JOBIM 2020

Montpellier, 30 juin - 3 juillet

Posters

Poster 4

SAMBA: a flexible automated workflow for the reproducible and standardized processing of eDNA metabarcoding data

Laure Quintric*[1], Cyril Noël*[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

* These authors contributed equally to this work

Corresponding Authors: samba-sebimer@ifremer.fr

Environmental DNA (eDNA) metabarcoding has become a powerful method for assessing the diversity and dynamics of microbial communities from various environmental samples. This approach notably involve in-depth bioinformatics and biostatistics analyses and interpretations of next generation sequencing data. Today, such metabarcoding approaches are carried out in many projects including marine observatory networks. However, this approach can remain complicated to use for researchers with a biological background and therefore requires the implementation of automated, standardized and user-friendly solutions.

In this context, we have developed a modular workflow called SAMBA (Standardized and Automated MetaBarcoding Analyses) using the NextFlow workflow manager [1]. SAMBA automates the analysis of metabarcoding data from any project by producing robust, reproducible and standardized results. It is built around three main parts: data checking, bioinformatics processes and statistical analyses. The SAMBA checking process allows to verify the integrity of the raw data. All SAMBA bioinformatics processes are mainly based on the use of the next-generation microbiome bioinformatics platform QIIME 2 [2] and on the approach of grouping sequences in ASV (Amplicon Sequence Variants) using DADA2 [3]. Biostatistical analyses are performed using the R package phyloseq [4]. SAMBA conducts alpha-diversity analysis (boxplot, barplot and table) using a set of diversity index (Observed, Chao1, InvSimpson, Shannon, Pielou) and beta-diversity analysis (ordinations, hypothesis testing) also using different ecological distances (Jaccard, Bray-Curtis, UniFrac and Weighted UniFrac). All of these statistical analyses are performed on data normalized using three different methods (rarefaction, DESeq2 and CSS). All the results are presented in an automatically edited html report. SAMBA offers a real alternative to the complex use of a suite of command line bioinformatics tools while providing access to state-of-the-art methods and tools in the field. It will be released in an IFREMER GitLab repository and is written with collaborative development and the continuous addition of features in order to meet users' needs.

References

1. Di Tommaso P., Chatzou M., Floden E.W., Barja P.P., Palumbo E. & Notredame C. (2017). Nextflow enables reproducible computational workflows. *Nature biotechnology*, **35**(4):316-319
2. Bolyen E., Rideout J.R., Dillon M.R., Bokulich N.A., Abnet C.C., Al-Ghalith G.A., *et al*. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature biotechnology*, **37**(8): 852-857
3. Callahan B.J., McMurdie P.J., Rosen M.J., Han A.W., Johnson A.J.A. & Holmes S.P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nature methods*, **13**(7): 581
4. McMurdie P.J & Holmes S. phyloseq: an R package for reproducible interactive analysisand graphics of microbiome census data. *PLoS One,* **8**(4):e61217

Poster 5

TimeNexus identifies dynamic pathways from gene expression time-series with temporal networks

Michaël PIERRELÉE[1], Fabrice LOPEZ[2], Laurent TICHIT[3] and Bianca HABERMANN[1]

[1] Aix Marseille Univ, CNRS, IBDM, Campus de Luminy, 13009 Marseille, France
[2] Aix Marseille Univ, INSERM, TAGC, Campus de Luminy, 13009 Marseille, France
[3] Aix Marseille Univ, CNRS, I2M, Campus de Luminy, 13009 Marseille, France

Corresponding Author: michael.pierrelee@univ-amu.fr

1 Introduction

According to common definition, a « biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in the cell » [1]. In other words, a stimulus leads to actions within the cell. The response to a stimulus comes from interactions between molecules within the cell. We model this system by a network in which molecules are nodes and their interactions are edges. The nodes have a status (active or not). This status evolves over time and may affect the status of interacting neighbors. This dynamic has to be considered when determining an active biological pathway.

Active biological pathways can be found using expression data, e.g. from RNA-sequencing. In a time-course experiment, a series of snapshots of gene expression is taken, enabling to detect all differentially expressed genes at a given time or condition. We can represent this change over time on interaction networks and extract these activated pathways within these networks.

Several algorithms exist to exploit these multidimensional data within networks. A standard approach is to calculate gene-wise correlations and draw a network from them [2]; but this model fails to represent the dynamics of the network. To overcome that, a method was proposed that categorizes genes according to the time of their highest fold change [3]. It then finds paths from the 'early' genes to the 'late' genes. However, this algorithm excludes the genes which are involved at other time points in the pathways.

2 TimeNexus applies temporal networks

In this project, we modeled the evolution of the cellular response over time as a temporal network by making use of multi-layer networks [4]. High-quality databases are used to build the initial gene-interaction network. Each layer in the multi-layer network represents the gene interactions at a given time. The layers differ because the set of active genes depends on the time, while layer structures are identical. As a layer at one time is determined by the layer at the previous time, there is a causality link between layers. We model this causality as directed "inter-layer edges". They connect the nodes from one layer to their counterpart in the next layer. We obtain active pathways by extracting active subnetworks from our multi-layer network. *TimeNexus* was developed as a Cytoscape app [5] and we are testing it with an RNA-sequencing dataset of the yeast cell-cycle with the goal to identify known genes involved in the cell-cycle as given by the dataset.

3 Conclusion

TimeNexus models the dynamics of cellular responses directly within the structure of the network, by considering that nodes change over time. This is contrary to other methods which loose this dynamics or only partially model it. Thus, TimeNexus enables users to determine dynamic pathways and visualize them.

References

[1] National Human Genome Research Institute (NHGRI). (2019). Biological Pathways Fact Sheet. [online] Available at: https://www.genome.gov/27530687/ [Accessed 13 Mar. 2019].
[2] van Dam, S., et al. (2017). Gene co-expression analysis for functional classification and gene–disease predictions. Brief. Bioinform. 19, bbw139.
[3] Patil, A., and Nakai, K. (2014). TimeXNet: identifying active gene sub-networks using time-course gene expression profiles. BMC Syst. Biol. 8, S2.
[4] Kivelä, M., et al. (2014). Multilayer networks. J. Complex Networks 2, 203–271.
[5] Shannon, P., et al. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res., 30:2498-2504.

Poster 9

Whole metagenome analysis with metagWGS

Joanna FOURQUET[1], Céline NOIROT[2], Christophe KLOPP[2], Philippe PINTON[3], Sylvie COMBES[1], Claire HOEDE[2], Géraldine PASCAL[1]

[1] GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France
[2] INRAE, UR875 MIAT, PF Bioinfo GenoToul, F-31326, Castanet-Tolosan, France
[3] Toxalim, Université de Toulouse, INRAE, ENVT, INP-Purpan, UPS, F-31027 Toulouse, France

Corresponding Author: geraldine.pascal@inrae.fr

Whole DNA shotgun sequencing of environmental samples allows to study their taxonomic composition and their functional profiles. However, the biological process from collecting data to sequencing and bioinformatics analysis are still very tricky [1].

We are developing a complete, scalable, easy-to-use and reproducible workflow, metagWGS, with Nextflow [2] and Singularity [3] that processes short Illumina reads from shotgun metagenomics data. It delivers (i) contig assemblies, (ii) syntactic and functional annotations of genes, (iii) taxonomic affiliations of reads and contigs, (iv) count table of reads per genes and (v) contig binning to obtain metagenome species.

The workflow begins by preprocessing steps that clean adapters, low quality reads and the host reads. We control the quality of the reads with FastQC [4]. The taxonomic classification of reads uses Kaiju [5] in order to have a first overview of reads. The assembly step uses metaSPAdes [6] or megahit [7] to generate contigs for each sample. These contigs are annotated by Prokka [8]. Then, with CD-HIT [9] we remove redundancy and generate a gene catalog by clustering ORFs at sample level and globally with a 95% sequence identity cutoff. We map reads back to contigs and we use featureCounts [10] to count the reads overlapping annotated genes. The raw count table gathers the number of reads aligned on each gene for each sample. We use DIAMOND [11] for the taxonomic affiliation of contigs versus nr database. We include contig binning processes from nf-core/mag pipeline. We generate a single result report with MultiQC [12].

MetagWGS is available on https://forgemia.inra.fr/genotoul-bioinfo/metagwgs. We will apply it on sequences from ExpoMycoPig project that aims to study gut microbiota of pigs exposed to mycotoxins [13].

Acknowledgements

ExpoMycoPig project and JF are funded by France Futur Elevage.

References

1. C. Quince et al. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol*. Vol 35, No 9, 2017.
2. P. Di Tommaso et al. Nextflow enables reproducible computational workflows. *Nat Biotechnol.* 35:316-319, 2017.
3. GM Kurtzer et al. Singularity: Scientific containers for mobility of compute. *PLoS One.* Vol. 12, 2017.
4. S. Andrews. FastQC. *Available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc*. 2010.
5. P. Menzel et al. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat Commun.* Vol. 7, No 11257, 2016.
6. S. Nurk et al. MetaSPAdes: a new versatile metagenomic assembler. *Genome Res.* 27:824-834, 2017.
7. D. Li et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics.* 31:1674-1676, 2015.
8. T. Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics.* 30:2068-2069, 2014.
9. L. Fu et al. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics.* 28:3150-3152, 2012.
10. Y. Liao et al. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics.* 30:923-930, 2014.
11. B. Buchfink et al. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*. 12:59–60, 2015.
12. P. Ewels et al. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics.* 32:3047-3048, 2016.
13. P. Pinton and I.P. Oswald. Effect of deoxynivalenol and other Type B trichothecenes on the intestine: a review. *Toxins*, 6:1615-1643, 2014.

Poster 12

RNA-seq Nextflow pipeline

Souhila AMANZOUGARENE[1], Alaa BADREDDINE[1], Lionel BERBERIAN[1], Mehdi DERHOURHI[1], Stefan GAGET[1], Amélie BONNEFOND[1] and Philippe FROGUEL[1]

[1] UMR INSERM/CNRS 1283/8199, (Epi)génomique Fonctionnelle et Physiologie Moléculaire Du Diabète et Maladies Associées. European Genomic Institute for Diabetes (E.G.I.D.), Pasteur Institute of Lille, University of Lille, Lille University Hospital, 1 place de Verdun, 59045 LILLE CEDEX, FRANCE

Corresponding Author: Souhila.amanzougarene@cnrs.fr

Next-generation sequencing (NGS) technology, including Transcriptome Sequencing (RNA-seq), Whole-Genome and Whole-Exome Sequencing (WGS/WES), is progressing rapidlyand generates huge amounts of data. The analysis of sequencing data is computationally intensive and involves multiple steps with several sequential operations using bioinformatics tools.
When analyzing this data, it is important to automate different steps and increase resource utilization efficiency, which can be achieved through parallel execution using an automated pipeline.
In order to meet this optimization objective, we have built an RNA-seq pipeline with Nextflow.

Nextflow is a bioinformatics workflow manager system that run tasks across multiple compute infrastructures in a portable manner and with high level parallelization [1].
Nextflow is easy to install, to launch and simplifies development of complex pipelines.
A Nextflow pipeline is made up by putting together several independent processes. Each process can be written in any scripting language (such as bash, python or perl).
Parallelization and task dependencies are implicitly defined by process input/output declarations. The output of a process can be passed as input to another process by using channels. Channels are unidirectional asynchronous queues that enable the processes to communicate and improves latency tolerance and third-party independence [1].

Here, we present the RNA-seq nextflow pipeline now routinely used in our laboratory. This pipeline manages differential gene expression analysis and SNP calling for Paired End and Single End data. It is a suite of 23 processes capable of managing large sets of NGS software. The script runs with slurm job schedulers system and can be deployed on clusters as well as on local machines.
This pipeline is launched into a single command line with few arguments (SampleSheet file, genome version).

The main steps of this RNA-seq pipeline consist of : ***demultiplexing*** (converts bcl files to fastq files), ***Quality Control*** on the raw data with fastQC, ***cutadapt*** (removes adapters and reads with low quality), ***mapping*** reads with RSEM tool using star aligner, ***differential gene expression*** analysis (DESseq2), ***mapping*** reads for calling SNP (star), ***Post-processing mapping*** (PicardTools/GATK4), ***base quality recalibration*** (GATK4), ***variant calling*** (GATK4 haplotypeCaller), ***annotation*** (Ensembl Vep), ***multiQC***, ***statistics***, where the majority of these steps are run in parallel.
If an error occurs, even after running the nextflow pipeline for hours or sometimes days, it is possible to resume on this error and start again at the last completed step.
With nextflow monitoring system (Nextflow Tower), we can track pipeline execution of jobs in real time by providing information with execution metrics (time, cores, CPU, memory) [1]. This allows us to better assist in the optimization of the pipeline. The complete analysis of 15 samples takes 1d 10h on a server with 500 GB and 120 CPU, and there is no limit to the number of samples to be analyzed simultaneously.
In addition, this pipeline offers an easy way to manipulate the various options of the different software or update them through the pipeline, either by modifying the parameters part at the beginning of the script or by using a single configuration file, which can be changed without having to change the code itself.

References

1. Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. Nature Biotechnology, 35(4), 316–319. doi:10.1038/nbt.3820

Poster 13

Exome and CNV Sequencing Analysis pipeline with Nextflow

Alaa BADREDDINE[1], Souhila AMANZOUGARENE[1], Lionel BERBERIAN[1], Mehdi DERHOURHI[1], Stefan GAGET[1], Amélie BONNEFOND[1] and Philippe FROGUEL[1]

[1] UMR INSERM/CNRS 1283/8199, (Epi)génomique Fonctionnelle et Physiologie Moléculaire Du Diabète et Maladies Associées. European Genomic Institute for Diabetes (E.G.I.D.), Pasteur Institute of Lille, University of Lille, Lille University Hospital, 1 place de Verdun, 59045 LILLE CEDEX, FRANCE

Corresponding Author: Alaa.Badreddine@cnrs.fr

As technology advances rapidly, it is now possible to output data in terabytes scale through a single run on next-generation sequencing (NGS) machines. Moreover, hospitals are steadily migrating towards the new era of e-health as it is becoming cheaper to sequence the genome of individuals whether they are inpatients or in ambulatory care. To keep pace with the huge amount of information generated and provide precise diagnostic report for patients, it is necessary to develop a new approach to maximize the use of resources and thus reduce the time needed to analyze the data.

Here, we report a novel pipeline that uses Nextflow[1] as backbone and integrates diagnostic analysis. Our pipeline, written from scratch, allows us to analyze a run from raw files from NGS systems to the final results such as the annotation of mutations, gene expression, identification of epigenetic markers, CNV detection, splicing predictions, mutation pathogenicity determination, quality checks, and plot generation while keeping tracks of every version of tools and input files used. Additionally, it works with the job scheduler and manager SLURM.

We chose Nextflow due to its multiple important aspects: reproducibility, parallelization, portability, coding languages integration and most importantly the ease of manipulating complex data by chaining processes altogether. Here, we bring a focus on *W*hole *E*xome *S*equencing (WES) and on *Co*py number variation *D*etection and *E*xome *seq*uencing[2] (CoDE-seq) pipelines.

For WES, the discovery of variants is made through HaplotypeCaller of GATK while performing all the necessary pre-processes for a sample. Then, VEP from Ensembl is used to annotate the variants while using various external databases such as dbNSFP, MasterMind, dbscSNV, GeneSplicer, MaxEntScan, SpliceRegion.

CoDE-seq has been developed in the laboratory to allow us the detection of CNVs along the mutations at the same cost of a regular WES. The pipeline is similar to WES and so does the analysis time, since the two additional processes for CoDE-seq are parallelized along the other running processes.

We are able to analyze an S4 flowcell (~ 150 samples) in less than three days on average while running on three servers with a total of 260 CPUs and 1 terabyte of RAM from raw files to final results. In total, 41 processes are executed in the pipeline, with some including connections to internal databases such as MySQL/MongoDB.

Moreover, it is possible to classify the mutations according to ACMG criteria and provide diagnostic information by using HGMD database which gives path to the state-of-art of Personalized Medicine.

References

1. P. Di Tommaso, et al. Nextflow enables reproducible computational workflows. Nature Biotechnology 35, 316–319 (2017) doi:10.1038/nbt.3820
2. Montagne, Louise, et al. "CoDE-seq, an augmented whole-exome sequencing, enables the accurate detection of CNVs and mutations in Mendelian obesity and intellectual disability." Molecular metabolism 13 (2018): 1-9.

Poster 14

HiC DOC: detecting and comparing genomic compartments

Cyril Kurylo[1], Sylvain Foissac[1] and Matthias Zytnicki[2]

[1] GenPhySE (INRAE), 24 Chemin de Borde Rouge, 31320, Auzeville Tolosane, France
[2] MIAT (INRAE), 24 Chemin de Borde Rouge, 31320, Auzeville Tolosane, France

Corresponding Author: **cyril.kurylo@inrae.fr**

Genomic compartmentalization is a biological factor affecting cell functionality. Different compartments can be observed in the nucleus of eukaryotic cells, grouping genomic regions into clusters. Active compartments are usually associated with open chromatin and gene expression while inactive compartments are usually associated with closed chromatin and gene repression [1]. Analysis of data produced by the Hi-C protocol reveals compartmentalization of chromatin in the nucleus, which can vary as a tissue develops. Today, existing methods to detect genomic compartmentalization are limited in at least one of the following ways: detecting compartments qualitatively with no confidence measure, ignoring experimental biases, and/or dismissing replicate variability.

We propose an improvement over existing methodology to detect compartments and compare compartmentalization between conditions. First, we properly correct the diverse biases inherent to Hi-C data, using cyclic loess normalization [2] to reduce technical biases and Knight-Ruiz matrix balancing [3] to mitigate biological biases. Then, we correct interaction counts with a loess regression to clearly expose the compartmentalization information captured by the data. Finally, we use an unsupervised learning method, constrained K-means [4], to computationally detect compartments from the normalized data. This method enables us to produce quantitative "concordance" values for each genomic region in each replicate, supporting our compartment predictions. Finally, we use these concordance values for differential analysis of compartmentalization between conditions. From their distributions, we obtain p-values revealing the significance of each predicted compartment change.

The method is implemented in an R package available at *github.com/mzytnicki/HiCDOC*, and is applied to Hi-C data originating from fetal pig muscles. Our data consists of three biological replicates at 90 days of gestation and three biological replicates at 110 days of gestation [5]. The detected compartment changes open a way towards a better understanding of neonatal mortality affecting piglets.

References

[1] Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, Michael A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, (326):289–293, 2009.

[2] John C Stansfield, Kellen G Cresswell, and Mikhail G Dozmorov. multiHiCcompare: joint normalization and comparative analysis of complex Hi-C experiments. *Bioinformatics*, (35):2916–2923, 2019.

[3] Philip A Knight, and Daniel Ruiz. A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis* (33): 1029–1047, 2012.

[4] Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schrödl. Constrained K-means Clustering with Background Knowledge. In *18th International Conference on Machine Learning*, pages 577–584. ICML, 2001.

[5] Maria Marti-Marimon, Hervé Acloque, Matthias Zytnicki, Sarah Djebali Quelen, Nathalie Villa-Vialaneix, Ole Madsen, Yvette Lahbib Mansais, Diane Esquerre, Florence Mompart, Laurence Liaubet, Martien Groenen, Martine Bouissou-Matet Yerle, and Sylvain Foissac. Characterization of 3D genomic interactions in fetal pig muscle. In *36th International Society for Animal Genetics Conference*, page 43. ISAG, 2017.

Poster 15

Metage2Metabo: metabolic complementarity applied to genomes of large-scale microbiotas for the identification of keystone species

Arnaud Belcour*[1], Clémence Frioux*[1,2,3], Méziane Aite[1], Anthony Bretaudeau[1,4,5], Falk Hildebrand[1,6] and Anne Siegel[1]

[1] Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France
[2] Gut Microbes and Heath, Quadram Institute, NR4 7GJ, Norwich, United Kingdom
[3] Inria Bordeaux Sud-Ouest, France
[4] INRA, UMR IGEPP, BIPAA, 35042 Rennes, France
[5] Inria, IRISA, GenOuest Core Facility, 35042 Rennes, France
[6] Digital Biology, Earlham Institute, NR4 7GJ, Norwich, United Kingdom

Corresponding Author: clemence.frioux@inria.fr

Recent advances in metagenomics and amplicon sequencing enables the characterisation of microorganisms inhabiting environments. Yet, understanding their roles in the large communities forming microbiotas is a more difficult task. Metabolic modelling is widely applied to individual organisms and small communities [1] but no method addresses the scaling to large microbiotas. In this work we propose a workflow, Metage2Metabo (M2M), to build genome-scale metabolic networks (GSMNs) for communities of hundreds or thousands of microorganisms in order to identify keystone species among them.

M2M encompasses a wrapper for Pathway Tools [2] orchestrating the automatic and parallel reconstruction of GSMNs starting from annotated genomes. Once reconstructed, individual GSMNs are analysed before being considered collectively [3]. This enables the identification of the added-value of cooperation in a large community, that can be used to decipher keystone species. The objective for community selection is a set of metabolites whose production has to be ensured by the selected communities. It by default the added-value of cooperation but can be customised.

We applied M2M to a set of 1,520 reference genomes from the gut microbiota and 913 metagenomic-assembled genomes of the cow rumen. For each dataset we select minimal communities enabling the production of compounds requiring metabolic cooperation. By using logic programming, we can efficiently explore the whole search space of solutions and suggest keystone species. The latter are distinguished between essential symbionts, present in every optimal community, and alternative ones. We enumerated all minimal communities associated to the gut bacteria to characterize associations between keystone species using graph compression methods, enabling an efficient visualisation of thousands of communities.

Altogether, M2M performs a thorough comparison of the metabolism of symbionts in large communities. It aims at reducing the complexity of communities by suggesting important members. Each step of the pipeline can be run independently, ensuring the flexibility of the proposed protocol.

Acknowledgements

The authors acknowledge the GenOuest bioinformatics core facility for providing the computing infrastructure and the Bioinformatics Research Group of SRI International for their help regarding Pathway Tools. This research was supported in part by the NBI Computing infrastructure for Science (CiS) group.

References

[1] Bosi E, Bacci G, Mengoni A, Fondi M. Perspectives and Challenges in Microbial Communities Metabolic Modeling. Front Genet 2017;8:88. https://doi.org/10.3389/fgene.2017.00088.

[2] Karp PD, Latendresse M, Paley SM, Krummenacker M, Ong QD, Billington R, et al. Pathway tools version 19.0 update: Software for pathway/genome informatics and systems biology. Brief Bioinform 2016;17:877–90. https://doi.org/10.1093/bib/bbv079.

[3] Frioux C, Fremy E, Trottier C, Siegel A. Scalable and exhaustive screening of metabolic functions carried out by microbial consortia. Bioinformatics 2018;34:i934–43. https://doi.org/10.1093/bioinformatics/bty588.

Poster 16

MetIDfyR, an open-source R script to decipher small-molecule drug metabolism through high resolution mass spectrometry: Application in doping control

Agnès BARNABÉ, Vivian DELCOURT, Benoît LOUP, Patrice GARCIA, François ANDRÉ, Benjamin CHABOT, Yves MOULARD, Marie-Agnès POPOT and Ludovic BAILLY-CHOURIBERRY

GIE LCH, 15 rue de Paradis, 91370, Verrières-le-Buisson, France

Corresponding Author : a.barnabe@lchfrance.fr

Once administered to humans or animals, small-molecule drugs are modified by the endogenous enzymatic machinery, available in the liver, through phase I and phase II metabolism, enabling compounds elimination. Regarding the extensive variety of drug chemistries, metabolites prediction can be complex to achieve. It leads to a challenging detection in biological matrices for drug discovery and development [1], drug-testing analysis [2,3], forensics and pharmaco-kinetic studies [4]. Usually, metabolites from new molecules are often predicted and sought manually in samples processed by LC-HRMS/MS methods. However, manual processing is time-consuming and can lead to errors especially in case of complex and/or consecutive bio-transformations. *In-silico* metabolism prediction tools already exist, some are freely available such as Biotransformer or SyGMa for humans and others are commercial such as Mass-MetaSite. Since there is no freely available and customizable solution in the wide R-mass spectrometry toolbox, there is a need to develop a tool for *prediction*, *detection* and *evaluation* of metabolites from LC-HRMS/MS data.

Here, we present MetIDfyR, an *open-source*, *cross-platform* and *versatile R script* for *in-silico* drug phase I/II bio-transformations prediction and automated mass spectrometry data mining. Based on the raw formula of the parent drug of interest and few parameters, MetIDfyR can detect the small-molecules and their putative metabolites from LC-HRMS/MS data from either *in-vitro* or *in-vivo* drug studies. This software uses an universal file format (mzML) allowing the data-processing from a wide range of mass spectrometry technologies.

In order to assess the efficiency of MetIDfyR, model compounds have been identified as compounds thoroughly studied from the scientific literature. Metabolite identifications have been performed on those model compounds from *in-vitro* and *in-vivo* experiments. This software successfully identified all known metabolites present in the extract, even for compounds undergoing extensive metabolization. MetIDfyR demonstrated a reliable open-source solution for prediction and detection of metabolites to assist the scientific community.

References

1. J. Kirchmair *et al*. Prediction drug metabolism: experiment and/or computation? *Nature reviews Drug discovery*, (6):387-404, 2015.
2. J.P. Scarth *et al*. Drug metabolism in the horse: a review. *Drug Testing and Analysis*, (1):19-53, 2011.
3. Y. Moulard *et al*. Use of benchtop exactive high resolution and high mass accuracy orbitrap mass spectrometer for screening in horse doping control. *Analytica chimica acta*, (1/2):126-136, 2011.
4. M-A. Popot *et al*. Pharmacokinetics of tiludronate in horses: A field population study. *Equine veterinary journal*, (4):488-492, 2018.

Poster 19

MEMHDX v2: implementation of new functionalities and redesign of MEMHDX - An interactive tool to expedite the statistical validation and visualization of large HDX-MS datasets.

Stevenn Volant[1], Rachel Torchet[1], Véronique Hourdel[2] and Sébastien Brier[3]

[1]Hub de Bioinformatique et Biostatistique – Département Biologie Computationnelle, Institut Pasteur, USR 3756 CNRS, Paris, France

[2]Environment and Infectious Risks Unit, Department of Infection and Epidemiology, Institut Pasteur, Paris, France.

[3]Biological NMR Technological Platform, Center for Technological Resources and Research, Department of Structural Biology and Chemistry, Institut Pasteur, CNRS UMR3528, Paris, France

Corresponding Author: stevenn.volant@pasteur.fr

Hydrogen/Deuterium exchange measured by Mass Spectrometry (HDX-MS) is a powerful biophysical approach able to probe the structure and conformations of proteins and complexes. HDX-MS is currently combined with other classical structural tools such as X-ray, SAXS, EM or NMR, to complement the general structural picture of biological systems. Over the last decade, numerous HDX-MS technical challenges have been addressed and some of them have been met. HDX-MS is now able to handle very complex biological systems hence increasing the complexity and the size of datasets. It is therefore no more conceivable to manually analyze datasets acquired with modern HDX-MS protocols.

In 2016, we developed MEMHDX (Mixed-Effect Models for HDX), a web application to help users analyze, validate and visualize very large HDX-MS datasets acquired in two distinct experimental conditions [1]. The application is based on a linear mixed-effects model for each peptide and allows to analyse simultaneously changes in dynamics and magnitude, taking into account the time dependency of the exchange reaction. In some cases, however, the linearity assumption appears too strong and the mathematical model does not fit well to the data.

Herein, we implemented two new modelisations (log and power) that are more suitable for complex dynamics to improve the quality of the fitting. For each peptide, the best model is automatically selected using the log-likelihood. We used User Centered Design methodologies to integrate those new functionalities and improve user experience. We conducted different rounds of stakeholder and user interviews which helped us define UX milestones and objectives. The resulting new version of MEMHDX provides more statistical options such as selection of the model and multiple corrections methods. Furthermore two experimental conditions are now allowed and peptides with multiple charge states are handled. We make the most of this update to enhance some graphical elements and visualization tools.

MEMHDX is freely available as a web tool at the project home page http://memhdx.c3bi.pasteur.fr

References

1. Hourdel V, Volant S, O'Brien DP, Chenal A, Chamot-Rooke J, Dillies MA, Brier S. *MEMHDX: An interactive tool to expedite the statistical validation and visualization of large HDX-MS datasets*, 2016.

Poster 20

In silico characterization of the gene repertoires of immunoglobulins and T cell receptors of the various inbred laboratory strains of *Mus musculus*

Anna Tran, Géraldine Folch, Véronique Giudicelli and Sofia Kossida
Institut de Génétique Humaine, IGH, UMR 9002, CNRS - Université de Montpellier, 141 rue de la Cardonille, 34090, Montpellier, France

Corresponding author: anna.tran@igh.cnrs.fr

Abstract

The laboratory mouse is the most widely used animal model in the life sciences for the study of disease and human development. Mouse strains are known for their differences in the adaptive immune response[1], but the genomic repertoires of genes that code for antigen receptors, immunoglobulins or antibodies (IG) and T cell receptors (TR) are far from having been fully and precisely sequenced and/or characterized in each strain despite the existence of Mouse Genome Informatics resources dedicated to the species [2].

IG (proteins composed of 2 heavy chains or IGH, and 2 light chains IGK or IGL) and TR (composed of chains Alpha and Beta, or chains Gamma and Delta) are encoded by 4 types of genes, variable (V), diversity (D), joining (J), constant (C) belonging to multigene families and are very polymorphic. The synthesis of these molecules results from complex mechanisms including rearrangements of the V, D and J genes at the DNA level, the mechanisms of N-diversity and, for IGs, of somatic hypermutations. These mechanisms are at the origin of an extreme diversity of IG and TR (potentially more than 2.10^{12} IG and 2.10^{12} different TR per individual) and the effectiveness of the adaptive immune system.

Knowing and understanding the organization of these repertoires in the different strains is therefore essential for understanding the reactions of the adaptive immune system and for the choice of mouse models in biology. For example on IGH locus, the most widely used inbred strains C57BL/6 and BALB/c have only a few sequences in common, which means that their IGH locus are probably a mosaic of very disparate genes. It is highly probable that the same holds true for the loci of other inbred strains of mice. It is important to document this diversity in order to understand the variation within as well between strain models of antibody-mediated diseases, among other things[3].

IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org)[4], is a unique source of knowledge in immunogenetics and immunoinformatics and is recognized as the international reference. IMGT® engages in the precise and detailed characterization of the IG and TR loci by mouse strain according to IMGT® standards in order to establish their genomic repertoires and allow their comparison. This thesis aims to design and develop and/or adapt high-performance software tools and a methodology which implement the standards and carry out the annotation of the loci IG and TR of the mouse strains with a "Gold standard" quality (equivalent to the manual annotation). This will allow enrichment of IMGT® databases and implementation of strain-specific research and analysis in IMGT® software tools.

References

[1] RS Sellers, CB Clifford, PM Treuting, and Cory Brayton. Immunological variation between inbred laboratory mouse strains: points to consider in phenotyping genetically immunomodified mice. *Veterinary pathology*, 49(1):32–43, 2012.

[2] Cynthia L Smith, Judith A Blake, James A Kadin, Joel E Richardson, Carol J Bult, and Mouse Genome Database Group. Mouse genome database (mgd)-2018: knowledgebase for the laboratory mouse. *Nucleic acids research*, 46(D1):D836–D842, 2018.

[3] Corey T Watson, Justin T Kos, William S Gibson, Leah Newman, Gintaras Deikus, Christian E Busse, Melissa L Smith, Katherine JL Jackson, and Andrew M Collins. A comparison of immunoglobulin ighv, ighd and ighj genes in wild-derived and classical inbred mouse strains. *Immunology and cell biology*, 97(10):888–901, 2019.

[4] Marie-Paule Lefranc. Immunoglobulin and t cell receptor genes: Imgt® and the birth and rise of immunoinformatics. *Frontiers in immunology*, 5:22, 2014.

Poster 21

VCFProcessor: a complete toolbox for improved VCF file analysis

Thomas E. Ludwig[1,2], Gaëlle Marenne[1] and Emmanuelle Génin[1,2]
[1] Inserm, Univ Brest, EFS, UMR 1078 GGB, F-29200 Brest, France
[2] CHU Brest, F-29200 Brest, France

Corresponding author: thomas.ludwig@inserm.fr

1 Introduction

The VCF file format is the standard for storing genetic variants. Like VCFtools [1] and BCFtools [2], VCFProcessor handles such files and proposes several variant selection filters and functions to transform or annotate a VCF file. More than that, VCFProcessor offers a complete set of tools that are relevant when working on Human data.
VCFProcessor is implemented in Java and will run on Unix based system as well as Windows, making use of multiple processing cores when available. It requires no installation nor configuration and is available at: `http://lysine.univ-brest.fr/VCFProcessor/`

2 Variant Filtering

VCFProcessor proposes 150 command line arguments, that can be combined to finely pick desired variants from a VCF file. Those arguments trigger filters selecting variants on their position, local frequency or properties, as well as selecting only certain samples or setting genotypes to *missing* depending on the value of their annotations. Every command line filters from VCFtools and BCFtools are implemented as well as several new ones.

3 Analyses

VCFProcessor relies on the concept of *Functions*. Those functions are of several types: VCF Annotations that add information to a VCF file, VCF Transformations that modify values within the file, VCF Filters that perform complex filtering operations. The main category of functions is called Analysis. Those analyses pertain to quality control, case control data comparison and population genetics. Several dozens of functions are implemented to performs various analyses, such as comparing VCF files, measuring the quality of imputation, getting summary statistics per variant or per sample, measuring differences between groups of samples, measuring frequency distributions...
Furthermore, VCFProcessor is flexible and offers the possibility to users to develop their own functions, when a particular need is not answered by the preexisting ones.

4 Visualization Tools

Most analyses produce complex table-like results that need to be visualized in a particular manner. VCFProcessor proposes a set of tools to produces adapted graphs in the PNG format from those outputs.

5 Graphical User Interface

VCFProcessor was mainly developed as a Unix tool, thus allowing to pipe its output to subsequent programs. A more user-friendly version of the program is also available that uses a graphical user interface (GUI) and puts all VCFProcessor tools at the click of the mouse.

References

[1] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, and Richard Durbin. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, August 2011.

[2] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–2079, August 2009.

Poster 22

The BIBS service (Bioinformatics and Biostatistics) at the Centre International de Recherche en Infectiologie (CIRI)

Marie Cariou[1], Omran Allatif[1], Antoine Corbin[1] and The BIBS Steering Committee*
CIRI, Inserm U1111, CNRS UMR5308, Université Claude Bernard Lyon 1, École Normale Supérieure de Lyon, Univ Lyon, F-69007, Lyon, France

Corresponding author: marie.cariou@ens-lyon.fr

The **BIBS (Bioinformatics and biostatistics) service** is a platform of the CIRI (*Centre International de Recherche en Infectiologie*) in Lyon. Its aim is to provide support to the research teams regarding bioinformatics, business intelligence, statistics and data analysis.

The CIRI was created on 1st January 2013 by the **Inserm**, **CNRS**, **ENS Lyon** and **University Claude Bernard Lyon 1**. It comprises over 20 teams behind one goal: the fight against infectious diseases. The ambition of the CIRI is to promote an integrated approach on a broad range of host-pathogen interactions. It aims at being internationally recognized as a top-level center dedicated to fundamental research on **microbiology** and **immunology**, and to translate scientific discoveries into clinical or industrial applications, at least up to early proofs of concept.

In this context, the missions of the BIBS service involve:

1. Data analysis for the research teams : The service provides expertise in bioinformatic data processing, notably from high-throughput sequencing technics, and biostatistics (e.g. multivariate analyses, data mining, omics data analysis, descriptive and inferential statistics, data visualisation)

2. Consulting and support : The service supports teams from experiment design to analysis choices, through short or long term collaborations.

3. Training : Courses and tutorials can also be provided following the needs of the researchers: data handling, database interrogation, programming or use of common tools and pipelines.

In its missions, the service benefits from close relationships with the other research units in Biology at the ENS Lyon, **LBMC** (*Laboratoire de Biologie et Modélisation de la Cellule*), **IGFL** (*Institut de Génomique Fonctionnelle de Lyon*) and **RDP** (*Reproduction et Developpement des Plantes*), through frequent **joint work sessions and meetings**. It also relies on the support of the **high performance computing facilities** of the ENS Lyon, the *Pôle Scientifique de Modélisation Numérique* (PSMN) and Blaise Pascal Center (CBP).

This poster will provide an overview of the activities of the BIBS service, whose aim is to favor the deployment of bioinformatics to support various infectious disease and immunology research programs of the CIRI.

*The steering committee of the BIBS service: Nathalie Alazard-Dany, Francois-Loïc Cosset, Lucie Etienne, Christophe Ginevra, Marie-Paule Gustin, Jacqueline Marvel, Helena Paidassi, Thierry Walzer and Emiliano Ricci.

Poster 23

A pipeline for quality control of target-enrichment DNA sequencing data

Chloé Beaumont[1], Sophie Gallina[1], Vincent Castric[1] and Mathieu Genete[1]

[1] CNRS, Univ. Lille, UMR 8198 - Evo-Eco-Paleo, F-59000 Lille, France

Corresponding Author: chloe.beaumont@univ-lille.fr

The characterization of sequence variations has been accelerated by high throughput sequencing technologies (NGS). However, achieving sufficient read coverage with whole-genome sequencing is costly and time-consuming. Hence, target-enrichment strategies for NGS are commonly used to efficiently access to genomic regions of interest. DNA-RNA hybridization, an effective tool to capture specific genomic sequences, consists in using synthetic RNA oligonucleotide probes linked to magnetic beads, corresponding to the sequences of interest, thereby allowing DNA fragments from genomic libraries to hybridize based on sequence complementarity. Hybridized DNA molecules are then captured, amplified and sequenced by NGS sequencing.

Several tools exist to control the quality of sequence[1,2], however, target-enrichment data need specific quality controls to evaluate whether the capture of regions of interest was successful. In order to meet the routine needs of the laboratory for specific sequences capture, we have developed an automated quality control pipeline to measure enrichment based on several metrics such as the enrichment factor [3], the number of reads aligned to the targets over the total number of reads sequenced or over the total number of reads aligned.

Our pipeline uses reads aligned to an input reference genome in which repeated regions have been masked. We eliminated duplicated reads and counted reads aligned on regions of interest and reads aligned on the total of the reference genome in order to produce a quality control report with data summary including the total number of individuals, the total number of reads per individuals, the size of the covered targets, the size of targets without repeated regions, the reference coverage rate, the enrichment factor, the coverage plot, the depth plot and issuing warnings on outliers for metrics like coverage, depth, total number of reads, targets coverage percentage. This pipeline is designed to be used on a large number of individuals. The efficiency of our quality assessment pipeline was estimated based on simulated and real capture data. The simulated data are used to assess accuracy of the pipeline in specific situations such as when coverage is low or when there is variation in size and structure of the regions of interest.

[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] Chen, S. *et al.* AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. *BMC Bioinformatics* 18, 80 (2017).

[3] Dapprich, J. *et al.* The next generation of target capture technologies - large DNA fragment enrichment and sequencing determines regional genomic variation of high complexity. *BMC Genomics* 17, 486 (2016).

Poster 24

Exome sequencing allows detection of relevant pharmacogenetics variants in epileptic patients.

Simon VERDEZ[1,2], Philippine GARRET[1,4] , Emilie TISSERANT[1,2] , Céline VERSTUYFT[6] , Antonio VITOBELLO[1,2] , Frédéric TRAN MAU-THEM[1,2] , Christophe PHILIPPE[1,2] , Marc BARDOU[5] , Maxime LUU[5] , Juliette ALBUISSON[3] , Abderamem BOURREDJEM[5] , Patrick CALLIER[1,2] , Christelle THAUVIN-ROBINET[1,2] , Nicolas PICARD[7] , Laurence FAIVRE[1,2] and Yannis DUFFOURD[1,2]

[1] Univ. Bourgogne Franche-Comté, UMR 1231 GAD team, Genetics of Developmental disorders, F-21000 Dijon, France

[2] INSERM, UMR 1231 GAD team, Genetics of Developmental disorders ; FHU TRANSLAD, F-21000 Dijon, France ;Centre de référence Anomalies du Développement et Syndromes Malformatifs ; UF Innovation en diagnostic génomique des maladies rares ;GIMI Institute, F-21000 Dijon, France

3 CGFL, GIMI Institute, F-21000 Dijon, France

4 Laboratoire CERBA, Saint-Ouen l'Aumone, France

5 CIC, Centre Hospitalier Universitaire et Université de Bourgogne-Franche Comté,Dijon,France

6 Service de génétique moléculaire, pharmacogénétique et hormonologie, Hopital Bicêtre, Groupe Hospitalier Paris-Sud, AP-HP, Le Kremlin Bicêtre (France) univ Paris-Sud Unité UMR 1184 Faculté de médecine

7 Inserm U1248, service de pharmacologie et toxicologie, université de Limoges, CHU de Limoges, 87042 Limoges, France

Corresponding Author: simon.verdez@chu-dijon.fr

Abstract

Each genome sequence is specific to each individual and numerous variants are potentially relevant for clinical care. Relevant variants can be divised into two separate groups, those which are suspected of responsible for the initial presentation (primary variants) and the others (secondary variants). Detection of secondary variants can thus be conducted to improve future health outcomes in patients and their at-risk relatives, such as predicting late-onset genetic disorders accessible to prevention or treatment or identifying differential drug efficacy and safety. Pharmacogenetic is the study of link between drug and genetic variants, many examples of pharmacogenetic interaction are described in literature [1]. These studies consider the possibility of detecting secondary pharmacogenetic variants before administration of drugs.

To evaluate the interest of pharmacogenetics information, we designed an "in house" pipeline to determine the status of 122 PharmGKB (Pharmacogenomics Knowledgebase [2]) variants in 31 genes, including structural variants and HLA genotyping. Three tools were specifically used during this work, SnpSift, xHMM and HLAscan. This pipeline was applied on a cohort of 90 epileptic patients who had a previous exome sequencing (ES) to determine the frequency of pharmacogenetics variants of interest. Retrospective analysis of plasma concentrations and treatment efficacy of those which had been administered in the presence of at least one relevant PharmGKB variant has been performed.

We extracted 1717 substitution variants described in PharmGKB, 7 Copy Number Variations involving genes of interest and determined 360 alleles for HLA groups. CYP2C9 status for phenytoin's prescription was the only relevant information. 19/90 patients were treated by phenytoin in our cohort. Among the patients treated, four intermediate metabolizers and zero low metabolizers were reported. While being treated with a standard protocol , each identified intermediate metabolizers was associated with a plasma concentration above the toxic range.

In summary, based on the current knowledge and the result of this study, there is evidence that justify the detection and the analysis of pharmacogenetic variants belonging to incidental findings. This study shows that we can extract more informations from the initial data with appropriate bioinformatics method dedicated to determine alleles for genes of interest.

References

[1] Caudle KE, Rettie AE, Whirl-Carrillo M, Smith LH, Mintzer S, Lee MTM, et al. Clinical Pharmacogenetics Implementation Consortium Guidelines for CYP2C9 and HLA-B Genotypes and Phenytoin Dosing. Clin Pharmacol Ther. nov 2014;96(5):542-8.

[2] Alan D Sokal. Transgressing the boundaries: Toward a transformative hermeneutics of quantum gravity. *Social text*, (46/47):217–252, 1996.

Poster 25

Pipeline for ribosome profiling analysis (RiboSeq)

Pauline François[1], Stéphane Demais[1] and Olivier Namy[1]

[1] Institute for Integrative Biology of the Cell (I2BC), UMR 9198 CEA, CNRS, Univ. Paris Sud, Bâtiment 400, 91405 Orsay, France.

Corresponding Author : olivier.namy@i2bc.paris-saclay.fr

1. Introduction

Works about genetic expression are frequently restricted to transcriptome studies. Even if we have learned a lot about transcriptional regulations thanks to this technique, a poor correlation between mRNA and protein levels (~ 40%) was revealed by mass spectrometry approaches [1]. This difference can be explained especially by translational regulations which would be the most predominant way to modify the intracellular protein level [2]. In 2009, a new technique named Ribosome profiling (RiboSeq) was published [3]. This allows the genome-wide analysis of translation. It consists in isolating and sequencing ribosome footprints called ribosome protected fragments or RPF. Although this is an omic approach, RPFs are obtained with a high precision. This gives us the opportunity to map the position of each ribosome at one nucleotide resolution. So it is possible to perform a precise translation qualitative analysis (uORF mapping, identification of the reading frame).

RNAseq bioinformatics worflows are well described and implemented. However, the RiboSeq bioinformatics analysis face several specific challenges that must be properly addressed to avoid misinterpretations. Unfortunately there is currently no gold-standard workflow to guide users during their RiboSeq analysis. This leads to issues in reproducibility and also in misinterpretations. Indeed, several studies do not show quality controls which are necessary to prove that the signal come from bona-fide active ribosome and not from ribonucleoproteins fixed on mRNA [4]. Two methods can be used : i) demonstrating an increased signal in CDS vs UTRs; ii) identifiying a 3 nucleotides periodicity signal (ribosomes move codon by codon). Finally, due to the non homogeneous repartition of RPF alongside the genome, it is necessary and important to have a high sequencing depth.

Since eight years, our team realizes and analyzes RiboSeq data. After around twenty RiboSeq data analysis in different organisms like yeasts, plant, virus and human cells [4, 5, 6, 7] we want to share this knowledge with the french bioinformatics community. Our poster will show the good practices for a proper RiboSeq analysis.

2. Citations

[1] Vogel C, et al. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet. 2012 Mar 13;13(4):227-32.
[2] Schwanhäusser B, et al. Global quantification of mammalian gene expression control. Nature. 2011 May 19;473(7347):337-42.
[3] Ingolia NT, et al. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009 Apr 10;324(5924):218-23.
[4] Planchard N, et al. The translational landscape of Arabidopsis mitochondria. Nucleic Acids Res. 2018 Jul 6;46(12):6218-6228.
[5] Blanchet S, et al. Deciphering the reading of the genetic code by near-cognate tRNA. Proc Natl Acad Sci U S A. 2018 Mar 20;115(12):3018-3023.
[6] Legendre R, et al. RiboTools: a Galaxy toolbox for qualitative ribosome profiling analysis. Bioinformatics. 2015 Aug 1;31(15):2586-8.
[7] Baudin-Baillieu A et al. Genome-wide translational changes induced by the prion [PSI+]. Cell Rep. 2014 Jul 24;8(2):439-48.

Poster 26

Automated solutions for big genomic data treatment in the context of the medical diagnosis platform SeqOIA

Adrien LEGENDRE[1], Virginie SAILLOUR[1], Mario NEOU[1], Nicolas DERIVE[1], Camille BARETTE[1,2], Mouhamadou NIANG[1], Jocelyn BRAYET[2] and Alban LERMINE[1,2]

[1] SeqOIA-IT, 33 bd Picpus, 75012, Paris, France

[2] MOABI, AP-HP Bioinformatics platform, 33 bd Picpus, 75012, Paris, France

Corresponding Author : adrien.legendre-ext@aphp.fr

The SeqOIA (Sequencing, Omics, Information Analysis) project is part of the French national plan called : “Plan France Medecine Genomique 2025” (FMG 2025) and supported by the public assistance – Hopitaux de Paris (AP-HP), Curie institute and Gustave Roussy institute. The aim of FMG 2025 project is to establish high-throughput sequencing platforms in order to promote the access conditions for genetic data in terms of diagnosis, prognosis and therapeutic in France. There is also scientific and technological goals, in order to have a better understanding of pathologies such as cancer and rare diseases. The SeqOIA-IT platform is one of the two platforms selected in the FMG 2025 project. In terms of data volume, the SeqOIA platform will analyze 18 000 equivalent genome sequencing data each year. This platform will also integrate its own data storage, data calculation infrastructure, prescription and visualization solution.

In this abstract, we will present our solution to handle a big amount of genomic data in a context of medical diagnosis platform. We will begin to briefly describe our prescription solution which allows us to recover patients informations’s. In a second part, we will focus our poster on interoperability solutions and pipeline description, in order to show the way we choose to treat reproducibility and automation challenges in the framework of big data treatment. Our interoperability solutions allows us to recover all needed informations such as: family pedigree, HPOs (Human Phenotype Ontology), sequencing informations, to run our analysis pipelines. Those informations will be used and sent to a Redis [1] database in order to automate our Whole Genome Sequencing Rare Disease pipeline launching. This pipeline, as all pipelines of the project fit within the scope of reproducibility and stability. As a consequence, we choose to use several technologies which we are going to describe :

→ Docker [2] which allows us to eliminate tools dependency issues and allow us to set a version tool in an image, then each Docker image is created for a specific tool’s version.

→ Snakemake [3] is a workflow management system based on implicit rule implementation (input/output logic) which allow us to handle automatic chromosome parallelization and sequencing well parrallelization during alignment phase to enhance pipeline’s running time.

→ GitLab-CI [4] is a tool built into GitLab [5] for software development through the continuous methodologies. It provides us with continuous integration features and builds, automatic run test of the Docker container for each new tool.

The workflow is composed of 40 different steps (about 2000 jobs for a trio) allowing to perform the alignment step (BWA), complete quality control (FastQC, SAMtools, PicardTools), variant calling (GATK4) and annotation steps using public database (SNPEff, SNPSift).

Finally, the VCF (Variant Call Format) will be imported into our solution for visualization and help in interpreting results called Gleaves. This will allow us to analyze the equivalent of a NovaSeq run in less than 30h which represent 24 genomes or 8 trio. This poster is a typical example of those technologies utilization in order to emphasize the automation of information treatment and pipelines launching in the context of big genomic data analysis.

References

[1] Web site : https://redis.io

[2] Web site : https://www.docker.com/

[3] Johannes Köster, Sven Rahmann ; Snakemake a scalable bioinformatics workflow engine. Bioinformatics 2012 ; 28 (19) : 2520-2522. doi : 10.1093/bioinformatics/bts480

[4] Web site : https://about.gitlab.com/gitlab-ci/

Poster 28

The Migale bioinformatics facility

Valentin Loux[1,2],Mouhamadou Ba[1,2] , Damien Berry[1,2], Hélène Chiapello[1,2], David Christiany[1,2], Sandra Dérozier[1,2], Olivier Inizan[1,2], Mahendra Mariadassou[1,2], Véronique Martin[1,2], Cédric Midoux[1,2,3], Olivier Rué[1,2], Claire Vincent[1,2,] Valérie Vidal[1,2] and Sophie Schbath[1,2]

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

2 *Université Paris-Saclay, INRAE, BioinfOmics, MIGALE bioinformatics facility, 78350, Jouy-en-Josas, France*

3 *Université Paris-Saclay, INRAE, PROSE, 92160, 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.

The Migale platform 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 has been labelled ISC ("Infrastructure Scientifique Collective") by INRAE. It is also one of the four INRAE 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.

Poster 30

SingleCellSignalR: Reconstructing stromal molecular networks from tumor single cell transcriptomes

Simon CABELLO-AGUILAR[1], Jacques COLINGE[1]

[1] IRCM, 208 avenue des Apothicaires, 34090, Montpellier, France

Corresponding Author: jacques.colinge@inserm.fr

Multicellular organism cells engage in a large number of interactions with adjacent or remote cells to coordinate their fate and behavior from early development stages to mature tissues, in healthy and diseased conditions. Although other mechanisms play a role such as tunneling nanotubes, secreted vesicles, ion fluxes, etc., cellular communication is carried over by ligand-receptor (LR) interactions and physical contacts predominantly. The particular case of tumors, where cancer cells are able to reprogram stromal cells through secreted factors, turning neutral or anti-tumoral cells into tumor supportive partners or inactive cells constitutes an extreme example.

A number of recent reports [1, 2] demonstrated the interest of mapping putative LR interactions taking place in tumors to drive fundamental discoveries in cancer biology.

We present SingleCellSignalR, a comprehensive framework to obtain cellular network maps from scRNA-seq data. SingleCellSignalR comes with a complete pipeline integrating existing algorithms to cluster transcriptomes and detect cell type-enriched genes. It then implements a suite of original features allowing to identify the tumor cell subpopulations, calculate cell type-specific internal molecular networks, and infer LR interactions across the subpopulations. SingleCellSignalR generates multiple tabular and graphic reports. All the generated networks can be imported in Cytoscape for improved graphical presentation or subsequent functional analyses.

References

[1] Costa et al., Cancer Cell, 2018
[2] Puram et al., Cell, 2017

Poster 31

Functional inference integrated in the FROGS suite

Moussa SAMB[1], Maria BERNARD[2] and Géraldine PASCAL[1]

[1] GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France
[2] Univ. Paris-Saclay, INRAE, AgroParisTech, GABI, SIGENAE, F-78352, Jouy-en-Josas, France

Corresponding Author: geraldine.pascal@inrae.fr

The high-throughput sequencing of biomarkers has opened new horizons in the study of microbial communities. To help biologist in their studies, several years ago, we developed FROGS [1] is a metabarcoding analysis pipeline. It gives, among other information, the abundance table, the taxonomic affiliation of operational taxonomic units (OTUs) and statistics data. In addition to command line mode, it can be used as a Galaxy workflow, focused on user-friendliness, so it does not require bioinformatics or command lines skills.

To go further in their analyses, biologists generally need metabolic data associated with the microbial composition of the environment they are studying. There are currently several solutions: (i) analysis by RNASeq, (ii) analysis by metagenomics sequencing (iii) analysis by metabolomics and (iv) analysis by functional inference. Despite the ever-increasing accessibility of metagenomics sequencing, functional inference from data obtained from amplicons remains very useful. Indeed, this strategy is important for samples with high host contamination, low biomass and when metagenomic sequencing is not economically feasible. In any case, all produced results by these kind of methods should primarily be used for hypothesis generation.

The principle of functional inference is to infer the metabolic pathways of organisms based only on their taxonomic affiliation. Several tools can do this, MACADAMExplore [2], Tax4Fun [3], PAPRICA [4], Piphillin [5] and PICRUSt2 [6] for example. PICRUSt2 bases on marker gene from sequencing profiles. First, it relies on an algorithm that insert marker sequence into an existing phylogenetic tree thank to short-read placement tools. After, it infers gene family of OTUs. Then, it determines gene family abundance per sample. Finally, it infers pathway abundances to predict sample pathway abundances. It enables functional predictions from 16S or 18S or ITS amplicon profiling.

