed in their high throughput experiments.

Acknowledgements

The IBENS genomics core facility is supported by the France Génomique national infrastructure, funded as part of the "Investissements d'Avenir" program managed by the Agence Nationale de la Recherche (contract ANR-10-INBS-09).

References

[1] http://genomique.bio.ens.psl.eu
[2] Twitter @Genomique_ENS
[3] https://github.com/GenomicParisCentre/
[4] https://hub.docker.com/r/genomicpariscentre/
[5] https://github.com/GenomicParisCentre/toulligQC
[6] Lebrigand, K., *et al.* High throughput error corrected Nanopore single cell transcriptome sequencing. *Nat Commun* **11,** 4025 (2020)

Reproducibility in bioinformatics: take care of your code with Software Heritage

Pierre Poulain[1], Morane Gruenpeter[2] and Roberto Di Cosmo[3]
[1] Université de Paris, CNRS, Institut Jacques Monod, F-75006, Paris, France
[2] Software Heritage, Inria, France
[3] Software Heritage, Inria and University of Paris, France

Corresponding author: pierre.poulain@u-paris.fr

Reproducibility is an ongoing effort in the bioinformatics community[1]. Open science helps toward this goal with open access to the scientific literature, open data and open source research software. In 2018, more than 36% of yearly published papers were published under open access conditions[2]. In biology and bioinformatics, the recent development of preprints has acted as a leverage towards open access.

Raw data deposit in public international repositories of genomics and proteomics data is now well established and enforced by most journal editorial policies. Availability of all-purpose data repositories such as Zenodo or Figshare also fostered open data.

It is now important to establish good practices also for scientific software, going beyond the common approach of depositing code in development platforms such as GitHub or GitLab, where long-term preservation is not guaranteed.

This poster aims to present to our community the Software Heritage (SWH) [1] archive[3]: it collects, preserves, and makes available all source codes, from the one that ran on the Apollo 11 Guidance Computer to the source code of the Gromacs molecular dynamics engine, the Bowtie 2 genomics read aligner, the Cytoscape network visualization software... Software Heritage can also archives smaller programs like the scripts commonly used in bioinformatics.

Software Heritage regularly collects source code from a growing list of code hosting platforms, and provides a powerful "Save code now" functionality [2] that allows to trigger archival for any public repository based on the Git, Mercurial or Subversion version control systems, free of charge. Any object archived in Software Heritage is assigned an intrinsic persistent identifier [3] called the SWHID[4], that can be independently verified.

We will present actionable recommendations for better referencing and indexing research source code, including best practices for providing metadata files in the code repository (AUTHOR(s) file with the list of authors, LICENSE file with the applicable license to the source code, README file with the description of the software and other valuable information) and for making it citable [4], including pointers to appropriate bibliographic styles[4].

