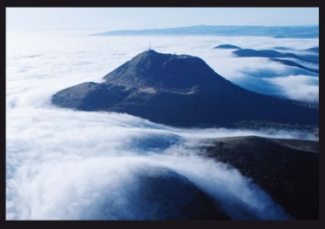
JOBIM

Clermont-Ferrand - 6-9 juillet 2015

Journées
Ouvertes
en Biologie,
Informatique et
Mathématiques



JOBIM 2015

UdA | Université d'Auvergne

INRA
SCIENCE & IMPACT

UNIVERSITE
BLAISE PASCAL
CLERMONT-FERRAND
UBP

JOBIM 2015

Journées Ouvertes en Biologie Informatique & Mathématiques

Polydôme, Clermont-Ferrand, 6-9 Juillet 2015

Editeurs
Philippe Leroy
Pierre Peyret

Journées Ouvertes de Biologie, Informatique et Mathématiques

Par Philippe LEROY – philippe.leroy@clermont.inra.fr – UMR 1095 INRA/UBP, Clermont-Ferrand, France

Réalisation : Philippe LEROY (INRA GDEC) & Pierre PEYRET (Université d'Auvergne EA CIDAM)
Conception Logo : Audrey LENOBLE-CHAMERLIN (auto-entrepreneur)
Gestion des fichiers PDF sous EasyChair (http://www.easychair.org/) : Marie PAILLOUX (LIMOS, Université Blaise Pascal)
Mise en page : Philippe LEROY (INRA GDEC)
Mise en forme : Philippe LEROY & Nicolas GUILHOT (INRA GDEC)
Couverture : Philippe LEROY (INRA GDEC) et Audrey LENOBLE-CHAMERLIN (auto-entrepreneur)
Photos : Jodie WAY (Photographe) – Ville de Clermont-Ferrand

Editeur : INRA & Université d'Auvergne – **version numérique**

ISBN : 2-7380-1377-5
Code EAN : 978 273 801 3774

Préface

Marie-France SAGOT, INRIA, Lyon
Invitée d'honneur JOBIM 2015

Clermont-Ferrand a le plaisir d'accueillir la seizième édition des Journées Ouvertes en Biologie, Informatique et Mathématiques. Depuis 2000, la conférence rassemble, dans le cadre d'un chaleureux rendez-vous annuel de partages et d'échanges, une communauté francophone toujours aussi active en bioinformatique, biomathématique et biostatistique. Comme chaque année, la conférence est placée sous l'égide de la Société Française de Bioinformatique. JOBIM aura aussi l'immense plaisir d'accueillir en amont de la conférence le séminaire de JeBiF, l'Association des Jeunes Bioinformaticiens de France qui contribue pour beaucoup à la vivacité de la communauté.

JOBIM recouvre un foisonnement de plus en plus riche de disciplines, allant des plus traditionnelles telles la génomique, la bioinformatique structurale, l'évolution et la phylogénie, où cependant d'innombrables problèmes demeurent ouverts, à la biologique systémique à travers l'étude des réseaux métaboliques et de régulation, l'épigénétique et l'épigénomique, la génétique y compris des populations, la biologie translationelle et la pharmacogénomique, les méta-omiques et la génomique environmentale, l'imagerie et le traitement de l'image. Ces multiples sujets seront abordés en particulier par les 11 invités de renommée internationale que JOBIM est fier d'accueillir cette année : ANA CONESA, ANTHONY COX, ROB FINN, CHRISTOPH GRUNAU, KATHARINA HUBER, LAURENCE MOREAU, CHRISTINE ORENGO, STEPHANE ROMBAUTS, DANIEL SEGRE, THOMAS WALTER, ZIHENG YANG. Nous les remercions tous chaleureusement d'avoir accepté de participer à la réussite de ces journées en nous exposant quelques unes de leurs réussites scientifiques récentes.

La nouvelle explosion des données dues aux techniques à haut débit en constante innovation, ainsi que le dynamisme de la communauté et les échanges extrêmement fructueux entre disciplines ont aussi fait que de plus en plus de services, ressources et infrastructures pour la bioinformatique sont mises à la disposition des chercheurs. Ceux-ci seront présentés tout au long de la conférence à travers des Annonces, des Présentations de Faits Marquants, des Posters, et des Démonstrations.

En tout, JOBIM aura reçu 193 soumissions, et nous profitons pour remercier l'ensemble des relecteurs sollicités dans le comité de programme et au-delà pour leur travail important de révision. Nous espérons que leurs commentaires auront aidé le plus grand nombre à améliorer la qualité des contributions, dont 22 ont été acceptées pour une présentation orale (soumissions originales ou faits marquants), et 15 pour une démonstration. Par ailleurs, 143 soumissions seront présentées sous forme d'affiches pour être discutées tout au long de ces journées.

Un grand merci à nos partenaires académiques et industriels, aux collectivités territoriales locales, à tous les membres des comités de programme et d'organisation, ainsi qu'à l'ensemble des bénévoles qui ont largement oeuvré à la réussite de JOBIM 2015.

Benvolença a totes en auvèrnha, e subrebona conferença !

Pour le Comité Scientifique
Pierre PEYRET, EA-CIDAM, Université d'Auvergne
Jérôme SALSE, GDEC, INRA

Pour le Comité Logistique
Eric PEYRETALLADE, EA-CIDAM, Université d'Auvergne
Philippe LEROY, GDEC, INRA

Comité logistique

Philippe LEROY - Inra (**co-président**)
Éric PEYRETAILLADE - Université d'Auvergne (**co-président**)

Violaine ANTOINE - Université d'Auvergne
Julien BOILEAU - Office du Tourisme Clermont-Ferrand
Delphine BOUCHER - Université d'Auvergne
Delphine BOYER - Inra
Agnès COHADE - Inra
Isabelle DELPIT - Université d'Auvergne
Sabrina GASSER - Inra
Françoise GRAIVE - Office du Tourisme Clermont-Ferrand
Adnane HITMI - Université d'Auvergne
Aurélie LAMBERT - Université d'Auvergne
Nicolas MAINETTI - Université d'Auvergne
Manon MARTINET - Université d'Auvergne
Marie PAILLOUX - Université Blaise-Pascal
Nils PAULHE - Inra
Pierre PEYRET - Université d'Auvergne
Valérie POLONAIS - Université d'Auvergne
Matthieu REICHSTADT - Inra
Cathy RESSOT - Université d'Auvergne
Karine RIBEYRE - Inra
Sébastien RIMOUR - Université d'Auvergne
Isabelle ROBINET - Office du Tourisme Clermont-Ferrand
Jérôme SALSE - Inra
Sébastien TEMPEL – Université Marseille
Patricia TIXIER - Inra
Sylvie TOILLON – Inra

Et avec l'aide ponctuelle mais très efficace et essentielle de Nicolas GUILHOT – Inra GDEC

Comité Scientifique

Pierre PEYRET - Université d'Auvergne (**co-président**)
Jérôme SALSE - Inra (**co-président**)

Violaine ANTOINE - Université Blaise-Pascal

Denis ARDID - Université d'Auvergne

Xavier BAILLY - Inra

Vincent BARRA - Université Blaise-Pascal

Virginie BERNARD - Institut Curie

Christophe BLANCHET - IFB-core

Valentina BOEVA - Institut Curie

Richard BONNET - Inserm / Université d'Auvergne

Emilie BRASSET - Université d'Auvergne

Vincent BRETON - Cnrs / Université Blaise Pascal

Céline BROCHIER - LBBE

Gisèle BRONNER - Université Blaise-Pascal

Marie CHABBERT - Université d'Angers

Hélène CHIAPELLO - Inra

Frédéric CHOULET - Inra

Sébastien COLIN - Cnrs

Didier DEBROAS - Université Blaise-Pascal

Guillaume DEFFUANT - IRSTEA

François ENAULT - Université Blaise-Pascal

Marc FERRARA - Inra

Franck GIACOMONI - Inra

Jean-François GIBRAT – Inra

David HILL - Université Blaise-Pascal

Dominique JOLY - Cnrs

Daniel KAHN - Inra

Charles KERVRANN - Inria

Philippe LABEL - Inra

Philippe LEROY - Inra

Antoine MAHUL - CRRI

Gilles MAILHOT - Cnrs/Université Blaise-Pascal

Jean-Pierre MARTINANT - Limagrain

Engelbert MEPHU NGUIFO - Université Blaise-Pascal

Marie PAILLOUX - Université Blaise-Pascal

Éric PELLETIER - Génoscope

Bruno PEREIRA - CHU Clermont-Ferrand

Guy PERRIERE - Cnrs-LBBE

Pierre PETERLONGO - Inria

Éric PEYRETAILLADE - Université d'Auvergne

Valérie POLONAIS - Université d'Auvergne

Flora PONELLE - Centre Jean Perrin

Éric RIVALS - LIRMM

Nathalie RIVIÈRE - Biogemma

Marie-France SAGOT - INRIA **–** ***Invité d'Honneur***

Laurent SARRY - Université d'Auvergne

Sophie SCHBATH - Inra

Dominique TESSIER - Inra

Hélène TOUZET – Lifl, Lille

Cristina VIEIRA – Université Lyon 1

Annegret WAGLER - Université Blaise-Pascal

Partenaires

Organisateurs Principaux

Autres Organisateurs

Soutiens Académiques

Soutiens Collectivités Locales

Soutiens Industriels

Types de contributions

Hormis les résumés proposés par les conférenciers invités, les contributions présentées dans ce recueil sont de trois types :

- **Articles originaux** ou *Proceeding* : (2 pages minimum, 8 pages maximum – en Anglais ou en Français) : réservés aux résultats originaux non publiés par ailleurs. Ce type de contribution donne lieu à une présentation orale lors de la conférence JOBIM.

- **Résumés étendus** ou *Highlight* (2 pages maximum, *a priori* en Anglais) : présentation de résultats récents, publiés dans l'année (été 2014 – 2015). Ce type de contribution donne lieu à une présentation orale lors de la conférence JOBIM.

- **Résumés courts** ou Affiches/Poster & Démo : (1 page maximum en Français et en Anglais) : concernant des travaux récents ou en cours, publiés ou non. Ce type de contribution donne lieu à une présentation orale sous forme de démonstration lors des sessions parallèles ou un affichage sous forme d'affiche/Poster lors de la conférence JOBIM.

PROGRAMME SYNTHETIQUE

	Lundi 6 Juillet	Mardi 7 Juillet	Mercredi 8 Juillet	Jeudi 9 Juillet
08h10__			Session VI Meta-omics & Env. Genomics Topic-07 keynote: Rob FINN	
09h00__				
09h30__		Session III Organization & Genome Expression Topic-01 keynote: Stéphane ROMBAUTS		Session IX Service, Resources & Infra. Bioinfo. Topic-09 keynote: Katharina HUBER
10h30__			Session VII Translational Bio. & Pharmaco. Topic-08 keynote: Ana CONESA	
10h40__			Coffee Break	Coffee Break
11h00__		Coffee Break		
11h10__			Session IVb Genetics & Pop. Genet. Topic-03 keynote: Laurence MOREAU	Session IX Services, Ress. & Infra Bioinfo
11h40__		Session IVa Imaging & Image process. - Topic-10 keynote: Thomas WALTER Network, Reg. & Mod. - Topic-06 keynote: Daniel SEGRE		
12h30__			parallele Sessions Session D - Topic-11 Session E - Topic-11 / 03 / 09 Session F - Topic-09	Awards + JOBIM 2016 Clôture
12h40__	ACCUEIL CONFERENCE JOBIM 2015	Announcement J.F. Gibrat – IFB-core		Lunch Box Red ticket
13h00__			Lunch Green ticket	
13h30__		Lunch Blue ticket		
14h00__			Poster Session Nombres Impairs / Odd numbers	
14h30__	Ouverture JOBIM 2015			
15h10__	Session I Biochemistry, Struc. Bio.& Bioinfo Topic-02 keynote: Christine ORENGO	parallele Sessions Session A: Topic-06 Session B: Topic-07 / 09 Session C: Topic-11	Assemblée Générale SFBI	
15h30__				
16h00__		Poster Session Nombres Pairs / Even numbers		
16h10__			Coffee Break	
16h30__		Coffee Break	Session VIII Epigenetics & Epigenomics Topic-04 keynote: Cristoph GRUNAU	
16h40__				
17h00__	Coffee Break	Session V Method. Seq. Ana. & Omics Data Topic-11 keynote: Antony COX		
17h30__	Session II Evolution, Phylo. & Paleo. Topic-05 keynote: Ziheng YANG		Announcement Biogemma / Limagrain	
18h00__			Group Photo	
18h30__				
18h40__			18h30 Bus Departure Vulcania	
19h00__	Announcement Réseaux 3E & GDER GE	Announcement DUT IUT Aurillac / Master Bioinfo UBP		
19h10__			Gala Dinner Vulcania	
19h20__	Announcement JeBiF & bioinfo-fr.net			
19h30__		Cocktail - *Entre 2 Villes* - Polydome		

PROGRAMME DETAILLE

Lundi 6 Juillet 2015 – Monday, July 6th 2015

12h30 – 14h00 : Hall d'Accueil, Rez-de-Chaussez – Polydome
Accueil JOBIM 2015 / Pause-café (petits fours sucrés)

14h00 - 14h30 : Amphitéâtre/Amphitheater - Ouverture de la Conférence JOBIM 2015 – JOBIM 2015 Opening

Présidence Université d'Auvergne – **Philippe DULBECCO**
représenté par **Alain Eschalier** Vice Président Recherche – Université d'Auvergne
Présidence Université Blaise Pascal – **Mathias BERNARD**
représenté par **Khalil DRISSI** Vice Président Valorisation – Université Blaise Pascal
Présidence Centre INRA Auvergne-Rhône-Alpes – **Jean-Baptiste COULON**
représenté par **Thierry LANGIN** Directeur de l'Unité de Recherche
UMR INRA/UBP 1095 GDEC
Invitée d'Honneur – Marie-France SAGOT

Session Plénière I – Amphithéâtre/Amphitheater

Biochimie, Biologie Structurale & Bioinformatique Structurale
Biochemistry, Structural Biology & Structural Bioinformatics

Chair: Jean-François GIBRAT (Institut Français de Bioinformatique - IFB, France)

14h30 - 15h10 Keynote-01 **Christine ORENGO** - University College London (UCL) – UK
A Structural Perspective on the Evolution of Protein Functions

15h10 - 15h30 Proc-01 **Yasaman Karami** – Iran University of Science and Technology, Iran
Accelerating Protein Structure Prediction using Particle Swarm Optimization on GPU

15h30 - 15h50 Proc-02 **Isaure Chauvot de Beauchene** – Technical University Munich, Germany
Modelling ssRNA-protein complexes at atomic resolution

15h50 - 16h10 Proc-03 **Seyed Ziaeddin Alborzi** – Université de Lorraine - LORIA, France
EC-PSI: Associating Enzyme Commission Numbers with Pfam Domains

16h10 - 16h30 High-01 **Marco Pasi** – CNRS UMR5086/Université Lyon I, France
Sequence determines the structure and microsecond-scale dynamics of B-DNA and its bound cations

16h30 - 17h00 - Pause-café petits fours sucrés (*installation Affiches*) / Coffee break (*Posters installation*)

Session Plénière II – Amphithéâtre/Amphitheater

Evolution, Phylogénie & Paléogénomique
Evolution, Phylogeny & Paleogenomics

Chair: Guy PERRIERE (CNRS, LBBE, France)

17h00 - 17h40 Keynote-02 **Ziheng YANG** - University College London (UCL), UK
Estimation of species divergence times incorporating fossil and molecular information.

17h40 - 18h00 Proc-04 **Elodie Cassan –** UMR5506 CNRS, Université de Montpellier, France
Evolutionary analyses strongly support that ASP (Anti Sense Protein) overlapping ORF is the 10th gene of HIV-1 M pandemic group

18h00 - 18h20 High-02 **Blerina Sinaimeri** – INRIA Grenoble Rhône-Alpes, France
Cophylogeny Reconstruction via an Approximate Bayesian Computation

18h20 - 18h40 High-03 **Gabriel Markov** – Max Planck Institute for Developmental Biology, Germany
The same or not the same: Lineage-specific gene expansions and homology relationships in multigene families in nematodes

18h40 - 19h00 : *Announcement (2x10 min)*
- **Christophe GRUNAU** - Réseaux thématiques pluridisciplinaires 3E «Epigénétique en Ecologie et Evolution »
- **Dominique JOLY** - Groupement De Recherche en Génomique Environnementale (GDR GE)

19h00 - 19h20 : *Announcement (2x10 min)*
- **Julien FUMEY** - JeBIF
- **Yoann MOUSCAZ** - BioInfo

Soirée Libre / Free Evening

Mardi 7 Juillet 2015 - Tuesday, July 7th 2015

Session Plénière III - Amphithéâtre/Amphitheater

Organisation & Expression des Génomes
Organization & Genome Expression

Chair: Gisèle BRONNER (LMGE, Université Blaise Pascal, France)

09h00 - 09h40 Keynote-03 **Stéphane ROMBAUTS** – VIB, Gand, Belgique
The next challenge: sequencing, assembling and annotation of complex genomes with NGS

09h40 - 10h00 Proc-05 **Jérémy Tournayre** – INRA, UMR1213 Herbivores, France
Fat&MuscleDB: A database to understand tissue growth processes contributing to body or muscle composition

10h00 - 10h20 High-04 **Alexandre Cormier** – CNRS-UPMC UMR8227 LBI2M, Station Biologique de Roscoff, France
Sexual dimorphism and the evolution of sex-biased gene expression in the brown alga Ectocarpus

10h20 - 10h40 High-05 **Frédéric Choulet** – INRA/UBP UMR1095 GDEC, France
A Reference Sequence of the Bread Wheat Chromosome 3B

10h40 - 11h10 - Pause-Café viennoiseries / Coffee Break

Session Plénière IVa - Amphithéâtre/Amphitheater

Spéciales Conférences Invités
Special Invited speakers

11h10 - 11h50 Keynote-04 **Imagerie & Traitement de l'Image** - *Imaging & Image Processing*
Chair: Valérie POLONAIS (Université d'Auvergne – IUT Aurillac)

Thomas WALTER – Institut Curie, France
Bioimage Informatics for Phenomics

11h50 - 12h30 Keynote-05 **Réseaux, Régulations & Modélisation** - *Network, Regulation & Modeling*
Chair: Franck GIACOMONI (INRA, France)

Daniel SEGRÈ – Boston University, USA
Spatio-temporal models of metabolism in microbial communities

12h30 - 13h00 - *annonces diverses (2x20 min)*
- **Jean-François GIBRAT** - *IFB-core*

13h00 - 14h30 - Déjeuner - Ticket Bleu / Lunch - Blue Ticket

Parallel Sessions - Démos/Demos

	Session A	Session B	Session C
	Espace Restauration Niveau 2	**Amphitéâtre / Amphitheater**	**Salles Commission 11-12-13 - Niveau 2**
Chair	**Gabriel MARKOV**	**Benoit BELY**	**Hélène CHIAPELLO**
	Réseaux, Régulation & Modélisation	*Méta-omiques & Génomique Environnementale*	*Méthodologies pour l'analyse des séquences et des données omiques*
14h30 - 14h50	**Demo-A1** Jacques van Helden	**Demo-B1** Géraldine Pascal	**Demo-C1** Nicolas Kaspric
		Services, Ressources & Infrastructures pour la Bioinformatique	
14h50 - 15h10	**Demo-A2** Laurent Bulteau	**Demo-B2** Bryan Brancotte	**Demo-C2** Isabelle Guigon

- Demo-A1 : Regulatory Sequence Analysis Tools (RSAT)
- Demo-A2 : DINGHY: Dynamic Interactive Navigator for General Hypergraphs in Biology
- Demo-B1 : FROGS: Find Rapidly OTU with Galaxy Solution
- Demo-B2 : Interrogation de bases de données biologiques publiques par reformulation de requêtes et classement des résultats avec ConQuR-Bio
- Demo-C1 : Proteome data mining using ProteINSIDE online tool
- Demo-C2 : Finding and analysing microRNAs in plant genomes with miRkwood

15h10 - 16h10 - **Session Affiches / Posters (even poster numbers / posters nombres pairs)**

16h10 - 16h40 - Pause-Café petits fours sucrés / Coffee Break

Session Plénière V - Amphithéâtre/Amphitheater

Méthodologies pour l'analyse des séquences et des données omiques
Methodologies for Sequence Analysis & Omics Data

Chair: Pierre PETERLONGO (INRIA, France)

16h40 - 17h20 Keynote-06 **Anthony COX** - Illumina Cambridge Ltd, Essex, UK
The power of populations: methods that scale to thousands of human genomes

17h20 - 17h40 Proc-06 **Aymeric Antoine-Lorquin** – IRISA UMR 6074, Inria et Université de Rennes1, France
Comparison of the targets obtained by a scoring matrix and by a regular expression. Application to the search for LXR binding sites

17h40 - 18h00 Proc-07 **Mathilde Le Boudic-Jamin** - IRISA UMR 6074 et Université de Rennes1, France
De novo detection of structure repeats in Proteins

18h00 - 18h20 High-06 **Morgane Thomas-Cholier** - Institut de Biologie de l'Ecole Normale Supérieure de Paris, France
ExoProfiler: a motif-based approach to analyse ChIP-exo signal

18h20 - 18h40 High-07 **Nicolas Terrapon** – UMR 7257 CNRS Aix-Marseille Université, France
How does the human gut microbiota breakdown complex polysaccharides?

18h40 - 19h10 *annonces diverses (2x15 min)*
- **Valérie POLONAIS** - *IUT Aurillac*
- **Gisèle BRONNER** - Master Bioinformatique Université Blaise Pascal

19h30 - 22h00 **Cocktail Polydome – Espace *Entre 2 Villes***

Mercredi 8 Juillet 2015 - Wednesday, July 8th 2015

Session Plénière VI - Amphithéâtre/Amphitheater

Méta-omiques & Génomique Environnementale
Meta-omics & Environnemental Genomics

Chair: Eric PELLETIER (CEA / Genoscope, France)

08h10 - 08h50 Keynote-07 **Rob FINN -** European Bioinformatics Institute (EMBL-EBI)**,** UK
Towards understanding the functional and taxonomic repertoire of a metagenome

08h50 - 09h10 Proc-08 **Guy Perrière** - UMR CNRS 5558, Université Claude Bernard - Lyon 1, France
RecStat: a set of R utilities for coding DNA sequences prediction in metagenomes

09h10 - 09h30 Proc-09 **Faouzi Jaziri** – Université d'Auvergne EA 4678 CIDAM, France
Propositional Logic for Efficient Microbial Community Assessment by DNA Microarray Data Analysis

Session Plénière VII - Amphithéâtre/Amphitheater

Biologie Translationnelle & Pharmacogénomique
Translational Biology & Pharmacogenomics

Chair: Nathalie RIVIERE (Biogemma, France)

09h30 - 10h10 Keynote-08 **Ana CONESA** – Centro de Investigaciones Príncipe Felipe, Spain
Lessons and results of integrative multi-omics data analysis

10h10 - 10h30 High-08 **Jacques Colinge** – IRCM Inserm U1194, University of Montpellier, France
Computational Analysis of Promiscuous Drug Protein Target Spectra

10h30 - 11h00 - Pause-Café viennoiseries / Coffee Break

Session Plénière IVb - Amphitéâtre/Amphitheater

Spéciales Conférences Invités
Special Invited speakers

11h00 - 11h40 Keynote-09 **Génétique & Génétique des Populations -** *Genetics & Population Genetics*
Chair: Dominique JOLY (CNRS, France)

Laurence Moreau - INRA, France
New prospects and challenges raised by the implementation of Genome-wide association mapping and Genomic Selection in plants. Illustrations in maizer

Parallel Sessions - Démos/Demos

	Session D	Session E	Session F
	Salles Commission 11-12-13 - Niveau 2	**Espace Restauration Niveau 2**	**Amphitéâtre / Amphitheater**
Chair	**Mahendra MARIADASSOU**	**Morgane THOMAS-CHOLLIER**	**Valentin LOUX**
	Méthodologies pour l'analyse des séquences et des données omiques	*Méthodologies pour l'analyse des séquences et des données omiques*	*Services, Ressources & Infrastructures pour la Bioinformatique*
11h40 - 12h00	**Demo-D1** Magali Berland	**Demo-E1** Marie-Laure Franchinard	**Demo-F1** Laurent Jourdren
		Génétique & Génétique des Populations	
12h00 - 12h20	**Demo-D2** Bérénice Batut	**Demo-E2** Lydia Ait Braham	**Demo-F2** Stéphane Télétchéa
		Services, Ressources & Infrastructures pour la Bioinformatique	
12h20 - 12h40	**Demo-D3** Alexandra Louis	**Demo-E3** Benoit Bely	**Demo-F3** Delphine Steinbach

- Demo-D1 : Demo of metagenomic data analysis: from reads to biomarkers with METEOR and MetaOMineR
- Demo-D2 : ASAIM: an intuitive and adjustable pipeline to process metatranscriptomic data from intestinal microbiota
- Demo-D3 : Genomicus: fast and intuitive comparative genomics in eukaryotes
- Demo-E1 : View and synchronize several genotypes using IGV
- Demo-E2 : BioMercator : A complete framework to integrate QTL, meta-QTL, genome annotation and genome-wide association studies
- Demo-E3 : How UniProtKB/TrEMBL Tackles High Redundant Proteomes
- Demo-F1 : Eoulsan 2: Facilitating expansion to new NGS tools and execution platforms
- Demo-F2 : DockNmine, a web portal to compare virtual and experimental interaction data
- Demo-F3 : GnpIS-Asso : a new generic tool for managing and exploiting genetic association studies results using high throughput genotyping and phenotyping data

12h40 - 13h30 – Déjeuner - Ticket Vert / Lunch - Green Ticket

13h30 - 14h30 – **Session Affiches (Affiches nombres impairs / odd poster numbers)**

14h30 - 15h30 ***Assemblée Générale de la Société Française de Bioinformatique - SFBI***

15h30 - 16h00 – Pause-Café petits fours sucrés / Coffee Break

Session Plénière VIII - Amphithéâtre/Amphitheater

Epigénétique & Epigénomique

Epigenetics & Epigenomics

Chairman : Emilie BRASSET (GReD, Université d'Auvergne, France)

16h00 - 16h40 Keynote-10 **Christophe GRUNAU** – IHPE, Université de Perpignan, France

Environmental and evolutionary epigenomics of non-model organisms: challenges and solutions

16h40 - 17h00 High-09 **Daniel Jost** – Université Joseph Fourier, UMR5525 CNRS, France

Interactions between chromatin states shape the epigenome nuclear organization

17h00 - 17h30 - **Présentation Recherche privée en Bioinformatique**

- **Nathalie RIVIERES** - Biogemma
- **Jean-Pierre MARTINANT** - Limagrain

17h30 : Rassemblement pour la Photo de Groupe / Group photo

18h30 : Départ en Bus / Bus departure to Vulcania

19h00 : Soirée Gala à Vulcania / Gala Dinner at Vulcania

24h00 minuit) : Retour aux Hôtels / Back to Hotels

Jeudi 9 Juillet 2015 - Thursday, July 9th 2015

Session Plénière IX - Amphithéâtre/Amphitheater

Services, Ressources & Infrastructures pour la bioinformatique
Services, Resources & Infrastructures for Bioinformatics

Chair: Christophe BLANCHET (IFB, France)

09h00 - 09h40 Keynote-11 **Katharina HUBER** - Computational Biology Laboratory at the University of East Anglia, UK
Dealing with big evolutionary data

09h40 - 10h10 Proc-10 **Imene Boudellioua** – King Abdullah University of Science and Technology, Kingdom of Saudi Arabia
Association Rule Mining for Metabolic Pathway Prediction

10h10 - 10h30 Proc-11 **Nicolas Daccord** – UMR8071 CNRS, Université Evry-Val d'Essonne IBGBI, France
TopoIBase: a comprehensive database dedicated to type IA DNA-topoisomerases

10h30 - 11h00 - Pause-Café viennoiseries / Coffee Break

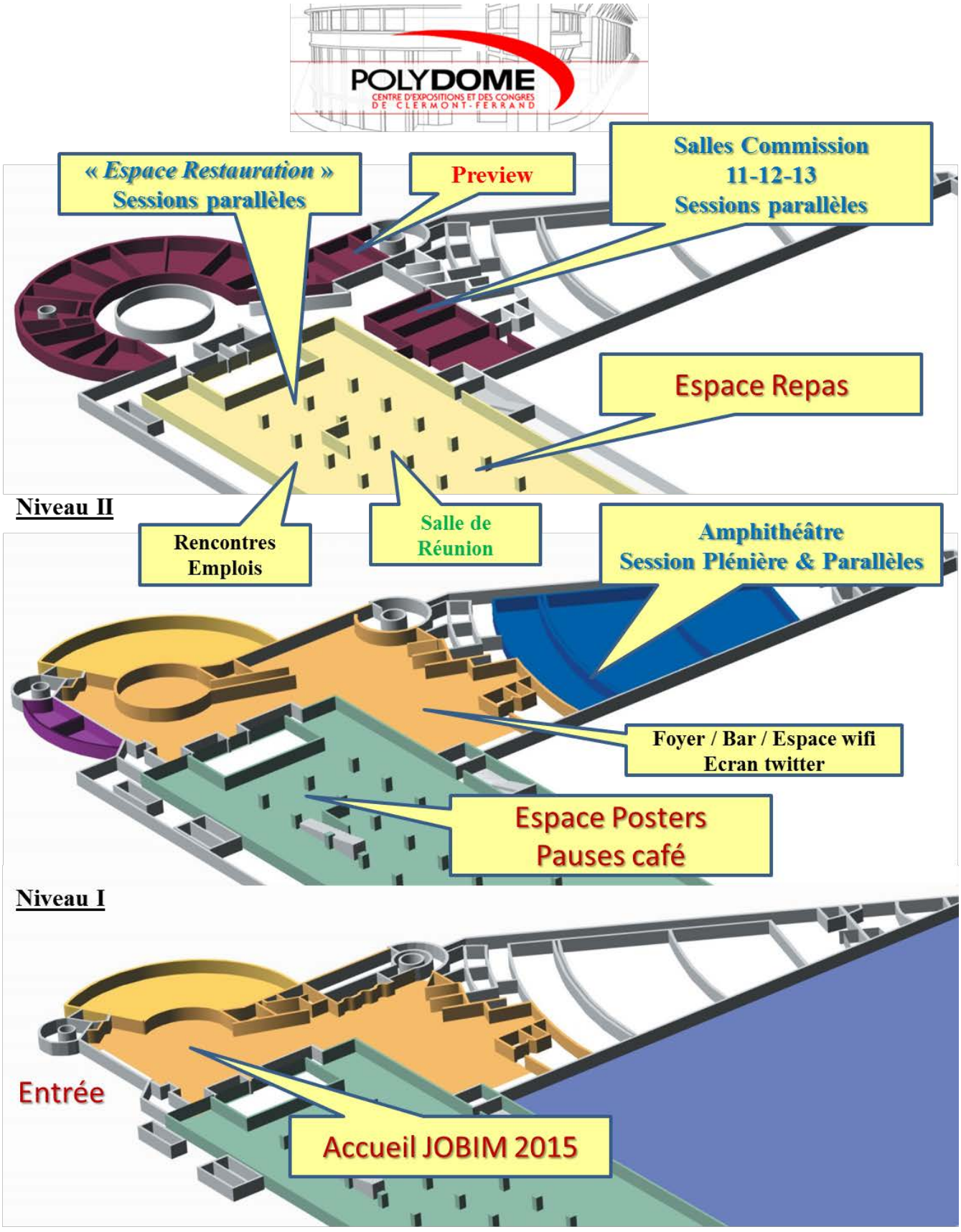
11h00 - 11h20 High-10 **Vincent Henry** – STATSARRAY / CISMeF, TIBS, LITIS EA 4108, CHU et Université de Rouen, France
OMICtools: A workflow for multi-omic data analysis

11h20 - 11h40 High-11 **Mélanie Pétéra** - PFEM, UMR1019 INRA, France
Workflow4Metabolomics: A collaborative research infrastructure for computational metabolomics

11h40 - 12h30
- Remise des récompenses par **S. Schbath** & **M-F Sagot**
  - Meilleure présentation orale
  - Deux meilleurs posters
- Annonce JOBIM 2016 par **F. Picard**
- Clôture par **P. Peyret**

12h30 – Distribution des Lunch-Box - Ticket Rouge (à réserver le jour de l'accueil) / Lunch box distribution - Red Ticket (to be reserved at the frontdesk the first day)

PLAN POLYDOME



SOMMAIRE

- Avant-Propos -

- Contributions -

Session I

BIOCHIMIE, BIOLOGIE STRUCTURALE & BIOINFORMATIQUE STRUCTURALE

BIOCHEMISTRY, STRUCTURAL BIOLOGY & STRUCTURAL BIOINFORMATICS

Thématique - Topic 02

Conférence Invitée
Keynote Lecture

CHRISTINE ORENGO

University College London, UK

A Structural Perspective on the Evolution of Protein Functions

Powerful tools for comparing protein structures and protein sequences have allowed us to analyse proteins from more than 7000 completed genomes and identify 2700 evolutionary domain superfamilies. These superfamilies cover nearly 70% of domains from all kingdoms of life and are captured in our resource (CATH-Gene3D). More detailed analyses of the highly populated superfamilies, accounting for nearly two thirds of all known domains, identified some particularly promiscuous superfamilies that can be traced back to the last universal common ancestor (LUCA) and in which relatives can diverge considerably to acquire modified structures and functions. Some structural frameworks seem particularly suited to supporting diverse residue arrangements in the active sites, and considerable structural variations on the surfaces of the domains. We also find a surprising number of examples of convergent evolution within superfamilies where very different catalytic machineries are associated with similar enzymatic chemistries, showing that these scaffolds enable multiple routes to the same function. Phylogenetic analyses of protein families can also yield insights into evolution of novel chemistries or substrate specificities and functional analyses can be combined with thermodynamic analyses to reveal the energetic considerations associated with functional divergence.

Accelerating Protein Structure Prediction using Particle Swarm Optimization on GPU

Hamed KHAKZAD[1], **Yasaman KARAMI**[1] and Seyed Shahriar ARAB[2]

[1] Computer Engineering Department, Iran University of Science and Technology, Tehran, IRAN
[2] Tarbiat Modares University, Tehran, IRAN

Corresponding author: h_khakzad@comp.iust.ac.ir

Protein tertiary structure prediction (PSP) is one of the most challenging problems in bioinformatics. Different methods have been introduced to solve this problem so far, but PSP is computationally intensive and belongs to the NP-hard class. One of the best solutions to accelerate PSP is the use of a massively parallel processing architecture, such graphical processing unit (GPU), which is used to parallelize computational algorithms. In this paper, we have proposed a parallel architecture to accelerate PSP. A bio-inspired method, particle swarm optimization (PSO) has been used as the optimization method to solve PSP. We have also performed a comprehensive study on implementing different topologies of PSO on GPU to consider the acceleration rate. Our solution belongs to ab-initio category which is based on the dihedral angles and calculates the energy-levels to predict the tertiary structure. Indeed, we have studied the search space of a protein to find the best pair of angles that gives the minimum free energy. A profile-level knowledge-based force field based on PSI-BLAST multiple sequence alignment has been applied as a fitness function to calculate the energy values. Different topologies and variations of PSO are considered here and the experimental results show that the speedup gain using GPU is about 34 times faster than CPU implementation of the algorithm with an acceptable precision. The energy values of predicted structures confirm the robustness of the algorithm.

Link PDF

Modelling ssRNA-protein complexes at high resolution

Isaure CHAUVOT DE BEAUCHENE[1], Sjoerd DE VRIES[1] and Martin ZACHARIAS[1]

[1] Physics Dep., Technical University Munich, James-Franck Str. 1, 85748 Garching, Germany

Corresponding author: zacharias@tum.de

RNA-protein specific binding underlies a large variety of fundamental cellular processes. An atomistic description of such binding processes would aid the rational conception of pharmaceutical modulators of those functions. However, while the field of protein-protein docking has achieved considerable improvements in the last decade, protein-RNA docking encounters specific difficulties. This is mainly due to the flexibility and the large conformational space of RNAs compared to proteins, and especially single-stranded RNAs (ssRNAs). Here, we present a novel fragment-based approach to tackle this problem, capable of accurate prediction of the structure of a ssRNA bound to a protein, starting from the structure of the protein and the sequence of the RNA. As a proof-of-principle, we tested the method on one complex containing a ssRNA of uniform sequence. Without any information on specific contacts or the RNA structure, our method permitted to define accurately the binding site on the protein with 10 Å precision, through the use of a comprehensive fragment library. Moreover, the bound conformation of the ssRNA could be sampled with ~1.5 Å RMSD on heavy atoms, a precision never reached so far. In future research, the method will be tested on more cases, and extended to dock ssRNA of arbitrary sequence

Link PDF

EC-PSI: Associating Enzyme Commission Numbers with Pfam Domains

Seyed Ziaeddin ALBORZI[1,2], Marie-Dominique DEVIGNES[3] and David W. RITCHIE[2]

[1] Université de Lorraine, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, 54506, France
[2] INRIA, Villers-lès-Nancy, 54600, France
[3] CNRS, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, 54506, France

Corresponding author: dave.ritchie@inria.fr

With the growing number of protein structures in the protein data bank (PDB), there is a need to annotate these structures at the domain level in order to relate protein structure to protein function. Thanks to the SIFTS database, many PDB chains are now cross-referenced with Pfam domains and enzyme commission (EC) numbers. However, these annotations do not include any explicit relationship between individual Pfam domains and EC numbers. This article presents a novel statistical training-based method called EC-PSI that can automatically infer high confidence associations between EC numbers and Pfam domains directly from EC-chain associations from SIFTS and from EC-sequence associations from the SwissProt, and TrEMBL databases. By collecting and integrating these existing EC-chain/sequence annotations, our approach is able to infer a total of 8,329 direct EC-Pfam associations with an overall F-measure of 0.819 with respect to the manually curated InterPro database, which we treat here as a "gold standard" reference dataset. Thus, compared to the 1,493 EC-Pfam associations in InterPro, our approach provides a way to find over six times as many high quality EC-Pfam associations completely automatically.

Link PDF

High-01 (#165) - Marco PASI - Sequence determines the structure and microsecond-scale dynamics of B-DNA and its bound cations

Sequence determines the structure and microsecond-scale dynamics of B-DNA and its bound cations

Marco PASI[1], John H MADDOCKS[2] and Richard LAVERRY[1]

[1] Bases moléculaires et Structurales des Systèmes Infectieux, CNRS UMR5086/Université Lyon I, IBCP, 7 passage du Vercors, 69367 Lyon
[2] Section de Math´ematiques, Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland

Corresponding author: marco.pasi@ibcp.fr

Paper reference: **M. Pasi, J.H. Maddocks, R. Lavery. Analyzing ion distributions around DNA: sequence-dependence of potassium ion distributions from microsecond molecular dynamics. *Nucleic Acids Res*. 43:2412-2423, 2015** - http://dx.doi.org/10.1093/nar/gkv080

Specifically binding proteins recognise their target sites in the genome both by “read-ing” the sequence and by probing the deformability of DNA; these interactions are fundamental in many important cellular processes, including DNA replication, genome organization and tran-scription regulation. The sequence dependence of the mechanical properties of DNA is at the basis of these “indirect recognition” processes. Understanding this second layer of genetic information requires extensive structural and dynamical knowledge of B-DNA, which is currently unavailable from experiment. We analysed a large database of microsecond-scale MD simulations of a set of B-DNA oligomers with sequences designed to allow the comprehensive study of base-sequence effects, resulting from the collaborative effort of an international consortium of laboratories. Our results elucidate the molecular details of how sequence affects the structure and dynamics of DNA, modulating the relative stability of its conformational substates and shaping the distributions of its tightly bound couterions, and how this in turn influences the interactions of DNA with proteins and other molecules.

Link PDF

Session II

EVOLUTION, PHYLOGÉNIE & PALÉOGÉNOMIQUE

EVOLUTION, PHYLOGENY & PALEOGENOMICS

Thématique - Topic 05

Conférence Invitée
Keynote Lecture

ZIHENG YANG

Department of Genetics, Evolution & Environment
University College London, UK

Estimation of species divergence times incorporating fossil and molecular information

Knowledge of absolute divergence times between species is extremely useful as it allows us to place species divergences in the correct geological and palaeoclimatol context. Yet molecular sequences are informative about distances only, not about times and rates individually. This confounding effect causes major difficulties and many counterintuitive results in molecular clock dating studies, such as extreme sensitivity to the prior and persistent uncertainty in time estimates even in analysis of genome-scale datasets. In this talk I will review the Bayesian method for divergence time estimation, which integrates information from the molecules and fossils, with an emphasis on the limit of estimation and impact of prior.