Thus, to allow the FROGS community to deduce the potential metabolic functions of the targeted environment, we have developed a series of applications in python using PICRUSt2 in particular. The Galaxy interface of these applications also allows non-expert users to use easily these new features.

Acknowledgements

MS are funded by ENVT.

References

1. F. Escudie, et al., FROGS: Find, Rapidly, OTUs with Galaxy Solution. *Bioinformatics*, 2018. 34(8): p. 1287-1294.
2. M. Le Boulch, et al., The MACADAM database: a MetAboliC pAthways DAtabase for Microbial taxonomic groups for mining potential metabolic capacities of archaeal and bacterial taxonomic groups. *Database (Oxford)*, 2019. 2019.
3. F. Wemheuer, et al., Tax4Fun2 : a R-based tool for the rapid prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene marker gene sequences. *bioRxiv*, 2018.
4. J. S. Bowman and H. W. Ducklow, Microbial Communities Can Be Described by Metabolic Structure: A General Framework and Application to a Seasonally Variable, Depth-Stratified Microbial Community from the Coastal West Antarctic Peninsula. *PLoS One*, 2015. 10(8): p. e0135868.
5. N. R. Narayan, et al., Piphillin predicts metagenomic composition and dynamics from DADA2-corrected 16S rDNA sequences. *BMC Genomics*, 2020. 21(1): p. 56.
6. G. M. Douglas, et al., PICRUSt2 : An improved and extensible approach for metagenome inference. *bioRxiv*, 2019.

Poster 32

Auvergne Bioinformatics platform at UCA Mesocentre

Nadia Goué[1], David Grimbichler[1], Pierre Peyret[2] and Antoine Mahul[1]

[1] Auvergne Bioinformatic Platform, Mesocentre, Clermont Auvergne University, 7 avenue Blaise Pascal, TSA 60026, 63 178 Aubière Cedex, France

[2] UMR454 MEDIS Microbiologie Environnement Digestif Santé, UMR UCA - INRAE, CBRV 26 Place Henri Dunant, 63 001 Clermont-Ferrand, France

Corresponding Author: nadia.goue@uca.fr

The Mesocentre as part of Clermont Auvergne University (UCA) is delivering services in high performance computing for sciences data and short-term storage through a network of technology core facilities. These offers are done to assist multi-disciplinary scientists in their computing projects. At that time, we are hosting a computer farm with about 800 cores; 40 nodes for moderate memory usage (<256 Gb); a SMP supercomputer made of 384 cores and 12 To memory; plus a GPU technology (8 GPU of 5120 cores each); a cloud platform - based on Openstack Technology - with a total of 960 physical cores and 9To memory; and a CEPH storage of at least 1 To capacity by user.

Hosted by the Mesocentre, the Auvergne bioinformatics (AuBi) platform is a member of the French Bioinformatics Institute (IFB, https://www.france-bioinformatique.fr/en/platforms/AUBI). AuBi platform aims at sharing expertises and knowledge in large-scale data treatments and analysis by supplying a complete computing environment with hardware and software infrastructures for 9 research laboratories. AuBi platform is then involved in various projects belonging to genomics, metagenomics, transcriptomics, modeling and imaging fields amongst others [1,2,3]. Furthermore, we provide support to UCA laboratories and Associates in their effort to maintain and enhance their scripts and pipelines used on our infrastructure as well as an easy access to public databanks mirrored through BioMAJ [4].

Another aspect of AuBi platform work is to facilitate computing access by the way of Galaxy [5] and an image metadata storage facility through an OMERO [6] server. We are also organizing training sessions to help our users, either biologists or bioinformaticians to optimize computing resources usage through command line interface or Galaxy environment.

References

[1] Pierre Amato, Ludocic Besaury, Muriel Joly, Benjamin Penaud, Laurent Deguillaume and Anne-Marie Delort. Metatranscriptomic exploration of microbial functioning in clouds. Scientific Reports 9: 4383, 2019.

[2] François Balfourier, Sophie Bouchet, Sandra Robert, Romain De Oliveira, Hélène Rimbert, Jonathan Kitt, Frédéric Choulet, IWGS Consortium, BreedWheat Consortium and Etienne Paux. Worldwide phylogeography and history of wheat genetic diversity. Science Advances 5(5): eaav0536, 2019

[3] Caroline Pont, Thibault Leroy, Michael Seidel, Alessandro Tondelli, Wandrille Duchemin, David Armisen, Daniel Lang, Daniela, Bustos-Korts, Nadia Goué, François Balfourier, Márta Molnár-Láng, Jacob Lagen Benjamin Kilian, Hakan Özkan, Darren Waite, Sarah Dyer, Letellier Thomas, Michael Alaux, WHEALBI consortium, Joanne Russel, Beat Keller, Fred van Eeuwijk, Manuel Spannagl, Klaus Mayer, Robbie Waugh, Nils Stein, Kuigi Cattivelli, Georg Haberer, Gilles Charmet and Jérôme Salse. Tracing the ancestry of modern bread wheats. Nature Genetics, (51): 905-911, 2019

[4] Olivier Filangi, Yoann Beausse, Anthony Assi, Ludovic Legrand, Jean-Marc Larré, Véronique Martin, Olivier Collin, Christophe Caron, Hugues Leroy, David Allouche. BioMAJ: A flexible framework for databanks synchronization and processing. Bioinformatics, 24(16): 1823-1825, 2008.

[5] Enis Afgan, Dannon Baker, Bérénice Batut, Marius van den Beek , Dave Bouvier, Martin Cech, John Chilton, Dave Clements, Nate Coraor, Björn 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. 46(W1) : W537-W544, 2018.

Poster 33

Sex-specific transcriptomic signatures of Non-Alcoholic SteatoHepatitis (NASH)

Jimmy Vandel[1], Julie Dubois-Chevalier[1], Céline Gheeraert[1], Bruno Derudas[1], Violetta Raverdy[2], Dorothée Thuillier[2], Luc Van Gaal[3,5], Sven Francque[4,5], François Pattou[2], Bart Staels[1], Jérôme Eeckhoute[1] and Philippe Lefebvre[1]

[1] Univ. Lille, Inserm, CHU Lille, Institut Pasteur de Lille, U1011-EGID, Lille, France.
[2] Univ. Lille, Inserm, CHU Lille, Institut Pasteur de Lille, U1190-EGID, Lille, France.
[3] Dept. of Endocrinology, Diabetology and Metabolism, Antwerp University Hospital, Antwerp, Belgium.
[4] Dept. of Gastroenterology and Hepatology, Antwerp University Hospital, Antwerp, Belgium.
[5] Laboratory of Experimental Medicine and Paediatrics (LEMP), University of Antwerp, Antwerp, Belgium.

Corresponding author: jimmy.vandel@inserm.fr

Abstract

Non-Alcoholic Fatty Liver Disease (NAFLD) is initiated by lipid accumulation in liver to which an inflammatory and ballooning reaction later superimposes to reach the Non-Alcoholic SteatoHepatitis (NASH) stage. NASH may progress further toward fibrosis, cirrhosis and possibly hepatocarcinoma in predisposed individuals. The etiology of the human pathology is multi-factorial and identifying molecular players and/or biomarkers requires large-scale studies which are hampered by the limited availability of representative NASH patient cohorts with associated liver biopsies. Epidemiological studies have pointed to important risk factors in NAFLD such as altered glucose homeostasis, obesity, genetic factors and ethnicity. Gender differences have also been observed [1]. However, no sex-specific signatures for NASH have been proposed nor compared in a robust statistical framework so far.

Hepatic transcriptomic profiles were obtained using microarrays from a 910 obese patient cohort, which was stratified to define "No NASH" and "NASH" patients based on histological characteristics. To minimize biases in the analysis, a subset of 170 patients was selected to compose a sex-balanced learning cohort where men and women were matched according to age, BMI, insulin resistance and fibrosis. Differentially expressed (DE) genes between "No NASH" and "NASH" livers were identified for each sex group in the learning cohort with LIMMA [2]. Due to natural transcriptomic heterogeneity in human samples, bootstrapping was employed to assess the stability of the differential analysis. Then, using random forest method in a recursive feature elimination strategy [3], NASH transcriptomic signatures were learnt from the learning cohort. The classification power of defined sex-specific signatures was evaluated by predicting the NASH status of the remaining patients and further validated using 3 independent cohorts. Sex-specific signatures were also compared to a global signature built independently of sex and to randomly-determined signatures to evaluate the weight of sex factor as a feature driving NASH signature definition.

Sex was identified as the main factor of data heterogeneity in this 910-patient cohort. A bootstrap procedure identified reliably DE genes participating to distinct biological processes in NASH as a function of sex. Generated signatures were evaluated in 3 independent cohorts for their ability to detect NASH patients and reached AUCs in the 0.62-0.76, 0.83-0.93 and 0.80-0.89 ranges respectively.

Acknowledgements

This work was supported by grants from Agence Nationale pour la Recherche (ANR-16-RHUS-0006-PreciNASH and ANR-10-LBEX-46), the European Union (FP6 Hepadip FP6-018734 and FP7 Resolve, FP7-305707), Fondation de France (Grant 2014 00047965) and Fondation pour la Recherche Médicale (Equipe labellisée).

References

[1] A. Lonardo, F. Nascimbeni, S. Ballestri, D. Fairweather, S. Win, T. A. Than, M.F. Abdelmalek, and A. Suzuki. Sex differences in nonalcoholic fatty liver disease: State of the art and identification of research gaps. *Hepatology*, 70(4):1457–1469, 2019.

[2] G. K. Smyth. *Limma: Linear Models for Microarray Data*, pages 397–420. Springer New York, New York, NY, 2005.

[3] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1):389–422, 2002.

Poster 34

PanGBank: exploring GenBank genomes through a resource of pangenome graphs

Paul Amours[1], Adelme Bazin[1], Guillaume Gautreau[1], Mathieu Dubois[1], Laura Burlot[1], Alexandra Calteau[1] and David Vallenet[1]

[1]Laboratoire d'Analyses Bioinformatiques pour la Génomique et le Métabolisme, Génomique Métabolique, CEA, Genoscope, Institut François Jacob, CNRS, Université d'Évry, Université Paris-Saclay, Evry, France

Corresponding Author: pamours@genoscope.cns.fr

Many comparative genomic studies try to get a grasp on the overall gene content of a species. However, with the current explosion of available genomic data, it becomes more complex to use all-vs-all genome comparison approaches. Over the last years, the concept of pangenome emerged, whose goal is to capture the overall diversity of a taxonomic group [1]. A pangenome was described first as two-part components divided into core and accessory. The core genome is defined as the set of genes shared by all the organisms of a taxonomic unit (generally a species). The accessory part contains all other genes, it is crucial to understand the adaptive potential of bacteria and contains genomic regions that are exchanged between strains by horizontal gene transfer (HGT). Currently, pangenome databases (*i.e.* panWeb [2], panX [3], PanGFR-HM [4],...) do not use all available genomes from large genomic databases such as GenBank. Furthermore, no existing databases have, to our knowledge, an API that allows bioinformaticians to extract information automatically and thus make large scale analyses possible.

PanGBank collects pangenomes for all genomes of the NCBI GenBank database (>500K genomes). It uses the PPanGGOLiN bioinformatics method [5] that computes pangenomes not based on the core and accessory paradigm but on multiple statistically inferred partitions. PPanGGOLiN also brings a graph-based solution to represent thousands of genomes in a partitioned pangenome graph. In order to build a PanGBank release, GenBank genomes are first filtered based on various assembly metrics. Then, we use Mash genomic distances [6] combined with the GTDB taxonomy [7] to assign genomes to species with the purpose of building a more homogeneous and correct classification than the one from the NCBI. Then, we build pangenomes for each species that has a sufficiently high number of affiliated genomes (≥15) to apply the PPanGGOLiN statistical partitioning method. Afterward, PanGBank can be used as a reference resource to predict Regions of Genome Plasticity (RGP) and identify insertion hotspots, evaluate the completeness of genomes, visualize and analyze pangenomes according to their gene content among the different partitions.

References

[1] H. Tettelin *et al.*, Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome", *Proc. Natl. Acad. Sci.*, vol. 102, nº 39, p. 13950‑13955, 2005

[2] Y. Pantoja *et al.*, PanWeb: A web interface for pan-genomic analysis, *PLOS ONE*, vol. 12, nº 5, p. e0178154, 2017

[3] W. Ding, F. Baumdicker, et R. A. Neher, panX: pan-genome analysis and exploration, *Nucleic Acids Res.*, vol. 46, nº 1, p. e5‑e5, 2018

[4] N. M. Chaudhari, A. Gautam, V. K. Gupta, G. Kaur, C. Dutta, et S. Paul, PanGFR-HM: A Dynamic Web Resource for Pan-Genomic and Functional Profiling of Human Microbiome With Comparative Features, *Front. Microbiol.*, vol. 9, 2018

[5] G. Gautreau *et al.*, PPanGGOLiN: Depicting microbial diversity via a Partitioned Pangenome Graph, *bioRxiv*, p. 836239, 2019

[6] B. D. Ondov *et al.*, Mash: fast genome and metagenome distance estimation using MinHash, *Genome Biol.*, vol. 17, nº 1, p. 132, 2016

[7] P.-A. Chaumeil, A. J. Mussig, P. Hugenholtz, et D. H. Parks, GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database, *Bioinformatics*, 2019

Poster 35

Systematic Analysis of Protein Post-translational Modifications at a Proteomic Scale in the pathogenic yeast *Candida albicans*

Thomas DENECKER[1], Nicolas SENECAUT[2], Pierre POULAIN[2], Gaëlle LELANDAIS[1] and Jean-Michel CAMADRO[2]

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

[2] Mitochondria, Metals and Oxidative Stress, CNRS, Institut Jacques Monod (IJM), Univ. Paris Diderot, Paris, France

Corresponding Author: thomas.denecker@gmail.com

Bottom-up proteomics consists in identifying proteins in a biological sample using mass spectrometry. It relies on enzymatic digestion of proteins (generally using trypsin) giving rise to complex peptide mixtures that are separated by reversed phase chromatography and injected to a mass spectrometer. As a result, fragmentation mass spectra are obtained. They represent the "peptide signatures" depending of the sequences of amino acids. To identify peptides from spectra, comparisons are performed with theoretical spectrum banks, which are specific of the studied organism proteome. Therefore, if an observed spectrum is highly similar to a theoretical spectrum in the bank, a peptide sequence is inferred and can be used to identify the original protein (before trypsin digestion).

Today, all proteins of a given proteome are not systematically identified. For example, from the ~7000 proteins of *Candida albicans*, only ~2500 proteins are usually detected. One explanation for the lack of identifications can be the absence of systematic consideration of post-translational modifications (PTMs). Indeed, a peptide with a PTM will exhibit a spectrum that is different from the spectrum of the same peptide without the PTM.

Currently, protein identification algorithms consider by default, a limited number of PTMs in their search, *i.e.* oxidation (Met), phosphorylation (Ser, Thr, Tyr), acetylation (N-term of protein) and carbamidomethylation (Cys). These are indeed the most biologically relevant PTMs that are often investigated. They represent only 6 PTMs from the 1500 which are referenced in UNIMOD [1].

Using the software RAId (Robust Accurate Identification), a protein identification algorithm [2], we performed a systematic search for the 237 PTMs on mass spectrometry dataset obtained for the hyphae and the yeast forms of *Candida albicans* (30 mass spectrometry raw files). We observed that the protein identification rate was multiplied by 2.5. From this data set, we are presently (*i*) defining an original list of PTMs of particular interest to be systematically queried with identification software and (*ii*) studying the dynamics of PTMs between two cellular forms of *C. albicans* (hyphae and yeast).

References

1. Creasy, D. M.; Cottrell, J. S. Unimod: Protein modifications for mass spectrometry. Proteomics 2004, 4, 1534– 1536, DOI: 10.1002/pmic.200300744
2. Ogurtsov AY, Alves G, Yu YK (2019) RAId: Knowledge Integrated Proteomics Web Service with Accurate Statistical Significance Assignment. Proteomics, Mar 25;

Poster 37

DeepOmics, a Digital Environmental Engineering Platform for meta-omics data

Ariane BIZE[1], Guillaume PERREAL[2], Aurélie GRAMUSSET[2], Marion PREDHUMEAU[2], Cédric MIDOUX[1,3,4], Valentin LOUX[3,4], Yannick FAYOLLE[1], Patrick DABERT[5], Théodore BOUCHEZ[1], Nicolas RAIDELET[2]

[1] Université Paris-Saclay, INRAE, PROSE, 92761 Antony, France

[2] INRAE, DSI-Solutions Applicatives, 69625 Villeurbanne, France

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

[4] Université Paris-Saclay, INRAE, BioinfOmics, MIGALE bioinformatics facility, 78350, Jouy-en-Josas, France

[5] INRAE, OPAALE, F-35044 Rennes, France

Corresponding Author: ariane.bize@inrae.fr

In the field of environmental biotechnologies, meta-omics approaches enable to finely understand microbial communities acting as catalysts. They thereby offer the possibility to develop more cost-effective processes, through ecological engineering approaches or through the design of operational biomarkers. However, the variety of processes and operating conditions worldwide is very high. It is therefore difficult to extend the conclusions gained on one specific system: achieving a critical mass of data appears as a key-issue to extract relevant and robust information. It highlights the need for data collection and organization at a large scale. We present a new software tool developed to favor the collection, requesting and sharing of such data: DeepOmics is a data warehouse dedicated to environmental biotechnologies. Its main objective is to promote the sharing of meta-omics data and high quality associated metadata.

Deep-omics is an n-tier web application: the user interface is a single page application built using the Angular framework (https://angular.io/). It accesses the data using a RESTful API. The latter is built with the Symfony framework (https://symfony.com/) and its API-platform extension (https://api-platform.com/); it connects the different server-side processes. Data are stored in a relational database (PostgreSQL, https://www.postgresql.org/) as well as in an indexation and search engine (open-source version of Elasticsearch, https://www.elastic.co/fr/community).

DeepOmics offers the possibility to upload, request and export 16S metabarcoding data from several environmental biotechnologies, together with many relevant associated metadata, especially regarding operating conditions and process design. Standard data formats (csv, biom, fastq, …) were preferentially selected. DeepOmics also enables the graphical monitoring of analytical data. In DeepOmics, each project coordinator manages the rights associated to its project data (e.g. private, public). A test server has been released, with restricted access. We plan to open a pilot version to a wider community in the near future.

DeepOmics is a user-friendly tool which should help to take full advantage of next generation sequencing data and to turn knowledge into operational outputs. It will also promote the better agreement of the collected data with the FAIR data principles.

Acknowledgements

We acknowledge E. Le Quéméner, V. Rossard and E. Latrille (INRAE-LBE) for helpful discussions and feedback.

Poster 38

MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis

David VALLENET[1], Alexandra CALTEAU[1], Paul AMOURS[1], Jérôme ARNOUX[1], Adelme BAZIN[1], Mylène BEUVIN[1], Laura BURLOT[1], Xavier BUSSELL[1], Mathieu DUBOIS[1], Stéphanie FOUTEAU[1], Aurélie LAJUS[1], David ROCHE[1], Zoé ROUY[1] and Claudine MÉDIGUE[1].

[1] Laboratoire d'Analyses Bioinformatiques pour la Génomique et le Métabolisme, Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France

Corresponding Author: vallenet@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, the LABGeM group at Genoscope has developed the MicroScope platform (https://mage.genoscope.cns.fr/microscope). Launched in 2005, the MicroScope platform has been under continuous development and provides analysis for prokaryotic genome projects together with metabolic network reconstruction and post-genomic experiments allowing users to improve the understanding of gene functions [1].

In summary, MicroScope supports free-of-charge external submissions of assembled genomes and metagenomes (i.e. Metagenome-Assembled Genomes, MAGs) generated by any sequencing technology or collected from public databanks. Submission of reads for transcriptomics and variant analysis is also available for genomes already integrated into MicroScope. All genomes are analyzed through several workflows for the syntactic and functional annotation that rely on more than 50 tools and databases. Results of computational inferences, including the prediction of metabolic pathways from KEGG or MetaCyc, are loaded into the MicroScope database and made accessible to biologists through a Web user interface (via authenticated or anonymous connections) that provides a variety of analytical and visualization tools. Recents improvements focus on new tools and pipelines developed to perform comparative analyses on thousands of genomes based on pangenome graphs.

To date, MicroScope contains data for >12 600 microbial genomes, part of which are manually curated and maintained by microbiologists (>4800 personal accounts in March 2020). Moreover we propose bi-annual professional trainings as well as on-demand external training sessions. More than 500 users have been trained since 2008 in France or abroad. The platform enables collaborative work in a rich comparative genomic context and improves community-based curation efforts.

References

1. David Vallenet, Alexandra Calteau, Mathieu Dubois, Paul Amours, Adelme Bazin, Mylène Beuvin, Laura Burlot, Xavier Bussell, Stéphanie Fouteau, Guillaume Gautreau, Aurélie Lajus, Jordan Langlois, Rémi Planel, David Roche, Johan Rollin, Zoe Rouy, Valentin Sabatet, and Claudine Médigue. MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. Nucleic Acids Res., Jan 8;48(D1):D579-D589, 2020.

Poster 39

Testing markers to characterize microbial communities in food ecosystems using metagenomics approach

Claire Vincent[1], Serge Casaregola[2], Monika Coton[3], Céline Delbès[4], Hugo Devillers[2], Éric Dugat-Bony[5], Stéphane Guezenec[6], Kate Howell[6], Françoise Irlinger[5], Jean-Luc Legras[6], Valentin Loux[1], Élisa Michel[6], Jérôme Mounier[3], Cécile Neuveglise[6], Thibault Nidelet[6], Audrey Pawtowski[3], Pierre Renaud[7], Delphine Sicard[6], Monique Zagorec[8] and Olivier Rué[1]

[1] UNITE1404 MaIAGE, INRAE Domaine de Vilvert, 78350, Jouy-en-Josas, France
[2] UMR1319 MICALIS, AgroParisTech - site de Grignon, 78850, Thiverval-Grignon, France
[3] UR3882 LUBEM, 970 avenue du Technopôle, 29280, Plouzané, France
[4] UR0342 URTAL, INRAE Poligny, 30 route de Versailles, 39800, Poligny, France
[5] UMR782 SayFood, Avenue Lucien Brétignières, 78850, Thiverval-Grignon, France
[6] UMR1083 SPO, INRAE - Campus Supagro Montpellier, 2 place Viala, 34060, Montpellier, France
[7] UMR1319 MICALIS, INRAE Domaine de Vilvert, 78350, Jouy-en-Josas, France
[8] UMR1014 SECALIM, INRAE Oniris - site de la Chantrerie, route de Gachet La Chantrerie, 44307, Nantes, France

Corresponding author: claire.vincent@inrae.fr

Next generation sequencing offers several ways to study microbial communities. For agricultural sciences, being capable of identifying species in food ecosystems is quite important in a context of sustainable food system development and monitoring. The aim of this study was to test different metabarcoding methods in order to determine which barcode could be best to characterize fungal species diversity in food ecosystems. Four markers were tested for each mock: D1D2, RPB2, ITS1 and ITS2. Mock communities of fungal species were processed in regards of informations provided by previous studies and literature. Each mock corresponds to a different food ecosystem: bread, wine, cheese and meat. A drosophila mock was also added as drosophila may be a vector of yeast dispersion between food ecosystems. In addition, real samples coming from sourdough, wine must, meat and cheese were studied. Some bioinformatics tools were used using their guidelines or with customized parameters to deal with sequence specificities. After the analysis, composition was computed for each community to determine which marker seems to be the 'best' among the four markers tested. To assess this, expected diversity (in the mocks) was compared to the observed. Furthermore, this analysis allowed to improve fungal taxonomic knowledge and enrich fungal-specific databases.

Poster 43

Comparison of plasmid prediction tools to assess the plasmidome of *Salmonella enterica*

Pauline BARBET[1,2], Pierre-Emmanuel DOUARRE[1],
Nicolas RADOMSKI[1], Laurent LALOUX[1], Arnaud FELTEN[1]

[1] Anses, 14 rue Pierre et Marie Curie, 94700, Maisons-Alfort, France
[2] Université de Rouen Normandie, 1 rue Thomas Becket, 76130, Mont-Saint-Aignan, France

Corresponding Author: pauline.barbet@anses.fr

Plasmids are mobile genetic elements that are ubiquitous in bacterial genomes. They play an important role in adaptation and evolution by transferring genes conferring selective advantages to their hosts. Identifying plasmids is critical to understand and control the dissemination of antibiotic resistance [1] and virulence [2] genes in pathogenic bacteria. Several tools, based on various approaches (coverage analysis, k-mer-based classification, replicon detection…) are currently available to assess the plasmidome from draft assemblies. However, prediction with high sensitivity and specificity is difficult to achieve and recovering plasmid sequences from genome assemblies is still challenging. Consequently, the aim of this study is to evaluate the performance of plasmid prediction tools to assess the plasmidome of *Salmonella enterica.*

To that end, a dataset of 56 known genomes of *S. enterica* including 51 genomes carrying from one to five plasmids (n=86) of various sizes was created. This dataset was tested against four recent prediction tools PlaScope [3], PlasFlow [4], HyAsP [5] and MOB-recon (MOB-suite) [6] with different reference plasmid databases we created by compiling either broad or targeted-species (*S. enterica*) plasmid sequences. The performance of the plasmid prediction tools were evaluated by calculating various scores (based on the presence of true/false positive/negative contigs) such as the precision, specificity and recall. The combination of several prediction tools to improve plasmid detection was also investigated.

Overall, the performance of HyAsP and MOB-recon was lower than the other tools and their plasmid prediction was more dependent on the database used as reference. The results showed that PlasFlow was able to detect more plasmid contigs (highest recall) than the other tools and identified at least one contig for all 86 plasmids. However, PlasFlow also identified more non-plasmid contigs (lowest precision and specificity). In contrast, PlaScope demonstrated the highest specificity and precision (less false positive contigs detected). The combination of the two latter tools showed promising results for the accurate prediction of the plasmidome of *S. enterica* and will be tested on a more extensive dataset.

References

1. Alessandra Carattoli, Plasmids and the Spread of Resistance, *International Journal of Medical Microbiology*, 303 (6-7):298-304, 2013.
2. I Rychlik, D Gregorova, H Hradecka, Distribution and Function of Plasmids in Salmonella enterica, 112 (1), 1-10, *Veterinary Microbiology*, 2006.
3. Royer G et al., PlaScope : a targeted approach to assess the plasmidome from genome assemblies at the species level, *Microbial Genomics*, 4, 2018.
4. Krawczyk PS et al., PlasFlow: predicting plasmids identification, *Nucleic Acids research*, Vol.46 No.6, 2018.
5. Robert Müller and Cedric Chauve, HyAsP, a greedy tool for plasmids identification, *Bioinformatics*, 1-4, 2019.
6. James Robertson and John H E Nash, MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies, *Microbial Genomics*, 4, 2018.

Poster 44

Pseudo mitosis during differentiation in Multiciliated cells to drive *de novo* centriole amplification

Tran Bich Ngoc Cao[1,2], Cyril Matthey-Doret[2], Michella Khoury Damaa[3], Agnes Thierry[2], Alice Meunier[3] and Romain Koszul[2]

[1] International Master in Life Sciences, Department of Biology, Ecole Normale Superieure de Paris, 75005, Paris, France

[2] Spatial Regulation of Genome, Institut Pasteur, 75015, Paris, France

[3] Institut de Biologie de l'École Normale Supérieure (IBENS), Paris Sciences et Lettres (PSL), Research University, 75005, Paris, France

Corresponding author: tran-bich-ngoc.cao@pasteur.fr; romain.koszul@pasteur.fr; ameunier@biologie.ens.fr

Abstract

Multiciliated cells (MCCs) differentiation involves the differentiation of progenitor cells into cells displaying hundreds of cilia produced via centriole-independent pathway - the process in which *de novo* cilia synthesis is initiated from a mysterious organelle called *deuterosome* but not the classical *mother* centriole. Much less is known about how the cells harness energy or cellular machinery to manipulate cell cycle towards differentiation by exiting normal mitosis - thus escaping the centriole-dependent pathway. Recently, our collaborators have shown that differentiating MCCs exhibit dependency on mitotic regulators: CDK1 in amplification phase (S-like phase) and CDK2 in growth and disengagement phase (M-like phase). Moreover, their nuclei show the presence of condensation markers such as Ki-67, Histone 3 phosphorylation of serine residue and lamin A/C [1]. The similarity between normal mitosis and MCC differentiating state has inspired us to look at their chromatin architecture in 4D, in space and time along their differentiation stages.

Recent advancements in chromosome conformation capture techniques such as Hi-C have revolutionized the way we investigate the organization of genomes, and its potential interplay with DNA functions. In mammalian interphase, chromatin is thought to organize into self-interacting domains called topologically associating domains (TADs) [2]. This folding is proposed to occur through loop extrusion, a dynamic process mediated by the a ring-shaped molecular motor cohesin. During mitotic compaction, staircase model of chromatin for mitotic pathway has been deducted by comparison of HiC map and polymer modeling. This model proposes that mitotic chromosomes are compacted into arrays of small loops nested into larger loops by the coordinated action of 2 condensin proteins, by which their loop bases, in turn, stemmed and organized in an imaginary staircase, helicoidal backbone [3].

We want to investigate whether chromosomes of differentiating MCCs during interphase exhibit pseudo-mitotic behaviors, and to what extent differentiated MCCs genome folding differ from progenitor cells. To test this hypothesis, we are generating high-resolution HiC maps for differentiating progenitor and differentiated MCCs as well as FACS-sorted differentiating MCCs in their S-like and M-like phases. The role of condensation in gene expression will also be investigated through RNA-seq. Variations between the different cell types may point at pathways and regulators involved both in the differentiation process, but also potentially in the regulation of mitosis. In-house HiC processing pipeline is available in our GitHub: https://github.com/koszullab/hicstuff

References

[1] Adel Al Jord, Asm Shihavuddin, Raphaël Servignat d'Aout, Marion Faucourt, Auguste Genovesio, Anthi Karaiskou, Joëlle Sobczak-Thépot, Nathalie Spassky, and Alice Meunier. Calibrated mitotic oscillator drives motile ciliogenesis. *Science*, 358(6364):803–806, 2017.

[2] Elphège P Nora, Bryan R Lajoie, Edda G Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L van Berkum, Johannes Meisig, John Sedat, Joost Gribnau, Emmanuel Barillot, Nils Blüthgen, Job Dekker, and Edith Heard. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485(7398):381–385, apr 2012.

[3] Johan H. Gibcus, Kumiko Samejima, Anton Goloborodko, Itaru Samejima, Natalia Naumova, Johannes Nuebler, Masato T. Kanemaki, Linfeng Xie, James R. Paulson, William C. Earnshaw, Leonid A. Mirny, and Job Dekker. A pathway for mitotic chromosome formation. *Science*, 359(6376):1–29, 2018.

Poster 46

Shallow shotgun metagenomics workflow validation

Younous ADROUJI[1], Yao AMOUZOU[1], Thomas CARTON[1], Sophie LE FRESNE-LANGUILLE[1], Morgane PIERRE[1], Erwann SCAON[1], Pauline VAISSIE[1] and Sébastien LEUILLET[1]

[1] Biofortis Mérieux NutriSciences, 3 route de la Chatterie, 44800, Saint-Herblain, France

Corresponding Author: younous.adrouji@mxns.com

Introduction

Microorganism biodiversity can be apprehended by metagenomic study of microbiota. This approach relies on high-throughput sequencing technologies, bioinformatic workflows and has applications in clinical investigations, food industry... Shotgun sequencing is the most powerful approach available to investigate a microbiome. Sequencing depth needed is often set high, but some papers start to show that even with limited number of sequences (= the shallow level) the microbiome description appears to be very similar [1]. The main objectives of our study were to validate a complete workflow, from sample to taxonomic and functional compositions, and evaluate the relevance of producing fewer sequences per sample.

Methods

DNA samples from different matrices (stool, skin, vaginal flora) were sequenced, using next generation sequencing technologies, at different sequencing depths. The analysis of raw data was performed with an internally developed bioinformatic workflow based on published tools and algorithms. The shallow level was studied by using sub-sampling on high-depth sequenced samples and calculating diversity and similarity indices between deep sequenced and in silico sub-samples. Multiple replicates were finally used to evaluate the performance of the laboratory methods in terms of repeatability and reproducibility.

Results

As expected, sub-sampling method showed decreasing similarity values, on taxonomic composition at species level and functional composition based on GO terms, when lowering the number of sequences. However, this diminution was not linear and a plateau was observed. This validates the use of fewer sequences to obtain the major part of the information. A threshold was set and sample replicates were sequenced at this targeted number of sequences. The taxonomic and functional compositions were the same between biological replicates and validate the use of limited number of sequences for shotgun metagenomics.

Conclusion

The aim of the project was to validate a complete workflow based on shallow shotgun metagenomics. It succeeds and proves that relatively low number of sequences per sample can be set for this application. However, the sequencing depth depends on the aim of the analysis. For instance, accessing strain-level resolution can be challenging with fewer sequences. Another important angle is the quantity of analyzed data and its quality, both of which depend on the samples and the laboratory methods, such as host contamination and sequencing technology.

References

1. Knights D. Evaluating the Information Content of Shallow Shotgun Metagenomics. mSystems. 2018;3(6):e00069-18.

Poster 47

A shotgun metagenomics analysis applied to the ecological engineering of anaerobic microbial communities for the production of hydrogen from Citrus Peel Waste

Cédric MIDOUX[1, 2, 3], Franciele CAMARGO[4], Olivier RUÉ[2, 3], Mahendra MARIADASSOU[2, 3], Valentin LOUX[2, 3], Maria Bernadete AMÂNCIO VARESCHE[4] and Ariane BIZE[1]

[1] Université Paris-Saclay, INRAE, PROSE, 92160, Antony, France

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

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

[4] Department of Hydraulics and Sanitation, School of Engineering of São Carlos, University of São Paulo, Av. Trabalhador São Carlense, 400, 13566-590 São Carlos, SP, Brazil

Corresponding Author: cedric.midoux@inrae.fr

Brazil, the first Citrus fruit producer worldwide, produces annually 88 million tons of this fruit; a large part of it is processed, generating waste, composed mainly by peels and bagasse, which accounts for half of the processed mass. More than half of this waste is not yet valorized, although it could be used in biorefineries to produce various invaluable compounds, including biogas. In this project, we focused on the production of hydrogen from Citrus Peel Waste by anaerobic microbial communities. We used a shotgun metagenomics approach to identify key strains and biological functions for this process, by comparing the inoculum (not yet adapted) with the microbial communities sampled after hydrogen production in optimized conditions.

We developed a bioinformatics pipeline for shotgun metagenomics data analysis. After a preprocessing step, reads were co-assembled using metaSPADES [1]. Coding regions were predicted with Prodigal [2]. The taxonomic assignments of the predicted genes were obtained with kaiju [3] against NCBI nr database. For the functional annotation, the predicted genes were affiliated with ghostKoala [4] against KEGG database and with dbCAN [5] against the CAZY database. For each dataset individually, reads were mapped on the co-asssembled contigs. It was thus possible to build a table, which included, for each predicted gene, the counts by sample as well as the functional and taxonomic annotations.

This pipeline was developed with snakemake [6] to favour reproducible and scalable data analysis. It was run on the cluster of the INRAE MIGALE bioinformatics platform. It is accessible to the community on GitLab (https://gitlab.irstea.fr/cedric.midoux/workflow_metagenomics).

We were able to identify key strains and biological functions selected during this process, especially those related to lignocellulose deconstruction and to hydrogen production. A strong and reproducible microbial selection was observed, leading to the dominance of *C. beijerinckii* and *C. butyricum*, known to be hydrogen producers. The glycoside hydrolase 48 family, thought to be critical for cellulose hydrolysis, was in low proportion and it was associated to *C. puniceum*, a cellulosome-producing bacterium. *C. puniceum* could thus in the present case be key for some steps of the lignocellulose deconstruction, despite its low proportion.

1. Bankevich, A., et al., *SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.* J Comput Biol, 2012. **19**(5): p. 455-77.
2. Hyatt, D., et al., *Prodigal: prokaryotic gene recognition and translation initiation site identification.* BMC Bioinformatics, 2010. **11**: p. 119.
3. Menzel, P., K.L. Ng, and A. Krogh, *Fast and sensitive taxonomic classification for metagenomics with Kaiju.* Nat Commun, 2016. **7**: p. 11257.
4. Kanehisa, M., Y. Sato, and K. Morishima, *BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences.* J Mol Biol, 2016. **428**(4): p. 726-731.
5. Zhang, H., et al., *dbCAN2: a meta server for automated carbohydrate-active enzyme annotation.* Nucleic Acids Res, 2018. **46**(W1): p. W95-W101.
6. Koster, J. and S. Rahmann, *Snakemake--a scalable bioinformatics workflow engine.* Bioinformatics, 2012. **28**(19): p. 2520-2.

Poster 48

Comparison of Bacterial Metabarcoding Analysis Pipelines for Species-level Identification

Yao AMOUZOU[1], Damien CHAUVEAU[1], Younous ADROUJI[1], Xuwen WIENEKE[2], Erwann SCAON[1], Sarita RAENGPRADUB WHEELER[2] and Sébastien LEUILLET[1]

[1] Biofortis Mérieux NutriSciences, 3 route de la Chatterie, 44800, Saint-Herblain, France
[2] Corporate Microbiology R&D, Mérieux NutriSciences, 3600 Eagle Nest Drive, South Building, Crete IL, 60417, USA

Corresponding Author: yao.amouzou@mxns.com

Introduction

NGS metabarcoding technology is a powerful tool for analysis of the composition and diversity of microbiota. However, various bioinformatic approaches are available but perform differently to assess microbiota taxonomic profiles.

Methods

This study evaluated several clustering and classification methods/algorithms to determine which can characterize bacterial communities to the species level. Different matrices (urine, skin, stool, vaginal and oral flora), plus a bacterial simulated data sample as a positive control, were profiled by targeting the V3-V4 region of the 16S ribosomal RNA and using the Illumina MiSeq. DADA2, Deblur, VSEARCH implemented in QIIME2 pipeline [1] and MED [2] were evaluated for data clustering. BLAST, ScikitLearn, and VSEARCH were evaluated for operational taxonomic unit (OTU) classification. A total of 11 pipelines (various combinations of the data clustering and classification approaches), including a validated Mothur [3] workflow as a reference, were tested. The pipeline performance score was calculated based on computing efficiency and OTU classification results (precision, recall, F-measure, specificity, taxonomic ranking).

Results

All the tools tested provided good results for genus level classification. However, at species level, the combination of MED+BLAST and Mothur proved to be the best pipelines. They demonstrated the highest efficiency and the OTU accuracy, for all the analyzed matrices.

Conclusion

This study highlights that some bioinformatic methods outperform others in terms of classification robustness and computational resources. These pipelines could enable surveying bacterial communities to the species level, although a significant rate of unclassified taxonomy is to be expected.

References

1. Bolyen, E., Rideout, J.R., Dillon, M.R. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37, 852–857 (2019). https://doi.org/10.1038/s41587-019-0209-9

2. A. Murat Eren, Hilary G. Morrison, Pamela J. Lescault, Julie Reveillaud, Joseph H. Vineis, and Mitchell L. Sogin. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. 9(4) :968–979.

3. Schloss, P.D., et al., Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol, 2009. 75(23):7537-41

Poster 49

CGST : New Core-Genome Sequence Typing Tool

Aurélien Birer[1] and Richard Bonnet[1]

[1] Centre National de Référence “Résistances aux antibiotiques”, CHRU de Clermont-Ferrand
Laboratoire de Bactériologie, 63003, Clermont-Ferrand, France

Corresponding Author: abirer@chu-clermontferrand.fr

1 Summary

The typing of bacteria is key activity in microbiology, which allows to discriminate among bacterial strains from the same species. The main goals are the investigation of outbreaks and the surveillance of pathogenic and/or antibiotic-multiresistant bacteria to understand of the transmission, pathogenesis and phylogeny of bacteria.

With the emergence of the new generation of sequencing (NGS) technologies, it became possible to identify allele variations from whole genomes and to classify bacterial isolates at different levels of resolution. Sequence types (STs) can be determined for a set of 7 to 15 housekeeping genes accordingly to the multi locus sequence typing (MLST)[1] approach or for all conserved genes accordingly to the core genome MLST (cgMLST) approach.

Here is proposed a pipeline typing tool based on NGS data and designated CGST. CGST aims two goals: (i) the detection of allelic variants using ARIBA[3] for MLST, MentaLiST[2] for cgMLST and species-specific typing tools (i.e. ClermonType[4] and FimTyper[6] for Escherichia coli and Kleborate[5] for Klebsiella pneumoniae), and (ii) the analysis of variants to assign the strains to clusters and investigate their clonality using phylogenetic and unsupervised clustering approaches.

Overall, CGST is a typing tool integrating multiple approaches for the classification of bacteria in coherent genetic groups using different scales of resolution. CGST would be available on GitHub.

1.1 KeyWords

Multi Locus Sequence Typing ; Core-Genome ; Sequence Typing

References

[1] Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci USA. 1998;95:3140-3145. doi: 10.1073/pnas.95.6.3140.

[2] Feijao P, Yao HT, Fornika D, et al. MentaLiST - A fast MLST caller for large MLST schemes. Microb Genom. 2018;4(2):e000146. doi:10.1099/mgen.0.000146

[3] Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J, et al. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. Microb Genom. 2017;3:e000131. doi: 10.1099/mgen.0.000131.

[4] Clermont O, Christenson JK, Denamur E, Gordon DM. The Clermont Escherichia coli phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. Environ Microbiol Rep. 2013;5:58–65. doi: 10.1111/1758-2229.12019.

[5] Lam MCC, Wick RR, Wyres KL. et al. Kleborate: comprehensive genotyping of Klebsiella pneumoniae genome assemblies. 2018. https://github.com/katholt/Kleborate.

[6] Roer L, Tchesnokova V, Allesoe R, Muradova M, Chattopadhyay S, Ahrenfeldt J, Thomsen MCF, Lund O, Hansen F, Hammerum AM, et al. Development of a web tool for Escherichia coli subtyping based on fimH alleles. J Clin Microbiol. 2017;55(8):2538–2543. doi: 10.1128/JCM.00737-17.

Poster 51

A graph-based approach to classify A-minor motifs in RNA structures according to their structural context

Coline Gianfrotta[1,2], Vladimir Reinharz[3] , Olivier Lespinet[4], Dominique Barth[1], Alain Denise[2,4]

[1] DAVID, Univ. Versailles-Saint-Quentin, Univ. Paris-Saclay, Versailles, France

[2] LRI, Univ. Paris-Sud, UMR CNRS 8623, Univ. Paris Saclay, Gif-sur-Yvette, France

[3] Département d'informatique, Université du Québec à Montréal, Québec, Canada

[4] I2BC, Univ. Paris-Sud, UMR CNRS 9198, CEA, Univ. Paris-Saclay, Gif-sur-Yvette, France

Corresponding Author: coline.gianfrotta@ens.uvsq.fr

1 Introduction

Functions of RNA are numerous and many of them are essential to any living organism. These functions rely on a good folding of the molecule, through the formation of bonds between non-consecutive nucleotides. Consequently, it is necessary to predict the three-dimensional structure of an RNA to determine its function. Nowadays, effective algorithms based on dynamic programming are able to predict a first level of folding, called secondary structure, composed of canonical interactions. However, more complex interactions exist, and some of them stay currently unpredictable. *A-minor like* motifs fall into this category. These motifs are the most frequent interactions binding distant regions of the molecule, and have been proven to play a crucial role in folding [1,2] and cellular mecanims [3].

In this work, we study structural context of the A-minor interactions, e.g. the set of canonical and non-canonical bonds around the motif, in order to determine how the local context is involved in the formation of the interactions. We would like to leverage this context to classify A-minor and predict their location. For this purpose, we are looking for common structural characteristics in structural contexts of real occurrences stored in PDB.

2 Method

We represented structural contexts by graphs, where vertices represent nucleotides or groups of nucleotides and edges represent canonical and non-canonical bonds. To compare these graphs, we set up a similarity measure, called contextual graph similarity, based on the number of canonical and non-canonical edges in a largest common subgraph. Using this measure, we developed comparison graph and clustering algorithms in order to obtain a first classification.

3 Preliminary results

Our first results show that occurrences, clustered according to our similarity measure, often share close 3D-structures. Moreover, not only homologous occurrences are clustered together, but also some non-homologous occurrences, which means not evolutionnary related occurrences. This tends to show that a classification of A-minor motifs according to the structural context can exist independently of homology. We are currently refining our similarity measure, in order to have a better correspondance between our similarity and the 3D-structure.

References

[1] C. Geary, A. Chworos, and L. Jaeger. Promoting RNA helical stacking via A-minor junctions. *Nucleic Acids Res*, (39/3):1066–1080, 2011

[2] K. N. Sripathi, P. Banáš, K. Réblová, J. Šponer, M. Otyepka, and N. G. Walter. Wobble pairs of the HDV ribozyme play specific roles in stabilization of active site dynamics. *Phys. Chem. Chem. Phys, (*17/8):5887–5900, 2015

[3] Aurélie Lescoute and Eric Westhof. The A-minor motifs in the decoding recognition process, *Biochimie*, (88/8):993–999, 2006

Poster 52

Chromatin Conformation Capture unveils mechanisms for DNA Double-Strand Breaks repair foci formation

Coline Arnould[1], Vincent Rocher[1], Thomas Clouaire[1], Pierre Caron[1], Philippe. E. Mangeot[2], Emiliano. P. Ricci[3], Raphaël Mourad[1], Daan Noordermeer[4] and Gaëlle Legube[1]

[1] LBCMCP, Centre de Biologie Intégrative (CBI), CNRS, Université de Toulouse, UT3

[2] CIRI - International Center for Infectiology Research, Université Claude Bernard Lyon 1

[3] Laboratoire de Biologie et Modélisation de la Cellule, Université de Lyon

[4] Institute for Integrative Biology of the Cell (I2BC), Université Paris-Saclay, Gif-sur-Yvette

Corresponding author: gaelle.legube@univ-tlse3.fr, raphael.mourad@ibcg.biotoul.fr

Within the nucleus, DNA is associated with proteins called histones, to form a complex structure, the chromatin, which can adopt several levels of condensation. The properties of the chromatin fiber are very heterogeneous along the chromosome and regulated by epigenetic marks like DNA methylation and post-translational modifications of histones. It can also modulate its structure, in term of epigenome and 3D structure. When DNA is damaged by double-strand breaks (DSBs), local chromatin structure greatly influences its repair. In Legube's team, we are interested in the relationship between DNA repair and chromatin. We have already shown in a previous study that DSBs have the ability to form clusters in 3D[1]. In this study, we seek to understand the local changes in the 3D structure of chromatin around the breaks, and particularly how the chromatin loops evolve near the break.

Our team has developed the DIvA cell line, which is a powerful experimental system that enables the induction of multiple DSBs at specific locations across the genome in various chromatin contexts[2].We investigated the 3D environment of induced DSBs, to gain high resolution around breaks, using 4C-seq, as well as low resolution structure of the whole genome with Hi-C, to study DNA Damage Response (DDR).

By integrating 4C-seq, and Hi-C with ChIP-seq, we found that the recruitment of repair components is governed by pre-existing high-order chromatin, established before DNA damage induction. We found that the 3D profile around the break (4C-seq) is very similar to the spreading of repair components (ChIP-seq). To generalize this finding, we computed Topologically Associating Domain (TADs) boundaries (Hi-C) on the whole genome, and found that these TADs are functional units governing DDR chromatin domain establishment. Also, differences of spreading between kinases and phosphorylated histones suggested that 3D conformation constrains and mediates the spreading of DDR components.

In summary, our work[3] showed that TADs are templates for the spreading of many DSB repair components, which allow DSB signaling at the megabase scale. We showed that TADs play a major role in genome stability, and provide an efficient way to signal DNA damages and create specific repair prone chromatin compartment.

References

[1] F. Aymard, M. Aguirrebengoa, E. Guillou, B. M. Javierre, B. Bugler, C. Arnould, V. Rocher, J. S. Iacovoni, A. Biernacka, M. Skrzypczak, K. Ginalski, M. Rowicka, P. Fraser, and G. Legube. Genome-wide mapping of long-range contacts unveils clustering of DNA double-strand breaks at damaged active genes. *Nat. Struct. Mol. Biol.*, 24(4):353–361, 04 2017.

[2] J. S. Iacovoni, P. Caron, I. Lassadi, E. Nicolas, L. Massip, D. Trouche, and G. Legube. High-resolution profiling of gammaH2AX around DNA double strand breaks in the mammalian genome. *EMBO J.*, 29(8):1446–1457, Apr 2010.

[3] Coline Arnould, Vincent Rocher, Thomas Clouaire, Pierre Caron, Philippe. E. Mangeot, Emiliano. P. Ricci, Raphael Mourad, Daan Noordermeer, and Gaëlle Legube. Loop extrusion as a mechanism for dna double-strand breaks repair foci formation. *bioRxiv*, 2020.

Poster 55

Where are the dual C/E domains in NRPS originated from?

Clémentine CAMPART[1], Loïc COUDERC[2], Matthieu DUBAN[3], Valérie LECLERE[3] and Maude PUPIN[1]

[1] Univ. Lille, CNRS, Centrale Lille, UMR 9189–CRIStAL–Centre de Recherche en Informatique Signal et Automatique, F-59000 Lille, France

[2] Bilille, Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, US 41 - UMS 2014 - PLBS, F-59000 Lille, France

[3] Univ. Lille, INRAE, ISA, Univ Artois, Univ. Littoral Côte d'Opale, EA 7394–ICV–Institut Charles Viollette, F-59000 Lille, France

Corresponding Author: clementine.campart@univ-lille.fr

NonRibosomal Peptides (NRPs) are microbial secondary metabolites. NRPs are composed of more than 500 different building blocks that can be proteogenic and non proteogenic amino acids, their derivatives, fatty acids, carbohydrates, and many other monomers. Not all NRPs are linear. They can contain branches and/or cycles. Such diversity in structure and composition leads to a diversity in function (antibiotics, siderophores...) and class (lipopeptides, glycopeptides…).

NRPs are built up by huge multimodular enzymatic complexes called NonRibosomal Peptide Synthetases (NRPSs). NRPSs exploit a modular concept for the synthesis of NRPs in which each module is responsible for the activation and incorporation of one monomer into a growing peptide chain [1]. The modules are themselves divided into distinct enzymatic domains. There are 4 main domains: 1) adenylation domain (A), responsible for the selection and activation of a monomer, 2) thiolation domain (T), responsible for the transfer and tethering of the corresponding adenylate to the NRPS-bound 4'-phosphopantetheinyl cofactor, 3) condensation domain (C) responsible for the peptide bond formation, and 4) thioesterase domain (Te) responsible for the release of the peptide. Additional domains are sometimes present. Their role is to modify some monomers during biosynthesis. The most frequent secondary domain is the epimerization one (E) which modifies an L-monomer into its D-isomer.

Subtypes of C-domain have been identified such as the dual C/E domain that combine both functions leading to an epimerization and the linkage between the new D-monomer and the L-monomer of the next module. This dual C/E domain is only found in lipopeptide NRPSs present in bacteria belonging to *Pseudomonas*, *Xanthomonas* and *Burkholderia* genera. These 3 genera also produce siderophores thanks to NRPSs containing E-domains distinct from C-domains. With such particularity, we propose a phylogenetic study to better understand the occurrence of the dual C/E domain in the lipopeptide NRPSs while a combination of E and C-domains is present on other NRPSs in this 3 genera. The aim is to find the origin of these dual C/E domains and to detect presence of potential event of horizontal gene transfer (HGT) between this 3 genera.

We collect all genomes and proteomes available in the *Xanthomonas*, *Burkholderia* and *Pseudomonas* genera, representing more than 1000 strains. Florine [2] pipeline was used to extract the dual C/E, C and E-domains. This pipeline uses antiSMASH [3] and NapDos [4] tools. Regarding the production of phylogenetic trees (gene tree, species tree) and the identification of xenolog genes (*ie.* gene involved in HGT), several tools and pipelines will be tested.

References

[1] Süssmuth RD, Mainz A (2017) Nonribosomal peptide synthesis-principles and prospects. Angew Chem Int Ed Engl 56:3770–3821

[2] Caradec T, Pupin M, Vanvlassenbroeck A, Devignes M-D, Smaïl-Tabbone M, Jacques P, Leclère V (2014) Prediction of monomer isomery in Florine: a workflow dedicated to nonribosomal peptide discovery. PLoS One 9:e85667

[3] Blin, K.; Shaw, S.; Steinke, K.; Villebro, R.; Ziemert, N.; Lee, S.Y.; Weber, T. antiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res. 2019, 47, 81–87

[4] Ziemert, N., Podell, S., Penn, K., et al. (2012). The natural product domain seeker NaPDoS: A phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLoS One, 7*, e34064

Poster 56

COBRA: an automated RNA-seq data analysis pipeline.

Skander HATIRA, Frederic GRANDJEAN and Bouziane MOUMEN

Écologie et Biologie des Interactions, UMR7267, 5, rue Albert Turpain TSA 51106, 86000, Poitiers, France.

Corresponding Author: bouziane.moumen@univ-poitiers.fr

RNA-seq has become almost a routine technique in biological laboratories, widely used to quantify the expression of genes in tissues and cells permitting to take a whole picture of the transcriptional pattern of an organism and so deciphering its genomic repertoire without having to sequence the entire genome especially for non-model organisms.

Honestly, there is no shortage of tools to analyze RNAseq data. RNAseq blog [1] gives an idea of the multitude of tools, programs and pipelines, that we can use to deal with RNAseq data, each with its advantages and disadvantages. For a biologist the most important is the results and its interpretation, not the details of the processes used to generate it. In these situations, the automation of analysis processes and steps becomes very desirable, especially for people lacking bioinformatics skills, but also for projects that generate a lot of data (samples, conditions). Reproducibility, sharing and publishing the results is another aspect of the utility of pipelines [2].