References

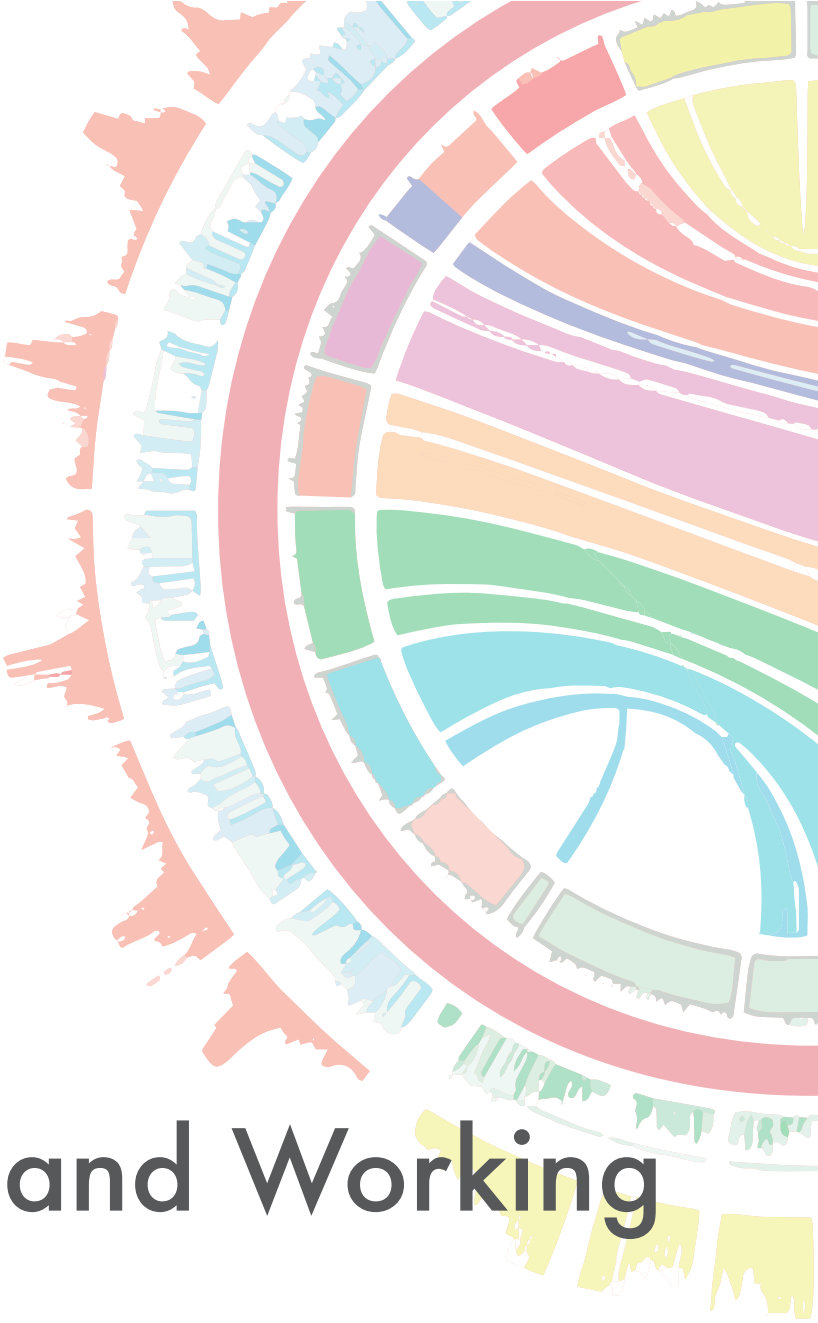
[1] Y.-M. Kim, J.-B. Poline, and G. Dumas. "Experimenting with Reproducibility: A Case Study of Robustness in Bioinformatics". en. In: *GigaScience* 7.7 (2018).

[2] European Commission. *Trends for open access to publications*. Website. Available from https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/open-science-monitor/trends-open-access-publications_en. 2018.

[3] R. Di Cosmo and S. Zacchiroli. "Software Heritage: Why and How to Preserve Software Source Code". In: *Proceedings of the 14th International Conference on Digital Preservation, iPRES 2017, Kyoto, Japan*. Available from https://hal.archives-ouvertes.fr/hal-01590958. 2017.

[4] R. D. Cosmo. "Archiving and Referencing Source Code with Software Heritage". In: *ICMS*. Volume 12097. Lecture Notes in Computer Science. Springer, 2020, pages 362–373.

1. https://www.softwareheritage.org/
2. https://archive.softwareheritage.org/save/
3. https://www.softwareheritage.org/2020/07/09/intrinsic-vs-extrinsic-identifiers/
4. https://codemeta.github.io/



> Networks and Working Groups

Educational innovation: The BioInformatics Learning Lab project (BILL)

Students of the Bioinformatics master2, Students of the IMHE master1, Emira Cherif3, Marie-Ka Tilak3, Aida Diouf4, Fabrice Merezegue4, Jonathan Kirszling5, Colin Durand4, Nicolas Moureau4, Fabien Dubois4, Olivier Guirado4, Mélanie Debiais Thibaud3, Jean-Christophe Avarre3, Anne-Sophie Gosselin-Grenet6* and **Anna-Sophie Fiston-Lavier3***

1 Master Interactions Microorganismes-Hôtes-Environnements, Faculté des Sciences, Université de Montpellier, Montpellier, France.

2 Master Bioinformatique, Connaissances, Données, Faculté des Sciences, Université de Montpellier, Montpellier, France.

3 ISEM, IRD, CNRS, EPHE, Université de Montpellier, Montpellier, France

4 Faculte des Sciences, Université de Montpellier, Montpellier, France

5 Direction du Systeme d'Information et du Numerique (DSIN, Université de Montpellier, Montpellier, France

6 DGIMI, Université de Montpellier, INRAE, Montpellier, France.

* corresponding authors

Corresponding Author: anna-sophie.fiston-lavier@umontpellier.fr

The BILL (for BioInformatics Learning Lab) project has enabled the implementation of an innovative interdisciplinary learning lab on the Montpellier Faculty of Science Campus. This learning lab aims to foster collective intelligence and training through research for students.