Evolutionary analyses support that *ASP* (Anti Sense Protein) overlapping ORF is the 10th gene of HIV-1 M pandemic group

Elodie CASSAN[1,2,3], Anne-Muriel ARIGON CHIFOLLEAU[1,2], Antoine GROSS[3] and Olivier GASCUEL[1,2]

[1] LIRMM, Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier, UMR5506 CNRS, Université de Montpellier, France
[2] IBC, Institut de Biologie Computationelle, Montpellier, France
[3] CPBS, Centre d'études d'agents Pathogènes et Biotechnologies pour la Santé, FRE3689 CNRS, Montpellier, France

Corresponding author: elodie.cassan@lirmm.fr

The existence of overlapping genes encoded by the anti-sense strand of retroviruses is an old concept, which has been supported for the last ten years by the discovery of HBZ in HTLV- 1, and recently by a series of results demonstrating the expression of the ASP protein of HIV-1. As a consequence, several research directions are followed to understand the function and expression of this protein. We use here bioinformatics analyses to study the appearance and conservation of ASP. We show that the ASP ORF is present only in sequences of HIV-1 M group and in one sequence of the SIVcpz Ptt group which is basal regarding HIV-1 M group. The creation of the ASP ORF is thus concomitant with the emergence of the HIV pandemic. As ASP is an overlapping gene (within the env gene), measuring the selection pressure acting on both strands requires specific models and methods. We propose such a method, which indicates that the presence of the ASP ORF is not just caused by the mechanical constraints induced by the sense strand (env gene), but that there is a selection pressure induced by ASP protein, thus supporting that ASP ORF could be the 10th gene of HIV-1.

Link PDF

Cophylogeny Reconstruction via an Approximate Bayesian Computation

Christian BAUDET[1,2], Béatrice DONATI[1,2,3], **Blerina SINAIMERI**[1,2], Pierluigi CRESCENZI[3], Christian GAUTIER[2], Catherine MATIAS[4] and Marie-France SAGOT[1,2]

[1] INRIA Grenoble Rhône-Alpes, 38330 Montbonnot Saint-Martin, France
[2] Université de Lyon, 69000 Lyon; Université Lyon 1; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, 69622 Villeurbanne, France
[3] Università di Firenze, Dipartimento di Sistemi e Informatica, 50134 Firenze, Italy
[4] Laboratoire Statistique et Génome, UMR CNRS 8071 & USC INRA, Université d'Evry, France

Corresponding author: blerina.sinaimeri@inria.fr

Paper reference: **C. Baudet, B. Donati, B. Sinaimeri, P. Crescenzi, C. Gautier, C. Matias, and M.-F. Sagot. Cophylogeny Reconstruction via an approximate bayesian computation, Systematic Biology, 2014** – doi 10.1093/sysbio/syu129 - http://DX.DOI:10.1093/SYSBIO/SYU129

A crucial issue in the cophylogeny reconstruction problem is that from a biological point of view, reasonable cost values for an event-based reconciliation are not easily chosen. We designed and developed an algorithm, COALA, that, for a given pair of host and parasite trees, estimates the frequency of the events based on an approximate Bayesian computation approach. The algorithm we propose on one hand provides more confidence in the set of costs to be used for a given pair of host and parasite trees, while on the other hand it allows to estimate the frequency of the events in cases where the dataset consists of trees with a large number of taxa. The software is freely available at http://COALA.GFORGE.INRIA.FR/.

Link PDF

High-03 (#34) - Gabriel V. MARKOV - The same or not the same: Lineage-specific gene expansions and homology relationships in multigene families in nematodes

The same or not the same: Lineage-specific gene expansions and homology relationships in multigene families in nematodes

Gabriel V. MARKOV[1], Praveen BASKARAN[1] and Ralf J. SOMMER[1]

[1] DEPARTMENT FOR EVOLUTIONARY BIOLOGY, Max Planck Institute for Developmental Biology, Spemannstrasse 37, 72076, Tübingen, Germany

Corresponding author: ralf.sommer@tuebingen.mpg.de

Reference paper: **Markov *et al.* The same or not the same: Lineage-specific gene expansions and homology relationships in multigene families in nematodes, 2015 -** http://dx.doi.org/10.1007/s00239-014-9651-y

Homology is a fundamental concept in comparative biology and a crucial tool for the analysis of character distribution. Introduced by Owen in 1843 in a morphological context, homology can similarly be applied to protein-coding genes. However, in molecular biology the proper distinction between orthology and paralogy was long limited by the absence of whole- genome sequencing data. By now, genome-wide sequencing allows comprehensive analyses of the homology of genes and gene families at the level of an entire phylum. Here, we analyze a manually curated dataset of more than 2000 proteins from the genomes of 11 nematode species of seven different genera, including free-living and animal and plant parasites to study the principles of homology assignments in gene families. Using all sequenced species as an extensive outgroup, we specifically focus on the two model species *Caenorhabditis elegans* and *Pristionchus pacificus* and compare enzymes involved in detoxification of xenobiotics and synthesis of fatty acids. We find that only a small proportion of genes in these families are one- to-one orthologs and that their history is shaped by massive duplication events. Of a total of 349 and 528 genes from *C. elegans* and *P. pacificus*, respectively, only 39 are one-to-one orthologs. Thus, frequent amplifications and losses are a widespread phenomenon in nematode lineages. We also report variation in birth and death rates depending on gene families and nematode lineages. The near absence of one-to-one orthology in related organisms has important consequences regarding the transfer of functional annotations from one species to another. Concerning our enzyme dataset, we argue that broad assignation of general types of catalytic activities seems reasonable at least in some functionally homogenous families. However, the high level of lineage-specific duplications makes it impossible to precisely infer the structure of metabolic pathways from *P. pacificus* just based on genomic comparisons with enzyme-encoding genes that are functionally characterized in *C. elegans*. This implies that additional types of data, such as metabolic profiles, have to be generated and new bioinformatic tools have to be implemented to further integrate those genomic and metabolomic data in a fully unified framework.

Link PDF

Session III

ORGANISATION & EXPRESSION DES GENOMES

ORGANIZATION & GENOME EXPRESSION

Thématique - Topic 01

Conférence Invitée
Keynote Lecture

STEPHANE ROMBAUTS

VIB, Belgium

The next challenge: sequencing, assembling and annotation of complex genomes with NGS

New Generation Sequencing (NGS) unleashed the genomes of numerous organisms, including plants like pinus, and wheat with large and complex genomes. But also the genomes of pet-organisms for which earlier, no sufficient budgets could ever been collected.

The access, thanks to technological advances with NGS, has given us insight into genomes with features never seen before. With genes imbricated in each other, genes in introns of others, compactness of the genomes leaving a priori no space for regulatory element and thus questioning our understanding of how a genome “looks like”, with all the consequences at gene prediction level.

On the other hand, NGS lead to endeavors that were doomed to fail. The short reads, no matter how much coverage generated, are not suited for large, repeat-rich genomes like those of some plants. Luckily technology advances so fast, bringing new approaches, but also new challenges.

Proc-05 (#73) - Jérémy TOURNAYRE - Fat&MuscleDB: A database to understand tissue growth processes contributing to body or muscle composition

Fat&MuscleDB: A database to understand tissue growth processes contributing to body or muscle composition

Jérémy TOURNAYRE[1,2], Isabelle CASSAR-MALEK[1,2], Matthieu REICHSTADT[1,2], Brigitte PICARD[1,2], Nicolas KASPRIC[1,2] and Muriel BONNET[1,2]

[1] INRA, UMR1213 Herbivores, Site de Theix, 63122, Saint-Genès-Champanelle, France
[2] Clermont Université, VetAgro Sup, UMR1213 Herbivores, BP 10448, 63000, Clermont-Ferrand, France

Corresponding authors: jeremy.tournayre@clermont.inra.fr & muriel.bonnet@clermont.inra.fr

To minimise unnecessary redundancy in research efforts by a better use of available data, we present a web-based database and data-mining platform named Fat&MuscleDB. Genomics on muscle and adipose tissue growth has generated huge amount of data which are available in journals and in databases. Unfortunately these data are scattered on the Internet in a heterogeneous format. Thus, it is difficult to exploit them efficiently. We hypothesise that these data can allow identifying genes or proteins involved in adipose and muscle tissues development contributing to body or muscle composition, two key criteria of carcass and meat quality. Currently, Fat&MuscleDB contains genomic expression data and differential abundance data from about 100 publications and 75 GEO datasets. These data can be queried, visualised, and downloaded in different ways: the data visualisation of each reference, the search of transcripts or proteins in references, and the data aggregation based on criteria of adipose and muscle growth. The aggregation function of Fat&MuscleDB is illustrated through two questions: "What are the proteins secreted by muscles?" and "What are the transcripts and proteins involved in the growth of muscle tissue from genetic origins of bovine?".

Link PDF

High-04 (#54) - Alexandre CORMIER - Sexual dimorphism and the evolution of sex-biased gene expression in the brown alga Ectocarpus

Sexual dimorphism and the evolution of sex-biased gene expression in the brown alga *Ectocarpus*

Alexandre CORMIER[1,2], Agnieszka LIPINSKA[1,2], Rémy LUTHRINGER[1,2], Akira PETERS[3], Erwan CORRE[4], Claire GACHON[5], Mark COCK[1,2] and Susana COELHO[1,2]

[1] Sorbonne Universités, UPMC Univ Paris 06, UMR 8227, Integrative Biology of Marine Models, Station Biologique de Roscoff, CS 90074, 29688, Roscoff cedex, France
[2] CNRS, UMR 8227, Integrative Biology of Marine Models, Station Biologique de Roscoff, CS 90074, 29688, Roscoff cedex, France
[3] Bezhin Rosko, 29250 Santec, France
[4] Abims Platform, CNRS-UPMC, FR2424, Station Biologique de Roscoff, CS 90074, 29688 Roscoff, France
[5] Scottish Marine Institute, Microbial and Molecular Biology Department, Scottish Association for Marine Science, Oban, United Kingdom

Corresponding author: acormier@sb-roscoff.fr

Reference paper: **A. Cormier, A. Lipinska *et al.* Sexual dimorphism and the evolution of sex-biased gene expression in the brown alga *Ectocarpus*. Molecular Biology and Evolution, 2015** - http://dx.doi.org/10.1093/molbev/msv049

Males and females often have marked phenotypic differences, and the expression of these dissimilarities invariably involves sex differences in gene expression. Sex-biased gene expression has been well characterized in animal species, where a high proportion of the genome may be differentially regulated in males and females during development. Male-biased genes tend to evolve more rapidly than female-biased genes, implying differences in the strength of the selective forces acting on the two sexes. Analyses of sex-biased gene expression have focused on organisms that exhibit separate sexes during the diploid phase of the life cycle (diploid sexual systems), but the genetic nature of the sexual system is expected to influence the evolutionary trajectories of sex-biased genes. We analyse here the patterns of sex-biased gene expression in Ectocarpus, a brown alga with haploid sex determination (dioicy) and a low level of phenotypic sexual dimorphism. In Ectocarpus, female-biased genes were found to be evolving as rapidly as male-biased genes. Moreover, genes expressed at fertility showed faster rates of evolution than genes expressed in immature gametophytes. Both male- and female-biased genes had a greater proportion of sites experiencing positive selection, suggesting that their accelerated evolution is at least partly driven by adaptive evolution. Gene duplication appears to have played a significant role in the generation of sex-biased genes in Ectocarpus, expanding previous models that propose this mechanism for the resolution of sexual antagonism in diploid systems. The patterns of sex-biased gene expression in Ectocarpus are consistent both with predicted characteristics of UV (haploid) sexual systems and with the distinctive aspects of this organism's reproductive biology.

Link PDF

A Reference Sequence of the Bread Wheat Chromosome 3B

Frédéric CHOULET[1,2]

[1] INRA UMR1095 Genetics, Diversity and Ecophysiology of Cereals, 5 chemin de Beaulieu, 63039 Clermont-Ferrand, France
[2] University Blaise Pascal UMR1095 Genetics, Diversity and Ecophysiology of Cereals, 5 chemin de Beaulieu, 63039 Clermont-Ferrand, France

Corresponding author: frederic.choulet@clermont.inra.fr

Reference paper: **F. Choulet *et al.* Structural and Functional Partitioning of Bread Wheat Chromosome 3B. *Science* 345:1249721, 2014 -** http://dx.doi.org/10.1126/science.1249721

We produced a reference sequence of the 1-gigabase chromosome 3B of hexaploid bread wheat. By sequencing 8452 bacterial artificial chromosomes in pools, we assembled a sequence of 774 megabases carrying 5326 protein-coding genes, 1938 pseudogenes, and 85% of transposable elements. The distribution of structural and functional features along the chromosome revealed partitioning correlated with meiotic recombination. Comparative analyses indicated high wheat-specific inter- and intrachromosomal gene duplication activities that are potential sources of variability for adaption. In addition to providing a better understanding of the organization, function, and evolution of a large and polyploid genome, the availability of a high-quality sequence anchored to genetic maps will accelerate the identification of genes underlying important agronomic traits.

Link PDF

Session IVa

SPECIAL CONFERENCES INVITES

SPECIAL INVITED SPEAKERS

Conférence Invitée
Keynote Lecture

Imaging & Image Processing

Thématique - Topic 10

THOMAS WALTER

Institut Curie, France

Bioimage Informatics for Phenomics

While we have the technologies and computational tools to analyze entire genomes, transcriptomes and proteomes, the computational description of all aspects of the phenotypes resulting from this molecular basis is still lagging behind. Yet, the quantitative description of all aspects of the phenome is a prerequisite for understanding the complex genotype-phenotype relationships in living systems.

High Content Screening (HCS) allows the collection of phenotypic responses of cellular populations to perturbations, such as alterations of gene expression or drug exposure. In this presentation, I will show how we can decipher the molecular basis of fundamental biological processes by the computational analysis of HCS data. I will also show that these image data sets are formidable scientific resources that can be remined in order to increase our knowledge on other cellular processes than what they were initially designed for.

Cellular phenotyping is not only informative about gene function and the mechanism of action of drugs, but it is also disease relevant. I will present tools to automatically quantify cellular phenotypes in Hematoxylin & Eosin stained tumor tissue sections. While technically more challenging, I believe that such tools will become increasingly important in cancer research as they ideally complement omics approaches.

Conférence Invitée
Keynote Lecture

Network, Regulation & Modeling

Thématique - Topic 06

DANIEL SEGRÈ

Boston University, USA

Spatio-temporal models of metabolism in microbial communities

Metabolism, in addition to being the “engine” of every living cell, plays a major role in the cell-cell and cell-environment relations that shape the dynamics and evolution of microbial communities, e.g. by mediating competition and cross-feeding interactions between different species. Despite the increasing availability of metagenomic sequencing data for numerous microbial ecosystems, fundamental aspects of these communities, such as the unculturability of many isolates, and the conditions necessary for taxonomic or functional stability, are still poorly understood. Our lab develops mechanistic computational approaches for studying the interactions between different organisms based on the knowledge of their entire metabolic networks. In particular, we have recently built a new open source platform for the Computation of Microbial Ecosystems in Time and Space (COMETS), which combines metabolic models with diffusion equations to simulate the 3D spatio-temporal dynamics of metabolism in microbial communities. COMETS has been experimentally tested on small artificial communities, and is in principle scalable to hundreds of species in complex environments. I will discuss recent developments and challenges towards the implementation of models for complex microbiomes.

Session V

METHODOLOGIES POUR L'ANALYSE DES SEQUENCES ET DES DONNEES OMIQUES

METHODOLOGIES FOR SEQUENCE ANALYSIS & OMICS DATA

Thématique - Topic 11

Conférence Invitée
Keynote Lecture

ANTHONY COX

Illumina Cambridge ltd, UK

The power of populations: methods that scale to thousands of human genomes

The rapid evolution of sequencing technology has meant that it is now feasible to perform whole-genome sequencing on cohorts of 10000 or more human genomes. Although challenges remain, reads from an individual single human WGS sample can be accurately aligned and many classes of genetic variant can be reliably detected from them, all of which can be achieved with a computational efficiency that is viable in a high-throughput setting.

Here we will focus more on what happens once these steps are completed and the variant calls from a large cohort of individuals are available. We will discuss how such datasets can be stored and indexed, crucial similarities and differences between WGS data and array data, how common population genetic analyses can be performed efficiently and how population-scale analysis can improve the accuracy of variant calls.

Comparison of the targets obtained by a scoring matrix and by a regular expression. Application to the search for LXR binding sites

Aymeric ANTOINE-LORQUIN[1], Sandrine LAGARRIGUE[2,3], Frédéric LECERF[2,3], Jacques NICOLAS[1] and Catherine BELLEANNÉE[1]

[1] Irisa/Inria/Université de Rennes1, Campus de Beaulieu, 35042 Rennes, FRANCE
[2] INRA, UMR 598 Génétique Animale, 35000 Rennes, France
[3] Agrocampus Ouest, UMR 598 Génétique Animale, 35000 Rennes, France

Corresponding author: aymeric.antoine-lorquin@irisa.fr

In bioinformatics, it is a common task to search for new instances of a pattern built from a set of reference sequences. For the simplest and most frequent cases, patterns are represented in two ways: regular expression or scoring matrix. In the first case, the acceptance of a sequence is a binary decision. In the second case, the quality of the sequence is indicated by a score. Since both representations seem to be used indifferently in pratice, one may wonder if they have any impact on the result. Is there a best representation? What is the accurate threshold value for a scoring matrix? Allowing mutations in a regular expression is it comparable to moving the score of acceptance of a matrix? These are questions adressed in this paper, through a test case on binding site search. This study compares hits obtained with scoring matrices or by regular expressions allowing up to two substitutions. The study shows that, in our LXR study, sequences found by a scoring matrix are closer to the targeted hits than sequences found by a regular expression.

Link PDF

De novo detection of structure repeats in Proteins

Mathilde LE BOUDIC-JASMIN[1] and Rumen ANDONOV[1]

[1] IRISA UMR 6074 et Université de Rennes 1, Campus de Beaulieu, 263 avenue du Général Leclerc 35 042 RENNES cedex, France

Corresponding author: mathilde.le_leboudic-jamin@inria.fr

Almost 25% of proteins contain internal repeats; these repeats may have a major role in the protein function. Furthermore some proteins actually are the same substructure in- cluded many times, these proteins are solenoids. However, only few repeat detection programs exist. Here, we present *Kunoichi*, a simple and efficient tool for discovering protein repeats. *Kunoichi* is based on protein fragment comparison and clique detection. As first results, we show that *Kunoichi* can find different levels of repetitions and successfully identify protein tiles. *Kunoichi* is available on request from the authors.

Link PDF

ExoProfiler: a motif-based approach to analyse ChIP-exo signal

Stephan R. STARICK[1], Jonas IBN-SALEM[1,2], Marcel JURK[1], Céline HERNANDEZ[2], Michael I. LOVE[1], Ho-Ryun CHUNG[1], Martin VINGRON[1], Morgane THOMAS-CHOLLIER[2] and Sebastiaan H. MEIJSING[1]

[1] Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73 14195, Berlin, Germany.
[2] Institut de Biologie de l'Ecole Normale Supérieure, INSERM U1024, CNRS UMR8197, 75005 Paris, France

Corresponding authors: mthomas@biologie.en.fr & meijsing@molgen.mpg.de

Reference paper: **S.R. Starick, J. Ibn-Salem, M. Jurk *et al.* ChIP-exo signal associated with DNA-binding motifs provide insights into the genomic binding of the glucocorticoid receptor and cooperating transcription factors. *Genome Research*, 2015 -** http://dx.doi.org/10.1101/gr.185157.114

The classical DNA recognition sequence of the glucocorticoid receptor (GR) appears to be present at only a fraction of bound genomic regions. To identify sequences responsible for recruitment of this transcription factor (TF) to individual loci, we turned to the high-resolution ChIP-exo approach. We exploited this signal by determining footprint profiles of TF binding at single base pair resolution using ExoProfiler, a computational pipeline based on DNA binding motifs. When applied to our GR and the few available public ChIP-exo datasets, we find that ChIP-exo footprints are protein- and recognition sequence-specific signatures of genomic TF association. Furthermore, we show that ChIP-exo captures information about TFs other than the one directly targeted by the antibody in the ChIP-procedure. Consequently, the shape of the ChIP-exo footprint can be used to discriminate between direct and indirect (tethering to other DNA-bound proteins) DNA association of GR. Together, our findings indicate that the absence of classical recognition sequences can be explained by direct GR binding to a broader spectrum of sequences than previously known, either as homodimer, or as a heterodimer binding together with a member of the ETS or TEAD families of TFs, or alternatively by indirect recruitment via FOX or STAT proteins. ChIP-exo footprints also bring structural insights and locate DNA:protein cross-link points that are compatible with crystal structures of the studied TFs. Overall, our generically applicable footprint-based approach uncovers new structural and functional insights into the diverse ways of genomic cooperation and association of TFs.

Link PDF

How does the human gut microbiota breakdown complex polysaccharides?

Nicolas TERRAPON[1], Vincent LOMBARD[1], Harry J. GILBERT[2] and Bernard HENRISSAT[1]

[1] ARCHITECTURE ET FONCTION DES MACROMOLÉCULES BIOLOGIQUES, UMR 7257 CNRS Aix-Marseille Université, 163 Avenue de Luminy, 13288 Marseille, France

[2] INSTITUTE FOR CELL AND MOLECULAR BIOSCIENCES, The Medical School of Newcastle University, NE2 4HH Newcastle Upon Tyne, United Kingdom

Corresponding author: terrapon@afmb.univ-mrs.fr

Reference paper: **N. Terrapon *et al.* Automatic prediction of polysaccharide utilization loci in Bacteroidetes species. *Bioinformatics*, 2015 -** http://dx.doi.org/10.1093/bioinformatics/btu716

A bacterial polysaccharide utilization locus (PUL) is a set of physically-linked genes in supra-operons that orchestrate the breakdown of a specific polysaccharide. PULs are prevalent in the Bacteroidetes phylum and are keys to the digestion of complex polysaccharides, notably by the human gut microbiota. A Bacteroidetes genome can encode dozens of different PULs whose boundaries and precise gene content are difficult to predict. We recently published a fully-automated approach for PUL prediction, using as input the genomic-context and domain annotation alone. By combining the detection of a pair of marker genes with operon prediction using intergenic distances, and detection of carbohydrate-active enzymes and regulators, our predictor achieved above 86% sensitivity in two Bacteroidetes species with extensive experimental PUL characterization. PUL predictions in 69 Bacteroidetes genomes are presented in our database accessible at www.cazy.org/PULDB/.

Link PDF

Session VI

META-OMIQUES & GENOMIQUE ENVIRONNEMENTALE

META-OMICS & ENVIRONNEMENTAL GENOMICS

Thématique - Topic 07

Conférence Invitée
Keynote Lecture

ROB FINN

European Bioinformatics Institute (EMBL – EBI) UK

Towards understanding the functional and taxonomic repertoire of a metagenome

In this talk, I will outline the rapidly developing field of metagenomics and some of the different applications of the technique. I will then describe the EBI (European Bioinformatics institute) metagenomics analysis portal, a free to use community resource for the archiving and interrogation of metagenomic data. While some systems perform assembly of the DNA sequence reads, this analysis platform uses the InterPro database to determine the functional potential of the metagenome. The current pipeline still uses 16S ribosomal RNA for taxonomic characterization of the bacteria present, alternative approaches are needed to characterize both viruses and eukaryotic DNA that are present. I will present some preliminary results on an alternative system that we have been investigating, which permits the linking of taxonomy and function. Finally, I will present our some of the latest updates to the EBI metagenomics website, which permits the basic comparison of samples from a metagenomic project.

RecStat: a set of R utilities for coding DNA sequences prediction in metagenomes

Olivier CLERC[1], Gwennaële FICHANT[2] and **Guy PERRIERE**[1]

[1] Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, Université Claude Bernard - Lyon 1, 43 bd. du 11 Novembre 1918, 69622 Villeurbanne Cedex, France
[2] Laboratoire de Microbiologie et Génétique Moléculaires, UMR CNRS 5100, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex 9, France

Corresponding author: guy.perriere@univ-lyon1.fr

While coding DNA sequences prediction in complete prokaryotic genomes is a relatively easy task, this is not the case for metagenomic data, especially when the sequencing methods used produce short reads. Indeed, the main algorithms used for coding sequences prediction are based on statistical learning, meaning that the first step of any analysis requires the availability of a set of known genes in order to train the program. As metagenomes may contain sequences that come from hundreds or even thousands of different organisms, it is usually not possible to apply those learning algorithms to this kind of data. In that context, we reimplemented in R the RecStat algorithm, an intrinsic approach initially aimed at the prediction of coding sequences in eukaryotes. This approach is based on the use of correspondence analysis, a multivariate method that has been frequently used in genomics for studying codon usage or codon biases in various organisms. We tested this method on a simulated metagenome and compared its performances to those obtained with three other programs. It appears that RecStat shows, on average, better results than the other programs. It shows a lower sensitivity but a much better specificity, while all the four programs have a similar precision.

Link PDF

Propositional Logic for Efficient Microbial Community Assessment by DNA Microarray Data Analysis

Faouzi JAZIRI[1,2], Céline RIBIÈRE[1], Cyrielle GASC[1], Nicolas PARISOT[1], Réjane BEUGNOT[1], Clémence DEFOIS[1], Owen K. APPADOO[3], Eric PEYRETAILLADE[1], David R.C. HILL[2] and Pierre PEYRET[1]

[1] EA 4678 CIDAM, Université d'Auvergne, 63000 Clermont-Ferrand, France
[2] ISIMA/LIMOS, UMR 6158 CNRS, Université Blaise Pascal, 63171 Aubière, France
[3] Institut PASCAL, UMR6602 CNRS, Campus des Cézeaux 63171 Aubière, France

Corresponding authors: faouzi.jaziri@udamail.fr & pierre.peyret@udamail.fr

DNA microarrays represent high-throughput molecular tools able to study the presence, or the expression levels of several thousands of genes, combining qualitative and quantitative aspects in only one experiment. However, the analysis of DNA microarrays, with huge amount of data to process, is a complex but crucial step. In this article, we present an efficient parallel algorithm for DNA microarray data analysis. Our software uses parallel computing and the concepts of propositional logic to determine the microbial composition of a hybridized biological sample. We present its performance on real biological datasets.

Link PDF

Session VII

BIOLOGIE TRANSLATIONNELLE & PHARMACOGENOMIQUE

TRANSLATIONAL BIOLOGY & PHARMACOGENOMICS

Thématique - Topic 08

Conférence Invitée
Keynote Lecture

ANA CONESA

Professor, Dept. Microbiology and Cell Science, University of Florida, USA.
Head Genomics of Gene expression Lab, Principe Felipe Research Center, Valencia, Spain.

Lessons and results of integrative multi-omics data analysis

Next generation sequencing has speed up genome analysis and brought omics research closer to many organisms and biological scenarios. Today an increasing number of research projects propose the combined use of different omics platforms to investigate diverse aspects of genome functioning. These proposals ideally seek to provide complementary sources of molecular information that eventually can be put together to obtain systems biology models of biological processes. Hence, it is not rare anymore to find experimental designs involving the collection of genome, transcriptome, epigenome and even metabolome data on a particular system. However, standard methodologies for the integration of diverse omics data types are not yet ready and researchers frequently face post-experiment question on how to combine data of different nature, variability, and significance into an analysis routine that sheds more light than the analysis of individual datasets separately. The STATegra project has been conceived to address these problems and provide the genomics community with user-friendly tools for the integration of different omics data types. STATegra targets several sequencing based functional genomics methods, proteomics and metabolomics. A first level of results deals with experimental design and power analysis issues in the multi-omics context. Furthermore, the project has develop statistical methods for explorative analysis of multi-omics datasets, inferring transcriptional networks, defining gene regulatory programs, identify significant pathways based on multi-omics evidence and combine experimental data with publics datasets. These methods, together with software implementations, will be presented.

Computational Analysis of Promiscuous Drug Protein Target Spectra

Jacques COLINGE[1]

[1] IRCM Inserm U1194, University of Montpellier, ICM, Montpellier, France

Corresponding author: jacques.colinge@inserm.fr

Reference paper: **Müllner *et al.* Targeting a cell state common to triple-negative breast cancers. Molecular Systems Biology, 2015** - http://dx.doi.org/10.15252/msb.20145664

Continuous developments in the field of chemical proteomics as well as the emergence of new methods of mapping drug targets in physiological conditions (cellular thermal shift assay) provide the scientific community with unprecedented tools to understand drugs. Still, some compounds such as tyrosine kinase inhibitors interact with a rather large number of proteins thus making the interpretation of their target spectra an exciting bioinformatics challenge. In a recent paper (Müllner et al., Mol Syst Biol, 2015), we introduced a new computational method to identify targets likely to induce drug response from chemical proteomic data. This was demonstrated with midostaurin, a kinase inhibitor that we identified as specifically toxic to basal-like cells representing triple negative breast cancer cells. Detailed validation supported our findings including in a PDX. Perspectives of drug target mapping in the context of personalized medicine will be briefly discussed as well.

Link PDF

Session IVb

SPECIAL CONFERENCES INVITES

SPECIAL INVITED SPEAKERS

Conférence Invitée
Keynote Lecture

Genetics & Population Genetics

Thématique - Topic 03

LAURENCE MOREAU

INRA, France

New prospects and challenges raised by the implementation of Genome-wide association mapping and Genomic Selection in plants. Illustrations in maize

The availability of high-throughtput genotyping techniques and the possibility to directly sequence at a low cost at least part of the genome for a growing number crop species modified the approaches used to identify locus (or QTL) involved in quantitative traits and methods used to integrate marker-trait associations in breeding. In crop species, such as maize, during the last years approaches moved from QTL detection experiments in single biparental populations to approaches involving much higher level of diversity and higher number of molecular markers such as genome wide association mapping (GWA) and genomic selection (GS). After describing the evolutions in the data and approaches used in quantitative trait analysis we will present some applications on Maize and discuss the new challenges and prospects raised by these changes. Both GWA and GS rely on mixed models and their implementation combined with the large amount of data available makes it necessary to develop efficient models and algorithms to reduce computation time. In addition to the increase of molecular data available, there is also a substantial modification in the nature of this information. Some approaches such as genotyping by sequencing generate a high number of missing values that can be partly completed using imputations but this adds an additional level of uncertainty in the data that must be taken into account. Sequence data gives access to new sources of genotypic variation such as structural variations that are likely to impact phenotypic variation. All these evolutions clearly call for the development of tools and methods to analyze data but also to allow efficient integration of all these new sources of information and results.

Session VIII

EPIGENETIQUE & EPIGENOMIQUE

EPIGENETICS & EPIGENOMICS

Thématique - Topic 04

Conférence Invitée
Keynote Lecture

CHRISTOPH GRUNAU

IHPE, Univeristé de Perpignan, France

Environmental and evolutionary epigenomics of non-model organisms: challenges and solutions

Christoph GRUNAU[1,2], Julie MJ. LEPESANT[1,2], Marion AL. PICARD[1,2], Michael FREITAG[3], Hugues PARRINELLO[4], Rémi EMANS[1,2], Céline COSSEAU[1,2] and Cristian CHAPARO[1,2]

[1] Université de Perpignan Via Domitia, Perpignan, 66860 France
[2] CNRS, UMR 5244, Interactions Hôtes-Pathogènes-Environnements (IHPE), Perpignan, 66860 France
[3] Oregon State University, Corvallis, OR 97331-7305, USA
[4] CNRS, GenomiX IBiSA, Montpellier, 34094 France

Epigenetics is terntatively described as the study of heritable changes in gene expression that is not based on modifications of the DNA sequence. The role of epigenetics in adaptive evolution is a hot topic in the field. Deep-sequencing techniques have helped to overcome many of the difficulties that originated from the obligation to focus on a small number of genes and/or model species. However, studying the epigenome of non-model organisms and its modifications in response to environmental conditions poses still both conceptual and technical challenges. While the principal questions are almost trivial (Is there a difference? Is it related to an environmental change? Does it produce a phenotype with a fitness advantage?) , designing experiments and analyzing the data is far more complex. Our laboratory has been working in the last few years on a number of different invertebrates that are of economical and/or medical importance. In my presentation I would like to share our experiences with designing experimental approaches to identify the dynamic of epigenetic marks. I will also talk about the way to avoid many of the pitfalls when dealing with the analysis of deep-sequencing data for the characterization of histone modification and DNA-methylation (BS-Seq, ChIP-Seq).

Interactions between chromatin states shape the epigenome nuclear organization

Daniel JOST[1], Noelle HADDAD[2] and Cédric VAILLANT[2]

[1] LABORATOIRE TIMC-IMAG, Université Joseph Fourier, UMR5525 CNRS, Domaine de la Merci, 38706 La Tronche, Cedex, France
[2] LABORATOIRE DE PHYSIQUE, ENS Lyon, UMR5672 CNRS, 46 Allée d'Italie, 69007 Lyon, France

Corresponding author: daniel.jost@imag.fr

Reference paper: **D. Jost *et al.* Modeling epigenome folding: formation and dynamics of topologically-associated chromatin domains. *Nucleic Acids Res*. 42:9541-9549, 2014.**

Cellular differentiation occurs during the development of multicellular organisms and leads to the formation of many different tissues where gene expression is modulated without modification of the genetic information. These modulations are in part encoded by chromatin-associated proteins or biochemical tags that are set down at the chromatin level directly on DNA or on histone tails. These markers are directly or indirectly involved in the local organization and structure of the chromatin fiber, and therefore may modulate the accessibility of DNA to transcription factors or enzymatic complexes, playing a fundamental role in the transcriptional regulation of gene expression. Statistical analysis of the repartition of this epigenomic information along the chromosomes have shown that genomes of higher eukaryotes are linearly partitioned into domains of functionally distinct chromatin states. In particular, experimental evidence has shown that the pattern of chromatin markers along chromosomes is strongly correlated with the 3D chromatin organization inside the nucleus. This suggests a coupling between epigenomic information and large-scale chromatin structure. Recently, using polymer physics and numerical simulations, we showed that attractive interactions between loci of the same chromatin state might be the driving forces of the folding of chromatin inside the nucleus. Here we propose an efficient pipeline to infer the values of such epigenomic-dependent interactions from experimental measurements of pairwise contact frequencies between chromatin loci. Learning the parameters from the epigenome and contactome data of drosophila using Bayesian inference, we analyze the relation between these interactions and the corresponding regional epigenomic compositions. We build a statistical model for this relation and finally we test the predictive power of the full model. Our approach provides a general framework to improve our understanding of chromatin folding during cell-cycle and differentiation and its relation to epigenetics.

Link PDF

Session IX

SERVICES, RESSOURCES & INFRASTRUCTURES POUR LA BIOINFORMATIQUE

SERVICES, RESOURCES & INFRASTRUCTURE FOR BIOINFORMATICS

Thématique - Topic 09

Conférence Invitée
Keynote Lecture

KATHARINA HUBER

University of East Anglia, UK

Dealing with big evolutionary data

High through-put sequencing has allowed the Life Sciences to embark on exciting albeit challenging endeavors ranging from constructing the Tree of Life to the International HapMap Project and the prospect of Personalized Medicine. Hand in hand with this technological advance come large amounts of data meaning that many of the current tools for analysing it struggle due to the limitations of the computational approaches they rely upon. By focusing on big data arising in evolutionary studies, we first illustrate some of these problems and then present novel tools aimed at tackling them.

Association Rule Mining for Metabolic Pathway Prediction

Imene BOUDELLIOUA[1], Rabie SAIDI[2], Maria MARTIN[2] and Victor SOLOVIEV[1]

[1] Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences and Engineering Division (CEMSE), King Abdullah University of Science and Technology (KAUST), Thuwal, Kingdom of Saudi Arabia, 23955

[2] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge CB10 1SD, UK

Corresponding author: rsaidi@ebi.ac.uk

Prediction of chemical reactions and pathways is among the most challenging prob- lems of systems biology. In this work, we present a system that utilizes rule mining techniques to predict metabolic pathways across wide range of various prokaryotes. Our system was successfully applied to gain knowledge that can be applied to annotate protein pathways involvement in UniProtKB-TrEMBL entries. The resulting rules represent models for pathways prediction of UniProtKB-TrEMBL entries. We carried out an experimental study of the performance of the system on UniProtKB-TrEMBL reference proteome set of prokaryotes entries to demonstrate the robustness of our system. Our prediction models touched 551,418 UniProtKB-TrEMBl entries where 371,265 of them lacked any previous pathway annotations. Moreover, using cross-validation tech- nique, we found that our system achieved a very high accuracy of pathway identification of 99.99% with F-measure of 0.987, precision of 0.991, and recall of 0.982.

Link PDF

TopoIBase: a comprehensive database dedicated to type IA DNA-topoisomerases

Nicolas DACCORD[1], Eduardo COREL[1,2], Damien CORREIA[1,4], Anaïs LOUIS[1], Hélène DEBAT[3,4,5], Vladimir DARIC[4], Marc NADAL[4,5], Claudine DEVAUCHELLE[1] and Franck SAMSON[1]

[1] LaMMe, UMR8071 CNRS, 23 bvd de France, Université Evry-Val d'Essonne IBGBI, 91037 Evry, France
[2] IBPS, Equipe AIRE, UMR 7138, Université Pierre et Marie Curie, 7 quai Saint Bernard, 75252 Paris, France
[3] Université Versailles St-Quentin, 45 avenue des Etats-Unis, 78035 Versailles, France
[4] IGM, UMR 8621 CNRS, Université Paris-Sud, Bât. 409, 91405 Orsay Cedex, France
[5] Institut Jacques Monod, UMR 7592 CNRS, Université Paris Diderot, 15 rue Hélène Brion, 75205 Paris Cedex 13, France

Corresponding author: nicolas.daccord@free.fr

Topoisomerases form a complex family of proteins whose function is to maintain the DNA topology during cellular processes in which can appear the under- or over-winding of DNA (for example the forming of a DNA replication fork). Topoisomerases IA have a special place because they are conserved in all the living world and are essential for the genomes stability. They are involved in the cellular aging process and in the development of tumors.

The topoisomerase IA studies are faced to several difficulties. They are divided into different func- tional sub-families distributed in a lot of very different organisms. The role and the mode of action of topoisomerases vary depending on the sub-family as well as the interactions in which they are involved. Despite their importance, these proteins remain largely unknown. Therefore, it is interesting to build a topoisomerases IA dedicated database, annotated and validated by topoiso- merases' experts.

In order to easily visualize one or one group of topoisomerases, a database of more than 1800 IA topoisomerases has been associated with an user friendly interface. These topoisomerases were grouped into ten categories with a comparison method without alignment (VLD). The database takes into account these categories and the taxonomy. Moreover, the database allows annotations on topoisomerases. The objective of the user interface is to facilitate the browsing and to retrieve informations from the database (categories, taxa and annorations, sequences), but also to offer the use of tools on these data.

By grouping and connecting numerous informations about topoisomerases in a dedicated database, we should simplified research on these enzymes, especially on their functions by highlighting their distribution in VLD categories. It is planned to extend the scope of this database to all topoiso- merases by adding data on topoisomerase IB and II.

Link PDF

OMICtools: A workflow for multi-omic data analysis

Vincent J. HENRY[1,2,3], Anne-Sophie PÉPIN[1], OMICtools Community[4], Bruno J. GONZALES[5] and Arnaud DESFEUX[1]

[1] STATSARRAY, Sotteville-lès-Rouen, France
[2] CiSMeF, LITIS-TIBS EA 4108, Rouen University Hospital, Normandie Université, France
[3] Master de bioinformatique de Rouen, Normandie Université, Mont-Saint Aignan, France
[4] OMICTools community, names and affiliations: http://omictools.com/curators.html
[5] NéoVasc, Haute-Normandie-INSERM, ERI-28, Rouen, France

Corresponding author: vincent.henry@omictools.com

Reference paper: **V.J. Henry *et al.* MICtools: an informative directory for multi-omic data analysis. *Database*, 2015** - http://database.oxfordjournals.org/content/2014/bau069

Abstract Recent advances in 'omic' technologies have created unprecedented opportunities for biological research, but current software and database resources are extremely fragmented. OMICtools is a manually curated metadatabase that provides an overview of more than 7800 web-accessible tools related to genomics, transcriptomics, proteomics, metabolomics. All tools have been classified by omic technologies (next-generation sequencing, microarray, mass spectrometry and nuclear magnetic resonance) associated with published evaluations of tool performance. Information about each tool is derived either from a diverse set of developers, the scientific literature or from spontaneous submissions. OMICtools is expected to serve as a useful didactic resource not only for bioinformaticians but also for experimental researchers and clinicians. Database URL: http://omictools.com/

Link PDF

High-11 (#132) – Mélanie PÉTÉRA - Workflow4Metabolomics: A collaborative research infrastructure for computational metabolomics

Workflow4Metabolomics: A collaborative research infrastructure for computational metabolomics

Mélanie PETERA[1], Gildas LE CORGUILLE[2], Marion LANDI[1], Misharl MONSOOR[2], Marie TREMBLAY-FRANCO[3], Christophe DUPERIER[1], Jean-François MARTIN[3], Daniel JACOB[4], Yann GUITTON[5,6], Marie LEFEBVRE[4], Estelle Pujos GUILLOT[1], Franck GIACOMONI[1], Etienne A. THEVENOT[7] and Christophe CARON[2]

[1] PFEM, UMR1019 INRA, Centre Clermont-Ferrand-Theix, 63122 Saint Genes Champanelle, France
[2] ABiMS, FR2424 CNRS-UPMC, Station Biologique, Place Georges Teissier, 29680 Roscoff, France
[3] PF MetaToul-AXIOM, UMR 1331 Toxalim INRA, 180 chemin de Tournefeuille, 31027 Toulouse, France
[4] Metabolome Facility of Bordeaux Functional Genomics Center, INRA, IBVM, 33140 Villenave d'Ornon, France
[5] CNRS, IRISA UMR 6074, Campus de Beaulieu, 35042 Rennes, France
[6] Université de Nantes, LINA UMR 6241, 44322 Nantes, France
[7] CEA, LIST, Laboratory for Data Analysis and Smart Systems (LADIS), MetaboHUB Paris, 91191 Gif-sur-Yvette France

Corresponding authors: melanie.petera@clermont.inra.fr & etienne.thevenot@cea.fr

Paper reference: **F. Giacomoni *et al.* Workflow4Metabolomics: A collaborative research infrastructure for com- putational metabolomics. *Bioinformatics*, 2014 -** http://dx.doi.org/10.1093/bioinformatics/btu813

In the context of an emergent and fast evolving science, the development of various tools dedicated to metabolomic data processing and data analysis increased. Because metabolo- mic analyses require a variety of steps involving various disciplines from analytical chemistry to statistics and bioinformatics, it requires many skills and expertise. However, despite this abun- dance of tools, standardization is lacking in these diversity of programs, as well as infrastructure to handle and link the different steps of metabolomic analyses. We recently implemented Work- flow4Metabolomics (W4M), a collaborative online platform hosting and providing a full pipeline for metabolomics from data preprocessing to annotation including statistical analysis. It is not designed to respond to only one specific type of metabolomic analysis, but to cover a maximum range of possible approaches - as metabolomics is a complex science that can be studied through various complementary analytical techniques. Thus, more than just gathering programs, W4M provides relevant combinations of generic and specific tools, a large part of which being develo- ped and sustained by the partners providing this virtual research environment (VRE). Moreover, using Galaxy, a web-based platform technology, W4M provides modules from various sources and of various types. This platform allows hosted tools to be run and linked together via an instinc- tive and ergonomic interface, which is beneficial for both beginners and experts in metabolomics. W4M gets its strength from the collaboration of complementary teams from bioinformatics and me- tabolomics environment. Initiated by the collaboration between two platforms, it gathers today six research teams and platforms with a higher diversity in skills and expertise. It allows a continuous enrichment in the service provided, with addition of new modules and new possible workflows de- dicated to cover a large scope of the increasing needs of the metabolomic community. Moreover, the 'open-source' aspect of this platform allows to open it to new collaborators bringing specific expertise that can be highlighted and disseminated in the metabolomic community.