To address these aspects, we developed Cobra, a Snakemake-based workflow [3] that allows preprocessing, assembly, annotation, differential gene expression analysis and identification of genetic variants from RNA-seq data from raw data to the final results. COBRA has the particularity of being easy to use, with preset parameters for almost steps.

We made COBRA in production to build and annotate crayfish *Austropotamobius pallipes* [4] and identify the differentially expressed immune genes in crayfish *Astacus astacus* infected with plague.

References

1. RNA-Seq Blog : https://www.rna-seqblog.com/.
2. Joël Simoneau, Simon Dumontier, Ryan Gosselin, Michelle S Scott, Current RNA-seq methodology reporting limits reproducibility, Briefings in Bioinformatics, , bbz124, https://doi.org/10.1093/bib/bbz124.
3. Köster, Johannes and Rahmann, Sven. “Snakemake - A scalable bioinformatics workflow engine”. Bioinformatics 2012.
4. Frederic Grandjean, Han Ming Gan, Bouziane Moumen, Isabelle Giraud, Skander Hatira, Richard Cordaux, Christopher M. Austin, Dataset for sequencing and de novo assembly of the European endangered white-clawed crayfish (Austropotamobius pallipes) abdominal muscle transcriptome, Data in Brief,Volume 29,2020,https://doi.org/10.1016/j.dib.2020.105166.

Poster 57

Gut microbiota and clinical efficacy of immunotherapy in renal cell carcinoma

Laurie Alla[1]*, Lisa Derosa[2]*, Valerio Iebba[2], Emmanuelle Le Chatelier[1], Laurence Zitvogel[2]
[1] MetaGenoPolis, INRAE - Université Paris-Saclay, Domaine de Vilvert, Bâtiment 325, 78350 Jouy-en-Josas, France
[2] Gustave Roussy, 114 Rue Edouard Vaillant, 94800 Villejuif, France
Corresponding author : laurie.alla@inrae.fr

In adults, renal cell carcinoma (RCC) is the most common type of kidney cancer representing about 80% of renal cancer. Immunotherapy is a type of cancer treatment that helps the immune system to fight cancer. One strategy involves immune checkpoint inhibitors (ICIs) such as the binding of PD-L1 (on tumor cells) to PD-1 (on T cells) that keeps T cells from killing tumor cells. Blocking that binding with an immune checkpoint inhibitor (anti-PD-L1 or anti-PD-1) allows the T cells to kill tumor cells. However, patient's response to immune checkpoint blockade is heterogeneous. Indeed, between 20 to 40% of patients respond to this kind of therapy. Gut microbiota has been shown to have a significant role in health and response to various disease treatment. More specifically, it is known that an abnormal gut microbiota composition can be linked to primary resistance to ICIs.

Using shotgun metagenomic, we analysed the stool samples at diagnosis of 69 patients with advanced RCC. As expected, we showed that treatment response was affected by antibiotic (ATB) intake. Indeed, patients that received antibiotics within 60 days before ICIs (n = 11, 16%) had a lower objective response rate (ORR) compared to the no antibiotic (noATB) subgroup (9% versus 28%, $p < 0.03$) and lower progression free survival (PFS) and overall survival (OS). This compromised clinical efficacy of ICB goes with an alteration of intestinal microbiota composition. Indeed, *Eubacterium rectale* ($p = 0.02$) seems to be significantly more abundant in noATB stools samples while bacteria such as *Erysipelotrichaceae bacterium_2_2_44A* ($p = 0.02$) and *Clostridium hathewayi* ($p < 0.02$) seem to be overrepresented in ATB fecal samples. We showed that metagenomic profiles of baseline stools samples could predict PFS (at 3, 6, 9 and 12 months) for responders (R) versus non-responders (NR) patients. Indeed, higher gene and species richness metrics were found correlated with the clinical response defined by the absence of progression of the tumor at 12 months after initiation of ICB.

Finally, to perform a robustness test across at least 3 clinical parameters, we found 27 reliable Metagenomic Species (MGS) (out of 1347) contrasting between R (n=21) and NR (n=6) (based on the cliff delta for each MGS recovered in >50% tests). Among these selected MGS, four were in common with microbiome profiles associated with lung cancer patients response to ICIs. *Akkermansia muciniphila* and *Bacteroides salyersiae* were found associated with favorable outcome during anti-PD-1 blockade while others species such as *Eggerthella lenta*, *Clostridium bolteae* seem to be overrepresented in NR patients.

Altogether, we conclude that analysis of the gut microbiota before treatment, according to both microbial richness and composition could be used to identify the RCC population likely to respond to anti-PD-1 treatment and predict patients with PFS longer than 12 months. This paves the way to the identification of microbial partners or communities that could favor patient response to ICI treatment.

Poster 58

Unveiling cancer cell and tumor microenvironment interactions at the single cell level in colorectal-cancer liver metastases

Ambre GIGUELAY[1,2,3], Ander CHURRUCA SCHUIND[1,2,4], Simon CABELLO-AGUILAR[1], Barbara CHIAVARINA[1], Pierre-Emmanuel COLOMBO[5], Lakhdar KHELLAF[5], Didier POURQUIER[5], Andrei TURTOI[1] and Jacques COLINGE[1,2,3]

[1] Institut de Recherche en Cancérologie de Montpellier, Inserm U1194, 208 avenue des apothicaires, 34298 Cedex 5, Montpellier, France

[2] Université de Montpellier, 163 rue Auguste Broussonnet, 34090, Montpellier, France

[3] Labex EpiGenMed, 141 rue de la Cardonille, 34396 Cedex 5, Montpellier, France

[4] Department of Molecular and Translational Medicine, Viale Europa 11, 25123, Brescia, Italy

[5] Institut régional du Cancer de Montpellier, 208 avenue des apothicaires, 34298 Cedex 5, Montpellier, France

Corresponding Authors: andrei.turtoi@inserm.fr, jacques.colinge@inserm.fr

Despite the fact that liver metastases are the major cause of death from colorectal, lung and pancreatic cancers, their biology is far from understood. While it is generally accepted that the tumor microenvironment plays a crucial role in the process of metastases, little is known about the molecular crosstalk between stromal and cancer cells. In the present work and within the context of colorectal cancer liver metastases, we sought to explore the molecular interactions between cancer, immune and mesenchymal cells at the single cell level. Owing to our collaboration with the Surgery and Pathology Departments of ICM, we started a unique collection of fresh liver metastases from colorectal patients undergoing surgery and following neoadjuvant chemotherapy. To date over 50 liver metastasis were collected and 7 of those have been utilized in the present project for single cell RNAseq analysis. For the latter metastases, fresh samples were subjected to enzymatic digestion following FACS isolation of cancer, immune and mesenchymal cells. In total, around 30,000 cells have been subjected to scRNAseq using 10X Chromium technology, providing an unprecedented map of tumoral heterogeneity and highlighting notably subpopulations of mesenchymal cells. Current works are underway to develop and apply algorithms enabling the creation of a comprehensive map of molecular interactions between the different cellular types.

Poster 60

PanFam: a pangenomic workflow for the iterative construction of gene families on thousands of genome

Jérôme Arnoux[1], Mylène Beuvin[1], Adelme Bazin[1], Alexandra Calteau[1] and David Vallenet[1]
[1]Laboratoire d'Analyses Bioinformatiques pour la Génomique et le Métabolisme, Génomique Métabolique, CEA, Genoscope, Institut François Jacob, CNRS, Université d'Évry, Université Paris-Saclay, Evry, France

Corresponding author: jarnoux@genoscope.cns.fr

Large-scale genome sequencing and the increasingly massive use of high-throughput approaches produces a vast amount of new information that completely transforms our understanding of microbial species. Hence performing comparative genomics studies on hundreds to thousands of genomes has become a challenge as relying on millions of pairwise sequence comparisons is too computationally intensive.

To overcome this problem we have decided to develop a new high throughput strategy to compare and cluster protein sequences using the MMSeqs2 software suite[1]. We designed a workflow for the construction of homologous protein families at different similarity levels (80%, 50% and 30% amino acid identity), with an alignment coverage of 80% on both target and query protein sequences. To identify the best strategy we decided to evaluate different clustering methods. Those approaches need to support an iterative construction process to update families with new proteins as well as the identification of protein fragments and fusion events. To evaluate the functional consistency of the obtained protein families, i.e. the homogeneity of the functional annotations inside of each cluster, we used KofamKOALA[2] which assigns KEGG[3] ortholog groups.

This gene family resource will be used in the MicroScope platform[4] for different comparative genomic functionalities such as synteny computation and gene phylogenetic profiles. Moreover, it will be distributed through the PPanGGOLiN software[5] for the computation of microbial pangenomes.

References

[1] Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol*, 35(11):1026–1028, 2017.

[2] Takuya Aramaki, Romain Blanc-Mathieu, Hisashi Endo, Koichi Ohkubo, Minoru Kanehisa, Susumu Goto, and Hiroyuki Ogata. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*, 11 2019. btz859.

[3] Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res*, 44:D457–D462, 2016.

[4] David Vallenet, Alexandra Calteau, Mathieu Dubois, Paul Amours, Adelme Bazin, Mylène Beuvin, Laura Burlot, Xavier Bussell, Stéphanie Fouteau, Guillaume Gautreau, Aurélie Lajus, Jordan Langlois, Rémi Planel, David Roche, Johan Rollin, Zoe Rouy, Valentin Sabatet, and Claudine Médigue. MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. *Nucleic Acids Research*, page gkz926, 2019.

[5] Guillaume Gautreau, Adelme Bazin, Mathieu Gachet, Rémi Planel, Laura Burlot, Mathieu Dubois, Amandine Perrin, Claudine Médigue, Alexandra Calteau, Stéphane Cruveiller, Catherine Matias, Christophe Ambroise, Eduardo PC Rocha, and David Vallenet. Ppanggolin: depicting microbial diversity via a partitioned pangenome graph. *bioRxiv*, 2020.

Poster 62

Sequencing, Assembly and Annotation of 40 Brown Algae Genomes

Nachida Tadrent[1], Corinne Cruaud[1], Arnaud Couloux[1], Corinne Da Silva[1], Benjamin Noel[1], Olivier Godfroy[2], Zofia Nehr[2], Erwan Corre[2], Susana Coelho[2], France Denoeud[1], Patrick Wincker[1], Jean-Marc Aury[1] and Mark J. Cock[2]

[1] Genoscope, Institut de biologie François-Jacob, Commissariat à l'Energie Atomique (CEA), Université Paris-Saclay, 2 Rue Gaston Crémieux, 91000, Évry, France

[2] UMR 8227 CNRS-UPMC Integrative Biology of Marine Models, Station Biologique, Place Georges Teissier, 29680, Roscoff, France

Corresponding author: ntadrent@genoscope.cns.fr

The Phaeoexplorer project aims to generate transcriptomic data and annotated genome assemblies for a broad range of brown algal species at different phylogenetic distances from the model *Ectocarpus* species in order to address a number of key questions about the biology and evolutionary history of this poorly characterised but important group of marine eukaryotes. The knowledge generated by the project will be exploited to develop new techniques and products for the macroalgal mariculture and processing industries.

This comparative genomic project is a large sequencing project (40 brown algal species and four species closely related to the brown algae in male and female). Genomic sequencing was performed with two different strategies, Illumina short reads to generate draft assemblies, and long reads from Oxford Nanopore Technology (ONT) for high-quality genome assemblies. As brown algae live in symbiosis with other organisms, mainly bacteria, efforts have been made upstream of sequencing to isolate the organism of interest and get axenic cultures. Despite this, extracted DNA may come from several organisms, which can in most cases be compared to metagenomic samples. Long reads obtained with ONT allow us to more easily separate the sequences of the brown alga from those of the symbionts to obtain high quality and symbiont-less genome assemblies.

An automated annotation pipeline has been developed to define the exon / intron structures and their positions in the genome assemblies of different species, whether close or distant from the reference, and manages the fact that these are weakly known lineage with potential horizontal gene transfer. This pipeline, combine biological data from different resources (transcripts, conserved proteins or *ab initio* gene models) using Gmove, an in house software, to predict coding gene models without prior training.

The generated data represent an important new resources for global research on brown algae. The analyses of these data will allow significant advances in our understanding of the biology of a currently poorly characterised group of marine organisms, for example, the evolution of complex multicellularity, the mechanism of speciation or the evolution of sexual systems.

Poster 64

RGCCA Shiny, a graphical package for multimodal omics data analysis: application to Parkinson's disease

Etienne CAMENEN[1], Caroline PELTIER[1], François-Xavier LEJEUNE[1], Arthur TENENHAUS[1,2] and Ivan MOSZER[1]

[1] iCONICS Core Facility, Institut du Cerveau, Inserm U 1127, CNRS UMR 7225, Sorbonne Université, F-75013, Paris, France

[2] Laboratoire des Signaux et Systèmes (L2S), CentraleSupélec, Université Paris-Saclay, F-91190, Gif-sur-Yvette, France

Corresponding Author: caroline.peltier@icm-institute.org

In cohort studies, the variety of biological and health data collected on the same individuals is of high interest to improve our understanding of complex diseases. To this end, the use of Canonical Correlation Analysis (CCA) is a popular approach for heterogeneous data integration, which has been conveniently generalized to allow the use of various CCA-related methods and variable selection with sparsity-inducing norms (S/RGCCA) [1,2,3]. The current version of the S/RGCCA package [4] proposes algorithms able to efficiently solve the optimization problems. In the latest release of this package described here, new functionalities were made available: (i) missing data strategies (NIPALS or imputation procedures), even with blockwise missing structure; (ii) the tuning parameters of S/RGCCA can now be automatically selected using a permutation or cross-validation scheme; (iii) the robustness of the selected variables can be assessed using bootstrap resampling techniques; (iv) several graphical functions were implemented to ease the interpretation of S/RGCCA outputs (individuals/variables maps, bootstrap confidence intervals, etc.); (v) graphical user interfaces were developed to promote usage of these tools by end-users (biologists, clinicians).

The usefulness of these developments was evaluated in the context of the Nucleipark project. The clinical and omics data of 57 advanced Parkinson's Disease (PD) patients were investigated in order to identify potential biomarkers of disease severity. Prior to analysis, data were organized in 4 sets of variables: clinical data (34 variables), metabolomics (3,530 metabolites), lipidomics (1,019 lipids) and transcriptomics (10,729 genes). Out of the 57 PD patients, 32 subjects (56.1%) have missing values in at least one modality. The new features of RGCCA allowed us to take advantage of all the available information and avoid a substantial loss of valuable data. Graphical display of the S/RGCCA outputs were easily plotted. For instance, the correlation between the variable selection within each modality and the two first components of the clinical modality could be observed on a correlation circle. We identified genetic and chemical biomarkers of interest that were found to be associated with the motor status of PD patients. The S/RGCCA framework applied to independent and larger cohorts is expected to validate and refine our multiomics signatures toward a better comprehension of PD progression.

The R package RGCCA is available in a new CRAN update and in a Shiny application.

Acknowledgements

We warmly thank Claire EWENCZYK, Farid ICHOU, Fanny MOCHEL and Marie VIDAILHET for giving access to the Nucleipark dataset and fruitful discussions about the clinical and biological objectives and results. CP is funded by the iMAP program (ANR-16-RHUS-0001). EC is funded by the Institut Français de Bioinformatique (ANR-11-INSB-0013) in the framework of the pilot project "IntegrParkinson". This work was also supported by the IHU-A-ICM program ANR-10-IAIHU-06.

References

1. Tenenhaus A and Tenenhaus M. Regularized generalized canonical correlation analysis. *Psychometrika*, 76:257-284, 2011.
2. Tenenhaus M, Tenenhaus A and Groenen PJF. Regularized generalized canonical correlation analysis: A framework for sequential multiblock component methods. *Psychometrika*, 82(3):737–777, 2017.
3. Tenenhaus, A., Philippe, C., Guillemot, V., Le Cao, K. A., Grill, J., and Frouin, V. Variable selection for generalized canonical correlation analysis. *Biostatistics*, 15(3):569-583, 2014.
4. Tenenhaus, A. and Guillemot V. RGCCA: Regularized and Sparse Generalized Canonical Correlation Analysis for Multiblock Data. R package version 2.1.2. *https://CRAN.R-project.org/package=RGCCA*, 2017.

Poster 69

Circular RNAs detection and cross-species conservation

Chloé Cerutti[1], Annie Robic[1] and Thomas Faraut[1]
GenPhySE, Université de Toulouse, INRAE, ENVT, 31326, Castanet Tolosan, France

Corresponding author: Thomas.Faraut@inrae.fr

For many years, circular RNAs (circRNAs) were considered as splicing byproducts because they were associated with low levels of expression. However, the development of high-throughput RNA sequencing and circRNAs-specific computational tools highlighted the abundance of these circRNAs in eukaryotic cells. These single-stranded RNAs are made of closed continuous loops lacking free ends and are generated during the splicing of pre-mRNA [1]. According to recent studies, circRNAs are mainly produced by exonic sequences of coding genes (exonic circRNAs) through an alternative splicing mechanism called "backsplicing". More precisely, the end of a "donor" exon is linked to the beginning of an upstream "acceptor" exon. More rarely, circRNAs can also be produced from intronic sequences (intronic circRNAs) from intronic lariat [1]. The specific conformation of circRNAs makes their detection, quantification and functional characterization difficult [2]. Several circRNA mechanisms of action were already identified. According to existing studies, circRNAs would mainly act as microRNAs sponges or by protein interactions [2]. Recent studies have also shown that some circRNAs are evolutionarily conserved [3].

To screen the exhaustive circRNAs genomic content, we analyzed Total-RNAseq data obtained from pig (sus scrofa) and bovine (bos taurus) testicular and liver tissues. In an attempt to obtain a comprehensive overview of circRNAs in those tissues we developped an approach with a detection step agnostic to genome annotation [4]. This approach enables us to quantify the relative proportion of exonic and intronic circRNAs from coding and non-coding genes, the variability between individuals, tissues and species. Differential recruitement of splice junctions for circular transcripts versus linear transcript is also addressed.

Acknowledgements

We thank the genotoul bioinformatics facility for providing computing ressources and the institute of genome biology of FBN (Dummersdorf, Germany) for providing bovine datasets.

References

[1] A. Robic, T. Faraut, S. Djebali, R. Weikard, K. Feve, S. Maman, and C. Kuehn. Analysis of pig transcriptomes suggests a global regulation mechanism enabling temporary bursts of circular RNAs. *RNA Biol*, 16(9):1190–1204, 09 2019.

[2] L. S. Kristensen, M. S. Andersen, L. V. W. Stagsted, K. K. Ebbesen, T. B. Hansen, and J. Kjems. The biogenesis, biology and characterization of circular RNAs. *Nat. Rev. Genet.*, 20(11):675–691, 11 2019.

[3] S. Xia, J. Feng, L. Lei, J. Hu, L. Xia, J. Wang, Y. Xiang, L. Liu, S. Zhong, L. Han, and C. He. Comprehensive characterization of tissue-specific circular RNAs in the human and mouse genomes. *Brief. Bioinformatics*, 18(6):984–992, Nov 2017.

[4] Y. Gao and F. Zhao. Computational Strategies for Exploring Circular RNAs. *Trends Genet.*, 34(5):389–400, 05 2018.

Poster 73

PYTHIA: Deep Learning Approach For Local Protein Conformation Prediction

Gabriel Cretin[1,2], Tatiana Galochkina[1,2], Charlotte Perin[1,2], Alexandre G. de Brevern[1,2] and Jean-Christophe Gelly[1,2]

[1] Université de Paris, Biologie Intégrée du Globule Rouge, UMR_S1134, BIGR, INSERM, F-75015, Paris, France

[2] Laboratoire d'Excellence GR-Ex, Paris, France

Corresponding Author: jean-christophe.gelly@univ-paris-diderot.fr

1. Introduction

Protein Blocks (PBs) constitute a structural alphabet which describe the local structure of a protein more accurately than the classical secondary structures [1]. PBs are composed of 16 structural conformations of five consecutive amino acids able to encode complex protein structures (3D) into a sequence vector (1D). PBs have been used in several applications such as protein structure alignment and protein structure prediction [2,3]. We present here a deep inception-inside-inception convolutional network, called PYTHIA (**P**redicting An**y** Conforma**t**ion at **Hi**gh **A**ccuracy), to predict the local structure of a given protein sequence in terms of Protein Blocks.

2. Material and methods

The objective of our method is to predict one of the 16 PB for each position of a protein sequence. Each amino acid is encoded by a vector of features: 58 physico-chemical properties (AA-Index [4]), a 20-dimensional position-specific substitution matrix (PSSM) generated by PSI-BLAST and finally gaps as a single encoded value. Our dataset of 9638 protein chains is composed of a non redundant set of resolved protein structures extracted from the PDB. We divide this dataset into independant train, validation and test datasets to perform a 10 fold cross validation. We implemented PYTHIA, a deep inception-inside-inception convolutional neural architecture, with Tensorflow. We compare PYTHIA results to the reference method LOCUSTRA [5] on test dataset and CASP 13 targets of the free modelling category.

3. Results

The mean accuracy of PYTHIA on the test set measured as Q_{16} is near 70% compared to LOCUSTRA with a Q_{16} of 61%. PYTHIA greatly outperforms LOCUSTRA on every PB class even for the rarest PBs if we compare MCC values such as with 'g' 0.209 vs. 0.154 and 'j' 0.315 vs. 0.223 for PYTHIA and LOCUSTRA respectively. We observe the same tendency on CASP targets.

Acknowledgments

The authors were granted access to high performance computing (HPC) resources at IDRIS (Institut du développement et des ressources en informatique scientifique) under grant no. AD011011381 funded by the GENCI (Grand Equipement National de Calcul Intensif).

References

1. de Brevern, A., Etchebest, C. and Hazout, S. (2000), Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins, 41: 271-287.
2. Gelly J-C, Joseph AP, Srinivasan N, de Brevern AG. iPBA: a tool for protein structure comparison using sequence alignment strategies. Nucleic Acids Res. 2011;39: W18–W23.
3. Ghouzam Y, Postic G, Guerin P-E, de Brevern AG, Gelly J-C. ORION: a web server for protein fold recognition and structure prediction using evolutionary hybrid profiles. Sci Rep. 2016;6: 1–10.
4. van Westen GJP, Swier RF, Wegner JK, IJzerman AP, van Vlijmen HWT, Bender A. Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets. J Cheminform. 2013;5: 41.
5. Olav Z. and Ulrich H. E. Hansmann. LOCUSTRA: Accurate Prediction of Local Protein Structure Using a Two-Layer Support Vector Machine Approach. Journal of Chemical Information and Modeling 2008 48 (9), 1903-1908

Poster 75

Industrial NGS analysis processes from sequencing to variant interpretation on MOABI platform

Aminata NDIAYE[1], Jocelyn BRAYET[1], Maxime CHEVAILLOT, Mathieu BARTHELEMY[1], Romain DAVEAU[1], Vivien DESHAIES[1], Laurent FROBERT[1], Guillaume MEURICE[1], and Alban LERMINE[1,2]

[1]MOABI (Bioinformatic platform of AP-HP), 33 boulevard Picpus, 75012, Paris, France
[2]SeqOIA-IT, 33 boulevard Picpus, 75012, Paris, France

Corresponding Author: aminata.ndiaye2@aphp.fr

The Assistance Publique – Hôpitaux de Paris (AP-HP) is a teaching hospital groupment with a European dimension globally recognized. The AP-HP is organized into twelve hospital groups, for a total of 39 hospitals localized in Paris and its region. Currently, those hospitals attend each year 8 millions patients.

Three years ago, MOABI, a new bioinformatics platform was created for multiple missions: the progressive storage centralization for genomic data routinely produced by hospitals, their analyses in controlled and standardized workflows and the provisioning of tools for results exploitation.

In this abstract, we present two softwares designed to analyze NGS diagnosis data: G-route and Leaves. The first tool is a java n-tier rich client web application that provides to users raw sequencing data loading, traceability metadata definition, analysis pipelines running and data files browsing. Data files management relies on the adaptative middleware iRODS [1]. At the present time, G-route contains 12600 analyzed patients, 247 users, 14 skeletons and 31 versions of pipelines for 61 different gene panels and exome analysis. This makes it possible to propose more than 100 different pipeline combinations. The second program, Leaves, is an open source tool that aims to help biologists for genetic alterations interpretation and biological report generation by associating detected alterations with different annotations and scores and performing reproducible filters combination and ranking. Leaves is a web interface mainly developed with python 3 and javascript. Currently, Leaves's database contains 146 users, 61 projects, 3.800.830 variants and 70 variants classifications. Leaves also permit the sharing of AP-HP expertise, in a standard way, between biologists, promoting human interaction over artificial intelligence. Users can run analyses from G-route through more than 100 different pipelines that end up inserting variant calling results into Leaves. Pipelines are written in Snakemake [2], that use Docker [3] containers as version fixed tools. Docker allows to eliminate tool dependencies problems and sets a version tool in an image. Presently, we have nearly 128 tools integrated in that way. Snakemake is a workflow management system with implicit rule implementation (input and output logic). The advantage over Nextflow [4] is the capability to share rules between pipelines which allowed us to create a rule library (194 rules) that can be shared between pipelines. As other pipeline frameworks, error recovery, automatic parallelization and workflow integrity features are included in Snakemake. To ease medical diagnosis routine, AP-HP's scientists are able to execute tagged workflows with their data and to consult results through G-route and Leaves interfaces.

Key words: Industrial NGS analysis, variants interpretation, iRODS, Docker and Snakemake

References

[1] Web site: https://irods.org/
[2] Johannes Köster, Sven Rahmann; Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 2012; 28 (19): 2520-2522. doi: 10.1093/bioinformatics/bts480
[3] Web site: https://www.docker.com/
[4] Jeremy Leipzig; A review of bioinformatic pipeline frameworks. *Brief Bioinform* 2017; 18 (3): 530-536. doi: 10.1093/bib/bbw020

Poster 76

ERA-BIO-IT: A new 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 recently created 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] simple pipelines) 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 and data storage 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.

Poster 77

Bioinformatc evaluation of the impact of base quality score recalibration (BQSR) over variant discovery from NGS data

C. Fournier[1,2], C. Chapusot[1,2], B. Tournier[1,2], R. Aucagne[1,2], E. Tisserant[3,4], L. Faivre[3,4,5], M. Callanan[1,2], Y. Duffourd[3,4]

[1] Université Bourgogne-Franche Comté, UMR1231 Inserm, F-21000 Dijon, France
[2] Unité fonctionnelle Innovation génétique et épigénétique en oncologie, Plateforme de biologie hospitalo-universitaire, CHU Dijon Bourgogne, F-21000 Dijon, France
[3] INSERM, UMR1231 GAD team, Genetics of Developemental disorders, F-21000 Dijon, France
[4] CHU Dijon, FHU TRANSLAD, F-21000 Dijon, France
[5] CHU Dijon, Centre de référence Anomalies du Développement et Syndromes Malformatifs, F-21000 Dijon, France

Corresponding Author: cyril.fournier@chu-dijon.fr

1

Nowadays, next-generation sequencing (NGS) technologies are faster and cheaper than ever, enough for paving the road to a precision medicine. However, we know that these techniques are not flawless and bring their share of errors at different stages of the data generation process. Therefore, bioinformatic analysis must take into account these errors and correct them to generate the most possible precise and realistic result. Meanwhile in a diagnostic context, the running time of analysis must remain reasonable in order to meet the time requirements imposed by this type of examination. Among the many steps driving a bioinformatic analysis of NGS data, the base quality score recalibration (BQSR) is challenging in terms of time and computing resources, no matter the type of analysis – germline or somatic.

BQSR is a data pre-processing step focusing on detecting the recurrent errors generated by the sequencer when it computes an estimation of the quality score of a base call. The base quality score, known as phred score, define the probability of a base call to be correct. This score is yet biased by the sequencing context, the machine model, the sequencing technology used, etc … Therefore, the BQSR step goal is to detect and fade out those different biases to obtain the most realistic quality score.

We benchmarked several recalibration tools in order to assess their impact on the variant calling. We tested « GATK » [1] versions 2, 3 and 4, « bamUtils » and « LACER », in addition to « Kbbq » and « Stampy ». Those tools use slightly different algorithms to do the job.

The dataset used come from the Genome in a Bottle (GIAB) consortium [2], which provides sequencing data on human genomes to allow benchmarking analysis and variant calling pipeline validations. Our dataset is composed of a mix of 2 genomes, NA12878 carrying somatic mutations and NA24385, from which a set of verified variations has been generated by the GIAB and then can be used to assess our bioinformatics tools. The base calling pipeline is based on the « *Best practices workflow* » from GATK.

Results obtained indicates a very small variation in the number of detected variants, and a similar sensitivity among methods (GATK4 = 98.36 %, GATK2/3 = 98.25 %, bamUtils = 98.34 %, No BQSR = 98.36%). Results tends to show that recalibration doesn't affect pipeline sensitivity to call variant, but on the other hand doubles the overall running time in most cases.

Base quality score recalibration seems to have a minor impact on variant validation by variant calling algorithm in the conditions of our experiment. Regarding these results, the BQSR step, which is a time-consuming step, could therefore be removed from our pipeline to save precious time in the diagnostic process.

References

[1] McKenna A, DePristo MA et al. *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data*. GenomeRes. 2010 Sep;20(9):1297-303. doi: 10.1101/gr.107524.110

[2] Zook, J., Catoe, D., McDaniel, J. et al. *Extensive sequencing of seven human genomes to characterize benchmark reference materials*. Sci Data 3, 160025 (2016). https://doi.org/10.1038/sdata.2016.25

Poster 78

New insights into cow holobiont in relation to health

Mahendra MARIADASSOU[1], Xavier NOUVEL[3], Diego MORGAVI[4], Lucie RAULT[2], Sophie SCHBATH[1], Sarah BARBEY[5], Frederic LAUNAY[5], Olivier SANDRA[6], Pierre GERMON[7], Emmanuelle HELLOIN[7], Rachel LEFEBVRE[8], Yves LE LOIR[2], Christine CITTI[3] and Sergine EVEN[2]

[1] Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France
[2] INRAE, Institut Agro, UMR1253 STLO, 35042, Rennes, France
[3] INRAE UMR1225 IHAP, [4] INRAE UMR1213 Herbivores,[5] INRAE DEP,
[6] INRAE UMR1198 BDR, [7] INRAE UMR1282 ISP, [8] INRAE UMR1313 GABI

Corresponding Author: mahendra.mariadassou@inrae.fr

1. Context

Infectious diseases have been traditionally considered as the result of the bipartite interaction between a given pathogen and its host. Recent advances in high-throughput sequencing technology have uncovered the complexity of the various microbiomes associated with the host and symbiotic microbiota have emerged as a main player of the infectious process.

In cattle, major efforts have been devoted to the characterization of the microbiome associated to different anatomical sites in relation to animal performance and health. However, these have mainly focused on the comparison of microbiomes of healthy versus diseased animals. Issues that remained to be addressed include the role in disease development of the microbiome associated to the affected organ as well as the impact of microbiome located at remote body sites.

Here, we propose to explore the structure, diversity and dynamics of microbiomes associated to 4 anatomical sites in cows before and after calving. The interdependence of these microbiomes in relation to animal health and genetics will also be investigated.

2. Material and methods

Over one thousand samples were collected from 45 primiparous prim' Holstein cows selected from two divergent lineages that are respectively more or less susceptible to mastitis. Sampling were performed at 4 time points from 1week pre-partum to 7 months post-partum and from 4 anatomic sites: nasal, genital, buccal (as a proxy for rumen), and foremilk (as a proxy for internal teat microbiome). For each sample, we performed 16S and ITS metabarcoding sequencing and used DADA2 [1] to characterize the communities using Amplicon Sequence Variant (ASV, analogue of OTU) table. We then performed diversity analyses on the ASV count table using Phyloseq [2].

3. Results

We first show that the 4 sites harbor different microbiota both in terms of richness and ASV repertoire. We then use alpha- and beta-diversity analyses to show how the composition evolves in time and differs between lineages. Finally, we will present the results obtained while performing differential abundance analysis within each site to identify the differentially abundant taxa, i.e. taxa whose abundances differ between lineages.

Acknowledgements

This work is part of the MICROCOSM project funded by the INRAE Metaprogramme MEM.

References

1. Callahan, B., McMurdie, P., Rosen, M. et al. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* 13, 581–583 (2016). https://doi.org/10.1038/nmeth.3869
2. McMurdie PJ, Holmes S (2013) phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* 8(4): e61217. https://doi.org/10.1371/journal.pone.0061217

Poster 79

Multi-omics integrative approaches for biological meaningful signatures

Florian Jeanneret and Stéphane Gazut
CEA, LIST, 91191 Gif-sur-Yvette cedex, France

Corresponding author: florian.jeanneret@cea.fr

1 Abstract

High throughput technologies(e.g genomic, transcriptomic, proteomic) have greatly enhanced physiological insights discovery. Last years, several integrative multi-omic methods have been developed to handle biological complexity seen as interdependent combinations of molecular heterogeneous actors [1]. Especially, some of these tools showed relevant uses on large cancer omic datasets provided by TCGA consortium with new underlying pathways actors assumptions[2][3]. In fact, dimension reductions, samples clustering and multi-omic features selection could lead to better understanding of the correlated structures between phenotype aspects. Supervised sPLS-DA[4] or unsupervised modified Consensus PCA(CPAC)[5] results are promising and highlight the interest of these multiblock multivariate analysis[6]. The first one have been assessed on several omics data types from TCGA.

Our work addresses histological type contrasts between two types of breast cancers : invasive lobular or invasive ductal carcinomas with exhaustive comparisons between single-omic datasets multivariate analysis — SVM and Random forest — from BioSigner and DIABLO multi-omic integrative method. Gene and protein expressions with miRNA-seq and phenotypic data from 359 patients(TCGA) have been used to reveal few multi sources discriminant features and such non-invasive predictive signatures for diagnostic and prognostic in clinical cases. Finally, prior knowledge integration to improve result relevance is a main issue in our project.

References

[1] Indhupriya Subramanian, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. Multi-omics data integration, interpretation, and its application. 14:1177932219899051.

[2] Laura Cantini, Pooya Zakeri, Celine Hernandez, Aurelien Naldi, Denis Thieffry, Elisabeth Remy, and Anaïs Baudot. Benchmarking joint multi-omics dimensionality reduction approaches for cancer study. page 2020.01.14.905760.

[3] Morgane Pierre-Jean, Jean-François Deleuze, Edith Le Floch, and Florence Mauger. Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration.

[4] Amrit Singh, Casey P. Shannon, Benoît Gautier, Florian Rohart, Michaël Vacher, Scott J. Tebbutt, and Kim-Anh Lê Cao. DIABLO: from multi-omics assays to biomarker discovery, an integrative approach.

[5] Chen Meng, Dominic Helm, Martin Frejno, and Bernhard Kuster. moCluster: Identifying joint patterns across multiple omics data sets. 15(3):755–765.

[6] Raphael Ployet, Mônica T. Veneziano Labate, Thais Regiani Cataldi, Mathias Christina, Marie Morel, Hélène San Clemente, Marie Denis, Bénédicte Favreau, Mario Tomazello Filho, Jean-Paul Laclau, Carlos Alberto Labate, Gilles Chaix, Jacqueline Grima-Pettenati, and Fabien Mounet. A systems biology view of wood formation in eucalyptus grandis trees submitted to different potassium and water regimes. 223(2):766–782.

Poster 80

JOBIM 2020 Poster

Improving probe design for smFISH experiments

Elodie SIMPHOR[1], Edouard BERTRAND[1] and Charles-Henri LECELLIER[1,2]

[1] Institut de Génétique Moléculaire de Montpellier, University of Montpellier, CNRS, Montpellier, France

[2] LIRMM, Université Montpellier, CNRS, Montpellier, France, 5 Université Paul-Valery-Montpellier 3, Montpellier,

Corresponding Author: elodie.simphor@etu.umontpellier.fr, charles.lecellier@igmm.cnrs.fr

Imaging gene expression at the level of single cells has revolutionized our perceptions of this fundamental processes. In particular, detection of single mRNAs in fixed cells by fluorescent microscopy (such as single molecular Fluorescent *In Situ* Hybrization or smFISH), allows exquisitely precise quantification of gene expression while at the same time retrieving spatial information at the cellular and sub-cellular levels [1]. Recent developments of these techniques allow multiplexed labelling and simultaneous detection of up to 1000 different RNA species, and these techniques offer great promise for both fundamental research and as diagnostic tools [2]. One underappreciated limitation of smFISH experiments lies in the design of the oligonucleotide probes that hybridize to the target RNA. A high signal-to-noise ratio requires the use of multiple oligos that all binding to the same target RNA. Indeed, the noise arises from the non-specific binding of single oligonucleotide, while the signal scales linearly with the number of probes used. Typically, 24 oligonucleotides are used to detect a single mRNA by smFISH [3]. However, large scale experiments revealed a great variability in the efficiency of RNA detection, with good signal-to-noise sometimes obtained with as little as 10 oligos, and other times requiring 10 times more probes. Yet, probe design is still made with algorithms that only equalize the Tm or DG° of the probes, and it is clear that such algorithms can be vastly improved [4]. Here, I present our approach aimed at automatically improving the design of smFISH probes using a set of existing data and a machine learning-based classification trained to discriminate efficient from non-efficient probes.

Keywords: smFish, machine learning, probe

References

1. Nikolay Tsanov, Aubin Samacoits, Racha Chouaib, Abdel-Meneem Traoulsi, Thierry Gostan, Christian Weber, Christophe Zimmer, Kazem Zibara, Thomas Walter, Marion Peter, Edouard Bertrand and Florian Mueller smiFISH and FISH-quant – a flexible single RNA detections approach with super-resolution capability. *Nucleic Acid Research 2016.*
2. Li,G.-W. and Xie,X.S. (2011) Central dogma at the single-molecule level in living cells. *Nature*, **475**, 308–315.
3. Raj,A., van den Bogaard,P., Rifkin,S.A., van Oudenaarden,A. and Tyagi,S. (2008) Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods*, **5**, 877–879.
4. Mueller,F., Senecal,A., Tantale,K., Marie-Nelly,H., Ly,N., Collin,O., Basyuk,E., Bertrand,E., Darzacq,X. and Zimmer,C. (2013) FISH-quant: Automatic counting of transcripts in 3D FISH images. *Nat. Methods*, **10**, 277–278.

Poster 81

pgSNP : Bacterial DNA pangenomic workflow

Madeleine DE SOUSA VIOLANTE[1,2], Valérie MICHEL[1], Carole FEURER[3], Nicolas RADOMSKI[2], Michel-Yves MISTOU[4] and Ludovic MALLET[2]

[1] Actalia, 310 Rue du Père Popiełujko,50000, Saint-Lô, France
[2] ANSES, 14 Rue Pierre et Marie Curie, 94700, Maisons-Alfort, France
[3] IFIP-Institut du Porc, La Motte au Vicomte B.P. 35104, 35651, Le Rheu Cedex, France
[4] INRAE, MaIAGE, Université Paris-Saclay, F-78352, Jouy-en-Josas, France

Corresponding Author: madeleine.desousaviolante@anses.fr

Salmonella is one of the most common bacterial pathogen worldwide in human and animal infections [1]. Each year, gastroenteritis cases due to non-typhoidal *Salmonella* were estimated up to 93.8 million including 155,000 deaths [2]. To provide new insights on *Salmonella* epidemic investigations, whole genome sequencing (WGS) methods have been developed, especially focusing on detection of outbreaks and estimation of genetic relationships between isolates [3,4,5,6].

The current trend to leverage WGS data is to detect highly informative markers such as Single Nucleotide Polymorphisms (SNPs) in the core genome. This high-resolution comparison method discriminates sequences and is able to reveal evolutionary histories, as well as tracing back the source of an outbreak. However, in the context of food safety control, the network-like nature of contamination and the short evolutionary time lead to compare highly related samples, often diverging only by few mutations. Coregenome SNP-based approaches exclude accessory genome, which may be present in sub-clusters of the outbreak of interest and increase genetic distances between outbreak related genomes harboring similar core-genome. Facing the increase in size of datasets in which each sample exhibits genome plasticity and diversity, coregenome size tends also to shorten.

Here, we developed a pangenome-based analysis workflow including variants aiming at evaluating the entire genome of bacterial samples, including accessory genome fractions not shared by all samples. A pangenome reference was built from samples to showcase all existing sequences by a cumulative iterative blast approach [7]. Then, detection of pangenomic variants against the pangenome reference was performed by Snippy [8]. All variants from all samples were concatenated using an in-house pipeline to produce different alignments where one tree was built for each alignment. Finally, a supertree was inferred from all trees.

This workflow was applied on a large *Salmonella* Typhimurium and its monophasic variant dataset. The results showed that half of the data are not taken into account in coregenome-based analysis, underscoring the importance of developing new methods that include the information associated to the dispensable genome. We also displayed that the accessory genome provided further phylogenetic signal and brought higher resolution for strain discrimination. This work demonstrated the importance of the pangenomic variants-based analysis.

References

[1] EFSA. The European Union One Health 2018 Zoonoses Report. *EFSA Journal*, 17(12):5926, 2019
[2] Shannon E. Majowicz and al. The Global Burden of Nontyphoidal Salmonella Gastroenteritis. *Clinical Infectious Diseases*, (50):882–889, 2010.
[3] Philip M. Ashton and al. Identification of Salmonella for public health surveillance using wholegenome sequencing. *PeerJ4*, e1752, 2016
[4] Sophie Octavia and al. Delineating Community Outbreaks of Salmonella enterica Serovar Typhimuriumby Use of Whole-Genome Sequencing: Insights into Genomic Variability within an Outbreak. *Journal of Clinical Microbiology*, (53.4):1063-1071, 2015
[5] Henk C. den Bakker and al. Rapid Whole-Genome Sequencing for Surveillance of Salmonella enterica Serovar Enteritidis. *Emerging Infectious Diseases*, (20.8): 1306–1314, 2014
[6] Tim Dallman and al. Phylogenetic structure of European Salmonella Enteritidis outbreak correlates with national and international egg distribution network. *Microbial Genomics*, 2016
[7] Stephen F. Altschul and al. Basic local alignment search tool. *Journal of Molecular Biology*, (215.3): 403–410, 1990
[8] Seeman T. Snippy : fast bacterial variant calling from NGS reads. https://github.com/tseeman/snippy.

Poster 83

MetExplore: Solutions to access, extend and exploit knowledge of metabolism

Florence Vinson[1], Ludovic Cottret[2], Maxime Chazalviel[3], Amina Taoui[1], Nathalie Poupin[1], Clément Frainay[1] and Fabien Jourdan[1]

[1] UMR1331, Toxalim (Research Centre in Food Toxicology), Université de Toulouse, INRAE, ENVT, INP-Purpan, UPS, 31300 Toulouse, France

[2] LIPM, Université de Toulouse, INRAE, CNRS, 31326 Castanet-Tolosan, France

3 Medday Pharmaceuticals, 75008, Paris, France

Corresponding Author: fabien.jourdan@inrae.fr

1 Introduction

Metabolism is a complex system involving thousands of metabolites interconnected through biochemical reactions. With the rise of metabolomics technologies, the study of this system is now done on a large scale. This calls for the development and distribution of dedicated tools aimed at managing the overwhelming amount of information needed to interpret metabolomics results.

2 Service description

MetExplore is a freely available web server dedicated to the study of metabolism. It has been continuously maintained and extended with new features for 10 years[1]. MetExplore centralizes 297 published metabolic models from various organisms and sources (xml files attached to publications and public databases). The MetExplore platform provides collaborative edition and curation tools for those models[2]. It also offers a graphical interface to easily navigate and filter the information contained in those models, including an interactive network view that can be embedded in other websites via an open source javascript library[3]. This network view has been recently extended to propose an innovative visualization of metabolism aimed at minimizing information overload. Beyond information processing and visualization, several features have been added over the years to contextualize and enhance metabolomics results, providing data mapping, identifier harmonization, pathway enrichment, flux modeling and network analysis, including a unique metabolite recommender system[4].

3 Availability

The MetExplore web server is freely accessible at https://metexplore.toulouse.inrae.fr

Acknowledgements

The MetExplore team thanks the GenoToul bioinformatics facility for the hosting and support of the MetExplore service. MetExplore is supported by MetaboHub French metabolomics and fluxomics infrastructure.

References

[1] Ludovic Cottret, David Wildridge, Florence Vinson, Michael P. Barrett, Huber Charles, Marie-France Sagot et Fabien Jourdan. MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks. Nucleic Acids Research. 1;38 Suppl:W132-7, 2010

[2] Ludovic Cottret, Clément Frainay, Maxime Chazalviel, Floréal Cabanettes, Yoann Gloaguen, Etienne Camenen, Benjamin Merlet et al. MetExplore: collaborative edition and exploration of metabolic networks. Nucleic acids research, (46):W495–W502, 2018.

[3] Maxime Chazalviel, Clément Frainay, Nathalie Poupin, Florence Vinson, Benjamin Merlet, Yoann Gloaguen, Ludovic Cottret, and Fabien Jourdan. MetExploreViz: web component for interactive metabolic network visualization. Bioinformatics, (34):312–313, 2017.

[4] Clément Frainay, Sandrine Aros, Maxime Chazalviel, Thomas Garcia, Florence Vinson, Nicolas Weiss, Benoit Colsch et al. MetaboRank: network-based recommendation system to interpret and enrich metabolomics results. Bioinformatics, (35):274−283, 2018.

Poster 84

ESKRIM: accurate reference-free comparison of microbial richness in shotgun metagenomic samples by k-mers counting

Florian PLAZA OÑATE[1] and Emmanuelle LE CHATELIER[1]

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

Corresponding Author: florian.plaza-onate@inrae.fr

Microbial richness is the ecological measure of the number of taxa encountered in a given microbial community. Many studies revealed that richness of the host microbiome is associated with health and disease. For instance, richness of the human gut microbiome is significantly lower in individuals with metabolic or immune disorders compared to healthy controls [1–3].

Shotgun metagenomics is a non-targeted sequencing technique particularly suitable for estimating microbial richness. Indeed, it detects microorganisms reluctant to culture and achieves species-level resolution contrary to amplicon sequencing. A typical bioinformatics analysis first consists in mapping reads against a non-redundant gene catalog. Then, sample microbial richness is inferred from the total number of detected genes (gene richness) [1] or from the number distinct species traced with specific marker genes (species richness) [4].

However, this method provides accurate results only if the reference is representative of samples microbial composition. References can be continuously improved by adding genes from newly sequenced samples. Yet, frequent updates are computationally intensive and lead to traceability issues. As the number of samples grows, gene catalogs may become very large and difficult to manipulate, more specifically when they combine data from different hosts and bodysites [5].

We introduce ESKRIM, a tool that accurately compare microbial richness in shotgun metagenomic samples without relying on any reference. It implements an innovative algorithm computing the number distinct k-mers in a sample (k-mers richness). On real metagenomic data for which a highly representative reference is available, we show that k-mers richness strongly correlates with gene and species richness (Pearson's R = 0.96). In addition, ESKRIM provides better results when the available references are representative only of a fraction of the studied samples. As illustrations, we compare microbiome richness across populations (industrialized and rural), bodysites (gut, skin, vagina and mouth) and hosts (human, mouse, chicken and pig).

EKSRIM is implemented in Python3 and relies on the Jellyfish2 API for k-mers counting. The tool achieves good performance while we did not focus on that aspect at first. A metagenomic sample with 10M reads is processed in about ten minutes (RAM disk, Intel Xeon E5-2680 CPU, 4 threads).

ESKRIM is open-source and freely available at https://forgemia.inra.fr/florian.plaza-onate/eskrim

References

1. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, et al. Richness of human gut microbiome correlates with metabolic markers. Nature. 2013;500:541–6.
2. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, et al. Alterations of the human gut microbiome in liver cirrhosis. Nature [Internet]. 2014;513:59–64.
3. Wen C, Zheng Z, Shao T, Liu L, Xie Z, Le Chatelier E, et al. Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. Genome Biol. 2017;
4. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol . 2014
5. Global Microbial Gene Catalog [Internet]. Available from: http://gmgc.embl.de/

Poster 85

Mobidetails, interprétation en ligne des variations génétiques

David Baux[1], Charles Van Goethem[1], Thomas Guignard[2], Olivier Ardouin[3], Michel Koenig[1] Anne-Françoise Roux[1]

[1] Laboratoire de Génétique Moléculaire, CHU de Montpellier, Université de Montpellier, IURC 641 av Giraud, 34093, Montpellier, France

[2] Unité de Génétique Chromosomique, CHU de Montpellier, Université de Montpellier, Hôpital Arnaud de Villeneuve 371 av Giraud, 34090, Montpellier, France

[3] Plateau de Médecine Moléculaire et Génomique, CHU de Montpellier, Université de Montpellier, 371 av Giraud, 34090, Montpellier, France

: d-baux@chu-montpellier.fr

L'utilisation en clinique des technologies de séquençage haut-débit et les études génomiques de cohortes ont profondément modifié la pratique du diagnostic moléculaire, discipline s'efforçant d'établir les causes moléculaires de pathologies. L'interprétation clinico-biologique des variants d'ADN identifiés dans ce cadre constitue l'une des étapes chronophages de l'analyse. De multiples outils d'aide ou de prédiction existent, et les sources de données sont nombreuses et variées, mais y accéder de manière efficace requiert l'utilisation de suites logicielles commerciales ou soumet l'utilisateur à diverses contraintes notamment l'utilisation de données personnelles.
MobiDetails est une application web libre d'accès pour un usage académique permettant sur une seule page d'agréger un nombre important de données par variant. L'utilisateur a la possibilité de créer dans le système des variants (via VariantValidator) dans le gène de son choix (en utilisant la nomenclature HGVS transcrit (c.)). Le variant peut être de tout type (exonique, intronique, substitution, délétion, duplication...). Une fois créé, le variant est connu du système et pourra par la suite être rappelé via le moteur de recherche intégré. Pour chaque gène, une page permet d'afficher les variants déjà connus par catégories. L'outil présente pour chaque variant une synthèse incluant notamment :

— les différentes nomenclatures HGVS (génomiques, protéique...)

— l'interprétation des positions dans le gène (exon/intron), la protéine, le contexte nucléotidique, la proximité de sites d'épissage...

— les fréquences globales dans gnomAD exome/genome, l'identifiant dbSNP

— l'interprétation Clinvar

— des liens directs vers les instances LOVD et les résumés pubmed recensant le variant (via LitVar)

— les scores dbscSNV et spliceAI pour les prédictions d'épissage pour les substitutions

— divers scores de prédictions (SIFT, Polyphen2, metaSVM, FATHMM, REVEL...) pour les faux-sens, données issues de dbNSFP

Cette synthèse est exportable au format pdf. Une API associée permet la création de variants par lots. Les utilisateurs ont la possibilité de créer des comptes autorisant notamment l'ajout de classification ACMG propres aux variants ou de suivre des variants particuliers. Un aspect communautaire est intégré puisqu'il est possible de contacter les autres utilisateurs via l'application.
L'application est disponible à l'adresse suivante :
https://mobidetails.iurc.montp.inserm.fr/MD/
Exemple de variant
Code source

Poster 86

IMGT/RNAseq-ImmunoProfile

Benjamin Viart[1], Veronique Giudicelli[1], Elina Alaterre[1], Jerome Moreaux[12], Sofia Kossida[1]

[1]IGH, Univ Montpellier, CNRS, Montpellier, France

[2] CHU Montpellier, Laboratory for Monitoring Innovative Therapies, Department of Biological Hematology, Montpellier, France

1 Introduction

ImmunoGlobulin (IG) and T cell receptors (TR) genes are specific to immunogenetics and are composed of different gene types including : variable (V), diversity (D), joining (J), constant (C). The V(D)J recombination is a unique mechanism of genetic recombination that occurs only in developing lymphocytes during the early stages of T and B cell maturation. It involves somatic recombination and is part of the process creating the highly diverse repertoire of IG and TRs found in B cells and T cells, respectively. This process is a defining feature of the adaptive immune system.

High-throughput profiling of immune receptors has become an important tool for studies of adaptive immunity and for the development of diagnostics, vaccines and immunotherapies. However it is still a challenge due to the recombination process. RNA sequencing has been rapidly adopted for transcriptome's profiling of normal and tumoral cells in cancer studies, including multiple myeloma. This hematological malignancy is characterized by an accumulation of plasma cells in bone marrow. Plasma cells represent the terminal stage of B cell differentiation and normally produce [1]. Of particular interest is the discovery of clonality and V,D and J alleles diversity and expression levels.

IMGT®, the international ImMunoGeneTics information system is the global reference in immunogenetics and immunoinformatics. IMGT® [2] is a high-quality integrated knowledge resource specialized in the immunoglobulins, T cell receptors, major histocompatibility of human and other vertebrate species. We have developed a pipeline that uses Mixcr [3] and IMGT/HighV-QUEST [4] to extract from bulk RNA sequencing data immune system related information including clonotypes, genes and alleles frequency, CDR3 sequences and CDR3 length statistics. This pipeline is being tested using multiple myeloma patient data in order to investigate possible diagnostic and prognostic value within the results.

3 Bibliography

1. Raab MS, Podar K, Breitkreutz I, Richardson PG, Anderson KC. Multiple myeloma. Lancet. 2009;374: 324–339.

2. Lefranc M-P, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, et al. IMGT®, the international ImMunoGeneTics information system® 25 years on. Nucleic Acids Res. 2015;43: D413–22.

3. Bolotin DA, Poslavsky S, Davydov AN, Frenkel FE, Fanchi L, Zolotareva OI, et al. Antigen receptor repertoire profiling from RNA-seq data. Nat Biotechnol. 2017;35: 908–911.

4. Alamyar E, Duroux P, Lefranc M-P, Giudicelli V. IMGT(®) tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. Methods Mol Biol. 2012;882: 569–604.

Poster 88

Structural prediction of macromolecular interactions using evolutionary information

Chloé QUIGNOT[1], Aravindan Arun NADARADJANE[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

Macromolecular interactions are central to most biological processes. Protein docking aims to predict the most likely structural binding modes of interacting protein partners. Understanding how binding partners coevolved can provide essential clues to improve the structural prediction of protein interfaces. In the past few years, our team has contributed to the improvement of protein docking methods by combining evolutionary information with more traditional approaches.