In the collaborative BILL space, composed of a molecular biology lab and a computer lab, both modular and dedicated to DNA sequencing and bioinformatics, students with diverse backgrounds are encouraged to interact and share their mutual skills. Supported by a dynamic multidisciplinary team of professors, researchers and technicians, they generate and analyze the sequencing data themselves and participate in the valorization (writing of international scientific publications) and dissemination of their results (communications in congresses, popularization among the general public). The students, confronted with a real research laboratory, thus manage small research projects together and become actors in their training.

Thanks to the BILL project, students of the 'Microorganism/Host and Environment Interactions' (IMHE) (https://bioagro.edu.umontpellier.fr/master-biologie-agrosciences/interactions-microorg-hotes/) and 'Bioinformatics, Knowledge, Data' (BCD) (https://sns.edu.umontpellier.fr/fr/master-sciences-numerique-pour-la-sante-montpellier/bcd/) masters programs at the University of Montpellier have been conducting research projects related to viral evolution and species hopping for the past three years by carrying out high-throughput genomic analyses. In these projects, students perform viral DNA extraction from field or laboratory samples, construction of DNA libraries for high-throughput sequencing and bioinformatics analyses of the obtained sequencing data. The first two years (2018-2020), the students studied the evolution of the carp herpesvirus CyHV3 by sequencing short reads (MiSeq Illumina®) of viral genomes from serial passages in cellulo [1]. This year (2020-2021), students studied CyHV3 cross-sepecies transmission by analyzing viral variants isolated from other fish species than common carp using long reads sequencing (MinIon; Oxford Nanopore Technologies). This pedagogical and scientific effort will allow the publication of scientific articles of which the students of the IMHE and BCD masters are co-authors [1].

References

[1] Klafack *et al.* Cyprinid herpesvirus 3 Evolves In Vitro through an Assemblage of Haplotypes that Alternatively Become Dominant or Under-Represented. *Viruses* **2019**, *11*(8), 754; https://doi.org/10.3390/v11080754

BiogenHack 2021: a hackathon for enhancing interoperability within biology and health community

Sofia STRUBBIA[1], Audrey BIHOUÉE[1], Philippe BORDRON[2], Aurélien BRIONNE[3], Erwan CORRE[4], Olivier DAMERON[5], Adrien FOUCAL[1], Isabelle HUE[3], Gabriel MARKOV[6], Camille MAUMET[7], Aline FOURY[8], Jeanne GOT[5], Marie-Pierre MOISAN[8], Richard REDON[1], Anne SIEGEL[5], Alban GAIGNARD[1]

[1]Institut du Thorax, Inserm UMR1087, CNRS UMR 6291, Univ Nantes - [2]Univ Nantes, Inserm, TENS, The Enteric Nervous System in Gut and Brain Diseases, IMAD, Nantes, France - [3]LPGP UR 1037, INRAE, Rennes -[4]ABiMS FR2424 CNRS/Sorbonne Université - [5]Dyliss,IRISA, Rennes - [6]LBI2M UMR 8227 CNRS/Sorbonne Université - [7]Inria, Univ Rennes, CNRS, Inserm – IRISA, Empenn, Rennes - [8]NutriNeuro UMR 1286 INRAE/Université de Bordeaux

Corresponding Author: sofia.strubbia@univ-nantes.fr ; alban.gaignard@univ-nantes.fr

Supported by the Biogenouest network, the MoDaL (Multi Scale Data Links) project aims to decompartmentalize life science resources and enhance collaborations across various actors through interoperable datasets as well as analysis tools and workflows. Indeed, life science research is increasingly facing the need for linking multi-scale data (genes, cells, organs, individuals, populations) and heterogeneous data, such as phenotypes (including micro and macroscopic imaging) and omics data.

To facilitate knowledge sharing and co-design of innovative computational solutions, we organized a hackathon for a large scientific community, spanning the fields of computer science, marine and plant biology, as well as human - animal health and biology. Although this type of event is well known in the (bio)informatic community, it is still new for biologists in the midst of a digital transition.

The hackathon took place remotely, on January 25th and February 1st, 2021. During the two days, the 30 participants had the opportunity to meet in a virtual space [1], attend brief conferences and work together on collaborative projects. In this poster, we briefly introduce the projects proposed by the participants which are also available online on the hackathon's GitHub repository [2]. The projects addressed the data science aspects of different questions, ranging from biological issues such as the search for a transcriptomic signature of aggressivity level in sport horses or the integration of genome-wide knowledge of metabolomic networks, to more computationally-oriented issues such as facilitating the sharing of data analysis workflows (including the portability of environments or the version of software, packages, annotations, the genome or ontologies used).