Link PDF

Sessions Parallèles

DEMONSTRATIONS ACADEMIQUES

ACADEMIC DEMONSTRATIONS

RESEAUX, REGULATION & MODELISATION

NETWORK, REGULATION & MODELING

Thématique - Topic 06

Demo-A1 (#158) - Alejandra MEDINA-RIVERA - Regulatory Sequence Analysis Tools (RSAT)

Regulatory Sequence Analysis Tool (RSAT)

Alejandra MEDINA-RIVERA[1], Matthieu DEFRANCE[2], Olivier SAND[3], Carl HERRMANN[4,5], Jaime A. CASTRO-MONDRAGON[4], Jeremy DELERCE[4], Sébastien JAEGER[6], Christophe BLANCHET[7], Pierre VINCENS[8], Christophe CARON[9], Daniel M. STAINES[10], Bruno CONTRERAS-MOREIRA[11,12], Marie ARTUFEL[4], Lucie CHARBONNIER-KHAMVONGSA[4], Céline HERNANDEZ[8], Denis THIEFFRY[8], Morgane THOMAS-CHOLLIER[8] and Jacques VAN HELDEN[4]

1. SickKids Research Institute, Toronto, Canada.
2. Laboratory of Cancer Epigenetics, Université Libre de Bruxelles, Belgium.
3. CNRS-UMR8199 Institut de Biologie de Lille. Génomique Intégrative et Modélisation des Maladies Métaboliques. France.
4. Technological Advances for Genomics and Clinics (TAGC), INSERM U1090, Aix-Marseille Université (AMU), France.
5. Institute of Pharmacy and Molecular Biotechnology, and Bioquant Center, University of Heidelberg, German Cancer Research Center (DKFZ), Germany.
6. Centre d'Immunologie de Marseille-Luminy (CIML), Aix-Marseille University, Centre National de la Recherche Scientifique (CNRS), UMR7280, Marseille, France.
7. CNRS, UMS 3601; Institut Français de Bioinformatique, IFB-core, France.
8. Ecole Normale Supérieure, Institut de Biologie de l'ENS, IBENS, Paris, France; CNRS, UMR 8197, France.
9. Station Biologique / Service Informatique et Bio-informatique, Roscoff, France.
10. European Bioinformatics Institute (EMBL-EBI). Wellcome Trust Genome Campus Hinxton. United Kingdom.
11. Estación Experimental de Aula Dei / CSIC, Zaragoza, Spain. **12.** Fundación ARAID, Zaragoza, Spain.

Corresponding authors: mthomas@biologie.ens.fr & Jacques.van-Helden@univ-amu.fr

RSAT (Regulatory Sequence Analysis Tools) is a modular software suite for the analysis of *cis*-regulatory elements in genome sequences. Its main applications are (i) motif discovery, appropriate to genome-wide datasets like ChIP-seq, (ii) transcription factor binding motif analysis (quality assessment, comparisons and clustering), (iii) comparative genomics, (iv) analysis of regulatory variations. Nine new programs have been recently added to the RSAT Web site, including a tool to extract sequences from a list of coordinates (*fetch-sequences from UCSC*), novel programs dedicated to the analysis of regulatory variants from GWAS or population genomics (*retrieve-variation-seq* and *variation-scan*), a program to cluster motifs and visualize the similarities as trees (*matrix-clustering*). To deal with the drastic increase of sequenced genomes, RSAT public sites have been reorganised into taxon-specific servers. The suite is well-documented with tutorials and published protocols. The software suite is available through Web sites, SOAP/WSDL (Simple Object Access Protocol/Web Services Description Language) web services, virtual machines and stand-alone programs at http://www.rsat.eu/.

[1] M. Thomas-Chollier, M. Defrance, A. Medina-Rivera, O. Sand, C. Herrmann, D. Thieffry, J. van Helden. RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Research.* 39:W86-W91, 2011.

Link PDF

Demo-A2 (#96) - Laurent BULTEAU - DINGHY: Dynamic Interactive Navigator for General Hypergraphs in Biology - A visualization tool for small metabolic networks

DINGHY: Dynamic Interactive Navigator for General Hypergraphs in Biology A visualization tool for small metabolic networks

Laurent BULTEAU[1], Alice JULIEN-LAFERRIERE[1], Vincent LACROIX[1], Delphine PARROT[1] and Marie-France SAGOT[1]

[1] Inria et LBBE, UMR5558 CNRS, Université Claude Bernard Lyon 1,43 bd du 11 novembre 1918, 69622 Villeurbanne Cedex, France

Corresponding author: laurent.bulteau@inria.fr

DINGHY is a light-weight visualization tool, in order to rapidly display small subnetworks of a larger metabolic system. A pre-processing step is necessary to parse the data, but the visualization tool itself is a simple html page which can be opened by any modern browser.

The input consists of an SBML file representing the whole system and of a list of reactions (up to ~150) which should be represented. Most of the meta-data available in the SBML file of the system can be integrated in the display (metabolite names, formulas, metabolic pathways of the reactions, etc.). The list of reactions can also include “intensity” values associated to each reaction. Furthermore, it is possible to read a list of cofactors (whose nodes will then be uncoupled in the graph), and a list of metabolites of interest (whose nodes will be highlighted). Such list may typically be obtained from a metabolomics experiment.

Dinghy is based on the d3.js library distributed under BSD license [1]. The main feature is that the user can interact with the layout; by dragging nodes, the whole neighborhood stretches to keep the graph readable. Several parameters are adjustable to adapt, e.g., to denser or sparser graphs.

We aim at providing this software freely (under the CeCILL licence), via a web-site or to be used locally. We also hope to make the program read a wide range of data types, including, hopefully, any kinds of hypergraphs.

This work was supported by the BachBerry project (No. FP7-613793)

[1] M. Bostock, V. Ogievetsky, J. Heer, *D^3 Data-Driven Documents*. IEEE Transactions on Visualization and Computer Graphics, 17(12) :2301–2309, Dec. 2011

Link PDF

META-OMIQUES & GENOMIQUE ENVIRONNEMENTALE

META-OMICS & ENVIRONNEMENTAL GENOMICS

Thématique - Topic 07

FROGS: Find Rapidly OTU with Galaxy Solution

Lucas AUER[1*], Frédéric ESCUDIE[2*], Maria BERNARD[3], Laurent CAUQUIL[4], Katia VIDAL[4], Sarah MAMAN[4], Mahendra MARIADASSOU[5], Guillermina HERNANDEZ-RAQUET[1] and Géraldine PASCAL[4]

1 Université de Toulouse; INSA, UPS, INP; LISBP, 31077 Toulouse, France
2 INRA UMR792, CNRS, UMR5504, 31400 Toulouse, France
3 INRA, MIAT, 31326 Castanet-Tolosan, France
4 INRA, UMR1313, SIGENAE, 78352 Jouy-en-Josas, France
5 INRA, INPT ENSAT, INPT ENVT, UMR1388, 31326 Castanet-Tolosan, France
6 INRA, Unité MaIAGE, 78352 Jouy-en-Josas, France
(*) These authors have contributed equally to the present work.

Corresponding author: geraldine.pascal@toulouse.inra.fr

High-throughput sequencing of 16S/18S RNA amplicons has opened new horizons in the study of microbe communities. With the sequencing at great depth the current processing pipelines struggle to run rapidly and the most effective solutions are often designed for specialists. These tools are designed to give both the abundance table of operational taxonomic units (OTUs) and their taxonomic affiliation. In this context we developed the pipeline FROGS*: «Find Rapidly OTU with Galaxy Solution»*. Developed for the Galaxy platform, FROGS was designed to be run in two modes: with or without demultiplexed sequences. A preprocessing tool merges paired sequences into contigs with flash, cleans the data with cutadapt, deletes the chimeras with UCHIME and dereplicates sequences with a home-made python script. The clusterization tool runs with SWARM that uses a local clustering threshold, not a global clustering threshold like other softwares do. This tool generates the OTU's abundance table. The affiliation tool returns taxonomic affiliation for each OTU using both RDPClassifier and NCBI Blast+ on Silva SSU 119. And finally, the postprocessing tool allows users to process this table with the user-specified filters and provides statistical results and graphical illustrations of these data. FROGS has been developed to be very fast even on large amounts of MiSeq data in using cutting-edge tools and an optimized design, also it is portable on all Galaxy platforms with a minimum of informatics and architecture dependencies. FROGS was tested on several simulated data sets. The tool has been extremely rapid, robust and highly sensitive for the detection of OTU with very few false positives compared to other pipelines widely used by the community.

Link PDF

SERVICES, RESSOURCES & INFRASTRUCTURES POUR LA BIOINFORMATIQUE

SERVICES, RESOURCES & INFRASTRUCTURES FOR BIOINFORMATICS

Thématique - Topic 09

Demo-B2 (#57) - Bryan BRANCOTTE - Interrogation de bases de données biologiques publiques par reformulation de requêtes et classement des résultats avec ConQuR-Bio

Interrogation de bases de données biologiques publiques par reformulation de requêtes et classement des résultats avec ConQuR-Bio

Bryan BRANCOTTE[1], Bastien RANCE[2], Alain DENISE[1,3] and Sarah COHEN-BOULAKIA[1,4]

[1] Laboratoire de Recherche en Informatique (LRI), CNRS UMR 8623, Univ. Paris-Sud, France
[2] HEGP, AP-HP, Paris, France; INSERM UMRS 1138, CRC, Paris-5, Paris, France
[3] Institut de Biologie Intégrative de la Cellule (I2BC), CEA, CNRS, Univ. Paris-Sud, France
[4] INRIA et Institut de Biologie Computationnelle (IBC), Montpellier

Corresponding authors: brancotte@lri.fr & cohen@lri.fr

L'analyse d'expériences bioinformatiques comprend la comparaison des nouveaux résultats obtenus aux données existantes. Durant ces trente dernières années, les scientifiques ont dû faire face à une avalanche de données, de différents types, et présentes dans une multitude de bases de données publiques. L'accès aux données publiques se fait par l'interrogation de portails (tels que le portail Entrez du NCBI) au moyen de mots clés. Cependant, deux requêtes très similaires peuvent fournir des ensembles de réponses différents conduisant l'utilisateur à devoir tester différentes reformulations de ses requêtes (termes synonymes, variantes orthographiques, abréviations).... Les résultats obtenus doivent ensuite être filtrés, comparés... En outre, chaque ensemble de résultats est classé par le portail (en utilisant le nombre d'occurrences du mot-clé dans chaque résultat). Cependant, lorsque plusieurs reformulations sont considérées, il n'est pas simple de produire un classement triant par ordre de pertinence l'ensemble des résultats recueillis séparément, d'autant que ces résultats peuvent être fournis par centaines.

Dans cette démonstration, nous présentons ConQuR-Bio (http://conqur-bio.lri.fr) qui permet aux utilisateurs d'interroger les bases de données publiques du NCBI tout en générant automatiquement toutes les reformulations possibles et fournit des réponses triées en utilisant des techniques de *consensus de classement* (ou *agrégation de classements*). Notre démonstration montrera l'intérêt de notre approche pour des requêtes biomédicales, lors de la recherche de gènes issus d'EntrezGene et impliqués dans des maladies.

Ce travail est financé en partie par le PEPS FACSIDO 2015 (Projet RankaBio). Ce travail a été publié à la conférence Data Integration in the Life Science (Lecture Note in Bioinformatics 8574, Springer, ISBN 978- 3-319-08589-0).

Link PDF

METHODOLOGIES POUR L'ANALYSE DES SEQUENCES & DES DONNEES OMIQUES

METHODOLOGIES FOR SEQUENCE ANALYSIS & OMICS DATA

Thématique - Topic 11

Proteome data mining using ProteINSIDE online tool

Nicolas KASPRIC[1,2], Matthieu REICHSTADT[1,2], Brigitte PICARD[1,2], Jérémy TOURNAYRE[1,2] and Muriel BONNET[1,2]

[1] INRA, UMR1213 Herbivores, 63122 Saint-Genès-Champanelle, France
[2] Clermont Université, VetAgro Sup, UMR1213 Herbivores, 63000, Clermont-Ferrand, France

Corresponding authors: nicolas.kaspric@clermont.inra.fr & muriel.bonnet@clermont.inra.fr

Genomic and proteomic experiments are widely acknowledged to produce a huge amount of data to be analyzed. The challenge for scientist working on ruminant species is to extract meaningful biological context for proteins or genes to generate new hypothesis. This challenge is currently difficult because of the lack of an integrative workflow that hinders the efficiency and the robustness of data mining, thus scientists must use several tools often not designed for ruminant studies.

We designed the online tool ProteINSIDE that provides an overview of the biological information stored in public databases (NCBI and UniProt) or provided by annotations according to the Gene Ontology. It also predicts proteins that are secreted to search for proteins that mediate signalization between cells or tissues, and analyses protein-protein interactions to identify proteins contributing to a process or to visualize functional pathways. Using lists of proteins or genes as a unique input, ProteINSIDE is an original all-in-one tool to mine genomic and proteomic data from cattle, sheep, goat, human, rat, and mouse. A user-friendly web interface provides access to create analyzes and view results that can be sorted online or directly downloaded as several file formats (image, pdf, tabular, spreadsheet, fasta…). ProteINSIDE's database collects and stores the biological information required to the efficient functioning of ProteINSIDE and is monthly updated to gather the last available knowledge. ProteINSIDE was successfully bench tested with 1000 protein identifiers from each species by comparison with DAVID, BioMyn, and AgBase designed for information retrieval and annotation, as well as with PrediSi, and Phobius that predict secreted proteins.

ProteINSIDE is freely available using a simple internet browser at www.proteinside.org and an example of results get using ProteINSIDE is provided on the home page of website.

Link PDF

Finding and analysing microRNAs in plant genomes with miRkwood

Isabelle GUIGON[1,2], Sylvain LEGRAND[3], Jean-Frédéric BERTHELOT[1,2], Mohcen BENMOUNAH[1] and Hélène TOUZET[1,2]

1 CRIStAL, UMR CNRS 9189 University of Lille, Cité Scientifique, 59655 Villeneuve d'Ascq cedex, France
2 Inria Lille Nord Europe, Parc scientifique de la Haute-Borne, 40 avenue Halley, 59 650 Villeneuve d'Ascq, France
3 Evolution-Ecologie-Paléontologie, UMR CNRS 8198 University of Lille, Cité Scientifique, 59655 Villeneuve d'Ascq cedex, France

Corresponding author: helene.touzet@univ-lille1.fr

MicroRNAs (miRNAs) play a crucial role in the post-transcriptional regulation of eukaryotic gene expression, in plants and animals. Many aspects of the biogenesis and evolution of miRNAs in animals and plants differ. For example, unlike miRNAs of animals, which are mainly found in introns or exons from protein coding genes, most plant miRNAs are encoded by discrete genes. Moreover, miRNAs are released from their precursors using distinct pathways in the two kingdoms. Also, miRNA precursors are more heterogeneous in plants than in animals, varying greatly in size and structure. These differences have justified dedicated approaches for miRNA gene finding. However although several prediction tools are available for metazoan genomes, the number of tools dedicated to plants is relatively limited. Considering this gap, we have developed miRkwood, a user-friendly web server specifically designed for plant miRNAs. miRkwood is able to face the diversity of plant pre-miRNAs and allows the prediction of precursors of both conserved and non-conserved miRNAs. miRkwood can deal with both full small RNA sequencing reads and short genomic sequences (up to 100 000 nt). Moreover, it offers an intuitive and comprehensive user interface to navigate in the data, as well as many export options (GFF, CSV, FASTA, ODT) to allow the user to conduct further analyses on a local computer. It is accessible at http://bioinfo.lifl.fr/mirkwood.

Link PDF

Demo-D1 (#152) - Magali BERLAND - Metagenomic data analysis: from reads to biomarkers with METEOR and MetaOMineR

Metagenomic data analysis: from reads to biomarkers with METEOR and MetaOMineR

Magali BERLAND[1], Emmanuelle LE CHATELIER[1], Edi PRIFTI[2], Nicolas PONS[1], Franck GAUTHIER[1], Amine GHOZLANE[1], Mathieu ALMEIDA[3], Pierre LEONARD[1], Jean-Michel BATTO[1], Pierre RENAUD[4] and S. Dusko EHRLICH[1,5]

[1] Metagenopolis, US1367 INRA, Domaine de Vilvert, 78352 Jouy-en-Josas CEDEX, France
[2] Institute of Cardiometabolism and Nutrition (ICAN), Paris, France
[3] Department of Computer Science, Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA
[4] MICALIS, UMR 1319 INRA, 78352 Jouy-en-Josas, France
[5] King's College London, Centre for Host-Microbiome Interactions, Dental Institute Central Office, Guy's Hospital, United Kingdom

Corresponding authors: emmanuelle.lechatelier@jouy.inra.fr & nicolas.pons@jouy.inra.fr

Metagenomic data analysis faces a critical challenge: the unprecedented amount of data generated by new sequencing technologies requires unified and user-friendly tools for data management and analysis. We present here the first ready-to-run pipeline able to deal with millions of reads and to mine in a huge amount of sparse variables of unknown interdependency. This pipeline is composed of:

- *METEOR*, a software suite dedicated to the primary processing of sequencing data (Illumina, SOLiD or Ion Proton) for quantitative metagenomic applications. The primary processing is composed of several modules for (i) metagenomic data indexation, (ii) quality controls (cleaning and filtering), (iii) mapping the reads and counting the genes in very large reference catalogues (millions of genes) and (iv) aggregation of counting profiles of hundreds of samples. METEOR comes with helpful tools for the integration of new reference catalogues, the management of sequencing data and the creation of running workflows.
- *MetaOMineR*, a R package allowing laptop analysis of the large counting profiles generated by METEOR. Normalization and down-sampling routines reduce technical variability between samples. Dimension reduction can be achieved by clustering, projecting the data on MetaGenomic Species (MGS) [1] or different filtering procedures. A variety of statistical routines allow to identify genes, MGS and functions associated to a given trait or phenotype.

This pipeline handles many important issues with the analyses of metagenomics data and has played an important role in successful projects [2-4]. It offers the possibility to identify important and promising biomarkers in the quest for understanding complex ecosystems and treating human disease.

[1] H.B. Nielsen, M. Almeida, *et al*. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature biotechnology*, 32:822-828, 2014.
[2] E. Le Chatelier, *et al*. Richness of human gut microbiome correlates with metabolic markers. *Nature*, 500:541-546, 2013.
[3] A. Cotillard, *et al*. (2013). Dietary intervention impact on gut microbial gene richness. *Nature*, 500:585-588, 2013.
[4] N. Qin, *et al*. (2014). Alterations of the human gut microbiome in liver cirrhosis. *Nature*, 513:59-64, 2014.

Link PDF

Demo-D2 (#111) - Bérénice BATUT - ASAIM: an intuitive and adjustable pipeline to process metatranscriptomic data from intestinal microbiota

ASAIM: an intuitive and adjustable pipeline to process metatranscriptomic data from intestinal microbiota

Bérénice BATUT[1], Clémence DEFOIS[1], Céline RIBIERE[1], Cyrielle GASC[1], Jean-François BRUGIERE[1], Eric PEYRETAILLADE[1], CPER consortium Environnement Digestif and Pierre PEYRET[1]

[1] EA 4678 CIDAM, Université d'Auvergne, BP 10448, 63001 Clermont-Ferrand, France

Corresponding authors: berenice.batut@udamail.fr & pierre.peyret@udamail.fr

We present here ASAIM (Auvergne Sequence Analysis of Intestinal Microbiota), a pipeline to process metatranscriptomic data from intestinal microbiota. This pipeline is structured in four modules. The first one deals with data pretreatment: quality control, complexity reduction with similar read clusterisation, RNA sorting and read assembly. In second and third modules, RNAs are taxonomically and functionally assigned using databases dedicated to intestinal microbiota. Taxonomic, functional and abundance information is combined, treated and analyzed in last module.

The pipeline, implemented in Python, is built to be adjustable in treatment and tools. It has been tested on metatranscriptomic data from human intestinal microbiota. With its documentation and tutorials, ASAIM is freely available at http://g2im.u-clermont1.fr/asaim/ and can be downloaded or directly interrogated through the web interface.

[1] M.M Leimena, *et al.* A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets. *BMC Genomics* **14:**530, 2013.

[2] G. Xu, *et al.* RNA CoMPASS: A Dual Approach for Pathogen and Host Transcriptome Analysis of RNA-Seq Datasets. *PLoS ONE* **9:**e89445, 2014.

Link PDF

Genomicus: fast and intuitive comparative genomics in eukaryotes

Alexandra LOUIS[1,2,3], Nga Thi Thuy NGUYEN[1,2,3], Matthieu MUFFATO[4] and Hugues Roest CROLLIUS[1,2,3]

[1] Ecole Normale Supérieure, Institut de Biologie de l'ENS, IBENS, Paris, 75005 France
[2] Inserm, U1024, Paris, 75005 France.
[3] CNRS, UMR 8197, Paris, 75005 France.
[4] European Molecular Biology Laboratory - European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton Cambridge CB10 1SD, UK

Corresponding author: alexandra.louis@biologie.ens.fr

Since 2010, the Genomicus web server [1,2] is available online at http://genomicus.biologie.ens.fr.

This genome visualization tool allows comparative genomics in four different phyla (Vertebrate, Plants, Fungi, and non- vertebrate Metazoan). It provides access to genomic information from extant species, as well as ancestral gene content and order for vertebrates and flowering plants.

As comparative genomics combined with phylogenetic reconstructions are powerful approaches to study the evolution of genes and genomes, recent new features have been developed to integrate and synthetize comparative genome data in a fast and intuitive way.

In this tutorial we will show how whole genome duplications, gene duplications, gain and loss, chromosomal rearrangements, synteny conservation at all scales and over any phylogenetic distances can be studied.

Genomicus makes it easier to observe new patterns by exploring genomic data, and to answer questions by removing a large burden of data processing and transformation usually associated with comparative genomics.

[1] A. Louis, F. Murat, J. Salse, H.R. Crollius, H. GenomicusPlants: A Web Resource to Study Genome Evolution in Flowering Plants. *Plant Cell Physiol.* 56:e4, 2015.
[2] A. Louis, N.T. Nguyen, M. Muffato, H.R. Crollius. Genomicus update 2015: KaryoView and MatrixView provide a genome-wide perspective to multispecies comparative genomics. *Nucleic acids research* 43:D682-689, 2015.

Link PDF

View and synchronize several genotypes with IGV

Marie-Laure FRANCHINARD[1], Frédéric SAPET[2], Sandra DEROZIER[1], Franck SAMSON[1] and Jean-François GIBRAT[1]

[1] INRA, UR1404 Mathématiques et Informatique Appliquées du Génome à l'Environnement, 78352 Jouy-en-Josas, France

[2] BIOGEMMA - Upstream Genomics Group, Centre de Recherche de Chappes, CS 90126, 63720 Chappes, France

Corresponding author: Marie-Laure.Franchinard@jouy.inra.fr

IGV (Integrative Genomics Viewer) [1] is a very efficient genome browser written in Java, allowing users to visualize and explore a large variety of genomic data types, but limited to a single genome. However, to perform comparative genomic studies, it is very useful to be able to observe different types of data simultaneously on several genotypes.

As part of the BioDataCloud project (Investments for the Future Initiative), a collaboration between the INRA Migale platform and the Biogemma company was established to tackle this issue. According to the technical specifications set by Biogemma, a new feature has been added to IGV that allows users to jump to a new genotype from different types of data (genes, regions in genomic sequence, genetic markers) selected by the user on the reference genome. This jump results in the opening of a new IGV window on these data, if they are available for the new genotype. This window retains all IGV features and synchronizes simultaneously with the main window. The number of jumps achievable and therefore the number of simultaneously observable genotypes is unlimited and depends only upon the available hardware capabilities and the availability of the corresponding data. All jumps can be saved in an IGV session file allowing users to quickly restore already used genotypes and data or to share them with other.

With this new feature, the user can now compare different genotypes with the reference genome and navigate between them synchronously while keeping the IGV performance.

[1] H. Thorvaldsdóttir, J.T. Robinson and J.P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* 14:178-192, 2013.

Link PDF

GÉNÉTIQUE & GÉNÉTIQUE DES POPULATIONS

GENETICS & POPULATION GENETICS

Thématique - Topic 03

Demo-E2 (#136) - Lydia AIT BRAHAM - BioMercator: A complete framework to integrate QTL, meta-QTL, genome annotation and genome-wide association studies

BioMercator: A complete framework to integrate QTL, meta-QTL, genome annotation and genome-wide association studies

Lydia AIT BRAHAM[1], Yannick DE OLIVIERA[1], Olivier SOSNOWSKI[1] and Johann JOETS[1]

[1] Génétique Quantitative et Evolution – Le Moulon, UMR320 INRA, Ferme du Moulon, 91190 Gif-sur-Yvette, France

Corresponding author: joets@moulon.inra.fr

Compilation of genetic maps combined to QTL meta-analysis has proven to be a powerful approach contributing to the identification of candidate genes underlying quantitative traits. One of the most interesting properties of meta-QTL (or consensus QTL) is its confidence interval (IC) often shorter than IC of corresponding QTLs, decreasing the number of candidate gene to consider. As map compilation and QTL meta-analysis do not rely on genotyping raw data or trait measure, they can be easily achieved even if user holds maps from the literature or genetic databases.

BioMercator was the first software offering a complete set of algorithms and visualization tool covering all steps required to perform QTL meta-analysis. The fourth version of BioMercator proposes additional methods and improves graphical representation of large datasets. In this version, user may import sequence and genome annotations datasets within the software in order to display and mine functional annotation related to QTL and meta-QTL.

In order to improve candidate genes detection, we aim to include genetic association approach in the release of BioMercator. Association genetics allow to build a relationship between molecular polymorphism and phenotypic variation so, Genome-Wide Association Studies (GWAS) present a good potential for QTL's sharpening. We integrated GWAS results in Biomercator and provided new functionalities to display and exploit them.

BioMercator V4 is freely available from: http://moulon.inra.fr/biomercator and Biomercator V5 will be available soon.

Link PDF

SERVICES, RESSOURCES & INFRASTRUCTURES POUR LA BIOINFORMATIQUE

SERVICES, RESOURCES & INFRASTRUCTURES FOR BIOINFORMATICS

Thématique - Topic 09

How UniProtKB/TrEMBL Tackles High Redundant Proteomes

Benoit BELY[1], Ramona BRITTO[1], Borisas BURSTEINAS[1], Maria MARTIN[1,2,3] and UNIPROT CONSORTIUM[1,2,3]

[1] EMBL - European Bioinformatics Institute (EMBL-EBI), Welcome Trust Genome Campus, CB10 1SD Hinxton, Cambridge, United Kingdom
[2] SIB Swiss Institute of Bioinformatics (SIB), Centre Medical Universitaire, 1, rue Michel Servet, Geneva, Switzerland
[3] Protein Information Resource (PIR), Georgetown University Medical Center, 3300 Whitehaven Street NW, Washington, United States of America

Corresponding authors: benoit.bely@ebi.ac.uk & help@uniprot.org

The UniProt Knowledgebase (UniProtKB) has witnessed exponential growth in the last few years with a two-fold increase in the number of entries in 2014. This follows the vastly increased submission of multiple genomes for the same or closely related organisms. This increase was accompanied by a high level of redundancy in unreviewed UniProtKB (TrEMBL), and many sequences were over-represented in the database. This was especially true for bacterial species where different strains of the same species have been sequenced and submitted (e.g. 1,692 strains of Mycobacterium tuberculosis, corresponding to 5.97 million entries). High redundancy led to an increase in the size of UniProtKB, and thus to the amount of data to be processed internally and by our users, but also to repetitive results in BLAST searches for over-represented sequences.

To reduce this redundancy, we have developed a procedure to identify highly redundant proteomes within species groups using a combination of manual and automatic methods. We applied this procedure to bacterial proteomes (which constituted 82% of UniProtKB/TrEMBL as of release 2015_03) beginning in the 2015_04 release. Sequences corresponding to redundant proteomes (47.0 million entries) were removed from UniProtKB. From release 2015_04 on, we no longer create new UniProtKB/TrEMBL records for proteomes identified as redundant. The redundant sequences removed from UniProtKB (in 2015_04) or never added to UniProtKB (from 2015_05) are still available in the UniParc sequence archive dataset. All proteomes remain searchable through the Proteomes pages. The history (i.e. previous versions) of redundant UniProtKB records remains available.

Link PDF

Eoulsan 2: Facilitating expansion to new NGS tools and execution platforms

Sandrine PERRIN[1], Céline HERNANDEZ[2], Morgane THOMAS-CHOLLIER[2], Stéphane LE CROM[1,*] and Laurent JOURDREN[1,*]

[1] Plateforme Génomique, École normale supérieure, Institut de biologie de l'ENS, UMR8197 CNRS, U1024 INSERM, 46 rue d'Ulm, 75005, Paris, France
[2] Laboratoire Biologie computationnelle des systèmes, École normale supérieure, Institut de biologie de l'ENS, UMR8197 CNRS, U1024 INSERM, 46 rue d'Ulm, 75005, Paris, France
[*] Co-auteurs

Corresponding author: jourdren@biologie.ens.fr

Eoulsan [1] is a versatile open source framework that can reproducibly analyse huge amount of sequencing data. Especially dedicated to service platforms, this workflow manager automates the analysis of a large number of samples by using a first file containing the list of the steps to complete in the analysis (and their parameters) and a second file that describe the experimental design. Finally, it can be used on a large choice of computational infrastructure: Hadoop clusters, cloud-computing or any standard workstation. We present here Eoulsan 2 [2], a major update of our tool that enhances the original Eoulsan concepts with a new workflow manager. It allows the reuse of wrappers developed for the Galaxy platform [3], eventually enhanced with an optional link to a Docker [4] image that package the tool to execute. Moreover, besides RNA-Seq, Eoulsan 2 natively comes with many new modules especially committed to ChIP-Seq analyses. With all these new features, Eoulsan is now able to easily use and catalogue a wide choice of bioinformatic tools. Besides these improvements, in order to deploy Eoulsan on more diverse computational infrastructures, we plan to support before the end of 2015 Condor, TORQUE and other cluster schedulers (including the TGCC supercomputer centre scheduler) in addition to Eoulsan's current Hadoop features. Our framework provides an integrated and flexible solution for high throughput sequencing data analyses, from standalone workstations to clusters and cloud computing. With its modular structure and its parallel data processing, Eoulsan takes up the challenges from the massive data amount production in high throughput sequencing and brings a simple solution for reproducibility of analyses in bioinformatics.

Remerciements: Consortium « France Génomique » (ANR-10-INBS-0009).

[1] Jourdren *et al*. *Bioinformatics* 2012.
[2] http://transcriptome.ens.fr/eoulsan2
[3] Goecks *et al*. *Genome Biol*. 2010.
[4] http://docker.com

Link PDF

DockNmine: a web portal to compare virtual and experimental interaction data

Jean LETHIEC[1] and Stéphane TELETCHEA[1]

[1] UFIP, UMR 6286 CNRS, University of Nantes, 2 rue de la Houssinière, 44322 Nantes, Cedex03, France

Corresponding author: stephane.teletchea@univ-nantes.fr

Scientists have to perform multiple experiments producing qualitative and quantitative data to determine if a compound is able to bind to a given target, although their absolute value may vary significantly. Due to the large diversity of the potential ligand chemical space, the possibility of experimentally exploring a lot of compounds on a target rapidly becomes out of reach. Scientists therefore need to use virtual screening methods to determine the putative binding mode of ligands on a protein, and then post-process the raw docking experiments with a dedicated scoring function in relation with experimental data.

Two of the major difficulties for comparing docking predictions with experiments come mostly from the lack of transferability of experimental data and the lack of standardization in molecule names. Although large portals like Pubchem [1] are available for general purpose, there is no service allowing a formal expert annotation of both experimental data and docking studies. To address these issues, researchers build their own collection of data in flat files, often in spreadsheets, with limited possibilities of extensive annotations or standardization of ligand descriptions allowing cross-database retrieval.

We have conceived the dockNmine platform to provide a service allowing an expert and authenticated annotation of ligands and targets. First, this portal allows a scientist to incorporate controlled information in the database using standardized identifiers for the protein (Uniprot ID) and the ligand (SMILES description), the data and the publication associated to it. Second, it allows the incorporation of docking experiments using forms parsing automatically useful parameters and results. Last, the web interface provides a lot of pre-computed outputs to assess the degree of correlations between docking experiments and experimental data. The expert can control the visibility of his data using different level of access.

The demonstration of the dockNmine platform will be presented at JOBIM 2015.

[1] Y. Wang *et al.* PubChem's BioAssay Database. *Nucleic Acids Res.* 40:D400–D412, 2012.

Link PDF

Demo-F3 (#70) - Guillaume MERCERON - GnpIS-Asso: a new generic tool for managing and exploiting genetic association studies results using high throughput genotyping and phenotyping data

GnpIS-Asso: a new generic tool for managing and exploiting genetic association studies results using high throughput genotyping and phenotyping data

Guillaume MERCERON[1], Célia MICHOTEY[1], Raphael FLORES[1], Cyril POMMIER[1], Nacer MOHELLIBI[1], Christopher SAUVAGE[3], Mathilde CAUSSE[3], Stéphane NICOLAS[2], Alain CHARCOSSET[2], Vincent SEGURA[9], Frédéric SAPET[4], Pierre DUBREUIL[4], Jacques LEGOUIS[5], Patrice THIS[6], Catherine DELAITRE[7], Hadi QUESNEVILLE[1] and Delphine STEINBACH[1]

[1] LABORATOIRE, URGI, INRA, RD 10 Route de St Cyr 78000 Versailles, France
[2] LABORATOIRE UMR GQE-Le Moulon, INRA, 91190 Gif-sur-Yvette, France
[3] LABORATOIRE UMR GAFL, INRA, 67 allée des chênes 84143 MONTFAVET, France
[4] Biogemma compagny, Route d'Ennezat 63720 Chappes, France
[5] LABORATOIRE UMR GDEC, INRA, Site de Crouel 63000 Clermont-Ferrand, France
[6] LABORATOIRE UMR AGAP, INRA Campus Supagro Montpellier, 2 place Viala 34060 Montpellier, France
[7] LABORATOIRE UMR Agroécologie, INRA, 17 rue de Sully 21065 Dijon, France
[8] LABORATOIRE, UR AGPF, INRA 2163 avenue de la pomme de pin 45075 Orléans Cedex 2, France

Corresponding author: delphine.steinbach@versailles.inra.fr

GnpIS, http://dx.doi.org/10.1093/database/bat058 is an information system for genomic, genetic and phenomic data for plants and its bioagressors. It was extended in the frame of the French project ANR GnpAsso, whose aim was to develop in collaboration with scientists and breeders working on several species (wheat, maize, tomato, pea, grape), a complete workbench to manage and exploit results from association genetic studies (GWAS). We will present a demo of the third component, GnpIS-Asso, (D. Steinbach et al. paper in prep. 2015), available at: https://urgi.versailles.inra.fr/gnpis or directly at https://urgi.versailles.inra.fr/association.This tool allows mining into associations found between traits and markers obtained by setting several filters such as markers, traits, genetic resources or year, site, treatment or statistical model. Results are visualized graphically (Manhattan plots, Boxplots…) or can be exported in file. The IS contains today data from 2 papers on tomato and maize. Other data are in preparation on wheat, pea, rapeseed and poplar species in other projects. The tool was transferred to two breeding companies for use on their own data.

Link PDF

AFFICHES / POSTERS

PREMIER APPEL

FIRST CALL

Les résumés présentés dans cette session ont été reçus lors du premier appel à communication. Ils ont fait l'objet d'un processus de relecture par le Comité de Programme.

BIOCHIMIE, BIOLOGIE STRUCTURALE & BIOINFORMATIQUE STRUCTURALE

BIOCHEMISTRY, STRUCTURAL BIOLOGY & STRUCTURAL BIOINFORMATICS

Thématique - Topic 02

Post-001 (#45) - Ikram ALLAM - On the Interest of Semi-Supervised Approaches with Spatial Dependence in Structural Alphabet Encoding

On the Interest of Semi-Supervised Approaches with Spatial Dependence in Structural Alphabet Encoding

Ikram ALLAM[1,2,3], Delphine FLATTERS[1], Leslie REGAD[1], Anne-Claude CAMPROUX[1] and Grégory NUEL[2]

[1] MTi, UMR-S 973 INSERM, University Paris Diderot, 35 rue Hélène Brion, Paris 7, 75205, Paris Cedex 13, France
[2] LPMA PSB, UMR 7599 CNRS INSMI, University Pierre et Marie Curie, 4 Place Jussieu, Paris 6, 75252, Paris Cedex 05, France
[3] MAP5, UMR 8145 CNRS INSMI, University Paris Descartes, 45 Rue des Saints Pères, Paris 5, 75270, Paris Cedex 06, France

Corresponding author: iram.alli1@gmail.com

L'Alphabet Structural (SA) est une bibliothèque de fragments représentatifs de protéines, appelés lettres structurales (SL), obtenus par classement des fragments protéiques commun à un ensemble de conformations tridimensionnelle (3D) de protéines. Le SA est un outil très puissant pour simplifier les conformations 3D en une représentation unidimensionnelle (1D) par le codage des fragments protéiques en séquence structurale de SL. Le SA est pertinent pour l'analyse des protéines et des applications telles que la classification, l'alignement, la comparaison rapide des structures 3D, l'extraction des motifs fonctionnels, etc. Le développement d'un SA qui intègre la flexibilité de la structure 3D des protéines et de leurs cavités (poches) capable de lier les médicaments, est un défi crucial dans les domaines de «Drug Discovery» et «Drug Design». C'est actuellement en parti possible grâce aux nombre croissants de structures 3D disponible dans la protein data bank PDB (≈ 99 642 structures). Dans ce travail, nous abordons la construction des différents SA existants, basés sur des modèles de mélange, de Markov cachés (HMM) ou de classification… . Nous évaluons ensuite l'intérêt de deux concepts fondamentaux pour le développement des SA : a) la classification non supervisée versus semi-supervisée ; b) la pertinence de tenir en compte ou non de la dépendance spatiale entre les fragments de protéines. Les résultats montreront l'intérêt de l'approche semi-supervisée par rapport au non supervisé et de l'intégration de la dépendance spatiale entre les fragments, par le biais HMM, pour améliorer la précision de la classification. Cette combinaison d'approche pourrait être très prometteuse pour les applications en « Drug Design ».