2. Results and perspectives

We analyzed the way interface structures coevolved [1] and developed InterEvDock2, a server for protein-protein docking designed to integrate evolutionary information in the docking process [2,3]. InterEvDock2 uses one of the most successful interface sampling algorithms to date, which relies on fast Fourier transforms, followed by a consensus scoring scheme to discriminate correct from incorrect interfaces. We benchmarked InterEvDock2 on a large dataset of docking targets based on unbound homology models [4]. We successfully applied this pipeline to targets of the international CAPRI assembly prediction challenge [5,6] and to various biological applications [7]. Perspectives of this work include considering evolutionary information at the atomic scale through explicit modeling of homologous complex structures, integrating other types of constraints such as mutant screening using deep mutational scanning approaches and extending our structural prediction pipeline to protein-RNA interactions.

Acknowledgements

This work was supported by the French National Research Agency under grant ANR-15-CE11-0008-01 CHIPSeT to R.G., by CEA doctoral funding to A.N. and H.B., by IDEX Paris-Saclay doctoral funding to C.Q. It was granted access to the HPC resources of CCRT under allocation 2017-7078 by GENCI (Grand Equipement National de Calcul Intensif).

References

1. Jessica Andreani, Guilhem Faure and Raphael Guerois. Versatility and invariance in the evolution of homologous heteromeric interfaces. *PLoS Comput Biol*, 8(8):e1002677, 2012.
2. Jinchao Yu, Marek Vavrusa, Jessica Andreani, Julien Rey, Pierre Tufféry and Raphael Guerois. InterEvDock: a docking server to predict the structure of protein-protein interactions using evolutionary information. *Nucleic Acids Res*, 44(W1):W542-9, 2016.
3. 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.
4. 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.
5. Jinchao Yu, Jessica Andreani, Françoise Ochsenbein and Raphael Guerois. Lessons from (co-)evolution in the docking of proteins and peptides for CAPRI Rounds 28-35. *Proteins,* 85(3):378-390, 2017.
6. 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*, in press, 2019.
7. Alessandro Berto, Jinchao Yu, Stéphanie Morchoisne-Bolhy, Chiara Bertipaglia, Richard Vallee, Julien Dumont, Françoise Ochsenbein, Raphael Guerois and Valérie Doye. Disentangling the molecular determinants for Cenp-F localization to nuclear pores and kinetochores. *EMBO Rep*, 19(5), 2018.

Poster 89

Using positional information for predicting transcription factor binding sites

Raphaël Romero[1,2], Jean-Michel Marin[1], Sophie Lèbre[1,4], Charles-Henri Lecellier[3] and Laurent Bréhélin[2]

[1] IMAG, Univ. Montpellier, CNRS, Montpellier, France
[2] LIRMM, Univ Montpellier, CNRS, Montpellier, France
[3] Institut de Génétique Moléculaire de Montpellier, Univ. Montpellier, CNRS, Montpellier, France
[4] Univ. Paul-Valéry-Montpellier 3, Montpellier, France

Corresponding author: raphael.romero@umontpellier.fr

Transcription factors (TF) play a central role in the mechanism of tran- scription. These proteins bind the DNA sequence at particular binding sites. Binding sites are resumed in probabilistic model known as binding motifs or Position Weight Matrix [1]. Such motif can be used to compute binding affini- ties and to identify potential binding sites of the associated TF. However this approach has usually low accuracy, with lot of false positives.

In order to provide more accurate predictions of TF binding sites, we recently proposed a method that uses the fact that TFs do not bind DNA in an isolated way but in combination with others TFs. This method, named TFcoop [2], bases its prediction upon the binding affinity of the target TF as well as any other TF identified as cooperating with the target. Given a set of positive and negative sequences obtained from ChIP-seq experiments, TFcoop uses a logistic model trained with a LASSO regularization [3] for selecting the cooperating TFs. The approach outperforms the classical TF binding prediction methods and allows the identification putative cooperating TFs.

Here we introduce a more refined method that complements the TF binding affinity with positional information of the potential binding sites. Our aim is to study the importance of such information for predicting TF binding. In order to consider several subsequences of the original sequence while avoiding prohibitive computing time, we developed a segmentation algorithm based on a lattice. The selected subsequences are used to create new features that are added to the logistic model.

In addition, by centering the sequences on the binding site of the targeted TF, this segmentation algorithm enables us to consider the relative position between TFs' binding sites. This information is particularly relevant for specific biological questions, and can be used in different classification problems like cell type specific TF binding. The relative position information already pointed out cell-type specific cooperativity between TFs in some experiments. We are currently exploring 12 TFs on 90 ChIP-seq experiments.

References

[1] Wyeth W. Wasserman and Albin Sandelin. Applied bioinformatics for the identification of regulatory elements. Nature Reviews Genetics, 5(4):276-287, April 2004.

[2] Jimmy Vandel, Oceane Cassan, Sophie Lebre, Charles-Henri Lecellier, and Laurent Brehelin. Probing transcription factor combinatorics in different promoter classes and in enhancers. March 2018.

[3] Trevor Hastie, Sami Tibshirani, and Jerome Friedman. Elements of Statistical Learning: data mining, inference, and prediction. 2nd Edition., 2009.

Poster 91

Detection of m⁶A RNA modifications using Nanopore direct RNA sequencing

Charlotte BERTHELIER[1], Médine BENCHOUAIA[1], Corinne BLUGEON[1], Rana JURDAK[2], Naira NAOUAR[3], Christophe ANTONIEWSKI[3] and Christophe BAILLY[2]

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

[2] Sorbonne Université, CNRS, Laboratoire de Biologie du Développement, F-75005 Paris, France

[3] Sorbonne Université , CNRS, Institut de Biologie Paris Seine (IBPS),ARTbio Bioinformatics Analysis Facility, Paris, France

Corresponding Author: charlotte.berthelier@bio.ens.psl.eu

Summary

The study of RNA modifications is crucial to better understand gene expression. N6-methyladenosine (m⁶A) is the most prevalent modification in mRNA ; it has a key role at several levels of mRNA post-transcriptional regulation (splicing, translation, stability…). We aim at detecting m⁶Ausing direct RNA sequencing from Oxford Nanopore Technologies (ONT), and evaluate the bioinformatics tools available to identify this RNA modification. Our experimental design is based on the comparison of two conditions: RNA from seeds of a wild-type control sample of the model plant *Arabidopsis thaliana* treated by ethylene and of a mutant which is supposed to contain low level of methylated mRNA.

As of today, only a handful of software have been developped to identify m⁶A methylation on direct RNA sequencing data. EpiNano, developped by ONT, identifies RNA modifications directly from raw sequencing FAST5 files. It is supposed to predict m⁶A RNA modifications with high accuracy thanks to its training with m6A-modified and unmodified synthetic sequences [1]. We are also considering testing various basecallers (Guppy, Nanoflit and Tombo) on raw sequencing data to find the most appropriate basecaller for m⁶A methylation detection. This was because it was shown that base-callers increased mismatch frequencies in m⁶A-modified data-sets (with the largest increased in A positions) and decreased qualities [2]. Our final goal is to provide the developed m⁶A methylation identification pipeline to the community on GitHub [3], and offer this service to the IBENS genomics core facility users.

This work is a collaborative project between two teams from Sorbonne Université (Christophe Bailly's Seed Biology team [4] and ARTbio bioinformatics analyses platform [5]) and the genomics core facility of the Institut de Biologie de l'École normale supérieure (IBENS) [6,7].

References

[1] Parker, Matthew T., et al. "Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification." *eLife* 9 (2020).
[2] Liu, Huanle, et al. "Accurate detection of m 6 A RNA modifications in native RNA sequences." *Nature communications* 10.1 (2019): 1-9.
[3] https://github.com/GenomicParisCentre/
[4] https://www.ibps.upmc.fr/en/research/developmental-biology-laboratory/seed-biology
[5] https://www.artbio.fr/
[6] http://genomique.bio.ens.psl.eu
[7] Twitter @Genomique_ENS

Poster 92

Evaluation of isoform characterisation tools using long cDNA sequences

Sophie LEMOINE[1], Ammara MOHAMMAD[1], Corinne BLUGEON[1], Laurent JOURDREN[1]

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

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

Precise transcript and isoform identification are a real challenge with short read sequencing. As a sequencing facility, we sequence full-length cDNAs to directly access isoforms using Oxford Nanopore Technologies (ONT) sequencing. In addition to sequencing, our platform provides the associated bioinformatics analyses. We thus need to evaluate the available software to integrate in our analysis pipeline.

We rely on a benchmark dataset, obtained with samples that we have repetitively used over the years, each time we needed to test a new protocol or a technology enhancement. The samples are Egr2 KO mouse sciatic nerve and WT mouse sciatic nerve; the design involves triplicates to perform differential analyses on the dataset. We have a deep knowledge of our dataset at the gene level and we want to go further at the transcript level.

We performed RNA-seq with the ONT MinION. In the past years, we optimized a cDNA protocol, allowing us to sequence very long cDNAs from mouse samples. The sequence mean length is about 2.6kb, what we think very close from the expected length (2.7 kb) for mouse transcripts. The 5'-3' transcript coverage appears good and homogeneous, even compared to short-read RNA-Seq data.

These data were used to make an evaluation of the software available in the literature (StringTie2[1], Pinfish[2], FLAIR[3], TALON[4], SQANTI2[5], UNAGI[6]…). These algorithms do not have the same aims. Some focus on building transcripts for get a better annotation, while others build transcripts to quantify expression and perform differential analyses at the isoform level (FLAIR). To achieve isoform discovery and identification, these algorithms use different strategies. Some build consensus transcripts without any other information than alignments on the genome (Pinfish and FLAIR) whereas some aggregate any supplementary data and analyses to validate and reinforce the resulting transcripts (TALON, StringTie2 and SQANTI2).

This evaluation will allow us to define the best tool(s) to provide isoform characterization and analyses to the IBENS Genomic Core facility[7] users.

References

1. Kovaka, S. et al. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. 34, 748–29 (2019).
2. https://github.com/nanoporetech/pinfish
3. Workman, R. E. et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat. Methods 16, 1297–1305 (2019).
4. Wyman, D. et al. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. 19, A68–48 (2019).
5. https://github.com/Magdoll/SQANTI2
6. Kadi, Al, M. et al. UNAGI: an automated pipeline for nanopore full-length cDNA sequencing uncovers novel transcripts and isoforms in yeast. Funct. Integr. Genomics 7, 11706 (2020).
7. http://genomique.bio.ens.psl.eu

Poster 93

The IBENS Genomics core facility

Laurent JOURDREN[1], Médine BENCHOUAIA[1], Charlotte BERTHELIER[1], Corinne BLUGEON[1], Karine DIAS[1], Sophie LEMOINE[1], Catherine SENAMAUD-BEAUFORT[1], and Stéphane LE CROM[1,2]

[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 have been focused on **eukaryotes** and specifically on **functional genomics** analyses since the beginning. We handle classical model organisms and also 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 remaining half being involved on the data analysis part. Our goal is to help 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 we have been following the **ISO 9001** quality international standard since March 2013.

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 a lot of our 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. Our concern is to develop workflows to achieve **reproducible and transparent data analysis** of our high throughput experiments.

Since 2016, our facility has been developing two new technologies. The first one is devoted to **single cell RNA-seq** with the buying of a **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.

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/

Poster 94

Reconstruction of hidden data in chromosome contact maps

AXEL BREUER[1,3] AXEL COURNAC[1,2]

[1] Institut Pasteur, Unité Régulation Spatiale des Génomes, 28 rue du Docteur Roux, 75015,Paris, France

[2] Institut Pasteur, Computational Biology Department (CBD), Paris, France

[3] ENGIE, Global Energy Management, Paris, France

Corresponding Authors: axel.breuer@engie.com, acournac@pasteur.fr

Abstract *The spatial organisation of chromosomes may impact or be impacted by major biological functions such as gene expression, replication or chromosome segregation. To observe and study 3D structure of chromosomes, so-called contact techniques (3C, Hi-C, ChIA-PET) are being developed in parallel with microscopy. They are based on the capture and quantification of physical contact between different loci within a genome and bring a new type of information to an unprecedented spatial resolution. These techniques generate millions pairs of short sequences (~ 50 nucleotides), a certain proportion of which cannot be located directly due to their repetition in the sequence of the reference genome (several alignments are possible). To overcome this limitation, we propose the Apollo method, which uses statistical inference and inpainting methods to predict the contacts of the repeated sequences and thus reveal the hidden side of chromosomes. Unpublished results will be presented with simulated data and applications on micro-organisms contact maps.*

Keywords Chromosome organisation- repeated sequences – statistical inference – contact data – Hi-C – Microbiology -

Poster 95

PEWO: a set of procedures to benchmark species identifications based on phylogenetic placement

Benjamin Linard[1,2], Nikolai Romashchenko[1], Fabio Pardi[1] and Eric Rivals[1]

[1] LIRMM, Univ. Montpellier, equipe MAB, 860 rue de St Priest, 34095 Montpellier, France
[2] SPYGEN, 17 rue du Lac Saint-André, 73051 Le Bourget du Lac, France

Corresponding Author: benjamin.linard@lirmm.fr

Metagenomic and metabarcoding projects, whether related to ecological studies, biodiversity exploration or medical diagnostics, are concerned by the critical step of species identifications. Identifications are generally operated via 1) sequence clustering or 2) local alignment of sampled markers to very large references databases (NCBI or marker-specific databases). In an attempt to overcome limitations related to incomplete reference databases, results are often refined by contextualization in a taxonomy [1] but ignore the potential resolution brought by more advanced phylogenetic models and reference phytogenetic trees.

An alternative remains in using phylogenetic placement (PP) [2], in which query reads are "placed" on the branches of a reference phylogeny. Recently, it attracted much attention. New implementations [3] or novel algorithms [4,5] made PP scalable to current sequencing volumes (10^6 reads placed on a tree in <30 min). With already 7 proposed algorithms, 5 different implementations and 3 fundamentally different approaches (distance-based, alignment-based, alignment-free), today it may be hard to judge which PP solution best fits a particular study.

For this reason, we developed PEWO (Placement Evaluation WOrkflows). This set of experimental procedures aims to answers the classical questions that arise when phylogenetic placement s chosen as a solution for taxonomic identification :

1) For end-user: which PP accuracy can be expected for different reference phylogenies based on different taxonomic markers (16S, cox1) ? PEWO can highlight which locus will likely produce the best identifications under different PP approaches.

2) For data analysts: Which PP solution and which parameter combinations should be selected to attain the desired [accuracy vs computational cost] trade-off. Some solutions are more adapted to repeated analyses, some other to longer reads... PEWO facilitates the benchmarking of existing solutions.

3) For developers: PEWO provide a library for testing and comparing new and old PP solutions under the same framework. It also aims to become a community effort to support future evaluation procedures and future PP implementations.

In my poster, I will present the PEWO package and describe some applications on real datasets such as mitochondrial datasets for Coleopteran species identification.

References

[1] *MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data*. Huson DH,et al. PLoS Comput Biol. 2016
[2] *pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree.* Matsen FA, et al. BMC Bioinformatics. 2010
[3] *EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences.* Barbera et al. Syst Biol. 2019.
[4] *Rapid alignment-free phylogenetic identification of metagenomic sequences.* Linard B et al. Bioinformatics. 2019
[5] *APPLES: Scalable Distance-based Phylogenetic Placement with or without Alignments.* Balaban M, et al. Syst Biol. 2019

Poster 97

Draft genome sequences of *Blastocystis* subtypes 1 and 8, and comparative analysis.

Ivan Wawrzyniak[1], Amandine Cian[2], Céline Nourrisson[1], Aldert Bart[3], Magali Chabé[2], Philippe Poirier[1], Tom van Gool[3] , Eric Peyretaillade[1] , Eric Viscogliosi[2] , Frédéric Delbac[1].

[1] Laboratoire Microorganismes Génome et Environnement, Université Clermont Auvergne, CNRS, UMR 6023, 1 impasse Amélie Murat, F 63000, Clermont-Ferrand, France

[2] Center of Infection and Immunity of Lille, Institut Pasteur de Lille, Inserm U1019, CNRS, UMR 9017, University of Lille, CHU of Lille, F 59019, Lille, France

[3] Department of Medical Microbiology,Amsterdam Medical Center, 1105 AZ, Amsterdam, The Netherlands

Corresponding Author: ivan.wawrzyniak@uca.fr

Blastocystis is a highly prevalent anaerobic eukaryotic parasite found in the intestinal tract of human and various animals. Although the role of *Blastocystis* as human pathogen remains unclear, it can cause acute or chronic digestive disorders and some studies have suggested an association with irritable bowel syndrome. Seventeen subtypes (ST1-ST17), among which ten (ST1 to ST9 and ST12) are found in human with varying prevalence, have been identified based on the small-subunit ribosomal RNA. The genomes of three isolates belonging to ST1 [1], ST4 [2] and ST7 [3] have been previously sequenced and annotated. Using illumina HiSeq 2000 system technology, we conducted genome sequencing and annotation of a new *Blastocystis* ST1 isolate and one ST8 isolate. Assembly of reads generated genomic sequence of 15.01 Mbp possessing 6604 putative genes and 13.9 Mbp for 5579 putative genes in ST1 and ST8, respectively. The Mitochondrion-Like Organelle (MLO) genome full sequences were obtained for both ST1 (28.3 kbp) and ST8 (27.9 kbp). These circular mitochondrial genomes encompass 10 NADH subunits, 13 ribosomal proteins and 4 ORFs with unknown functions and present a highly conserved gene synteny. Comparative analysis of whole proteomes identified the core and the specific proteins between the five sequenced ST. Analysis of the core proteins revealed an over representation of gene families coding for proteins that may be involved in virulence including hydrolases. Some of these proteins are predicted to be secreted or targeted to the plasma membrane and may play important roles in processes such as cytoadherence, host cell invasion, molecule degradation, or host immune response evasion. A whole genome phylogeny of *Blastocystis* confirmed the *Blastocystis* lineage based on the complete SSU rDNA [4]. Our genomic comparative analysis also revealed a highly conserved gene synteny between the 5 ST. Finally, we demonstrated for the first time the presence of transposable elements in *Blastocystis* using TransposonPSI software and a tblastX approach.

References

[1] Eleni Gentekaki, Bruce A Curtis, Courtney W stairs, Vladimir Klimes, Marek Elias, Dayana E Salas-Leiva, Emily K Herman, Laura Eme, Marias C Arias, Bernard Henrissat et al. Extreme genome diversity in the hyper-prevalent parasitic eukaryote *Blastocystis*. *Plos Biology*, (46/47):217–252, 2017.

[2] Ivan Wawrzyniak, Damien Courtine, Marwan Osman, Christine Hubans-Pierlot, Amandine Cian, Céline Nourrison, Magali Chabé, Philippe Poirier, Aldert Bart, Valérie Polonais et al. Draft genome sequence of the intestinal parasite *Blastocystis* subtype 4-isolate WR1. *Genomic Data*. Feb 2; 4:22-3, 2015.

[3] France Denoeud, Michael Roussel, Benjamin Noël, Corinne Da Silva, Marie Diogon, Eric Viscogliosi, Céline Brochier-Armanet, Arnaud Couloux, Julie Poulain, Béatrice Segurens et al. Genome sequence of the stramenopile Blastocystis, a human anaerobic parasite. *Genome Biology*, 12(3): R29, 2011.

[4] Hisao Yoshikawa, Yukiko Koyama, Erika Tsuchiya, Kazutoshi Takami. Blastocystis phylogeny among various isolates from humans to insects. *Parasitology International*, Dec;65:750–759, 2016.

Poster 98

Application of the Modular Response Analysis method to highly stable biological systems

Meriem MEKEDEM[1,2], Gabriel JIMENEZ DOMINGUEZ[1,2], Patrice RAVEL[1,2] and Jacques COLINGE[1,2]

[1] Institut de Recherche en Cancérologie de Montpellier, 208 avenue des apothicaires, 34298 Cedex 5, Montpellier, France

[2] Université de Montpellier, 163 rue Auguste Broussonnet, 34090, Montpellier, France

Corresponding Author: jacques.colinge@inserm.fr

Cell biology is governed by an intricate network of interactions between various molecules that establish various biochemical fluxes and regulatory mechanisms. To investigate and eventually understand the emergent global behavior arising from such networks, we seek to use the computational approach termed Modular Response Analysis (MRA) [1]. MRA allows reconstructing network topologies with information regarding edge orientations and strengths from systematic perturbation experiments.

However, MRA, like any method for solving reverse engineering problems, faces difficulties with poorly conditioned problems. In such cases, noise and potential measurement biases induce massive errors in the reconstructed solutions. Application of MRA to highly-stable biological systems, such as certain metabolic pathways, may cause poorly conditioned linear algebraic equations since perturbation experiments induce very small changes in the observed data. We tried to extend applicability of MRA to such cases by introducing matrix preconditioning, which we illustrate by investigating a simplified tricarboxylic acid (TCA) cycle for autotrophic tissues [2].

References

1. Boris N Kholodenko *et al.* Untangling the wires: a strategy to trace functional interactions in signaling and gene networks. *Proc Natl Acad Sci*, *U S A,* 99, 12841-12846, 2002.
2. Ralf Steuer *et al.* From structure to dynamics of metabolomic pathways: application to the plant mitochondrial TCA cycle. *Bioinformatics,* 1378-1385, 2007.

Poster 99

Towards exhaustive characterization of structural variation in animals: example of cattle.

Arnaud Di Franco[1], Sarah Djebali[2], Camille Eche[3], Denis Milan[3], Didier Boichard[4], Christine Gaspin[1], Cécile Donnadieu[3], Carole Iampietro[3], Christophe Klopp[1] and Thomas Faraut[2]

[1] MIAT, INRAE, 24 Chemin de Borde Rouge, 31320 Auzeville-Tolosane, Toulouse, France
[2] GenPhyse, INRAE, 24 Chemin de Borde Rouge, 31320 Auzeville-Tolosane, Toulouse, France
[3] US 1426, GeT-PlaGe, Genotoul, INRAE, 24 Chemin de Borde Rouge, 31320 Auzeville-Tolosane, Toulouse, France
[4] GABI, INRAE, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France.

Corresponding author: arnaud.di-franco@inrae.fr

Structural variations (SVs) have been proved to be of high importance in both medicine and molecular biology. They are now recognized as the major source of interindividual genetic variation [1] and are major actors in various diseases, gene regulation, and consequently in genome evolution [2,3]. Structural variation are defined as variations ranging from 50 bp up to over megabases of sequence [4] and, as such, can be difficult to infer with traditional sequencing technologies (i.e. short-read) [5].

The landscape of sequencing technologies is quickly moving and several solutions already exist to provide sequence over many kilobases (i.e. Oxford Nanopore and PacBio) or to gather long range information between molecules (Linked reads, optical mapping, HiC). These new technologies improve significantly our ability to detect and characterize structural variations [6]. We proposed, in the context of the SeqOccIn project, aiming at addressing the added value of long read technologies for the understanding of genome variability, epigenetics and metagenomic in agronomy (https://get.genotoul.fr/seqoccin), to study the structural variability of two important species, bovine and maize.

In the line of the giab project (https://jimb.stanford.edu/giab) we propose to take advantage of the large number of sequencing technologies used in the SeqOccIn project (Nanopore, PacBio, Chromium 10X, Optical mapping) to comprehensively characterize the genomic structural variability of a few individuals. Here, we show the results of our analysis on SVs detection using data sequenced from a trio of bovine individuals. First, we present our study of bioinformatic pipelines within and between technologies, explaining the pros and cons for the different types of SVs. Then, with the help of all the SVs detected previously, we refined a set of genuine variants to be used as benchmark for bovine SV detection. Finally, we test our benchmark by comparing statistical values obtained with various pipelines to those obtained on comparable human benchmark [7].

Acknowledgements

This work was supported by the SeqOccIn project funded by the Occitanie Region through a FEDER funding, four INRAE laboratories as well as 14 companies.

References

[1] Donald F. Conrad, Dalila Pinto, Richard Redon, et al. Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–712, apr 2010.

[2] Peter H. Sudmant, Tobias Rausch, Eugene J. Gardner, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81, oct 2015.

[3] Steve S. Ho, Alexander E. Urban, and Ryan E. Mills. Structural variation in the sequencing era. *Nature Reviews Genetics*, 21(3):171–189, mar 2020.

[4] Monya Baker. Structural variation: the genome's hidden architecture. *Nature Methods*, 9(2):133–137, feb 2012.

[5] Medhat Mahmoud, Nastassia Gobet, Diana Ivette Cruz-Dávalos, et al. Structural variant calling: the long and the short of it. *Genome Biology*, 20(1):246, dec 2019.

[6] Mark J. P. Chaisson, Ashley D. Sanders, Xuefang Zhao, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications*, 10(1):1784, dec 2019.

[7] Justin M. Zook, Nancy F. Hansen, Nathan D. Olson, et al. A robust benchmark for germline structural variant detection. *bioRxiv*, page 664623, jul 2019.

Poster 101

The antibiotic resistance mobilome in biofilms: Bioinformatic strategies.

Sophia Achaibou[1], Elena Buelow[2], Sean Kennedy[3], Olivier Chesneau[4], Catherine Dauga[1, 3]

[1] Biomics Pôle – Institut Pasteur, 25-28 rue du Docteur Roux, 75015 Paris, FRANCE
[2] Université Limoges, INSERM – CHU Limoges, UMR 1092, Limoges, FRANCE
[3] Department of Computational Biology – Institut Pasteur, 25-28 rue du Docteur Roux, 75015 Paris, France
[4] Collection de l'Institut Pasteur - Institut Pasteur, 25-28 rue du Docteur Roux, 75015 Paris, France

Corresponding Author: sophia.achaibou@pasteur.fr

Today, antibiotic resistance poses a serious threat to global health. Reservoirs, sources and distribution of Antibiotic Resistance genes (ARGs), as the effect of human activities on the selection and dissemination of ARGs, are not yet well comprehended.

Here, we studied the expression of ARGs and mobile elements (MGEs) in response to stress due to antibiotic exposure in hospital and urban wastewater biofilms.

The metatranscriptomic approach used in this study involves extraction and analyzing messenger RNA (mRNA) providing information about genes actively expressed in complex microbial communities. This approach, in contrast to metagenomics, allows to explore active microbial *in situ* functions. Moreover, it may be able to assess subtle changes in gene expression in different environmental contexts [2].

We developed strategies to detect expression of low abundant and allelic variants of ARGs and MGEs. The characterization of ARGs was improved by separating active genes on mobilome, potentially acquired by transfer, and genes inherited by descent, linked more closely to variation of microbiome composition.

After identifying critical steps in the bioinformatics process, we developed a pipeline for ribosomal RNA subtraction *in silico* to reduce the size of samples, decreasing informatics resources required and improving analysis time. For assembly, we chose the SPAdes pipeline allowing to obtain contigs large enough for all the samples [1]. The mapping process was assessed on a specialized database of 88 genes targeted by a high-throughput RT-qPCR assay. We also developed a pipeline analyzing SAM files to make link between mobilome and resistome data. Finally, the Snakemake pipelines were applied on the metranscriptomes of wastewater biofilms.

We found that hospital wastewater contained high abundance of ARGs (4.3%) compared to urban wastewater (0.01%). Nature of ARGs expressed in hospital effluent and urban effluent clearly differed. A greater proportion of total ARGs was associated with plasmids in hospital compared to urban wastewater.

1. Bankevich *et al. J Comput Biol.* (2012) 19(5): 455–477

2. Tsementzi *et al.*, *Environmental Microbiology Reports* (2014) 6(6), 640–655.

Poster 102

An evaluation of binning methods to recover human gut microbial pan-genomes from non-redundant reference gene catalog

Marianne BORDERES[1,2,3], Cyrielle GASC[1], Emmanuel PRESTAT[1], Mariana FERRARINI[2,4], Susana VINGA[5], Lilia BOUCINHA[1] and Marie-France SAGOT[2,3]

[1] MaaT Pharma, 317 Avenue Jean Jaurès, 69007 Lyon, France

[2] Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Évolutive UMR 5558, F-69622 Villeurbanne, France

[3] Erable team, INRIA Grenoble Rhône-Alpes, 655 Avenue de l'Europe 38330 Montbonnot Saint-Martin, France

[4] INSA-Lyon, 20 Avenue Albert Einstein, 69100 Villeurbanne, France

[5] INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Lisbonne, Portugal

Corresponding Authors: marie-france.sagot@inria.fr and mborderes@maat-pharma.com

Human gut microbiota exerts functions essential for the maintenance of host physiology. However, the characterization of host-microbiota interactions remains challenging, notably due to the difficulty of describing the functioning of microbial communities in reference-based quantitative metagenomics analyses. Indeed, taxonomic and functional analyses being realized independently, there is no link between microbial genes and species. Although a first set of species-level bins (metagenomics species – MGS) was built by clustering co-abundant genes [1], no reference MGS set is established based on the most used gut microbiota gene catalog – the Integrated Gene Catalog (IGC) [2]. Published benchmarking results focusing on the reconstruction of individual genomes from contigs have highlighted best-performing binning solutions but do not include methods clustering co-abundant genes.

In order to identify the best suitable and most accurate approach to group IGC genes, we benchmarked 9 taxonomy-independent binners implementing abundance-based, hybrid (abundance-based and composition-based) or integrative approaches. To this end, we built a simulated non-redundant gene catalog composed of 41 gut-associated microbial species and adapted a quality assessment tool [3] to evaluate binners on a non-redundant gene set.

The quality assessment results show that no hybrid or abundance-based binner performs best on all metrics. Overall, the best trade-off between the average purity and completeness per bin is achieved by an integrative method. With the aim to further explore the obtained binning results, we selected the best-performing binner for each category of approaches, and compared their results with our expected community structures, taking into account the characteristics of the genes and corresponding genomes included in our simulated catalog. Eventually, the three selected binners are distinguished by specific advantages, but also limitations inherent to their approach or in terms of scalability.

References

1. Nielsen, H.B *et al*. (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol*.,32(8):822-8.
2. Li, J *et al*. (2014) An integrated catalog of reference genes in the human gut microbiome. *Nat Biotechnol*., 32(8):834-41.
3. Meyer, F *et al*. (2018) AMBER: Assessment of Metagenome BinnERs. *GigaScience,*7(6):1-8

Poster 103

Evaluate the impact of abiotic stressors (heat and suboptimal diet) on the gut microbiota of two chicken breeds diverging in feed efficiency, using the 16S rRNA high-throughput sequence technology.

Alexandre LECOEUR[1], Maria BERNARD[1, 2], Sandrine LAGARRIGUE[3] and Tatiana ZERJAL[1]

[1] Univ. Paris-Saclay, INRAE, AgroParisTech, GABI, 78350, Jouy-en-Josas, France

[2] INRAE, SIGENAE, 78350, Jouy-en-Josas, France

[3] INRAE, Agrocampus-Ouest, PEGASE, 35590, Saint-Gilles, France

Corresponding Author: maria.bernard@inrae.fr

The globalization of poultry production exposes the animals to a variety of climatic and feeding constraints. For example, compared to the European production conditions, the chickens produced in the Southern Asian are reared under higher ambient temperatures and with a feeding regime characterized by a reduced energy formula. Although the impact of heat stress and suboptimal diets on chicken production and quality traits has been largely studied [1], little is known about the impact of these stressors on the gut microbiota composition of adult laying chickens. A large amount of evidences indicates that the maintenance of the gut microbiota homeostasis is essential to guarantee physical health and a functional immune system [2]. Several studies seem to indicate that abiotic stressors can alter the gut microbiota homeostasis inducing intestinal injury and epithelial barrier dysfunction, ultimately affecting the makeup of intestinal flora leading the animals to a higher susceptibility to gut pathogens.

In this study, we use the FROGS [3] metabarcoding pipeline to analyze the 16S rRNA sequence data obtained from the caecum of two experimental chicken lines divergent for feed efficiency [4] that were exposed to chronic heat stress or fed with a reduced energy feeding.The aim of this study is to quantify the genetic and abiotic stress impact on the composition of the caecal microbial community and test for genetic by environment (GxE) interactions. The availability of a large number of production and metabolic traits for these birds allows us also to correlate them with the caecal microbiota composition.

Preliminary results, show that the microbiota structure is different between the two chicken lines and that the stress impact on the microbiota is stress dependent.

Acknowledgements

This study is part of the ANR project "Chickstress" (https://anr.fr/Projet-ANR-13-ADAP-0014) and of the European project “Feed-a-Gene” (https://www.feed-a-gene.eu/).

References

1. Carrasco, Juan M. Diaz, et al, Microbiota, Gut Health and Chicken Productivity: What Is the Connection ?, *Microorganisms*, vol. 7, no. 10, Oct. 2019
2. Wu HJ, Wu E, The role of gut microbiota in immune homeostasis and autoimmunity. *Gut Microbes*. 2012;3(1):4–14.
3. F. Escudie, et al., FROGS: Find, Rapidly, OTUs with Galaxy Solution. *Bioinformatics*, 2018. 34(8): p. 1287-1294.
4. A. Bordas, et al., Direct and correlated responses to divergent selection for residual food intake in Rhode island red laying hens. *British Poultry Science*. 1992. 33: p. 741–54

Poster 104

On the inference of complex phylogenetic networks by Markov Chain Monte-Carlo

Charles-Elie Rabier[1,2], Vincent Berry[3], Jean-Christophe Glaszmann[4], Fabio Pardi[3] and Céline Scornavacca[1]

[1] ISE-M, Univ.Montpellier, CNRS, EPHE, IRD, Montpellier, France
[2] IMAG, Univ.Montpellier, CNRS, Montpellier, France
[3] LIRMM, Univ.Montpellier, CNRS, Montpellier, France
[4] AGAP, CIRAD, Montpellier, France

Corresponding author: charles-elie.rabier@umontpellier.fr

Complete genomes for numerous species in various life domains (Denoeud et al. 2014, Badouin et al. 2017, Garsmeur et al. 2018), and even for several individuals for some species (Hapmap Consortium 2003, 3000 Rice Genome Project 2014) are nowadays available thanks to next generation sequencing. To process such a large amount of data, methods need not only to be accurate, but also time efficient. We present here an efficient method dedicated to phylogenetic network inference.

In phylogenetics, species tree inference has been studied extensively for many years, and the theory behind it is relatively well known. However, a species tree is unable to model complex biological events such as horizontal gene transfer (e.g. procaryotes, Koonin et al. 2001, but also among eucaryotes, Szollhosi et al. 2015), hybridization (plants and animals, Mallet 2007), introgression (e.g. citrus, Minamikawa et al. 2017) and recombination. In contrast, phylogenetic networks, that differ from species trees because of reticulate edges, are able to capture all those phenomenons.

We present here a novel way to compute the likelihood of biallelic markers given a phylogenetic network. This computation is at the heart of a Bayesian network inference method – called SNAPPNET, as it extends the SNAPP method (Bryant et al., 2012). SNAPPNET is available as a package of the well-known Beast 2 software (Bouckaert et al., 2014 and 2019). This package partly relies on code from SNAPP method (Bryant et al., 2012) to handle sequence evolution and on code from SPECIESNETWORK (Zhang et al., 2018) to modify the network during the MCMC as well as to compute network priors.

Our approach differs from that of Zhang et al. (2018) in that SNAPPNET takes a matrix of biallelic markers as input while SPECIESNETWORK expects a set of nucleotide alignments for which it samples possible gene trees as part of its process. Thus, the considered substitution models differ but more importantly, our method does not need to consider gene tree inference as an intermediary step. Following SNAPP, SNAPPNET's computations integrate over all possible tree histories for a locus, while SPECIESNETWORK considers only a sample of locus trees from the infinite number of possible topologies and branch lengths.

SNAPPNET is much closer to the MCMCBiMarkers method of Zhu et al. (2018), which also extends the SNAPP method (Bryant et al., 2012) to network inference. Both methods take biallelic markers as input, rely on the same model of evolution and also both sample networks in a Bayesian framework. However, SNAPPNET is exponentially more efficient in computing likelihoods for non-trivial networks. Also, the methods differ in the way the Bayesian inference is conducted.

In this poster, we will describe SNAPPNET and compare its performances with MCMCBiMarkers on simulated data. We will also give an illustration on rice data.

Acknowledgements

This work was supported by the Key Initiative Muse Data Science (I-SITE MUSE: ANR-16-IDEX-0006) and by the project Genome Harvest ref. ID1504-006 ("Investissements d'avenir", ANR-10-LABX-0001-01).

References

D. Bryant et al. (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular biology and evolution*, 29(8), 1917-1932.

C. Zhang et al. (2018). Bayesian inference of species networks from multilocus sequence data. *Molecular biology and evolution*, 35(2), 504-517.

J. Zhu et al. (2018). Bayesian inference of phylogenetic networks from bi-allelic genetic markers. *PLoS computational biology*, 14(1), e1005932.

Poster 106

X!TandemPipeline new version brings native support of timsTOF raw data

Thomas Renne[1,2], Filippo Rusconi[1] Michel Zivy[1], Olivier Langella[1]

1. PAPPSO, Génétique Quantitative et Évolution, Ferme du Moulon, 91190, Gif-sur-Yvette, France

2. Master BioInformatique Modélisation et Statistiques, Université de Rouen Normandie, 76130, Rouen, France

Poster JOBIM

30^{th} June – 3^{rd} July

Poster 107

Automatic annotation of ICE and IME with the ICEscreen tool

Julie Lao[1,2], Thomas Lacroix[1], Gérard Guédon[2], Nathalie Leblond-Bourget[2] and Hélène Chiapello[1]

[1] Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France
[2] Université de Lorraine, INRAE, DynAMic, 54000, Nancy, France

Corresponding Author: julie.lao@inrae.fr

Abstract

ICEs (Integrative Conjugative Elements) and IMEs (Integrative Mobilizable Elements) are bacterial mobile that play a key role in horizontal transfers such as the dissemination of antibiotic resistance genes. They have the ability to integrate, excise and transfer themselves by conjugation from one bacteria to another.

Automatic identification of these highly prevalent but poorly known elements is challenging. Thus, they are currently not annotated in almost all public genomes.

So far, only 2 bioinformatics approaches allow the automatic detection of ICEs and the detection of IMEs, but with a low reliability [1,2]. All of them first co-locate ”Signature Proteins” (SPs) that are essential for a functional element. Search of element’s boundaries is then carried out, either by using closely related genomes of the same species to delineate ICEs with surrounding core genes [1] or at the nucleotide level by searching DNA repeats at both ends of ICEs and IMEs [2].

None of these approaches can detect accurately nested or tandem ICEs and IMEs, which are frequently observed in bacterial genomes.

Thus, we designed a 4-steps bioinformatics strategy implemented in the *ICEscreen tool*, that can detect both single ICE and IME, but also complex ICE/IME regions. Our approach co-localize up to 4 types of SPs that are part of either the transfer module or the integration module of an ICE or IME:

(i) Detection of SPs of of transfer and integration modules of ICEs and IMEs;

(ii) Co-localized SPs of potential transfer modules by a ”seed-and-extend” strategy applied to regions of co-localized SPs;

(iii) Recursive merging of partial potential transfer modules that allow to solve complex regions;

(iv) Aggregation of integration module to detected potential transfer modules that enables the classification of detected elements as ICEs or IMEs.

We will present the first results of the ICEscreen tool which detected 225 elements including 85 in complex regions in a set of 40 bacterial firmicutes genomes.

References

[1] Cury, J. *et al.* (2020). Identifying Conjugative Plasmids and Integrative Conjugative Elements with CONJscan. In *Horizontal Gene Transfer* (pp. 265-283). Humana, New York, NY.

[2] Liu, M. *et al.* (2019). ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. *Nucleic acids research*, 47: D660-D665.

Poster 108

Evaluation of tools and strategy for the assembly of an allopolyploid banana genome using nanopore technology

Anaïs LOUIS[1], Cédric MARIAC[2], François SABOT[2] and Mathieu ROUARD[1]

[1] Bioversity International, Parc Scientifique Agropolis II, 34397, Montpellier, Cedex 5, France

[2] Institut de Recherche pour le Développement (IRD), UMR DIADE, BP 64501, 34394 Montpellier, France

Corresponding Author: m.rouard@cgiar.org

Genome assembly will rapidly become a standard procedure for many organisms but still remain challenging for polyploid species, in particular for allopolyploid genomes (polyploids which evolved by the merger of two or more distinct species) such as Banana, Brassica, Cotton, Strawberry or Wheat. The emergence of third generation sequencing provides a solution to resolve assembly of complex genomes due to the production of longer reads than the second-generation technologies.

Recent studies have released polyploid genomes using a combination of technologies such as Illumina, 10X Genomics, PacBio, Bionano and Hi-C [1–4]. Here, we implemented a strategy to assemble an allopolyploid banana genome (ABB) based on Oxford Nanopore sequencing strategy only and the availability of ancestral A and B diploid genomes. We thus tested and compared several *de novo* assemblers [5, 6] with haplotypes-phasing software and RaGOO [7] for Reference-Guided Scaffolding to obtain the most possible contiguous genome assembly. An optimized SnakeMake pipeline (see Poster CulebrONT) will be used and adapted to polyploid genomes. This poster is an opportunity to present the result of our investigations that can be useful for any project planning to generate assemblies with Nanopore data in polyploid organisms.

Acknowledgements

We acknowledge Sebastien Carpentier for providing the plant material, Catherine Breton for some help with SNP calling with GATK. This work was supported by the iTrop high-performance cluster from IRD as part of the South Green Bioinformatic platform (www.southgreen.fr).

References

1. Edger PP, Poorten TJ, VanBuren R, Hardigan MA, Colle M, McKain MR, et al. Origin and evolution of the octoploid strawberry genome. Nat Genet. 2019;51:541–7.
2. Yang J, Liu D, Wang X, Ji C, Cheng F, Liu B, et al. The genome sequence of allopolyploid Brassica juncea and analysis of differential homoeolog gene expression influencing selection. Nat Genet. 2016;48:1225–32.
3. Bertioli DJ, Jenkins J, Clevenger J, Dudchenko O, Gao D, Seijo G, et al. The genome sequence of segmental allotetraploid peanut Arachis hypogaea. Nat Genet. 2019;51:877–84.
4. Zhang X, Zhang S, Zhao Q, Ming R, Tang H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nat Plants. 2019;5:833–45.
5. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37:540–6.
6. Vaser R, Šikić M. Yet another de novo genome assembler. bioRxiv. 2019;:656306.
7. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biology. 2019;20:224.

Poster 109

KmerTool, a new tool to quickly explore large RNA-Seq datasets

Sébastien Riquier[1,2], Benoit Guibert[1,2], Chloé Bessiere[1,2], Anne-Laure Bougé[1,2], Anthony Boureux[1,2,3], Florence Rufflé[1,2], Nicolas Gilbert[1,2], Daniel Gautheret[4,5] and Thérèse Commes[1,2,3]

[1] Institut de Médecine Régénératrice et de Biothérapie, INSERM U1183, Montpellier, France
[2] Plateforme bioinformatique Bio2M, Montpellier, France
[3] University of Montpellier, Montpellier, France
[4] Institute for Integrative Biology of the Cell, CEA, CNRS
[5] Université Paris-Sud, Université Paris-Saclay, Gif-sur-Yvette, France

Corresponding author: therese.commes@inserm.fr

The huge body of publicly available RNA-seq libraries is a treasure of functional informations allowing to explore RNAs tissue expression by quantification of known RNA species or emerging novel transcript variants. However, transcript quantification using classical approaches relies on alignment methods that require a lot of computational resources, long processing time and produce possible bias due to alignment issues. Such whole transcriptome approaches can not be easily adapted to the quantification of small sets of candidate transcripts in large datasets. Recent studies have demonstrated that k-mer decomposition constitutes a new way to process RNA-Seq data for the identification of transcriptional signatures as k-mers (or tags) and can be used to quantify gene expression in a more specific and less resource-consuming way than classical approaches. However, applying this method to a candidate gene approach will rely on the high specificity of the k-mers set that will be quantified. Here, we present KmerTool that includes: i/ a specific k-mers design [1], based on the decomposition of transcript sequences into k-mers, ii/ a subset selection of these k-mers, regarding their specificity into the reference genome and transcriptome, iii/ a counting step of the selected k-mers into RNA-seq datasets.

We propose to use our strategy to set-up a pipeline for RNA-seq data quality analysis. Indeed, using well defined sets of k-mers, we are able to predict metadata from public RNA-Seq data such as library orientation, sample gender, Mycoplasma contamination or RNA ribodepletion usage. Finally, we show that k-mer analysis can also be used to test known genomic and transcriptomic modifications (mutations, splice events, fusion genes, etc...) as well as for the discovery of new ones.

Acknowledgements and fundings

This work was supported by the Agence Nationale de la recherche for the projects "Computational Biology Institute" and "Transipedia" [grant numbers 18-CE45-0020-02, ANR-10-INBS-09] and the Canceropole Grand-Sud-Ouest "Trans-kmer" project [grant number 2017-EM24] and "SuRicare" [Région Occitanie]

References

[1] Kmerator. https://github.com/Transipedia/kmerator.

Poster 112

The South Green Genome Hubs

Gaëtan DROC[1], Xavier ARGOUT[1], Stéphanie BOCS[1], Aurore COMTE[2], Alexis DEREEPER[3], Jean-François DUFAYARD[1], Olivier GARSMEUR[1], Anestis GKANOGIANNIS[4], Valentin GUIGNON[4], Chantal HAMELIN[1], David LOPEZ[1], Nicolas OUBDA[1], Guillaume MARTIN[1], Sébastien RAVEL[5], Manuel RUIZ[1], Marilyne SUMMO[1], Coline THOMAS[1], Christine TRANCHANT-DUBREUIL[2], Mathieu ROUARD[4]

[1] UMR AGAP, Univ Montpellier, CIRAD, INRA, SupAgro, Montpellier, France
[2] UMR DIADE, IRD, Univ Montpellier, 34394 Montpellier, France
[3] UMR IPME, Univ Montpellier, CIRAD, IRD, Montpellier, France
[4] Bioversity International, Parc Scientifique Agropolis II, 34397, Montpellier, France
[5] UMR BGPI, Univ Montpellier, CIRAD, IRD, Montpellier, France

Corresponding Author: droc@cirad.fr

Mediterranean and tropical crops are sources of products of macroeconomic importance. The revolution in next-generation sequencing has led to the generation of reference genome sequences for multiple crops and their pathogens. Our participation in genome sequencing projects allowed us to develop crop-specific information systems, so called Genome Hubs (www.southgreen.fr/genomehubs), that enable centralized access to multi-omics data and analytical tools to facilitate translational and applied research. We opted for the CMS Drupal with GMOD components (i.e. Tripal, Chado, JBrowse) that are open source, modular and benefiting from a large community support. Additionally, we plugged in-house tools such as SNiPlay, Gigwa, GreenPhyl and DiffExDB. User-friendly web interfaces provide search functionalities (Blast, Gene Search, Tree search patterns, Primer Designer) and interfaces for phylogeny and microsynteny and gene families context viewers (*e.g.* Genomicus). Several Genome Hubs were released on Banana**[1]**, Cassava, Cocoa, Coffee**[2]**, Rice, Sugarcane some of them still being finalized (Palm, Grass, magnaporthe, Hevea, Myrtaceae). The Hubs are part to the South Green bioinformatics platform**[3]** and are supported by the French bioinformatics Institute (IFB) via the ELIXIR bio.tools registry for Service Delivery Plan. Future plans include integration of visualization tools dedicated to GWAS, mosaic genomes, and pangenomes.

Acknowledgements

We are grateful to Angélique D'Hont, Philippe Lashermes, Claire Lanaud, Luc Baudouin, Norbert Billotte, David Pot, Anne Dievart, Christophe Perin, Fabienne Morcillo that contributed to genome hubs through their respective projects and crop expertise. We thank Alexandra Louis for help with Genomicus and Guilhem Sempéré on GIGWA. This work was supported by the AGAP high-performance cluster as part of the South Green Bioinformatic platform.

References

1. Droc G, Lariviere D, Guignon V, Yahiaoui N, This D, Garsmeur O, et al. The Banana Genome Hub. Database. 2013;2013:bat035–bat035.
2. Dereeper A, Bocs S, Rouard M, Guignon V, Ravel S, Tranchant-Dubreuil C, et al. The coffee genome hub: a resource for coffee genomes. Nucleic Acids Research. 2015;43:D1028–35.
3. collaborators SG. The South Green portal: a comprehensive resource for tropical and Mediterranean crop genomics. Current Plant Biology. doi:10.1016/j.cpb.2016.12.002.

Poster 113

Proteomics guided detection of abnormal NGS genomes and proteomes

Karen CULOTTA[1], Virginie JOUFFRET[1], Jean ARMENGAUD[1], Olivier PIBLE[1]

[1] CEA-Marcoule, DRF/JOLIOT/DMTS/SPI/Li2D, Laboratoire Innovations technologiques pour la Détection et le Diagnostic, BP17171, 30207 Bagnols sur Cèze, France

Corresponding Author: olivier.pible@cea.fr

Background

Understanding the functioning of microbial consortia and more sophisticated ecosystems through the analysis of their structure and biological interactions is gaining momentum. Metaproteomics has recently emerged as a powerful analytical tool for studying the protein content of complex biological systems. High-throughput shotgun metaproteomic approaches on environmental or medical microbiomes are producing huge amounts of tandem mass spectrometry data. These can be interpreted either with a general protein sequence database comprising tens of thousands of sequenced genomes or with a more customized database such as those obtained after sequencing of the DNA or mRNA material extracted from the same sample. However, not all entries in a nucleotide or protein sequence database are of equal quality and this can critically impact metaproteomic data interpretation.

Results

First, either genome or transcriptome data interpretation due to inaccurate contig assembly and gene prediction may be erroneous. For its mitigation, the metaproteogenomic strategies could have an interesting perspective. Errors in sample handling and taxonomical characterization may also be problematic, as well as taxonomy consistency issues. Cross-contamination of genome sequences is also underestimated while frequent. As a consequence of these structural errors regarding protein sequences and additional problems due to homology-based functional annotation of proteins, specific efforts for better interpretation of metaproteomic data are required.

Conclusions

We propose the development of new bioinformatic pipelines devoted to detection and correction of errors and contaminations to improve the overall quality of sequence and taxonomy databases for metaproteomics.

Acknowledgements

We thank the Commissariat à l'Energie Atomique et aux Energies Alternatives, the NRBC-E transversal program, and the Languedoc-Roussillon region for their financial support, as well as Guylaine Miotello, Jean-Charles Gaillard, and Gérard Steinmetz for their technical support.

References

1. Olivier Pible, François Allain, Virginie Jouffret, Karen Culotta, Guylaine Miotello and Jean Armengaud. Estimating relative biomasses of organisms in microbiota using "phylopeptidomics". *Microbiome*, (8):30, 2020.
2. Olivier Pible and Jean Armengaud. Improving the quality of genome, protein sequence, and taxonomy databases: a prerequisite for microbiome meta-omics 2.0. *Proteomics*, (15):3418-23, 2015.
3. Olivier Pible, Erica M.Hartmann, Gilles Imbert and Jean Armengaud. The importance of recognizing and reporting sequence database contamination for proteomics. In *EuPA Open Proteomics*, volume 3 pages 246–249. ResearchGate, 2014.
4. Jean Armengaud. Microbiology and proteomics, getting the best of both worlds!. *Environmental Microbiology*, (15):12-23, 2013.

Poster 115

SysMics Integrative Research Cluster: Toward Systems Medicine based on Genomics

Audrey BIHOUEE[1,2], Stéphanie BONNAUD[1,2], Jérémie BOURDON[3], Laurent DAVID[4], Stéphane MINVIELLE[5], Michel NEUNLIST[6], Patricia PARNET[7], Pierre-Antoine GOURRAUD[8], Antoine MAGNAN[1], Françoise LE VACON[9] and Richard REDON[1]

[1]Université de Nantes, CHU Nantes, CNRS, INSERM, l'institut du thorax, F-44000, Nantes, France
[2]Université de Nantes, CHU Nantes, Inserm, CNRS, SFR Santé, F-44000, Nantes, France
[3]Université de Nantes, CNRS, LS2N, Nantes, F44322, France
[4]CRTI, INSERM, Université de Nantes, F-44000, Nantes, France
[5]CRCINA, INSERM, CNRS, Université d'Angers, Université de Nantes, Nantes, France
[6]INSERM, Univ Nantes, IMAD, TENS, F-44000, Nantes France
[7]Nantes Université, INRA, UMR1280, PhAN, F-44000 Nantes, France
[8]CHU de Nantes, INSERM, CIC 1413, 44000 Nantes, France
[9]Biofortis Mérieux NutriSciences, 44000 Nantes, France

Corresponding Author: audrey.bihouee@univ-nantes.fr

1. Context

The democratisation of next-generation sequencing technologies has transformed our capabilities to characterise complex living systems such as microbial communities and organisms, down to the single cell level. Personalized medicine is expected to benefit from combined large-scale information with regular monitoring of physiological states. Longitudinal integrative personal profiling has been proven effective to interpret healthy and diseased states by connecting genomic information with additional dynamic –omics activity. SysMics is an interdisciplinary cluster led by the University of Nantes regrouping *1) every joint laboratory in biomedical research* based in Nantes and already applying population-scale genomics at any level, *2) The LS2N laboratory*, which federates all teams working on computational sciences in Nantes, *3) Biofortis Mérieux NutriSciences*, which develops clinical investigations based on microbiome analysis.

2. Objectives

SysMics aims at federating the scientific community in Nantes toward a common objective: anticipate the emergence of systems medicine by co-developing 3 approaches in population-scale genomics: genotyping by sequencing, cell-by-cell profiling and microbiota meta-omics. SysMics has begun first to set up all necessary resources to implement/consolidate these 3 approaches on-site. Combining these approaches in the context of pilot projects in immunology, haematology and pathophysiology of cardiovascular, metabolic, respiratory, brain and gut disorders will result in integrative personal profiles, from the newborn to the adult, which will be instrumental in better understanding cascades of events leading to disease. Whenever possible, our translational research will be adapted and its outputs transferred to molecular diagnosis.

3. Ground resources

SysMics relies on the developing infrastructures based in Nantes (Genomics and Bioinformatics BiRD core facilities), which includes: *1) Clinical research organizations* and sample storage facilities in CHU Nantes; *2) The data warehouse & clinics* at the CHU Nantes to help exploiting large population-scale information arising from healthcare and clinical research; *3) Next-Generation Sequencers* with ancillary equipment for automated library preparation and sequencing; *4) Computing and storage resources*, directly connected to the sequencers, and distantly accessible to all SysMics members; *5) An open space dedicated to computer biology*, aiming to share their skills and experiences in population scale genomics.