Thanks to the feedback from three participants interviewed at the end of the hackathon [3], we know that what was most appreciated were the variety of the proposed projects, the smooth communication between the biological and (bio)-informatics communities and the spontaneity of the exchanges, enhanced by the software chosen and by a meeting between projects leaders organized before the hackathon. Furthermore, collaborative work was timed by collective moments of feedback on the projects, thus encouraging discussion and the transmission of knowledge. Finally, during the two days, participants had the opportunity to sign up for flash presentations: brief interventions to talk about a tool, a technology or to propose a discussion topic to be deepened. The hackathon took place in a friendly and collaborative atmosphere and was an opportunity to broaden/federate the community, to identify people and skills and to experience a moment of exchange between the different scientific communities.

References

[1] https://gather.town
[2] https://github.com/Biogenouest/biogen-hack-2020
[3] https://www.biogenouest.org/article/projet-modal-retour-sur-le-hackathon/

FAIR-Checker, a web tool to support the findability and reusability of digital life science resources

Thomas Rosnet[1,5], Vincent Lefort[2,5], Marie-Dominique Devignes[3,5] and Alban Gaignard[4,5]
[1] TAGC/INSERM U1090, Univ Aix-Marseille, Marseille, France
[2] LIRMM, Univ Montpellier, CNRS, Montpellier, France
[3] LORIA, Université de Lorraine, CNRS, Inria, Nancy, France
[4] L'institut du thorax, INSERM, CNRS, University of Nantes, Nantes, France
[5] Institut Français de Bioinformatique, CNRS UMS 3601, France

Corresponding author: thomas.rosnet@france-bioinformatique.fr

1 Problem statement

With a major increase of life science data production over the last decade, it is becoming more and more important to better share and reuse biological digital resources (datasets, bioinformatics tools or workflows, training materials, etc.). To that end, FAIR principles [1] have recently been proposed and are currently being adopted by large communities. However, assessing how much a resource is FAIR is nowadays challenging since answering human-oriented questionnaires is time-consuming and computational evaluations (FAIRMetrics, RDA Maturity Indicators) often require technical expertise. In this work we plan at easing the monitoring of the FAIRness of life science digital resources such as datasets, computational tools, training material or publications.

2 Approach and results

We propose a web interface [https://fair-checker.france-bioinformatique.fr] aimed at empowering scientists to progress in the FAIRification of their resources. This tool benefits from FAIRMetrics APIs to provide a global assessment and recommendations. We also leverage semantic technologies to help users in annotating their resources with high-quality metadata. For each evaluated resource, we build a knowledge graph based on embedded RDF triples (microdata, json-ld), as well as external knowledge (public SPARQL endpoints). To evaluate metadata quality, we check that used ontology terms are already known in reference registries. Finally, we leverage Bioschemas, the extension of Schema.org for life sciences, to automatically generate SHACL constraints. Their evaluation informs users on missing metadata, required or recommended for specific types of resources (genes, proteins, training material, computational tools, etc.). This also results in a form to annotate and produce enriched metadata.

3 Demonstration scenario

As an example, we will showcase how *FAIR-Checker* can help in assessing the FAIRness of PhyML [2]. (Step 1) We launch a remote evaluation of 22 FAIR metrics and show recommendations in case of failure. Then in (Step 2) we investigate the use of standard vocabularies or ontologies in metadata. We show how metadata can be scrapped from web resources, and which ontology terms they rely on. Further, we search if these terms are known in major vocabulary registries such as Linked Open Vocabularies (LOV) Ontology Lookup Service (OLS) or BioPortal. Finally, in (Step 3) we demonstrate how *Bioschemas*, can be exploited to (i) evaluate metadata quality through *required*, *recommended* and *optional* annotations, and (ii) capture missing annotations through an auto-generated user form.

4 Future works

As future work, we aim at extending our tool to (i) support multiple resource types in line with the different released Bioschemas profiles and (ii) provide a common and synthetic view on other FAIR recommendations such as the RDA maturity indicators, as well as the forthcoming EOSC FAIR metrics.

References

[1] Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016).

[2] Guindon S., Dufayard J.F., Lefort V. *et al.* New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology, 59(3):307-21, 2010.