Link PDF

Post-002 (#90) - Leslie REGAD - SA-conf : un outil d'analyse et de comparaison de la variabilité de séquence et de structure d'un ensemble de conformères d'une protéine

SA-conf : un outil d'analyse et de comparaison de la variabilité de séquence et de structure d'un ensemble de conformères d'une protéine

Leslie REGAD[1], Jean-Baptiste CHERON[2], Caroline SENAC[3], Delphine FLATTERS[1] and Anne-Claude CAMPROUX[1]

[1] MTi, Univ Paris Diderot, Sorbonne Paris Cité, INSERM UMRS 973, Paris, France
[2] Institut de Chimie de Nice, UMR 7272, Université de Nice-Sophia Antipolis, CNRS, Parc Valrose, 06108 Nice Cedex 2, France
[3] Sorbonne Universités, UPMC Univ Paris 06, CNRS, INSERM, Laboratoire d'Imagerie Biomédicale, 75006, Paris, France

Corresponding author: leslie.regad@univ-paris-diderot.fr

La Protein Data Bank (PDB) inclut plusieurs conformères qui sont des structures d'une même protéine obtenues dans différentes conditions expérimentales (cristallographie ou RMN, complexée ou non avec un partenaire (protéine, ligand, ADN/ARN), sous une forme mutante ou sauvage). Ces conformères présentent des variations structurales subtiles pouvant illustrer la « plasticité » de la protéine. Actuellement aucun outil n'est disponible pour effectuer une analyse fine, systématique et rapide d'un ensemble de conformères d'une protéine. Or cette analyse permettrait d'identifier les régions structurellement variables qui sont associées à la plasticité de la protéine et pourrait aider à déterminer les régions importantes pour sa fonction biologique. Dans ce contexte, nous avons développé un outil, nommé SA-conf, qui permet d'étudier et de comparer un ensemble de conformères d'une protéine et analyser leur diversité de séquences et de structures. Pour cela, SA-conf construit un alignement multiple des séquences (AMS) extraites des différents conformères. L'analyse de ce AMS permet d'identifier les séquences mutantes, celles avec des insertions/délétions ou des régions non résolues et de différencier les séquences provenant de différents organismes. Afin de proposer une comparaison entre la variabilité des séquences et de structures locales des conformères, SA-conf construit un alignement des structures locales (ASL) extraites des différents conformères grâce à l'alphabet structural HMM-SA (Camproux et al., 2004). L'analyse de ce ASL aboutit à une analyse rapide de la plasticité de la protéine en identifiant les conformères ayant des particularités structurales, les régions structurellement variables de la protéine. Cette variabilité structurale est ensuite analysée en fonction des différentes conditions expérimentales (présence d'un partenaire (ligands, protéines, etc…), présence de mutations) afin d'identifier des premiers éléments de réponse sur les raisons de cette plasticité.

Link PDF

Computational analysis of the Gibberellin signaling in moss

Anne BADEL[1,2], Juliet COATES[3], Brian C. KING[3], Henrik T. SIMONSON[4] and Olivier TABOUREAU[1,2]

[1] Univ Paris Diderot, Sorbonne Paris Cité, 75013, Paris, France
[2] MTI, INSERM UMR-S 973, 75013, Paris, France
[3] School of Biosciences, University of Birmingham, Birmingham, Great Britain
[4] Dpt of plant and environmental science, University of Copenhagen, Copenhagen, Denmark

Corresponding author: anne.badel@univ-paris-diderot.fr

Gibberellins (GAs) comprise a large family of tetracyclic, diterpenoid plant hormones that interact with GID1-DELLA receptors and play diverse biological roles in plant growth. It is assumed that bioactive GID1 receptors protein pathway does not exist in moss. However, recently, Yasumura et al (Current Biology (2007), 17:1225) identified 2 GID1-like and 2 DELLA proteins in Physcomitrella. We assumed that some Gibberellin substrates (GAs) have an effect on the moss growth by interacting with these enzymes. To evaluate from a structural point of view the reason of GA activity in moss, (i) we developed a homology model of the two GID1-like and DELLA proteins based on the X-ray structure of the complex GID1-GA3 (2ZSH, Arabidopsis thaliana) and (ii) we docked a series of GA ligand analogs on both of these models. In conclusion, from our computational study, GAs can interact in the two GID1-like pockets from moss with similar mode of binding depicted in the X-ray structure.

Link PDF

Smiles2Monomers: Tool for monomeric structure search

Yoann DUFRESNE[1,2], Laurent NOE[1,2], Valérie LECLERE[1,2,3] and Maude PUPIN[1,2]

[1] CRIStAL, UMR CNRS 9189, Univ Lille, Cité Scientifique 59650 Villeneuve d'Ascq, France
[2] Inria Lille Nord-Europe, Inria Lille Nord-Europe, 59650 Villeneuve d'Ascq, France
[3] ProBioGEM, Institut Charles Viollette, University of Lille 1, Polytech'Lille, 59650 Villeneuve d'Ascq, France

Corresponding author: yoann.dufresne@ed.univ-lille1.fr

Creating methods for discovering new active chemical compounds is an active field of research. Based on the principle "similar molecules have similar activities", several methods rely on structure similarity to predict the activities of unknown compounds. Many similarity criteria exist, mainly extracted from atomic structures. In case of macromolecules like RNAs or proteins, sequence analysis is pertinent to study their function. By extension to small biopolymers, we call monomeric composition the list of building blocks composing a polymer and its monomeric structure, the chemical bonds between them. Unfortunately, compounds databases such as PubChem or ChEBI only contain atomic structures of the molecules. We present a new method to infer the monomeric structure of polymers from their atomic structure.

Our method is based on a two-step algorithm and needs a predefine set of monomers. First, we search for all the given monomers in the studied polymers using a fast and efficient graph matching algorithm. The second step of the software tries to find a subset of the matched monomers that covers all atoms of the polymer without overlap. If gaps remain (atoms of the polymer not matched by a monomer), the graph matching is extended to allow mismatched hydrogen atoms or bonds types (single or multiple covalent bonds). This tool uses a greedy algorithm optimized by a branch and bound algorithm to perform the second step.

All the method is implemented and has been tested on annotated data from NORINE database (non-ribosomal peptide database). Test shows more than 95% of correct rediscover of known structures of this database. Moreover, with an average execution time under 1/10 second per peptide, the software is very fast.

Link PDF

Salt bridges in protein ligand interactions

Alexandre BORREL[1,2], Regad LESLIE[1], Camproux ANNE-CLAUDE[1] and Henri XHAARD[2]

[1] INSERM UMRS 973, University of Paris Diderot, France
[2] Faculty of pharmacy, University of Helsinki, Finland

Corresponding author: alexandre.borrel@helsinki.fi

In contrast to their well-studied role in protein structure, salt bridges in protein ligand complexes have received so far little attention. Here, we analyzed the environment of ligands extracted from the Protein Data Bank, concentrating on those ligands that contains well represented acidic and basic groups: primary, secondary and tertiary amines; imidazole and imidazoline; guanidium; and carboxylic acids.

As a result, filters set to sequence identity, 50%, and resolution < 3.00 Å led us to extract 4392 complexes containing 814 unique ligands. Salt bridges were evaluated by the presence of a counter ion nearby. The involvement of acid and basic groups from those ligands in salt bridges varies significantly between 20% (tertiary amines) to 61% (primary amines). The functional groups interacting with ligand's acid and basic groups were characterized in terms of nature and geometry of interactions. Interaction with metals was found to affect the distribution of molecular contacts and needed to be filtered. We constructed a smaller dataset of very high resolution complexes (less than 1.5 Å). Water molecules were found to be very present near acid and basic groups.

Link PDF

Post-006 (#142) - Seyed Ziaeddin ALBORZI - EC-PSI: Associating Enzyme Commission Numbers with Pfam Domains

EC-PSI: Associating Enzyme Commission Numbers with Pfam Domains

Seyed Ziaeddin ALBORZI[1,2], Marie-Dominique DEVIGNES[3] and David W. RITCHIE[2]

[1] Université de Lorraine, LORIA, UMR 7503, Vandœuvre-lès-Nancy, 54506, France
[2] INRIA, Villers-lès-Nancy, F-54600, France
[3] CNRS, LORIA, UMR 7503, Vandœuvre-lès-Nancy, 54506, France

Corresponding author: dave.ritchie@inria.fr

With the growing number of three-dimensional protein structures in the protein data bank (PDB), there is a need to annotate these structures at the domain level in order to relate protein structure to protein function. Thanks to the SIFTS database, many PDB chains are now cross-referenced with Pfam domains and enzyme commission (EC) numbers. However, these annotations do not include any explicit relationship between individual Pfam domains and EC numbers. This article presents a novel statistical training-based method called EC-PSI that can automatically infer high confidence associations bet- ween EC numbers and Pfam domains directly from EC-chain associations from SIFTS and from EC-sequence associations from the SwissProt, and TrEMBL databases. By collecting and integrating these existing EC-chain/sequence an- notations, our approach is able to infer a total of 8,329 direct EC-Pfam as- sociations with an overall F-measure of 0.819 with respect to the manually curated InterPro database, which we treat here as a "gold standard" reference dataset. Thus, compared to the 1,493 EC-Pfam associations in InterPro, our approach provides a way to find over six times as many high quality EC-Pfam associations completely automatically.

Link PDF

EVOLUTION, PHYLOGÉNIE & PALÉOGÉNOMIQUE

EVOLUTION, PHYLOGENY & PALEOGENOMICS

Thématique - Topic 05

Post-007 (#5) - Julien FUMEY - Tempo and mode in Astyanax mexicanus cavefish evolution: a population genomic reappraisal

Tempo and mode in *Astyanax mexicanus* cavefish evolution: a population genomic reappraisal

Julien FUMEY[1,2], Hélène HINAUX[3], Céline NOIROT[4], Sylvie RETAUX[3] and Didier CASANE[1,5]

[1] RÉSEAUX DE GÈNES, DÉVELOPPEMENT ET ÉVOLUTION, EGCE UMR 9191 CNRS, Avenue de la Terrasse, 91190 Gif-sur-Yvette, France
[2] UNIVERSITÉ PARIS SUD, Université Paris Saclay, Orsay, France
[3] DECA GROUP, Paris Saclay Institute of Neuroscience, UMR 9197 CNRS, Avenue de la Terrasse, 91190 Gif-sur-Yvette, France
[4] PLATEFORME BIOINFORMATIQUE TOULOUSE MIDI-PYRÉNÉES, UBIA, INRA, Auzeville Castanet-Tolosan, France
[5] UNIVERSITÉ PARIS DIDEROT, Sorbonne Paris Cité, Paris, France

Corresponding author: didier.casane@egce.cnrs-gif.fr

Populations of blind cavefish belonging to the Mexican tetra species *Astyanax mexicanus* are outstanding models to study evolution because the phenotypic convergence of independently-evolved and cave-adapted populations allows questioning whether the evolution of similar phenotypes involved the fixation of standing genetic variation and/or de novo mutations. In order to estimate the time frame of the evolution of the Pachón cave population which is considered as one of the "oldest" and most isolated cave population, we applied a population genomics approach. We compared polymorphisms and substitution rates in the transcriptomes of Pachón cavefish and surface fish from San Solomon Spring in Texas, using the Buenos Aires tetra (*Hyphessobrycon anisitsi*) as an outgroup. The higher polymorphism in the surface population suggests, as expected, that its effective population size is larger. We found a higher substitution rate in cavefish than in surface fish, also in accordance with a smaller cavefish population size that allowed a more rapid fixation of derived alleles present in the ancestral population, but it implies that the Pachón cave population is much "younger" than previously estimated. The comparison of these data with simulations suggests that this cavefish population has probably spent less than 30,000 years underground. This new time frame, together with other evidence, indicate that evolution of cave phenotypes mainly involved the fixation of cryptic genetic variants present in surface fish populations within a short period of time.

Link PDF

Post-008 (#29) - Dominique GUYOT - Using HMMs to Improve Sequence Clustering Method

Using HMMs to Improve Sequence Clustering Method

Dominique GUYOT[1], Simon PENEL[2], Laurent DURET[2] and Guy PERRIERE[1,2]

[1] Pôle Rhône-Alpes de Bioinformatique, Université Lyon 1, Bât. Grégor Mendel, 43 boulevard du 11 Novembre 1918, 69622 Villeurbanne Cedex, France
[2] Laboratoire de Biométrie et Biologie Evolutive, UMR 5558 CNRS, Université Lyon 1, Bât. Grégor Mendel, 43 boulevard du 11 Novembre 1918, 69622, Villeurbanne Cedex, France

Corresponding author: dominique.guyot@univ-lyon1.fr

Data clustering is a widely used method for sequence analysis in comparative genomic including molecular phylogeny. There are two major classes of algorithms: those that use *k*-mers, very fast but not very sensitive and those that use a global homology comparison, much more sensitive but slow. The exponential growth of sequences data induces a very high increase of global homologies computing time which have a quadratic complexity. Thus the construction of gene families such as those present in the HOGENOM database becomes extremely difficult even using thousands of processor cores. In this context we have developed a method that uses the two classes of algorithms to reduce the overall calculation time. Our approach proposes to use an intermediate object between the sequence and the family which is the pre-family, it is generated by a method of the type *k*-mers. We then use those objects in conjunction with a hidden Markov model states. These models induce a faster global homologies computation with more sensitivity which accelerates and improves the quality of the clustering.

Link PDF

Post-009 (#32) - Emma SAULNIER - Inferring epidemiological parameters from summary statistics of phylogenetic trees

Inferring epidemiological parameters from summary statistics of phylogenetic trees

Emma SAULNIER[1,2], Olivier GASCUEL[2] and Samuel ALIZON[1]

1 Laboratoire MIVEGEC, UMR5290 CNRS, IRD 224, Université de Montpellier, 911 avenue Agropolis, BP 64501, 34394, Montpellier, Cedex 5, France
2 Institut de Biologie Computationelle, LIRMM, UMR5506 CNRS, Université de Montpellier, 860 rue de St Priest, 34095, Montpellier, Cedex 5, France

Corresponding author: emma.saulnier@ird.fr

With the advent of affordable sequencing techniques, phylogenies are routinely generated in epidemiological studies on viral epidemics.

Recent works in the field of phylodynamics lead to the development of methods that perform epidemiological parameter inference by maximum likelihood, and which are essentially based on simple epidemiological models. Indeed, a main difficulty in these models is to write (but also compute) the likelihood function.

Approximate Bayesian Computation (ABC) allows to perform parameter inference by bypassing the computation of the likelihood function. ABC methods are based on simulation and comparison between target data and simulated data using summary statistics.

We created a flexible simulation system implementing an event-driven model of construction of transmission trees, which we assume to be equivalent to phylogenies, from epidemiological models. Then we designed 43 summary statistics that summarize the epidemiological information of phylogenies. These were chosen to be as diverse (and as numerous) as possible and involve objects such as the Lineage-Through-Time (LTT) curve (9 statistics), the branch lengths (26 statistics) or the tree topology (8 statistics).

Inference precision by our method is really close to the precision obtained using the likelihood method. This work shows that phylogenies of viral sequences and ABC can inform us on epidemiological parameters and is a first step towards the analysis of more detailed epidemiological scenarios.

Link PDF

Post-010 (#36) - Franck CERUTTI - Evolutionary dynamics of *Listeria monocytogenes* sRNA

Evolutionary dynamics of *Listeria monocytogenes* sRNA

Franck CERUTTI[1], Anaïs PAINSET[1], Annick MOISAN[1], Claire HOEDE[1], Christine GASPIN[1] and Hélène CHIAPELLO[1]

[1] Unité MIAT, UR0875 INRA, 24 chemin de Borde Rouge, Auzeville, 31326, Castanet Tolosan, France

Corresponding author: hchiapello@toulouse.inra.fr

Small regulatory RNAs (sRNAs) are ubiquitary widespread in all domains of life. In bacteria, sRNAs include crucial gene expression regulators, acting in a wide range of physiological processes. Studying evolutionary scenarios of sRNAs gains and losses is important to better understand the origin of sRNAs- mediated regulation in bacteria. The aim of this study consists in analyzing the evolutionary dynamics of *Listeria monocytogenes* EGD-e strain sRNAs among all sequenced genomes of the *Listeria* genus.

For this purpose, we first built a presence/absence matrix of *L. monocytogenes* EGD-e sRNAs among all sequenced genomes of the *Listeria* genus. A phylogenetic tree of strains was then built using a supertree approach named *Matrix Representation with Parsimony* (MRP). This tree was used as a reference in order to infer sRNAs ancestral states using Sankoff parsimony with DELTRAN optimization. Evolutionary patterns of sRNAs were then analyzed according to *Listeria* strain phenotypes and host specificities.

First results of this study show that the majority of *Listeria monocytogenes sRNAs* are conserved in all strains of the *Listeria* genus. The results obtained on some strain-specific sRNAs suggest different evolutionary scenarios depending on the type of sRNA. By considering broader evolutionary scale like the entire firmicute phylum, we will be able to better understand the evolutionary history of all *Listeria* sRNAs, including those conserved in all strains.

Link PDF

Post-011 (#38) - Gabriel V. MARKOV - Evolution of metabolic pathways in nematodes: integrating genomics and metabolomics

Evolution of metabolic pathways in nematodes: integrating genomics and metabolomics

Gabriel V. MARKOV[1], Jan M. MEYER[1], Oishika PANDA[2], Marc CLAASSEN[1], Hanh WITTE[1], Frank C. SCHROEDER[2] and Ralf J. SOMMER[1]

[1] DEPARTMENT FOR EVOLUTIONARY BIOLOGY, Max Planck Institute for Developmental Biology, Spemannstrasse 37, 72076, Tübingen, Germany
[2] BOYCE THOMPSON INSTITUTE and DEPARTMENT OF CHEMISTRY AND CHEMICAL BIOLOGY, Cornell University, Ithaca, NY 14853, USA

Corresponding author: ralf.sommer@tuebingen.mpg.de

Metabolic pathways are an important subject for evolutionary questions, because enzymes provide a direct link between genotype and phenotype. To evaluate the conservation of a given pathway between species, we have to test for the conservation of both, the enzymes and the metabolites. In addition, the catalytic activity for each step in the pathway has to be tested at the level of individual enzymes. Our detailed comparative genomic approach in nematodes indicates that there is a very high level of lineage-specific gene duplications among enzymes, with one-to-one orthologs being the exception rather than the rule. Complementary metabolomic approaches also revealed the existence of lineage-specific metabolites. Some of those specific metabolites, from the ascaroside family, are involved in a conserved biological process, the induction of a conserved life-history transition (dauer entry) in *Caenorhabditis elegans* and *Pristionchus pacificus*. Ascaroside synthesis in *C. elegans* requires *daf-22* encoding for an enzyme involved in beta-oxidation of fatty acids. We demonstrate that the shared precursors from *C. elegans* and *P. pacificus* ascarosides are not synthesized by orthologous enzymes in *P. pacificus* and *C. elegans*. The paralogous genes involved also differ in their protein domain structure. Using differential metabolomic profiling of novel CRISPR-Cas9 generated mutants for the two *P. pacificus* paralogous copies of *Cel-daf-22*, we show that the structurally more divergent paralog is more conserved at the functional level, and that the shift in substrate specificity also correlates with a difference in the length of the fatty-acid-derived lateral chain of the dauer-inducing ascarosides in both nematodes.

Link PDF

Post-012 (#61) - Magali RICHARD - Identification of fitness advantage in dynamic environment by modelisation of BAR- seq data

Identification of fitness advantage in dynamic environment by modelisation of BAR- seq data

Magali RICHARD[1], Jérôme SALIGNON[1] and Gael YVERT[1]

[1] CNRS / Ecole Normale Supérieure de Lyon, France

Corresponding author: magali.richard@ens-lyon.fr

Experimental approaches to understand the molecular mechanisms of evolution are generally restricted to environmental conditions that are standardized. This limits our comprehension of gene-environment interactions. Several studies indicate that cell-to-cell variability between genetically identical organisms can contribute to fitness in dynamic natural environments. However, such studies are sparse, and our understanding of underlying genetic mechanisms is considerably low.

How fitness varies in natural dynamic environments, how genes deal with external fluctuations, and whether variability is a natural mechanism to anticipate changes remain largely unclear. As a systematic and unbiased approach, we try to decrypt how cells cope with fluctuating environments by i) characterizing the genetic roots of adaptability to dynamic environments and the potential genotype-environment interactions and ii) studying the potential fitness benefit of cell-to-cell heterogeneity in natural contexts.

We have developed an experimental screen to identify genotypes improving proliferation in fluctuating environments. Genetically heterogeneous yeast populations are grown in competition, in various dynamic environments and quantified at several time points. All the yeast strains that we are using have a barcode tag inserted in the genome that allows us to quantify them by PCR and sequencing (BAR-seq). Custom multiplexing allows us to analyse 1,000 competing populations in a single run. To normalize our BAR-seq data, we used the method proposed in the DESeq package for RNA-seq. We are currently implementing a Generalized Linear Model method to identify strains that have a fitness advantage specifically in dynamic environment. This project is supported by grant SiGHT StG-281359 from the E.U.

Link PDF

Post-013 (#63) - Rémi PLANEL - Web tools for the phylogenetic exploration of genomes history

Web tools for the phylogenetic exploration of genomes history

Rémi PLANEL[1], Pierre DUPUIS[1], Guillaume GENCE[1], Guy PERRIERE[1,2], Vincent DAUBIN[1] and Laurent DURET[1]

[1] LABORATOIRE DE BIOMÉTRIE ET BIOLOGIE ÉVOLUTIVE, UMR5558 CNRS, Université Lyon 1, Bât. Grégor Mendel, 43 boulevard du 11 Novembre 1918, 69622, Villeurbanne Cedex, France
[2] PÔLE RHÔNE-ALPES DE BIOINFORMATIQUE, Université Lyon 1, Bât. Grégor Mendel, 16 rue Raphaël Dubois, 69622, Villeurbanne Cedex

Corresponding author: pierre.dupuis@univ-lyon1.fr

The Ancestrome project develops tools and databases that allow us to reconstruct the evolutionary history of genomes and organisms. The visualization of data requires handling complex concepts and objects such as phylogenetic trees, reconciliation of gene trees and species trees (representing the history of duplications, losses and gene transfer), or the reconstruction of ancestral chromosomes and coevolution links between genes.

We are developing a dedicated web interface that allows the user to access the data generated by the Ancestrome project and gathered for instance in the HOGENOM database: phylogenetic trees (annotated speciation events, duplication, loss and transfer), taxonomy, synteny, protein domains and sequence annotations. We develop several tools to browse phylogenetic trees and genomes (e.g. focusing on a set of taxa, a specific sequence or a topological pattern). The visualizations and interactions are developed using the SVG file format manipulated by the D3.js library with the Angular.js framework on top of that.

Link PDF

ORGANISATION & EXPRESSION DES GENOMES

ORGANIZATION & GENOME EXPRESSION

Thématique - Topic 01

Conservation of recurrent chimeras between human and mouse

Sarah DJEBALI[1], Bernardo RODRÍGUEZ MARTIN[1], Dmitri D. PERVOUCHINE[1], Alessandra BRESCHI[1], Carrie A. DAVIS[2], Julien LAGARDE[1], Alex DOBIN[2], Thomas R. GINGERAS[2] and Roderic GUIGÓ[1]

[1] Centre for Genomic Regulation and UPF, Doctor Aiguader, 88 Barcelona 08003, Spain
[2] Cold Spring Harbor Laboratory, Functional Genomics, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA

Corresponding author: sarahqd@gmail.com

The formation of chimeric transcripts (chimeras) has been widely reported [1,2,3]. Some of these reflect underlying chromosomal rearrangements [4] or are the results of the propensity of reverse transcriptase to engage in template switching [5]**,** however, a proportion of cases genuinely appear to correspond to trans-splicing of RNAs, as it has previously been described [6,7].

We used CSHL RNAseq data from 18 human and 30 mouse samples and the ChimPipe program, to identify chimeras that occur in multiple biological samples (recurrent) and are conserved between the two species, since these chimeras are likely to be transcriptionally induced and functional.

As a criterion for identifying recurrent RNA chimeras, we considered only chimeric exons supported by more than two staggered reads mapping across the junction sites in more than two cell types, and as a criterion for identifying conserved chimeras we further require that when two genes are connected in human their orthologs are also connected in mouse.

Recurrent chimeras exhibit a number of characteristic features: they tend to connect gene pairs located on the same chromosome, and when on the same chromosome they tend to be relatively near to each other (less than 100kb). These recurrent chimeras also tend to maintain an open reading frame. Notably, in the vast majority of the recurrent chimeras, the junction site involves orthologous splice junctions and thus, not only the gene-to-gene connection is conserved, but strikingly so are the specific junction sites. The genes connected in conserved chimeras tend to be involved in morphogenesis and body plan formation; consistently they tend to be detected in cell lines of embryonic origin. Finally, the genomic distance between orthologous gene pairs appears to be conserved between human and mouse.

[1] T.R. Gingeras. Implications of chimaeric non-co-linear transcripts. *Nature* 461:206-211, 2009.
[2] H. Kaessmann. Origins, evolution, and phenotypic impact of new genes. *Genome Res* 20:1313-1326, 2010.
[3] S. Djebali, S. *et al*. Evidence for transcript networks composed of chimeric RNAs in human cells. *PLoS One* 7:e28213, 2012.
[4] C.A. Maher, *et al*. Transcriptome sequencing to detect gene fusions in cancer. *Nature* 458:97-101, 2009.
[5] J. Houseley and D. Tollervey. Apparent non-canonical trans-splicing is generated by reverse transcriptase in vitro. *PLoS One* 5:e12271, 2010.
[6] H. Li, J. Wang, G. Mor, and J. Sklar. A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells. *Science* 321:1357-1361, 2008.
[7] C.S. Wu, *et al*. Integrative transcriptome sequencing identifies trans-splicing events with important roles in human embryonic stem cell pluripotency. *Genome Res* 24:25-36, 2013.

Link PDF

Post-015 (#15) - Pierre DELPECH - Analyse du transcriptome de Lactococcus garvieae durant l'inhibition de Staphylococcus aureus par RNA-seq

Analyse du transcriptome de Lactococcus garvieae durant l'inhibition de Staphylococcus aureus par RNA-seq

Pierre DELPECH[1], Etienne RIFA[1], Sabine NIDELET[3], Emeric DUBOIS[3], Geneviève GAGNE[2], Marie-Christine. MONTEL[1], Céline DELBES[1] and Stéphanie BORNES[2]

[1] UNITE DE RECHERCHES FROMAGERES (INRA), 20 côte de Reyne, 15000 Aurillac, France
[2] LABORATOIRE DE BIOLOGIE, Clermont Université (IUT Génie Biologique), 100 rue de l'égalité, 15013 Aurillac, France
[3] MGX, c/o Institut de Génomique Fonctionnelle, 141 rue de la Cardonille, 34094 Montpellier Cedex 5, France

Corresponding author: pierre.delpech@clermont.inra.fr

La croissance du pathogène Staphylococcus aureus est inhibée par Lactococcus garvieae dans les produits laitiers et in vitro. Exceptée l'implication du peroxyde d'hydrogène (H2O2) produit par L. garvieae en aération, les mécanismes cellulaires impliqués sont peu caractérisés (Delbes-Paus et al., 2010). Afin de mieux les comprendre, l'expression des gènes de L. garvieae a été déterminée après 9 h de culture in vitro (inhibition maximale). Les ARNm ont été extraits en présence ou absence de S. aureus et en aération ou non, puis séquencés par un HiSeq 2000 (Illumina, reads de 50nt). Après alignement sur les génomes de référence avec Bowtie (Langmead et al., 2009), puis comptage du nombre de reads aligné par gène avec HTSeq Count (Anders et al., 2015), trois méthodes, DESeq, DESeq2 et EdgeR (Anders and Huber, 2010; Love et al., 2014; Robinson et al., 2010), ont été confrontées pour réaliser l'analyse différentielle de l'expression des gènes. Nous avons observé que S. aureus n'induisait aucune réponse de L. garvieae pouvant expliquer cette inhibition. En revanche, l'expression de 314 gènes de L. garvieae était affectée par l'aération, principalement des gènes impliqués dans le métabolisme primaire et l'adaptation au milieu. Ces analyses ont permis de proposer deux nouvelles hypothèses pour expliquer l'inhibition de croissance de S. aureus : une compétition nutritionnelle pour la thréonine et l'excrétion de molécules potentiellement inhibitrices. Les hypothèses seront confirmées par RT-PCRq et validées biologiquement.

Link PDF

Post-016 (#16) - MARIE GARAVILLON-TOURNAYRE - High throuput workflow for RNAseq data treatment linking laboratory data server and remote parallel calculation platform

High throuput workflow for RNAseq data treatment linking laboratory data server and remote parallel calculation platform

Marie GARAVILLON-TOURNAYRE[1,3], David LOPEZ[1,3], Marion LANDI[2], Corentin HOCHART[1], Jean-Stéphane VENISSE[1,3], Boris FUMANAL[1,3], Aurélie GOUSSET-DUPONT[1,3] and Philippe LABEL[1,3]

[1] PIAF, UMR-547 Université Blaise Pascal Clermont2, 8 Avenue Blaise Pascal, TSA 60026, CS 60026, 63178 AUBIERE Cedex, France
[2] PFEM, INRA, Site de Theix, 63122, Saint Genes Champanelle, France
[3] PIAF, UMR-547, INRA, 5 chemin de Beaulieu, Site de Crouël, 63039, Clermont-Ferrand, cedex 2, France

Corresponding author: philippe.label@clermont.inra.fr

RNAseq analysis for transcriptomic spreads in biology at an increasing pace. At the same time, in the laboratory, biologists are facing challenges for data treatment where integrated software platforms, e.g. the Galaxy environment, provide mid-range throughput data treatment. To increase data treatment throughput, we developed a workflow integrating bash, perl, python and C++ tools, either open-sourced or home-made, to manage RNAseq data volumes at rates of 200 Mseq.h $^{-1}$ (millions of sequences per hour) optimizing local laboratory resource for automatic scripts generation and use of distant clustered parallel calculation platform available through high-speed network. This provides a solution allowing modular steps of data cleaning and filtering, mapping onto reference transcriptome and delivery of mapping results within the next day typically. Dataset of 3,2 Gseq, representing 300 Gb of compressed sequences were treated typically within 2 days. Half of this processing time was used for data file transfer between servers. We illustrate this workflow using poplar RNAseq data and poplar genome database reference (version 3) set up in our laboratory from public resources (Phytozome.jgi.doe.gov). Integration of local PIAFdb server (Clermont-Ferrand, France) and remote clustered calculation server Genotoul (Toulouse, France) improves workflow feasibility and data management for the biologist. Automatized tasks parallelization is the cornerstone of this workflow enabling maximization of calculation resources. Mapping results are then downloaded back on PIAFdb and datamined by the biologist through classical R procedures.

Link PDF

Post-017 (#18) - Isabelle LESUR - Local adaptation and genetic adaptive potential of forest trees

Local adaptation and genetic adaptive potential of forest trees

Isabelle LESUR[1,2], Christophe PLOMION[1] and Ivan SCOTTI[3]

[1] INRA, UMR1202, BIOGECO, F-33610, Cestas, France; Univ. Bordeaux, BIOGECO, UMR1202, 33170, Talence, France

[2] HelixVenture, Mérignac, France

[3] INRA, UR 0629, 84914, Avignon, France

Corresponding author: ilesur@pierroton.inra.fr

The FLAG ANR project aims at understanding how forest tree populations may respond to future local and global environmental changes linked to global warming. To do so, we need to understand how trees can adapt to their environment along ecological gradients. We focus here on water availability gradients in particularly sensitive forest areas, such as the Guiana shield, the Mediterranean basin, Sub-Saharan Africa and the Brazilian Cerrado where populations are concerned with seasonal limitations in soil water availability and seasonal soil hypoxia. Our study aims at investigating the complex genome-wide effects of local adaptation in eight tree species (5 gymnosperm species + 3 angiosperm species): maritime pine (*Pinus pinaster*), aleppo pine (*Pinus halepensis*), fir (*Abies alba*), cedar (*Cedrus atlantica*), larch (*Larix decidua*), beech (*Fagus sylvatica*), wapa (*Eperua falcata*) and manil (*Symphonia globulifera*).For each species, we wish: i/ to identify several thousand molecular markers (SNPs) localized in genomic area of interest, ii/ to genotype a large number of individuals among populations localized along gradients, iii/ to identify genes involve in adaptation to local environment. First, we generated a large amount of genomic and transcriptomic NGS sequences. Then, for each species, we constructed a unigene to identify SNPs in the coding part of the genome. Finally, we constructed or used genomic references required for capture of these polymorphic areas through probe design. Genotyping by sequencing specific to target regions of the genome should provide us thousands of SNPs for hundreds of individuals. These genomic approaches which require bioinformatics developments are presented in this poster. The association between SNP frequencies, phenotypic values and environmental gradients should help us to better understand micro-evolutive mechanisms implied in the adaptation of trees to their environment.

Link PDF

Sequence-driven in silico discovery of potential effectors involved in *Corynespora cassiicola* pathogenesis

David Lopez[1,2], Philippe Label[1,2], Sébastien Ribeiro[1,2], Emmanuelle Morin[3], Annegret Kohler[3], Francis Martin[3] and Valérie Pujade-Renaud[1,4]

[1] PIAF, UMR-547 University Clermont2 Blaise Pascal, 24 avenue des Landais, Campus des Cézeaux - BP 80026, 63171 Aubière, Cedex, France
[2] PIAF, UMR-547, INRA, 5 chemin de Beaulieu, Site de Crouël, 63039 Clermont- Ferrand, cedex 2, France
[3] IAM, UMR-1136, INRA, Rue d'Amance, 54280 Champenoux, France
[4] AGAP, UMR-1334, CIRAD, 03 Avenue Agropolis, 34398 Montpellier Cedex 5, France.

Corresponding author: lopez.dav@icloud.com

Corynespora cassiicola (Berk. & M. A Curtis) is an ascomycota fungus found in over 300 plant species and is mainly described as a pathogen. *C. cassiicola* causes important damages in crops (Soy, tomato, coton, etc.). In rubber tree, our study model, *C. cassiicola* is responsible for a Corynespora Leaf Fall disease, which results in decreased latex production. Pioneering work has stressed the role of cassiicolin, a secreted protein with a central role in pathogeny. Polymorphism study of cassiicolin genes of strains from different geographical origins have shown that about 60% of tested strains did not possess cassiicolin genes. This observation has led to the conclusion that other effectors must exist and was at the origin of the genome sequencing and annotation initiative by DOE Genome Institute (1000 Fungal Genome Project). We carried out an interspecific comparative genomic analysis to identify genes putatively involved in virulence and trophic modes. Proteomes of 45 fungi, essentially plant-pathogens, were analyzed in order to identify homologous proteins (ortholog and paralog clustering) and to outline evolutionary tendencies operating on related protein families (contracting / stable / expanding clusters). Proteomes were also analyzed to identify primary and secondary metabolism actors. This study was expanded by a similar comparative approach with Whole Genome Shotgun sequencing of 37 *C. cassiicola* strains. We finally searched for sequence polymorphism in putative effectors as well as genomic dynamics that occurred on chromosomal region carrying cassiicolin gene.

Link PDF

Post-019 (#35) - Marie GARAVILLON-TOURNAYRE - Transcriptome and ecophysiological data of Populus nigra genotypes during drought stress

Transcriptome and ecophysiological data of *Populus nigra* genotypes during drought stress

Marie GARAVILLON-TOURNAYRE[1,2], Aurélie DUPONT-GOUSSET[1,2], Boris FUMANAL[1,2], Jean-Stéphane VENISSE[1,2], Romain DE OLIVEIRA[1], Pierrick BENOIT[1], Benjamin ALARY[1], Adrien ROUZEAUD[1] and Philippe LABEL[1,2]

[1] PIAF, UMR-547 University Clermont2 Blaise Pascal INRA-UBP, 24 Avenue des Landais, Campus des Cézeaux – BP 80026, 63171 Aubière, Cedex, France
[2] PIAF, UMR-547, INRA, 5 chemin de Beaulieu, Site de Crouël, 63039 Clermont- Ferrand, cedex 2, France

Corresponding author: Marie.GARAVILLON@univ-bpclermont.fr

The increase in intensity and frequency of droughts excepted in the global warming context may lead to the loss of the less adapted woody species. It is therefore interesting to study the phenotypic and molecular responses emerging from the Genotype x Environment interaction during limited water availability.

The biostatistical approach presented in this poster allows (i) to characterize the phenotypes of Populus nigra in water stress through two biological issues (plasticity and ecological strategies) and (ii) to select integrative variables from tree responses which are combined with RNAseq results. The experimental plan includes two explanatory variables: treatment (2 modalities: well watered or water deficit) and genotype (6 modalities). The explained variables (133) in kinetic or point describe the hydraulic and physiologic status of trees across the whole plant and leaf. Molecular response to drought is studied by RNAseq on leaf blade using the Populus trichocarpa transcriptome as a reference. High quality of total RNA allowed the production of libraries and sequences by multiplexing and paired-end of 15 million reads by samples.

Among 8836 up-regulated and 9689 down-regulated transcripts in drought stress (FDR 0.1‰), transmembrane channels known to regulate water flux have been identified. Gene Ontology enrichment associated to biological conclusions gives an overview of regulation strategies of poplar in moderated drought stress.

Link PDF

Post-020 (#47) - Amandine CAMPAN-FOURNIER - Analyse de 32 souches de Legionella pneumophila de sévérités variables afin d'identifier les déterminants de la pathogénicité

Analyse de 32 souches de *Legionella pneumophila* de sévérités variables afin d'identifier les déterminants de la pathogénicité

Amandine CAMPAN-FOURNIER[1,2], Christine OGER[2], Christophe GINEVRA[3,4], Vincent NAVRATIL[2], Anne-Gaëlle RANC[3,4], Guy PERRIERE[1,2] and SOPHIE JARRAUD[3,4]

[1] Laboratoire de Biométrie et Biologie Evolutive, Université Lyon 1, UMR5558 CNRS, 43 bd du 11 novembre 1918, 69622 Villeurbanne Cedex, France
[2] Pôle Rhône-Alpes de Bio-Informatique, Université Lyon 1, 43 bd du 11 novembre 1918, 69622 Villeurbanne Cedex, France
[3] Hospices Civils de Lyon, Centre National de Référence des Légionelles, 59 bd Pinel, 69500 Bron, France
[4] Centre International de Recherche en Infectiologie, Université Lyon 1, UMR5308 CNRS, U1111 Inserm, Ecole Normale Supérieure de Lyon, 46 allée d'Italie, 69364 Lyon Cedex07, France

Corresponding author: amandine.fournier@univ-lyon1.fr

La légionellose, ou maladie des légionnaires, est une pathologie caractérisée par une pneumonie sévère ets'avère mortelle dans 10 à 15 % des cas en France. Cette infection est causée par une bactérie du genreLegionella, le plus souvent par l'espèce Legionella pneumophila. Des données épidémiologiques suggèrentque les souches n'ont pas toutes le même pouvoir pathogène. Afin d'identifier les déterminants du pouvoirpathogène, nous nous intéressons au génome et au transcriptome de 32 souches de L. pneumophilaimpliquées dans des pathologies de sévérités variables. Nous avons obtenu la séquence du génome de 16isolats cliniques provenant de patients décédés et de 16 isolats cliniques de patients ayant survécu àl'infection, avec la technologie Illumina (MiSeq). La méthode de typage utilisée habituellement pourdistinguer les différentes souches de l'espèce L. pneumophila consiste à séquencer 7 gènes par la techniquede Sanger. Nous proposons ici de typer les souches à partir des lectures produites par le séquençage desgénomes complets et de vérifier sa cohérence avec le typage réalisé par séquençage Sanger. Ensuite, nous présentons les résultats de la comparaison de plusieurs outils d'assemblage de novo appliqués à nos données. Les génomes assemblés seront ensuite annotés et analysés pour mettre en évidence les différences génétiques distinguant les isolats, selon la sévérité de la légionellose. Nous planifions également de séquencer le transcriptome de ces mêmes souches, afin d'étudier l'expression différentielle de leurs gènes et de pouvoir ainsi caractériser plus finement les déterminants de la pathogénicité de Legionella pneumophila.

Link PDF

Post-021 (#48) - Géraldine CAUMES - Microbial biomineralization processes as part of the environmental phosphorus cycle: deciphering the genomic diversity and evolution of candidate gene families

Microbial biomineralization processes as part of the environmental phosphorus cycle: deciphering the genomic diversity and evolution of candidate gene families

Géraldine CAUMES[1], Sara RIVAS[1], Gaëlle LELANDAIS[2], Fériel SKOURI-PANET[1], Karim BENZERARA[1] and Elodie DUPRAT[2]

[1] IMPMC, Sorbonne Universités-UPMC, CNRS UMR 7590, MNHN, IRD UR 206, 4 place Jussieu, 75005, Paris, France
[2] IJM, Université Paris Diderot, CNRS UMR 7592, Bâtiment Buffon, 15 rue Hélène Brion, 75205, Paris, Cedex 13, France

Corresponding author: elodie.duprat@upmc.fr

Phosphorus (P) is an essential element for life and a limiting nutrient in many ecosystems. The understanding of the formation of ancient and modern phosphate sedimentary deposits, in relation with the biogeochemical P cycle, constitutes a major environmental issue. Microorganisms are involved in the P cycle, by their ability of active storage or release of inorganic phosphate, then available for the biosphere or trapped by biomineralization of mineral phases of calcium or metal phosphates.

In this study, we selected a set of gene families according to their potential influence on calcium or metal phosphate biomineralization. We aimed to describe their diversity in relation with the taxonomic distribution or environmental adaptation of microbial genomes. We proceeded in the functional annotation of these genes in the complete microbial genomes available in the NCBI Genome database. We identified specific pattern of gene composition, genomic organization, and transcriptional regulatory sequences. In the context of the identification of microbial molecular processes of the P cycle, this study provides high quality reference sequence datasets of genes and regulatory elements from diverse microbial species; these data will be essential for further annotation and combined analysis of metagenomic sequences and geochemical parameters from modern phosphatogenesis environments.