SysMics *1) Animates and coordinates* all activities in population-scale genomics conducted in connection with the core facilities; *2) Promotes the international visibility* of SysMics members, by facilitating the development of collaborative network; *3) Supports on-site R&D*, to accelerate the emergence of novel methodological approaches and facilitate the development on new collaborative projects within SysMics.

Poster 116

Functional networks of co-expressed genes to explore iron homeostasis processes in the pathogenic yeast

Thomas DENECKER[1], Youfang ZHOU LI[2], Cécile FAIRHEAD[2], Karine BUDIN[1], Jean-Michel CAMADRO[3], Monique BOLOTIN-FUKUHARA[2], Adela ANGOULVANT[2] and Gaëlle LELANDAIS[1]

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

[2] Génétique Quantitative et Évolution Le Moulon, INRA, Univ. Paris-Sud / Université Paris Saclay, CNRS, AgroParisTech, Orsay, France.

[3] CNRS, Institut Jacques Monod (IJM), Univ. Paris Diderot, Paris, France.

Corresponding Author: gaelle.lelandais@universite-paris-saclay.fr

Infections due to *Candida* yeast species cause serious problems in aging populations and patients with compromised immunity. In this context, *Candida glabrata* has been reported as the second cause of candidiasis (1). Infections remain challenging to treat owing to delayed diagnosis, natural low susceptibility to azole antifungals and acquired resistance to echinocandins (2). During host infection, pathogens face abrupt physiological changes in their immediate environment. A major player is iron, as iron bioavailability is a key factor involved in the "nutritional immunity" host-defense mechanism (3). Remarkably, iron is a two-faced oligo-element for living organisms. On the one hand, iron is essential, as part of heme- and iron-sulfur cluster (ISC)-containing proteins involved in a variety of vital functions including oxygen transport, DNA synthesis, metabolic energy or cellular respiration and on the other hand, iron is toxic. Its excess triggers oxidative stress, lipid peroxidation and DNA damage that ultimately compromise cell viability and can promote programmed cell death. Iron homeostasis is therefore essential to allow pathogens to maintain a balance between iron utilization, storage, transport and uptake in the host environment.

The aim of the present work was to specifically study iron homeostasis in the pathogenic yeast *C. glabrata*. We performed transcriptomic experiments to monitor gene expression changes of *C. glabrata* to iron deficient and overload conditions, at 30°C and 37°C. The resulting dataset was analyzed to (*i*) clarify the potential effect of temperature on iron homeostasis, (*ii*) identify iron responsive genes, *i.e* genes significantly up- or down-regulated in at least one iron imbalanced situation and (*iii*) define a new set of genes, referred to as "iron homeostasis key genes" (iHKG). These genes are good candidates to be chief components of iron homeostasis. Our exploration of the datasets was facilitated by the inference of functional networks of co-expressed genes, which can be accessed through a web interface (https://thomasdenecker.github.io/iHKG/).

The philosophy of this work is to empower researchers by providing access to all transcriptomics data and by generating easily interpretable graphical outputs. This should facilitate deep exploration of genome-wide functional data in the pathogenic yeast *C. glabrata* to advance our global understanding of iron homeostasis.

References

1. Pfaller,M.A. and Diekema,D.J. (2007) Epidemiology of Invasive Candidiasis: a Persistent Public Health Problem. Clin. Microbiol. Rev., 20, 133–163.Barbara Gastel and Robert A Day. *How to write and publish a scientific paper*. ABC-CLIO, 2016.
2. Pfaller,M.A., Castanheira,M., Lockhart,S.R., Ahlquist,A.M., Messer,S.A. and Jones,R.N. (2012) Frequency of Decreased Susceptibility and Resistance to Echinocandins among Fluconazole-Resistant Bloodstream Isolates of Candida glabrata. J. Clin. Microbiol., 50, 1199–1203.
3. Sutak,R., Lesuisse,E., Tachezy,J. and Richardson,D.R. (2008) Crusade for iron: iron uptake in unicellular eukaryotes and its significance for virulence. Trends Microbiol., 16, 261–268.

Poster 117

Reconstruction and comparison of brown algal metabolic networks for identification of specific genes

Alexandre LAMBARD[1], Jeanne GOT[2], Anne SIEGEL[2], Gabriel MARKOV[3] and Erwan CORRE[1]

[1] CNRS - Sorbonne Université – Plateforme ABIMS - Station Biologique de Roscoff, Place Georges Teissier, 29680, Roscoff, France

[2] Université de Rennes 1, Institute for Research in IT and Random Systems (IRISA), Equipe Dyliss, 35052, Rennes, France

[3] CNRS - Sorbonne Université - Integrative Biology of Marine Models (LBI2M/UMR8227) - Station Biologique de Roscoff, Place Georges Teissier, 29680, Roscoff, France

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

Methods for reconstructing metabolic networks from genomic data are under development, particularly for marine organisms. Brown algae are photosynthetic organisms belonging to Stramenopiles. They are closely related to unicellular photosynthetic microalgae such as diatoms or eustigmatophytes, with which they form the ochrophyte clade. The majority of brown algae live in the marine environment, mostly in the intertidal zone and shallow depths, where they can form true underwater forests. Brown algae are being studied from an applied perspective, to understand how to use genetic diversity for varietal improvement of species grown in aquaculture. From this point of view, the study of metabolic pathways is of particular interest, as the enzymes encoded by genes make it possible to make a direct link between the genotype and a phenotype that can be described by metabolic profiling. As part of an international consortium, the Phaeoexplorer project [1] has sequenced forty brown algae whose genomes are currently being assembled or annotated.

In this project we start to explore the diversity of these genomes using a genome-scale metabolic network reconstruction approach [2]. The aim is to see how a pipeline for reconstruction and comparison of metabolic networks developed by the Dyliss team at IRISA (Rennes), called AuCoMe [3,4], behaves to compare 14 ochrophyte genomes.

In terms of biology, the challenge will be to reconstruct the evolutionary history of gene losses and variations in the structure of metabolic pathways, particularly in relation to the acquisition of an endophytic lifestyle in certain filamentous brown algae [5]. The comparison will focus both on the overall architecture of the metabolic network at the genome level and more specifically on specific metabolic pathways (sterols and carotenoids) for which additional biological knowledge is available.

Acknowledgements

This research received funding from the French Government via the National Research Agency investment expenditure program IDEALG (ANR–10–BTBR–04).

References

1. http://application.sb-roscoff.fr/project/phaeoexplorer/index.html
2. Gu, C.; Kim, G. B.; Kim, W. J.; Kim, H. U. & Lee, S. Y. Current status and applications of genome-scale metabolic models. Genome biology, 2019, 20, 121
3. Aite, M.; Chevallier, M.; Frioux, C.; Trottier, C.; Got, J.; Cortes, M. P.; Mendoza, S. N.; Carrier, G.; Dameron, O.; Guillaudeux, N.; Latorre, M.; Loira, N.; Markov, G. V.; Maass, A. & Siegel, A. Traceability, reproducibility and wiki-exploration for "à-la-carte" reconstructions of genome-scale metabolic models. PLOS Computational Biology, 2018, 14, 1-25
4. https://github.com/AuReMe/aucome
5. Bernard, M. S.; Strittmatter, M.; Murua, P.; Heesch, S.; Cho, G. Y.; Leblanc, C. & Peters, A. F. Diversity, biogeography and host specificity of kelp endophytes with a focus on the genera Laminarionema and Laminariocolax (Ectocarpales, Phaeophyceae). European Journal of Phycology, 2018, 1-13

Poster 118

Examples of bioanalysis activities on the ABiMS (Analysis and Bioinformatic for Marine Science) platform

Nahéma HECHT[1], Thomas ENJALBERT[1], Simon DITTAMI[1], Olivier GODFROY[1], Agnieszka LIPINSKA[1], Francois THOMAS[1] and Erwan CORRE[2]

[1] CNRS - Sorbonne Université – UMR8227 - Station Biologique de Roscoff, Place Georges Teissier, 29680, Roscoff, France

[2] CNRS - Sorbonne Université - Plateforme ABIMS - Station Biologique de Roscoff, Place Georges Teissier, 29680, Roscoff, France

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

The mission of the ABiMS platform is to assist researchers of the marine community and, more broadly, of the life sciences, in the bioinformatic analysis of their data as well as in the development of software and databases. It is one of the national platforms of the French Institute of Bioinformatics (IFB). It is also associated to the EMBRC (European Marine Biology Ressource Center) infrastructure, is part of the IBiSA network via the regional BioGenouest project and is ISO 9001: 2015 certified. Through its numerous interactions with research units, ABiMS is involved in several projects, with national and European impacts involving bioanalysis activities, software, and e-Infrastructures development. Through 2 examples of collaborative projects conducted by Bachelor students we wish to illustrate the bioanalysis activity conducted by the ABiMS platform in the field of marine data.

Algavor project (collaboration with F. Thomas – UMR8227): 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. As part of this project, we are exploring the presence of some of these pioneer bacteria of the genus *Zobellia* in marine, coastal or alga-associated metagenomes. We use available genomes of pure *Zobellia* strains to recruit reads and evaluate the distribution, abundance, and activity of *Zobellia* spp. in marine environments. Further, we attempt to build metagenome-assembled genomes (MAG) from recruited reads to gain insights into their biodiversity and catabolic functions.

HIGH-quality geNomE aSSembly of the *Ectocarpus subulatus* genome (collaboration with S. Dittami, O. Godfroy, A. Lipinska – UMR8227) *Ectocarpus subulatus* is a highly stress-tolerant species of brown algae frequently found in environments with high temperature, low salinity, or high variability in abiotic factors, e.g. driftwood. Currently only a very fragmented assembly of the *E. subulatus* genome is available, limiting comparative genomic analyses with other brown algae (1). The aim of this project is to generate a new, high-quality assembly and annotation of the *E. subulatus* genome combing Illumina sequencing data with 6Gb of newly generated long-read sequences (Oxford Nanopore sequencing, already accomplished) and a HiC data (to be generated). We test alternative base callers for the Nanopore data, the generation of draft assemblies with different assemblers, cleaning (removal of prokaryotic contaminants) and comparison of assemblies, structural annotation (repeated elements, genes, etc.), and functional annotations (predicted gene functions, possibly including the generation of a metabolic network). This will form the bases for updated structural and functional comparisons of *E. subulatus* and the *E. siliculosus*, which will also be initiated during this project.

References

1. Dittami et al. 2020. The genome of *Ectocarpus subulatus* - a highly stress-tolerant brown alga. Marine Genomics https://doi.org/10.1016/j.margen.2020.100740

Poster 120

Semi-automated noise removal tool for single-cell mass cytometry data

Maria-Fernanda SENOSAIN[1] and Pierre P. MASSION[2]

[1] Cancer Biology Graduate Program, Vanderbilt University, Nashville, TN 37232, USA
[2] School of Medicine, Vanderbilt University Medical Center, Nashville, TN 37232, USA

Corresponding Author: mariafernanda.senosain@vanderbilt.edu

Mass cytometry is a single cell proteomic technique that allows for measurement of the expression of ~40 proteins. Although signal overlapping is minimal due to the use of metal-conjugated antibodies, other sources of noise are present. Tissue samples such as solid tumors that had gone through dissociation steps and are cryopreserved present large amounts of debris and dead cells. Traditional gating for noise removal can be subjective and tedious when working with a large number of samples. Here we present denoisingCTF, a noise removal R package for CyTOF data. This software has two main modules, one to remove noise using our current trained models or user-customized models and a second module that allows the user to train its own models for noise removal. To train and test the classification models we used our previously generated CyTOF datasets (*unpublished*): human lung adenocarcinoma (n = 80) + beads, 1:1 A549 & Ramos cell lines (n = 20) + beads, Beads only (n = 3), 1:1 A549 & Ramos cell lines (n = 3) cells only. The noise removal function works as follows: A first step removes events with zero expression of mandatory markers (e.g. His H3 for nucleated cells or any intact-cell-marker of preference) as well as events not expressing any of the cell type specific markers. The second step removes normalization beads using a classification model which was trained on a dataset for which beads were detected and labeled in an unsupervised manner (GMM or k-means). We used tumor samples to build both training and test sets, and we validated the model using cell lines and beads run separately, which then were labeled and merged computationally. To ensure the quality of the training data, samples in which bead detection failed (CV<0.05) were not considered. A final step removes debris using a classification model trained on a labeled (noise= 0,1) dataset which was obtained by manually gating on the Gaussian Discrimination parameters and our marker for intact cells (Histone H3) per Fluidigm recommendations for noise removal. For both steps 2 and 3 we trained a Random Forest and a XGBoost classifier, both yielding similar results (accuracy > 0.9, sensitivity > 0.9, specificity > 0.9). In summary, this approach can provide an automated unbiased detection and removal of noise compared to the use of theoretical cutoff values or user-dependent gating. Our R package can be found here https://msenosain.github.io/denoisingCTF/index.html. This work was supported by CA196415.

Poster 123

ToulligQC 2: a nicer and faster quality control software for Oxford Nanopore sequencing data

Karine DIAS[1], Corinne BLUGEON[1], Charlotte BERTHELIER[1], Médine BENCHOUAIA[1], and Laurent JOURDREN[1]

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

Corresponding author: karine.dias@bio.ens.psl.eu

The sequencing devices developed by Oxford Nanopore Technologies (ONT) produce long DNA sequence (up to 200 kb) and full-length RNA. Sequencing and primary data acquisition are driven by MinKNOW, an ONT tool. MinKNOW produces Fast5 files to store raw data. Basecalling can be performed during the acquisition step or after it is over by Guppy, the official ONT basecaller. The output files are stored in FASTQ or Fast5 format.

The metrics and scales that were provided by MinKNOW when we launched RNA-Seq were not appropriated for gene expression applications (no barcode handling for example and unsuitable scales for RNA). It was necessary to develop a dedicated QC tool, flexible enough to handle both RNA and DNA sequencing, hence ToulligQC inspired by FastQC [1].

Used in production since 2017, ToulligQC allows researchers to quickly estimate the quality and homogeneity of their samples for further expression analysis RNA-Seq or DNA-Seq. Easy to use, this tool aims to give a detailed graphical output about the quality of Nanopore runs and exploratory data analysis.

This poster introduces you to ToulligQC 2, a new major version of our QC software. Faster than the previous version, ToulligQC 2 will provide an improved HTML report with modernised and interactive plots made with Plotly, Seaborn and Matplotlib libraries. Statistics about pass and fail reads, but also barcoding charts are also improved. In addition, new plots about read quality and read length over run time are added.

Because ONT protocols and tools are constantly evolving, ToulligQC 2 supports the latest version of Guppy and the latest sequencing protocols. It can be used with all the Oxford Nanopore sequencing devices and remains compatible with both 1D and $1D^2$ chemistries. It takes as input the sequencing summary file generated by the basecaller and the sequencing telemetry file too if available.

ToulligQC 2 is an *open source* software which can be freely downloaded on *Github* [2], as a *Docker image* (genomiquepariscentre/toulligqc), and as a *PyPy package* [3].

References

[1] https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
[2] https://github.com/GenomicParisCentre/toulligQC
[3] https://pypi.org/project/toulligqc/

Poster 125

DESCRIPTION's Axiom integration in IMGT-ONTOLOGY : Construction of a Knowledge Graph in Immunogenetics

Gaoussou SANOU[1,2], Véronique GIUDICELLI[1], Konstantin TODOROV[2], Sofia KOSSIDA[1] and Patrice DUROUX[1]

[1] IMGT®, Institut of Human Genetics, CNRS, University of Montpellier, Montpellier, France
[2] FADO, LIRMM, CNRS, University of Montpellier, Montpellier, France

---

Nowadays ontologies and knowledge graphs play an important role in the context of Open Data and Big Data applications. In life sciences, ontologies and knowledge graphs promise to provide the key to federated data sharing and reuse, however, the rich and diverse vocabularies and definitions in the field make it difficult to formalize the scientific terms and propose a unified structure.

Developed since 1989, the international ImMunoGeneTics information system® (IMGT®) regroups today several rich relational databases such as sequence, genome, structure and monoclonal antibody databases, software tools and multiple unstructured resources, such as HTML pages or pdf documents, accessible to the public through the IMGT portal (http://www.imgt.org). IMGT's strength is the provision of a standard vocabulary: the IMGT-ONTOLOGY [1,2,4] refined over the years, for data and tools in the system. A first publication in 1999 laid the foundation of IMGT-ONTOLOGY and a first implementation in RDF+OWL language became available in 2010 through the BioPortal (https://bioportal.bioontology.org/). However, this formalization takes into account only a small part of the data in the system. Therefore, we cannot currently describe a sequence with the IMGT-ONTOLOGY, due to the missing of description concepts in IMGT-ONTOLOGY. In fact, description's concepts allows us to describe sequences and their structure with a set of labels and relations.

Our work aims to propose a generalized description model which will cover all the system's data by integrating the IMGT DESCRIPTION's axiom. This will allow us to structure all the data in the form of a knowledge graph [3]. Subsequently, the goal is to apply machine learning (pattern mining, clustering, prediction) methods in order to discover new knowledge from the structured data.

Keywords: immunogenetics, immunoinformatics, ontologies, knowledge graphs, machine learning

REFERENCES

1. Duroux, P., Kaas, Q., Brochet, X., Lane, J., Ginestoux, C., Lefranc, M.-P. and Giudicelli, V. "IMGT-Kaleidoscope: the Formal IMGT-ONTOLOGY paradigm", Biochimie 90 (2008): 570-583. Epub 2007 Sep11. PMID: 17949886
2. Lefranc, M.-P. "From IMGT-ONTOLOGY DESCRIPTION Axiom to IMGT Standardized Labels: For Immunoglobulin (IG) and T Cell Receptor (TR) Sequences and Structures" Cold Spring Harbor Protocols 2011 (6) (2011). pii: pdb.ip83. doi: 10.1101/pdb.ip83. PMID: 21632791
3. Achichi, M., Ben Ellefin, M., Bellahsene, Z., Todorov, K. "Linking and disambiguating entities across heterogeneous RDF graphs." Journal of Web Semantics 55 (2019): 108-121.
4. Giudicelli V, Lefranc M.-P. "IMGT-ONTOLOGY 2012" Frontiers in genetics 3 (2012):79. doi: 10.3389/fgene.2012.00079. Epub 2012 May 23. PMID: 22654892

Poster 126

A Shiny application for RNA-seq data analysis and interpretation

Justine GUÉGAN[1], Beáta GYÖRGY[1], Mathilde BERTRAND[1], Thomas GAREAU[1] and Ivan MOSZER[1]

[1] iCONICS Core Facility, Institut du Cerveau, Inserm U 1127, CNRS UMR 7225, Sorbonne Université, F-75013, Paris, France

Corresponding Author: justine.guegan@icm-institute.org

The iCONICS core facility is part of the Paris Brain Institute (ICM), an organization dedicated to basic and clinical neuroscience research. Within the platform, a specialized team assists scientific and clinical teams (study design, data processing and analysis), and develops graphical tools to help in the interpretation of omics data. In particular, RNA-seq data analysis requires the use of several statistical methods and algorithms, which are often only accessible to users mastering computer tools such as R. A wide variety of software exists for normalization, differential analysis, or functional analysis of transcriptomic data. In addition to programming skills, most of them require an expert point of view to correctly apply the underlying methods. To address those limits, we propose an interactive graphical interface, which provides a guided, easy to use and comprehensive set of tools for RNA-seq data analysis. Based on the Shiny framework, this application allows end-users to easily manipulate and explore their gene expression experiment results.

The key steps of our Shiny application are (i) count data matrix import and normalization, (ii) primary exploratory analysis, (iii) differential gene expression (DE), (iv) functional enrichment analysis, (v) result reporting. The user starts by importing his count data matrix and his sample annotations. He can visualize the effect of normalization, the profile of the genes of interest (barplot or boxplot), and run a Principal Component Analysis (PCA) to check for batch effects or identify outlier samples; the latter can then be removed and the user can run the PCA again. At each step of the analysis, all the plots and tables that are generated can be downloaded. Differential gene expression analysis can be performed in an autonomous and flexible way. The user has the capability to define the groups of samples to be compared, choose the parameters of the analysis, exclude outliers if necessary and filter the results. The analysis returns a volcano plot, an MA plot, a heatmap and the result table. The results of the differential gene expression analysis can be further explored via functional enrichment analysis. Two types of methods are implemented in the application: over-representation using Fisher tests, and Gene Set Enrichment Analysis (GSEA) [1]. The first one is based on Reactome pathways [2] and Gene Ontology [3] while the second one is based on MSigDB collections [1] and WikiPathways [4]. Finally, the user can create a personalized HTML report of his analysis. He can choose which step to include and add comments using the R Markdown syntax.

This Shiny application has been widely used in the framework of RNA-seq project analyses run by iCONICS. Users particularly appreciate its ease of use and responsiveness. Thus, it constitutes a true added-value in the usability and understanding of their biological results. New developments are ongoing to offer the same kind of functionality for single-cell RNA-seq data exploration. The underlying code is now being adapted so that data not generated by our core facility can also be handled. The resulting software will be made available through standard repositories.

Acknowledgements

We thank Romain DAVEAU for his seminal work on the Shiny application, and Derya SEBUKHAN for her help in optimizing the source code. This work was supported by the IHU-A-ICM program ANR-10-IAIHU-06.

References

1. Aravind Subramanian, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545-50, 2005.
2. Bijay Jassal, *et al*. The Reactome Pathway Knowledgebase. *Nucleic Acids Res*, 48(D1):D498-D503, 2020.
3. The Gene Ontology Consortium. The Gene Ontology Resource: 20 Years and Still GOing Strong. *Nucleic Acids Res*, 47(D1):D330-D338, 2019.
4. Denise N Slenter, *et al*. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res*, 46(D1):D661-D667, 2018.

Poster 127

From reads to chromosomes: What is the optimal path?

Clément Birbes[1], Andreea Dréau[1], Camille Eche[2], Carole Iampietro[2], Cécile Grohs[3], Didier Boichard[3], Cécile Donnadieu[2], Christine Gaspin[1], Denis Milan[2,4], Christophe Klopp[1] and Matthias Zytnicki[1]

[1] INRAE, UR875 MIAT, F-31326 Castanet-Tolosan, France
[2] INRAE, US1426 GeT-PlaGe, F-31326 Castanet-Tolosan, France
[3] INRAE, GABI, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France
[4] INRAE, GenPhySE, ENVT, F-31326 Castanet-Tolosan, France

Corresponding author: andreea.dreau@inrae.fr

Sequencing advances in the last years led to a significant increase in the number and quality of published de novo genomes, each one of them being a result of a mix of technologies and assembly strategies. While the variety of existing read types helped with different challenges in the genome assembly problem, it also raised questions regarding the optimal combination of technologies. The purpose of our study is to identify the intake of each type of reads and their coverage in order to determine an optimal approach depending on the sequencing cost or the expected assembly quality. In our project we combine many cutting edge technologies as Oxford Nanopore, Pacific Bioscience (HiFi and CLR), 10x Chromium, Hi-C and Bionano optical mapping in a six step assembly pipeline: contig assembly, polishing, splitting, scaffolding, gap filling and final polishing. The tests were conducted on two trios of *Bos Taurus* for which we constructed chromosome level assemblies.

The contig assembly step is the most consuming in terms of CPU, memory and running time. It is also very sensitive to the length, quantity and quality of used reads. The best ratio between required computational resources and assembly quality was obtained with wtdbg2 [1]. The assembler needs a minimum coverage of 45x nanopore reads to obtain chromosome arm level contigs. Also, filtering reads shorter than 10kb can improve contiguity and reduce the running time. The role of polishing is to remove the contig sequence errors coming from the long reads, but it can also introduce new errors when a wrong combination of tools or number of iteration is used. We obtained the best BUSCO scores with one run of Racon [2] using long reads followed by one run of Pilon [3] with short reads. Then the connection errors introduced by the assembly process are corrected during the splitting step either by removing the regions with very low coverage or by splitting the contigs in case of coverage alterations. We studied the different types of errors that can be identified in this step depending on the algorithm and combination of types of reads.

Finally, contigs are regrouped into scaffolds and chromosomes using Hi-C reads. For this, we used 3d-dna [4] to build scaffolds, and then juicebox [5] to connect them into chromosomes. We tested different Hi-C protocols and identify the minimum coverage needed to obtain the chromosomes.

Acknowledgements

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

References

[1] Jue Ruan and Heng Li. Fast and accurate long-read assembly with wtdbg2. *Nature Methods*, 17(2):155–158, 2020.

[2] Robert Vaser, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, 27(5):737–746, 2017.

[3] Bruce J Walker, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K Young, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, 9(11), 2014.

[4] Olga Dudchenko, Sanjit S Batra, Arina D Omer, Sarah K Nyquist, Marie Hoeger, Neva C Durand, Muhammad S Shamim, Ido Machol, Eric S Lander, Aviva Presser Aiden, et al. De novo assembly of the aedes aegypti genome using hi-c yields chromosome-length scaffolds. *Science*, 356(6333):92–95, 2017.

[5] Neva C Durand, James T Robinson, Muhammad S Shamim, Ido Machol, Jill P Mesirov, Eric S Lander, and Erez Lieberman Aiden. Juicebox provides a visualization system for hi-c contact maps with unlimited zoom. *Cell Systems*, 3(1):99–101, 2016.

Poster 129

Mobilome conservation among Drosophila paralogs

Gabrielle Pozo[1], Carène Rizzon[2] and Emmanuelle Lerat[1]

[1] Laboratory Biométrie et Biologie Evolutive (LBBE), CNRS, Université de Lyon, Université Lyon 1, F-69622, Villeurbanne, France

[2] Mathématiques et Modélisation d'Evry (LaMME), Université d'Evry Val d'Essonne, UMR CNRS 8071, ENSIIE, USC INRA, 23 bvd de France, 91037, Evry Cedex, France

Corresponding Author: emmanuelle.lerat@univ-lyon1.fr

In a genome, duplicated genes (paralogs) can provide the opportunity for new functions to be developed in a species [1]. They can arise from different mechanisms such as retrotransposition, segmental duplications via homologous or non-homologous recombination, and chromosome or whole genome duplications [2]. After duplication, paralogs can encounter various fates. A large number are lost trough the process of pseudogenization due to the accumulation of deleterious mutations [2]. However, a significant proportion is maintained for which three possibilities of evolution exists [2]. Either one copy evolves to acquire a new function (neofunctionaliation) whereas the other copy conserves the ancestral function, or the duplicated genes can complement each other to provide the original function (sub-functionalization), or the two copies can keep the ancestral function (redundancy).

Eukaryotic genomes also contain numerous types of sequences among which protein-coding genes are often a minority. Among the non-coding part, transposable elements (TEs) may represent a substantial fraction. In that respect, the genome of *Drosophila melanogaster*, which contains 13 % of protein-coding genes is also composed of more than 20% of TEs [3,4]. TEs are middle-repeated DNA sequences that have the ability to move from one position to another along chromosomes. They typically encode for all the proteins necessary for their movement and possess internal regulatory regions, allowing their independent expression. Different categories of TEs have been identified, among which the LTR (Long Terminal Repeat)-retrotransposons, the non-LTR retrotransposons grouping the LINE and the SINE elements (standing for Long and Short Interspersed Nuclear Elements respectively) and the DNA transposons. TEs are not randomly distributed in the drosophila genome since they are mainly found outside genes and preferentially in regions with low recombination rates, suggesting that selection is acting against their insertions [5,6]. Due to their mobility and repeatedness, TEs can promote various types of mutations, which are expected to be mostly harmful for the host genome [7]. In *D. melanogaster*, it has been shown that the maintenance of TEs in the genome is the result of both host repression and purifying selection against deleterious insertions [8]. However, it is still possible to find some adaptive insertions when studying natural populations [9].

In this work, we will identify all gene families present in the *D. melanogaster* genome and we will test whether the presence of TEs around the genes is associated with particular features of the gene families like their size, the time since the duplication events, and the gene function. Moreover, we will determine if variations in TE neighborhood around members of a same gene family impact the gene epigenetic landscape as well as their expression level in different tissues. The question we want to address here is whether TE may have played a role in the evolution of duplicated genes in the Drosophila genome.

References

[1] Susumu Ohno. *Evolution by gene duplication*. Springer. 1970.
[2] Matthew Hurles. Gene duplication: the genomic trade in spare parts. *PLoS Biol* 2:E206, 2004.
[3] Mark D Adams *et al*. The genome sequence of *Drosophila melanogaster*. *Science* 287:2185-2195. 2000.
[4] Andrew P Dowsett and Michael W Young. Differing levels of dispersed repetitive DNA among closely related species of Drosophila. *Proc Natl Acad Sci USA* 79:4570–4574. 1982.
[5] Joshua S Kaminker et al. The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. *Genome Biol* 3:RESEARCH0084. 2002.
[6] Carène Rizzon et al. Recombination rate and the distribution of transposable elements in the *Drosophila melanogaster* genome. *Genome Res 12:*400–407. 2002.
[7] Margaret G Kidwell and Damon R Lisch. Transposable elements and host genome evolution. *Trends Ecol Evol* 15:95-99. 2000.
[8] Grace YC Lee, Charles H Langley. Transposable elements in natural populations of *Drosophila melanogaster*. *Philos Trans R Soc Lond B Biol Sci* 365:1219–1228. 2010.
[9] Josefa González *et al*. Genome-wide patterns of adaptation to temperate environments associated with transposable elements in Drosophila. *PLoS Genet* 6:33–35. 2010.

Poster 130

Convergent stepwise recombination suppression in mating-type chromosomes of anther-smut fungi

Marine Duhamel[1,2], Ricardo C.. Rodriguez de la Vega[1], Fantin Carpentier[1], Tatiana Giraud[1]

1 Ecologie Systématique Evolution, Université Paris-Sud, AgroParisTech, CNRS, Université Paris-Saclay, 362 Rue du Doyen André Guinier, 91400, Orsay, France
2 Ruhr-University Bochum, Evolution and Biodiversity of Plants Geobotany, Universitätsstraße 150, 44780, Bochum, Germany

Corresponding Author: marine.c.duhamel@gmail.com

Recombination suppression is a common feature of genomic regions involved in mating compatibility, such as sex chromosomes in animals and plants, self-incompatibility loci in plants and mating-type chromosomes in fungi [1]. While first occurring between several genes controlling gamete compatibility to maintain beneficial allelic combinations, the non-recombining regions (NRRs) often gradually extend beyond these genes, forming evolutionary strata of different ages. The resulting evolutionary strata can be visualized by plotting the synonymous divergence between the alleles along the mating-type chromosomes using the ancestral gene order, i.e., the gene order of the recombining chromosome for sex chromosomes. Although evolutionary strata on sex chromosomes have been described for long in many species, the evolutionary causes leading to stepwise recombination cessation beyond genes controlling gamete compatibility are still under debate. The dominant hypothesis is the progressive linkage of sexually antagonistic genes (i.e., genes with alleles beneficial for one sex and deleterious for the other).

In basidiomycete fungi, mating compatibility is controlled by two mating-type loci, the pheromone receptor (PR) loci and the homeodomain (HD) loci, mating only occurring between gametes carrying alternative alleles at both mating-type loci. In these fungi reproducing mainly between the products of a single meiosis, the linkage of the two mating-type loci increases odds of gamete compatibility. Mating-type loci linkage and progressive extension beyond mating-type loci occurred five times independently in anther-smut *Microbotryum* genus [2,3], despite the absence of sexual antagonism in fungi. Using comparative genomics in the *Microbotryum* genus, we identified additional independent events of mating-type gene linkage in different *Microbotryum* species. Because both mating-type chromosomes stopped recombining, they both accumulated different chromosomal rearrangements, rendering impossible to observe the ancestral gene order directly for assessing evolutionary strata as performed in sex chromosomes. Therefore, the ancestral gene order was inferred from a closely related species whose mating-type loci are still unlinked.

Further perspectives include the exploration of potential genomic mechanisms involved in such recombination suppression extension in the absence of sexual antagonism, e.g., sheltering against deleterious mutations [4] or the local spread of transposable elements [5].

References

1. Uyenoyama, M. K., 2005 Evolution under tight linkage to mating type. New Phytologist 165: 63-70.
2. Branco, S, Carpentier, F, Rodriguez de la Vega, RC, Badouin, H, Snirc, A, Le Prieur, S, Coelho, MA, de Vienne, DM, Hartmann, FE, Begerow, D, Hood, ME, Giraud, T (2018). Multiple convergent supergene evolution events in mating-type chromosomes. Nat Commun, 9, 1:2000.
3. Carpentier, F, Rodriguez de la Vega, RC, Branco, S, Snirc, A, Coelho, MA, Hood, ME, Giraud, T (2019). Convergent recombination cessation between mating-type genes and centromeres in selfing anther-smut fungi. Genome Res., 29, 6:944-953.
4. Bachtrog, D (2005). Sex chromosome evolution: molecular aspects of Y-chromosome degeneration in Drosophila. Genome Res., 15, 10:1393-401.
5. Kent, T. V., J. Uzunovic and S. I. Wright, 2017 Coevolution between transposable elements and recombination. Philosophical Transactions of the Royal Society B-Biological Sciences 372.

Poster 131

Benchmarking of Minion long read assembly tools to produce a high-quality genome of *Pseudogynoascus destructans*, a pathogen fungus with a high proportion of repetitive elements.

Matthieu Fournel[1,3], Sèverine Bérard[1], Nicola Fischer[1,2], Anna-Sophie Fiston-Lavier[1], Marie-Ka Tilak[1], Sébastien J. Puechmaille[1]

[1] ISEM, Université de Montpellier, CNRS, IRD, EPHE, Montpellier, France
[2] Zoological Institute and Museum, Greifswald University, Greifswald,Germany
[3] Master Science and Numerics for Health - Speciality: Bioinformatics, Knowledge, Data

Abstract:
White nose syndrome is one of the most devastating wildlife diseases, causing massive mass mortality in many bat species throughout North America. Here we present a new genome assembly of the fungal pathogen *Pseudogymnoascus destructans* to uncover the genetic basis of pathogenicity. Comparing *P. destructans* to other non-pathogenic members of the genus *Pseudogymnoascus* will provide us with important knowledge on the genetic basis and mechanisms of pathogenicity. The actual challenge in assembling this genome is the treatment of repetitive elements that represent more than a third of the 35 - 40Mb genome [1]. Repetitive elements are thought to be important drivers of diversity and innovation in fungal pathogens. While the long Minion reads (Oxford Nanopore Technologies) clearly provides some advantages over short reads, specifically taking into consideration repetitive elements is still required. Here we present a benchmark of Shasta [2], Flye [3] and Canu [4]. Shasta and Flye are two recent tools designed to quickly process Nanopore reads and designed to deal with repetitive elements and Canu is the reference tool for long reads assembly. Preliminary results suggest that a better assembly is obtained with Flye, the program with the most elaborate method to handle repetitive elements. We obtain a total genome size of 38.5Mb, 7.5% larger than the previous reference genome and an N50 around 2.6MB, more than twice as large as the N50 of the previous assembly.

Références:

1. *Kevin P Drees, Jonathan M Palmer, Robert Sebra, Jeffrey M Lorch, Cynthia Chen, Cheng-Cang Wu, Jin Woo Bok, Nancy P Keller, David S Blehert, Christina A Cuomo, et al. Use of multiple sequencing technologies to produce a high-quality genome of the fungus pseudogymnoascus destructans, the causative agent of bat white-nose syndrome. Genome Announc., 4(3) :e00445–16, 2016.*

2. *Kishwar Shafin, Trevor Pesout, Ryan Lorig-Roach, Marina Haukness, Hugh E Olsen, Colleen Bosworth, Joel Armstrong, Kristof Tigyi, Nicholas Maurer, Sergey Koren, et al. Efficient de novo assembly of eleven human genomes using promethion sequencing and a novel nanopore toolkit. BioRxiv, page 715722, 2019.*

3. *Mikhail Kolmogorov, Jeffrey Yuan, Yu Lin, and Pavel A Pevzner. Assembly of long, error-prone reads using repeat graphs. Nature biotechnology, 37(5) :540–546, 2019.*

4. *Sergey Koren, Brian P Walenz, Konstantin Berlin, Jason R Miller, Nicholas H Berg- man, 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.*

Poster 132

Artificial intelligence approach for predicting variations in low-coverage NGS sequences

Nguyet DANG[1,2] 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

Corresponding author: thi-minh-nguyet.dang@ird.fr

In Vietnam, rice is not only the main food staple but also the main agricultural export. Recently, it has been showed that rice production in Vietnam is vulnerable to climate change, which raises the necessity for crop yield and quality improvement. In this regard, a Vietnamese rice population of 250 individuals was developed and phenotyped on different aspects. To access the genomic content of this collection at a reasonable expense, each individual was sequenced at low-coverage by Illumina technology. Those data were used to generate SNP data, using inference approaches.

Due to reference bias, the current short-read analysis tools do not give access to large genomic variations in efficient ways, especially for low-coverage data. In this context, the pangenome concept, a combination of a common core-genome among all individuals and dispensable compartments, has been proposed [1,2]. Since the pangenome structure is complex, graph-based methods [3,4] are generally used for documentation and visualization of all the genomic information. We propose to explore new possibilities offered by artificial intelligence, particularly deep learning, to improve genomic variation detection on low-coverage data based on genome graph.

We will use 100 artificial genomes implementing known variants. Then, we will simulate short and long-read sequence, and use them to construct genome graphs. By applying artificial intelligence algorithms on those graph, we expect to identify the variations introduced into the dataset. Practically, once the test data generated, we will apply first a supervised approach, using a subset of 20 reference test genomes, and let the system learn and apply on the remaining 80 simulated one. Then we will apply an non-supervised model (that will identify by itself the underlying models), and finally a reinforcement learning, where the user will correct the system at each iteration

Once the models are validated on the artificial data, we will test the method on an already analyzed population of high-depth data, from which we will generate low coverage data, to validate the model. Finally, we will apply the validated models on the Vietnamese rice collection, in order to identify the specific genes and sequences from these rices explaining their specific adaptation.

References

[1] Hervé Tettelin, Vega Masignani, Michael J. Cieslewicz, Claudio Donati, Duccio Medini, Naomi L. Ward, Samuel V. Angiuoli, Jonathan Crabtree, Amanda L. Jones, A.S Durkin, and *et al.* 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.

[2] Christine Tranchant-Dubreuil, Mathieu Rouard, and Francois Sabot. Plant Pangenome: Impacts On Phenotypes And Evolution. *Annual Plant Reviews*, May 2019.

[3] Adam M. Novak, Glenn Hickey, Erik Garrison, Sean Blum, Abram Connelly, Alexander Dilthey, Jordan Eizenga, M. A. Saleh Elmohamed, Sally Guthrie, André Kahles, Stephen Keenan, Jerome Kelleher, Deniz Kural, Heng Li, Michael F. Lin, Karen Miga, Nancy Ouyang, Goran Rakocevic, Maciek Smuga-Otto, Alexander Wait Zaranek, Richard Durbin, Gil McVean, David Haussler, and Benedict Paten. Genome graphs. *bioRxiv*, 2017.

[4] Erik Garrison, Jouni Sirén, Adam M Novak, Glenn Hickey, Jordan M Eizenga, Eric T Dawson, William Jones, Shilpa Garg, Charles Markello, Michael F Lin, Benedict Paten, and Richard Durbin. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, 36(9):875–879, 2018.

Poster 133

NG-AR2T, an epidemiological analysis pipeline for *Neisseria gonorrhoeae* sequenced isolates

Manel MERIMÈCHE[1,2], Thibaut PONCIN[1,2,4], François CAMÉLÉNA[1,4], Mary MAINARDIS[1,2], Aymeric BRAILLE[1,2], Alban LERMINE[3], Béatrice BERÇOT[1,2,4], MOABI members[3]

[1] APHP, Saint Louis Hospital, Department of infectious diseases, 75010, Paris, France

[2] Associated expert laboratory for gonococci - National Reference Center for Bacterial STIs, St Louis Hospital, 75010, Paris, France

[3] MOABI (Bioinformatic platform of AP-HP), 33 boulevard Picpus, 75012, Paris, France

[4] Paris University, INSERM, IAME, 75018, Paris, France

Corresponding Author: manel.merimeche-ext@aphp.fr

Neisseria gonorrhoeae causes gonorrhoea, the second most prevalent global bacterial Sexually Transmitted Infection (STI). Untreated gonorrhoea can lead to severe sequelae including upper genital tract infections, ectopic pregnancy, infertility, and increased HIV transmission [1].

The associated expert laboratory for gonococci - National Reference Center for Bacterial STIs [2] at Saint-Louis Hospital, Paris, France performs the phenotypic and genotypic characterization of clinical gonococci strains to monitor the evolution of the resistance of gonococcal strains sent by the laboratories of the French national network.

Genomics and whole genome sequencing (WGS) have the capacity to greatly enhance knowledge and understanding of infectious diseases and clinical microbiology. In this abstract, we present the NG-AR2T pipeline (*Neisseria gonorrhoeae* - Assembly, Resistome, Typing & Tree) which generates complete epidemiological reports from *Neisseria gonorrhoeae* sequenced isolates.

The genomic analysis deployed on the portal G-route of MOABI [3] (Bioinformatic platform of AP-HP) begins with a quality check for coverage and contamination of each sample, then the short reads (raw or trimmed) are assembled *de novo* into contigs. Antimicrobial resistance determinants and sequence types are *in silico* determined by NG-MLST, NG-MAST, and NG-STAR techniques and assigned to each sample automatically. Finally, a phylogenetic tree is generated from the alignment of the assemblies with the detection of chromosomal mutations to show the relationship and genomic distance between the strains.

All this information is then combined with clinical and statistical information to assess the situation and the virulence of each isolate.

Keywords: Neisseria gonorrhoeae, genomic analysis, epidemiological study, whole genome sequencing, antimicrobial resistance, phylogeny

References

[1] Magnus Unemo, Catriona S Bradshaw, Jane S Hocking, et al. Sexually transmitted infections: challenges ahead : 2017. The Lancet Infectious Diseases Commission (2017), 17: e235–79 DOI:http://dx.doi.org/10.1016/ S1473-3099(17)30310-9

[2] https://www.cnr-ist.fr

[3] http://idfseqit.fr/

Poster 136

A Galaxy interface to facilitate multi-block analysis

Maxime BRUNIN*[1], Pierre PERICARD*[1] and Guillemette MAROT[2]

[1] Institut Pasteur de Lille, Univ. Lille, CNRS, Inserm, CHU Lille, US 41 - UMS 2014 - PLBS - Plateforme bililles, F-59000 Lille, France

[2] Univ. Lille, CHU Lille, ULR 2694 - METRICS, Inria, MODAL, F-59000 Lille, France

*These authors contributed equally to this work

Corresponding Author: guillemette.marot@univ-lille.fr

Multi-omics data analysis is one of the main challenges currently faced by integrative biology. Solving it requires combining competences from multiple domains like statistical analysis, computer science and experimental biology. Several types of analytical approaches have been proposed over the past few years in order to perform integrative biology, among them a lot of work around networks (inference or visualization) and more recently around multi-block analyses. The mixOmics package [1] implements multiple statistical analysis methods to integrate different types of omics data (e.g. transcriptomics, metabolomics, proteomics). We focus here on selecting variables in the context of discriminant analysis, where various blocks (each block corresponding to one type of omics) are provided as input.

We built upon the existing mixOmics *block.splsda* function, which performs feature selection simultaneously on several types of omic data measured on the same individuals, with an emphasis on prediction, and deals with the high number of variables. Sparse PLS-DA is a particular case of SGCCA [2] and therefore exploratory analysis relies on correlation circle plots. We provide additional tools in order to check the possibility of overlaying different correlation circles relative to several blocks. The user can also zoom in on the resulting plot to select subsets of relevant correlated variables. Finally, a network in graphml format is built from selected variables and additional variables of interests. Links are drawn between variables when they are correlated, and the network can be visualized externally using a platform like Cytoscape.

The entire pipeline has been integrated into Galaxy [3] and can be installed from the Toolshed (*viscorvar* repository). Galaxy XML wrappers and additional R functions source code are also available on GitLab (https://gitlab.com/bililles/galaxy-viscorvar).

During the JOBIM 2020 conference we will be presenting a live demonstration of our pipeline, from initial multi-omic data integration in Galaxy up to variables network visualization in Cytoscape.

References

1. Florian Rohart, et al. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS computational biology*, 13(11): e1005752, 2017.
2. Arthur Tenenhaus, et al. Variable selection for generalized canonical correlation analysis. *Biostatistics*, 15(3):569-583, 2014.
3. Enis Afgan, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, 46(W1):W537-W544, 2018.

Poster 140

Gene expression predictions using deep learning approach

Camille KERGAL[1], Catherine ANDRÉ[1], Marie-Dominique GALIBERT[1,2], Thomas DERRIEN[1] and Christophe HITTE[1]

[1] Univ Rennes 1, CNRS, IGDR – UMR 6290, F-35000 Rennes, France.
[2] 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

Statistical learning knows a growing number of applications in biology including genome-wide analysis of gene expression. More specifically, Deep Learning (DL) has shown to outperform other machine learning approaches for gene expression prediction [1]. Recently, a DL-based tool, called Basenji [2], was shown to predict gene expression levels in human solely based on the DNA sequence of the reference genome and using Convolutional Neural Network (CNN) techniques. The model learned by the algorithm led to high agreement between predicted expression levels and experimentally measured ones based on CAGE data (Cap Analysis Gene Expression [3]), with correlations coefficients up to r=0.62 in the different tissues used to train the model. The ultimate goal of such method consists in predicting the impact of genome variations on the level of gene expression and thus allowing prioritizing regulatory mutations.

First, as part of the team's work in comparative oncology between human and dog, we adapted the tool to be used with the dog reference genome and canine CAGE data. Using this species-specific strategy to train Basenji, we obtained high gene expression predictions in dog (mean r=0.61). Next, while Basenji employs large human genomic regions (131kb) to train a CNN model, in our work we adjusted its framework to focus on gene promoter regions (1024bp around the Transcription Start Sites) in order to assess whether it improves gene expression predictions. Indeed, promoters are non-coding essential genomic features which regulate gene expression profiles. Using the same CAGE transcriptomic data as input dataset, we were able to reach better predictions (mean r=0.75) as compared to the original Basenji method. Furthermore, Recurrent Neural Network (RNN) methods have demonstrated greater achievements than CNN in the tasks of classification of DNA sequences [4]. We are therefore developing a gene expression prediction tool, based on a RNN and more specifically Long Short-Term Memory strategies (LSTM) to be benchmarked with the CNN approach. We are also investigating the optimization of hyperparameters (dropout and learning rates, layers size, etc.) using bayesian optimization via the GPyOpt library to implement our RNN [5].

Overall, our work highlights deep learning frameworks for modeling gene expression data, based on experimental high-throughput transcriptomic data.

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 Cani-DNA biobank [ANR-11-INBS-0003], veterinarians, breeders and dog owners.

Poster 143

Improving microbial taxonomic profiling with long read technologies

Jean MAINGUY[1], Adrien CASTINEL[2], Olivier BOUCHEZ[2], Sylvie COMBES[3], Carole IAMPIETRO[2], Christine GASPIN[1], Denis MILAN[2,3], Cécile DONNADIEU[2], Claire HOEDE[1] and Géraldine PASCAL[3]

[1] INRAE, UR875 MIAT PF Bioinfo GenoToul F-31326 Castanet Tolosan, France
[2] INRAE, US 1426, GeT-PlaGe, Genotoul, F-31326, Castanet Tolosan, France
[3] GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France

Corresponding author: geraldine.pascal@inrae.fr

In recent years, sequencing technologies allowing the production of very long reads (from 1 to 100kb) have emerged. Offering new possibilities of research in various genomics fields, they are starting to be used alongside with or in addition to previous sequencing strategies. The advantage of long read sequencing could be especially useful in metabarcoding-based exploration of complex microbial community. In metabarcoding methodology, microbial diversity and taxonomic composition of an environment are assessed using a short read PCR amplicon marker (250-500 pb). A common limitation of metabarcoding is the difficulty to assign microbes at the genus or species taxonomic level. Often, barcodes of closely related organisms are too similar and thus cannot discriminate them. Long read sequencing technologies can address these limitations as longer sequences may include more discriminant information. Studies testing this approach by targeting the full 16S rRNA genes (1500 pb) show encouraging results [1] [2] although rRNA genes are present in several copies and might not be conserved within a given organism.

In order to overcome the current limitations, we developed a method that identifies new genomic regions that could be targeted for long read technologies. This method consists to identify regions bounded by two universal and single copy genes that are separated by a consistent length across genomes. First, the universal genes are identified using eggNOG v5.0 [3] and then retrieved in a selection of representative RefSeq genomes. Each candidate gene pair that presents a consistent length and the same orientation in more than 95% of the genomes is selected. Finally, we compute the taxonomic resolution for each selected region that is the percentage of species that can be unambiguously identified. Using this pipeline, we identified 84 regions with a sequence length ranging from 500pb to 5000pb and a taxonomic resolution from 71% to 93%. In order to perform a PCR amplification, we designed primers using the Degeprime [4] tool and assessed primer coverage using ecoPCR [5]. According to previous metrics, we have selected 8 primer pairs targeting 4 regions and we are currently testing them *in vitro* on a mock of 8 bacterial species.

Acknowledgements

This work is part of the SeqOccIn project supported by Region Occitanie and FEDER

References

[1] Jethro S Johnson, Daniel J Spakowicz, Bo-Young Hong, Lauren M Petersen, Patrick Demkowicz, Lei Chen, Shana R Leopold, Blake M Hanson, Hanako O Agresta, Mark Gerstein, et al. Evaluation of 16s rrna gene sequencing for species and strain-level microbiome analysis. *Nature communications*, 10(1):1–11, 2019.

[2] Andres Santos, Ronny van Aerle, Leticia Barrientos, and Jaime Martinez-Urtaza. Computational methods for 16s metabarcoding studies using nanopore sequencing data. *Computational and Structural Biotechnology Journal*, 2020.

[3] Jaime Huerta-Cepas, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, Ivica Letunic, Thomas Rattei, Lars J Jensen, Christian von Mering, and Peer Bork. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47(D1):D309–D314, 11 2018.

[4] Luisa W Hugerth, Hugo A Wefer, Sverker Lundin, Hedvig E Jakobsson, Mathilda Lindberg, Sandra Rodin, Lars Engstrand, and Anders F Andersson. Degeprime, a program for degenerate primer design for broad-taxonomic-range pcr in microbial ecology studies. *Appl. Environ. Microbiol.*, 80(16):5116–5123, 2014.

[5] Gentile Francesco Ficetola, Eric Coissac, Stéphanie Zundel, Tiayyba Riaz, Wasim Shehzad, Julien Bessière, Pierre Taberlet, and François Pompanon. An in silico approach for the evaluation of dna barcodes. *BMC genomics*, 11(1):434, 2010.

Poster 145

Protein multiple alignments: sequence-based versus structure-based programs.

Mathilde CARPENTIER[1] and Jacques CHOMILIER[2]

[1] Institut Systématique Evolution Biodiversité (ISYEB), Sorbonne Université, MNHN, CNRS, EPHE, Paris, France

[2] Sorbonne Université, CNRS, MNHN, IRD, IMPMC, BiBiP, Paris, France

Corresponding Author: mathilde.carpentier@mnhn.fr

Paper Reference: **Carpentier,M. and Chomilier,J. (2019) Protein multiple alignments: sequence-based versus structure-based programs. *Bioinformatics*, 35, 3970–3980. https://doi.org/10.1093/bioinformatics/btz236**

Multiple protein sequence alignments are used daily in bioinformatics to annotate and predict the characteristics of currently mass-produced sequences. The quality of their results has been assessed many times and have reached a plateau. Proteins fold into stable three-dimensional structures with a topology much more conserved than sequence. Consequently, it should be advantageous to use this other source of information to align the sequences, in order to find the homologous positions. Several programs have been developed to align proteins according to their structure or to their sequence and their structure. In this study, we wanted to assess the added value of structural information in multiple alignments and compared the results of these programs to the results of the sequence alignment programs.

We compared the multiple alignments resulting from 25 programs either based on sequence, structure, or both, to reference alignments deposited in five databases (BALIBASE 2[1] and 3[2], HOMSTRAD[3], OXBENCH[4] and SISYPHUS[5]). On the whole, the structure-based methods compute more reliable alignments than the sequence-based ones, and even than the sequence+structure-based programs whatever the databases. Two programs lead, MAMMOTH[6] and MATRAS[7], nevertheless the performances of MUSTANG[8], MATT[9], 3DCOMB[10], 3DCOFFEE[11] are better for some alignments. The advantage of structure-based methods increases at low levels of sequence identity, or for residues in regular secondary structures or buried ones. Concerning gap management, sequence-based programs set less gaps than structure-based programs. Concerning the databases, the alignments of the manually built databases are more challenging for the programs.

[1] J.D. Thompson, F. Plewniak, O. Poch, BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs, Bioinformatics. 15 (1999) 87–88.

[2] J.D. Thompson, P. Koehl, R. Ripp, O. Poch, BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark, Proteins. 61 (2005) 127–136. https://doi.org/10.1002/prot.20527.

[3] K. Mizuguchi, C.M. Deane, T.L. Blundell, J.P. Overington, HOMSTRAD: a database of protein structure alignments for homologous families., Protein Sci. 7 (1998) 2469–2471. https://doi.org/10.1002/pro.5560071126.

[4] G.P.S. Raghava, S.M.J. Searle, P.C. Audley, J.D. Barber, G.J. Barton, OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy, BMC Bioinformatics. 4 (2003) 47. https://doi.org/10.1186/1471-2105-4-47.

[5] A. Andreeva, A. Prlić, T.J.P. Hubbard, A.G. Murzin, SISYPHUS--structural alignments for proteins with non-trivial relationships, Nucleic Acids Res. 35 (2007) D253-9. https://doi.org/10.1093/nar/gkl746.

[6] A.R. Ortiz, C.E. Strauss, O. Olmea, MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison, Protein Sci. 11 (2002) 2606–2621.

[7] T. Kawabata, MATRAS: A program for protein 3D structure comparison, 31 (2003) 3367–3369.