ASPICov: an Automated Pipeline for Identification of SARS-Cov2 nucleotidic Variants

Valentin TILLOY[1,2], Pierre CUZIN[3], Laura LEROI[3] , Patrick DURAND[3] and Sophie ALAIN[1]

[1] Centre National de Référence des Herpèsvirus, CHU Dupuytren, 87000, Limoges, France
[2] UF9481 Bioinformatique, CHU Dupuytren, 87000, Limoges, France
[3] IFREMER-IRSI-Service de Bioinformatique, Centre Bretagne, 29280 Plouzane, France

Corresponding Author: valentin.tilloy@unilim.fr

Whole-genome sequencing (WGS) is used for clinical surveillance of SARS-Cov2 in order to detect emerging mutations, facilitate epidemiological studies and anticipate possible therapeutic/vaccinal escape.

In this context, we have developed a pipeline dedicated to identify SARS-Cov2 mutations from a broad range of samples (clinical, wastewaters,…).

ASPICov is a Nextflow [1] pipeline developed to provide a rapid, reliable and complete analysis of NGS SARS-Cov2 samples. ASPICov produces useful information such as quality reports, VCF files, consensus sequences and various plots. In order to ensure FAIR data analysis, the workflow follows nf-core guidelines and use Singularity [2] containers to wrap tool environments.

A succession of commonly used tools combined to an optimized configuration is a key for the robustness of the pipeline. Our workflow performs a trimming on raw-reads depending on sequencing technology (Illumina or Iontorrent), then a quality check is carried out in order to visualize data. An optimized alignment is done against a determined reference which can differ according to DNA preparation strategy (Capture or Ampliseq). Several data are collected and formatted (files, plots) from the alignments as sequencing indicators. Variant calling and normalization steps are done from alignments and nucleotidic variations are annotated. Specific filters are then applied in order to get only confident variants. Specific variants files and a plot comparing samples are done for a better interpretation of the dataset by the biologist.

ASPICov has been validated using a dataset and has demonstrated its efficiency and accuracy. Several new features are under development.

This pipeline (source code, documentation and full list of tool dependencies) is available at: https://gitlab.com/vtilloy/aspicov

This software is developed as part of the French National Obepine Network

References

[1] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, Cedric Notredame. Nextflow enables reproducible computational workflows. Nature Biotechnology volume 35, pages 316–319 (2017).

[2] Gregory M. Kurtzer, Vanessa Sochat, Michael W. Bauer. Singularity: Scientific Containers for Mobility of Compute. PLoS One; 12(5): e0177459 (2017).

The IFB e-Learning Working Group

Hélène CHIAPELLO[1,2], Thomas DENECKER[1], Gildas LE CORGUILLE[1,3], Pierre POULAIN[4], Denis PUTHIER[5], Olivier SAND[1], Julien SEILER[1], Claire TOFFANO-NIOCHE[6]

[1] CNRS, Institut Français de Bioinformatique, IFB-core, UMS 3601, Évry, France
[2] Université Paris-Saclay, INRAE, MaIAGE, 78350 Jouy-en-Josas, France
[3] Sorbonne Université, CNRS, FR2424, ABiMS, Station Biologique, Roscoff, France
[4] Université de Paris, CNRS, Institut Jacques Monod, F-75006, Paris, France
[5] Aix-Marseille Université, TGML, TAGC, INSERM, Marseille, France
[6] Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France
Corresponding Author: helene.chiapello@inrae.fr

The French Institute of Bioinformatics (IFB) provides a national infrastructure of bioinformatics services adapted to the needs of the life and health science communities, and accessible to all. However the appropriate use of these resources, in particular the network of HPC (High Performance Computer) clusters, requires a minimal level of computing knowledge that our users may not always possess. Aware of this need, the IFB proposes a wide range of training courses each year, based both on all the actions proposed by the IFB member and associated regional platforms, but also on actions carried out at a national scale in close collaboration with the NNCR task force. The most emblematic of those are the EBAII school [1], the DUBii university diploma [2], and the training courses around FAIR principles [3, 4].

The IFB initiated in 2020 the creation of a working group on the provision of e-learning bioinformatics resources at the national level. Three areas of work have been identified. The first one is a technical watch on digital environments dedicated to the provision of self-training resources including practical sessions adapted to our target audiences. The second concerns the development of a self-training resource on the basics of Unix, a need widely shared by the users of IFB services. Finally, the third axis is the provision of a unique national portal for the management of IFB resources, training courses, as well as the self-monitoring of the learner's path.