This work is supported by the French national program EC2CO-MicrobiEn and by the Labex MATISSE.

Link PDF

Post-022 (#53) - Valentin LOUX - Mutations and genomic islands can explain the strain dependency of sugar utilization in 21 strains of *Propionibacterium freudenreichii*

Mutations and genomic islands can explain the strain dependency of sugar utilization in 21 strains of *Propionibacterium freudenreichii*

Valentin LOUX[3], Mahendra MARIADASSOU[3], Sintia ALMEDIA[1,2,4], Hélène CHIAPELLO[3], Amal HAMMANI[3], Julien BURATTI[3], Annie GENDRAULT[3], Valérie BARBE[5], Jean-Marc AURY[5], Stéphanie-Marie DEUTSCH[1,2], Aurélie NICOLAS[1,2], Sandrine PARAYRE[1,2], Marie-Noëlle MADEC[1,2], Victoria CHUAT[1,2], Gwenaël JAN[1,2], Pierre PETERLONGO[6], Yves LE LOIR[1,2] and Hélène FALENTIN[1,2]

1 INRA, UMR 1253, Science et Technologie du Lait et de l'Oeuf, 35000 Rennes, France
2 AGROCAMPUS OUEST, UMR1253, UMR Science et Technologie du Lait et de l'Oeuf, 35000 Rennes, France
3 INRA Mathématique Informatique et Génome, 78352 Jouy en Josas, France Institute of Biological
4 Sciences, Federal University of Minas Gerais Belo Horizonte, Minas Gerais, Brazil
5 CEA Genoscope CNRS and université d'Evry, 91006 Evry, France
6 INRIA Campus de Beaulieu- Rennes, 35042 Rennes Cedex, France

Corresponding author: helene.falentin@rennes.inra.fr

Propionibacterium freudenreichii (PF) is an actinobacterium used in cheese technology and for its probiotic properties. *PF* can grow on a variety of carbon and nitrogen sources. The aim of this work was to discover the genetic basis for strain-dependent traits related to its ability to use specific carbon sources. Twenty-one strains were sequenced. Each gene was attributed to either the core genome or an accessory genome. The ability of the 21 strains to degrade 50 different sugars was evaluated. Thirty-three sugars were degraded by none of the sequenced strains whereas eight sugars were degraded by all of them. The corresponding genes were present in the core genome. Lactose, melibiose were only used by some strains. The presence/absence of genes responsible for carbon uptake and degradation correlated well with the phenotypes and was in line the metabolic pathways described previously in other species. We also considered the genetic origin (transduction, rearrangement) of the corresponding genomic islands. Ribose and gluconate were to a greater or lesser extent (quantitative phenotype) degraded by some strains. For these sugars, the phenotypes correlated with the premature appearance of a stop codon interrupting protein synthesis, preventing the catabolism of corresponding sugar. These results illustrate (i) the power of correlation studies to discover the genetic basis of binary strain-dependent traits, and (ii) the plasticity of *PF* chromosomes, probably resulting from horizontal transfers, duplications, transpositions and an accumulation of mutations.

Link PDF

Post-023 (#102) - Théophile KAZMIERCZAK - Bioinformatic analysis of NAC transcription factor involved in senescence regulation in *Medicago*

Bioinformatic analysis of NAC transcription factor involved in senescence regulation in *Medicago*

Théophile KAZMIERCZAK[1,a], Colette GENEIX[2,a], Anne-Claude CAMPROUX[2] and Véronique GRUBER[1]

[1] Institut of Plant Sciences Paris-Saclay (IPS2), UMR 9213/UMR1403, CNRS, INRA, Université Paris-Sud, Université d'Evry, Université Paris-Diderot, Sorbonne Paris Cité, Bâtiment 630, 91405 0rsay, France

[2] INSERM, UMRS 973, MTi, Paris, France, Université Paris Diderot, Sorbonne Paris Cité, UMRS 973, MTi, Paris, France

[a] These authors contributed equally to this work.

Corresponding authors: veronique.gruber@isv.cnrs-gif.fr

Crop growth is controlled by nitrate levels in soil responsible for nitrate and nitrite pollution at high levels. Among the crops, legume roots interact with soil bacteria named rhizobia to form symbiotic nitrogen-fixing root nodules, therefore reducing external nitrate supply. This association is controlled by environmental cues as salinity and high nitrate levels that lead to the nodule uselessness and senescence. Transcriptomic analyses in *Medicago truncatula*, the legume model plant, demonstrate the importance of transcription factors involved in regulation of senescence mechanism in plants. Recently, we identified the transcription factor *MtNAC969* from the NAC (NAM/ATAF/CUC) family, which is involved in root response to salt stress and in nodule senescence. To construct a coregulated gene network with *MtNAC969* in *Medicago truncatula*, we selected candidate genes showing an expression pattern similar to *MtNAC969* using multivariate analyzes based on Principal Component Analysis and hierarchical classifications on available transcriptomes of the *Medicago truncatula* Gene Atlas website. Expression of 8 genes was validated by qRT-PCR in nitrate-treated nodules and in dark-stressed leaves of *Medicago truncatula*. This study combines closely experimental approaches and bioinformatics analysis, and contributes therefore to understand senescence mechanisms in legume senescence.

Link PDF

Post-024 (#109) - Magalie LEVEUGLE - Rapsodyn Whole Exome capture and genotyping in the polyploid *Brassica napus*

Rapsodyn Whole Exome capture and genotyping in the polyploid *Brassica napus*

Magalie LEVEUGLE[1], Hervé DUBORJAL[1], Michaela WEST[1], Nelly CUBIZOLLES[1], Alexandra POTHIER[1], Adeline CLEMENTI[1], Nathalie NESI[2], Emmanuelle DYRSZKA[3], Jean-Philippe PICHON[1] and Nathalie RIVIERE[1]

[1] BIOGEMMA, Centre de Recherche de Chappes CS 90126, 63720 Chappes, France
[2] Institute for Genetics, Environment and Plant Protection (IGEPP), UMR 1349, INRA – Agrocampus Ouest – University Rennes1 Bat. 301, Domaine de la Motte BP35327, 35650 Le Rheu, France
[3] SYNGENTA SEEDS SAS, 12 Chemin de l'hobit – B.P. 27, 31790 Saint-Sauveur, France

Corresponding author: magalie.leveugle@biogemma.com

One of the aims of the "Investissement d'Avenir" RAPSODYN project is the identification of genomic regions and genes associated to higher oil yield and better Nitrogen Use Efficiency (NUE) in rapeseed for the development of new varieties. This goal can only be attained through genotyping of markers spanning the whole genome, on large collections of genotypes to perform association studies.

In order to cover the majority of the rapeseed genes and to lower the sequencing costs, without impairing the results quality and keeping missing data low, we proposed a new assay able to combine both SNP discovery and genotyping of the 206 lines of the RAPSODYN PANEL, using exome capture technology. We used the public genome sequence of *Brassica napus* (Chalhoub et al, 2014) and a *Brassica* EST collection to design the exome capture assay (Roche Nimblegen technology) and we set up a bioinformatics pipeline to parallelize the analysis and generate a single genotyping matrix directly suitable for association studies.

After sequencing and bioinformatics analysis, more than 580 000 SNPs have been genotyped on 60% of the genes, including 400 000 with less than 10% of missing value and 180 000 without missing value for all the RAPSODYN panel.

Link PDF

Post-025 (#114) - Valérie POLONAIS - Draft genome sequence of Tubulinosema ratisbonensis, a microsporidian parasite of *Drosophila melanogaster*

Draft genome sequence of *Tubulinosema ratisbonensis*, a microsporidian parasite of *Drosophila melanogaster*

Valérie POLONAIS[1,2], Sebastian NIEHUS[5], Abdel BELKORCHIA[1,2], Ivan WAWRZYNIAK[1,2], Nicolas PARISOT[3], Adrien FRANCHET[5], Patrick WINCKER[6], Pierre PEYRET[4], Eric PEYRETAILLADE[4], Frédéric DELBAC[1,2] and Dominique FERRANDON[5]

[1] CNRS, UMR 6023, LMGE, 63171 Aubière, France.
[2] Clermont Université, Université d'Auvergne, Laboratoire "Microorganismes: Génome et Environnement", 63171 Clermont-Ferrand, France
[3] UMR 203 BF2I, Biologie Fonctionnelle Insectes et Interactions, INRA, INSA-Lyon, Université de Lyon, 69621 Villeurbanne, France
[4] Clermont Université, Université d'Auvergne, EA 4678 CIDAM, BP 10448, 63001 Clermont-Ferrand, France
[5] UPR 9022 du CNRS, Université de Strasbourg, Institut de Biologie Moléculaire et Cellulaire 67084 Strasbourg Cedex, France
[6] Genoscope (CEA) and CNRS UMR 8030, Université d'Evry, 91057 Evry, France.

Corresponding authors: valerie.polonais@udamail.fr & D.Ferrandon@ibmc.u-strasbg.fr

Microsporidia are fungi-related unicellular eukaryotes, all obligate intracellular parasites that form environmentally resistant spores. These organisms are ubiquitous in the animal kingdom parasitizing a wide range of invertebrate and vertebrate hosts including insects, crustaceans, fish, birds and most mammal groups. Nowadays, microsporidia are an important problem in beekeeping and likely contribute to the decline of honeybee colonies. They are also responsible of major economic losses in aquaculture and, as a result of the HIV epidemic, they have become a public health problem. Although numerous genomic data are now available for several microsporidian species (17 genomes sequenced), little is known on the biology of these pathogens mainly because relevant models of infection are lacking.

Here, we present first functional annotation data of *Tubulinosema ratisbonensis*, a microsporidian species infecting *Drosophila melanogaster*. Genome was sequenced using 454 pyrosequencing technology. Using a microsporidian annotation specific pipeline (MicroAnnot) coding sequences were characterized. Comparative genomic analysis was then conducted with available microsporidian genomes leading to the identification of *T. ratisbonensis* specific genes. This genome offers new insights on key proteins involved in the interactions with the well-known genetic model organism, *D. melanogaster*.

Link PDF

Post-026 (#120) - Jonathan LORENZO - Application du système GenFam à la réponse au stress des plantes : intégration de l'identification d'éléments cis spécifiques

Application du système GenFam à la réponse au stress des plantes : intégration de l'identification d'éléments cis spécifiques

Jonathan LORENZO[1,2], Delphine LARIVIERE[2], Gaëtan DROC[2], Jean-Francois DUFAYARD[2], Dominique THIS[3] and Stéphanie BOCS[2]

[1] Master 2 STIC pour la santé, Faculté des Sciences, Place Eugène Bataillon, 34095 Montpellier cedex 5, France
[2] CIRAD, UMR AGAP, Montpellier 34398, France
[3] Montpellier SupAgro, UMR AGAP, Montpellier 34060, France

Corresponding author: jonath.lorenzo@gmail.com

GenFam est un système intégratif d'analyse de familles de gènes. Ce système permet (i) de créer des familles de gènes de génomes complets, (ii) d'exécuter une analyse phylogénétique de cette famille à travers le gestionnaire de workflows Galaxy afin de définir les relations d'homologie, (iii) d'étudier des événements évolutifs à partir de blocs de synténie précalculées avec le workflow SynMap de la plateforme de génomique comparative (CoGe) et (iv) d'intégrer ces résultats dans l'interface de visualisation synthétique. La première application de GenFam est d'identifier des gènes candidats pour la tolérance aux stress environnementaux. Il nécessite de mettre en évidence la présence de séquences régulatrices cis spécifiques de la réponse aux stress (de type ABRE, DRE). Dans ce contexte, nous avons besoin d'intégrer de nouveaux outils afin de découvrir et chercher des sites de fixation de facteurs de transcription (Transcription Factor Binding Sites, TFBS) dans les séquences promotrices des gènes membre de la famille étudiée. Ce workflow Galaxy va, d'une part, sélectionner les régions flanquantes en 5' ou en 3' des gènes d'intérêts selon le choix de l'utilisateur. D'autre part, les régions flanquantes sont analysées afin de découvrir et rechercher les motifs de séquences régulatrices cis spécifiques de la réponse aux stress avec des méthodes complémentaires comme MEME, STIF, PHYME. Ces résultats ainsi que l'annotation fonctionnelle des gènes étiquetés comme étant impliqués dans la réponse au stress seront intégrés dans l'interface de visualisation. Ce travail doit permettre une réflexion sur la notion d'orthologie fonctionnelle et effectuer une recherche translationnelle depuis les espèces modèles jusqu'aux espèces d'intérêt agronomique (i.e identifier des gènes candidats pour la réponse au stress du caféier à partir d'informations fonctionnelles connues chez Arabidopsis).

Link PDF

RiboTools: A Galaxy toolbox for qualitative Ribosome Profiling analysis

Rachel LEGENDRE[1], Agnès BAUDIN-BAILLIEU[1], Isabelle HATIN[1] and Olivier NAMY[1]

[1] Institute for Integrative Biology of The Cell, UMR CNRS 9198, University Paris-Sud, 91405 Orsay, Cedex, France

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

Ribosome Profiling provides genome-wide information about translational regulation. However, there is currently no standard tool for the qualitative analysis of Ribo-seq data. Here we describe RiboTools [1], an efficient Galaxy package that can be used for primary analyses of ribosome profiling (Ribo-seq) data. RiboTools address translation fidelity issues, such as the identification of translational reading frames, stop codon readthrough events, and ribosomal A, P, and E-sites codon occupancy. RiboTools provides user- friendly publication-grade graphical results (html report). All tools use alignment files and reference annotation in BAM and GFF3 format respectively. Scripts are available from the Galaxy bioinformatics platform via the Test Tool Shed (ribo_tools) and in public Galaxy server RiboGalaxy (http://ribogalaxy.ucc.ie). RiboTools will be useful to the research community as it facilitates qualitative analysis of ribosome profiling data. We recently applied RiboTools to analyse ribosome profiling from yeast [2].

[1] R. Legendre, A. Baudin-Baillieu, I. Hatin, O. Namy. RiboTools: A galaxy toolbox for qualitative ribosome profiling analysis. *Bioinformatics*. first published online March 25, 2015

[2] A. Baudin-Baillieu, R. Legendre, C. Kuchly, I. Hatin, S. Demais, C. Mestdagh, D. Gautheret, O. Namy. Genome-wide translational changes induced by the prion [PSI+]. *Cell Rep*. 8:439-48, 2014

Link PDF

Post-028 (#125) - Sarah FARHAT - Exploring the mystery that leads to a specific parasite infection for marine blooming dinoflagellates

Exploring the mystery that leads to a specific parasite infection for marine blooming dinoflagellates

Sarah FARHAT[1], Tsinda RUKWAVU[2], Corinne DASILVA[2], Benjamin NOEL[2], Jean-Marc AURY[2], Laure GUILLOU[3], Betina PORCEL[1] and Patrick WINCKER[1]

[1] LAGE/Genoscope/IG/CEA, UMR8030 CNRS, Université d'Evry, 2 rue Gaston Crémieux, 91000 Evry, Cedex France
[2] LIS/Genoscope/IG/CEA, 2 rue gaston crémieux, 91000 Evry, Cedex France
[3] Station Biologique de Roscoff, UMR7144 CNRS, 29682 Roscoff, Cedex France

Corresponding author: sfarhat@genoscope.cns.fr

Although little studied, many extremely virulent microeukaryotic parasites infecting microalgae have been detected in the marine plankton. Syndiniales (also called Marine Alveolates or MALV), which constitute a diverse and highly widespread group, are parasitoids (i.e. obligatorily killing their host to complete their life cycle). Because of their virulence and abundant offspring, such parasites have the potential to control dinoflagellate populations, and therefore toxic microalgal blooms. Noteworthy, Syndiniales are not the only group of protistean parasites infecting marine blooming dinoflagellates. However, Syndiniales is the only group known to gather specialised parasites. In culture, they exhibit a narrow host spectrum, generally infecting no more than 1-2 dinoflagellate species. In the field, the same parasitic species may infect the same host species, years after years. Coastal planktonic ecosystems are by nature characterized by strong environmental fluctuations. This should theoretically lead to the natural selection of generalistic parasites at the expense of specialists. Thus, the persistence and ecological success of specialists among marine planktonic parasites is an intriguing paradox and a potential limitation to dinoflagellate blooms. The overarching goal of this project is to unveil the molecular components, mechanisms and evolutionary forces that determine the ability of specialized parasites to infect their primary host and adapt to a novel host, and how frequent host-range variations could be in nature. This will be performed by the genomic analysis of two distinct Amoebophrya strains in terms of impact on the host. The classic Genome Annotation pipeline turned out to be unfit for the annotation of Amoebophrya strains. Analysis revealed that Amoebophrya's gene structure does not follow the usage. Moreover, this parasite is phylogenetically distant of known organisms. Therefore, we set up a new pipeline based on transcriptomic data covering the complete life cycle of the parasite, and that used GMOVE, an automatic gene finding software that overcome generalist gene structures. A study of the global impact of these parasites will be conducted through the collection of planktonic sequences obtained in the Tara Oceans project.

Link PDF

RÉSEAUX, RÉGULATION & MODÉLISATION

NETWORK, REGULATION & MODELING

Thématique - Topic 06

Post-029 (#6) - Zohra SADOUK-HACHAÏCHI - Modélisation de la cinétique de croissance d'un consortium bactérien cultivé sur le gazole

Modélisation de la cinétique de croissance d'un consortium bactérien cultivé sur le gazole

Zohra SADOUK-HACHAÏCHI[1,3], Mohamed S. HACHAÏCHI[2], Ammal TAZEROUTI[1] and Hocine HACENE[3]

[1] Laboratoire de chimie organique appliquée-Faculté de Chimie, Université des Sciences et de la Technologie Houari Boumediene, Alger-Algérie
[2] Laboratoire d'analyse complexe-Faculté de Mathématiques, Université des Sciences et de la Technologie Houari Boumediene, Alger-Algérie
[3] Laboratoire de Microbiologie-faculté des Sciences Biologiques, Université des Sciences et de la Technologie Houari Boumediene, Alger-Algérie

Corresponding author: z.sadouk@hotmail.com

Notre étude porte sur la capacité d'une communauté bactérienne (composée de six souches isolées de boues actives) à utiliser le gazole comme seule source de carbone, à des concentrations élevées. En effet, une concentration de 84 g/L a été atteinte. Parallèlement à la croissance, des biosurfactants ont été produits par ce consortium durant la phase exponentielle. Ces substances abaissent la tension superficielle de l'eau de 72 mN/m à 31 mN/M avec une concentration micellaire critique évaluée à 81 mg/L.par ailleurs, les taux de croissance spécifiques obtenus à partir des courbes de croissance ont été déterminés dans le but de montrer que les résultats expérimentaux sont bien décrits par le modèle inhibiteur d'ANDREWS .Les constantes cinétiques trouvées sont : μmax = 0.535d−1 ± 0.063, KS = 18.68 g/L ± 3.59 and KI = 29.02 g/L ±4.96, ceci indique que la biodégradation se fait de manière assez lente .A la concentration de 25.2g (proche de la concentration optimale de S^*= 23.28 g/L± 4.23). Le consortium a métabolisé le gazole plus rapidement que chaque souche prise individuellement ; ce qui suggère l'idée que le processus a été stimulé par une synergie entre les souches bactériennes composant cette communauté.

Link PDF

Individual-based modeling of bacterial biofilms' resilience

Gabriel CARVALHO[1], Christiane FORESTIER[2], Damien BALESTRINO[2] and Jean-Denis MATHIAS[1]

[1] UR LISC Laboratoire d'ingénierie pour les systèmes complexes, IRSTEA, 9 avenue Blaise Pascal, 63178 Aubière, France

[2] LMGE, UMR6023 CNRS, 28 place Henri Dunant, Université Clermont Auvergne, 63001 Clermont-Ferrand, France

Corresponding author: gabriel.carvalho@irstea.fr

We propose here to study and formalize the concepts of resistance, persistence and resilience of bacterial biofilms by using an individual-based model (IBM). Such concepts constitute an important societal challenge with broad applications. However, these concepts need to be clearly defined to be efficiently used.

We argue that resistance, persistence and resilience are parts of an adaptive cycle. Bacteria have developed mechanisms in those three domains to cope with environmental fluctuations. The formation of heterogeneous biofilms, sheltering dormant bacteria, has a major role in those mechanisms. The IBM we develop focus on the response of bacterial biofilms perturbed by different agents like antibiotics. Our biofilm's model includes the transition of active cells to dormant cells and inversely. Transition rates are influenced by micro-environmental conditions like starvation and low sub-inhibitory antibiotic concentration.

The model manages to identify different phases of resistance, persistence and resilience of mature biofilms subject to one or multiple perturbations. In fine, we aim to identify and represent what drives the adaptation capacities of bacterial populations in biofilms.

Link PDF

Post-031 (#31) - Nordine KHERAKI - A systems-approach for modeling and analyzing the Ubiquitin/De-ubiquitin balance during virus infection

A systems-approach for modeling and analyzing the Ubiquitin/De-ubiquitin balance during virus infection

Nordine KHERAKI[1,2] and Denis MESTIVIER[1,2]

[1] Institut Jacques Monod, CNRS UMR7592 - Université Paris Diderot, 15 rue Hélène Brion, 75205 Paris Cedex, France
[2] Institut Mondor de Recherche Biomédicale, INSERM U955 - Université Paris-Est Créteil, 8 rue du Général Sarrail, 94010 Créteil, France

Corresponding author: denis.mestivier@u-pec.fr

Post-translational modification of proteins by ubiquitin (Ub) and degradation by the ubiquitin proteasome system (UPS) has emerged as a major regulatory process [1]. In particular, several viruses are challenged by Ub-mediated mechanisms in different steps of their life cycle and have evolved in order to exploit and interfere with the UPS [2,3]. Turnip yellow mosaic virus (TYMV) serves as a model system for positive- stranded RNA virus, and recent researches report that the TYMV-encoded replication protein bears a de- ubiquination (Dub) activity, which promotes this virus as an interesting model system to explore the balance of Ub/Dub events in the setting of a precise temporal and/or spatial control of viral infectivity [4,5].

We address the relationship between the Ub/Dub balance and viral infectivity using a system-wide modeling approach. We built a molecular model (22 species / 37 interactions) of the TYMV infection taking into account the Ub/Dub interactions with its hosts, using an exhaustive search of the literature. To cope with the lack of quantitative data in this specific field, we selected a boolean modeling approach.

We validate and analyze our model using (boolean) stochastic simulations [6,7,8]. We shows that: 1) our approach is well adapted to model host-virus interaction event with heterogeneous and scarce data, 2) we can explore the Ub/Dub balance and its consequences on viral infectivity.

[1] M.H. Glickman *et al*. *Physiol Rev* 82:373–428, 2002.
[2] M . K . Isaacson *et al*. *Host Microbe* 5:559–570, 2009.
[3] F . Randow *et al*. *Nat Cell Biol* 11:527–534, 2009.
[4] L. Camborde *et al*. *Plant Cell*. 22:3142-52, 2010.
[5] M. Chenon *et al*. *EMBO J*. 31:741-53, 2012.
[6] D.T. Gillespie. *The Journal of Physical Chemistry*, 81:2340-2361, 1977
[7] D.T. Gillespie *et al*. *The Journal of Chemical Physics*, 138:170901, 2013
[8] F. Achcar *et al*. *BMC Syst Biol*. 5(51), 2011

Link PDF

Metabolic network annotation with MetExplore

Floréal CABANETTES[1,2], Benjamin MERLET[3], Florence VINSON[3], Clément FRAINAY[3], Maxime CHAZALVIEL[1], Yoann GLOAGUEN[4], Sanu SHAMEER[3], Nathalie POUPIN[3], Fabien JOURDAN[3],Céline LAVIRE[5] and Ludovic COTTRET[1,2]

[1] INRA, Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR441, 24 Chemin de Borde Rouge – Auzeville, CS52627, 31326 Castanet Tolosan Cedex, France
[2] CNRS, Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR2594, 24 Chemin de Borde Rouge – Auzeville, CS52627, 31326 Castanet Tolosan Cedex, France
[3] INRA, UMR1331, TOXALIM, Université de Toulouse, Toulouse, France
[4] University of Glasgow, Glasgow, Scotland, G12 8QQ, UK
[5] University of Lyon, Université Lyon1, Ecologie Microbienne, UMR CNRS 5557/USC INRA 1364, 16 rue R. Dubois, 69622 Villeurbanne Cedex, France

Corresponding author: ludovic.cottret@toulouse.inra.fr

Metabolic network reconstruction consists in defining the list of the biochemical reactions involved in the metabolism of an organism. In order to facilitate, accelerate and improve this process, we have enhanced the MetExplore web server (http://metexplore.toulouse.inra.fr/metexplore2) with new features of metabolic network sharing and curation. Registered users can now create a project and share it with other MetExplore users so that they can also contribute to the curation. Each project can contain several metabolic reconstructions, facilitating the propagation of the annotations. The owner of the project defines the rights of the users (read, write...) on the metabolic reconstructions. Addition of new entities (reactions, metabolites, genes...) and their editing are facilitated thanks to automatic completions based on the names already present in the network. Metabolites can now be automatically linked to the different metabolic databases. Publications can also be assigned to reactions to justify their presence. Lastly, to facilitate the collaborative annotation, we set a vote system. For instance, MetExplore users can vote for the presence or absence of a reaction, or can point out some erroneous attributes. They can also add comments, with an option to attach a file to support their inference. This enables the project owner to make a final decision based on the votes and comments in order to better the quality of reconstruction.

Link PDF

Comparaison visuelle de sous-réseaux métaboliques avec MetExploreViz

Maxime CHAZALVIEL[1], Florence VINSON[1], Clément FRAINEY[1], Benjamin MERLET[1], Floréal CABANETTES[2,3], Yoann GLOAGUEN[4], Sanu SHAMEER[1], Ludovic COTTRET[2,3], Nathalie POUPIN[1] and Fabien JOURDAN[1]

[1] INRA, UMR1331 TOXALIM, Université de Toulouse, 31027 Toulouse Cedex 3, France

[2] INRA, Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR441, 24 Chemin de Borde Rouge – Auzeville, CS52627, 31326 Castanet Tolosan Cedex, France

[3] CNRS, Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR2594, 24 Chemin de Borde Rouge – Auzeville, CS52627, 31326 Castanet Tolosan Cedex, France

[4] University of Glasgow, Glasgow, Scotland, G12 8QQ, UK

Corresponding author: fabien.jourdan@toulouse.inra.fr

Le réseau métabolique d'un organisme est constitué de l'ensemble des réactions biochimiques potentiellement réalisables chez celui-ci indépendamment des conditions. Plusieurs méthodes existent pour extraire d'un réseau initial des sous-réseaux métaboliques spécifiquement associés à une perturbation ou actifs dans une condition donnée. Afin de comparer ces sous-réseaux, nous avons ajouté au serveur Web MetExplore (http://metexplore.toulouse.inra.fr/metexplore2) une interface qui permet de visualiser simultanément plusieurs réseaux et de mettre en évidence les parties communes et spécifiques. Un algorithme d'identification de sous-réseaux, basé sur le calcul des chemins les plus légers, permet d'identifier des sous-réseaux reliant des métabolites d'intérêt et de les mettre en évidence sur le dessin du réseau. L'interface permet également de mettre en évidence sur le réseau des données omiques importées grâce aux fonctions de « mapping ». Par ailleurs, plusieurs fonctionnalités ont été implémentées pour faciliter la visualisation et la lecture des réseaux. Notamment, MetExplore propose un affichage dynamique du réseau (contrairement aux cartes métaboliques traditionnelles statiques), avec la possibilité de déplacer des nœuds (ou groupes de nœuds) et des réarrangements automatiques des nœuds reliés pour éviter ainsi les superpositions. L'utilisateur peut également choisir de ne pas afficher les métabolites auxiliaires des réactions (e.g., H_20, H^+ …) pour alléger la visualisation du réseau. Enfin, la compartimentation du réseau est visible grâce à un code couleur, et un menu contextuel permet, pour chaque métabolite et réaction, d'afficher les données complémentaires présentes dans la base de données MetExplore. Le dessin du réseau est réalisé via la librairie javascript d3.js et est intégré dans l'interface MetExplore, développée via la librairie javascript ExtJs.

Link PDF

Post-034 (#112) - Shérazade BRAHAM - Inference of the protein interaction network between Fusobacterium nucleatum putative secretome and the human host

Inference of the protein interaction network between *Fusobacterium nucleatum* putative secretome and the human host

Shérazade BRAHAM[1,2], Christine BRUN[1,2,3] and Andreas ZANZONI[1,2]

[1] TAGC, UMR_S1090 INSERM, Campus de Luminy, 13288 Marseille, France
[2] TAGC, UMR_S1090 Aix-Marseille Université, Campus de Luminy, 13288 Marseille, France
[3] CNRS, 13402, Marseille, France

Corresponding author: zanzoni@tagc.univ-mrs.fr

It has been shown that *Fusobacterium nucleatum*, a gram-negative anaerobic species, is prevalent in colorectal cancer patients and may exert its tumorigenic capacity by inducing a pro-inflammatory response, an established cancer hallmark. Nevertheless, the causal role of *F. nucleatum* in colorectal cancer and the mechanistic details of host cell functions subversion are not fully understood. To date, only a handful of experimentally verified pathogenic effectors are known and the availability of interaction data with the host is very limited. In order to tackle these problems, we performed a comprehensive computational analysis to identify putative *F. nucleatum* virulence factors and to infer their interactions with human proteins. By using state-of-the-art tools (i.e. SignalP, SecretomeP), we predicted 237 *F. nucleatum* proteins as secreted (12% of the proteome). We observed an overrepresentation of functional domains commonly present in known bacterial virulence factors among those proteins. We next sought to identify molecular mimicry events (i.e. functional domains and short linear motifs, SLiMs) in the *F. nucleatum* putative secretome that might mediate the interaction with human proteins [1]. We found that 98 *F. nucleatum* secreted proteins have at least one potential mimicry event. For 89 of them, we inferred 1642 interactions with 412 human proteins, which are involved in immune system activation, inflammatory response, cell adhesion and cancer-related signaling pathways. Further exploration of this interaction network between the *F. nucleatum* secretome and the host should bring new insights on the role of the bacterium in colorectal cancer development.

[1] A. Via, B. Uyar, C. Brun, A. Zanzoni. How pathogens use linear motifs to perturb host-cell networks. *Trends in Biochemical Sciences.* 40:36-48, 2015.

Link PDF

Structural simplification of chemical reaction networks

Guillaume MADELAINE[1,2], Cédric LHOUSSAINE[1,2] and Joachim NIEHREN[1,3]

[1] CRIStAL, UMR 9189, 59650 Villeneuve d'Ascq, France
[2] University of Lille, France
[3] INRIA Lille, France

Corresponding author: guillaume.madelaine@ed.univ-lille1.fr

We study structural simplification of chemical reaction networks with kinetic functions, as used in systems biology, while preserving the deterministic behavior of the molecules of interest.

Our simplification aims to abstract from intermediate molecules and to preserve the dynamics of some specified molecules of interest. Despite the fact that reaction networks can be seen as input-output system, si- milarly to programming languages, we cannot use here its usual observational semantics based on convergence predicates, since they do not work for deterministic dynamics.

Therefore we define a new formal notion of model equivalence that says that two networks, in all possible contexts and under some partial equilibrium conditions, have the same deterministic dynamics modulo an ob- servation function. Our simplification is sound for this equivalence. And since it is based on a static analysis of network structure, we do not need to compute its ordinary differential equations system, neither its solu- tion, to perform the simplification. We can for instance merge molecules with symmetric behaviors, or delete intermediate molecules, as for instance in a generalized Michaelis-Menten simplification.

Finally, our simplification allows us to drastically reduce the size of reaction networks, as we show by applying it on a concrete biological model.

Link PDF

Post-036 (#147) - Marc LEGEAY - Construction et Analyse de Réseau de Gènes Contextuels dans le Domaine Végétal

Construction et Analyse de Réseau de Gènes Contextuels dans le Domaine Végétal

Marc LEGEAY[1,2], Béatrice DUVAL[1], Jean-Pierre RENOU[2] and Julie BOURBEILLON[2]

[1] LERIA, LERIA - Université d'Angers - UNAM, 2 boulevard Lavoisier, 49045 Angers, Cedex 01, France
[2] INSTITUT DE RECHERCHE EN HORTICULTURE ET SEMENCES (IRHS), UMR1345, INRA-Université d'Angers-AgroCampus Ouest, Centre Angers-Nantes, 42 rue Georges Morel BP 60057, 49071 Beaucouzé, Cedex, France

Corresponding author: marc.legeay@univ-angers.fr

Un des problèmes actuels en bioinformatique est de comprendre les mécanismes de régulation au sein d'une cellule. Notre travail concerne l'étude des réseaux de gènes chez le pommier, avec la particularité d'y intégrer les acteurs encore mal connus que sont les ARN anti-sens. Pour cela, nous étudions leur comportement dans deux contextes expérimentaux d'un processus biologique de stress (la maturation du fruit). Nous avons tout d'abord identifié un ensemble de gènes d'intérêt caractérisés par l'expression des couples sens/anti-sens dans les deux contextes. Parmi ces gènes d'intérêt, nous avons retrouvé des acteurs de voies biologiques bien connues pour leur implication dans la réponse au stress, ce qui soutient notre hypothèse d'une régulation de ce mécanisme par les anti-sens. Pour chaque contexte, nous construisons des réseaux de gènes grâce à des méthodes d'inférence. Nous utilisons le logiciel WGCNA (Weighted Gene Co-expression Network Analysis) pour analyser les propriétés topologiques des réseaux et pour détecter des modules. Dans un réseau de co-expression, un module est un ensemble de gènes fortement corrélés. Notre analyse des modules est orientée pour prendre en compte l'action des anti-sens (score d'activité des anti-sens au sein d'un module notamment), afin d'identifier, dans les contextes expérimentaux qui sont comparées, quelles fonctions biologiques et quelles régulations sont impactées par l'expression des anti-sens.

Link PDF

Post-037 (#153) - Hugo PEREIRA - Interface development of a new clustering tool for co-regulated genes identification (CORGI)

Interface development of a new clustering tool for co-regulated genes identification (CORGI)

Hugo PEREIRA[1], Sylvain GAILLARD[1], Marie-Laure MARTIN-MAGNIETTE[2], Herman HÖFTE[3], Jean-Pierre RENOU[1] and Sandra PELLETIER[1]

[1] INSTITUT DE RECHERCHE EN HORTICULTURE ET SEMENCES (IRHS), UMR1345 INRA-Université d'Angers-AgroCampus Ouest, Centre Angers-Nantes, 42 rue Georges Morel – BP 60057, 49071 Beaucouzé, Cedex, France
[2] MATHÉMATIQUES ET INFORMATIQUES APPLIQUÉES (MIA), UMR518 AgroParisTech-INRA, Saclay Plant Sciences, 16, rue Claude Bernard, 75231 Paris, Cedex 05, France
[3] INSTITUT JEAN-PIERRE BOURGIN (IJPB), UMR1318 INRA-AgroParisTech, Saclay Plant Sciences, Route de St-Cyr, 78026 Versailles, Cedex, France

Corresponding author: sandra.pelletier@angers.inra.fr

Transcriptomic data integration might give some insights on gene function inference, provided that we can establish reliable relationships of gene clusters with specific developmental or environnemental conditions. However classical clustering tools show severe limits to address this question with non- supervised approaches. We have developped a gene clustering software combining expression data discretisation and statistics based on the convergence of the binomial law to the normal law, hereheinafter called CORGI : CO-Regulated Genes Identification.

Basically, this tool identifies groups of individuals having the highest probability to cluster in the highest number of circumstances, consequently to identify the most meaningfull circumstances which constitute this core of conditions. Regarding transcription profiling it allows to identifiy the cluster of genes having the highest probability to be co-regulated amongst a list of differentially expressed genes, and the experiments which are the most explicative of this cluster. In a preliminar step, the statistically significant differences of expression are transformed in -1,0 or +1. We hypothezise that the sense of the difference (up or down-regulated between two conditions) is more important that the absolute value of this difference (the value of the ratio). This tool could be applied to any biological dataset which could be discretised in such a way.

Link PDF

Post-038 (#160) - Charles LECELLIER - Probing microRNA regulation

Probing microRNA regulation

Charles LECELLIER[1] and Chloé BESSIERE[2]

[1] Institut de Génétique Moléculaire de Montpellier, UMR5535 CNRS, 1919 Route de Mende, 34293 Montpellier, Cedex 5, France
[2] MASTER STIC Santé, Faculté des Sciences de Montpellier, 2 place Eugène Bataillon, 34090 Montpellier, France

Corresponding author: chloe.bessiere@gmail.com

The microRNA are small RNAs involved in the adjustment of protein production in response to various stimuli. Their expression must be accurately controlled to ensure plethora of cellular processes. Conversely, their deregulation is often, if not invariably, associated with human diseases and can be used as diagnostic biomarkers. Deciphering the mechanisms governing the control of microRNA expression is hence of prime importance. While most studies focus on the consequences of microRNA (de)regulations, we study their causes. The miRNA biogenesis can be controlled (i) at the transcriptional level by regulatory motifs present in their promoters and able to bind transcription factors and (ii) at the post-transcriptional level by proteins able to bind motifs present in their precursors. We are developing several statistical approaches using the expression of these key regulators as predictive variables to model miRNA expression in cancers. Our models are further validated with known miRNA regulations and protein-protein interaction data.

Link PDF

Post-039 (#161) - Andreas ZANZONI - Integration of quantitative proteomics data and interaction networks: identification of dys-regulated cellular functions during cancer progression

Integration of quantitative proteomics data and interaction networks: identification of dys-regulated cellular functions during cancer progression

Andreas ZANZONI[1,2,] and Christine BRUN[1,2,3]

[1] TAGC, UMR_S1090 INSERM, Campus de Luminy, 13288 Marseille, France
[2] TAGC, UMR_S1090 Aix-Marseille Université, Campus de Luminy, 13288 Marseille, France
[3] CNRS, 13402, Marseille, France

Corresponding author: brun@tagc.univ-mrs.fr

We have developed a computational network-based framework that, by combining quantitative proteomics data with protein interaction analysis, aims to identify dys-regulated cellular functions that can be involved in cancer progression. Indeed, the originality of our approach is three-fold: (1) it exploits protein expression levels, which better describe cell phenotypes; (2) it provides a proper cellular context by taking into account the expression profile at the network module level instead of individual proteins; (3) it detects statistically dys-regulated network modules taking into account the cancer evolution through different stages. We have applied our framework on two datasets: a high-resolution proteomic profile that recapitulates breast cancer (BC) progression [1], and an ensemble of normal colon and cancer samples from the CPTA Consortium [2]. In both datasets, the majority of the network modules show an increased up-regulation during cancer progression and represents functions related to cell cycle, transcription and splicing. We observed a significant down-regulation of modules associated to cell adhesion in the BC dataset, whereas down-regulated modules in the CPTA dataset are mainly involved in immune response evasion and protein degradation. Overall, our network-based framework provides a proper biological context to interpret high-resolution cancer proteomics data and to investigate the mechanistic details of cancer progression. Since it has been developed as a general framework, it can be apply on any time-resolved, or stage-based, proteomics profiling experiments.

[1] T. Geiger, S.F. Madden, V.M. Gallagher, J. Cox, M. Mann. Proteomic portrait of human breast cancer progression identifies novel prognostic markers. *Cancer Research*. 72:2428-2439, 2012
[2] B. Zhang, J. Wang, X. Wang, J. Zhu, Q. Liu, Z. Shi, e*t al*. Proteogenomic characterization of human colon and rectal cancer. *Nature*. 513:382-387, 2014

Link PDF

Multiscale Modeling of cancer in *C. Elegans*

Bertrand LAFORGE[1] and Micks MORVAN[1]

[1] LPNHE, UMR7585, 4 Place Jussieu, 75252 Paris Cedex 05, France

Corresponding author: micks.morvan@gmail.com

C. Elegans is a simple model organism to compare its computational model with the real in vivo worms. We know exactly the lineage tree describing all the steps of mitosis and apoptosis during all the stages of embryogenesis and "OpenWorm" (http://www.openworm.org) is the first computational model of *C. Elegans* existing in open-source. We will model the 2 types of cancer in *C. Elegans*: Vulvar cancer and teratoma (Kirienko *et al.* 2010 Dev Dyn) and try to identify all possible parameters in the progression dynamics of cancer with a multiscale model based on multi-agent architecture. In the extended computational models of OpenWorm we will apply stochastic mutations in each cell, but taking into account main DNA repair processes as well as some intercellular communications involved in local resistance to cancer in some tissues. To compare the 2 computational models of cancers in extended OpenWorm with cancers in real worms in different situations, we need to design microfluidic chips with many channels to test easily various experimental parameters, as temperature, pH, UV, Electric Fields (continuous or pulsed), cytokines with control worms. These chips should be reusable and allow easy microscopy analysis of tumors in worms with a temperature dependant (12°/ 22°C) hydrogel "Lab On Chip" (Guillaume Aubry, Mei Zhan and Hang Lu).