[8] A. Konagurthu, J. Whisstock, P. Stuckey, A. Lesk, MUSTANG: a multiple structural alignment algorithm, Proteins. 64 (2006) 559–574. https://doi.org/10.1002/prot.20921.

[9] M. Menke, B. Berger, L. Cowen, Matt: Local Flexibility Aids Protein Multiple Structure Alignment, PLoS Comput Biol. 4 (2008) e10. https://doi.org/10.1371/journal.pcbi.0040010.

[10] S. Wang, J. Peng, J. Xu, Alignment of distantly related protein structures: algorithm, bound and implications to homology modeling, 27 (2011) 2537–2545. https://doi.org/10.1093/bioinformatics/btr432.

[11] O. O'Sullivan, K. Suhre, C. Abergel, D.G. Higgins, C. Notredame, 3DCoffee: combining protein sequences and structures within multiple sequence alignments., J Mol Biol. 340 (2004) 385–395. https://doi.org/10.1016/j.jmb.2004.04.058.

Poster 148

Efficiency and strategies of model training to call methylation from Oxford Nanopore reads

Paul Terzian[1], Remy-Felix Serre[2], Clément Birbes[1], Christine Gaspin[1], Denis Milan[2], Cécile Donnadieu[2], Carole Iampietro[2], Christophe Klopp[1] and Céline Vandecasteele[2]

[1] MIAT, PF Bioinfo GenoToul, Université de Toulouse, INRA, Chemin de Borde Rouge, 31320 Castanet-Tolosan, France

[2] INRA, US 1426, GeT-PlaGe, Genotoul, 31320 Castanet-Tolosan, France

[3] LIPM INRA/CNRS, chemin de Borde Rouge, 31320 Castanet-Tolosan,France

Corresponding author: pl.terzian@gmail.fr, celine.vandecasteele@inrae.fr, christophe.klopp@inrae.fr

SeqOccIn is a 3 years project started in 2019 which aims at gaining experience in longs reads sequencing technologies applications. Regarding DNA methylation calling, we have an interest in the possibility to detect 5mC (5-Methylcytosine) in large eukaryotes genomes, such as birds or mammals, from native DNA reads.

Oxford Nanopore Technologies sequencers produce an electrical signal enabling to call both long DNA reads and their chemical modifications. Despite its great capability, this technology is still unstable and so are the analysis methods. The available tools to detect nucleotide modifications are model-based (HMM, deep learning), classically using generic models trained for different modifications like 5mC (5-Methylcytosine) for human and 6mA (6-Methyladenine) for *E. Coli* reads. There is little feedback on the impact of using these generic models to detect 5mC or 6mA modifications on reads from other species.

Tools such as Tombo[1] and DeepSignal[2] enable model training. In a previous study on Ralstonia Solanacearum, we compared GTWWAC motifs based 6mA modification detection using an *E. Coli* GATC motif trained model versus a specifically trained model. Results showed that the model specific to *R. solanacearum* GTWWAC motifs[3] was more accurate than the generic model.

However, training such models requires control data that can be tedious to produce when dealing with large genomes like mammals. Two strategies can be used, first producing two datasets of fully methylated and fully unmethylated reads, second using BS-Seq Illumina reads as ground truth to find high confidence CpG sites. These sites are then extracted from nanopore reads to train a model. Today, we focus on evaluating the efficiency of these two approaches to train methylation calling models.

At the same time, collaborators in the Seqoccin project are producing long read genome assemblies. Because they are using the same read sets as we do to produce their references, we also aim at comparing 5mC modification detection accuracy using a generic species reference (Ensembl or NCBI) versus a newly specific reference closer to the genotypes studied.

Finally, we present a Nanopore long read Nextflow[4] processing pipeline built to easily call methylation and train models.

References

[1] Marcus H Stoiber, Joshua Quick, Rob Egan, Ji Eun Lee, Susan E Celniker, Robert Neely, Nicholas Loman, Len Pennacchio, and James B Brown. De novo identification of dna modifications enabled by genome-guided nanopore signal processing. *BioRxiv*, page 094672, 2016.

[2] Peng Ni, Neng Huang, Zhi Zhang, De-Peng Wang, Fan Liang, Yu Miao, Chuan-Le Xiao, Feng Luo, and Jianxin Wang. Deepsignal: detecting dna methylation state from nanopore sequencing reads using deep-learning. *Bioinformatics*, 35(22):4586–4595, 2019.

[3] Ivan Erill, Marina Puigvert, Ludovic Legrand, Rodrigo Guarischi-Sousa, Céline Vandecasteele, João C Setubal, Stephane Genin, Alice Guidot, and Marc Valls. Comparative analysis of ralstonia solanacearum methylomes. *Frontiers in plant science*, 8:504, 2017.

[4] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature biotechnology*, 35(4):316–319, 2017.

Poster 150

GLeaves : Bringing clinical variant interpretation to genome scale

Nicolas DERIVE[1], Laurent FROBERT[2], Mouhamadou NIANG[1], Virginie SAILLOUR[1], Mario NEOU[1], Adrien LEGENDRE[1], Vivien DESHAIES[2], Camille BARETTE[1,2], Alban LERMINE[1,2] & Le groupe expert bioinfo SeqOIA

1. SeqOIA-IT, Paris, France,
2. MOABI AP-HP, Paris, France

Corresponding Author: nicolas.derive-ext@aphp.fr

PFMG2025 is rising and Auragen & SeqOIA genome scale diagnosis oriented sequencing platforms emergence has led to novel questions and challenges in the precision medicine field in France. How to analyze such an important data volume, within a clinically compatible delay, in a decentralized manner (as clinicians and biologists are based in multiple places)? SeqOIA choosed to develop by ourselves new medical prescription and variant interpretation tools focused towards these stakes.

SeqOIA-IT bioinformatics platform was able to take advantage of its strong interactions with MOABI, AP-HP bioinformatics platform, to develop GLeaves, a web-based genome sequencing results interpretation tool. It is based on Leaves MOABI's sequencing results interpretation tool focused on gene panels and exome, introducing big data technologies needed by the genome scale shift, like Elasticsearch and MongoDB. Its features have been adapted to the specifics of SeqOIA organization, matching sequencing results to computerized medical prescription notably, and the need for remote collaboration.

Thus, two main disease categories can be interpreted through GLeaves, « Rare diseases » and « Cancers », including all the steps from pinpointing the variants of interest to generating the interpretative report. This highlighting is made possible by the ability to apply a large set of filters in real time dynamically chosen by the biologist. These filters cover all possible aspects of interpretation: sequence quality (GATK[1] HaplotypeCaller) variants frequency in general population (gnomAD[2]), their presence in various disease oriented databases (COSMIC[3], OMIM[4]...) or global databases (for example dbSNP[5]). They also include prediction scores (CADD[6], SIFT[7], PolyPhen-2[8], SpliceAI[9]…), and position filters (chromosome(s), gene(s)…). Finally, the complete computerization of the medical prescription process allowed us to require upstream HPO patient's clinical signs inclusion[10], permitting automatic import towards their use in results interpretation phase.

Report generation is also simplified and computerized from the interpretation phase in the app, leading to reducing errors while its being written.

Finally, linked to interpreting biologist number and multi-site organization, we built a collaboration focused feature bundle: a « chat » like interface on prescriptions, the ability to take or give a prescription results interpretation to someone, to save and share interpretative filters or variants lists from a given prescription.

These are leading us toward maximizing service to patients by making the most of this new technological step, while at the same time allowing a quick and fluid analysis of the results.

References

[1] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al., *Genome Res.*, *20*, 1297–1303, 2010.
[2] K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, et al., *bioRxiv*, 531210, 2019.
[3] J. G. Tate, S. Bamford, H. C. Jubb, Z. Sondka, D. M. Beare, N. Bindal, H. Boutselakis, C. G. Cole, C. Creatore, E. Dawson, et al., *Nucleic Acids Res.*, *47*, D941–D947, 2019.
[4] J. S. Amberger, C. A. Bocchini, A. F. Scott, A. Hamosh, *Nucleic Acids Res.*, *47*, D1038–D1043, 2019.
[5] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, K. Sirotkin, *Nucleic Acids Res.*, *29*, 308–311, 2001.
[6] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, M. Kircher, *Nucleic Acids Res.*, *47*, D886–D894, 2019.
[7] P. C. Ng, S. Henikoff, *Nucleic Acids Res.*, *31*, 3812–3814, 2003.
[8] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, S. R. Sunyaev, *Nat. Methods*, *7*, 248–249, 2010.
[9] K. Jaganathan, S. Kyriazopoulou Panagiotopoulou, J. F. McRae, S. F. Darbandi, D. Knowles, Y. I. Li, J. A. Kosmicki, J. Arbelaez, W. Cui, G. B. Schwartz, et al., *Cell*, *176*, 535-548.e24, 2019.
[10] S. Köhler, L. Carmody, N. Vasilevsky, J. O. B. Jacobsen, D. Danis, J.-P. Gourdine, M. Gargano, N. L. Harris, N. Matentzoglu, J. A. McMurry, et al., *Nucleic Acids Res*, *47*, D1018–D1027, 2019.

Poster 151

CSTB : a web service to identify CRISPR sequences to differentially target bacteria

Cécile Hilpert[1], Sophie Lematre[1], Audrey Reuter[1], Sarah Bigot[1], Christian Lesterlin[1] and Guillaume Launay[1]

[1] University of Lyon, CNRS, UMR 5086 Molecular Microbiology and Structural Biochemistry, 7 passage du Vercors F-69367, Lyon, France

Corresponding author: cecile.hilpert@ibcp.fr

CRISPR-cas9 systems contain a small guide RNA (sgRNA) of a hundred of nucleotides followed by a variable specific sequence of around 20 nucleotides that hybrids to the target genome next to a PAM motif. This specific sequence triggers the recruitment of the Cas9 protein which can have different effects on genome, like inducing DNA double-stranded breaks or inhibiting gene transcription or replication [1]. Our main purpose is to identify specific sequences that will selectively hybridize to some bacterial genomes, in order to give to CRISPR systems strains specificity.

To do that, we released the CSTB web service to identify all CRISPR target sequences present in a pool of target genomes and absent from another. The search for CRISPR target sequences can be genome-wide or restricted to the homologous genes of a query sequence.

To quickly achieve the required search and comparisons we set up the storage of pre-detected CRISPR target sequence across all genomes. We also implemented a 2bits per base encoding scheme of the kmer for fast sequence comparisons, which also allows for efficient detection of degenerated CRISPR motifs, by considering mismatches between kmers.

This tool is available as a user-intuitive web service and uses libraries that we implemented : a C library for heavy computation integrations and Python libraries for database interrogation and post-processing of results.

The web service uses a back-end database of 2914 representative and complete genomes. In the client input interface, users can select target genomes, excluded genomes, target sequence length and configure the PAM motif. In specific gene features, they also provide a gene sequence. The client output interface displays the CRISP target sequences in tabular and graphical formats. The tabular format lists all target sequences with its number of occurrences in each organism. For each organism, the graphical format represents the repartition of all target sequences and target sequences coordinates. When targeting homologous genes, an additional representation displays the repartition of target sequences along the homologous genes. To assess the specificity of a single sgRNA, its number of degenerated CRISPR motifs can be reported across all excluded genomes.

Acknowledgements

We thank IBCP CRI for infrastructure, ShangNong Hu, Brice Letcher and Timothée Sluys for the first version of this project and Louis Béranger for some development and help with web interface.

References

[1] Fuguo Jiang and Jennifer A Doudna. Crispr–cas9 structures and mechanisms. *Annual review of biophysics*, 46:505–529, 2017.

Poster 152

Genome-Scale Metabolic Networks of *Penicillium chrysogenum*: evolution, combination and new reconstruction

D. NEGRE[1], A. LARLHIMI[2], E. WATIER[1], J. BOURDON[2], E. LORTHEAU[1,2], A. SIEGEL[3], J. NICOLAS[3], L. MESLET-CLADIERE[4], C. ROUILLIER[1], E. GENTIL[1], N. RUIZ[1], O. GROVEL[1], Y. F. POUCHUS[1], S. BERTRAND[1]

[1] MMS – Groupe Mer, Molécules et Santé – EA 2160 – Université de Nantes, 9 Rue Bias, 44035, Nantes, France

[2] LS2N – Laboratoire des Sciences du Numérique de Nantes – UMR CNRS 6004 – Université de Nantes, 2 Chemin de la Houssinière, 44322, Nantes, France

[3] IRISA – Institut de Recherche en Informatique et Recherches Aléatoires, Campus de Beaulieu, 263 avenue du Général Leclerc, 35042, Rennes, France

[4] LUBEM – Laboratoire Universitaire de Biodiversité et Ecologie Microbienne – EA 3882 – Université de Bretagne Occidentale, Technopôle Brest-Iroise, 29280, Plouzané, France

Corresponding Author: delphine.negre@univ-nantes.fr

Filamentous fungi are chemical factories and natural product producer. Modelling metabolism through Genome-Scale Metabolic Network (GSMN) reconstruction represents one way to better understand their biosynthesis mechanisms. Since the first *Penicillium chrysogenum* GSMN was published in 2008 [1], evolution and optimization of reconstruction methods has led to the emergence of various automatic tools. Two other networks resulting from many fungal species simultaneous reconstruction were published in 2016 [2] and 2018 [3]. As those available GSMNs are subject to intra- and inter-database variability [4], their direct comparison remains complicated. The observed differences often reflect databases evolution and enrichment. Furthermore, they also mainly reflect the method used for reconstruction instead of the biological reality of the organism. Even if the data presented in the GSMNs are based on genome knowledge, these approaches do not provide *sensu stricto* identical results. Thus, understanding what constitutes false positive or complementarity within GSMNs comparison will be a key step in the manual curation.

The increase in data availability points out once again the need for standardization of model reporting. Focusing initially on the MetaCyc database, a draft network representing the topology of all the reactions already published was obtained. Then, this draft was enriched by the contribution of intermediate networks from the functional annotation of the *P. chrysogenum* genome and by the search for homology with pre-existing models. Finally, the mandatory steps of manual curation allow to extend this GSMN with reactions from other database or created *de novo*. The traceability and reproducibility of the resulting GSMN was made possible by the use of AuReMe (AUtomatic REconstruction of MEtabolic models), a pipeline dedicated to "à la carte" reconstruction of GSMNs [5].

As a result, the combination of these various resources, associated with metabolomic approaches, led to a new GSMN model of *P. chrysogenum*, as close as possible to reality, by pooling existing knowledge. *In fine*, GSMN exploration, in relation to various physical and biological constraints, is expected to allow us to understand natural products biosynthesis regulation.

Keywords: genome-scale metabolic network (GSMN), data integration, *Penicillium chrysogenum*

Acknowledgements

This work was supported by the ANR-18-CE43-0013 FREE-NPs.

References

[1] Agren, R. *et al.* The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for *Penicillium chrysogenum*. *PLoS Computational Biology* 9, 2013.

[2] Castillo, S. *et al.* Whole-genome metabolic model of *Trichoderma reesei* built by comparative reconstruction. *Biotechnology for Biofuels* 9, 252, 2016.

[3] Prigent, S. *et al.* Reconstruction of 24 *Penicillium* genome scale metabolic models shows diversity based on their secondary metabolism. *Biotechnology and Bioengineering* 115, 2604–2612, 2018.

[4] Pham, N. *et al.* Consistency, Inconsistency, and Ambiguity of Metabolite Names in Biochemical Databases Used for Genome-Scale Metabolic Modelling. *Metabolites* 9, 28, 2019.

[5] Aite, M. *et al.* Traceability, reproducibility and wiki-exploration for “à-la-carte” reconstructions of genome-scale metabolic models. *PLoS Computational Biology* 14, e1006146, 2018.

Poster 153

Use of small RNA sequencing to shed light on host-parasite interactions during infection by the microsporidian *Anncaliia algerae*

Cédric BICEP[1], Ivan WAWRZYNIAK[1], Valérie POLONAIS[1], Fréderic DELBAC[1], Jean-Melaine SION[1] and Éric PEYRETAILLADE[1]

[1] Laboratoire Microorganismes : Génome et Environnement, Université Clermont Auvergne, CNRS, UMR 6023, 1 Impasse Amélie Murat, F 63000, clermont-Ferrand, France
Corresponding Author: cedric.bicep@uca.fr

Abstract

Microsporidia are known to be obligate intracellular parasites that have undergone a genome reduction throughout their evolution from their fungal ancestor[1]. Microsporidia are opportunistic parasites known to infect humans and animals, especially immune-deficient individuals, even though individual species usually have a narrow host range[2]. In contrast, *Anncaliia algerae* has the broadest host range and harbors a unique expended pool of transposable element families which are suspected to promote bidirectional lateral gene transfer with their host [3].

Microsporidia infection impacts the host's cell cycle and reduces apoptosis[4], but the molecular mechanisms involved are still poorly understood. To gain new insights into host parasite cross talk to develop strategies to control microsporidian spreading, we investigated the role of small non-coding RNA (ncRNA). NcRNA are short non-coding RNA molecules (~21-30nt) that can regulate gene expression via post-transcriptional gene silencing. In intracellular pathogens, ncRNA are also known to be exported to host cell cytosol to control host gene expression. This mechanism appears as a new virulence strategy.

We took advantage of small RNA sequencing to identify the host's differentially expressed small ncRNAs (miRNAs and piRNAs) and characterize the pathogen's specific small ncRNAs that could impact key host's metabolic processes. Human Foreskin Fibroblast (HFF) cells were infected with *A. algerae* and total human and parasitic small RNA were sequenced 72 hours post-infection. Small RNA sequencing data were analyzed using a bioinformatic workflow comprising Cutadapt to trim off the adapter, Bowtie to perform a sequential alignment strategy on miRBase and piRNABank for identification of miRNAs and piRNAs. Differential expression analysis was performed using DESeq, and deregulated miRNA's target genes were predicted using TargetScan. Overall 747 host's miRNAs and 9 piRNAs were over-expressed, whereas 14 miRNAs and 75 piRNAs were under-expressed during infection, when using a fold-change threshold of 6 or higher. To study the impact of *A. algerae* on its host, we generated a multipartite network depicting the relationships between deregulated miRNAs, their predicted target genes and the metabolic pathways they are involved in.

Our first investigations on the human side allowed us to find out that major regulators of the host's cell cycle and apoptosis could be targeted by deregulated miRNAs. Some deregulated miRNAs and their putative targets were also validated by qRT-PCR. Concerning parasites ncRNA, we found out that *A. algerae* expresses a range of small ncRNAs targeting its own transposable elements, suggesting that *A. algerae* represses them 72 hours post-infection. To characterize the specific parasitic miRNA repertoire, we also initiate a predictive method based on microsporidian specific translational initiation signal search associated with secondary structure prediction.

All the data collected bring new insights into the role and regulation of the expended pool of transposable elements of *A. algerae* and the molecular mechanisms allowing parasites to control their host cell.

References

[1] T. Y. James *et al.*, "Reconstructing the early evolution of Fungi using a six-gene phylogeny," *Nature*, vol. 443, no. 7113, pp. 818–822, Oct. 2006.

[2] J. Vávra and J. Lukeš, "Microsporidia and 'The Art of Living Together,'" *Adv. Parasitol.*, vol. 82, pp. 253–319, Jan. 2013.

[3] N. Parisot *et al.*, "Microsporidian Genomes Harbor a Diverse Array of Transposable Elements that Demonstrate an Ancestry of Horizontal Exchange with Metazoans," *Genome Biol. Evol.*, vol. 6, no. 9, p. 2289, Aug. 2014.

[4] R. Martín-Hernández *et al.*, "Microsporidia infection impacts the host cell's cycle and reduces host cell apoptosis," *PLoS One*, vol. 12, no. 2, p. e0170183, Feb. 2017.

Poster 155

Facilitating the connection between local datasets and neXtProt with Semantic Web technologies and AskOmics

Xavier Garnier[1], Anthony Bretaudeau[1,2], Alain Gateau[3], Lydie Lane[3], Fabrice Legeai[1,2], Pierre-André Michel[3], Anne Siegel[1]and Olivier Dameron[1]

[1] Univ Rennes, CNRS, Inria, IRISA - UMR 6074, F-35000 Rennes, France
[2] IGEPP - INRAE : UMR1349, Agrocampus Ouest - Agrocampus Ouest, UMR1349 IGEPP, F-35 042 Rennes, France
[3] CALIPHO Group, SIB - Swiss Institute of Bioinformatics, CMU, 1211 Geneva 4, Switzerland

Corresponding author: xavier.garnier@irisa.fr

1 Introduction

The study of biological mechanisms requires the production of large and heterogeneous datasets. *omics* datasets are produced routinely in labs, and are also available from public databases. Each of them has its own format and linking them require a lot of time. Semantic Web technologies such as RDF and SPARQL are one of the key elements for combining datasets, which has led to the emergence of linked data. An increasing number of biological databases, such as neXtProt, provide RDF versions of their data or expose them via SPARQL endpoints. Combining these resources with project-specific data typically requires both to convert local datasets into RDF, and to build federated SPARQL queries covering multiple SPARQL endpoints. However, this requires both SPARQL proficiency and a good knowledge of the of the various endpoints' data schema. The latest release of AskOmics allows (i) the integration of local datasets into a local triplestore (embedded within AskOmics), (ii) the intuitive composition of queries over multiple sources, and (iii) the automatic generation of the corresponding SPARQL code and its transmission to the query engine.

2 Integrate and query local and remote data with AskOmics

AskOmics is a web software that uses the semantic web technologies (RDF/SPARQL) to integrate multiple data formats, and query them through a user-friendly interface. During data integration, the user provides input files in usual formats (CSV, GFF or BED). AskOmics internally generates the corresponding RDF triples and load them into a triplestore. Two kinds of information are generated, the *content*, corresponds to the original data, and the *abstraction*, describes how the raw data is organized and interlinked.

AskOmics can also be used to explore remote SPARQL endpoints. For this, AskOmics needs the *abstraction* of the distant endpoint. The *abstraction* will be stored in the local triplestore embedded within AskOmics. Users can explore the data schema locally and perform queries on the remote endpoint. We developed Abstractor, which scans distant endpoints and generates their RDF *abstraction* that can be imported into AskOmics.

Once AskOmics contains local datasets and distant endpoints *abstraction*, the query builder can be used to build queries and then. It then generates automatically the corresponding local or federated SPARQL code. Federated queries are sent to a federated query engine (embedded within AskOmics) which takes care of splitting the query and sending the subqueries to the right SPARQL endpoint(s).

3 Save, redo and share datasets and queries

AskOmics is a multi-user web platform that can store each user's datasets and queries privately or publicly. Datasets and queries can also be shared between users. As some query patterns often recur in studies, sharing a set of example queries guides users in the construction of their own queries. In cooperation with neXtProt team, we deployed https://nextprot.askomics.org, an instance containing neXtProt description and a set of queries inspired from neXtProt examples.

4 Ongoing work

AskOmics is still under development. Next features will bring support of UNION and NOT SPARQL keyword through the query builder interface. This will allow a greater coverage of queries used in life science.

Poster 157

Galaxy Genome Annotation implementation in the European Open Science Cloud

Romain Dallet[1], Loraine Brillet-Guéguen[1,2], Gildas Le Corguillé[1], Gianluca De Moro[3], Cymon Cox[3] 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

[3] Centro de Ciências do MAR (CCMAR), Universidade do Algarve, Campus de Gambelas, 8005-139, Faro, Portugal

Corresponding Author: romain.dallet@sb-roscoff.fr

The EOSC-Life project aims to create an open collaborative digital space for life science in the European Open Science Cloud (EOSC). 13 Biological and Medical Research Infrastructures in Europe join forces for this project. 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 (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. This e-infrastructure provides a highly integrated set of "dockerized" GMOD tools (JBrowse, Apollo, Tripal, Chado, etc.). Galaxy [1] is used as a data loading orchestrator for administrators, with dedicated Galaxy tools and workflows, and Python libraries to make all tools work together.

As part of the EOSC-Life WP2, we are implementing the GGA environment in the EOSC and adding a new use case for genome annotation into the GGA resources. The workflow, developed by the CCMAR, aims to transfer genome annotations between closely related marine species – as a test case, pelagic fishes - using genome synteny relationships. The workflow consists of three parts: i) alignment of genomes and extraction of synteny relationships, ii) visualization of the synteny blocks and selection of possible missing annotations, and iii) injection of the new potential annotation on the ORCAE [2] portal.

References

1. Enis Afgan et al. *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
2. Sterck, L., Billiau, K., Abeel, T., Rouzé, P., Van de Peer, Y. *ORCAE: online resource for community annotation of eukaryotes*. Nat. Methods (2012) 9, 1041.

Poster 158

An integrative pipeline for region-based rare variant association studies : quality control, qualifying variant selection and association tests

Gaëlle MARENNE[1], Thomas LUDWIG[1,2], Ozvan BOCHER[1] and Emmanuelle GÉNIN[1,2]

[1] Inserm, Univ Brest, EFS, UMR 1078, GGB, 22 av. Camille Desmoulins, F-29200 Brest, France
[2]CHU Brest, Brest France

Corresponding Author: gaelle.marenne@inserm.fr

Abstract

Sequencing data are widely generated to improve our understanding of underlying biological mechanisms involved in the development of complex traits. This approach allows the detection of all variants across all the allele frequency spectrum and make possible rare variant association studies by collapsing them by genomic regions, usually genes. Unlike genotyping data for which quality control process is well described and provide highly reproducible data, sequencing data are subject to several sources of bias and technical artefacts. Because such studies require the sequencing of a large number of individuals, both cases and controls, collaborative projects usually bring together case and control data generated separately. In this context, quality control (QC) is a crucial and challenging step to limit technical bias and ensure good quality association results.

Here we propose an easy and user-friendly integrative pipeline that runs all the steps of variant and sample quality control, qualifying variant selection, association test, and top results detail table, in an automatic manner. Main strengths of the pipeline are 1) to propose standard QC measures for sequencing data; 2) to compare quality measures between groups to ensure comparability of data for the association test; 3) to manage multi-allelic variants and perform QC by allele; 4) to allow more than 2 groups (eg severe cases – mild cases – control) in order to empower case-control association studies in the presence of case heterogeneity; and 5) to run all steps sequentially and automatically in a quick and efficient way. All intermediate results are described and accessible, and the pipeline is flexible enough to allow user to customize intermediate files according to the specificity of its data.

This integrative pipeline is an interesting tool for researchers to conduct all required steps to achieve a good quality gene-based rare variants association test. Pipeline is wrapped up in R which is a widely used language in statistical genetics. Java and Python scripts are called for more computation efficiency. All scripts are open source and pipeline is freely available on GitLab.

Acknowledgements

We thank Elisabeth Tournier-Lasserve and Chaker Aloui for their fruitful inputs in improving the quality control measures.

The research leading to these results has received funding from 1- the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement n. PCOFUND-GA-2013-609102, through the PRESTIGE programme coordinated by Campus France 2- the National Agency for Research through the RHU programme TRT_cSVD (ANR-16-RHUS-004) ; 3- Region Bretagne, France, through the programme Strategie d'Attraction Durable (SAD); 4- Association Gaétan Saleün, France.

Poster 160

Deployment of Genome Databases for Brown Algae Using Galaxy Genome Annotation

Loraine Brillet-Guéguen[1,2], Arthur Le Bars[2], Delphine Nègre[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

1. Context

As part of the France Génomique Phaeoexplorer, the PIA IDEALG and the ERC SEXSEA projects, more than 50 genomes of brown algae have been sequenced. To scientifically exploit these new genomic resources, the community needs effective tools to present, analyze and value these data.

This can be achieved with the GMOD suite, a collection of open-source applications for visualizing, annotating, and managing genomic data (JBrowse, Apollo, Tripal, Chado, etc.). Relying on tools from this suite, several genome browsers have already been deployed on the ABiMS bioinformatics platform on a small number of marine model organisms these last years (*Ectocarpus siliculosus*, *Phaeodactylum tricornutum*, *Ostreococcus taurii*, etc.). However, manual deployment of all applications is not reasonable as the number of brown algae genomes resources to integrate is growing (error-prone, time-consuming, cost of maintenance in operational condition).

2. Genome databases

Based on the Galaxy Genome Annotation (GGA) project (https://galaxy-genome-annotation.github.io) and in partnership with the BIPAA/GenOuest bioinformatics platforms, we have deployed a new integrated environment dedicated to the management of genomic data. It offers the possibility of making genomes and their annotations available to the community through user-friendly interfaces in an automated way. This e-infrastructure uses lightweight virtualization technologies with Docker containers and is based on the GMOD suite and the Galaxy web portal.

We have implemented the first algal genomes (*Ectocarpus siliculosus*, three species of *Ectocarpus sp.*, *Saccharina japonica*, *Cladosiphon okamuranus*) using the GGA environment. To facilitate the deployment of the up-coming fifty two brown algae genomes, we are automatizing the process by using Docker Swarm as a container orchestrator and the BioBlend Python library to script the data loading through Galaxy.

3. Web portal

We are developing a web portal to provide the community a hub for accessing, visualizing and analyzing all algae genomes and resources: http://application.sb-roscoff.fr/project/phaeoexplorer. This portal is being designed to give access to the GGA environments with genomic and transcriptomic data, with features to visualize or download various datasets, and with hyperlinks to external database resources.

Acknowledgements

This work is supported by the French "Agence Nationale de la Recherche" programs IDEALG (ANR–10–BTBR–04) and France Genomique (AAP 2015), and by the European Research Council program SEXSEA (grant ID 638240). We thank the Genoscope for providing the sequence assembly and the structural annotation of the Phaeoexplorer genomes and Olivier Godfroy, Agnieszka P. Lipinska from LBI2M unit for primary data analysis.

Poster 161

Plasma: e-learning platform for massive data analysis

Jérémy TULOUP[1], Claire VANDIEDONCK[2], Pierre POULAIN[3] and Sandrine CABURET[3]

[1] QuantStack, F-75011 Paris, France.
[2] Université de Paris, INSERM UMRS 1138, Centre de Recherche des Cordeliers, F-75006 Paris, France.
[3] Université de Paris, CNRS UMR 7592, Institut Jacques Monod, F-75013 Paris, France.

Corresponding Authors: claire.vandiedonck@inserm.fr, pierre.poulain@u-paris.fr, sandrine.caburet@ijm.fr

Plasma [1], aka in French "Plateforme d'eLearning pour l'Analyse de données Scientifiques MAssives", aims at creating an interactive tool to teach computational analysis of massive scientific data. Plasma was born out of the need to offer a reproducible and high-performance analysis environment to our students.

Our previous experiences of teaching genomics were not satisfying. Because of the limited availability of computational resources, studied samples were restricted to very small datasets, far from what is nowadays routinely analyzed in research labs. Furthermore, remote access to computational resources was not always possible and the user experience provided by the classical Unix terminal was somewhat intimidating for the students.

Plasma aims at providing an authentic experience of the actual bioinformatic analyses performed in research labs. Jupyter notebooks will be used to describe, implement and teach such analyses. These notebooks are interactive numerical notebooks that integrate computer code in several programming languages (Python, R, Bash, C++...), text, mathematical equations and the visualization of analysis results in the form of graphics or tables. This technology is gradually becoming a standard for data analysis, as evidenced by more than 7.8 millions notebooks on the GitHub collaborative development platform [2] and recent publications on the subject [3-6].

We also wanted a web-based solution that could be easily deployed on bare-metal servers or virtual machines, able to handle numerous, simultaneous and specific analysis environments (supporting any programming languages), with a simple and intuitive management interface.

This project is carried out in collaboration with QuantStack, a company strongly involved in the development of the Jupyter ecosystem. Notebooks will be hosted on high-performance computer servers using the JupyterHub open source and highly customizable technology. Students will be able to connect remotely and carry out their analysis in a user-friendly and powerful environment. Data will be centralized on the servers and readily available for analysis.

The first instance of Plasma is designed for the needs of teachers and students of the European Master of Genetics at Université de Paris. Ultimately, this project is a proof of concept and the implemented solution will be fully documented and freely available to the community (see https://plasmabio.org/ and https://github.com/plasmabio/).

References

1. Jeremy Tuloup. *Plasma: A learning platform powered by Jupyter*. https://blog.jupyter.org/plasma-a-learning-platform-powered-by-jupyter-1b850fcd8624 Published 11/05/2020.
2. Jupyter team, *Estimate of Public Jupyter Notebooks on GitHub*, https://nbviewer.jupyter.org/github/parente/nbestimate/blob/master/estimate.ipynb Viewed on 09/06/2020.
3. Bernadette M. Randles, Irene V. Pasquetto, Milena S. Golshan, Christine L. Borgman. *Using the Jupyter Notebook as a Tool for Open Science: An Empirical Study*. 2017 ACM/IEEE Joint Conference on Digital Libraries, 2017. doi : 10.1109/JCDL.2017.7991618
4. Min Ragan-Kelley, Carol Willing, and Jason Grout. *Jupyter: Tools for the Life Cycle of a Computational Idea*. Siam News, 01/03/2018. https://sinews.siam.org/Details-Page/jupyter-tools-for-the-life-cycle-of-a-computational-idea Viewed on 04/03/2020.
5. Helen Shen, *Interactive notebooks: Sharing the code*, Nature, 515: 151-152, 2014. doi: 10.1038/515151a
6. Jeffrey M. Perkel, *Why Jupyter is data scientists' computational notebook of choice*, Nature, 563: 145-146, 2018. doi: 10.1038/d41586-018-07196-1

Poster 162

Title: Systems Immunology approach using tranSMART: challenges and solutions in integrating new type of data

Adaptive Immune Receptor Repertoire (AIRR) sequencing provides nowadays a valuable set of data to better understand health and disease conditions. Indeed, AIRR, which includes both T-cell receptor (TCR) and B-cell receptor (BCR), are key features of the adaptive immune response. Still AIRR-Seq is a new field of research lacking data management and analysis standards. The AIRR Community, launched in 2005, aims, among others, at developing those standards and recommendations data storage, analysis and sharing. In this latter aspect, the iReceptorPlus consortium aims at developing a federated database of currently available AIRR-Seq databases. Importantly, AIRR data can be combined with other omics data, which will provide additional information on the biology of an individual. Such multi-omics integration is the ground for systems biology approaches, which will offer new avenues for biomarker discovery and new therapeutic target identification. We applied such strategy to better understand autoimmune and inflammatory diseases (AID), through the Transimmunom Project, in which 500 patients with one out of 19 AID and 100 healthy volunteers where recruited. For all the patients, we recorded more than 4000 clinical variable and obtained data from high dimensional flow cytometry, cytokine and autoantibodies serum expression, whole blood RNAseq as well as AIRR-Seq and gut microbiome. Such project being at the interface between immunologists, bioinformaticians and medical doctors, we choose to integrate all the data in the tranSMART data warehouse, which provides a user-friendly database for data exploration and analysis. Initially developed for genomics and transcriptomics data, we will here (i) introducing the Transimmunom database, (ii) describe the specification and integration of AIRR-Seq data in tranSMART and (iii) the process to link AIRR-Seq databases together with tranSMART.

Poster 163

What are the functions of upstream open reading frames (uORFs) in dendritic cells?

Sébastien A. CHOTEAU[1,2], Audrey WAGNER[1], Andreas ZANZONI[1], Lionel SPINELLI[1,2], Philippe PIERRE[2], Christine BRUN[1,3]

[1] Aix-Marseille Univ, INSERM, TAGC, Turing Centre for Living Systems, Marseille, France
[2] Aix-Marseille Univ, CNRS, INSERM, CIML, Turing Centre for Living Systems, Marseille, France
[3] CNRS, Marseille, France

Corresponding Author: pierre@ciml.univ-mrs.fr & christine-g.brun@inserm.fr

The development of high-throughput technologies revealed the existence of non-canonical short open reading frames (sORFs) on most eukaryotic RNAs. Upstream ORFs (uORFs) have been defined as sORFs preceding the main coding sequence (CDS). They are ubiquitous elements conserved across species that may be key players of the translational regulation. To date, uORFs have been essentially reported to be gene expression *cis*-regulatory elements. By reducing the efficiency of translation initiation of the main CDS, uORFs participate to the translational regulatory mechanisms, notably during cellular stress [1]. In mammals, dendritic cells (DCs) play a pivotal role in the immune system by orchestrating both the innate and adaptive responses. While upon infection, the sensing of the pathogen by the DCs may be responsible of a global translational arrest, the concomitant up-regulation of the expression of some proteins has been highlighted. The uORFs may be responsible of the preferential expression of some particular CDSs. Moreover, the discovery of uORF-encoded peptides (sPEPs) led to the assumption that they may also play functional roles in *trans*. Indeed, DCs process peptides to be loaded on major histocompatibility complex (MHC) molecules. These peptides could be encoded by uORFs [2].

In this study, we aim to *(i)* build a resource database of sORFs identified in the human and mouse genomes, to explore *(ii)* the *cis*-regulatory potential and *(iii)* the *trans* functions of the uORFs in these species.

(i) Publicly available data has been gathered to characterize the sORFs [2-7]. The curation of data from computational predictions, Ribo-seq and proteomic experiments and the merging of the redundant information into unique entries represent the added value of this database. This notably enables analysis at gene level.

(ii) The database will be exploited to investigate the possible regulatory role of the uORFs in the translation of stress-induced transcription factors, and to propose a model of translation regulation by the uORFs that will then be assessed by experimental validations.

(iii) A pipeline recently developed by our team [8] allows inferring peptide-protein interactions. It will be used in order to build the first sPEPs-protein interactome that will be explored to scrutinize sORFs *trans* functions.

References

1. Starck *et al.*, Protein translation from open reading frames with alternative initiation codons occurs during induction of cellular stress responses. *Science*, 2016.
2. Erhard *et al.*, Improved Ribo-seq enables identification of cryptic translation events. *Nature Methods*, 2018.
3. Olexiouk *et al.*, An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Research*, 2018.
4. Samandi *et al.*, Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. *eLife*, 2017.
5. Mackowiak *et al.*, Extensive identification and analysis of conserved small ORFs in animals. *Genome Biology*, 2015.
6. Laumont *et al.*, Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nature Communications*, 2016.
7. Johnstone *et al.*, Upstream ORFs are prevalent translational repressors in vertebrates. *The EMBO journal*, 2016.
8. Zanzoni *et al.*, Perturbed human sub-networks by Fusobacterium nucleatum candidate virulence proteins. *Microbiome*, 2017.

Poster 164

Long-term dynamics of protist paleocommunities in a coastal marine ecosystem (Bay of Brest, NW France) revealed by DNA metabarcoding

Pierre CUZIN[1], Malwenn LASSUDRIE DUCHESNE[2], Laure QUINTRIC[1], Cyril NOEL[1], Patrick DURAND[1] and Raffaele SIANO[3]

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

[2] IFREMER-ODE-Laboratoire Environnement Ressources de Bretagne Occidentale (LERBO), Station de Concarneau - Place de la Croix, BP 40537 - 29185 CONCARNEAU, FRANCE

[3] IFREMER-ODE-Laboratoire d'Ecologie Pélagique (PELAGOS), Centre Bretagne - ZI de la Pointe du Diable, CS 10070 - 29280 PLOUZANE, FRANCE

Corresponding author: Raffaele.Siano@ifremer.fr

This study aims at exploring marine protists community dynamics in an estuarine ecosystem over the last centuries by using an ancient DNA metabarcoding approach in sedimentary archives of the Bay of Brest (France). We tested the hypotheses that (1) protist ancient DNA originates mainly from resting stages, and (2) temporal shifts in reconstructed protist communities could reveal anthropogenic and climatic changes.

Three sediment cores were collected in the Bay of Brest (NW Atlantic, France) and isotope dating validated the suitability of sedimentary archives for paleo-ecological analyses. The longest core cover dates back up to 3000 BC. Total, intracellular and extracellular DNA fractions were discriminated using different extraction methods. Illumina Mi-Seq sequencing of two barcode regions, the V4 (400bp) and the V7 (260bp) of the 18SrDNA, were obtained to conduct species diversity analyses and test if shorter sequences, amplifiable from degraded DNA were more adapted to diversity study of paleocommunities. Bioinformatics analyses were conducted with FastQC and MultiQC (data quality control), DADA2 [1] (ASV inference) and RDP along with PR2 [2] database (taxonomy assignment). Statistical analysis and diversity data visualization was carried out using various R packages (Phyloseq, Fantaxtic) along with home-made improvements. A first analysis was carried out with the aims of estimating the overall species diversity (richness and eveness) as well as different ecological distances with the two selected barcodes. A second analysis was carried out using Bayesian change point [3] to reveal timing changes in the community composition metrics over the time. In addition, a chronological clustering was used to suggest the timing and hierarchy of marked breakpoints in the community structure with a multivariate regression tree (MRT) according to (Borcard et al.) [4].

Overall this study validates the possibility of using ancient DNA approach for paleoecology studies in coastal marine ecosystem as well as a tool to highlight the effect of anthropogenic pressures on biological communities.

References

[1] Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. DADA2: High resolution sample inference from Illumina amplicon data. *Nature methods*, 13(7):581–583, July 2016.

[2] Laure Guillou et al. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, 41(D1):D597–D604, November 2012.

[3] Chandra Erdman and John W. Emerson. bcp: An R Package for Performing a Bayesian Analysis of Change Point Problems. *Journal of Statistical Software*, 23(1):1–13, December 2007.

[4] Numerical Ecology with R | Daniel Borcard | Springer.

Poster 167

CD-hit-medoid: Optimization of a nucleotide sequence-clustering tool

Oscar GITTON-QUENT[1], Florian PLAZA OÑATE[1], Emmanuelle LE CHATELIER[1], Nicolas PONS[1] and Mathieu ALMEIDA[1].

[1] MGP MetaGenoPolis, INRA, Université Paris-Saclay, Jouy-en-Josas, France.

Corresponding authors: oscar.gitton-quent@inrae.fr; mathieu.almeida@inrae.fr

The central role of the intestinal microbiota in host health has led to an explosion of studies characterizing this relationship in human and in mice, a convenient model regarding logistic and financial aspects. Therefore, the increasing number of metagenomic studies makes it necessary to create representative non-redundant microbial gene catalog. To answer that need, a 2.57 million microbial gene catalog (here called **Xiao** catalog) was published using 184 whole metagenomic mice faecal samples [1] and a 4.49 million microbial gene catalog called **iMGMC** (for **I**ntegrated **M**ouse **G**ut **M**etagenome **C**atalog) was published in 2019, using 298 whole metagenomic samples [2].

However, the bioinformatics methods used to create these catalogues are based on methodologies presenting major biases to the design of a microbial catalog. For instance, the **CD-HIT** gene-clustering tool [3], commonly used in these approaches promote rapid clustering processing to the detriment of selected gene representativeness.

In this study, we propose a new method **cd-hit-medoid** (https://forgemia.inra.fr/Oscar.Gitton-Quent/cd-hit-medoid) that will select the medoid of a gene cluster to improve representativeness. To test this new method, we applied it to build a 2.67 million gene catalog called **M**ouse **I**ntestinal **M**icrobiota **I**terative **C**atalog (or **MIMIC**), using 451 whole metagenomics murine gut samples.

This new MIMIC catalog was compared to the **Xiao** and **iMGMC** catalog using 92 independent whole metagenomics samples [4]. The comparisons revealed a slight improvement in representativeness for the **MIMIC** catalog (70.45% average mapping) compared to the **Xiao** and **iMGMC** catalog (67.58% and 67.90% average mapping respectively). Furthermore, as the representative genes tend to be more central, we observed a preserved data compression with 2.67 million genes in **MIMIC** vs 2.57 million genes in **Xiao** and 4.49 million genes in **iMGMC**. The **MIMIC** catalog, thus, demonstrates the interest of the cd-hit-medoid strategy for future catalog constructions.

References

1. Xiao, L., Feng, Q., Liang, S. *et al*. A catalog of the mouse gut metagenome. Nat Biotechnol**33,** 1103ï 1108, 2015
2. Till R. Lesker.et al. An Integrated Metagenome Catalog Reveals New Insights into the Murine Gut Microbiome. Cell ReportsP2909-2922.e6, 2020
3. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics, 28(23):3150ï 3152, 2012.
4. Cabral et al., Microbial Metabolism Modulates Antibiotic Susceptibility within the Murine Gut Microbiome. 2019, Cell Metabolism 30, 1ï 24, 2019

Poster 168

South Green Bioinformatic Resources For Tropical and Mediterranean Crops

Mathieu ROUARD[4], Stéphanie BOCS[1], Catherine BRETON[4], Aurore COMTE[2], Frederic DE LAMOTTE[1], Alexis DEREEPER[3], Gaëtan DROC[1], Jean-François DUFAYARD[1], Valentin GUIGNON[4], Chantal HAMELIN[1], Pierre LARMANDE[2], David LOPEZ[1], Frédéric MAHE[5], Guillaume MARTIN[1], Julie ORJUELA-BOUNIOL[2,3], Bertrand PITOLLAT[1], Sébastien RAVEL[5], Manuel RUIZ[1], François SABOT[2], Gautier SARAH[1], Guilhem SEMPERE[6], Maryline SUMMO[1], Ndomassi TANDO[2] and Christine TRANCHANT-DUBREUIL[2]

[1]UMR AGAP, Univ Montpellier, CIRAD, INRA, SupAgro, Montpellier, France
[2]UMR DIADE, Institut de Recherche pour le Développement, 34394 Montpellier, France
[3]UMR IPME, Univ Montpellier, CIRAD, IRD, Montpellier, France
[4] Bioversity International, Parc Scientifique Agropolis II, 34397, Montpellier, France
[5]UMR BGPI, Univ Montpellier, CIRAD, IRD, Montpellier, France
[6]UMR Intertryp, Univ Montpellier, CIRAD, IRD, Montpellier, France

Corresponding Author: manuel.ruiz@cirad.fr

South Green is a bioinformatics platform dedicated to the genetics and genomics of tropical and Mediterranean plants of agronomic interest and their related pathogens. It federates a network of bioinformaticians belonging to different units and institutes of Montpellier (Alliance Bioversity CIAT, CIRAD, INRAE and IRD) with a multidisciplinary expertise ranging from the data integration, bioinformatics software development, sequencing data analyses and high-performance computing. Exchange and collaborative developments are fostered through regular hands-on sessions on synergistic themes (information systems, pangenomic methods, graphical visualizations and workflows managers).

The South Green web portal (www.southgreen.fr) gathers all the information systems and tools developed and supported by the platform. Indeed, South Green ensures the development of original information systems such as GreenPhyl, SNiPlay, Gigwa, AgroLD or the Genome Hubs, and offers sequencing data analysis pipelines through two workflow managers: Galaxy and TOGGLe.

Overall, South Green's mission is to promote these original tools as well as their interoperability. A significant part of activities also comprises hands-on trainings that are regularly offered in the local community as well as with partners in the Africa and Asia on the following topics: Galaxy, NGS analyses, R, Perl, Linux, HPC administration (southgreenplatform.github.io/trainings/). Besides, South Green provides access to computing facilities for both users and developers engaged in this scientific area. South Green is part of the network of platforms of the French Institute of Bioinformatics (IFB).

Poster 171

A comparison of different approaches to estimate disease similarities

Maxime Delmas[1], Fabien Jourdan[1], Yoann Pitarch[2] and Clément Frainay[1]

[1] UMR1331, Toxalim (Research Centre in Food Toxicology), Université de Toulouse, INRAE, ENVT, INP-Purpan, UPS, 31300 Toulouse, France
[2] IRIT, Université de Toulouse, CNRS, Toulouse, France

Corresponding Author: maxime.delmas@inrae.fr

Establishing similarity between diseases has become an important challenge in the last few years. Innovative measures have been proposed to categorize and characterize groups of diseases in order to improve our global knowledge by sharing information between similar diseases. The underlying hypothesis is that similar diseases may be caused by similar molecules or mutations, can be diagnosed using similar biomarkers and phenotypes and may be cured using a similar therapy. However, there is a large heterogeneity in terms of diseases representation and applied statistical methods, making necessary to compare these measures with each others.

Recent reviews [1,2] have listed these measures and organized them into three major classes, based on the features used to represent a disease and establish similarities. In molecular-based approaches [3,4,5], a disease is represented by a gene (or protein) set associated to the disease from associated genotypes or gene differential expressions. Phenotypic approaches [6,7] represent diseases by a set of symptoms or by the vocabulary used in disease-related articles in the literature. Finally, in hierarchical methods [8] a disease is defined as a semantic concept (or a set of concepts) in a structured ensemble such as an ontology.

Eight similarity measures with publicly available data or code were chosen among these three classes and compared on a common set of 247 diseases. A comparison of their distribution, correlation and clustering have been performed. Our results illustrate some fundamental differences between these approaches, revealing divergent and consensual clusters, with some that have been characterized. Finally, the in-depth study of these approaches reveals their complementary, but also allow to identify qualities, bias and limits associated to each.

References

[1] Cheng, L., Zhao, H., Wang, P., Zhou, W., Luo, M., Li, T., Han, J., Liu, S., Jiang, Q., 2019. Computational Methods for Identifying Similar Diseases. Molecular Therapy - Nucleic Acids 18, 590–604.
[2] Yu, Y.-K., 2016. Mechanism-based disease similarity. J Rare Dis Res Treat 1, 1–4.
[3] Paik, H., Heo, H.-S., Ban, H., Cho, S., 2014. Unraveling human protein interaction networks underlying co-occurrences of diseases and pathological conditions. J Transl Med 12, 99.
[4] Suthram, S., Dudley, J.T., Chiang, A.P., Chen, R., Hastie, T.J., Butte, A.J., 2010. Network-Based Elucidation of Human Disease Similarities Reveals Common Functional Modules Enriched for Pluripotent Drug Targets. PLoS Comput Biol 6, e1000662.
[5] Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., Wang, S., 2010. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics 26, 976–978.
[6] Zhou, X., Menche, J., Barabási, A.-L., Sharma, A., 2014. Human symptoms–disease network. Nat Commun 5, 4212.
[7] Caniza, H., Romero, A.E., Paccanaro, A., 2016. A network medicine approach to quantify distance between hereditary disease modules on the interactome. Sci Rep 5, 17658.
[8] Yu, G., Wang, L.-G., Yan, G.-R., He, Q.-Y., 2015. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics 31, 608–609.

Poster 174

Hybrid *de novo* genome assembly using Oxford Nanopore Technology, 10X Genomics Linked-Reads sequencing, and Bionano optical map of a non-model species: the case of the melon-cotton aphid *Aphis gossypii*

Jacques LAGNEL[1], Rafael FERICHE-LINARES[1], Pierre BERTRAND[1], William MARANDE[4], Anne LOISEAU[2], Amalia SAYEH[3], Maxime MANNO[3] and Nathalie BOISSOT[1]

[1] INRAE PACA, UR1052, GAFL, 67 Allée des Chênes, 84140, Avignon, France

[2] INRAE, UMR INRAE/IRD/Cirad/Montpellier SupAgro, CBBGP, 755 avenue du campus Agropolis, 34988, Montferrier-sur-lez cedex, France

[3] GeT-PlaGe (Plateforme Génomique) Campus INRAE. 24 chemin de borde rouge – Auzeville, CS 52627, 31326, CASTANET-TOLOSAN Cedex, France

[4] French Plant Genomic Resource Center Campus INRAE, 24 chemin de borde rouge – Auzeville, CS 52627, 31326, CASTANET-TOLOSAN Cedex, France

Corresponding Author: jacques.lagnel@inrae.fr

Despite the considerable number of insect species, few have been sequenced, primarily because of their small size. In addition, insect genomes can be difficult to assemble due to the combination of high polymorphism, heterozygosity, the presence of repeat regions, and the pooling of polymorphic individuals to form libraries [1]. There are no highly-resolved genomes available for aphids, major pests of cultivated plants. We propose to meet this challenge for *Aphis gossypii,* a major pest of Cucurbits, cotton, citrus. Genome size was estimated to 339Mbp by flow cytometry [2] in 4 chromosomes.

To minimize heterozygosity, we selected a lineage with a low estimated heterozygosity, and we pooled individuals deriving from clonal reproduction.

Our sequencing strategy is to combine data from different technologies, combining linked-reads (10X Genomics + Illumina) and long reads (Oxford Nanopore: ONT), and optical map, in order to take advantage of these methods and overcome the disadvantages of each (size of readings, type of error). From preliminary results obtained from 3 flow cells MinIon (ONT), we get 50X coverage and a N50 of 12kbp. The 10X Genomics sequencing (expected more than 50X coverage) and Bionano optical mapping are in progress.

We performed a de novo assembly from ONT sequencing data. This assembly generated 1202 contigs with a N50 of 1.5Mbp and assembly size of 375Mbp (canu/SMARTdenovo). The contiguity of this first assembly already gave a four times improvement regarding the recently published *Aphis gossypii* genome and even any aphid genomes [3]. The improvement might be due to the inclusion of repetitive sequences that were unplaced in previous assemblies deriving from short-reads sequencing.

The hybrid assembly will be performed combining ONT data and linked-reads (10X Chromium) followed by “super scaffolding” on the optical map (Bionano). We expect to reduce the number of gaps and incorporate a substantial amount of additional sequences into the assembled chromosomes. Results will be presented in the poster.

Acknowledgements

We are grateful to the genotoul bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul) for providing help and computing and storage resources.

References

1 Richards, S. and S. C. Murali. Best practices in insect genome sequencing: what works and what doesn't. Current Opinion in Insect Science 7: 1-7, 2015.

2 Wenger, J. A. *et al.* Whole genome sequence of the soybean aphid, *Aphis glycines*. Insect Biochem. Mol. Biol. doi:10.1016/j.ibmb.01.005, 2017.

3 Quan, Q. *et al.* Draft genome of the cotton aphid *Aphis gossypii*. Insect Biochem Mol Biol 105, 25–32, 2019.

Poster 175

ELTerm: a terminology module for a plant data management system

Lysiane Hauguel[1], Tanguy Lallemand[1], Rayan Eid[1], Fabrice Dupuis[1], Sylvain Gaillard[1], Florian Blessing[1], Sandra Pelletier[1] and Julie Bourbeillon[1]
IRHS, Agrocampus-Ouest, INRAE, Université d'Angers, SFR 4207 QuaSaV, 49071, Beaucouzé, France

Corresponding author: julie.Bourbeillon@agrocampus-ouest.fr

1 Introduction

IRHS is a French plant research laboratory. Its BIDefI (for BIoinformatics plant Defense Investigations) team develops methods and tools to help other teams manage and analyse their datasets:

- Development of a data management system which links research projects, experiments and the biological material being used (along with its characteristics) and produced results,
- Development of various data analysis tools and in particular tools aiming at meta-analysis

In order to perform meta-analyses, it is mandatory to use a unified vocabulary across experiments to describe the datasets, that is to say use shared terminologies to fill in meta-data fields.