We present in this poster a first version of our e-learning resource entitled "Unix Introduction" [5] developed by the members of the working group and made available via the e-learning platform katacoda [6]. This resource proposes an interactive bioinformatics-oriented course that introduces beginners to the use of the Unix command line, a critical skill in the future usage of the IFB's computer resources. This material may then be proposed as a prerequisite for any training action using Unix resources.

We chose the katacoda platform to deliver this basic Unix training as it allows an interactive and progressive learning experience without necessitating anything more that the use of the common web browser: no software installation is required. All the learning scenarios are developed in Markdown and publicly accessible on the IFB GitHub account [7]. Three scenarios are currently online and have been proposed as prerequisites for the DUBii 2021 learners. Due to the good feedback they will also be proposed to the learners of the next session of the EBAII in November 2021.

References

1. https://www.france-bioinformatique.fr/ebaii/ [available the 2021/05/21]
2. https://www.france-bioinformatique.fr/dubii/ [available the 2021/05/21]
3. https://ifb-elixirfr.github.io/IFB-FAIR-data-training/index.html [available the 2021/05/21]
4. https://ifb-elixirfr.github.io/IFB-FAIR-bioinfo-training/ [available the 2021/05/21]
5. https://www.katacoda.com/ifb-elixirfr/courses/ifb-unix [available the 2021/05/21]
6. https://www.katacoda.com [available the 2021/05/21]
7. https://github.com/IFB-ElixirFr/katacoda-scenarios [available the 2021/05/21]

JeBif – Association for the Young Bioinformaticians of France

Emma Corre, Xavier Bussell, Slim El Khiari, Mathias Galati, Athénaïs Vaginay and Victor Grentzinger
Association des Jeunes Bioinformaticiens de France RSG France -- JeBiF,
4 rue des Arènes, 75005, Paris, France

Corresponding author: contact@jebif.fr

1 Abstract

JeBiF, which stands for "Jeunes Bio-Informaticiens de France" [Young Bioinformaticians of France], is a french non-profit organisation created in 2008. This association is the French branch of a bigger network handled by the International Society for Computational Biology Student Council (ISCB-SC). There are 26 other local branches. Such local branches are called Regional Student Groups (RSG), hence the full denomination of the association: RSG France – JeBiF.

The main goal is to promote the development of the next generation of bioinformaticians. In order to reach this goal, we mainly provide networking opportunities and career advice. We are also actively advertising computational biology and bioinformatics to general audience by the mean of science-popularisation events.

In the poster, we present with more details the different activities developed by RSG France – JeBiF. In particular:

— **JeBiF-Pub**: every month, in different cities, JeBiF sympathisers are invited to meet at a bar for a drink (paused for now due to sanitary constraints).
— **Table Ouverte en Bioinfo (TOBi)**: every month in a bar of Paris, a professional bioinformatician is invited to present their studies and career. Due to the sanitary constraints, TOBi have been replaced by **TOBirtuelles** which happen online and are opened to everyone, and not only Parisians.
— **Table Ronde**: once a year with the volunteering Master of Bioinformatics, several alumni of the Master are invited to present their path since they finished the Master. This year, due to the sanitary constraints, we replaced these local events by one big event organised with the support of the human resources team of the Pasteur Institute dedicated to career development and support for scientists (Mission Accueil, Accompagnement et Suivi des Carrières des Chercheurs – MAASQ). We received more than a hundred attendees.
— **JeBiF@JOBIM**: annual workshop with scientific presentations, open-table with various subjects, and flash-talks for people who have been accepted at JOBIM and would like to advertise their poster.
— **Fête de la science** and **Pint of Science**: two yearly events of science popularisation. JeBiF volunteers are encourage to create and animate activities, or give a talk.

The events proposed by RSG France – JeBiF are opened to everyone. The adhesion, which is *free* of charge, gives access to the mailing list and allows you to vote at the general assembly. It is also a way to quantify our impact. In particular, it gives us more weight in our funding applications. More funding allow us to maintain the association in long term, to propose more events and of larger scale. To date, JeBiF has 111 adherents. But its actions rely on the *participation* of its volunteers. We are always happy to meet new faces, and there is space for everybody to develop their ideas. Do not hesitate to join us if you would be part of the adventure!

JeBiF.RSG.France JeBiF rsg-france-jebif

Acknowledgements

RSG France – JeBiF's activities are funded among others by the GDR BIM and the ISCB. We are also thankful to all the volunteers who helped JeBiF this year, despite the sanitary context.

Index