Link PDF

GENETIQUE & GENETIQUE DES POPULATIONS

GENETICS & POPULATION GENETICS

Thématique - Topic 03

Genome Wide Association Studies on sunflower flowering time

Fanny BONNAFOUS[1,2], Nicolas LANGLADE[1,2] and Brigitte MANGIN[1,2]

[1] INRA, Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR441, 31326 Castanet-Tolosan, France
[2] CNRS, LIPM, UMR2594, 31326 Castanet-Tolosan, France

Corresponding author: Fanny.Bonnafous@toulouse.inra.fr

Most of agronomical traits are dependent on the action of many genes and their interaction with the environment. Understanding these mechanisms is a key issue in plant breeding. Recent advances in genotyping technologies allow the use of millions of genetic markers (SNPs) in GWAS (Genome Wide Association Studies). Using genotypic and phenotypic data, we look for chromosomal regions linked to the flowering time in a related sunflower hybrid population, in three environments. Several effects such as position in the field were taken into account in order to keep the variation related to the genotype only, using a mixed model. Once we predicted this value, GWAS studies were conducted to understand the determinism of the trait of interest, taking only into account relatedness between hybrids. The association between markers and trait was tested first marker by marker and then in a second multi-locus approach. The principle of this second method is to increase the model by incorporating the most associated SNP to the trait as regressor, until a model that explains nearly all the phenotypic variance. Selecting the best model was performed by retaining SNPs that are significant according to the False Discovery Rate in the model with the most regressors. Unlike marker by marker approach that selects many SNPs, many of which are in linkage disequilibrium, multi-locus analysis keeps only few SNPs associated with flowering date. One SNP stands out as being associated to the trait in the three sets of markers identified by multi-locus approach. This SNP is positioned on linkage group 9 and correspond to a locus already identified by a previous study, who already pointed this genomic region as important to control flowering time in a core-collection. In addition to this locus, 6 or 7 loci as involved in flowering time using this device, according to the different environments.

Link PDF

Genomic prediction of agronomic traits in soybean

Alexandra DUHNEN[1,2,3], Bruno CLAUSTRES[3], Jean DAYDE[4] and Brigitte MANGIN[1,2]

[1] INRA, Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR441, 31326 Castanet-Tolosan, France
[2] CNRS, Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR2594, 31326 Castanet-Tolosan, France
[3] RAGT 2n, BP 3336, 12033 Rodez, France
[4] Université de Toulouse, INP, École d'ingénieurs de Purpan, Équipe Agro-physiologie et Agro-molécules, 75 voie du TOEC, 31076 Toulouse, France

Corresponding author: Alexandra.Duhnen@toulouse.inra.fr

Genomic selection is a promising approach to predict the genetic value of individuals for polygenic traits. Because it allows the inclusion of the whole genotypic data available to calibrate a pre- diction model, genomic selection is expected to give more accurate predictions than traditional marker assisted prediction. However, its adequacy depends on the magnitude of genetic effects captured by the model to explain phenotypic variations. In a recently published study, a satisfactory prediction accuracy was obtained for grain yield in a soybean population. Our objective is to optimize the application of this approach to assist in selecting for yield and seed protein content in soybean populations.

A large population comprising early and late soybean lines is being studied. The whole population is

202 markers spread across the genome. Phenotypic data consist of multi-environment trials from the RAGT breeding program from 2003 to 2014. Based on phenotypic data we computed a genetic value per genotype, which represents the mean of its performance across its tested environments. These genetic values were used to calibrate a GBLUP model. Accuracy (the correlation between computed and predicted genetic values) was estimated by cross validations.

Our first results suggest that genomic selection could be applied to predict with good accuracy additive genetic values for yield in a population of late lines. Current work is made in order to optimize the prediction accuracy and to apply genomic selection to seed protein content.

Link PDF

Post-043 (#46) - Florent CHUFFART - Nouvelle approche en génétique quantitative : détection de loci contrôlant finement la distribution d'un caractère cellulaire probabiliste

Nouvelle approche en génétique quantitative : détection de loci contrôlant finement la distribution d'un caractère cellulaire probabiliste

Florent CHUFFART[1], Daniel JOST[2], Magali RICHARD[1], Hélène BOTTIN-DUPLUS[1] and Gael YVERT[1]

[1] LBMC/GY Lab. CNRS - ENS Lyon - UCB Lyon - HCL, UMR 5239, Lyon, France
[2] TIMC-IMAG Lab. CNRS - UJF Grenoble, UMR 5525, Grenoble, France

Corresponding author: gael.yvert@ens-lyon.fr

La variabilité de cellule à cellule peut avoir des conséquences importantes sur l'état général d'un individu. On s'intéresse habituellement à la valeur moyenne d'un caractère cellulaire (trait) d'un individu et la plupart du temps cela suffit pour appréhender son état physiologique. Cependant, il existe des situations où la dispersion d'un caractère cellulaire ou encore des évènements cellulaires rares sont à l'origine de troubles pathologiques ou de graves dysfonctionnements. Par exemple, il suffit que quelques cellules expriment un oncogène pour favoriser la survenue de cancers. Dans ces deux situations, si on ne s'intéresse qu'à la valeur moyenne du caractère cellulaire on manque la corrélation avec la pathologie. En revanche, en prenant en compte la variance ou mieux encore distribution complète du trait de cellule à cellule, on est en mesure de caractériser finement le phénotype des cellules et par conséquence de l'associer avec l'état général de l'individu. Nous avons développé une méthode statistique permettant de trouver les origines génétiques de variations quantitatives probabilistes de caractère cellulaire. Nous avons évalué notre méthode sur un jeu de données simulées et sur un jeu de données expérimentales. Nous avons constaté que si un locus impacte la moyenne ou la variance d'un caractère cellulaire quantitatif la méthode classique reste plus sensible. A contrario, notre méthode est capable de trouver l'origine génétique de variations phénotypiques fines dans des contextes où la méthode classique échoue. Potentiellement, notre méthode est capable de détecter des loci probabilistes dans tous types d'organismes et ainsi de mettre en évidence l'origine génétique de pathologies.

Nous remercions le PSMN et le CBP (Lyon, France) ainsi que Grid'5000 (www.grid5000.fr) pour les ressources informatiques (calcul et forge). Ce travail est financé par l'ERC (FP7/2007-2013 SiGHT-281359).

[1] G. Yvert, 'Particle genetics' : treating every cell as unique. *Trends in Genetics.*, 30/2 :49-56, 2014.
[2] S. Nogami, Y. Ohya and G. Yvert, Genetic Complexity and Quantitative Trait Loci Mapping of Yeast Morphological Traits. *PLoS Genetics*,3/2:e31, 2007.

Link PDF

In silico detection of causal mutation in beef cattle

Romain PHILIPPE[1,2], Gilles RENAND[3], Sébastien FRITZ[4], Pascal CROISEAU[3], Romain SAINTILAN[4], Didier BOICHARD[3] and Véronique BLANQUET[1,2]

[1] INRA, UMR 1061 Génétique Moléculaire Animale - GMA, 123 Avenue Albert Thomas, 87060 Limoges cedex, France
[2] Université de Limoges, UMR 1061 Génétique Moléculaire Animale – GMA, 123 Avenue Albert Thomas, 87060 Limoges cedex, France
[3] INRA, UMR 1313 Génétique Animale et Biologie Intégrative – GABI, Domaine de Vilvert, 78352 Jouy-en-Josas cedex, France
[4] Allice, 149 rue de Bercy, 75595 Paris cedex 12, France

Corresponding author: romain.philippe@clermont.inra.fr

Muscular and skeletal development and meat quality are commercially important traits in beef cattle industry. However, the genetic control of these traits is mainly unknown and few genetic markers are actually used in beef cattle selection. Association studies, using BayesCπ methods on phenotypic data of 21 traits and genotyping data of around 39 000 SNP (Illumina array bovineSNP50) from 2155 cattles from two of the three main French beef cattle breed (Charolaise and Blonde d'Aquitaine) lead to the detection of 600 Quantitative Trait Loci (QTL). Twenty seven cattles were selected based on their inferred QTL status (homozygous or heterozygous) and their whole genomes were sequenced by Illumina technology with paired-end. After alignment on the UMD3.1 bovine reference genome using BWA software, filtering on the alignment quality and elimination of PCR duplicats, the mean depth of coverage was 11.6X. To optimize the variant detection (SNP and Indel), Indel were realigned and base alignment quality was recalibrated using GATK software. The variant detection is in process using three tools: haplotypeCaller (GATK), unifiedGenotyper (GATK) and platypus. After recalibration of the SNP, a concordance analysis between QTL status and variant status, coupled to an *in silico* functional analysis of the variant using Variant Effect Predictor software, will identify candidate mutations. *In vivo* and/or *in vitro* functional analysis would validate or invalidate these mutations. Validated causal mutations would be used to develop some diagnostic tools for muscular and skeletal development and meat quality in beef cattle industry.

Link PDF

Thaliadb: A database dedicated to association genetics in plants

Guy-Ross ASSOUMOU-ELLA[1], Yannick DE OLIVIERA[1], Johann JOETS[1] and Alain CHARCOSSET[1]

[1] Génétique Quantitative et Evolution – Le Moulon, UMR320 INRA, Ferme du Moulon, 91190 Gif-sur-Yvette, France

Corresponding author: ydeoliveira@moulon.inra.fr

Diversity and association genetics studies lead to manipulate a large number of individual, lines, clones and/or populations. Moreover, emergence of high-throughput technologies for both genotyping and phenotyping generates a large amount of data. These need to be stored and managed in order to perform requests and organize datasets to conduct association genetics studies.

Thaliadb manages genetic resources, phenotyping and genotyping data, and also population structure information. Thaliadb enables data extraction in formats used by genetic association software.

The schema enables dynamic description of accession (an introduction within the collection) and seed lot types. Thus, users can describe accession types. Images can be linked to an accession. This dynamic description is also available for markers. Some recent improvement ensures NGS data management.

Phenotyping data are stored as expertised data relative to a seed lot observed in a given environment. Raw data are stored in compressed files.

Classifications are expertised results concerning the assignation of a seed lot to classes of a population structure analysis. All data in Thaliadb are managed in projects. A user can access to the data concerning projects in which he is involved. Some users can have an administrator status, which give them rights to insert data, and link data and users to projects.

Each user can request and extract data concerning projects he is involved in. Genotyping and phenotyping data can be requested separately, but it is also possible to cross those data to extract information for association genetics studies or to interact with SniPlay and GnPAsso databases.

A Google map viewer has been integrated to Thaliadb. For accessions with geographical coordinates, it makes it possible to display classification, allele frequencies, trait or accession repartition on the map.

Link PDF

Post-046 (#134) - Yannick DE OLIVIERA - SHiNeMaS: A database dedicated to seed lots history, phenotyping data and field practices

SHiNeMaS: A database dedicated to seed lots history, phenotyping data and field practices

Yannick DE OLIVIERA[1], Laura BURLOT[1,2], Marie LEFEBVRE[1], Darkawi MADI[1], Pierre RIVIERE[1,3] and Mathieu THOMAS[1]

[1] Génétique Quantitative et Evolution – Le Moulon, UMR320 INRA, Ferme du Moulon, 91190 Gif-sur-Yvette, France
[2] AgroBioPerigord, 20 rue du Vélodrome, 24000 Périgueux, France
[3] Réseau Semences Paysannes, 3 avenue de la Gare, 47190 Auguillon, France

Corresponding Author: ydeoliveira@moulon.inra.fr

In 2005, a collaboration started between the French National Institute for Agricultural Research (INRA) and the farmer organization Réseau Semences Paysannes (RSP). The aim was: (1) to study on-farm management of crop diversity; (2) to develop population-varieties adapted to organic and low inputs agricultures in the context of a participatory plant breeding program involving farmers, NGOs' faciltators and researchers. In this project, researchers needed to map the life cycle of seed lots using the network formalism. All this information needed to be centralized and stored.

Thus, we developed SHiNeMaS (Seeds History and Network Management System) a database with its web interface, dedicated to the management of the history of seed lots and the associated data. The tool has been developed to be flexible and to support multiple agronomic species. The schema of the database is organized around the seed lot. It manages several types of relations between seed lots: multiplication, cross, intra-varietal selection, seed lot mixture and diffusion, to which the users can associate data.

SHiNeMaS provides interfaces to massively load data in tabulated format, or to individually load data through the web interface. A file format was designed for each type of event, and contains the minimum information to describe the concerned event. SHiNeMaS also provides tools to correct data already recorded. A tool was developed in SHiNeMaS to provide a helpful assistant to the creation of the files used for massive data loading. SHiNeMaS also provides query interfaces to retrieve and extract data. The user can access to the profile of a seed lot or a population-variety.

Link PDF

Post-047 (#141) - Lydia AIT BRAHAM - BioMercator: A complete framework to integrate QTL, meta-QTL, genome annotation and genome-wide association studies

BioMercator: A complete framework to integrate QTL, meta-QTL, genome annotation and genome-wide association studies

Lydia AIT BRAHAM[1], Yannick DE OLIVIERA[1], Olivier SOSNOWSKI[1] and Johann JOETS[1]

[1] Génétique Quantitative et Evolution – Le Moulon, UMR320 INRA, Ferme du Moulon, 91190 Gif-sur-Yvette, France

Corresponding author: joets@moulon.inra.fr

Compilation of genetic maps combined to QTL meta-analysis has proven to be a powerful approach contributing to the identification of candidate genes underlying quantitative traits. One of the most interesting properties of meta-QTL (or consensus QTL) is its confidence interval (IC) often shorter than IC of corresponding QTLs, decreasing the number of candidate gene to consider. As map compilation and QTL meta-analysis do not rely on genotyping raw data or trait measure, they can be easily achieved even if user holds maps from the literature or genetic databases.

BioMercator was the first software offering a complete set of algorithms and visualization tool covering all steps required to perform QTL meta-analysis. The fourth version of BioMercator proposes additional methods and improves graphical representation of large datasets. In this version, user may import sequence and genome annotations datasets within the software in order to display and mine functional annotation related to QTL and meta-QTL.

In order to improve candidate genes detection, we aim to include genetic association approach in the release of BioMercator. Association genetics allow to build a relationship between molecular polymorphism and phenotypic variation so, Genome-Wide Association Studies (GWAS) present a good potential for QTL's sharpening. We integrated GWAS results in Biomercator and provided new functionalities to display and exploit them.

BioMercator V4 is freely available from: http://moulon.inra.fr/biomercator and Biomercator V5 will be available soon.

Link PDF

Post-048 (#150) - Ankit DWIVENDI - Characterization of *P.falciparum* sub-populations associated to artemisinin drug resistance in Cambodia

Characterization of *P.falciparum* sub-populations associated to artemisinin drug resistance in Cambodia

Ankit DWIVEDI[1], Christelle REYNES[2], Nimol KHIM[3], Didier MENARD[3] and Emmanuel CARNILLOT[1]

[1] Institut de Biologie Computationnelle, Montpellier, France
[2] Institut de Génétique Humaine, Montpellier, France
[3] Institut Pasteur du Cambodge, Phnom Penh, Cambodge

Corresponding author: aankitddwivedi@gmail.com

The undergoing WHO Malaria elimination program is threatened by emergence and the potential spread of the *Plasmodium falciparum* artemisinin resistant parasite. *P.falciparum* artemisinin resistant parasites have emerged in the western part of Cambodia, where chloroquine and pyrimethamine drug resistance emerged in the past. Recent reports have shown (1) presence of several P. falciparum sub- populations in Cambodia and (2) evidence that mutations in the propeller domain of the K13 gene are major determinants of artemisinin resistance in Cambodian parasite population.

To characterize the Cambodian parasite sub-population metabolic properties and identify genetic evidence associated to the acquisition and the transmission of artemisinin resistance, a reliable SNP variant calling pipeline based on analysis of signal parameters in comparison with 3D7 reference genome was applied to around 170 NGS Cambodian genome sequences recovered from ENA database. In addition, a barcode approach based on LUMINEX technology was used to screen for parasite population structure in Cambodia. In addition, a barcode approach based on LUMINEX technology was used to screen for parasite population structure in Cambodia.

Genome wide analysis revealed presence of major hot spots of variation and specific haplotypes among the Cambodian sub-populations. Parasite sub-populations differed in metabolic capacities and specific genes in some sub-populations were associated to various housekeeping functions including cytoskeleton and ubiquitination, likely involved in K13 protein interaction. Our findings question the origin and the persistence of the *P.falciparum* sub-populations in Cambodia.

Link PDF

Post-049 (#162) - Clément AGRET - Use of Illumina sequencing for accessing viral genomes Cocoa swollen shoot virus (CSSV)

Use of Illumina sequencing for accessing viral genomes Cocoa swollen shoot virus (CSSV)

Clément AGRET[1] and Emmanuelle MULLER[2]

[1] Master STIC & Santé, Faculté des Sciences de Montpellier, 2 Place Eugène Bataillon, 34090 Montpellier, France
[2] CIRAD, UMR BGPI, Campus international de Baillarguet, 34398 Montpellier, Cedex 5, France

Corresponding author: clement.agret@gmail.com

Cacao swollen shoot virus (CSSV) is a double-strand DNA virus occurring in all the main cacao-growing areas of West Africa and limiting the cocoa production. To better understand the origin and the development of epidemics, biodiversity of the virus was studied. The exhaustive knowledge of such biodiversity will allow to set up a versatile diagnosis of the virus, essential to prevent its spread.

The different molecular studies performed so far in the endemic area showed that CSSV was highly variable. To have access exhaustively to viral biodiversity, we want to use the high-throughput sequencing with the Illumina technology on symptomatic samples that did not allow amplification of the virus. The Illumina sequencing will be set up on DNA extracts from cocoa leaves infected and having been subjected to different enrichment strategies in viral DNA. Viral semi-purification and amplification by rolling circle technique (RCA) will be compared. Analyses of Illumina sequence data will be on a cluster using different analytical pipelines comprising steps of alignment to reference genomes and assembly using the different tools available. This preliminary study will choose the best strategy to be applied to the next samples before sequencing and the most effective analysis methods.

Link PDF

Breeding value estimation for Yield and Quality traits using BWGS pipeline

Gilles CHARMET[1], Van Giang TRAN[1], Delphine LY[1] and Jérôme AUZANNAU[2]

[1] INRA, UMR 1095 Genetics, Diversity and Ecophysiology of Cereals, GDEC, 5 Chemin de Beaulieu, 63039 Clermont-Ferrand cedex 2, France
[2] Agri-Optentions, Ferme de Gauvilliers, 78660 ORSONVILLE, France

Corresponding author: gilles.charmet@clermont.inra.fr

Genomic selection is nowadays a very powerful tool for improving wheat traits such as grain yield, protein and gluten content and thousand kernel weights. In this sense, a lot of statistical methodologies have been applied for estimating the breeding values (GEBVs) using the information from molecular markers. The Breed Wheat project in France (2011-2019, www.breedwheat.fr) aims to develop a new genotype-to-phenotype multi-model machine learning tool for this purpose, called BWGS (Breed Wheat Genomic Selection pipeline). This tool works currently on Windows and Linux. Many genomic data processing analyses and GEBV prediction models are integrated. This poster presents an analysis using this pipeline to estimate breeding value of yield and quality traits in a bread wheat training population of 760 breeding lines genotyped with the BREEDWHEAT 423 K SNP chip. Accuracy values of G- BLUP, RKHS, LASSO, E-NET, Multi-kernel RKHS, BayesA, BayesB, RandomForest and BayesC predictions for grain yield and protein content were found between 0.38 and 0.54, whatever the number of markers N when N >5000 (up to 100 000). No difference in accuracy was observed when using SNP derived from gene vs intergenic sequences. The effect of missing data, imputation methods and prediction models are discussed.

Link PDF

METHODOLOGIES POUR L'ANALYSE DES SEQUENCES ET DES DONNEES OMIQUES

METHODOLOGIES FOR SEQUENCE ANALYSIS & OMICS DATA

Thématique - Topic 11

View and synchronize several genotypes with IGV

Marie-Laure FRANCHINARD[1], Frédéric SAPET[2], Sandra DEROZIER[1], Franck SAMSON[1] and Jean-François GIBRAT[1]

[1] INRA, UR1404 Mathématiques et Informatique Appliquées du Génome à l'Environnement, 78352 Jouy-en-Josas, France

[2] BIOGEMMA - Upstream Genomics Group, Centre de Recherche de Chappes, CS 90126, 63720 Chappes, France

Corresponding author: Marie-Laure.Franchinard@jouy.inra.fr

IGV (Integrative Genomics Viewer) [1] is a very efficient genome browser written in Java, allowing users to visualize and explore a large variety of genomic data types, but limited to a single genome. However, to perform comparative genomic studies, it is very useful to be able to observe different types of data simultaneously on several genotypes.

As part of the BioDataCloud project (Investments for the Future Initiative), a collaboration between the INRA Migale platform and the Biogemma company was established to tackle this issue. According to the technical specifications set by Biogemma, a new feature has been added to IGV that allows users to jump to a new genotype from different types of data (genes, regions in genomic sequence, genetic markers) selected by the user on the reference genome. This jump results in the opening of a new IGV window on these data, if they are available for the new genotype. This window retains all IGV features and synchronizes simultaneously with the main window. The number of jumps achievable and therefore the number of simultaneously observable genotypes is unlimited and depends only upon the available hardware capabilities and the availability of the corresponding data. All jumps can be saved in an IGV session file allowing users to quickly restore already used genotypes and data or to share them with other.

With this new feature, the user can now compare different genotypes with the reference genome and navigate between them synchronously while keeping the IGV performance.

[1] H. Thorvaldsdóttir, J.T. Robinson and J.P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* 14:178-192, 2013.

Link PDF

Post-052 (#22) - Hugo VARET - SARTools: a DESeq2- and edgeR-based R pipeline for comprehensive differential analysis of RNA-seq data

SARTools: a DESeq2- and edgeR-based R pipeline for comprehensive differential analysis of RNA-seq data

Hugo VARET [1,2], Jean-Yves COPPEE [1] and Marie-Agnès DILLIES [1,2]

[1] PLATE-FORME TRANSCRIPTOME & EPIGENOME, CENTER FOR INNOVATION & TECHNOLOGICAL RESEARCH, INSTITUT PASTEUR, 28 rue du Docteur Roux, 75015 Paris, France
[2] HUB BIOINFORMATIQUE & BIOSTATISTIQUE, CENTER FOR BIOINFORMATICS, BIOSTATISTICS AND INTEGRATIVE BIOLOGY, INSTITUT PASTEUR, 28 rue du Docteur Roux, 75015 Paris, France

Corresponding author: hugo.varet@pasteur.fr

We present the SARTools R package dedicated to the differential analysis of RNA-seq data for experiments with a simple design, i.e. experiments comparing several conditions of the same biological factor. SARTools provides tools to generate descriptive and diagnostic graphs, to run the differential analysis with one of the well-known DESeq2 [1] or edgeR [2] packages and to export the results into easily readable tab-delimited files. It also facilitates the generation of a final HTML report which displays all the figures produced, explains the statistical methods and gives the results of the differential analysis. Note that SARTools does not intend to replace DESeq2 or edgeR: it simply provides an environment to go with them. Moreover, the vignette of the package contains extensive help on how to use the workflow and gives some advices to detect potential problems such as the presence of a batch effect within the experiment or inversions of samples. SARTools is available on GitHub and is distributed with two R script templates (one for DESeq2 and one for edgeR) which use functions of the package. It should also be available on Galaxy very soon.

[1] M.I. Love, W. Huber and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15:550, 2014
[2] M.D. Robinson, D.J. McCarthy and G.K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:11, 2010.

Link PDF

Post-053 (#43) - Lise POMIES - Strategy for inferring gene regulatory network from kinetic expression with an unfavorable data-to-variables ratio

Strategy for inferring gene regulatory network from kinetic expression with an unfavorable data-to-variables ratio

Lise POMIES[1], Mélanie DECOUTEIX[1], Jérôme FRANCHEL[1], Bruno MOULIA[1] and Nathalie LEBLANC-FOURNIER[1]

[1] PIAF, UMR547 Inra - Université Blaise Pascal, 24 avenue des Landais, 63177 Aubière, France

Corresponding authors: lise.pomies@univ-bpclermont.fr & nathalie.leblanc@univ-bpclermont.fr

The study of dynamic regulatory network is more and more common to analyze complex biological process. The inference of such network is based on mathematical models that require equal (or more) data than variables to be relevant. In the case of dynamic gene regulatory networks, genes are considered as the variables and kinetic expression measurements are the data. For processes without prior information, global approaches are generally used, such as DNA-microarray or RNAseq. Those approaches allow obtaining a huge number of genes (or variables) but for few expressions conditions (or data).

In our case, to study the accommodation process of poplar to repeated mechanical loads, a transcriptomic analysis was performed with DNA-microarrays. We obtained 3600 candidates genes for the network with a kinetics composed of 5 measurement points, which means a really unfavorable data-to-variables ratio.

To reduce this unfavorable ratio, we built a strategy composed of three steps. (i) The "output" genes of the regulatory network were identified with other static transcriptomic results. (ii) A bioinformatics analysis of the transcriptomic data (clustering, Gene Ontology enrichment) allowed the selection of "representative" genes of a sub-set of expression modules. Such pre-analysis permitted the design of a complementary experiment less expensive as possible but as useful as possible for the network inference. (iii) The kinetics data were modelized with a kernel ridge regression, allowing the addition of theoretical measurement points (data) for each gene.

The relevance of our strategy should however be validated after the inference of the dynamic regulatory network. This work will be conducted in collaboration with bio-mathematicians specialized in those questions.

Link PDF

De novo assembly pipeline for transcriptomic analysis

Arnaud MENG[1], Lucie BITTNER[1], Fabrice NOT[2], Erwan CORRE[2] and Stéphane LE CROM[1]

[1] Institut de Biologie Paris Seine, FR3631 CNRS, 7-9 quai Saint Bernard, Université Pierre et Marie Curie, 75000 Paris, Cedex 05, France
[2] Station Biologique de Roscoff, UMR 7144, Place Georges Teissier, 29680, Roscoff, France

Corresponding Author: arnaud.meng@etu.upmc.fr

High-throughput sequencing technologies generate unprecedented amounts of genomic data. These recent methodological breakthrough are particularly interesting to study the biology of non-culturable organisms for which there is no reference genome. This is the case for many of the marine planktonic organisms, in which symbiotic interactions between microalgae and a predator host (photosymbiosis) are frequently observed but remains poorly known. Our works propose the establishment of an original *de novo* assembly method for RNA-Seq data and the exploration of transcriptome. More specifically we will apply our methods to the study of the photosymbiosis processes in the marine plankton. The RNA-Seq *de novo* assembly allows to reconstruct most of the transcripts of an organism from its sequenced transcriptome. Our pipeline is based on this technique and process the analysis to carry out the exploration of the transcriptome and the annotation of the studied species. The originality of our program lies in its ability to merge the transcriptomes reconstituted by two different assembly programs (Trinity and Velvet / Oases), in order to take the best of both assemblers for a given dataset of raw sequences. We have tested our pipeline on an existing transcriptome from yeast to validate its performance. The success of the validation step led us to its application on RNA-Seq data in the context of the investigation on photosymbiosis in the marine plankton. Using our program we reconstructed the transcriptomes of four symbiotic microalgaes. The exploration of transcriptomes from the holobiont (mix host + symbiont) is still in progress. We developed a new method to explore the transcriptomes of the symbiotic microalgae. This method, based on sequence similarity networks applied to RNA-Seq data aims at classify sequences and having a support for partitioning our downstream analyzes on restraint groups of sequences. Ultimately, our analysis will promote the understanding of processes involved in marine photosymbiosis.

Link PDF

Looking for SCPP transcripts in jaw transcriptomes

Xi LIU[1,2], Erwan CORRE[2], Christophe CARON[2] and Jean-Yves SIRE[1]

[1] Evolution Paris-Seine, Institut de Biologie Paris-Seine, UMR7138, Université Pierre et Marie Curie, quai St Bernard, 75005 Paris, cedex, France
[2] ABiMS, FR2424 CNRS UPMC, Station Biologique, 29680 Roscoff, France

Corresponding authors: xi.liu@sb-roscoff.fr & corre@sb-roscoff.fr

Skeleton mineralization (bones and teeth) was a major event in the emergence of vertebrates. The Secretory calcium-binding phosphoprotein (SCPP) family represents crucial molecules, which control mineralization of skeletal tissues. A total of 25 jaw transcriptomes were sequenced from representative species of the main lineages of non-mammalian vertebrates and one from Monodelphis domestica, in order to elucidate the origin and evolution of SCPPs (ANR project JAWS). A protocol of data cleaning, assembly and analysis was established for the opossum's transcriptome (for which a reference genome is accessible) to serve as the reference protocol for further analyses of the other transcriptomes. The transcriptomes were generated by Illumina 50 bp paired-end sequencing. Six assemblers (Trinity, CLC assembly cell, SOAP-denovoTrans, Velvet/Oases, EBARdenovo and TopHat/Cufflinks) were compared to select the most powerful. Assembly evaluations were subsequently realized on the data using different criteria (metrics, remapping, completeness, etc.). Trinity has generally shown the better assessment. The second objective of this study was to provide an exhaustive annotation of the transcriptomes for the scientific community. A functional annotation protocol was implemented using Blast2Go and Trinotate pipelines. A secondary annotation allowed us to identify contaminants and finally the candidate sequences (SCPPs) were identified by syntactic exploration of the annotation, by similarity research and sequences analysis of signal peptides of this protein family. This work is still ongoing for identifying possible new members of the SCPP family, which are previously unknown.

Link PDF

Post-056 (#60) - Aurélien QUILLET - miRabel: a web tool for an effective prediction of microRNA targets and their related biological pathways

miRabel: a web tool for an effective prediction of microRNA targets and their related biological pathways

Aurélien QUILLET [1], Chadi SAAD [1], Gaëtan FERRY [2], Youssef ANOUAR [1], Nicolas VERGNE [3], Thierry LECROQ [2] and Christophe DUBESSY [1]

[1] LABORATOIRE DE DIFFERENCIATION ET COMMUNICATION NEURONALE ET NEUROENDOCRINE (DC2N), UMR982 Inserm, Institut de Recherche et d'Innovation en Biomédecine (IRIB), Université de Rouen, Place E. Blondel, 76821 Mont-Saint-Aignan, Cedex, France

[2] LABORATOIRE D'INFORMATIQUE, DU TRAITEMENT DE L'INFORMATION ET DES SYSTEMES (LITIS), EA4108, Institut de Recherche et d'Innovation en Biomédecine (IRIB), Université de Rouen, Avenue de l'université, 76800 Saint-Etienne-du-Rouvray, Cedex, France

[3] LABORATOIRE DE MATHEMATIQUES RAPHAËL SALEM (LMRS), UMR6085 CNRS, Université de Rouen, Avenue de l'université, 76800 Saint-Etienne-du-Rouvray, Cedex, France

Corresponding author: christophe.dubessy@univ-rouen.fr

The bioinformatics identification of microRNA (miRs) targets remains a challenge because the performance of the algorithms developed in the past years remain weak. To address these issues, we developed miRabel, a new web tool to predict potential miRs targets and their associated biological pathways. Mirabel aggregates the predictions of three algorithms (miRanda, PITA and SVmicro) using the RobustRankAggreg package for R. By comparing the areas under the ROC curve of 104 miRs with experimentally known target mRNAs, we found that miRabel has better sensitivity and specificity than miRanda, PITA and SVmicro and its performance is comparable to other recently developed software. Furthermore, the target mRNAs predicted by miRabel are linked to metabolic and cellular pathways retrieved from KEGG which enable to assess their potential impact *in vivo*. MiRabel's database, developed with MySQL and InnoDB, actually includes 2 578 human miRs, 26 269 genes and 275 metabolic and cellular pathways which represents almost 15 million predicted interactions from which 54 165 are experimentally known. Freely available (http://bioinfo.univ- rouen.fr/mirabel/), miRabel is a new efficient tool for miRs' mRNA target prediction and their associated biological functions.

Link PDF

NGS Goes Automatic: From library preparation to data quality control

Gaëlle VILCHEZ[1], Claire KUCHLY[1], Jérôme MARIETTE[3], Frédéric ESCUDIE[3], Ibouniyamine NABIHOUDINE[3], Olivier BOUCHEZ[2], Diane ESQUERRE[2], Johanna BARBIERI[1], Céline JEZIORSKI[1], Sophie VALIERE[1], Clémence GENTHON[1], Marie VIDAL[1], Alain ROULET[1], Sandra FOURRE[2], Catherine ZANCHETTA[1], Adeline CHAUBET[1], David RENGEL[4], Denis MILAN[1], Cécile DONNADIEU[2] and Gérald SALIN[2]

[1] DGA, UAR1209 INRA, Plateforme GeT-PlaGe, 31326 Castanet-Tolosan, France
[2] GenPhySE, UMR1388 INRA, Plateforme GeT-PlaGe, 31326 Castanet-Tolosan, France
[3] MIAT, UR875 INRA, Plateforme bioinformatique, 31326, Castanet-Tolosan, France
[4] LIPM, UMR441 INRA, 31326 Castanet-Tolosan, France

Corresponding author: gaelle.vilchez@toulouse.inra.fr

Due to the constant increase of second generation sequencers (as HiSeq or MiSeq) throughput and the multiplication of their applications, a challenge is to sequence, at the same time, in a single run, ever more samples of different types, while being able to verify the quality of the produced data. At the INRA GeT- PlaGe facility, we have automated both library production and data quality control steps, in partnership with the Genotoul Bioinformatics facility, for protocols such as whole genome sequencing, Amplicon sequencing (e.g. 16S sequencing on MiSeq for metagenomics studies), stranded RNA-seq, Mate-Pair or whole genome bisulfite sequencing.

Having acquired a solid expertise in library preparation and data quality control of short fragments, our challenge for the coming months will be to integrate data from 3rd generation sequencers in our automated quality control processes, dealing with the specificity of long fragments, in partnership with the bioinformatics community of Toulouse and France Génomique.

Link PDF

Analysis of diploid genomes with Next-Generation Sequencing (NGS)

Varun KHANNA[1,2]

[1] Unité de Génétique moléculaire des levures, UMR3525 CNRS, Institut Pasteur, 25 rue du Docteur Roux, 75724 Paris, France

[2] Center of Bioinformatics, Biostatistics and Integrative Biology (C3BI), Institut Pasteur, 25 rue du Docteur Roux, 75724, Paris, France

Corresponding author: varun.khanna@pasteur.fr

The detection of genetic variations, with the NGS, can have a real impact in the understanding of complex diseases (such as cancer). By definition, a diploid organism has two sets of chromosomes (2n) in their core genomes. However, the analytical approach is delicate, especially in the detection of heterozygous SNPs/Indels, when copy number varies.

This poster presents a pipeline and illustrates important concepts and steps to better manage the analysis process, using standards methods: mapping and *de novo* assembly. This pipeline was developed and applied to our project analysis of a series of evolved yeast strains sequenced by NGS.

Link PDF

Post-059 (#69) - Bastien JOB - ASSESS.LINE: How good are my cell lines?

ASSESS.LINE: How good are my cell lines?

Bastien JOB[1,2], Florian RAMBOW[3], Cathy PHILIPPE[2,4], Guillaume MEURICE[2], Lionel LARUE[5] and Daniel GAUTHERET[1,2]

[1] INSERM AMMICA Bioinformatics Core, Villejuif, FRANCE
[2] Bioinformatics Core, Gustave Roussy, Villejuif, FRANCE
[3] VIB Center for the Biology of Disease, KU Leuven, BELGIUM
[4] CNRS UMR 8203, Université Paris-Sud - Gustave Roussy, Villejuif, FRANCE
[5] CNRS UMR 3347 - INSERM U1021, Institut Curie, Orsay, FRANCE

Corresponding author: bastien.job@gustaveroussy.fr

Cell lines derived from tumors have been the instrument of choice in experimental research in oncology for the last six decades. However, they suffer from genetic instability drift and lack the tumoral microenvironment. Therefore, the question of evaluating their ability to be effective models quickly emerged, with a recent focus on their similarity at the molecular level. Since 2005, several publications proposed classifiers to label cell lines as good or bad models. Unfortunately, these methods do not return a quantitative similarity metric, neither the features contributing to this similarity. We propose a novel method that both performs the scoring of tumoral cell lines and returns the contributing features. For this purpose, Principal Component Analyses are performed in a bootstrap procedure on a dataset composed of tumor samples from different pathologies. For each iteration, the candidate cell line is projected on the first PCs. Scoring corresponds to the fraction of generated spaces in which the candidate cell line was found closer to the centroid of its expected type. Further screening of spaces based on a minimal distance between the cell line projection and the expected centroid allows to rank the contributing features through a hypergeometric test comparing features occurrences in the selected spaces against in all generated spaces. The method was applied on several data measurements (mRNA, miRNA, methylation) from different technologies (microarray, HTS). Surprisingly, some cell lines appeared to be mislabeled, greatly derived from their original cell type, or putatively contaminated. Surprisingly, some cell lines appeared to be either mislabeled, greatly derived from their original cell type, or putatively contaminated. The method will be available as an R package, and a user friendly web tool.

Link PDF

Post-060 (#71) - Stevenn VOLANT - A novel statistical approach and R package to analyse HDX-MS of protein-protein interactions

A novel statistical approach and R package to analyse HDX-MS of protein- protein interactions

Stevenn VOLANT[1], Véronique HOURDEL[2], Darragh P. O'BRIEN[3], Sébastien BRIER[2] and Marie-Agnès DILLIES[1]

[1] Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris, France
[2] Unité de Spectrométrie de Masse Structurale et Protéomique, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris, France
[3] Unité de Biochimie des Interactions macromoléculaires, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris, France

Corresponding author: stevenn.volant@pasteur.fr

Hydrogen Deuterium Exchange detected by Mass Spectrometry (HDX-MS) is a powerful technique to probe the conformation and dynamics of proteins. Over the past 10 years, the HDX-MS workflow has been optimized and automatized, leading to a rapid expansion of the technology in both academic lab and pharmaceutical companies. Thanks to such improvements, modern HDX-MS can be applied to investigate more complex biological systems, including large protein complexes and membrane proteins. In this context, the larger the size of the protein under study the more complex the HDX-MS dataset.

Most of the existing tools for HDX-MS datasets are suitable to analyze data at one time point but do not account for time dependency. To address this issue, we propose a linear mixed-effects model which a generalization of the ANCOVA model developed by HDX-Analyser. By integrating replicates as random effect, the mixed-effects model accounts for both time dependency and the variability between replicates. Once the model parameters are estimated *via* the REML method, two Wald tests are derived to detect significant dynamics changes and masking effects. We then applied the Benjamini-Hochberg procedure to control the False Discovery Rate (FDR). We implemented this approach into an R package named MEMHDX (Mixed-Effects Model for HDX-MS experiments) which also provided an in-house data visualization of the output.

As an application, we sought to investigate the behavior of the adenylate cyclase (CyaA) catalytic domain (AC) upon interaction with calmodulin (CaM) in both the absence (apo-) and the presence (holo-) of CaCl2. Using our statistical approach, we were able to locate and validate those regions of CaM which are modified upon interaction with AC.

Link PDF

Looking for mutations in PacBio cancer data: an alignment-free method

Justine RUDEWICZ[1,2,3], Richard IGGO[1], Raluca URICARU[2,3], Jonas BERGH[4], Macha NIKOLSKI[2,3] and Hayssam SOUEIDAN[2]

[1] CENTRE RÉGIONAL DE LUTTE CONTRE LE CANCER DE BORDEAUX, U916 INSERM, 229 Cours de l'Argonne, 33000 Bordeaux, France
[2] CENTRE DE BIOINFORMATIQUE DE BORDEAUX, Université de Bordeaux, 146 rue Léo Saignat, 33000 Bordeaux, France
[3] LABORATOIRE BORDELAIS DE RECHERCHE EN INFORMATIQUE, UMR5800 CNRS, 351 cours de la Libération, 33405 Talence
[4] DEPARTEMENT ONCOLOGY, Université d'Upsala, Box 256, 751 05 Uppsala Suède

Corresponding author: j.rudewicz@bordeaux.unicancer.fr

The clinical study EORTC 10994 was set up to determine whether the status of p53 would select the therapy of patients with breast cancer. The identification of mutations in the TP53 gene through NGS sequencing of TP53 mRNA has proved effective in a pilot study. To determine the TP53 mutations present in all patients (~1500), tumor TP53 mRNA was sequenced thanks to the third generation sequencing technology "Pacific Bioscences". However, neither the pipelines of the pilot study nor the use of GATK tools has proven suitable for this type of data. Indeed, in addition to the high sequencing error rate generated by PacBio (~15%), there is contamination of tumor specimens by the healthy tissue. The low mutation rate expected for some samples makes it impossible to differentiate real mutations from sequencing errors with standard tools. To circumvent this problem, we have developed a methodology for detecting mutations using de Bruijn graphs that we will present in this poster.