2 Terminology management in ELVIS/PREMS

The data management system consists of a data management layer, named ELVIS (for Experiment and Laboratory on Vegetal Information System), and graphical interfaces, called PREMS (for Plant Ressource Management System) [1]. Both layers include a terminology module which contents are used to describe data in other modules. This module allows to store several terminologies related to various topics: organism taxonomy, experimental conditions (in particular stresses), plant development stages, plant anatomy, etc. The general underlying principal is similar to standard terminologies representations such as TermBase Exchange [2]. The objective is to both keep the representation simple without going into more complex representation such as those used in ontologies by knowledge management engineers while still being expressive enough to suit our needs.

A terminology regroups a set of concepts in a direct acyclic graph where nodes are concepts and edges are named and represent relationships between concepts. Each concept is associated to a set of terms which support it. A terminology is ideally generic: for instance "Plant Anatomy". Therefore concepts are represented by generic terms: for instance "Fruit". However biologists usually use specific words: for instance people working on *Malus domestica* refer to "Apple" and not "Fruit". In our representation such specific words are terms associated to the relevant concept. Moreover, in order to keep track of the relevant context of use of a given word, we introduced a context notion. For instance the term "Apple" is associated with the "Malus" context. In PREMS, it will only be presented when people are inputting "Malus" related data. Therefore this context notion allows us to represent both specific and generic information along with the equivalence between the two. It also allows us to map our local ontologies, designed to be close to the day to day use of the biologist teams, with reference ontologies such as the Plant Ontology [3].

Acknowledgements

This work was supported by the region Pays de Loire as part of the RFI "Objectif Végétal" program.

References

[1] Lysiane Hauguel, Fabrice Dupuis, Sylvain Gaillard, Julie Bourbeillon, Claudine Landes, and Sandra Pelletier. Mise en place d'un lims enrichi par une organisation harmonisée des métadonnées. In *Journées Ouvertes de Biologie, Informatique et mathématiques*, Nantes, France, 2-5 July 2019.

[2] ISO/TC37/SC3. Iso 30042:2019 management of terminology resources — termbase exchange (tbx). Technical report, International Organization for Standardization, 2019.

[3] Shulamit Avraham, Chih-Wei Tung, Katica Ilic, Pankaj Jaiswal, Elizabeth A. Kellogg, Susan McCouch, Anuradha Pujar, Leonore Reiser, Seung Y Rhee, Martin M Sachs, Mary Schaeffer, Lincoln Stein, Peter Stevens, Leszek Vincent, Felipe Zapata, and Doreen Ware. The plant ontology database a community resource for plant structure and developmental stages controlled vocabulary and annotations. *Nucleic Acids Research*, 36(suppl 1):D449–D454, 01 2008.

Poster 176

My RNA-tailor is rich : fine modeling of alternative transcripts from long reads

Cyprien BORÉE[1], Aymeric ANTOINE-LORQUIN[2], Jean-Stéphane VARRÉ[2] and Hélène TOUZET[2]

[1] bilille, Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, US 41 UMS 2014 - PLBS, 59000 Lille, France

[2] Univ. Lille, CNRS, Centrale Lille, Inria, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, 59000 Lille, France

Corresponding Authors: {jean-stephane.varre,helene.touzet}@univ-lille.fr

When studying transcriptomes, the principle of RNA alternative splicing is a key phenomenon that dramatically increases the diversity of the subsequent proteome. This phenomenon however is hard to describe with second generation sequencing platforms, because short reads fail to capture the combinatorics of exons. Even the best bioinformatics pipelines still struggle to model the diverse landscapes of variants and deconvoluate individual isoforms [1]. In the last few years, the advent of long read sequencing for both cDNA and direct RNA has changed the game, since these reads have proven to be able to entirely cover messenger RNAs [2]. Analysing this new data requires however to invent new bioinformatics algorithms that take full advantage of the lengths of the reads while accommodating the high error profile and the volume of the data. For example, tools have been recently proposed to map long cDNA/RNA reads on genomes [3] or to cluster them according to their gene family [4].

Here, we address a different problem: given a particular gene of interest whose genomic sequence is known, find all transcripts present in the sample that correspond to this gene and model its alternative transcripts: exon skipping, mutually exclusive exons, alternative donor site, alternative acceptor site, intron retention, multiple promoters and multiple polyadenylation sites. The goal is to give a precise and accurate picture of the gene structure and to quantify the presence of each variant. For that, we exploit the fact that reads might span the full messenger RNA and do not require prior assembly. There are still several sources of difficulty to deal with: presence of sequencing errors, of repeats, of low complexity regions, of partial reads, of pre-messenger RNAs. The method we propose relies on sequence similarity (computed with megablast [5] and Exonerate [6]), motifs for splice sites and consistency of junction breakpoints between the reads. It is implemented in a snakemake workflow developed in Python and Biopython, called RNA-tailor (RNA Alternative Transcripts and Long Reads). The output is available in a variety of formats: multiple sequence alignment, GFF, xlsx spreadsheet. It is also possible to provide a GFF file and compare the predicted exons and transcripts to existing annotations. We have successfully tested it on the mouse transcriptome of the ASTER consortium (ONT cDNA and RNA reads from Brain and Liver, ENA PRJEB25574 and ENA PRJEB27590).

Acknowledgements

This work was supported by ANR ASTER (ANR-16-CE23-0001).

References

1. Kanitz A., Gypas F., Gruber A. J., Gruber A.R., Martin A.R., Zavolan M., Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biology,* vol 16, 150, 2019.
2. Sessegolo C., Cruaud C., Da Silva C., Cologne A., Dubarry M., Derrien T., Lacroix V. and Aury JM., Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules., *Sci Rep,* 9, 2019.
3. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34:3094-3100, 2018.
4. Sahlin, K. and Medvedev, P., De novo clustering of long-read transcriptome data using a greedy, quality-value based algorithm. In International Conference on Research in Computational Molecular Biology, 2019.
5. Zhang Z., Schwartz S., Wagner L. and Miller W. ,A greedy algorithm for aligning DNA sequences. *J. Comp. Biol.*, 7:203–214, 2000.
6. Slater G.S. and Birney E., Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6:31, 2005.

Poster 178

A user-friendly R-Shiny interface for accurate viral and bacterial typing without alignment or assembly

Morgane Bruno[1], Lionel Bigault[1], Maud Contrant[1] and Fabrice Touzain[1]
[1] ANSES, 31 rue des Fusillés, 22440, Ploufragan, France

Corresponding author: morgane.bruno@etu.univ-rouen.fr

Viral and bacterial typing is essential to establish a diagnostic. The main difficulty encountered for typing is the runtime of the alignment over a large number of sequences that delay the identification. Hence, an R-shiny application [1] has been developed to enable biologists to very quickly identify the closest strain without aligning or assembling the reads. In addition, it provides an immediate view of the typing distribution of the strains contrary to the other methods, pointing ancestral nodes in a phylogenetic tree in case of ambiguity or divergent new strain.

This application is divided into three parts. The first part allows the user to download all the FASTA sequences retrieved from the NCBI Nucleotide Database for a given taxid. It can be used to download whole genome sequences or specific gene sequences. The second part consists of a pipeline allowing the construction of a representative phylogenetic tree with those sequences. These two first parts are ran periodically to ensure an up to date typing process. The third part consists in the placement of reads on the phylogenetic tree using RAPPAS [2]. This is the fast typing part to run on high throughput sequencing data in case of emergence. Those pipeline are built with Snakemake [3] and implemented with the R package reticulate [4].

At the moment, the application has only been tested with the Porcine Epidemic Diarrhea Virus, an alpha-coronavirus affecting pigs that results in watery diarrhea and vomiting. Two forms have been described, a moderately virulent form (InDel virus) characterized by an insertion or a deletion in the sequence coding for the spike protein [5] compared to the non-InDel virus, which is a hyper-virulent form [6]. There are no hyper-virulent strains in France, so it is very important to be able to identify them quickly if they emerge in France.

References

[1] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. *shiny: Web Application Framework for R*, 2019. R package version 1.4.0.

[2] Benjamin Linard, Krister Swenson, and Fabio Pardi. Rapid alignment-free phylogenetic identification of metagenomic sequences. *Bioinformatics*, 35(18):3303–3312, 01 2019.

[3] Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 34(20):3600–3600, 05 2018.

[4] Kevin Ushey, JJ Allaire, and Yuan Tang. *reticulate: Interface to 'Python'*, 2019. R package version 1.14.

[5] Béatrice Grasland, Lionel Bigault, Cécilia Bernard, Hélène Quenault, Olivier Toulouse, Christelle Fablet, Nicolas Rose, Fabrice Touzain, and Yannick Blanchard. Complete genome sequence of a porcine epidemic diarrhea s gene indel strain isolated in france in december 2014. *Genome Announcements*, 3(3), June 2015.

[6] Tomoichiro Oka, Linda J. Saif, Douglas Marthaler, Malak A. Esseili, Tea Meulia, Chun-Ming Lin, Anastasia N. Vlasova, Kwonil Jung, Yan Zhang, and Qiuhong Wang. Cell culture isolation and sequence analysis of genetically diverse us porcine epidemic diarrhea virus strains including a novel strain with a large deletion in the spike gene. *Veterinary Microbiology*, 173(3):258 – 269, 2014.

Poster 179

CulebrONT, a snakemake pipeline to benchmark Oxford Nanopore Technologies assemblies and to improve quality control

Julie Orjuela[1,2,5,6], Aurore Comte[1,5,6], Bao Tram Vi[1,4], Sébastien Ravel[3,5], Florian Charriat[3,5], François Sabot[2,5]and Sébastien Cunnac[1]

[1] IPME IRD, University of Montpellier, CIRAD, 911 Avenue Agropolis, 34934, Montpellier Cedex 5, France
[2] DIADE IRD, University of Montpellier, 911 Avenue Agropolis, 34934, Montpellier Cedex 5, France
[3] BGPI CIRAD, University of Montpellier, INRA, Montpellier SupAgro, Campus International de Baillarguet, 34398, Montpellier Cedex 5, France
[4] Agricultural Genetics Institute, National Key Laboratory for Plant Cell Biotechnology, LMI RICE, 00000, Hanoi, Vietnam
[5] South Green Bioinformatics Platform, Bioversity-CIAT Alliance, CIRAD, INRA, IRD, Montpellier, France
[6] These authors contribute equally to this work.

Corresponding author: julie.orjuela@ird.fr, aurore.comte@ird.fr

Genome assembly using long reads obtained by *Nanopore sequencing Technologies* could solve repeats and structural variants in prokaryotic as well as in eukaryotic genomes, resulting in increased contiguity and accuracy. Plenty of softwares and tools are released or updated every week, and a lot of genome are beeing assembled using those tools. *Which assembly tool could give the best results for my favorite organism?* CulebrONT can help you! CulebrONT is a scalar, modulable and traceable snakemake pipeline.
CulebrONT optionally handles base-calling[1] of arbitrarily multiplexed libraries across several *Minion* runs with sequencing quality control for subsequent assembly steps.
CulebrONT includes assembly (Canu[2], Flye[3] and Minipolish[4]), circularisation (Circlator[5]), polishing (Racon[6]) and correction (medaka[7] and nanopolish[8]) steps. These steps can be activated according to user's requests. The most relevant tools commonly used for each step were integrated, as well as at least five quality control tools such as quast[9] and busco [10]. CulebrONT also generates a report compiling information obtained in every step.
This snakemake workflow is an open source solution to help you compare news assemblies (https://github.com/SouthGreenPlatform/CulebrONT_pipeline).

Acknowledgements

This work was supported by the I-Trop High-Performance Cluster from IRD as part of the South Green Bioinformatic platform.

References

[1] Ryan R. Wick, Louise M. Judd, and Kathryn E. Holt. Performance of neural network basecalling tools for oxford nanopore sequencing. *Genome Biology*, 20(1):129, 2019.

[2] 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, May 2017. 28298431[pmid].

[3] https://github.com/fenderglass/flye.

[4] Wick RR and Holt KE. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *PLOS Computational Biology*, 2138(11):1–11, 8 2019.

[5] Martin Hunt, Nishadi De Silva, Thomas D. Otto, Julian Parkhill, Jacqueline A. Keane, and Simon R. Harris. Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biology*, 16(1):294, 2015.

[6] https://github.com/lbcb-sci/racon.

[7] https://github.com/nanoporetech/medaka.

[8] https://github.com/jts/nanopolish.

[9] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. Quast: quality assessment tool for genome assemblies. *Bioinformatics (Oxford, England)*, 29(8):1072–1075, Apr 2013. 23422339[pmid].

[10] Mathieu Seppey, Mosè Manni, and Evgeny M. Zdobnov. *BUSCO: Assessing Genome Assembly and Annotation Completeness*, pages 227–245. Springer New York, New York, NY, 2019.

Poster 181

Benchmark of multiple sequence alignment methods applied on third-generation sequencing

Coralie Rohmer[1], Antoine Limasset[1] and Hélène Touzet[1]
Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL, 59000 Lille, France

Corresponding author: coralie.rohmer@univ-lille.fr

Third-generation sequencing is radically changing the whole sequencing environment. The availability of long reads of dozens or hundreds of kilobases allows to resolve the structure of complex genomes or to deal with polyploidy. However, this data presents a high amount of erroneous bases, including deletions and insertions with systematic error patterns [1]. A range of new tools and methods have been specifically developed to handle this noise and obtain trustworthy sequences from those long reads [2,3,4,5,6]. Those tools are scalable, but they are not able to entirely remove the noise. They all leave a room for improvement. In this work, we investigate whether traditional multiple sequence alignment (MSA) tools that have been primarily designed to analyze families of homologous genes can handle such data. In other words: to what extent can "old" MSA tools adapt to the error profile and length of long reads.

To assess this problem, we developed a benchmark that allowed us to evaluate the performance of MSA tools under varying conditions: composition and length of the target region (from 50nt to 5000nt), sequencing coverage (from $\times 10$ to $\times 150$), error profile (up to 10%). First, reads are aligned to the target region with Minimap2 and truncated to obtain piles of partial reads covering the region. MSA is then performed on this selection of reads. Lastly, we compute a series of metrics: consensus sequence with identity rate, gap rate, ambiguous character rate, computational time. This workflow is developed in Snakemake.

We have used it to compare the most popular MSA tools with complementary alignment strategies (POA, Muscle, Clustal Omega, T-Coffee, Maft and Kalign) on real Nanopore sequencing reads (*E.coli* SRR10177137 and *COVID 19* SRR11267570) as well as simulated reads to monitor the error profile.

As a result, we observed several interesting behaviors. First, the time and memory resources required by the different MSA methods vary vastly and do not show the same evolution according to the different parameters. The resulting accuracy is also very dissimilar across the methods, some methods being unable to produce high accuracy or being not robust to small parameter changes. All those observations are of prime interest for further developments to handle the error rate of long reads and to develop novel correction or polishing methods.

Acknowledgements

This work is supported by Region Hauts-de-France.

References

[1] R. Krishnakumar et al. Systematic and stochastic influences on the performance of the MinION Nanopore sequencer across a range of nucleotide bias. *Scientific reports*, 8(1):1–13, 2018.

[2] S. Koren et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*, 27(5):722–736, 2017.

[3] C.L. Xiao et al. Mecat: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *nature methods*, 14(11):1072, 2017.

[4] R. Vaser et al. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome research*, 27(5):737–746, 2017.

[5] P. Morisse et al. Consent: Scalable self-correction of long reads with multiple sequence alignment. *BioRxiv*, page 546630, 2019.

[6] R. Warren et al. ntEdit: scalable genome sequence polishing. *Bioinformatics*, 35(21):4430–4432, 2019.

Poster 182

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: xmorin@biologie.ens.fr, fischer@biologie.ens.fr, mthomas@biologie.ens.fr

In recent years, single-cell RNA-seq (scRNA-seq) has fostered the understanding of complex processes (e.g. cell differentiation, tumorigenesis) and their underlying cell heterogeneity at a remarkable high resolution. By providing gene expression at single cell resolution, this technique allows for the identification of new cell subtypes and their corresponding markers. However 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 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. In particular, this limitation may cause a partial loss of signal in poorly-annotated genomes. For example, the annotation of the chicken *Gallus gallus* is not yet as complete as for other well-studied organisms, such as human or mouse (Kuo et al. 2017). We wondered to which extent this incomplete annotation affects the scRNA-seq data analysis. In this respect, we propose a novel approach to improve genome annotation and subsequent scRNA-seq analyses at a reasonable cost, using long-read transcriptome in matching cell sample.

In order to identify the key transcriptional switches that occur during the neurogenic transition of vertebrate neural progenitors, we produced scRNA-seq data (10x Genomics) 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 lost while generating the count matrix. Visualizing the aligned reads in a genome browser revealed that significant signal was located outside of several known genes, thus missing 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 genes. We thus concluded that loss of scRNA-seq signal was due to shortcomings in the reference annotation files. To address this issue, we generated bulk long-read RNA-seq (Oxford Nanopore Technologies, ONT) from samples matching our scRNA-seq data, in order to delineate the transcripts specific of these cells. ONT was chosen as the sequencing starts in 3', as in 10x Genomics. We exploited the long-reads data to expand the reference annotation (from NCBI and Ensembl), using Stringtie2 (Kovaka et al. 2019) to assemble the transcriptome and complete the existing annotation. We used the resulting new reference annotation for all our single-cell downstream analyses. Overall, we found 134 novel genes and 164 novel transcripts, while 507 genes were elongated in 3'. Most importantly for our analysis, only 17% of single-cell reads remained unassigned. We are currently evaluating the impact of this approach on the downstream scRNA-seq analyses and interpretation, with respect to our biological questions.

This approach could be used to improve single-cell transcriptomic analyses of any other poorly-annotated genome, provided that single-cell and long reads data (ideally in matching cells) are available.

References

1. Kovaka, Sam, Aleksey V. Zimin, Geo M. Pertea, Roham Razaghi, Steven L. Salzberg, and Mihaela Pertea. 2019. "Transcriptome Assembly from Long-Read RNA-Seq Alignments with StringTie2." *Genome Biology* 20 (1): 278.
2. Kuo, Richard I., Elizabeth Tseng, Lel Eory, Ian R. Paton, Alan L. Archibald, and David W. Burt. 2017. "Normalized Long Read RNA Sequencing in Chicken Reveals Transcriptome Complexity Similar to Human." *BMC Genomics* 18 (1): 323.

Poster 183

Comparative transcriptomics upon shutdown of a major player in human epitranscriptome regulation

Julie Ripoll[1], Sébastien Relier[2], Amandine Bastide[2], Alexandre David[2] and Eric Rivals[1]

[1] Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, CNRS et Université de Montpellier, 161 rue Ada, 34095 Montpellier Cedex 5, France

[2] IGF, CNRS, INSERM, Univ. Montpellier, F-34094 Montpellier, France

Corresponding Authors: jripoll@lirmm.fr, rivals@lirmm.fr

Analogous to the epigenome, epimodifications exist in major types of RNAs. For instance, N6-methyladenosine (m6A) is the most abundant methylation in eukaryotic mRNAs epitranscriptome. This m6A methylation may affect RNA processing, nuclear export, RNA translation and decay, as well as some targeted gene isoforms and response pathways to stressors. These misregulations of these pathways may contribute to cancer development [1,2]. The first identified m6A demethylase, the fat mass and obesity-associated protein (FTO) in cancer stem cells, was identified as a regulator of RNA splicing events where an overlap between m6A and splice sites was observed on its targets [3,4]. FTO knock-down leads to substantial changes in pre-mRNA splicing with exon skipping events [5] and appeared preferentially involved in $m6A_m$ demethylation next to the cap according to cellular RNA localisation [6]. Intriguingly, FTO may act as an oncogene [7], or as a tumor suppressor [8] or cancer stem cells repressor [9] according to cancer type.

The discrepancy in the consequences of FTO activity according to cancer type may reflect its impact on different regulation pathways, or the distinct roles of m6A vs $m6A_m$ on mRNA regulation. To address this question, we perform a comparative transcriptome study on 5 different cancer cell types using RNA-seq: VHL deficient ccRCC cells, CRC cells [9], HEK293T cell line [5], AML cells [10] and Hep-G2 cell line. To decipher the complement of RNA isoforms, we developed reproducible pipelines that adopt three alternative and complementary approaches: an assembly based approach, a mapping based one (as recommended in [11]), and a machine learning approach.

We analyse the results of these approaches in each dataset and between datasets. Preliminary results support the idea that FTO splicing events are cancer cell dependent with the largest number of variants detected in HEK293T cells. Crosstalk of regulation pathways involved in each cancer type will be realized next. Impact of FTO depletion on alternative splicing may also differ according to the depletion method (*i.e.* KO, SI or SH). This point should be of major interest for future splicing meta-analysis study by comparing one cell type with different depletion methods on FTO.

Acknowledgements

This work was generously supported by Ligue contre le Cancer, SIRIC Montpellier Cancer (INCa-DGOS-Inserm 6045) and Labex NumeV (GEM flagship project).

References

[1] Sun T, et al. The role of m6A RNA methylation in cancer. *Biomedicine & Pharmacotherapy*, (112):108613, 2019.
[2] Liu J, et al. Regulation of Gene Expression by N6-methyladenosine in Cancer. *Trends in Cell Biology*, (29/6):487-499, 2019.
[3] Dominissini D ,et al. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature*, (485/7397):201-206, 2012.
[4] Zhao X, et al. FTO-dependant demethylation of N6-methyladenosine regulates mRNA splicing and is required for adipogenesis. *Cell Research,* (24):1403-1419, 2014.
[5] Bartosovic M, et al. N6-methyladenosine demethylase FTO targets pre-mRNAs and regulates alternative splicing and 3'-end processing. *Nucleic Acids Research*, (45/19):11357, 2017.
[6] Mauer J, et al. Reversible methyaltion of m6Am in the 5'cap controls mRNA stability. *Nature*, (541/7637):371-375, 2017.
[7] Deng X, et al. Critical enzymatic functions of FTO in obesity and cancer. *Frontiers in Endocrinology*, (9):396, 2018.
[8] Zuang C, et al. N6-methyladenosine demethylase FTO suppresses clear cell renal cell carcinoma through a novel FTO-PGC-1α signalling axis. *J Cell Mol Med*, (23):2163-2173, 2019.
[9] Relier S, et al. FTO-mediated cytoplasmic m6Am demethylation adjusts stem-like properties in colorectal cancer cell. *Under review*.
[10] Huang Y, et al. Small-Molecule Targeting of Oncogenic FTO Demethylase in Acute Myeloid Leukemia. *Cancer Cell,* (35/4):677-691, 2019.
[11] Benoit-Pilven C, et al. Complementarity of assembly-first and mapping-first approaches for alternative splicing annotation and differential analysis from RNAseq data. *Scientific Reports*, (8):4307, 2018.

Poster 184

Epigenetics regulation of hematopoietic lineage plasticity and early programming of metabolic diseases

Alexandre PELLETIER[1], Fabien DELAHAYE[1] and Philippe FROGUEL[1]

[1] UMR 1283 EGID – CHRU de Lille, Faculté de Médecine HENRI-WAREMBOURG, Pôle Recherche : 1, place de Verdun, 59045 LILLE CEDEX – France

Corresponding Author: alexandre.pelletier.etu@univ-lille.fr

Perinatal exposure to environmental stress can have long-term impact, especially by increasing the risk to chronic diseases, like diabetes and obesity, in adults[1]. Defects in stem cells potency can explain the progressive development of such diseases, as seen during aging process [2]. Thus, the long-term alteration of stem cells capacities due to early exposure could explain the increase sensitivity to age-related and metabolic diseases later in life. Evidence for epigenetic remodelling after early exposure exists[3], but the underlying mechanisms and the long-term influences on stem cell fate decision need to be further elucidated.

This is why, the goal of my project is to characterize the impact of early exposure to over-nutrition on epigenetics mechanisms regulating hematopoietic lineage plasticity, thus offering a mechanism to link early exposure and increased sensitivity to age-related and metabolic diseases.

I propose first to identify hematopoietic stem and progenitor cells (HSPC) specific cis-regulatory elements (CRE) governing lineage commitment at cellular level by integrating scRNA-seq and scATAC-seq data. The data is being generated using the 10X platform on cord blood derived HSPCs from appropriately grown neonates (CTRL) and large for gestational age (LGA) neonates, as subjects to over-nutrition. Using scRNA-seq data and the Seurat workflow as reference, I perform normalization and clustering analysis to identify HSPC subpopulations. To adequately annotate the different cell population, an essential step to monitor hematopoietic plasticity, I propose 1) to generate an "hematopoietic reference map" that recapitulates the lineage distribution by integrating data generated from selected subset of hematopoietic populations (from early progenitor to differentiated cells) ; and 2) to optimize the standard Seurat pipeline by generating a composite score to identify markers, not only focusing on differential expression but also considering biological connectivity of the genes. Then, using scATAC-seq data, I will identify open chromatin region (OCR) at cellular level based on peak calling approach. Further data integration (e.g Signac workflow (Satijalab)) will allow me to assign cell population to specific OCR profiles, and then, correlate chromatin remodelling and RNA expression to identify lineage specific CREs.

The second aspect of my project is to assess the epigenetic influence of early exposure to over-nutrition on HSPC lineage decision. To do so, I am first looking at DNA methylation (data previously generated from bulk samples) changes between CTRL and LGA using a linear regression approach (limma) on the normalized data to find LGA-related CpG, and integrate these data with our previously identified CREs. I will also measure chromatin remodelling comparing scATAC-seq data from CTRL and LGA. These analyses will provide a comprehensive view of how early exposure may impact lineage decision through epigenetic mechanism.

Finally, I will validate the influence of these exposure associated CREs on HSPC plasticity using genome editing (CRISPR-Cas9) approaches.

Preliminary results, suggesting a lineage commitment bias towards B-cells progenitors in LGA, as observed in HSPCs from aged primate [4], thus supporting our hypothesis, will be further discuss in our poster as well as the promises and limitations of our computational approaches.

References

[1] D. J. P. Barker, « In utero programming of chronic disease », p. 14, 1998.

[2] N. S. Chandel, H. Jasper, T. T. Ho, et E. Passegué, « Metabolic regulation of stem cell function in tissue homeostasis and organismal ageing », *Nature Cell Biology*, vol. 18, nº 8, p. 823-832, août 2016, doi: 10.1038/ncb3385.

[3] S. J. van Dijk *et al.*, « DNA methylation in blood from neonatal screening cards and the association with BMI and insulin sensitivity in early childhood », *International Journal of Obesity*, vol. 42, nº 1, p. 28-35, janv. 2018, doi: 10.1038/ijo.2017.228.

[4] K.-R. Yu *et al.*, « The impact of aging on primate hematopoiesis as interrogated by clonal tracking », *Blood*, vol. 131, nº 11, p. 1195-1205, mars 2018, doi: 10.1182/blood-2017-08-802033.

Poster 185

RF4SV: A Random Forest approach for accurate deletion detection

Emira Cherif[1*], Roberto Xavier[2], Anna-Sophie Fiston-Lavier[1] and Ronnie C.O. Alves[2,3]
[1]ISEM, Université Montpellier, CNRS, UM, IRD, CIRAD, EPHE, Montpellier, France
[2]Federal University of Pará, R. Augusto Corrêa, 1, Belém, 66075-110, PA, Brazil
[3]Instituto Tecnológico Vale, R. Boaventura da Silva, 955, Belém, 66055-090, PA, Brazil
[*]Corresponding author

Efficiently detecting genomic structural variants (SVs) is a key step to grasp the “missing heritability” underlying complex traits involved in major evolutionary processes such as speciation, phenotypic plasticity, and adaptive responses. Yet, the SV-based genotype/trait association studies are still largely overlooked mainly due to the lack of reliable detection methods. Here, we present a random forest ensemble method for accurate deletion identification. We called this approach RF4SV. Several classic and ensemble learning strategies were carefully evaluated using proper benchmark data. To carry out the benchmark, the genome of the model species Drosophila melanogaster was chosen to detect large deletions, given that most SVs along this genome are deletions (8 962 out of 10 183). The RF4SV was thus trained and tested to detect specifically this type of SV. The model consisted of 12 features from the mapping from a BAM file generated by mapping the reads on a reference genomic sequence. We show that RF4SV outperforms established SV callers (DELLY, Pindel, etc) with higher overall performance (F1-score > 0.75; 6x-12x sequencing coverage) and is less affected by low sequencing coverage and deletion size variations. It is theoretically possible to "compile" a list of sequence patterns linked to a given type of SVs. Therefore, a model could learn to recognize them (distinct SVs patterns). Models that recognize a particular type of variation using DNA sequence patterns can then be combined to form a learning system able of detecting all types of SVs in a given genome, beyond the one used in our benchmark study.

Keywords: Structural variant, Deletion, Random Forest, NGS

Poster 187

ISSI : IRCAN's Satellite Signature Project

Joris ARGENTIN[1] and Olivier CROCE[1]
Institute for Research on Cancer and Aging, Nice (IRCAN), 28 Avenue de Valombrose, 06107, Nice, France

Corresponding author: jargentin@unice.fr

Tandems repeat sequences constitute an important part of the genomic content, especially in eucaryotes. Even though these repeats are not usually located in coding regions, these repetitive structure play major roles in cell biology. For example, telomeres are made by long repeats aim to protect from DNA shortening. Some tandem repeats located in the centromeres[1] are fundamental to cell division. Tandem repeat distribution varies a lot depending of organism species, age of cells through telomere shortening, or cell lineages[2]. Moreover, these repeats sequences can be involved in various diseases such as some cancers where extreme expansion and contraction phenomena have also been observed[3].

The IRCAN's Satellite Signature Project (ISSP) aims at exploring the links between these sources of variation and the distribution that can be observed in genome sequences. Sequence processing has a tendency to underrepresent repeats, hence the need to work on raw reads. We have developed a pipeline that processes and binds two sources of information : (i) we perform an exact-match repeats detection using the kmer-ssr[4] software from the reads contain in the raw FASTQ files ; (ii) we generate a file containing metadata such as information from to the sequencing method or every annotations related to the samples This pipeline outputs a formatted JSON file representing the sequences repeats ≪ signatures ≫ of each samples that can be re-used for inter samples comparisons or statistical analyses.

We have also developed a web interface using VueJS (called ≪ ISSI ≫) backed by a Django server to be usable by biologists without bioinformatics skills. It allows to visualise and manipulate data. A database stores the signatures processed by users that can be made private or freely available for other users. ISSP provides a single file, combining an exhaustive tandem repeat profile and biological metadata for a given sequencing run. This file can then be used by biologists to assess tandem repeat variation between individuals, populations or with a single individual given that two different sequences are provided. Tests on public data are currently being run to validate the method and set up descriptive, comparative and predictive statistical analyses.

Acknowledgements

This work was supported by JOBIM 2020.

References

[1] Simona Giunta and Hironori Funabiki. Integrity of the human centromere DNA repeats is protected by CENP-A, CENP-C, and CENP-T. *Proceedings of the National Academy of Sciences of the United States of America*, 114(8):1928–1933, February 2017.

[2] Guy-Franck Richard, Alix Kerrest, and Bernard Dujon. Comparative Genomics and Molecular Dynamics of DNA Repeats in Eukaryotes. *Microbiology and Molecular Biology Reviews : MMBR*, 72(4):686–727, December 2008.

[3] Sergei M. Mirkin. Expandable DNA repeats and human disease. *Nature*, 447:932–940, June 2007.

[4] ridgelab. Fast, Accurate, and Complete SSR Detection in Genomic Sequences: ridgelab/Kmer-SSR, April 2019. original-date: 2016-10-30T22:06:37Z.

Poster 188

Development of a Snakemake "framework/templates" to study the whole-genome transcriptional profiling from a blood-flesh trait (bf) and non blood-flesh trait (non-bf) cultivars in *Prunus persica*

Laure Heurtevin[1], Thierry Pascal[1], Bénédicte Quilot-Turion[1], Marie Laure Martin Magniette[2] and Jacques Lagnel [1]

[1]-INRAE, GAFL, 84140, Avignon, France, [2]-INRAE, IPS2, 91190, Gif-sur-Yvette, France

Introduction

Little is known about the mechanisms controlling anthocyanin biosynthesis in flesh of fruit. We explored the genetic pathways related to the elaboration of the blood-flesh trait in peach (*Prunus persica*). For this purpose, a comparative RNAseq study was carried out on flesh from a blood-flesh cultivar and a non blood-flesh cultivar at 4 fruit development stages, from 60 days after blooming up to fruit maturity. 40 libraries including biological replicates were sequenced by Illumina platform (Get-PlaGe) which generated 145Gbp (2.5Greads PE 150pb reads). The RNAseq pipeline was developed using Snakemake [1] and Singularity in order to ensure reproducibility and flexibility in the analysis, traceability of the samples, pipeline ease of use as well as facilitate the portability and the scalability to large data sets.
Here, we proposed a Snakemake "framework" based on a set of interoperable Snakemake rules as well as a set of templates (config, Slurm and samples sheet). This Snakemake framework/templates and Singularity recipes/images will be available on a public forge based on GitLab source code management software (https://forgemia.inra.fr/gafl).
Statistical analyses were performed by DiCoExpress (R workflow ML Martin-Magniette).

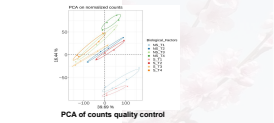
Biological Data and sequencing

Cultivars : 2 cultivars : blood-flesh trait and non blood-flesh trait
Samples : flesh from peach fruits
Reference genome: v2.0.1 (26873 genes)
Biological replicates : 2cultivarsx 4 stages x5 biological replicat
⇒ 5 biological replicates X 8 conditions (40 libraries)
⇒ Illumina paired ends 2x150pb sequencing

Sequencing Raw data summary

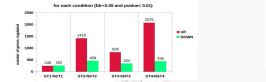
Total librairies	40
Average library reads count (Mreads)	30 +/- 5
Total Size (Gbp)	145
size of reads (bp)	2X150

Results



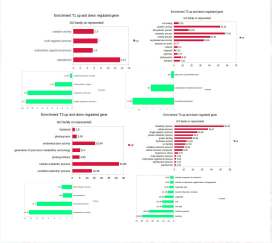
PCA of counts quality control

The PCA of counts shows that the biological replicates are well grouped together and separated between the different conditions. This quality control step allows us to validate our counts data in order to do the differential analysis.



Results of differential gene expression (DEG) profiles per time points

The histograms present the regulated genes level, up and down, between the blood-flesh fruits and non blood-flesh fruits in each time of development. We observe that despite restrictives variables fdr and pvalue, we have more than 2000 genes differentially expressed in our experimental conditions.



Enrichment results on 4 times

The enrichment analysis was carried out with AgriGO software (http://systemsbiology.cau.edu.cn/agriGOv2/)
We observed at time 2 and 3 an overrepresentation of many genes linked to photosynthesis, pigmentation and oxidoreduction activity in the blood-flesh cultivars.

RNAseq pipeline

samples.csv

SampleName	fq1	fq2
sample1	s1.R1.fq.gz	s1.R2.fq.gz
sample2	s2.R1.fq.gz	s2.R2.fq.gz
....	...	....

config.yaml
-Output path
-Ref path
-parameters
-containers path
- ....

cluster.json
Slurm Ressouces reservation for each rule

RNA-seq SnakeMake workflow
https://forgemia.inra.fr/gafl/pipelines_snakemake/rnaseq

Preproc (fastp)
Reads QC (fastqc/multiqc)
Mapping star 2 pass (star2)
bam QC (multiqc)
Build count matrix

Counts.csv

Target.csv
Annotation.csv
Enrichment.csv

R analysis using DiCoExpress pipeline
Quality control
Differential expression analysis (DEG)
CoExpression
GO Enrichment

Snakemake and Singularity framework

Snakemake
Snakemake module repository:
➢ 38 rules in 27 modules
https://forgemia.inra.fr/gafl/snakemake_modules
fastp, Utils, multiqc, fastqc, samtools, STAR, Features count, ...

Template files :
- Snakefile, config.yaml
- Samples list in csv
- cluster.json

Singularity
Singularity repository:
https://forgemia.inra.fr/gafl/singularity

Singularity repository:
➢ 80 containers recipes (versioned)
➢ gitlab CI/CD for images (.gitlab-ci.yml)
➢ Images accessible via OCI registry1

STAR v2.7.3, fastp v0.20, fastqc v0.11, multiqc v1.7, samtools v1.9, R_base v3.6.1, ..., DiCoExpress v2.0 R 3.6.1

Example STAR package:
https://forgemia.inra.fr/gafl/singularity/star
singularity pull star_v2.7.3.sif \
oras://registry.forgemia.inra.fr/gafl/singularity/star/star:latest

Conclusion

The proposed Snakemake "framework" and singularity repository facilitate 1) the pipeline construction using interoperable modules, 2) the bio-analyses by non bioinformatician 3) the scalability. The parallelisation is fully automated using Slurm. In order to run the pipeline, the user only needs to provide sample files (csv) and set few parameters (config.yaml). Furthermore, the use of Snakemake workflow manager and Singularity containers increases the bioanalysis reproducibility and facilitates the deployment across HPC platforms. The only requirement for the HPC platform is to provide Singularity (>3.3) and use Slurm as ressource manager.

References
1 Köster J and Rahmann S. Snakemake – a scalable bioinformatics workflow engine. Bioinformatics, 28:2520-2522, 2012.
2 Lambert I, Paysant-Le Roux C, Colella S, Martin-Magniette ML- DiCoExpress: a workspace to process multifactorial RNAseq experiments from quality controls to co-expression analysis through differential analysis based on contrasts inside GLM models. Plantmethods,10.1186/s13007-020-00611-7 https://forgemia.inra.fr/GNet/dicoexpress

Acknowledgements: We are grateful to the Genotoul Bioinformatics platform of Toulouse for providing help and computing and storage resources, to Sylvain Santoni for the librairies constructions and the GeT-PlaGe platform for the RNAseq experiments.

Poster 189

DECONbench: a benchmarking platform dedicated to deconvolution methods for tumor heterogeneity quantification

Clémentine DECAMPS[1*], Alexis ARNAUD[2*], Florent PETITPREZ[3], Mira AYADI[3], Aurélia BAURES[3], Lucile ARMENOULT[3], HADACA Consortium[‡], Rémy NICOLLE[3], Richard TOMASINI[4], Aurélien DE REYNIES[3], Jérome CROS[5], Yuna BLUM[3#+], Magali RICHARD[1#]

1 Laboratory TIMC-IMAG, UMR 5525, Univ. Grenoble Alpes, CNRS, Grenoble, France.
2 Data Institute, Univ. Grenoble Alpes, Grenoble, France.
3 Programme Cartes d'Identité des Tumeurs, Ligue Contre le Cancer, Paris, France.
4 INSERM U1068 CRCM, Marseille, France.
5 Beaujon Hospital, Dpt of Pathology - Univ. Paris-INSERM U1149; Clichy, France
‡ https://cancer-heterogeneity.github.io/data_challenges_HADACAconsortium.html
*Co-first #Co-last

+Presenting Author
Corresponding Authors: yuna.blum@ligue-cancer.net, magali.richard@univ-grenoble-alpes.fr

1. Background

Quantification of tumor heterogeneity is of utmost interest to the bioinformatics and biomedical research community, as it is related to tumor progression, clinical outcome and response to therapy. Advanced microdissection techniques to isolate a population of interest from heterogeneous clinical tissue samples are not feasible in daily practice. An alternative is to rely on computational deconvolution methods that infer cell-type composition. Bioinformatic tools to assess the different cell populations from bulk transcriptome [1] and methylome [2] samples have been recently developed, including reference-based and reference-free methods. However, their efficacy assessment has been impaired by the lack of dedicated benchmarking studies.

2. Results

Here we present DECONbench an innovative public digital benchmarking platform, open source, and freely available for the scientific community, aiming at comparing deconvolution methods for tumor heterogeneity quantification. DECONbench is hosted on the Codalab competition platform and is designed to execute methods developed in R environment, using a docker image. It includes both benchmarking datasets and computational methods to be evaluated. We have constructed benchmarking datasets composed of in silico simulated heterogeneous samples from transcriptome and methylome of primary cells isolated from pancreatic tumors. We recently used this unreleased datasets in a data challenge (https://tinyurl.com/hadaca2019). The best methods collectively discovered during the challenge are provided on DECONbench as a first set of reference benchmark methods. The benchmarking platform allows the submission of new methods. Performance scores for new methods and the set of reference methods are displayed on a leaderboard.

3. Conclusion

This platform is a unique opportunity to compare the performance of deconvolution methods between different omics data. It can be used to assess the performance of newly developed methods by applying them on high quality benchmark datasets in a user-friendly fashion.
The structure of DECONbench is open to evolution and extension are currently underway. Work is ongoing to generate new benchmark datasets that will be added to the platform.

References

1. Avila Cobos, F., et al. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics* 34, 1969–1979, 2018
2. HADACA consortium, Decamps, C., et al. Guidelines for cell-type heterogeneity quantification based on a comparative analysis of reference-free DNA methylation deconvolution software. *BMC Bioinformatics* 21, 16, 2020

Poster 190

A bioinformatics pipeline for urgent detection and characterization of emergent viral pathogens

Hourdel V, Kwasiborski A, Balière C, Matheus S, Manuguerra JC, Vanhomwegen J., Caro V

Institut Pasteur, Environment and Infectious Risks Unit, Laboratory for Urgent Response to Biological Threats

Global human health is challenged by emerging viral threats. Studies of these outbreaks have highlighted the impact that real-time sequencing of the viral genomes as part of a concerted outbreak response can have on public health, and called for genomic surveillance toolkits that can be deployed rapidly, especially in low-resource settings. The setup of more robust sequencing/bioinformatics pipelines with low turnaround times could provide crucial insights into the understanding of outbreaks through pathogen characterization and better monitoring of transmission, spread, and evolution.

The bioinformatics community has been very active to implement tools to perform NGS analysis, taking into account the nature of the data (error rate, slippage...). One of the big challenge still remains the development of rapid, easy-to-use software to analyse data and generate reports in real-time.

In this context, our laboratory has set up a workflow from clinical samples to data analysis, providing a real-time overview of genomic profile of the targeted pathogen and subsequent phylogenetic analysis using the consensus sequence. Illumina data were *de novo* assembled using SPADES and MAFFT tools. Phylogeographic analysis was performed with BEAST. ONT data were analyzed with RAMPART from the ARTIC network project. This application highlights insightful sequence data in few minutes and allows visualizing results along the genome.

In 2018, our workflow has been applied to investigate Rift Valley Fever outbreak in Mayotte. To decipher the origin of this emergence, we performed a phylogeographic analysis on the earliest samples. The study of the virus genomic epidemiology pointed to new introduction from Eastern African mainland.

The same approach has been implemented for the ongoing outbreak of coronavirus disease (COVID-19) caused by SARS-CoV-2. Indeed, as part of our 24/7 duty, our laboratory is involved in the French COVID-19 diagnosis and characterization of this pathogen. We were able to adapt our workflow to obtain full-length viral genomes with extremely rapid turnaround, directly from clinical samples.

Within the framework of our studies, we support that NGS combined with a robust bioinformatics pipeline is an essential tool to decipher the pathogen origin, and to track virus introduction and transmission networks, providing significant help to health authorities.

Poster 192

A rare variant burden test based on population frequencies for case-only rare disease study designs

Antoine Favier[1], Stefania Chounta[1] and Antonio Rausell[1]

[1] Université de Paris, Imagine Institute, Laboratory of Clinical Bioinformatics, INSERM UMR 1163, 24 boulevard du Montparnasse, F-75015, Paris, France

Corresponding Author: antonio.rausell@institutimagine.org

The genetic basis of approximately 50% of more than 4000 rare Mendelian disorders described to date remains uncharacterised. The identification of the underlying causal genetic variants from the genomes/exomes of patients is challenged by a large genetic heterogeneity. Low cohort sizes together with a large number of rare variants compromise the statistical power to associate a genotype with a phenotype. Moreover, incomplete penetrance and technical sequencing issues can further complicate the task [1]. In order to increase statistical power, rare variant burden tests have been proposed. In these frameworks, the aggregation of variants in specific genomic regions (*e.g.* genes or sets of genes) is evaluated either in case-control designs or in family-based studies [2]. However, case-only study designs, characterized by the absence of sequencing data from a matched control cohort or from relative individuals, represent a common scenario in the study of rare diseases for which currently available burden tests are not applicable. To overcome such limitation, we implemented a novel burden statistical test for rare variants analysis in case-only study designs using Whole Exome/Genome Sequencing. The test relies on parametric modeling of rare variants counts through binomial and Poisson distributions. It evaluates the null hypothesis H_0 stating that the number of rare variants within a given gene or set of genes observed in a cohort of patients originates from a random model of sequence neutral variation [3]. To calibrate random expectations, we took advantage of recent large-scale sequencing projects on the general population such as the Genome Aggregation Database (gnomAD) [4]. However, inferences done on such estimates are subject to both technical (*e.g.* sequencing technology, platforms and bioinformatics pipelines) and genomic confounding factors (segmental duplications, low complexity regions, highly polymorphic regions, and poorly resolved reference genome intervals). Such factors are susceptible to lead to varying sequencing quality, coverage and variant calling rates across genomic regions, and may translate in inaccurate random expectations of per-gene neutral sequence variation [5]. These errors may ultimately produce violations of model assumptions and/or an inflation of type I errors, *i.e.* falsely rejecting the previously defined null hypothesis H_0. To experimentally validate our framework we applied it to the assessment of per-gene burden missense and synonymous variants in 503 European individuals from the 1000 Genomes Project. We first validated our statistical modeling framework through evaluation of the per-gene mean-to-variance assumptions, goodness-of-fit and p-value distribution. Second, we evaluated the test's susceptibility to type I errors. To that aim we focused on rare synonymous variant analysis under the assumption that positive hits may be confidently considered as false rejections for most of the cases [6]. We thus created 1000 random sets each subsampling half of the 503 European individuals and assessed for each of them (i) our case-only per-gene burden test, and (ii) a prototypical case/control-like per-gene burden test against the remaining 503 individuals, for the sake of a comparative benchmark. Through such assessment, we identified recurrent sources of technical and genomic confounding factors leading to false positive signals and provide guidelines to control for them. Finally, we characterized the limits of statistical power of our test as a function of gene length and cohort size. The case-only statistical test is implemented in python and R as is available on the Github page of the lab.

References

[1] Kohane, I. S., Hsing, M., & Kong, S. W. (2012). Taxonomizing, sizing, and overcoming the incidentalome. Genetics in Medicine, 14(4), 399-404.

[2] Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. Am. J. Hum. Genet. 95, 5–23 (2014).

[3] Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., ... & Wall, D. P. (2014). A framework for the interpretation of de novo mutation in human disease. *Nature genetics*, *46*(9), 944.

[4] Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., ... & Gauthier, L. D. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv*, 531210.

[5] Isakov, O., Perrone, M., & Shomron, N. (2013). Exome sequencing analysis: a guide to disease variant detection. In *Deep Sequencing Data Analysis* (pp. 137-158). Humana Press, Totowa, NJ.

[6] Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., ... & Tukiainen, T. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285-291.

Poster 197

BiokoP: Bi-objective programming for RNA secondary structure prediction with pseudoknots

Audrey Legendre, Mandy Ibene, Eric Angel, Fariza Tahi

IBISC, Univ Evry, Université Paris-Saclay, 91000 Evry, France

Corresponding Author: fariza.tahi@univ-evry.fr

RNA structure prediction is an important field in bioinformatics, and numerous methods and tools have been proposed. Pseudoknots are specific motifs of RNA secondary structures that are difficult to predict. Almost all existing methods are based on a single model and return one solution, often missing the real structure. An alternative approach would be to combine different models and return a (small) set of solutions, maximizing its quality and diversity in order to increase the probability that it contains the real structure.

We propose an original method for predicting RNA secondary structures with pseudoknots, based on integer programming. We developed a generic bi-objective integer programming algorithm allowing to return optimal and sub-optimal solutions optimizing simultaneously two models. This algorithm was then applied to the combination of two known models of RNA secondary structure prediction, namely MEA (Maximum Expected Accuracy) and MFE. (Minimum Free Energy) The resulting tool, called BiokoP, is compared with the other methods in the literature. The results show that the best solution (structure with the highest F_1-score) is, in most cases, given by BiokoP. Moreover, the results of BiokoP are homogeneous, regardless of the pseudoknot type or the presence or not of pseudoknots. Indeed, the F_1-scores are always higher than 70% for any number of solutions returned.

The results obtained by BiokoP show that combining the MEA and the MFE models, as well as returning several optimal and several sub-optimal solutions, allow to improve the prediction of secondary structures. One perspective of our work is to combine better mono-criterion models, in particular to combine a model based on the comparative approach with the MEA and the MFE models. This leads to develop in the future a new multi-objective algorithm to combine more than two models.

BiokoP is available as a web server on the EvryRNA platform: https://EvryRNA.ibisc.univ-evry.fr .

Poster 198

International Society for Computational Biology Student Council Regional Student Group France (RSG France) : Association of Young Bioinformaticians of France (JeBiF)

Florence Jornod[1], Slim El Khiari[1] , Xavier Bussell[1], Julien Fumey[1], Mathias Galati[1], Athénaïs Vaginay[1] and Victor Grentzinger[1]

[1] Association des Jeunes Bioinformaticiens de France RSG France - JeBiF, 4 rue des Arènes, 75005, Paris, France

Corresponding Author: contact@jebif.fr

1. Abstract

The association of Young Bioinformaticians of France (RSG France - JeBiF) is the french regional group of the International Society for Computational Biology Student Council. Its main goal is to help building the community of young bioinformaticians in France. With this poster we will present the different activities that RSG France develop to reach its goal.

Acknowledgements

RSG France's activities are funded among others by the GDR BIM and the International Society for Computational Biology.

Poster 199

BiokoP: Bi-objective programming for RNA secondary structure prediction with pseudoknots

Ludovic Platon, Farida Zehraoui, Fariza Tahi

IBISC, Univ Evry, Université Paris-Saclay, 91000 Evry, France

Corresponding Author: fariza.tahi@univ-evry.fr

RNA structure prediction is an important field in bioinformatics, and numerous methods and tools have been on-coding RNAs (ncRNAs) play important roles in many biological processes and are involved in many diseases. Their identification is an important task, and many tools exist in the literature for this purpose. However, almost all of them are focused on the discrimination of coding and ncRNAs without giving more biological insight. In this paper, we propose a new reliable method called IRSOM, based on a supervised Self-Organizing Map (SOM) with a rejection option, that overcomes these limitations. The rejection option in IRSOM improves the accuracy of the method and also allows identifing the ambiguous transcripts. Furthermore, with the visualization of the SOM, we analyze the rejected predictions and highlight the ambiguity of the transcripts.

IRSOM was tested on datasets of several species from different reigns, and shown better results compared to state-of-art. The accuracy of IRSOM is always greater than 0.95 for all the species with an average specificity of 0.98 and an average sensitivity of 0.99. Besides, IRSOM is fast (it takes around 254 s to analyze a dataset of 147 000 transcripts) and is able to handle very large datasets.

IRSOM is available on our software platform EvryRNA (http://EvryRNA.ibisc.univ-evry.fr).

Poster 200

How to explain bioinformatics to non scientists? Feedback from the experience of RSG France - JeBiF

Slim El Khiari[1], Xavier Bussell[1], Stéphanie Chevalier[1], Julien Fumey[1], Fabien Genty[1], Victor Grentzinger[1], Julie Lao[1], Marylène Rugard[1], Athénaïs Vaginay[1], Amaury Vaysse[1] and Florence Jornod[1]

[1] Association des Jeunes Bioinformaticiens de France, RSG France - JeBiF, 4 rue des Arènes, 75005, Paris, France

Corresponding Author: contact@jebif.fr

1. Abstract

In 2016, we came to the conclusion that bioinformatics was largely unknown to the general public even though it is becoming more and more important for research in biology. Applications of bioinformatics are encountered more and more frequently in particular in health practice. Therefore we think that explaining our science to the public is a very important work.

In light of this observation, we initiated from scratch a completely new activity for the association consisting of popularizing bioinformatics to the general audience. The main goal of this activity is to make bioinformatics known and comprehensible for the public.

On this poster we will present different type of popularizing methods that we implemented since 2016: video contests, participation in science fairs (during the "Fête de la Science"), school workshops, partnership with Pint of Science.

Acknowledgements

This work benefits the support from the International Society for Computational Biology Student Council (ISCB-SC), BIOASTER, Cité des Sciences, École Polytechnique, Sorbonne Université, Île de Sciences and the Boullay-les-Troux city.

We also would like to thank all the volunteers that participated in the creation and animation of this activity.

Poster 202

Efficiency of gene regulatory network inference methods on genomic and transcriptomic

Lise Pomiès[1], Celine Brouard[1], Harold Duruflé[2], Nicolas Langlade[2] and Simon de Givry[1]
[1] MIAT, Université de Toulouse, INRAE, Castanet-Tolosan, France
[2] LIPM, Université de Toulouse, INRAE, CNRS, Castanet-Tolosan, France

Corresponding author: lise.pomies@inrae.fr

Network inference methods are powerful tools to study complex biological processes. However, a lot of inference algorithms exist, and it could be difficult to identify an algorithm adapted to a specific experimental dataset. We studied the resistance of sunflower to drought and heterosis. We collected transcriptomic and genomic data. We obtained the SNP measurements from 350 sunflower genotypes and the expression measurements of 173 genes (mostly transcription factors) on these genotypes. To evaluate the behaviour of different inference algorithms, we constructed 100 artificial datasets with biological properties close to the properties of our real dataset measured on sunflowers.

Five inference methods were tested (i) Lasso, (ii) Random Forest, (iii) Bayesian Network, (iv) Ordinary Least Square and (v) Pairwise Exponential Markov Random Field. Each method produced a list of ranked edges (from the most probable to the less). A meta-analysis was performed on the results of the five inferred methods.