Link PDF

Post-062 (#75) - Patrice BAA-PUYOULET - From SymbAphidBase to SymbAphidCyc: two companion databases to study and compare aphid symbionts from genomes to metabolic pathways

From SymbAphidBase to SymbAphidCyc: two companion databases to study and compare aphid symbionts from genomes to metabolic pathways

Patrice BAA-PUYOULET[1], Mathieu LABERNARDIERE[1], Nicolas PARISOT[1], Jean-Pierre GAUTHIER[1], Gérard FEBVAY[1], Federica CALEVRO[1], Yvan RAHBE[1], Hubert CHARLES[1], Jean-Christophe SIMON[2] and Stefano COLELLA[1]

[1] UMR203 BIOLOGIE FONCTIONNELLE INSECTES ET INTERACTIONS (BF2I), INSA LYON, 20 ave Albert Einstein, bât Pasteur, 69621 Villeurbanne, Cedex, France
[2] UMR 1349 INSTITUTE OF GENETICS ENVIRONMENT AND PLANT PROTECTION (IGEPP), INRA, Domaine de la Motte, BP 35327, 35653 Le Rheu, Cedex, France

Corresponding author: Patrice.Baa-Puyoulet@lyon.inra.fr

The next-generation sequencing methods evolution opened the way to unprecedented opportunities to study symbiotic bacteria through genomic comparisons and metabolic networks analyses to understand their contributions to their host biology. Aphids are an established model: they harbor an obligate primary endosymbiont, *Buchnera aphidicola*, and several facultative secondary symbionts contributing to the insect's physiology and adaptation. Recently we developed SymbAphidBase (http://symbaphidbase.cycadsys.org/): an *ad hoc* genome database to store and analyze aphid symbionts' genome sequences. At present SymbAphidBase includes the genomes of 17 strains of *B. aphidicola* from 8 different aphid species available in GenBank (see poster ECCB 2014 http://f1000.com/posters/browse/summary/1096898). In a second phase, we used CycADS (http://www.cycadsys.org/), an automated annotation management system that we developed during the pea aphid genome project, to generate an enriched functional annotation geared towards metabolic reconstruction. It is worth noting that the CycADS pipeline has been recently integrated into the Galaxy web-based platform to ease its use. The genomic data as well as the functional annotations obtained using different methods (such as KAAS, PRIAM, Blast2GO, InterProScan), were collected to easily build SymbAphidCyc (http://symbaphidcyc.cycadsys.org/), the database collection for the aphid symbionts. In conclusion, SymbAphidBase and SymbAphidCyc companion databases facilitate genome data storage and comparisons, as well as metabolism analysis for the better understanding of symbiosis physiology in aphids.

Link PDF

Decoupling RNAseq data from genomic context to help understanding the data

Bernd JAGLA[1], Manuel HOLTGREWE[2], Marie-Agnès DILLIES[1], Caroline PROUX[1], Odile SISMEIRO[1], Hugo VARET[1], Jean-Yves COPPEE[1] and Knut REINERT[2]

[1] Plate-forme Transcriptome et Epigénome, Département Génomes et Génétique, Institut Pasteur, Paris, France
[2] Institut für Informatik, Algorithmische Bioinformatik, Free University, Berlin, Germany

Corresponding author: bernd.jagla@pasteur.fr

The visual exploration of high-throughput sequencing experiments is largely limited to the use of genome-browsers. These browsers present the sequencing profiles in the context of the reference genome. We have decoupled the expression profiles from this constrain and present tools that allow for the visual inspection and exploration of “regions of interest” outside of the context of the reference sequence. In the context of quality control we can visually inspect RNA-seq experiments within seconds even for larger genomes. In addition, the visual exploration of NGS data proved to being useful in smallRNA-seq experiments and others. We show how to apply this technology to quality control, miRNA analysis, and transcription start site annotation/analyses. C++ version, R, and Galaxy integrations are available through http://www.seqan.de/projects/ngs-roi/ and links therein. Development version can be found at git-hub: https://github.com/PF2-pasteur-fr/seqan/apps/. And the KNIME integration can be found here http://tech.knime.org/community/next-generationsequencing and through the official update-site of KNIME.

Link PDF

Post-064 (#103) - François BARTOLO - Computational optimisation for mixOmics, the R package dedicated to 'omics' data integration

Computational optimisation for mixOmics, the R package dedicated to 'omics' data integration

François BARTOLO[1], Sébastien DEJEAN[1], Benoît GAUTIER[1], Ignacio GONZALES[1], Florian ROHART[3] and Kim-Anh LE CAO[2]

[1] Institut de Mathématiques, UMR5219 CNRS, Université Toulouse 3 Paul Sabatier 118, route de Narbonne, 31062 Toulouse, Cedex 9, France

[2] The University of Queensland Diamantina Institute, QLD 4102 Woolloongabba, Brisbane, Australia

[3] Australian Institute for Bioengineering and Nanotechnology, The University of Queensland, QLD St Lucia, Brisbane, Australia

Corresponding author: k.lecao@uq.edu.au

mixOmics is an R package dedicated to the exploration and integration of 'omics datasets. Its first release to the CRAN in 2009 proposed statistical methodologies to integrate two 'omics data sets. Since then numerous methodologies and variants have been implemented, and amongst those Generalized and Sparse Canonical Correlation Analysis (GCCA) to integrate more than two datasets. These latest developments require effective computational optimization and memory management. Indeed, some functions could use one CPU for a full on a standard desk computer on large biological biological studies.

We investigated three ways to address these computational challenges via

1/ sequential optimization (pre-compilation of functions)

2/ parallel computation (using the parallel package)

3/ enhancement of memory management (using the bigmemory package). Our first results obtained on a micro-benchmark showed computation times divided by at least 4.

The poster will present a global overview of the computational improvements made with these enhancements on real biological datasets.

Link PDF

Approche intégrée pour la découverte de nouvelles activités enzymatiques

Guillaume REBOUL[1,2,3,4], Maria SOROKINA[1,2,3], Jonathan MERCIER[1,2,3], Karine BASTARD[1,2,3], Mark STAM[1,2,3], David VALLENET[1,2,3] and Claudine MÉDIQUE[1,2,3]

[1] Commissariat à l'Energie Atomique et aux Energies Alternatives, Institut de Génomique, Genoscope, LABGeM, 2 rue Gaston Crémieux, 91057 Evry, France
[2] CNRS-UMR8030, 2 rue Gaston Crémieux, 91057 Evry, France
[3] UEVE, Université d'Evry Val d'Essonne, bd François Mitterrand, 9105, Evry, France
[4] Université de Rouen, 76183 Rouen cedex, France

Corresponding author: mstam@genoscope.cns.fr

Le séquençage massif des génomes induit une croissance exponentielle du nombre de protéines dans les banques de données publiques. Parallèlement, de nombreuses activités enzymatiques (~22%), expérimentalement démontrées demeurent orphelines de séquence (Sorokina *et al.*, 2014). L'accès à ce vaste répertoire de nouvelles séquences protéiques offre donc l'opportunité de découvrir des enzymes pouvant être associées à de nouvelles réactions. Nous présentons ici une approche intégrée de méthodes bioinformatiques ayant pour objectif de combler cette lacune de connaissances du métabolisme et proposant de nouvelles associations protéines/activités enzymatiques à valider expérimentalement. Dans cette optique, le groupe « Nouvelles Activités Enzymatiques » du laboratoire LABGeM développe plusieurs outils. La méthode CanOE combine les contextes génomiques et métaboliques pour la prédiction de candidats pour des activités orphelines (Smith *et al.*, 2012). Cette approche est actuellement étendue à la recherche de motifs de transformations chimiques conservés dans le métabolisme (Sorokina *et al.*, en préparation). D'un point de vue structural, la méthode ASMC recherche et compare les poches catalytiques pour classifier les enzymes d'une famille et détecter les résidus importants pour la spécificité de substrats (de Melo-Minardi *et al.*, 2010). Ces méthodes ont été appliquées avec succès dans l'étude d'une famille de fonction inconnue qui a permis d'en révéler sa diversité de fonctions enzymatiques (Bastard *et al.*, 2014). Leurs résultats, associés aux connaissances actuelles, doivent être ainsi unifiés dans une base de données et permettre l'élaboration de stratégies pour la sélection de familles d'enzymes d'intérêt. Ce travail bénéficie des données de la plateforme MicroScope d'analyse de génomes microbiens qui associe l'information génomique à la reconstruction de réseaux métaboliques (Vallenet *et al.*, 2013).

Link PDF

MicroAnnot: a pipeline for high quality microsporidian genomes annotation

Nicolas PARISOT[1], Abdel BELKORCHIA[2,3], Valérie POLONAIS[2,3], Sébastien RIMOUR[4], Sebastian NIEHUS[5], Dominique FERRANDON[5], Frédéric DELBAC[2,3], Pierre PEYRET[4] and Eric PEYRETAILLADE[4]

[1] UMR 203 BF2I, Biologie Fonctionnelle Insectes et Interactions, INRA, INSA-Lyon, Université de Lyon, 69621 Villeurbanne, France.

[2] CNRS, UMR 6023, LMGE, 63171 Aubière, France

[3] Clermont Université, Université d'Auvergne, Laboratoire "Microorganismes: Génome et Environnement", 63171 Clermont-Ferrand, France

[4] Clermont Université, Université d'Auvergne, EA 4678 CIDAM, BP 10448, 63001 Clermont-Fd

[5] UPR 9022 du CNRS, Université de Strasbourg, Institut de Biologie Moléculaire et Cellulaire 67084 Strasbourg Cedex

Corresponding author: eric.peyretaillade@udamail.fr

Microsporidia are a group of obligate intracellular parasites which have been recently associated with a phylum closely related to the fungal kingdom, the Cryptomycota. Microsporidian genomes are the leading models to understand the streamlining in response to a pathogenic lifestyle: they are gene-poor (1,800 to 2,600 protein coding genes) and often possess small genomes (from 2.3 to 24 Mb). Nevertheless, high-quality annotation of microsporidian genomes is essential for understanding the biological processes that govern the development of these parasites. Despite substantial progress in the past decade, current gene identification methods are not able to produce an *in extenso* catalogue of protein-coding genes. Transcriptomic analyses combined with RACE-PCR experiments have shown a high reduction of 5'UTR regions of microsporidian mRNAs (<20 nts). Based on this observation CCC-like, GGG-like sequence motifs or AT-rich regions upstream all transcriptional start sites have been identified among Microsporidia. Thus, these signals, located in close proximity of the translational initiation sites represent a tremendous tool for accurate microsporidian genome annotation. Using these signals, re-annotation of four previously annotated genomes has been carried out, which allow detecting (i) annotation errors and (ii) a significant number of unpredicted genes as small genes (Coding DNA Sequence <300 nt). Thus, taking advantage of these intrinsic regulation signals and using accurate annotation data from the four re-annotated genomes to implement an extrinsic approach, a pipeline dedicated to microsporidian genomes annotation have been developed. This tool named MicroAnnot was firstly used to ensure annotation of two microsporidian genomes infecting insects: *Anncaliia algerae* and *Tubulinosema ratisbonensis.*

Link PDF

Post-067 (#110) - Pierre-Julien VIALLY- GenerateReports: an IonTorrent plugin summarizing a whole NGS experiment for clinical interpretation

GenerateReports: an IonTorrent plugin summarizing a whole NGS experiment for clinical interpretation

Pierre-Julien VIALLY[1,2], Sylvain MARESCHAL[1], Philippe BERTRAND[1], Sydney DUBOIS[1], Elodie BOHERS[1], Catherine MAINGONNAT[1], Hervé TILLY[1] and Fabrice JARDIN[1]

[1] INSERM 918, Centre Henri Becquerel, Rouen, France
[2] Master in Bioinformatics, Normandy University, France

Corresponding author: pierrejulien.viailly@gmail.com

The increasing arrival of Next Generation Sequencing technologies in diagnostic laboratories creates a need to develop tools for rapid data interpretation. For example, targeted cancer sequencing allows biologists to focus on a selected range of known cancer-relevant genes and has become a choice strategy to quickly screen patients' mutational profiles but their interpretation remains difficult. The aim is now to develop tools to provide a quick under- standing of mutation profiles, highlighting the most impactful anomalies. Here, we present an integrated IonTorrent plugin called GenerateReports. This tool aggregates data in a single clinical report, enabling the clear visualization of the main results for each sample sequenced. Generate Reports is based on Coverage Analysis and Variant Caller Torrent Suite plugins results but goes even further by performing annotation of single nucleotide variants and searching for copy-number variations (CNVs). Thus, the biologist has access to sample identification, run and sample quality metrics, annotated and stratified variants, CNVs detected with statistical relevance and information about experimental and informatics traceability, all in a single PDF report for each sample sequenced. An associated interfaced database allows for further statistical studies. It could help to identify sequencing artifacts by Sanger validation storing and to stratify thousands of anomalies in several runs. We illustrate the results obtained using this plugin through the sequencing of a patient suffering from Diffuse Large B-Cell Lymphoma.

Link PDF

Post-068 (#115) - Guillaume GAUTREAU - Genome assembly using Nanopore-guided Long and Error-free DNA Reads

Genome assembly using Nanopore-guided Long and Error-free DNA Reads

Guillaume GAUTREAU[1], Mohammed-Amin MADOUI[1], Stefan ENGELEN[1], Corinne CRUAUD[1], Caroline BELSER[1], Laurie BERTRAND[1], Adriana ALBERTI[1], Arnaud LEMAINQUE[1], Patrick WINCKER[1,2,3] and Jean-Marc AURY[1]

1 Commissariat à l'Energie Atomique (CEA), Institut de Génomique (IG), Genoscope, BP5706, 91057 Evry
2 Université d'Evry Val d'Essonne, UMR 8030, CP5706, 91057 Evry, France.
3 Centre National de Recherche Scientifique (CNRS), UMR 8030, CP5706, 91057 Evry, France

Corresponding author: jmaury@genoscope.cns.fr

Long-read sequencing technologies were launched a few years ago, and in contrast with short-read sequencing technologies, they offered a promise of solving assembly problems for large and complex genomes. Moreover by providing long-range information, it could also solve haplotype phasing. However, existing long-read technologies still have several limitations that complicate their use for most research laboratories, as well as in large and/or complex genome projects. In 2014, Oxford Nanopore released the MinION® device, a small and low-cost single-molecule nanopore sequencer, which offers the possibility of sequencing long DNA fragments. The assembly of long reads generated using the Oxford Nanopore MinION® instrument is challenging as existing assemblers were not implemented to deal with long reads exhibiting close to 30% of errors. Here, we presented a hybrid approach developed to take advantage of data generated using MinION® device. We sequenced a well-known bacterium, *Acinetobacter baylyi* ADP1 and applied our method to obtain a highly contiguous (one single contig) and accurate genome assembly even in repetitive regions, in contrast to an Illumina-only assembly. Our hybrid strategy was able to generate NaS (Nanopore Synthetic-long) reads up to 60 kb that aligned entirely and with no error to the reference genome and that spanned highly conserved repetitive regions. The average accuracy of NaS reads reached 99.99% without losing the initial size of the input MinION® reads. Our method, based ideally on 20x and 50x of MinION® and Illumina reads respectively, provides an efficient and cost-effective way of sequencing microbial or small eukaryotic genomes in a very short time even in small facilities.

Link PDF

Post-069 (#124) - Elise LARSONNEUR - Evaluation of de novo assemblies in view of creating automated pipelines dedicated to core-genome bacterial typing

Evaluation of *de novo* assemblies in view of creating automated pipelines dedicated to core-genome bacterial typing

Elise LARSONNEUR[1,2,3], Eduardo P.C. ROCHA[3,4], Philippe GLASER[2] and Sylvain BRISSE[3,4]

[1] CNRS, UMS3601 IFB-Core, 91198 Gif-sur-Yvette Cedex, France
[2] INSTITUT PASTEUR, Genopole, PFBaG, 75724 Paris Cedex 15, France
[3] INSTITUT PASTEUR, Microbial Evolutionary Genomics, 75724 Paris Cedex 15, France
[4] CNRS, UMR3525, 75724 Paris, France

Corresponding author: elise.larsonneur@pasteur.fr

Genotyping of bacterial pathogens is of great importance for epidemiological surveillance and outbreak investigation. Widely used, multilocus sequence typing (MLST) indexed typically the variability of seven coding sequences, and thus lacked discriminatory power among isolates. This power is greatly improved with the extension of this approach to the core-genome (cgMLST) of bacterial species. We aim at providing a pipeline that will rapidly and accurately automate the process of assembling sequence data from short sequences, building core-genomes, and genotyping of isolates from hospitals and reference laboratories.

Core-genome construction relies on identifying conserved orthologous coding sequences and requires accurate *de novo* assemblies. However, the performance of a given assembler varies widely across datasets and parameter ranges. Recent evaluations (GAGE, Assemblathon) have concluded that there is no single best approach to genome assembly. Instead, it is preferable to generate multiple assemblies and validate them to determine which is most useful for the desired analysis.

As a consequence, we decided to compare different assemblers regarding their specific performance in terms of cgMLST. We used iMetAMOS, a pipeline which encapsulates the process of running and validating multiple assemblies. This program also selects the best assembly from multiple weighted assembly scores. Several *de novo* assemblers (CLC, IDBA-UD, MaSuRCA, Mira, SGA, SOAPdenovo2, SPAdes, Velvet, Velvet-SC) were tested for contiguity, accuracy and genotype variability metrics. We also estimated the dependency of these parameters on read length and coverage depth. This work allows defining the best combination of tools and parameters specifically adapted to cgMLST genotyping of different bacterial genomes, while guaranteeing reproducible and robust results for public health applications.

Link PDF

Post-070 (#127) - Dhouha. GRISSA - Knowledge Discovery based on Formal Concept Analysis for biomarker identification from metabolomic data

Knowledge Discovery based on Formal Concept Analysis for biomarker identification from metabolomic data

Dhouha. GRISSA[1,2], Jérémie BOURSEAU[2], Blandine COMTE[1], Amedeo NAPOLI[2] and Estelle PUJOS-GUILLOT[1,3]

[1] INRA, UMR1019, UNH-MAPPING, 63000 Clermont-Ferrand, France
[2] LORIA, B.P. 239, 54506 Vandoeuvre-lès-Nancy, France
[3] INRA, UMR1019, Plateforme d'Exploration du Métabolisme, 63000 Clermont-Ferrand, France

Corresponding author: dhouha.grissa@clermont.inra.fr

Metabolomics is a powerful phenotyping tool in nutrition research to better understand the biological mechanisms involved in the pathophysiological processes and identify biomarkers of metabolic deviations. It can be described as a global analysis of small molecules present in a biofluid, using data- driven approaches based on multiple analytical platforms, such as mass spectrometry, and chemometrics and bioinformatics. Such platforms generate massive and complex data that need analyses and integration to extract the biologically meaningful information and to enrich our knowledge and understanding of biological systems. In this context, our work consists in developing a workflow using Knowledge Discovery and Data Mining methodologies to propose advanced biomarker discovery solutions. Our choice focused on a technique called *Formal Concept Analysis* (FCA), introduced by Wille in 1982, which analyzes and describes the existing relationship between objects (individuals) and attributes (molecules). Nevertheless, the application of FCA on metabolomics data is not easy, and several pre-processing steps must be realized, as data cleaning, transformation and reduction. Furthermore, a post-processing phase is also essential to help interpreting FCA outputs. For this purpose, we propose to use the concept of *Emerging patterns* in order to discriminate subgroups of individuals with particular phenotypes. Finally, we plan to present the new extracted knowledge as visualized graphs.

Link PDF

Post-071 (#143) - Sacha BEAUMEUNIER - Rôle de l'apprentissage automatique dans le problème de détection d'ARN chimères

Rôle de l'apprentissage automatique dans le problème de détection d'ARN chimères

Sacha BEAUMEUNIER[3,2], Jérôme AUDOUX[1,2], Anthony BOUREUX[1,2], Thérèse COMMES[1,2], Ronnie ALVES[1,2], and Nicolas PHILIPPE[1,2]

[1] Institut de Médecine Régénératrice et de Biothérapie, INSERM, Hôpital Saint-Eloi, 80, avenue Augustin Fliche, 34295 Montpellier, Cedex 5, France
[2] Institut de Biologie Computationnelle, Université de Montpellier, 860 rue St Priest, Bâtiment 5 – CC05019, 34095 Montpellier, Cedex 5, France
[3] Master TIC&Santé, Université de Montpellier, 163 rue Auguste Broussonnet, 34090 Montpellier, Cedex 2, France

Corresponding author: ronnie.alves@lirmm.fr

Les technologies NGS sont intensivement utilisées pour étudier la complexité du transcriptome en utilisant la méthode de RNA-seq. Quelles que soient les questions biologiques, chaque analyse RNA-seq va permettre la prédiction de mutations, d'indels, de jonctions d'épissage, ou d'ARN de fusion. L'analyse est réalisée à l'aide des pipelines complexes impliquant de multiples outils pour l'alignement, la couverture et la prédiction des différents évènements biologiques. Nous utilisons Le programme, CRAC et les CracTools pour réaliser nos analyses. CRAC propose une nouvelle façon d'analyser les données en intégrant la localisation génomique et la couverture locale et permet toutes les prédictions biologiques en une seule étape, (crac.gforge.inria.fr/; Philippe et al, 2013) et les CracTools permettent d'effectuer des analyses RNA-seq diversifiées avec la possibilité d'intégrer des spécifications supplémentaires. Un des objectifs développé au laboratoire est la classification de nouveaux transcrits chimériques qui pourraient avoir un rôle dans les tumeurs. Pour cela nous avons tout d'abord développé un pipeline de benchmark intégrant une procédure de génération de génomes mutés et des données RNA-seq simulées par Flux Simulator. L'analyse de CRAC donne un set de chimères candidats contenant à la fois les chimères connues mais aussi des faux positifs correspondant à des artefacts méthodologiques. Nous utilisons des procédures de machine learning associée à la procédure de benchmark pour distinguer les caractéristiques permettant une classification rigoureuse et pertinente des reads couvrant des jonctions chimériques afin de discriminer les artefacts méthodologiques des candidats biologiques. Nous avons réalisé une évaluation des modèles par une stratégie d'échantillonnage progressif et observé un compromis entre la taille de l'échantillon et la précision de la classification. Nous proposons ainsi une classification des candidats chimères afin de faciliter la validation biologique et favoriser la découverte de nouveaux biomarqueurs pour le diagnostic, le pronostic et le suivi des cancers.

Link PDF

Semi-automatic biological data mining workflow

Imène CHENTLI[1], Arnaud CHARLEROY[1] and Catherine Marie BRETON[2]

[1] Spécialité « Bioinformatique, Connaissances, Données » du Master STIC pour la santé
[2] Université de Montpellier, 163 rue Auguste Broussonet, 34090 Montpellier, France
[3] Ifremer, UMR MARBEC CNRS, Bd Jean Monnet, BP 171, 34203 Sète Cedex, France

Corresponding author: chentli.imene@hotmail.fr

Biological data is constantly growing. Multiple results, files and sources are brought online. Thus, biological databases are continuously enriched demonstrating that the web has become a primary source of information. However, these sources are raw and difficult to be analyzed manually by biologists. They require organization and a well-defined structure for their interrogation, filtering and retrieval.

On that point, we propose a semi-automatic method to biological data structuring and knowledge extraction via a web service. The method is implemented in the portable language Julia, through three main steps (1) connection to the "Entrez" portal (Global Query Cross-Database Search System of NCBI) and retrieval omics data regardless of the studied species, (2) cleaning, formatting and preprocessing the datasets and (3) data clustering, analysis and visualization of the results by applying two data mining approaches carried out under the R and Weka tools.

In this study, we focused exclusively on Archaea proteome with 2.1 million data available on NCBI. Our interest is particularly focused on the correlation between 5 attributes: organism, locus and name of genes, the predicted or given protein function and its length. It is also possible to take into account additional attributes such as phyla.

The proposed method allows working with large amount of data. It is generic independent on the kingdom and studied macromolecules (DNA, RNA or proteins). Our approach also allows structuring and presenting data to experts in a format usable and supported by data mining tools such as R and Weka. As future work, our approach can be complemented by a dynamic interface to visualize the results. Python and R libraries can also be used for this purpose.

Link PDF

Long reads based assembly and impact of error correction

Amal Zine EL AABIDINE[1] and Eric RIVALS[1,2]

[1] L.I.R.M.M. U.M.R. 5506, CNRS et Université de Montpellier Montpellier, France
[2] Institut Biologie Computationnelle, Montpellier, France

Corresponding Author: rivals@lirmm.fr

Recent technological advances gave rise to the 3rd generation of se- quencing techniques, which generate long reads. The Pacific Biosciences tech- nology (’PacBio’) and Oxford Nanopore Technology (’MinIon’) are the most prevalent long-read technologies today. In theory, long reads should ease genome assembly by resolving the structure of repeated regions, and help distinguishing RNA isoforms in complex eukaryotic transcriptomes. Yet, this promise is ham- pered by a high error rate and lower coverage With error rates superior to 15%, long reads are not amenable to direct assembly.Hence, the challenge of long reads is to error correct them. Two error correction strategies have been proposed: ei- ther self-correction using only long reads, or hybrid correction using high quality set of short reads. LORDEC is a hybrid error correction tool for long reads that can handle very large sets of short and long reads. It aligns the long reads on the paths of a de Bruijn graphs of the short reads. LoRDEC achieves a good level of correction while it outperforms existing other error correction software in terms of time and memory. We set out three goals. First, we wanted to evaluate the accuracy of error correction of both PacBio and MinIon reads by comparing the results of mapping raw and corrected long reads on a reference sequence. Second, we sought adequate strategies to assemble long reads from genomic samples, and investigated the impact of error correction on the assembly.

Link PDF

Post-074 (#151) - Magali BERLAND - Metagenomic data analysis: from reads to biomarkers with METEOR and MetaOMineR

Metagenomic data analysis: from reads to biomarkers with METEOR and MetaOMineR

Magali BERLAND[1], Emmanuelle LE CHATELIER[1], Edi PRIFTI[2], Nicolas PONS[1], Franck GAUTHIER[1], Amine GHOZLANE[1], Mathieu ALMEIDA[3], Pierre LEONARD[1], Jean-Michel BATTO[1], Pierre RENAUD[4] and S. Dusko EHRLICH[1,5]

[1] Metagenopolis, US1367 INRA, Domaine de Vilvert, 78352 Jouy-en-Josas CEDEX, France
[2] Institute of Cardiometabolism and Nutrition (ICAN), Paris, France
[3] Department of Computer Science, Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA
[4] MICALIS, UMR 1319 INRA, 78352 Jouy-en-Josas, France
[5] King's College London, Centre for Host-Microbiome Interactions, Dental Institute Central Office, Guy's Hospital, United Kingdom

Corresponding authors: emmanuelle.lechatelier@jouy.inra.fr & nicolas.pons@jouy.inra.fr

Metagenomic data analysis faces a critical challenge: the unprecedented amount of data generated by new sequencing technologies requires unified and user-friendly tools for data management and analysis. We present here the first ready-to-run pipeline able to deal with millions of reads and to mine in a huge amount of sparse variables of unknown interdependency. This pipeline is composed of:

- *METEOR*, a software suite dedicated to the primary processing of sequencing data (Illumina, SOLiD or Ion Proton) for quantitative metagenomic applications. The primary processing is composed of several modules for (i) metagenomic data indexation, (ii) quality controls (cleaning and filtering), (iii) mapping the reads and counting the genes in very large reference catalogues (millions of genes) and (iv) aggregation of counting profiles of hundreds of samples. METEOR comes with helpful tools for the integration of new reference catalogues, the management of sequencing data and the creation of running workflows.
- *MetaOMineR*, a R package allowing laptop analysis of the large counting profiles generated by METEOR. Normalization and down-sampling routines reduce technical variability between samples. Dimension reduction can be achieved by clustering, projecting the data on MetaGenomic Species (MGS) [1] or different filtering procedures. A variety of statistical routines allow to identify genes, MGS and functions associated to a given trait or phenotype.

This pipeline handles many important issues with the analyses of metagenomics data and has played an important role in successful projects [2-4]. It offers the possibility to identify important and promising biomarkers in the quest for understanding complex ecosystems and treating human disease.

[1] H.B. Nielsen, M. Almeida, *et al*. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference gnomes. Nature biotechnology, 32:822-828, 2014.
[2] E. Le Chatelier, *et al*. Richness of human gut microbiome correlates with metabolic markers. *Nature*, 500:541-546, 2013.
[3] A. Cotillard, *et al*. (2013). Dietary intervention impact on gut microbial gene richness. *Nature*, 500:585-588, 2013.
[4] N. Qin, *et al*. (2014). Alterations of the human gut microbiome in liver cirrhosis. *Nature*, 513:59-64, 2014.

Link PDF

Post-075 (#155) - Fabrice TOUZAIN - Bioinformatics reconstruction of the first complete genome sequence of European Turkey coronavirus by NGS sequencing, DB matching, tblastn gap filling, assemblies merging and realignment

Bioinformatics reconstruction of the first complete genome sequence of European Turkey coronavirus by NGS sequencing, DB matching, tblastn gap filling, assemblies merging and realignment

Fabrice TOUZAIN[1*], Paul Alun BROWN[2*], François-Xavier BRIAND[2], Meriadeg AR GOUILH[3,4], Céline COURTILLON[2], Chantal ALLEE[2], Evelyne LEMAITRE[2], Claire DE BOISSESON[1], Yannick BLACHARD[1] and Nicolas ETERRADOSSI[2]

[1] GVB unit, Agence Nationale de Sécurité Sanitaire (Anses), laboratoire de Ploufragan-Plouzané, BP53 22440 Ploufragan, France
[2] VIPAC unit, Agence Nationale de Sécurité Sanitaire (Anses), laboratoire de Ploufragan-Plouzané, BP53 22440 Ploufragan, France
[3] Environment and Infectious Risks Research and Expertise unit, Institut Pasteur, 25-28 rue du docteur Roux, 75724 Paris cedex 15, France
[4] Microbial Risks Research unit, Université de Caen Basse-Normandie, France
* The two first authors contributed equally to the study

Corresponding author: fabrice.touzain@anses.fr

In France in 2008, a coronavirus (TCoV-080385d) was isolated from turkeys exhibiting clinical signs of enteritis and was later shown to form a sub lineage related too, but significantly different from coronaviruses isolated from turkeys in North America (NA TcoV).

Here, we describe the process to determine an almost complete viral sequence for TCoV-080385d from Mi-seq reads and bioinformatics software's among which were trimmomatic (Bolger et al. 2012), bowtie2 (Langmead et al. 2012), ABYSS (Simpson et al. 2009), Vicuna (Yang et al. 2013), mira (Chevreux et al. 1999), and minimus (Sommer et al. 2007).

The originality of the method lies in the assembly process, through in house Perl scripts, of unmatching reads obtained by a tblastn (Altschul et al. 1997) from viral protein portions missing in original alignments on the closest viral genomes. Merging assemblies, realigning reads and manual corrections led us to the first almost complete turkey coronavirus genome of 27 661 nucleotides (and two gaps smaller than 6 nucleotides).

The full genome sequence for TCoV-080385d has now been completed. Annotation was performed using VIGOR program (Wang et al. 2010).

Link PDF

META-OMIQUES & GENOMIQUE ENVIRONNEMENTALE

META-OMICS & ENVIRONNEMENTAL GENOMICS

Thématique - Topic 07

ePANAM and the phylogenetic diversity of microbial communities

Gisèle BRONNER[1,2], Najwa TAIB[1,2,3], Jean-Christophe CHARVY[1,2], Isabelle JOUAN-DUFOURNEL[1,2] and Didier DEBROAS[1,2]

[1] Université Blaise Pascal, Laboratoire «Microorganismes : Génome et Environnement», BP 10448, 63000 Clermont-Ferrand, France
[2] CNRS, UMR 6023, LMGE, 63171 Aubière, France
[3] Université Claude Bernard – Lyon 1, Laboratoire de Biométrie et Biologie Evolutive, 43 bd. du 11 Novembre 1918, 69622 Villeurbanne, France

Corresponding Author: gisele.bronner@univ-bpclermont.fr

Advancements in next-generation sequencing technologies, have had a major impact in the field of microbiology by enabling an in depth analysis of microbial genomes from various ecosystems. In this context, inferring microbial assemblages diversity relies on the clustering of sequences of ~300 to 500 bp length of SSU rDNA genes (16S or 18S) to identify Operational Taxonomic Units (OTU). OTUs are classically taxonomically assigned to representative sequence via sequences comparison or k-mer analyses. Although phylogenetic informations offer an improved taxonomic affiliation compared with the state of art of classification algorithms (Taib *et al*., 2013 PLoS One. 8:e58950), studies of microbial phylogenetic relatedness are rare. We have developed ePANAM a tool for the phylogenetic analysis of high-throughput sequences generated by modern sequencing platforms (454-pyro, Illumina MiSeq), and the comparison of community structure within and among multiple samples. In addition to traditional measures of species diversity (Chao, Shannon), a set of phylogenetic diversity measures (PD, NRI, NTI) were implemented. These indices are expected to be more effective at revealing underlying ecological patterns. The robustness of such indices in the context of high throughputs microbial amplification of SSU rDNA gene variable regions is investigated.

Link PDF

Post-077 (#40) - Yannick LAURENT - MetaBiote® Online: a user-friendly and dedicated tool for the taxonomic profiling of microbial communities

MetaBiote® Online: a user-friendly and dedicated tool for the taxonomic profiling of microbial communities

Yannick LAURENT[1], Jérémie DENONFOUX[2], Nathalie ADELE[2], Stéphanie FERREIRA[2] and Frédéric TEXIER[1]

[1] GENOSCREEN, Service Bioinformatique, 1 rue du Professeur Calmette, 59000 Lille, France
[2] GENOSCREEN, Service Recherche et Développement Santé-Environnement, 1 rue du Professeur Calmette, 59000 Lille, France

Corresponding author: stephanie.ferreira@genoscreen.fr

The subsequent development of next generation sequencing technologies has led metagenomics studies became a powerful tool for the analysis of microbial communities from various environments. To date, high-throughput sequencing of PCR-amplified conserved genes such as 16S genes remains the most accessible and widely used approach to describe bacterial organisms in a community. Nevertheless, these approaches generate huge amount of data that needs to be processed for the extraction of relevant information. Although several tools are now available, their use and configuration remains complicated to implement and require knowledge in bioinformatics and computer science. To make these analytical tools available to all, we developed a fully automatic tool implemented with a user-friendly web interface. Metabiote® Online appeared suitable for management of Roche/454 as well as Illumina sequencers. Here, we will present the complete pipeline based on QIIME software and internal developments comprising (i) quality control and read preprocessing steps, (ii) chimeras detection step, (iii) complete linkage clustering into Operational Taxonomic Units (OTUs), (iv) removal of singletons OTUs, (v) taxonomic affiliation against dedicated database Greengene and (vi) the computation of taxonomic relative abundances, diversity index associated with each sample as well as a graphical interpretation. Our tool being first developed for the management of Roche/454 sequencing data, we will focus on the improvements to the steps of reads preprocessing and chimera detection that are necessary for the management of MiSeq 2×300bp sequencing data.

Link PDF

Post-078 (#78) - Jimmy H.W. SAW - A novel phylum-level archaea characterized by combining single-cell and metagenomic approaches

A novel phylum-level archaea characterized by combining single-cell and metagenomic approaches

Jimmy H.W. SAW[1], Guillaume REBOUL[1,2], Anja SPANG[1], Anders E. LIND[1], Dan COLMAN[3], Cristina TAKACS-VESBACH[3], Lionel GUY[1] and Thijs J.G. ETTEMA[1,4]

[1] Department of Cell and Molecular Biology, Uppsala University, Box 596, 75123 Uppsala, Sweden
[2] Master de Bioinformatique, Normandie Université, 76821 Mont-Saint-Aignan cedex
[3] Department of Biology, University of New Mexico, Albuquerque, NM, USA
[4] Single-cell Genomics Platform, Science for Life Laboratory, Uppsala, Sweden

Corresponding Author: jimmy.saw@icm.uu.se

In the context of the "microbial dark-matter" (Rinke et al. 2013) and the next-generation sequencing (NGS), multiple current projects try to resolve this unknown part of the tree of life. In this perspective, two methods are available: single cell genomics (SCG) and metagenomics. We assume that both of these methods can and should be combined into a unique approach to obtain better results on de novo genome assemblies and we applied it on an unknown archaeon (named N21) found in an environmental sample. For that, a pipeline was developed to clean the reads from the Single-cell Amplified Genome (SAG) sequencing run and an assembly was performed using these reads. Then, multiple tools were used to find, in two steps, the metagenomics contigs that belong to the N21 strain in the metagenomics pool of contigs generated from a high- throughput sequencing of the environmental sample. Finally, an assembly was made using the SAG reads and the metagenomics reads identified during the last two steps. A phylogenetic tree was constructed using the archaeal cluster orthologous genes (arCOGs) to place the strain into the archaeal domain of life according to the predicted genes found in the assembly. Also, we are pursue further analysis on N21 to annotate and identify its metabolism.

Link PDF

Post-080 (#83) - Lucas AUER - How to design an efficient and robust pipeline for 16S rRNA-gene sequence analysis to improve our understanding on microbial communities?

How to design an efficient and robust pipeline for 16S rRNA-gene sequence analysis to improve our understanding on microbial communities?

Lucas AUER[1], Laurent CAUQUIL[2], Stéphane CHAILLOU[3], Céline DELBES[4], Eric DUGAT-BONY[5], Hélène FALENTIN[6], Guillermina HERNADEZ-RAQUET[1], Mahendra MARIADASSOU[7], Aurélie NICOLAS[6], Géraldine PASCAL[2], Etienne RIFA[4], Sophie SCHBATH[7], Anne-Laure ABRAHAM[3] and Sébastien TERRAT[8]

[1] LISBP UMR 792 INRA, CNRS, INSA, 31077 Toulouse, France.
[2] INRA, INPT ENSAT, INPT ENVT, UMR1388, 31326 Castanet-Tolosan, France.
[3] INRA, UMR1319 Micalis, AgroParisTech, 78352 Jouy-en-Josas, France.
[4] INRA, UR 545 URF, 15000 Aurillac, France.
[5] INRA, UMR 782 GMPA, AgroParisTech, 78850 Thiverval-Grignon, France.
[6] INRA, UMR 1253 STLO, Agrocampus Ouest, 35042 Rennes Cedex, France.
[7] INRA, Unité MaIAGE, 78352 Jouy-en-Josas, France.
[8] INRA, UMR 1347 Agroecologie, BP 86510, 21000 Dijon, France

Corresponding authors: sebastien.terrat@dijon.inra.fr & anne-laure.abraham@jouy.inra.fr

Microorganisms are considered one of the most important players involved in different environmental processes and services including nutrient cycling, pollutants attenuation as well as plant, animal and human health. In order to understand the functioning of microbial ecosystems and their impact on ecosystem processes it is need to accurately assess their composition and response to environmental constraints. The application of NGS targeting the 16S/18S rRNA-gene has revolutionized the characterization of complex microbial ecosystems. However, although it is now possible to generate hundreds of thousands of sequence reads at lower cost, the appropriate interpretation of the obtained data is still challenging: potential source errors including amplification biases, contamination, sequencing artefacts and mistakes in the taxonomic affiliation can lead to a misinterpretation of microbial community diversity. The recent sequencing Illumina Mi-Seq technology produce larger number of sequences at lower cost, but tools and pipelines should be adapted for huge dataset. With the objective of defining best practices to analyse 16S/18S rRNA-gene sequence data, the *Metagenomics, species identification, phylogeny* pole of INRA was created to put together experience of biologist, bioinformaticians and statisticians. We will discuss the different approaches proposed by the available bioinformatics tools. We will give some recommendations, based on recent bibliography studies, and on the comparative studies performed by the pole. Finally, we will propose training sessions mainly destined to microbiologist interested on analysing 16S rRNA-gene NGS data.

Link PDF

Post-081 (#84) - Charlie PAUVERT - Accurate taxonomy assignments in cheeses ecosystems via a metagenomic approach

Accurate taxonomy assignments in cheeses ecosystems via a metagenomic approach

Charlie PAUVERT[1,2], Anne-Laure ABRAHAM[1,2], Mathieu ALMEIDA[1,2], Nicolas PONS[3], Mahendra MARIADASSOU[4], Emna ACHOURI, Sophie SCHBATH[4] and Pierre RENAULT[1,2]

[1] INRA, UMR1319 Micalis, 78352 Jouy-en-Josas, France
[2] AgroParisTech, UMR Micalis, 78352 Jouy-en-Josas, France
[3] INRA, Metagenopolis, 78352 Jouy-en-Josas, France
[4] INRA, Unité MaIAGE, 78352 Jouy-en-Josas, France

Corresponding author: charlie.pauvert@jouy.inra.fr

The manufacturing process of cheese, as for most fermented food, involves a complex flora, which is composed of bacteria, yeast and filamentous fungi. The exact composition of most cheeses is not completely known. Further understandings of cheeses ecosystems and control of cheese product's constant quality both needs a better characterization of the cheese flora and a precise taxonomic identification. But the key points to tackle are low abundant species and identification up to the strain level. Hundreds of genomes extracted from dairy products are currently available in genomic databases. Precise taxonomic identification is one of the key issues of our project, hence metagenomic shotgun sequencing approach has been chosen and applied to 40 samples of cheese ecosystems. Several current metagenomic tools are based on a set of gene markers, or on the k-mer composition of the reads, but few are able to identify species up to the strain level. We develop an original approach able to identify precisely species, and strains if reference genomes are available. We can also identify species or genus present when genomes references have down to 90% of identity with the metagenomic reads. Our method is based on the mapping of metagenomic reads on a set of reference genomes, completed with the adequation to a statistical model of CDS coverage by metagenomic reads. We test this method on several datasets, including simulated metagenomic reads mapped on reference genomes belonging to close species. We will also present examples on cheese ecosystems.