We decided to keep only the top 75 edges to infer graphs with at least 80% of correctly predicted regulations between genes. As expected the meta-analysis provided better results than each method alone. However, the model has difficulties to predict the highly connected parts of biological networks. Graph structures as triangles and mutual edges often contain falsely predicted edges. Other errors concern edges predicted in the wrong orientation or missing an intermediate regulator gene.

Acknowledgements

This work was supported by the "SUNRISE" project of the French National Research Agency (ANR-11-BTBR-0005, 2012-2020).

References

Poster 203

Epigenetic Regulations of Inflammation in Cardiovascular Complications of Type 2 Diabetes

Diane GIRAD[1], Elena CARCARINO[1], Jean-Baptiste JULLA[1], Louis POTIER[1], Jean-Francois GAUTHIER[1], Ronan ROUSSEL[1], Nicolas VENTECLEF[1], and Claire VANDIEDONCK[1]

[1] Université de Paris, INSERM UMRS 1138, Centre de Recherche des Cordeliers, F-75006 Paris, France

Corresponding Authors: nicolas.venteclef@inserm.fr; claire.vandiedonck@inserm.fr

At the genome scale, non-coding DNA sequences, called enhancers, orchestrate the differentiation and specialization of cells by regulating the expression of genes implicated in these functions. These enhancers have a major role in driving the intensity of the inflammatory response through the modulation of the expression of genes controlling the activation of monocytes and macrophages [1].

Chronic inflammation due to an aberrant and uncontrolled activation of circulating monocytes is an important mechanism in the development of cardiovascular (CV) complications in type 2 diabetic (T2D) patients, such as stroke and myocardial infarction [2,3]. We speculate that T2D patients with extreme CV risks carry genetic variants modulating enhancer activity thus influencing monocyte activation. This could drive an uncontrolled inflammation and accelerate the development of CV complications in patients with high CV risk compared to those with low CV risk [1,4,5].

The main objective of our study will be to map the enhancers of circulating monocytes in a T2D cohort by both identifying the open-chromatin regions using ATACseq and by evaluating their activation state through ChIPseq of the H3K27ac mark [4]. Although enhancers in human circulating monocytes have already been mapped, the novelty of our study resides in the fact that we are working in a T2D context which is important when fine-mapping regulatory elements [8]. Our cohort is composed of a couple hundred T2D patients presenting various CV risks that were estimated based on their coronary artery calcium score, allowing us to stratify and compare the epigenetic pattern of low versus high CV risk T2D patients. The transcriptomic profile of the monocytes is also being studied using RNAseq to determine the impact of enhancer region activity changes on gene expression. Finally the genetic variants that could be contributing to these variations of enhancer activity will be identified through genetic association after performing a pangenomic genotyping of the patients [5,6,7,8]. By combining these multi-omics results, we hopefully will be able to determine an epigenetic signature driving the uncontrolled activation of circulating monocytes.

Our preliminary results on the transcriptomic profile of the circulating monocytes of 47 T2D patients have revealed a modulation of the expression of mitochondrial and inflammatory genes correlated to the increase of CV risk. These genes are implicated in oxidative stress and mitochondrial activity regulation, suggesting a deregulation of the metabolic activity that would drive the hyper-activation of monocytes.

References

1. Casey E Romanoski. Exploiting Genomics and Natural Genetic Variation to Decode Macrophage Enhancers. *Trends Immunol*, 36 (9), 507-18, 2015.
2. Josgua A Beckman. Diabetes and Atherosclerosis: Epidemiology, Pathophysiology, and Management. *JAMA*, 287 (19), 2570-81, 2002.
3. Josephine M Forbes. Mechanisms of Diabetic Complications. *Physiol Rev*, 93 (1), 137-88, 2013.
4. Kyle J Gaulton. Genetic Fine Mapping and Genomic Annotation Defines Causal Mechanisms at Type 2 Diabetes Susceptibility Loci. *Net Genet*, 47 (12), 1415-21, 2015
5. Rongrong Fan. Loss of the Co-Repressor GPS2 Sensitisezes Macrophage Activation Upon Metabolic Stress Induced by Obesity and Type 2 Diabetes. *Nat Med*, 22 (7), 780-91, 2016.
6. Weiping Sun. Integrative Analysis of Super Enhancer SNPs for type 2 Diabetes. *PLoS One*, 13 (1), e0192105, 2018.
7. Peng Gao. Risk Variants Disrupting Enhancers of Th1 and Treg Cells in Type1 Diabetes. *Proc Natl Acad Sci USA*, 116 (15), 7581-7590, 2019.
8. Claire Vandiedonck. Genetic association of molecular traits: A help to identify causative variants in complex diseases. *Clin Genet*. 93 (3), 520-32, 2018.

Poster 204

A method to classify α-satellites and other short sequences

Thomas Haschka, Loïc Ponger, Julien Mozziconacci and Christophe Escudé
Structure et instabilité des genomes - Sorbonne Universités - MNHN - CNRS 7196 / INSERM U1154 | ARChE
Muséum National d'Histoire Naturelle, 43 rue Cuvier, Paris 75005

Corresponding author: thomas.haschka@mnhn.fr

Abstract

In order tackle the classification of short sequenced datasets into families and subfamilies, we have created a in house toolset based around an adaptive method of the DBSCAN algorithm [1]. This method allows us to elucidate the evolutionary hierarchy of tandem repeats, or satellites which are abundant in all eukariotic genomes. Herein we present α-satellite families in particular.

Using this novel technique we shed light into proposed evolutionary mechanisms that govern the emergence of α-satellite sequence families, reviewed in [2]. α-Satellite sequences undergo rapid evolution and proposed models suggest successive waves of sequence amplification followed by mutation of amplified sequences. In the continuation of work previously published by our group [3] [4], we sequenced four cercopithecidae families using next generation, targeted, sequencing techniques. These four families contain short tandem repeats of alpha satellite sequences on the magnitude of up to 1.6 million sequences of approximately 172 bases pairs (bp) each.

Besides highlighting the results, and analyze the classification of α-satellites and the differences found in our in house cercopithecidae dataset between the four sequenced families. We undermine the effectiveness of our method by outlining how we correctly classify preaviously annotated α-satellite sequences on the human genome (hg38) dataset [5]. We further show that we are able to subclassify the previously annotated sequences. We outline how we find peculiar fine structures in this dataset by highlighting trees describing the phylogeny of α-satellite sequences, in a single individual, built from the hierarchy found by our tools.

In our poster/demo we will present the developed tools, show interesting applications on α-satellites found in the human genome hg38 and discuss how other groups can profit from our work in classifying sequences using our tools/methods.

Acknowledgements

We are grateful for the Romeo High Performance Computing Cluster of the Université de Reims Champagne Ardenne to have supported this project by granting us compute time. This work was supported by JOBIM 2020.

References

[1] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

[2] Shannon M. McNulty and Beth A. Sullivan. Alpha satellite dna biology: finding function in the recesses of the genome. *Chromosome Research*, 26(3):115–138, Sep 2018.

[3] Lauriane Cacheux, Loïc Ponger, Michèle Gerbault-Seureau, Florence Anne Richard, and Christophe Escudé. Diversity and distribution of alpha satellite DNA in the genome of an Old World monkey: Cercopithecus solatus. *BMC Genomics*, 17(916), 11 2016.

[4] Lauriane Cacheux, Loïc Ponger, Michèle Gerbault-Seureau, François Loll, Delphine Gey, Florence Anne Richard, and Christophe Escudé. The Targeted Sequencing of Alpha Satellite DNA in Cercopithecus pogonias Provides New Insight Into the Diversity and Dynamics of Centromeric Repeats in Old World Monkeys. *Genome Biology and Evolution*, 10(7):1837–1851, 06 2018.

[5] L.I. Uralsky, V.A. Shepelev, A.A. Alexandrov, Y.B. Yurov, E.I. Rogaev, and I.A. Alexandrov. Classification and monomer-by-monomer annotation dataset of suprachromosomal family 1 alpha satellite higher-order repeats in hg38 human genome assembly. *Data in Brief*, 24:103708, 2019.

Poster 205

Identifying novel sequence elements involved in transcription initiation in human genome

Yulia Rodina[1], Christophe Menichelli[1], FANTOM consortium, Sophie Lèbre[2,3], Laurent Bréhélin[1] and Charles-Henri Lecellier[1,4]

[1] LIRMM, Univ. Montpellier, CNRS, Montpellier, France
[2] IMAG, Univ. Montpellier, CNRS, Montpellier, France,
[3] Univ. Paul-Valéry-Montpellier 3, Montpellier, France
[4] Institut de Génétique Moléculaire de Montpellier, University of Montpellier, CNRS, Montpellier, France

Corresponding author: yulia.rodina@lirmm.fr

Using the cap analysis of gene expression (CAGE) technology, the FANTOM consortium built an atlas of transcription start sites (TSSs) in several species [1]. The CAGE data represents frequencies of transcription initiation across multiple samples (primary cells, tissues and cell lines) along the genome [2]. CAGE signal peaks were identified, and only part of them could be assigned to promoters or enhancers of known genes, while most of them initiate at unconventional regions.

In the presented work the CAGE human data was analysed at the genome level. It was shown that in many identified CAGE peaks transcription does not initiate always at a precise nucleotide position, but on the contrary, their transcription initiation profiles have a pronounced shape. The aim of the analysis was to predict the shape of the transcription initiation signal based on the sequence composition of human reference genome in order to identify novel elements involved in transcription initiation. At a first step, in order to explore transcription initiation architectures, we performed partitioning of CAGE peak shapes using the k-means clustering approach. The second step was to build a supervised classifier to distinguish the different shapes on the basis of the sequence. For this we explored sequence features, i.e. the composition of k-mers in the regions proximal to CAGE peaks and identified those that have impact on the transcription initiation architecture. These selected features were used to build a statistical model using machine learning techniques, which is used afterwards to predict the peak shapes of test data.

References

[1] A. R. Forrest et al. A promoter-level mammalian expression atlas. *Nature*, (507):462–470, 2014.

[2] M. Kanamori-Katayama et al. Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Res.*, (21):1150–1159, 2011.

Poster 207

Investigating DNA methylation in IDH mutant acute myeloid leukemia

Alexis Hucteau[1], Lucille Stuani[1], Ting Xie[1], Tony Kaoma[2], Jean-Emmanuel Sarry[1] and Vera Pancaldi[1]

[1] Cancer Research Center of Toulouse, 2 avenue Hubert Curien, 31037, Toulouse, France
[2] Luxembourg Institute of Health, 1A-B, rue Thomas Edison, L-1445 Strassen, Luxembourg

Corresponding author: vera.pancaldi@inserm.fr

1 Introduction

Acute Myeloid Leukemia is a blood cancer characterized by a blockage in myeloid differentiation and hyperproliferation of transformed myeloid progenitor cells. The mutation in the gene Isocitrate dehydrogenase 1 (IDH1) is implicated in this cancer as cells with the alteration abnormally produce an oncometabolite 2-hydroxyglutarate (2-HG). 2-HG has been found to inhibit enzymatic function of many alpha-ketoglutarate-dependent dioxygenases, including histone and DNA demethylases, causing widespread changes in histone and DNA methylation. These epigenetic changes lead to enhanced mitochondrial oxidative phosphorylation (OxPHOS) and fatty acid-oxidation (FAO). OxPHOS is the metabolic pathway in which cells utilize and oxidize nutrients to produce energy (adenosine triphosphate, ATP) and consume oxygen.The fatty acid-oxidation is the catabolic process by which fatty acid molecules are broken down in the mitochondria to generate acetyl-CoA, which enters the citric acid cycle, and NADH and FADH2 as major electron donors in OxPHOS process. IDH mutant AML cells have a higher activity in their mitochondria and accordingly, are more sensitive to inhibitors of mutated IDH1. This restored low levels of 2-HG and globally reversed the high methylated state of DNA and histones. Surprisingly and unexpectedly, cells maintain high OxPHOS and high levels of FAO upon IDH inhibitors, suggesting an adaptive mechanism [1]. The objective of this study is to find the compensatory pathway which decouples epigenomic and metabolic effects of the inhibitor.

2 Material and Methods

We analysed DNA methylation data (450k Illumina arrays) on an AML cell line (HL60) over-expressing WT or mutant R132H IDH1 with or without IDH inhibitor (AG-120) treatment. The data were preprocessed and 3 different types of normalization were applied. The R package ChAMP was used to perform differential methylation analysis, generating differentially methylated positions and regions (DMP/DMR) in DNA between cell lines with wild type or mutant IDH1 upon inhibitor. We looked for differential methylation on gene promoters and also on genomic regions that might be connected in 3D to gene promoters (through Promoter Capture HiC data collected across 17 human haematopoietic cell types as available on the GARDEN-NET webtool [2]. We investigated DMPs/DMRs for enrichment in biological pathways that are connected to metabolism metabolism.3. We have further integrated this methylation analysis with RNAseq data collected for the same cell line in similar experimental conditions.

3 Results

Preliminary results indicate the presence of methylation changes in specific pathways related to metabolism, development and pluripotency and sets of genes controlled by specific microRNAs.

4 Citations

References

[1] Lucille Stuani, Tony Kaoma, Jean-Emmanuel Sarry, and al. Combinatory therapy targeting mitochondrial oxidative phosphorylation improves efficacy of idh mutant inhibitors in acute myeloid leukemia. *bioRxiv*, 2019.

[2] Miguel Madrid-Mencía, Tran Bich Ngoc Raineri, Emanuele Cao, and Vera Pancaldi. Using garden-net and chaser to explore human haematopoietic 3d chromatin interaction networks. *Nucleic Acids Research (In press)*, 2020.

Poster 209

A repertoire of *Escherichia coli* K12 regulatory binding motifs

Claire Rioualen[1], Jacques van Helden[2,3], Julio Collado Vides[1,4]

1. Programa de Genómica Computacional, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, México.
2. Aix-Marseille Univ, INSERM, lab. Theory and Approaches of Genome Complexity (TAGC), Marseille, France.
3. Institut Français de Bioinformatique, UMS 3601-CNRS, Université Paris-Saclay, Orsay, 91403, France.
4. Department of Biomedical Engineering, Boston University, Boston, MA, USA.

Abstract

Regulation of gene expression is a fundamental process in all living organisms. One of the main mechanisms to achieve it is the binding of specific proteins, called transcription factors (TFs), to specific sites of the DNA. In order to identify those sites at a genome scale, ChIP-seq technology (Johnson et al., 2007; Robertson et al., 2007) can be used.

From an early time, *Escherichia coli* K12 has been a model organism of reference for the study of gene regulation (Jacob and Monod, 1961). Extensive information about TFs, their binding sites, target genes and operons has been manually curated and indexed for decades in dedicated databases such as RegulonDB (Santos-Zavaleta et al., 2019) and EcoCyc (Keseler et al. 2017), and its genome was one of the first to be entirely sequenced and published (Blattner et al, 1997). However, a lot of information remains unknown. In particular, about 70% of its 300 TFs still have unknown binding motifs. Surprisingly, the available ChIP-seq data targeting *E.coli*'s TFs is still rather scarce.

We propose to take advantage of RegulonDB's carefully curated information (Santos-Zavaleta et al., 2019) as well as new ChIP-seq data targeting *E.coli*'s 300 TFs (Gakagan et al., not published) in order to improve and complete the collection of regulatory binding motifs in *E. coli*, and explore the diversity and distribution of binding sites.

We show that ChIP-seq reveals a lot of previously unknown binding sites genome-wide, and can help us build a reliable collection of regulatory motifs. We also show that binding can occur outside of the typical upstream non-coding regions, challenging the canonical model of transcriptional regulation. Finally, we show that a few TFs seem to have various binding profiles.

References

Blattner, F. R. et al. (1997). The complete genome sequence of Escherichia coli K-12. Science (New York, N.Y.), 277(5331), 1453–1462.

Jacob, F., & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. Journal of Molecular Biology, 3, 318–356.

Johnson, D. S. et al. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science (New York, N.Y.), 316(5830), 1497–1502.

Keseler et al. (2017). EcoCyc: reflecting new knowledge about Escherichia coli K-12. Nucleic Acids Research 45:D543-50.

Robertson, G. et al. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nature Methods, 4(8), 651–657.

Santos-Zavaleta, A. et al. (2019). RegulonDB v 10.5: Tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. *Nucleic Acids Research*, *47* (Database issue), D212–D220. https://doi.org/10.1093/nar/gky1077

Poster 210

Investigating the taxonomic and functional diversity of benthic archaeal communities from two South Pacific hadal trenches using high-throughput sequencing data

Blandine Trouche[1], Feriel Bouderka[1], Clemens Schauberger[3], Julie Poulain[2], Patrick Wincker[2], Ronnie N Glud[3], Sophie Arnaud-Haond[4] and Loïs Maignien[1]

1 Laboratoire de Microbiologie des Environnements Extrêmes, Univ Brest/CNRS/IFREMER, F-29280, Plouzané, France

2 CEA–Institut de Génomique, GENOSCOPE, Evry, France

3 University of Southern Denmark, Department of Biology, Odense, Denmark

4 MARBEC, Ifremer/Univ. Montpellier/CNRS/IRD, Sète, France

Corresponding Author: Blandine.Trouche@univ-brest.fr

Despite recent advances in hadal research, hadal benthic microbial communities remain fairly understudied. They are however of particular interest due to their exceptional water depth, original trench ecosystems settings, and their situation at interface between hadopelagic waters and the deep biosphere. They are thus expected to harbor relatively original taxonomic and functional microbial diversity.

We explored benthic communities from Kermadec and Atacama Trenches in the South Pacific, focusing particularly on novel and evolutively interesting lineages such as Asgardarchaeota and DPANN superphylum members. To this end, we collected sediment cores about 30 cm deep from 11 locations in the two trench systems and adjacent abyssal plains and generated 56 metagenomic libraries.

Based on k-mer profiling, we found that sediment horizons are the main drivers structuring genomic sequences from these deep sea environments, followed by water depth (hadal vs. abyssal), with trench origin showing little influence despite their geographic separation. Based on 16S fragments (miTAGs) extracted from these metagenomes, we found a high proportion of Thaumarchaeota, as previously described in other trench ecosystems. We observed gradual importance of Nanoarchaeota (Woesearchaeales) with depth in the sediments. We could reconstruct 121 draft archaeal genomes, most of them affiliated to the order Nitrosopumilales, though partial genomes belonging to Nanoarchaeota and Asgardarchaeota were also recovered. In this study, we explore their metabolic capabilities and highlight trench-specific genes and pathways by comparing these reconstructed genomes to reference genomes recovered from environments such as hadopelagic waters and shallow sediments.

Poster 211

SNPer, a home made interrogation tool for variants and their effect throughout the genome

Frédérique Bitton[1] and Mathilde Causse[1]

[1] UGAFL 67 allée des chênes CS 60094 84143 MONTFAVET Cedex France

Corresponding Author: frederique.bitton@inrae.fr

With increasing number of resequencing projects and their publication, variants are easy to obtain from resequencing data, but it is not always easy to interrogate a list of variants with custom criteria like genes or intervals. I developped a tool with a web interface to find variants from a resequencing project, allowing the user to choose the kind of mutation (SNP or indel), the effect of the variant in genes (SnpEff [1] or ANNOVAR[2]), and in which sequenced individuals to search.
The collection of variants can be interrogated according to 3 different searches :

i) by a list of gene names ,

2) by interval on a chromosome,

3) according to a QTL results.

This tool is used at the GAFL for tomato and prunus variants.

References

[1] Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. Fly (Austin). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3., 2012 Apr-Jun;6(2):80-92. PMID: 22728672

[2] Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data Nucleic Acids Research, 38:e164, 2010

Poster 213

FAIR_bioinfo : a software perspective of reproducibility of bioinformatics analyses

Thomas Denecker[1], Céline Hernandez[2] and Claire Toffano-Nioche[3]

[1] Fungal Epigenomics and Development
[2] Next Generation Sequencing core facility
[3] RNA Sequence,S tructure, and Function
Institute for Integrative Biology of the Cell (I2BC)
Université Paris-Saclay, CEA, CNRS, 91198, Gif-sur-Yvette, France

Corresponding Author: claire.toffano-nioche@u-psud.fr

1 Reproducibility context

Recent studies have established that a reproducibility crisis challenges robustness of scientific results [1] including computational biology. In this context FAIR data principles are increasingly being used to ensure data integrity [2]. To complement this principles, we introduce FAIR_bioinfo, to apply similar guideline to code development and ensure reproducibility of results obtained from the same data set over time.

Computer tools do exist that can be applied in bioinformatics [3]. Convinced of their usefulness we propose an initiation to a range of software to make a complete bioinformatics analysis reproducible.

2 The FAIR_bioinfo Training

This training is based on a concrete example of classical data analysis, a differential gene expression between two RNAseq conditions. It deals with two of the know-how of the bioinformatician's job: the automated processing of raw data (through virtualization and pipeline development) and the analysis of processed data (with environment management and "notebooks" writing).

Improvement in reproducibility can be achieved in several steps, where each step brings an additional degree of reproducibility through a specific family of software tools.

We offer training for people from different backgrounds: on the one hand, over a long period of time, for learners who do not necessarily know programming (8 months, at a rate of 1h30 per month) so that they can replay the presentations at their own pace [4], and on the other hand, for bioinformaticians, in a format condensed into 2 days. In both solutions, the training is focused on general concepts with practical illustrations.

As the importance of reproducibility is no longer to be proven, the main interests of this training is to provide practical guidelines for its daily implementation with the long-term objective for everyone to gradually adopt good practices to overcome the challenge of reproducibility in science.

Acknowledgements

This work was supported by the French Infrastructure Institut Français de Bioinformatique (IFB) ANR-11-INBS-0013

References

[1] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452-4, 2016
[2] Mark D Wilkinson, Michel Dumontier, Ijsbrand Joan Aalbersberg, Gabirelle Appleton, *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, 3:160018, 2016
[3] Björn Grüning, John Chilton, Johannes Köster, Ryan Dale, *et al*. Practical Computational Reproducibility in the Life Sciences. *Cell Syst.* 6(6):631-63, 2018
[4] Thomas Denecker, Claire Toffano-Nioche, https://github.com/thomasdenecker/FAIR_Bioinfo, 2019

Poster 214

Feature selection and automatic machine learning identifies translationally regulated mRNAs by the mTOR pathway.

Costas Bouyioukos[1], Manuel Bulfoni[1], and Bertrand Cosson[1]

[1] Université de Paris, Epigenetics and Cell Fate CNRS UMR 7216, 35 rue Hélène Brion, F-75205, Paris, France

Corresponding Author: bertrand.cosson@u-paris.fr

1 Introduction

Pancreatic islet β-cells play a key role in maintenance of normoglycemia by synthesizing, storing and secreting insulin. In a previous work [1] we have exposed human pancreatic β-cells exhibiting glucose insulin secretion [2] to high glucose concentrations for 30 min. We observed a global protein synthesis increase, independent from transcription, through the activation of the mTORC1 pathway and the regulation of eIF2alpha phosphorylation. We identified 525 mRNAs that are differentially translated upon glucose stimulation by polysome profiling combined with mRNA sequencing and we describe their sequence specificities. Here we extract sequence and structural characteristic features from this set of differentially translated mRNAs, assigned translational response as a reference variable and build an automated machine learning model (auto-ML) in order to further identify more mTOR related mRNAs (topRNAs [3]) and devise a computational model to discover their sequence and structural specificities.

2 Characterisation of differentially translated mRNAs

A set of 525 differentially translated mRNAs between normal β-cells and cells exposed to high glucose for 20 minutes. Differential translation was determined by applying polysome profiling separation inthree different groups (monosomes, light and heavy polysomes) followed by RNA-seq, normalisation and differential expression analysis with the *limma* suite. A modelling cluster algorithm (Mclust) was employed to classify log-odd (high to low glucose) translation ratios into 6 distinct mRNA clusters [1]. Each cluster corresponds to different translational behaviour of mRNAs. The cluster with the largest and fastest increase in translation induction has been found to contain the vast majority of already known pre-described mTOR regulated mRNAs (topRNAs). An in-house developed software tool was used to extract and calculate 12 distinct sequence and structural features (including coding sequence length, length and structural complexity of each UTR) for each differentially translated mRNA [1]. After an one by one statistical analysis of each of the 12 features a subset of the 5 most significant ones were selected as input to an auto-ML model.

3 Identifying novel m-TOR related top-RNAs.

We identify the 5 most significant sequence and structural features of each of the 525 differentially translated mRNAs. And for each a measure of translational response was calculated based on the log-odds of the translation levels between high and low glucose. This measure was then used as a response variable to train an auto-ML model with these 5 features. The mRNAs identified as targets of the mTOR pathway, and found enriched in already known top-RNAs were used as positive examples and the cluster with the mRNAs exhibiting a drop in active translation was used as negative example. After the convergence of the auto-ML model the results were validated by using a second cluster of mRNAs with increased translational response as a set of "unknown" examples for validation. We report the performance of the model as well as the significance of each feature as tools to identify novel targets of the m-TOR pathway from polysome/ribosome profile data.

References

[1]

Poster 215

Genomic SNP detection by RNA-seq: lessons from multi-tissue & multi-population data analysis in chickens.

Jehl F[1$], Degalez F[1$], Bernard M[2$], Lecerf F[1], Coulee M[1], Zerjal T[3], Pitel F[4], Klopp C[2], Lagarrigue S[1]*

1 PEGASE, INRAE, Institut AGRO, 35590, Saint Gilles, France
2 SIGENAE, INRAE, 31326 Castanet-Tolosan - 78350 Jouy-en-Josas, France
3 GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France
4 GenPhySE, INRAE, Centre Occitanie-Toulouse, 31326 Castanet-Tolosan, France

§: Co-first authors
*: corresponding author

1. Introduction

The identification of a large number of variants (mostly SNP) in the genome is crucial to unravel the relationship between genotype and phenotype, to analyze the genetic diversity between populations or to detect variants with deleterious effect in protein-coding regions. The RNA-seq technology gives access to tissue transcriptomes and is mainly used for gene expression analysis. It can also be used to detect DNA variants in the expressed regions of the genome. However, the transcriptome is much more complex than the genome, thus requiring specific tools for sequence data analysis: it is characterized by mature transcripts corresponding to spliced genomic regions that can have highly variable expression levels, leading to depths of a few reads to tens of thousands of reads from one genomic position to another. Nevertheless, it is possible to detect genomic variants from RNA-seq data as shown by Piskol et al in 2013 [1]. However, these studies are still rare, few of them using the current reference tools and none of them have compared the variant detection performances between RNA-seq data and DNA-seq data (the latter constituting the reference), using the same individuals and tissues.

2. Results

Here we used the reference tools for RNA-seq sequence alignment (STAR-2-pass) and for variant calling (PicardTools and GATK pipeline). We analyzed 700 RNA-seq samples from 2-5 tissues of 300 chickens belonging to 10 populations. Using 2 populations for which DNAseq and RNAseq data were available on the same individuals and the same tissue (liver), we show the reliability of the SNP detection procedure with RNA-seq data (with 85% of SNP detected by DNA-seq procedure in the expressed regions) and comment the filters proposed by GATK. We show that the addition of tissues allows to increase the number of SNPs of ~+50% in average. We also provide a description of the ~10M SNPs detected from the 10 populations. We then determine appropriated filters for selecting SNP with reliable genotypes, and detect ~3M of SNPs with an intersection of 85000 SNP between the 10 populations. Finally, using genotype frequencies of this last SNP set, we analyze the genetic distances between the 10 populations and describe some interesting variations across populations

Acknowledgements

We thank M Tixier-Boichard, D Gourichon and B Bed'hom from INRAE and T Burlot from NOVOGEN for tissue accessibility; M Boutin, C Desert and L Lagoutte from PEGASE, Rennes and S Leroux from GenPhySe, Toulouse for RNA extraction; O Bouchez from GetPlaGe, Toulouse for RNA sequencing, F. Lecerf from PEGASE Rennes for the informatic support. This study is part of the ANR project "Chickstress" (https://anr.fr/Projet-ANR-13-ADAP-0014) and of the European project "Feed-a-Gene" (https://www.feed-a-gene.eu/).

References

1.Piskol et al. 2013, Reliable Identification of Genomic Variants from RNA-Seq Data. Am. J. Hum. Genet. 641-651

Poster 217

Using Neural Networks to Identify Biologically Relevant Features and Predict Cancer Types From Mass Spectrometry Data

Khawla Seddiki[1,3], Frédéric Precioso[2], Philippe Saudemont[3], Nina Ogrinc[3], Maxence Wisztorski[3], Michel Salzet[3], Isabelle Fournier[3] and Arnaud Droit[1]

[1] Centre de Recherche du CHU de Québec - Université Laval, Québec City, Canada

[2] Université Cote d'Azur, CNRS, I3S, Sophia Antipolis, France

[3] Univ. Lille, Inserm, CHU Lille, U1192 - Protéomique Réponse Inflammatoire Spectrométrie de Masse - PRISM, F-59000 Lille, France

Corresponding author: khawla.seddiki.1@ulaval.ca

Mass Spectrometry (MS) has become an increasingly important technology for metabolomic profiling. Advances in this domain promote its use for widespread and high-throughput cancer diagnosis purposes. Step-By-Step Machine Learning (ML) methods are generally applied to preprocessed MS data. Whereas building models with ML algorithms is time-consuming and can lead to errors, End-To-End Deep Learning (DL) are effective classifiers when applied to raw data, and hence eliminates the need for preprocessing. Convolutional Neural Networks (CNNs) are one of the most DL successful architectures. They have gained enormous traction in a variety of applications [1]. While their success in medicine and health care seems to be gaining ground, one major limitation hindering their acceptance is likely their lack of interpretability [2]. The "black box" nature of these models results in a poorly understood functioning and a difficult identification of the features that have influenced the classification decision [3]. Understanding the rationale behind the model's predictions would help users to evaluate the reliability of the predictions and potentially lead to the correction of its deficiencies. Such a transparency is fundamental in clinical contexts to provide insights on how conclusions are reached and aid to establish a relationship between pathological phenotype and properties revealed by the model [4].

In our study, we propose a CNN model to classify 1D-MS data and interpret which features in the input signal are class-discriminative during the classification. We firstly trained a 1D-CNN to classify two cancer datasets, one on gastric cancer (Healthy, Adenocarcinoma, and Poorly cohesive carcinoma types) and the other on ovarian cancer (Healthy and Adenocarcinoma types). To interpret the trained model, we developed a layer-wise propagation deconvolved network in 1D space (1D-DeconvNet) during the network knowledge extraction. We developed a feature importance heatmap (FIH) to illustrate the features degree of contributions highlighted by the 1D-DeconvNet. We selected the most significant features values (among positive and negative hidden weights) including some that were corroborated by the literature as they were described as predictive biomarkers (e.g. sulfatides in ovarian cancer [5] and aminophospholipids in gastric cancer [6]). The others are under oncologists validation and may constitute new therapeutic or diagnostic targets. Overall, our contribution substantially advances the current methodology used by clinicians as it provides an accurate and transparent tool to know how a CNN algorithm makes a cancer diagnosis decision.

References

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[2] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):93, 2018.

[3] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[4] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Towards medical xai. *arXiv preprint arXiv:1907.07374*, 2019.

[5] Ying Liu, Yanfeng Chen, Amin Momin, Shaner, et al. Elevation of sulfatides in ovarian cancer: an integrated transcriptomic and lipidomic analysis including tissue-imaging ms. *Molecular cancer*, 9(1):186, 2010.

[6] Livia S Eberlin, Robert J Tibshirani, Jialing Zhang, Teri A Longacre, Gerald J Berry, David B Bingham, Jeffrey A Norton, Richard N Zare, and George A Poultsides. Molecular assessment of surgical-resection margins of gastric cancer by mass-spectrometric imaging. *Proceedings of the National Academy of Sciences*, 111(7):2436–2441, 2014.

Poster 218

Impact of Cultural transmission of reproductive success on genetic diversity

Jeremy GUEZ [1], Romain LAURENT [1], Bruno TOUPANCE [1], Evelyne HEYER [1], Frederic AUSTERLITZ [1], Flora JAY [2]

[1] MNHN, 17 place du Trocadéro, 75016, Paris, France

[2] LRI, 6 Rue Noetzlin, 91190, Gif-sur-Yvette, France

Corresponding Author: frederic.austerlitz@mnhn.fr

Cultural transmission of reproductive success (CTRS) impacts the evolution of allelic frequencies [1,2]. This kind of transmission has been observed within several human populations around the World: people with many siblings tend to have more children, resulting in a higher variance in family-size. This transmission also leads to a typical signature in population genetic diversity: CTRS influences the shape of the coalescent trees, making them imbalanced [2,3]. This has been demonstrated in particular on Y chromosome and mtDNA data, allowing the inference of patrilinear and matrilinear transmission respectively [4,5].

We are now pursuing this research on autosomes for bilinear CTRS inference, which requests taking into account the impact of recombination. Because of recombination, each block within the genome has its own coalescent tree, resulting in a series of trees along the genome for each population. Our study will thus consists in: (a) simulating and analysing these tree series under different CTRS conditions, (b) constructing a simulation panel for later training of CTRS inference models, (c) caracterizing CTRS in various human populations.

To better understand the impact of CTRS on the genetic variation and in particular on the shape of coalescent trees along a recombining sequence, we integrated the CTRS model described by Brandenburg et al. (2012) into the forward-in-time simulator SliM [6]. Through this, we are able to evaluate the impact of numerous parameters such as population size, recombination rate, level of assortative mating and polygamy... In addition, we investigated the tree imbalance dynamics over time which is particularly challenging to observe in real data.

The analysis of real data will involve the inference of coalescent trees along the genome of several populations (e.g. with tsinfer [7]) and the computation of imbalance indices (from Brandenburg et al., 2012), as a starting point for CTRS inference in these populations.

References

1. Austerlitz, F., & Heyer, E. (1998). Social transmission of reproductive behavior increases frequency of inherited disorders in a young-expanding population. Proceedings of the National Academy of Sciences, 95(25), 15140–15144.

Poster 220

StatABio Facility: your user-friendly statistics

Manuela PASTORE[1], Emeric DUBOIS[2] and Christelle REYNES[1,2],
1)StatABio, Biocampus, Univ. Montpellier, CNRS, Inserm, 2)Institut de Génomique Fonctionnelle, Univ. Montpellier, CNRS, Inserm, 141, rue de la Cardonille; 34094 Montpellier Cedex 5, France

Nowadays it is difficult to read news or articles not mentioning artificial intelligence and statistics.
Why statistics is so famous and ...disliked?
Probably we all unconsciously know that we need it but we are a little afraid of what is not in our habits.
However, statistics

- provide methods to answer questions
- increase turnover
- help evaluate uncertainty and not rush into judgement
- help analyze and make sense of the data
- contribute to standardization and reproducibility

StatABio is part of Biocampus, a network of biological facilities, that

- interacts with other facilities and laboratories to enhance accomplishments,
- explores new methods to highlight results and make scientific research more robust

StatABio offers statistical analysis applied, mainly but not only, to biomedical data. It accessible to both academy and industry/biotech.

We follow an interactive process with the customer to summarize the needs through precise questions and find relevant answers together.
A proper data format is essential to rapidly start data analysis.
For project management we use the Redmine project management tool.

Services:

- Experimental design and statistical analysis
- Tailored forms for data records and ethics committees
- Speeding up publication and peer-review process for statistics
- Training

Field of expertise:

- Descriptive and inferential analysis
- Variable selection/ dimension reduction
- Model fitting
- Machine Learning
- Sample size determination & power calculation
- Speeding-up publication concerning statistical issues
- R language

Some examples of applications:

- Genetic Algorithms and Linear Discriminant Analysis to investigate resistance to chemotherapy in humans.
- GLM and Weighted Gene Coexpression Network Analysis to link clinics to genetics in Alzheimer Disease mouse models.
- Hierarchical Clustering to characterize the link between HIV infection and insulin-resistance.

Poster 221

How to choose a reference genome (while waiting for a pangenome)

Clothilde Chenal[12] *, Christine Tranchant-Dubreuil[1] *, Anna-Sophie Fiston-Lavier[3], Frédéric Simard[2], Yves Vigouroux[1], Michael Fontaine[2], François Sabot[1] and the *Anopheles gambiae* 1000 Genomes Consortium

[1] DIADE, Univ Montpellier, IRD, Montpellier, France
[2] MIVEGEC, Univ Montpellier, CNRS, IRD, Montpellier, France
[3] ISEM, Univ Montpellier, EPHE, CNRS, IRD, Montpellier, France

Corresponding authors: clothilde.chenal@ird.fr, christine.tranchant@ird.fr

Most bioinformatics pipelines require a reference genome, regardless of the data analyzed. However, choosing one can increase the quality and accuracy of the downstream analysis or can have a detrimental effect on it. In the same way, it has been shown that the choice of a particular gene annotation has a significant impact on the analysis of RNA-seq data [1,2]. Indeed, a reference genome is idiosyncratic and implies biases [3]. In alignment, it refers to the tendency for some reads with non-reference alleles to not be mapped or mapped at lower rates. Furthermore, variant callers might give more "variants" when the reference alleles are rare or could fail to call variants that are rare but also shared by the reference [4,5].

Therefore, it is crucial to have the -almost- perfect reference genome for a specific question and methodology. Selecting the right organism (species, strain,...), sequenced with the adapted technology, retrieving the proper version of this genome and finally picking the appropriate level of mask (soft-masked, hard-masked, unmasked) is not a trivial task yet. For this purpose, we will compare the results obtained under different conditions : reference genome types (type specimen vs. individual sampling), sequencing technologies (short reads vs. long reads) and species (both for cultivated and wild African rice, as for mosquitoes : species complex *Anopheles gambiae*).

This raises eventually the question of the ideal reference genome, taking into account all the individual variations (SNPs, CNVs, PAVs, ...). A pangenome may be able to capture all that complexity, since it is an inventory of genomic items shared or not (including genes) within a group of related individual [6]. It can be represented as an oriented graph [7], in which alternative paths replace both the structural variants and the unique variants [8]. By encompassing the complete genomic variation repertoire within a population, the pangenome-based approach offers unprecedented opportunities to uncover new genes, to fully explore genetic diversity and to better understand the impact of these variations in selection and adaptation processes.

References

[1] Davis J. McCarthy, Peter Humburg, Alexander Kanapin, Manuel A. Rivas, Kyle Gaulton, Jean-Baptiste Cazier, Peter Donnelly, and The WGS500 Consortium. Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine*, 6(3):26, 2014.

[2] Adam Frankish, Barbara Uszczynska, Graham RS Ritchie, Jose M. Gonzalez, Dmitri Pervouchine, Robert Petryszak, Jonathan M. Mudge, Nuno Fonseca, Alvis Brazma, Roderic Guigo, and Jennifer Harrow. Comparison of gencode and refseq gene annotation and the impact of reference geneset on variant effect prediction. *BMC Genomics*, 16(8):S2, 2015.

[3] Sara Ballouz, Alexander Dobin, and Jesse A. Gillis. Is it time to change the reference genome? *Genome Biology*, 20(1):159, 2019.

[4] Alberto Magi, Romina D'Aurizio, Flavia Palombo, Ingrid Cifola, Lorenzo Tattini, Roberto Semeraro, Tommaso Pippucci, Betti Giusti, Giovanni Romeo, Rosanna Abbate, and Gian Franco Gensini. Characterization and identification of hidden rare variants in the human genome. *BMC Genomics*, 16(1):340, 2015.

[5] Yury A. Barbitoff, Igor V. Bezdvornykh, Dmitrii E. Polev, Elena A. Serebryakova, Andrey S. Glotov, Oleg S. Glotov, and Alexander V. Predeus. Catching hidden variation: systematic correction of reference minor allele annotation in clinical variant calling. *Genetics in Medicine*, 20(3):360–364, 2018.

[6] Christine Tranchant-Dubreuil, Mathieu Rouard, and Francois Sabot. Plant Pangenome: Impacts on Phenotypes and Evolution. pages 453–478, 2019.

* both authors contributed equally to this manuscript

Poster 222

The Pole for data analysis of the Muséum National d'Histoire Naturelle

Jawad ABDELKRIM[1], Fayçal ALLOUTI[1], Amandine BLIN[1], Claire GACHON[1], Julien MOZZICONACCI[1], Nicholas TODA[1], Patricia WILS[1] and Anthony WOO[1]

[1] UMS 2AD: Acquisition et Analyse de Données pour l'Histoire Naturelle, Muséum National d'Histoire Naturelle, CP 26, 57, rue Cuvier, 75231 PARIS Cedex 05, FRANCE

Corresponding Author: claire.gachon@mnhn.fr

http://ums2700.mnhn.fr/analyse-de-donnees/presentation

Embedded in the UMS 2AD (*Acquisition and Analyse de Données*) of the Muséum National d'Histoire Naturelle, the pole for data analysis opened in early 2019. It aims to provide its users with multiple services, including: an assistance with project design, partial or total management of data analysis, methodological development, as well as animation and training actions.

Our areas of expertise are:

Statistical processing, such as modeling (regressions, generalised linear models…), tests, descriptive statistics, multidimensional statistics.

Genes & Genomes: Alignment and assembly, phylogeny, meta-genomics, epigenetics, transcriptomics, comparative genomics.

Image analysis: Segmentation, quantification, 3D modeling, visualization.

To carry out these analyzes, we master several programming languages (R, Python, C ++ ...). We can also provide access to software (Geneious, Galaxy, ImageJ, Aviso ...) Access can be requested via the completion of a form available on the pole's web page. Training workshops and seminars open to all are regularly organised and announced on the pole's web page.

Poster 223

Pilot study to screen alleles under selection related to drought tolerance in banana

Catherine BRETON[1], Alberto CENCI[1], Nicolas ROUX[1], Mathieu ROUARD[1] and Sebastien CARPENTIER[2,3]

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

[2] Bioversity International , Willem De Croylaan 42, B-3001 Leuven, Belgium.

[3] KULeuven mass spectomery facility SYBIOMA, Herestraat 49, B-3000 Leuven, Belgium

Corresponding Author: c.breton@cgiar.org

In banana, most of the edible cultivars are hybrids from one or both major wild diploid ancestors: *Musa acuminata* (A genome, 2n=*2x*=22) and *Musa balbisiana* (B genome, 2n=*2x*=22). These cultivars have been assigned to different genomic groups according to the number of chromosomes sets (subgenomes) and the species that donated them, e.g. AAA, AAB and ABB (2n=3x = 33). Variability in subgenome contribution to triploid clones makes pertinent to investigate the differential gene expression of *Musa* cultivars, in particular since genotypes with B subgenomes can be more tolerant to abiotic stresses than those containing only A subgenomes.

Due to the increasing power of high-throughput genotyping and transcriptomic methods, associated with the development of genetic and statistical analysis tools, significant knowledge gains have been generated for diploid species, [2], but on polyploids analysis it is still challenging [1,3]. Polyploidy can affect the phenotype creating complex interactions between loci or alleles, such as dominance or epistasis, or through allelic dosage. This represents an additive effect of multiple copies of the same allele. Collecting information of allelic dosage allow to describe a more realistic representation of the effect of each genotype.

The present pilot study describes the methodology to analyse and select Single Nucleotide Polymorphism (SNP) alleles in two conditions (drought stressed and control) on triploid *Musa acuminata* AAA, ABB and AAB genotypes. A SNP database developed under both conditions provided an opportunity for such allele detection and to study the nonsynonymous SNPs (substitutions altering the amino acid sequence of a protein) and synonymous SNPs (which do not affect amino acid sequence). These nonsynonymous substitutions could affect the response to a stress as drought. The nonsynonymous and synonymous substitution rates (Ka/Ks) indicate the direction of natural selection. The prediction of both positive (diverging) and negative (purifying) selection pressure on genes could provide the picture detail of evolutionary selection pressure linked to the allele dosage.

Musa, NGS, drought, allelic dosage, RNAseq, Ka/Ks

References

1. Cenci, A., Hueber, Y., Zorrilla-Fontanesi, Y., van Wesemael, J., Kissel, E., Gislard, M., & Rouard, M. (2019). Effect of paleopolyploidy and allopolyploidy on gene expression in banana. BMC Genomics, 20(1):244, 1-12.
2. Comai L., 2005 The advantages and disadvantages of being polyploid. Nat. Rev. Genet. 6: 836–846. doi:10.1038/nrg1711
3. Desta Z. A & Ortiz R., (2014) Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci. 19: 592–601. doi:10.1016/j.tplants.2014.05.006
4. Grandke F., Singh P., Heuven H. C., De Haan J. R., & Metzler D (2016) Advantages of continuous genotype values over genotype classes for GWAS in higher polyploids: a comparative study in hexaploid chrysanthemum. BMC Genomics 17: 672. doi:10.1186/s12864-016-2926-5
5. Sempéré G, Philippe F, Dereeper A, Ruiz M, Sarah G, Larmande P. (2016) Gigwa-Genotype investigator for genome-wide analyses. Gigascience. 5:25.

Poster 224

OMICSity, a database for Luminex immunoassay data storage and analysis

Corinne MENIGOT [1], Luigi RUSSO[1], Didier ZUGAJ[1]

Nicolas NOTTET[1], Romy RAMBERT [1], Bruno MEHUL[1], Francesca ZOLEZZI[1], Agnes PAQUET[1]

[1] Syneos Health, Integrated Biomarker Solutions, Translational Sciences
2400 route des Colles, 06410, Sophia-Antipolis, France
Corresponding Author: corinne.menigot@syneoshealth.com

Introduction

Multiplex immunoassays are widely used in clinical studies to quantify biomarkers, especially soluble mediators of inflammation (e.g., cytokines and chemokines) in serum or plasma, to profile drug efficacy. Our goal was to develop a bioinformatics tool to automate analyses of Luminex assays and centralize all research data generated by Syneos Health's proteomics platform.

Methods

OmicSity was developed using Windev from the design of the database to the creation of the interface. We designed a relational database (Merise) specifically tailored to targeted proteomics assay, and graphical interfaces for Windows. Our application allows importing raw data files exported from Bio-Plex manager with the associated plate design and sample/kit metadata, data normalization, and it performs quality control from dynamic graphs. Finally, a query interface was built to retrieve information across studies.

Results

Our tool is structured around four main functionalities:

1. Admin: to manage/update the different drop lists of metadata in the application (pathologies, tissue types ….)
2. Study: to create/update a study with the associated metadata. This window is used to import the raw data (excel file) with the associated design, and perform several types of analyses, such as normalization of the raw data (3 methods are implemented).
3. Graphics: 4 graphical visualizations are currently implemented for quality control of the assay data:

   (1) 5 PL equation plot (regression) with the standards, the controls, the blanks and the study samples,
   (2) The % CV,

   (3) 5PL Parameter 'a' Lower Asymptote

   (4) 5PL Parameter 'g' Asymmetry

   A report in Microsoft Word format can be automatically generated with all the QC graphs.
4. Query: Queries can be made on the study metadata to extract all data associated with a study. Metadata queries can also be combined with analytes queries to retrieve raw data/normalized data for a selected analyte for a specific pathology in all studies in the database that match the criteria.

Conclusion

OmicSity is now used in-house to manage our R&D targeted immunoassay data. Future development of the tool will include the management of untargeted proteomics and lipidomics datasets.

Poster 225

Génération de workflow d'analyse conteneurisé avec interface graphique

Mathieu Massaviol[1], Jimmy Lopez[1], Remy Dernat[2] et Khalid Belkhir[2]
[1] Labex CeMEB, Place Eugène Bataillon, Montpellier, 34095, France
[2] Institut des Sciences de l'Evolution (UMR 5554), Place Eugène Bataillon, Montpellier, 34095, France

Auteur référent: mathieu.massaviol@umontpellier.fr

L'évolution des techniques de séquençage et leur démocratisation au cours de ces 20 dernières années ont entraîné un important afflux de données à analyser. Des outils informatiques ont été développés pour traiter ces données. Ces outils de par leur multiplicité sont très souvent utilisés les uns à la suite des autres dans une chaîne de traitement des données appelée workflow ou pipeline. Plusieurs problématiques se posent quant à leur utilisation. Premièrement la facilité d'utilisation des outils et workflows. Ces derniers nécessitent souvent des compétences en ligne de commande, de l'installation à l'exécution, ce qui peut freiner leur utilisation. Deuxièmement on retrouve la problématique de la reproductibilité. L'évolution rapide des outils rends parfois difficile le fait de ré-effectuer une analyse dans les mêmes conditions (versions des outils et paramètres identiques). Des solutions pour faciliter la conception et l'utilisation de workflows ainsi que leur reproductibilité ont été développées comme par exemple Galaxy[1].

Nous avons conçu un framework de génération automatique de workflows. Ce système permet de générer un workflow ainsi que son environnement d'exécution et de l'utiliser avec ou sans interface graphique sur différentes infrastructures comme une machine personnelle, un serveur de calcul, un cluster, un cloud, etc, sans nécessiter de compétences système.

Ce framework intègre une soixantaine d'applications bioinformatiques et nous a permis de générer une dizaine de workflows.

Nous partons de fichiers détaillant d'une part les outils d'analyses avec leurs paramètres, leurs entrées et sorties, leur méthode d'installation, etc, et d'autre part la description des workflows avec l'enchaînement des étapes et le cheminement des données d'une étape à l'autre. Nous avons ensuite développé différents scripts pour générer l'environnement permettant l'utilisation du workflow : la recette du workflow en Snakemake[2], une recette de conteneur (Docker[3] ou Singularity[4]) ainsi qu'une application Shiny[5] permettant de paramétrer et exécuter le workflow. Le conteneur est alors autonome et contient les versions figées des outils ainsi que le workflow Snakemake et l'application Shiny.

L'exécution des workflows via l'interface Shiny ou en ligne de commande génère en plus des résultats un rapport (MultiQC[6]) reprenant les sorties des étapes du workflow ainsi que les versions des outils utilisés.

Références

[1] 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 Res.*, 46(W1) :W537–W544, 2018.

[2] Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19) :2520–2522, 08 2012.

[3] Dirk Merkel. Docker : Lightweight linux containers for consistent development and deployment. *Linux J.*, 2014(239), March 2014.

[4] Gregory M. Kurtzer, Vanessa Sochat, and Michael W. Bauer. Singularity : Scientific containers for mobility of compute. *PLOS ONE*, 12(5) :1–20, 05 2017.

[5] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. *shiny : Web Application Framework for R*, 2019. R package version 1.3.2.

[6] Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. MultiQC : summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19) :3047–3048, 06 2016.

Poster 230

Molecular mechanisms of the sodium ion binding to G protein coupled receptors

Asma Tiss [1,2], Rym Ben Boubaker [1], Linda Grimaud [1], Daniel Henrion [1], Hajer Guissouma [2] and Marie Chabbert [1]

[1] Laboratoire MITOVASC, UMR CNRS 6015 – INSERM 1083, University of Angers, Angers, France
[2] Laboratoire GIPH, Faculty des Sciences de Tunis, University of Tunis El Manar, Tunisie

Corresponding Author: asma.tiss@etud.univ-angers.fr

G protein coupled receptors (GPCRs) are the targets of 30% of available drugs [1]. Understanding their structural and functional properties is mandatory to develop more efficient drugs. The sodium ion is a negative allosteric regulator of GPCRs. Previous studies of allosteric regulation by sodium binding showed the importance of its implication in regulating the receptor functions [2].

In this study, we investigated the reactional pathway(s) by which the sodium ion reaches its allosteric binding site in three different receptors: the delta opioid receptor OPRD, the chemokine receptor CXCR4 and the angiotensin II receptor AT1. For this latter receptor, the presence of a sodium binding site is controversial [3]. To reach this objective, we carried out classical and accelerated molecular dynamic simulations of the apo-receptors embedded within a hydrated POPC membrane. The water layers contained a physiological concentration of sodium ions (0.15M).

In the three cases, we observed that a sodium ion could enter into the receptor internal cavity from the extracellular side and bind to the canonical sodium binding sites. For OPRD and CXCR4, sodium binding required less than 100 ns of classical MD simulations. For AT1, sodium binding was much slower and observed only upon accelerated conditions. Nevertheless, analysis of the sodium pathways to the allosteric site revealed similarities in the binding mechanisms for the three receptors. These results highlight the conserved role of the sodium ion in GPCR functions.

References

1. Rask-Andersen. Trends in the exploitation of novel drug targets. *Nat Rev Drug Discov*, (10):579–590, 2011.
2. Liu, W. et al. Structural Basis for Allosteric Regulation of GPCRs by Sodium Ions. *Science,* (80):232–236, 2012.
3. Wingler, L. M. et al. Angiotensin Analogs with Divergent Bias Stabilize Distinct Receptor Conformations. *Cell*, (176):468–478, 2019.

Poster 232

Meso@LR

Baptiste CHAPUISAT, Anne COUDERT and Anne LAURENT

Université de Montpellier, Place E.Bataillon, 34095, Montpellier, France

Corresponding Author: meso-lr-direction@umontpellier.fr

1. Introduction

Open to companies and academic, public and private actors, the Meso@LR platform (formerly HPC @ LR) offers shared resources and advanced architectures for high performance computing (High Performance Computing - HPC) and massive data processing (High Troughput Computing - HTC).

Acting as a proactive player of the French Tech in Montpellier, it also contributes to the strong dynamic of the Montpellier site around ISITE MUSE, especially in the fields of Agriculture, Environment and Health, for addressing the MUSE three major intertwined challenges: Feed, Protect, Care. The majority of Meso@LR users are located in Eastern Occitanie, with a national opening.

It currently offers 308 nodes (8624 cores + 2 SMP Nodes (3To RAM)). In 2020, a massive storage infrastructure will be implemented.

Computing resources can be accessed through SLURM. The Module Environment allows users to be provided with different versions of a same software. They can install softwares in their own environments. The technical resources of the mesocentre are hosted at CINES.

Meso@LR works closely on links to data science and artificial intelligence, especially in the framework of the Institut de Science des Données de Montpellier (ISDM).

2. Access Modes

The Resources are accessible in a flexible way, either on the fly, either by the hour or by providing dedicated and secure environments.

Particular interest is paid to teaching activities and capacity building, in the framework of practical work, workshops and a training cycle operated by the mesocentre.

Acknowledgements

Meso@LR, carried by the University of Montpellier, received support from the OCCITANIE / Pyrénées-Méditerranée Region and from Montpellier Méditerranée Métropole as part of the 2015/2020 CPER project.