Link PDF

Post-082 (#93) - Nicolas PARISOT - Selection of oligonucleotide probes for high-throughput exploration of complex environments

Selection of oligonucleotide probes for high-throughput exploration of complex environments

Nicolas PARISOT[1], Jérémie DENONFOUX[2], Céline RIBIERE[3], Cyrielle GASC[3], Faouzi JAZIRI[3], Pierre PEYRET[3] and Eric PEYRETAILLADE[3]

[1] UMR 203 BF2I, Biologie Fonctionnelle Insectes et Interactions, INRA, INSA-Lyon, Université de Lyon, 69621 Villeurbanne, France.

[2] GENOSCREEN, Campus de l'Institut Pasteur, 59000 Lille, France.

[3] Clermont Université, Université d'Auvergne, EA 4678 CIDAM, BP 10448, 63001 Clermont-Ferrand, France.

Corresponding authors: nicolas.parisot@insa-lyon.fr & eric.peyretaillade@udamail.fr

The evolution of high-throughput molecular tools such as DNA microarrays or gene capture methods coupled with next-generation sequencing opened the door to unprecedented opportunities in microbial ecology to appraise the taxonomic and functional diversity of microorganisms within complex environments. The success of these high-throughput strategies, however, strongly relies on the quality of designed probes. Consequently, probe design is of critical importance and therefore multiple parameters should be considered in order to ensure the selection of sensitive, specific and explorative probes.

The KASpOD (*K-mer based Algorithm for high-Specific Oligonucleotide Design*) software has been developed to design such probes. This multipurpose tool was implemented to design probes from the exponentially growing sequence datasets in microbial ecology by using highly parallel computing architectures and an innovative *k*-mers based strategy that allowed overcoming major limitations in this field. The high quality designed probe sets were used to develop innovative strategies in microbial ecology including a phylogenetic microarray and its probe database (PhylOPDb), a gene capture approach and a taxonomic binning algorithm for metagenomic data. These approaches can be carried out for various applications including better understanding of microbial ecosystems, bioremediation monitoring or identification of pathogens (eukaryotes, prokaryotes and viruses).

KASpOD is provided as both a web service (http://g2im.u-clermont1.fr/kaspod/) and a stand-alone package for large computations.

Link PDF

Post-083 (#97) - Lucas AUER - FROGS: Find Rapidly OTU with Galaxy Solution

FROGS: Find Rapidly OTU with Galaxy Solution

Lucas AUER[1*], Frédéric ESCUDIE[2*], Maria BERNARD[3], Laurent CAUQUIL[4], Katia VIDAL[4], Sarah MAMAN[4], Mahendra MARIADASSOU[5], Guillermina HERNANDEZ-RAQUET[1] and Géraldine PASCAL[4]

[1] Université de Toulouse; INSA, UPS, INP; LISBP, 31077 Toulouse, France
[2] INRA UMR792, CNRS, UMR5504, 31400 Toulouse, France
[3] INRA, MIAT, 31326 Castanet-Tolosan, France
[4] INRA, UMR1313, SIGENAE, 78352 Jouy-en-Josas, France
[5] INRA, INPT ENSAT, INPT ENVT, UMR1388, 31326 Castanet-Tolosan, France
[6] INRA, Unité MaIAGE, 78352 Jouy-en-Josas, France

(*) These authors have contributed equally to the present work.

Corresponding author: geraldine.pascal@toulouse.inra.fr

High-throughput sequencing of 16S/18S RNA amplicons has opened new horizons in the study of microbe communities. With the sequencing at great depth the current processing pipelines struggle to run rapidly and the most effective solutions are often designed for specialists. These tools are designed to give both the abundance table of operational taxonomic units (OTUs) and their taxonomic affiliation. In this context we developed the pipeline FROGS*: « Find Rapidly OTU with Galaxy Solution »*. Developed for the Galaxy platform, FROGS was designed to be run in two modes: with or without demultiplexed sequences. A preprocessing tool merges paired sequences into contigs with flash, cleans the data with cutadapt, deletes the chimeras with UCHIME and dereplicates sequences with a home-made python script. The clusterization tool runs with SWARM that uses a local clustering threshold, not a global clustering threshold like other softwares do. This tool generate the OTU's abundance table. The affiliation tool returns taxonomic affiliation for each OTU using both RDPClassifier and NCBI Blast+ on Silva SSU 119. And finally, the postprocessing tool allows users to process this table with the user-specified filters and provides statistical results and graphical illustrations of these data. FROGS has been developed to be very fast even on large amounts of MiSeq data in using cutting-edge tools and an optimized design, also it is portable on all Galaxy platforms with a minimum of informatics and architecture dependencies. FROGS was tested on several simulated data sets. The tool has been extremely rapid, robust and highly sensitive for the detection of OTU with very few false positives compared to other pipelines widely used by the community.

Link PDF

Gene capture by hybridization for accurate taxonomic affiliation

Cyrielle GASC[1*], Céline RIBIERE[1*], Clémence DEFOIS[1], Nicolas PARISOT[2], Thierry ALLARIO[1,3], Bérénice BATUT[1], Réjane BEUGNOT[1], Éric PEYRETAILLADE[1] and Pierre PEYRET[1]

[1] EA 4678 CIDAM, Université d'Auvergne, BP 10448, 63001 Clermont-Ferrand, France
[2] UMR 203 BF2I, INRA/INSA-Lyon/Université de Lyon, 69621 Villeurbanne, France
[3] UMR 547 PIAF, INRA/Université Blaise Pascal, 63171 Aubière, France
[*] These authors contributed equally to this work.

Corresponding authors: cyrielle.gasc@udamail.fr & pierre.peyret@udamail.fr

Microbial communities show the greatest organisms diversity on earth, which remains difficultly accessible even with current direct sequencing approaches. Therefore, reducing the complexity of metagenomic samples is necessary to explore microbial diversity. Through the PCR amplification of phylogenetic biomarkers, barcoding enables an easy access to microbial community structure. However, amplification biases and shorter targeted regions limit exhaustive diversity description and precise affiliations. Other approaches providing information on complete phylogenetic biomarkers and flanking regions are thus crucial to overcome those limitations.

Based on a solution hybrid selection method combined with next-generation sequencing, we developed an innovative gene capture approach to enable the enrichment in several kilobase pairs DNA fragments harbouring biomarkers of interest from complex environments. To benefit from capture advantages, a suitable bioinformatics data treatment according to two complementary workflows is necessary. Taxonomic affiliation can be restricted to reads spanning the complete phylogenetic biomarker. Phylogenetic assignation can also be improved using flanking regions thanks to contig assembly.

Analyses performed show that gene capture reveals a broader and more reliable taxonomic diversity than observed with barcoding approach. Moreover, gene capture enables to access the phylogenetic information of the complete targeted biomarker and its informative flanking regions, providing a reliable and resolvent affiliation until finest taxonomic ranks (species, strain). This innovative method combined with an original bioinformatics treatment thus enables a better exploration of microbial ecosystems.

Link PDF

Sensitive recovery of wild-type markers from metatranscriptomic data

Pierre PERICARD[1,2], Samuel BLANQUART[1,2] and Hélène TOUZET[1,2]

1 CRIStAL, UMR9189 CNRS, Université Lille 1, Bat M3, Cité Scientifique, 59655 Villeneuve d'Ascq, Cedex, France
2 Inria Lille Nord Europe, Parc scientifique de la Haute Borne, 40, avenue Halley, Bat A, Park Plaza, 59650 Villeneuve d'Ascq, France

Corresponding author: pierre.pericard@ed.univ-lille1.fr

Metatranscriptomics is an emerging domain with great potential for better understanding metagenomic communities dynamics. However, few methods are available today to exploit total RNA high-throughput sequencing datasets. Here, we propose a novel protocol to analyze those datasets by focusing on a given marker of interest. First, reads are sorted out to identify the ones matching the marker, which are mapped onto reference sequences from a taxonomically annotated database using SortMeRNA2. Then, the mapped reads are assembled within a graph model by taking advantage of sequence conservation and taxonomic information, while accounting for sequencing errors. The goal is to recover the full length transcripts corresponding to the marker, from as many wild-type species from the sample as possible. Resulting contigs can then be used for functional analysis or high-resolution taxonomic identification, using 16S rRNA or any other selection of markers.

Link PDF

panam2Pprobe, a tool for designing primers from high-throughput sequencing

Anne OUDART[1,2,3], Didier DEBROAS[1,2] and Najwa TAIB[1,2,3]

1 Laboratoire « Microorganismes : Génome et Environnement », Université Blaise Pascal, BP 10448, 63000 Clermont-Ferrand, France
2 LMGE, UMR 6023 CNRS, 63171, Aubière France
3 Laboratoire de Biométrie et Biologie Evolutive, Université Claude Bernard – Lyon 1, 43 bd. Du 11 Novembre 1918, 69622 Villeurbanne, France

Corresponding author: didier.debroas@univ-ubpclermont.fr

The development of molecular techniques has brought new insights into our understanding of the functioning and the structure of microbial communities. Indeed, the use of high throughput sequencing coupled with the golden standard for the microbial identification (16S and 18S rDNA), have enabled for the characterization of species up to now unknown and/or rare. These studies showed that the specific richness had been underevaluated for a long time, especially for the rare biosphere. According to these studies, phylogenetic affiliation is more accurate for taxonomic assignment than similarity research and probabilistic methods, and allows for the identification of complex and under studied microbial communities. Moreover, phylogenetic affiliations delineate monophyletic groups, which can be used to identify conserved regions.

In this context, we developed a program, panam2probe for designing specific primers to target clades of interest. As the universal primers were designed on known and abundant microorganisms, they usually miss the rare fraction. panam2probe enables to overcome this limitation. From a phylogenetic tree and a taxonomic affiliation, it extracts the sequences of a given clade with respect to its monophyly, and generates automatically a list of specific primers. panam2probe makes it possible to investigate typical clades allowing for the discovery of new lineages. We used the primers designed by this tool to investigate the rare taxa in lacustrine ecosystems.

Link PDF

Virome data analysis by cross-assemblage

Corentin HOCHART[1] and Didier DEBROAS[1]

[1] LMGE, UMR-6023 University Clermont2 Blaise Pascal, 24 avenue des Landais, Campus des Cézeaux - BP 80026, 63171 Aubière, Cedex, France

Corresponding author: Corentin.Hochart@etudiant.univ-bpclermont.fr

Metagenomic approaches by high-throughput sequencing are increasingly used in order to identify viral populations in ecosystems, either by direct affiliation of sequences obtained either by annotating viral genomes after assembly. These approaches by high-throughput sequencing producing millions of short sequences, afford a new angle of virome analysis, ie the viral fraction of metagenomic data.

However, when analyzing the viromes, numerous sequences (~ 70%) remain unaffiliated taxonomically or non-functionally characterized due to the non completeness of virus database or sequences size. Secondly, the assembly will only partially reconstituted (ie contig) viral genomes.

In order to optimize results of the assembly, it is possible to combine different genomes from the same environment. This cross-assembly approach should allow to assemble not individually assembled read and produce a greater number of contigs and / or longer contigs. The affiliation of assembled reads will be evaluated based on its genomic context, ie the reads that are affiliated contig, will be affiliated too.

The purpose of this study is to test the contribution of cross-assembly to the study of lacustri viromes to eventually later be able to apply to the analysis of viral dynamics in aquatic ecosystems.

We will apply this approach to the analysis of 6 viromes obtained by Illumina sequencing (MiSeq) of environmental sets sampled in aquatic ecosystems. The viromes were individually assembled and assembled as well together with the help of two assemblers (Meta Ray and IDBA). In this work, we studied the assembly quality and the cross-assembly contribution on reads annotation, comparing their individual affiliation and affiliation in a genomic context.

Link PDF

Post-088 (#149) - Catherine BRETON - Métabarcode et réseaux de co-occurence révèlent l'impact des perturbations sur les communautés associées aux herbiers de zostères

Métabarcode et réseaux de co-occurence révèlent l'impact des perturbations sur les communautés associées aux herbiers de zostères

Catherine BRETON[1] and Sophie ARNAUD-HAOND[1]

[1] Ifremer UMR MARBEC (Marine Biodiversity, Exploitation and Conservation), Bd Jean Monnet, BP 171, 34203 Sète Cedex, France.

Corresponding authors: sophie.arnaud@ifremer.fr & catherine.breton02@univ-montp2.fr

Dans le contexte actuel de changements globaux, l'érosion de la biodiversité représente un défi majeur du XXI[e] siècle. Les conséquences à court terme résident dans des évènements climatiques d'une extrême violence. Sur le fonctionnement des écosystèmes, elles sont difficiles à prédire, mais la structure des communautés sera impactée directement par l'environnement et les changements d'aire de distribution des espèces. L'avènement des NGS engendre un nombre important de données de séquences permettant d'établir une description plus exhaustive de la biodiversité constitutive des communautés par metaborcoding. Afin d'estimer l'impact des perturbations sur les communautés de métazoaires appartenant aux herbiers de zostères (*Zostera nolti*) dans deux localités, le Bassin d'Arcachon et Etang de Thau, nous avons adapté un protocole de metabarcoding basé sur le séquençage haut débit. Des échantillons de sédiments ont été collectés en triplicats le long d'un gradient de perturbation, afin de minimiser la dominance des organismes les plus gros par tamisage des carottiers par fractions (1mm, 500µm et 250µm). Deux marqueurs moléculaires mitochondrial (Cytochrome oxydase I) et nucléaire (petite sous-unité 18S de l'ARN ribosomal) ont été utilisés. La procédure bio-informatique d'analyse choisie repose sur le pipeline Qiime sous linux, utilisant des bases de données locales afin d'optimiser les procédures de blast. La description de la biodiversité étant dépendante de la qualité des bases de données de référence, les deux marqueurs choisis représentent un compromis car ils présentent des affinités différentes avec les grands taxa constitutifs des communautés. Les résultats ont montré une modification de la diversité alpha qui présente un profil similaire à celui attendu sous l'hypothèse de « perturbation intermédiaire », mais la complémentarité des marqueurs COI et 18S révèlent plus finement un remplacement des communautés accompagnant l'impact au-delà d'un certain seuil. Ce basculement des communautés est plus clairement mis en évidence par une analyse de réseaux de co-occurrence d'Unités Taxonomiques Opérationelles (MOTUs).

Link PDF

BIOLOGIE TRANSLATIONNELLE & PHARMACOGENOMIQUE

TRANSLATIONAL BIOLOGY & PHARMACOGENOMICS

Thématique - Topic 08

Post-089 (#81) - Stéphanie BORNES - Influence des procédés de fabrication industrielle sur l'expression génique du Lactobacillus rhamnosus Lcr35®

Influence des procédés de fabrication industrielle sur l'expression génique du Lactobacillus rhamnosus Lcr35®

Stéphanie BORNES[1], Caroline DAUSSET[2], Elina ALATERRE[2], Claudia MULLER[2], Philippe VEISSEIRE[1] and Adrien NIVOLIEZ[2]

[1] Laboratoire de Biologie, IUT Aurillac, Clermont Université, 63000 Clermont-Ferrand
[2] Département R&D - Probionov, Rue des frères Lumières 15130 Arpajon-sur-Cère, France

Corresponding author: a.nivoliez@probionov.com

Les « live biotherapeutic products» représentent une alternative aux traitements traditionnels lors de pathologies dues à un déséquilibre du microbiote. Les « live biotherapeutic microorganisms», sont principalement sélectionnés et caractérisés sur la base de leurs propriétés physiologiques natives lors de tests *in vitro*. Cependant, il a été démontré un impact des procédés de fabrication sur les propriétés probiotiques de ces souches bactériennes. La société Probionov a mis en évidence l'influence des formulations industrielles sur les propriétés anti-microbiennes de la souche *Lactobacillus rhamnosus* Lcr35. Pour étudier l'impact des formulations sur la souche Lcr35, une puce ADN à façon (Nimblegen®) a été développée à partir de l'ensemble des données de séquençage et d'annotation de 25 espèces appartenant au genre Lactobacillus. Son utilisation permet de suivre l'expression différentielle de l'ensemble des gènes de Lcr35 en fonction des conditions expérimentales. Les données brutes ont été normalisées avec l'algorithme RMA puis analysées à l'aide des modules affy, limma et stat du logiciel R. Les résultats significatifs ainsi obtenus ont ensuite été validés par des techniques ciblées de qRT-PCR à l'aide de primers spécifiques. Ces derniers ont été dessinés en utilisant un programme BioPerl, faisant appel aux modules primer3 et fuzznuc, puis leur spécificité a été validée par qRT-PCR. Les données, obtenues avec le logiciel Rotor-Gene Q (Qiagen), ont été analysées en utilisant la méthode des $2^{-\Delta\Delta Ct}$ avec comme référence le gène de la GAPDH. Ainsi, nous avons montré que le procédé de fabrication joue un rôle sur l'expression des protéines de surface du Lcr35. La formulation pour application vaginale, Lcr regenerans® (Gynophilus), induit une modification du profil d'expression génique de la souche Lcr35, dont la surexpression de molécules antimicrobiennes telles que les bactériocines. Au contraire, pour une même formulation, la forme galénique qu'elle soit en poudre, gélule ou comprimé, n'impacte pas les profils d'expression de la bactérie. En effet, l'étape de compression de la formulation Lcr regenerans n'entraîne aucune variation de l'expression génique du Lcr35 au sein d'un comprimé en comparaison de la poudre.

Link PDF

NutriQuantic : une Application Smartphone pour Déterminer l'Adéquation de la Prise Alimentaire vis-à-vis des Recommandations Nutritionnelles

Jules CISSOKO[1], Yves BOIRIE[1,2], Martine DUCLOS[1,2], Anthony FARDET[1], Romain GUIDOUX[3], Ludivine PARIS[1], Raksmey PHAN[4], Libo REN[5] and Sylvie ROUSSET[1]

[1] INRA Unité de Nutrition Humaine UMR 1019, CRNH d'Auvergne 63009 Clermont-Ferrand
[2] CHU G Montpied Services Nutrition Clinique et Médecine du Sport, 63003 Clermont-Ferrand
[3] 29bis rue du Sable d'Étampes, 63400 Chamalières
[4] LIMOS, UMR6025 CNRS 6158, Campus des Cézeaux, 63173 Aubière Cedex
[5] CRCGM, EA3849, 63000 Clermont-Ferrand

Corresponding author: sylvie.rousset@clermont.inra.fr

Le monde actuel est marqué par l'épidémie des maladies chroniques. A titre d'exemple, le surpoids et l'obésité touchent environ 50% de la population française (Obepi Roche, 2012). De plus, la surcharge pondérale prédispose à l'apparition du diabète, des cancers et des maladies cardio-vasculaires (Fardet et Boirie, 2014). Les coûts engendrés par la prise en charge des maladies chroniques sont considérables, à savoir 10 milliards d'euros/an en France (Lalanne, Caisse d'assurance Maladies, 2014). De ce fait, la prévention par la nutrition et l'activité physique est un enjeu majeur de santé publique. Parallèlement, le concept de « quantified-self », c'est-à-dire l'auto-évaluation de paramètres personnels, est en plein essor. Une multitude d'applications mobiles et d'objets connectés liés à la santé existent. Actuellement, plus d'un Français sur deux possède un smartphone (Baromètre du Marketing Mobile), expliquant ainsi l'intérêt grandissant pour le « quantified-self ». Cependant, la majorité des applications existantes ne sont pas validées scientifiquement (Loisel, 2014). Seulement 13% des applications concernant la prise alimentaire utilisent des données de composition nutritionnelles référencées. Dans ce contexte nous avons pensé l'élaboration de l'application NutriQuantic afin qu'elle soit originale, innovante, interactive et valide scientifiquement. Cette application a pour objectif d'évaluer la prise alimentaire de l'utilisateur et d'interagir avec lui afin de l'aider à atteindre un équilibre nutritionnel satisfaisant. Pour cela, des scores d'adhésion aux recommandations nutritionnelles, fournies par les guides français et internationaux, sont calculés à partir des catégories d'aliment consommés. Ce dernier point nous démarque des applications existantes puisqu'elles sont centrées sur la quantité (kcal) et non la qualité de l'apport. NutriQuantic permet donc de collecter les données de prises alimentaires reparties par catégorie d'aliment. Ces données sont alors envoyées sur la plateforme Web ActivCollector (https://activcollector.clermont.inra.fr/) pour le traitement algorithmique. Les résultats concernant l'équilibre nutritionnel quotidien et hebdomadaire sont retournés à l'utilisateur et intégrés dans des interfaces ergonomiques. Une première version de l'application a été testée en interne et sera testée au près d'une population plus large afin de la valider scientifiquement.

Link PDF

Post-091 (#137) - Nadia BESSOLTANE-BENTAHAR - RNAseq analysis: Evaluation of response to anastrozole and fulvestrant in postmenopausal women with estrogen receptor (ER)-positive breast cancer

RNAseq analysis: Evaluation of response to anastrozole and fulvestrant in postmenopausal women with estrogen receptor (ER)-positive breast cancer

Nadia BESSOLTANE-BENTAHAR[1], Virginie BERNARD[1], Patricia LEGOIX-NE[1], Thomas RIO-FRIO[1], Olivier DELATTRE[2], Celine CALLENS[3] and Florence LEREBOURS[4]

[1] Next Generation Sequencing platform, Institut Curie, 26 rue d'Ulm, Paris, France
[2] Genetics and Biology of Cancers Unit, Institut Curie, 26 rue d'Ulm, Paris, France
[3] Pharmacogenomics Unit, Department of Genetics, Institut Curie, 26 rue d'Ulm, Paris, France
[4] Department of Breast and Medical Oncology, Institut Curie, Hôpital René Huguenin, Saint-Cloud, France

Corresponding author: nadia.bessoltane-bentahar@curie.fr

Anastrozole (aromatase inhibitor) or fulvestrant (estrogen receptor (ER) antagonist) were used as neoadjuvant hormone therapy in postmenopausal women with ER-positive breast cancer included in CARMINA02 clinical trial. However some patients showed a resistance to these treatments. Thus we aim to identify, at the transcriptional level, the predictive markers of resistance or sensibility to anastrozole and fulvestrant, and describe the resistance mechanisms, for both treatments.

RNA sequencing was performed on ninety tumor samples from responder and non-responder patients treated with anastrozole or fulvestrant. These samples are from pre-treatment and post-treatment tumors (after 6 months of treatment). Three analyses are in progress. Differential gene expression is analyzed with the "Differential Expression analysis for Sequence count data" (DESeq2) package. Variant calling is processed following "Genome Analysis Toolkit" (GATK) recommendation for RNAseq data. Gene fusions detection is performed with "ChimeraScan, "TophatFusion" and "DeFuse" tools, followed by filtering to remove false positive fusions related to sequence similarities.

For each treatment, first, aiming to detect predictive markers of response, we compared predictions of gene fusions, gene expression, and variants, between responder and non-responder patients in pre-treatment biopsies. Then, we analysed the evolution of these variations over treatments by comparing samples before and after treatments.

This ongoing work will lead to the prediction of a large panel of variations for each sample. It will allow to better know whether variations are response specific or treatment specific.

Link PDF

SERVICES, RESSOURCES & INFRASTRUCTURES POUR LA BIOINFORMATIQUE

SERVICES, RESOURCES & INFRASTRUCTURE FOR BIOINFORMATICS

Thématique - Topic 09

Association Rule Mining for Metabolic Pathway Prediction

Imene BOUDELLIOUA[1], Rabie SAIDI[2], Maria MARTIN[2] and Victor SOLOVIEV[1]

[1] Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

[2] UniProt, European Bioinformatics Institute, Hinxton, United Kingdom

Corresponding author: imene.boudellioua@kaust.edu.sa

Prediction of chemical reactions and pathways is among the most challenging problems of systems biology. In this work, we are tackling the problem of metabolic pathway prediction in the context of metabolism. We developed ARBA (Association-Rule-Based Annotator), a system that utilizes machine learning methods, specifically rule mining techniques, to predict pathways associated with protein entries available in UniProtKB. Our system can be used to enhance the quality of automatically generated annotations as well as annotating unknown proteins. Moreover, this system will provide an insight into the conservation of pathways across prokaryotes that differ in their taxonomic classification. ARBA was successfully applied to gain knowledge about pathway annotation type in all UniProtKB-SwissProt entries with manual assertion evidence corresponding to a specified prokaryotic taxon. ARBA presents this knowledge in the form of association rules that takes into account the organism taxonomy and the InterPro signature matches of protein sequences. These rules are then filtered efficiently using the Skyline operator in order to select the best representative rules in terms of several interestingness metrics to effectively minimize false positives as well as eliminating rules generated out of pure randomness. The resulting rules could be used as models to infer pathways for poorly annotated TrEMBL entries. We carried out an experimental study of the performance of ARBA on real datasets representing various prokaryotic taxa to demonstrate the robustness of our system. We found that ARBA achieved an average overall accuracy as high as 99.99%, F-measure of 0.987, precision of 0.991, and recall of 0.982.

Link PDF

Post-093 (#9) - Cécile MONAT - TOGGLE: TOolbox for Generic nGs analyses

TOGGLE: TOolbox for Generic nGs analyses

Cécile MONAT[1], Christine TRANCHANT-DUBREUIL[1], Ayité KOUGBEADJO[2], Maryline SUMMO[2], Cédric FARCY[2], Mawussé AGBESSI[1] and François SABOT[1]

[1] UMR DIADE IRD/UM/CIRAD, 911 Avenue Agropolis, 34934 Montpellier Cedex 5, FRANCE
[2] UMR AGAP CIRAD/INRA/SupAgro, TA A-108/03 - Avenue Agropolis, 34398 Montpellier Cedex 5, FRANC

Corresponding author: christine.tranchant@ird.fr

In the 80s and 90s, access to the DNA sequence through Sanger sequencing have revolutionized the way we do genetic analyses. This revolution have been accentuated since 2005 and the appearance of the so-called Next Generation Sequencing (NGS) technologies. Those technologies indeed can generate very large amounts of data, in a very short period of time, and for a limited cost. Thus, they have spread very rapidly in the biological community to answer a large array of different questions. At the same time, a large number of Bioinformatics softwares and methods were developed and publicly released to work with those data. Some of those softwares are widely used, some were only in their initial publication. Moreover, the vast majority of those softwares are command-lines tools, designed to work on powerful computers and High-Performing Calculation infrastructures. Thus, they are generally not easy to use for wet-lab biologists, as more-than-basic informatics skills are requested to manipulate them. The TOGGLE suite (for TOols for Generic nGs anaLysEs) was developed to optimize the creation of new NGS analysis workflows. What we propose here is a set of 12 packages, gathering more than 125 modules (currently) designed for a fast design of robust and reliable pipelines. Each package represents either an NGS software manipulation or a set of dedicated tools. These UNIX packages are written in Perl, with unitary modules, and are the most generic possible. To ensure the stability of each module functions through the different installation and through the development, a series of unit tests, using a Test Perl framework, were developed. The whole tools and the test data are available on the GitHub of the project: https://github.com/SouthGreenPlatform/TOGGLE.

Link PDF

Wheat@URGI website for genomics, genetics and phenomics wheat data

Michael Alaux[1], Françoise ALFAMA[1], Véronique JAMILLOUX[1], Raphaël FLORES[1], Mathilde LAINÉ[1], Nacer MOHELLIBI[1] Cyril POMMIER[1], Erik KIMMEL[1], Célia MICHOTEY[1], Claire GUERCHE[1], Mikaël LOAEC[1], Delphine STEINBACH[1] and Hadi QUESNEVILLE[1]

[1] INRA, UR1164 URGI - Research Unit in Genomics-Info, INRA de Versailles, Route de Saint-Cyr, Versailles, 78026 France

Corresponding author: michael.alaux@versailles.inra.fr

URGI is a genomics and bioinformatics research unit at INRA dedicated to plants and crop parasites. We develop and maintain a genomic and genetic Information System : GnpIS (Steinbach *et al.*, Database 2013, doi: 10.1093/database/bat058). URGI sets up a website dedicated to wheat, called Wheat@URGI: http://wheat-urgi.versailles.inra.fr/

Wheat genomics, genetics and phenomics data could be searched by (i) a google-like tool, (ii) well-known tools like BLAST, GBrowse/JBrowse, Intermine, Biomart, (iii) homemade GnpIS interfaces using Java and GWT technologies.

Genomics data in Wheat@URGI are the data of the International Wheat Genome Sequencing Consortium (IWGSC) hosted in a dedicated section, the "Sequence Repository" (http://wheat-urgi.versailles.inra.fr/Seq-Repository). The data stored in the Sequence Repository are the wheat survey sequence, the chromosome reference sequence (at the moment the 3B), the genes and annotations, the physical maps, the RNA-Seq and the variations data. It allows the user to download, display and query the IWGSC sequences and physical maps. This resource is widely used at the international level with more than 180 000 BLAST searches performed and 200 000 wheat browsers pages viewed since 2012.

We developed a “wheat” filter in GnpIS to allow crossing the wheat genomic data with the germplasm, markers, QTLs, SNPs, expression and phenotypes data (http://wheat-urgi.versailles.inra.fr/Data). Moreover association and genomic selection data are in the process of integration in the information system.

Link PDF

Framboisine: A mini-cluster solution for the Southern Countries

Ndomassi TANDO[1], Christine TRANCHAND-DUBREUIL[1,2], Eric MOUTTET[3], Stephane BANON[3] and Francois SABOT[2]

[1] iTROP platform, UMR DIADE IRD/UM, 911 Avenue Agropolis BP64501, 34394 Montpellier, Cedex 5, France
[2] RICE team, UMR DIADE IRD/UM, 911 Avenue Agropolis BP64501, 34394 Montpellier, Cedex 5, France
[3] IS-DSI, IRD, 911 Avenue Agropolis BP64501, 34394 Montpellier, Cedex 5, France

Corresponding author: francois.sabot@ird.fr

L'objectif du projet Framboisine est de fournir une méthodologie d'assemblage et d'utilisation de mini- clusters de calcul formé de SingleBoard Computers SBC. Ces ordinateurs ultracompacts (format carte de crédit), de faible coût (de 35 à 100 U$D) et de faible consommation énergétique (de 2 à 5 W), sont basés sur des processeurs ARM de diverses générations (processeurs de téléphones portables), et avec des capacités de calcul pouvant aller aux QuadriCoeurs avec 2 Go de RAM. Il est possible de les mettre en cluster, i.e. de les monter en structure de calcul parallèle, et ainsi de pouvoir obtenir des machines de calcul certes à performance modérée, mais à très bas coût (moins de 2000 U$D par cluster). De telles machines pourront servir à la fois (i) dans le cadre de formation au Sud, où une des limites de ces formations est souvent l'accès à des ressources de calcul suffisantes et (ii) dans le cadre de projets de recherche réalisé par nos partenaires localement. Il existe des plate-formes de calcul distantes mais l'accès à une connexion réseau stable est parfois problématique. Une machine de ce type coûtant moins de 2000 U$D, il sera possible d'en emmener une à chaque fois qu'une école thématique ou une formation au Sud portant sur du calcul parallèle sera effectuée (génomique, épidémiologie, simulation). Ceci permettra de travailler directement sur la machine et de pouvoir la laisser sur place au bénéfice du partenaire ayant co-organisé la formation, et ainsi n'importe quel partenaire du Sud de pouvoir travailler sur des approches demandant du calcul parallèle (génomique, écologie, épidémiologie, simulation météo, astronomie et radioastronomie…) de manière autonome. Actuellement trois prototypes fonctionnels existent, basés sur le RaspberryPi, le CubieTruck et le OlinuXino-A20Lime2. L'assemblage final sera réalisé en utilisant des Lego Technics, seul matériel convenant à une telle utilisation (pas d'outils, pas de découpe, résistance à la chaleur, à l'humidité, pas de conduction électrique, disponibilité mondiale). L'avancée des travaux est régulièrement présentée sur le Facebook du projet https://www.facebook.com/FramboisineLego.

Link PDF

Development of a generic indexing tool to optimize the use of biological data

Luyen LE NGOC[1], Stéphane JOUANNIC[1,2], Pascal GANTET[1,2] and Pierre LARMANDE[2]

[1] LMI RICE, IRD, Hanoi, Vietnam
[2] UMR DIADE, IRD, Montpellier, France

Corresponding authors: pierre.larmande@ird.fr

This application is developed in the context of studies of genetic and phenotypic diversity in Asian and African rice (*Oryza sativa* and *Oryza glaberrima*). The objective of these studies is to identify by association genetics approaches some genes of interest in order to understand biological processes related to plant development and plasticity or disease resistance and their exploitation by breeding programs. These studies require handling large volumes of data that are heterogeneous and stored in different formats (Excel file, structured or semi-structured text, images, etc.). The volume and diversity of data can be a challenge for researchers for their optimal exploitation. The design of relational databases appears limited and not scalable. In this context, we have developed a tool for integration and generic indexing to navigate, share and annotate these data in order to exploit them. For this purpose, the project is based on a NoSQL management system Document-oriented - MongoDB, allowing data to be dynamically organized and modeled. The innovative aspect of this project is the development of a scalable system that allows users to perform all the steps from data integration to the query formulation.

Link PDF

Integration of RSAT tools in Galaxy with SOAP web services

Jocelyn BRAYET[1,2,3], Céline HERNANDEZ[4], Claire RIOUALEN[5,6], Alban LERMINE[1,2,3], Jacques VAN HELDEN[5,6], Morgane THOMAS-CHOLLIER[4] and Nicolas SERVANT[1,2,3]

1 Institut Curie, 26 rue d'Ulm, 75248 Paris, France
2 INSERM, UMR900, 26 rue d'Ulm, 75248 Paris, France
3 Mines ParisTech, 35 rue Saint-Honoré, 77300 Fontainebleau, France
4 Computational Systems Biology, CNRS UMR 8197, INSERM U1024, Institut de Biologie de l'Ecole Normale Supérieure, 46 Rue d'Ulm, 75005 Paris, France
5 INSERM, UMR1090 TAGC, 163 avenue de Luminy, 13288 Marseille, France
6 Aix-Marseille Université, 58 boulevard Charles Livon, 13284 Marseille, France

Corresponding author: jocelyn.brayet@curie.fr

Regulatory Sequence Analysis Tools (RSAT [1]) is a software suite proposing a series of modular computer programs dedicated to the detection of cis-regulatory element in genomic sequences. RSAT is accessible by three ways: RSAT website (http://rsat.eu), local installation, and remote programmatic access through SOAP/WSDL web services [2]. Galaxy is a web-based platform, designed to provide easy access to a versatile toolbox for users. Recently, the Nebula Galaxy instance [3] was proposed as a public server dedicated to ChIP-seq data analysis.

In order to expand the diversity of ChIP-seq bioinformatics solutions offered in Galaxy, we began to integrate the main RSAT tools in Galaxy. Galaxy wrapper acts as a client for the web services already supported by RSAT. Galaxy sends request to remotely execute the analysis on one of the available RSAT servers (local or away), and collects the results. This strategy can be extended to other web services.

In conclusion, this work expands the bioinformatics solutions publicly available with Galaxy for advanced ChIP-seq analysis and offers a reusable strategy to integrate web services in Galaxy.

[1] M. Thomas-Chollier, M. Defrance, A. Medina-Rivera, O. Sand, C. Herrmann, D. Thieffry and J. van Helden. RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Research*, 39:W86-W91, 2011.
[2] O. Sand, M. Thomas-Chollier, E. Vervisch.and J. van Helden. Analyzing multiple data sets by interconnecting RSAT programs via SOAP Web services: an example with ChIP-chip data. *Nature Protocols* 3:1604–1615, 2008.
[3] V. Boeva, A. Lermine, C. Barette, C. Guillouf, E. Barillot. Nebula--a web-server for advanced ChIP-seq data analysis. *Bioinformatics*, 28:2517-2519, 2012.

Link PDF

Post-098 (#25) - Martial BRIAND - Orthomcl-Companion : un outil d'aide à l'interprétation d'analyses de familles protéiques

Orthomcl-Companion : un outil d'aide à l'interprétation d'analyses de familles protéiques

Martial BRIAND[1], Corinne RANCUREL[2], Ludovic COTTRET[3] and Sébastien CARRERE[3]

[1] Institut de Recherche en Horticulture et Semences, UMR 1345, INRA, SFR 4207 QuaSaV, 42 rue Georges Morel, 49071 BEAUCOUZE CEDEX, France
[2] Institut Sophia Agrobiotech, UMR INRA-CNRS 1355/7254, 400 route des Chappes, 06903 Sophia-Antipolis, France
[3] Laboratoire des Interactions Plantes-Microorganismes, UMR INRA-CNRS 441/2594, Chemin de Borderouge, 31320 Castanet-Tolosan, France

Corresponding author: Sebastien.Carrere@toulouse.inra.fr

La génomique comparée est devenue un outil de base pour les biologistes afin de répondre à des questions aussi variées que la recherche de déterminants du pouvoir pathogène ou de la spécificité d'hôte de certaines espèces bactériennes, la duplication et l'expansion de famille de gènes dans un organisme. La multiplication des programmes de séquençage et l'automatisation des protocoles d'annotation structurale offrent aujourd'hui la possibilité d'effectuer de la génomique comparée à haut débit. Des outils, tel que OrthoMCL, existent afin de comparer les protéomes de dizaines d'organismes et de construire des familles de gènes homologues.

Cependant les résultats fournis par cet outil sont difficilement exploitables et interprétables pour un utilisateur ne maîtrisant pas les outils standards en ligne de commande ou un langage de script notamment quand un grand nombre de protéomes a été comparé. En effet, le fichier résultat classique est une matrice dont les lignes représentent les groupes de gènes homologues et les colonnes les identifiants des gènes constituants ce groupe et l'organisme dont ils sont issus, et ce dans des formats différents selon la version d' OrthoMCL utilisée.

Pour accompagner les utilisateurs biologistes dans l'exploitation de ces résultats, nous avons développé une application web qui permet entre autre d'accéder facilement au core-protéome et au pan-protéome, mais aussi aux protéines spécifiques de chaque organisme, ou encore de générer des matrices d'abondance et de présence/absence (profil phylogénétiques). En y associant des annotations InterPro, les données de sortie sont enrichies de tables d'occurrence de termes des différentes banques (ex: PFAM, GO). L'outil étant orienté utilisateur, les fichiers générés utilisent des formats standards (textes tabulés, fasta) pour une exploitation ultérieure facilitée. Des représentations graphiques sous forme d'histogrammes et de diagrammes de Venn interactifs sont proposées.

Link PDF

Post-099 (#27) - Aurélien BERNARD - TriAnnot: A workflow for plant genome automated structural and functional annotation – international distribution

TriAnnot: A workflow for plant genome automated structural and functional annotation – international distribution

Aurélien BERNARD[1,2], Nicolas GUILHOT[1,2] and Philippe LEROY[1,2]

1 INRA, UMR 1095 Genetics, Diversity and Ecophysiology of Cereals - GDEC, 5 Chemin de Beaulieu, 63039 Clermont-Ferrand cedex 2, France
2 Université Blaise Pascal, UMR 1095 Genetics, Diversity and Ecophysiology of Cereals - GDEC, 5 Chemin de Beaulieu, 63039 Clermont-Ferrand cedex 2, France

Corresponding author: philippe.leroy@clermont.inra.fr

Agronomical characters such as yield, biotic and abiotic resistance are determined by the genetic information carried by the plant genome. Within the framework of an international effort (IWGSC - http://www.wheatgenome.org) for obtaining a reference sequence of the bread wheat genome and to provide the scientific communities dealing with large and complex genomes a versatile, easy-to-use online annotation tool, we have developed the TriAnnot pipeline (Leroy *et al.* 2012 Frontiers in Plant Sciences, 3:1-14). TriAnnot has already been used to annotate the wheat chromosome 3B (Choulet *et al.* 2014 Science) and 4D (Helguera *et al.* 2015 Plant Science). Annotation of chromosome 1B is currently underway (INRA - GDEC), as well as the annotation of chromosome 7A in collaboration with the University of Murdoch (Australia). TriAnnot has the ambition to federate the international community around the annotation of the 21 wheat chromosomes and, in this perspective, the TriAnnot source code has been strongly improved to facilitate its deployment on external computing resources such as: IEB, Olomouc (Czech Republic); CSIRO, Kensington (Australia); National Research Council of Canada, Saskatoon (Canada); CRRI, Clermont-Ferrand and ABiMS CNRS bioinformatics platform, Roscoff (France). The development of a virtual machine is also underway in collaboration with ABiMS and the French Institute of bioinformatics (IFB). TriAnnot was also adapted for the annotation of other plant genomes such as barley, maize, rice and oak. A public instance of TriAnnot is currently deployed on the cluster of the URGI bioinformatics platform (Versailles) and usable through a user-friendly web interface or in command line. This poster presents the state of progress of all these installations as well as the technological challenges we had to overcome during the deployments.

Link PDF

Post-100 (#28) - Simon PENEL - Nouveaux développements dans la construction de HOGENOM une base de données de phylogénomique

Nouveaux développements dans la construction de HOGENOM une base de données de phylogénomique

Simon PENEL[1], Vincent DAUBIN[1], Pierre DUPUIS[1], Laurent DURET[1], Manolo GOUY[1], Dominique GUYOT[2], Vincent MIELE[1], Rémi PLANEL[1] and Guy PERRIERE[1,2]

[1] Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, Université Lyon 1, Bât. Grégor Mendel, 43 boulevard du 11 Novembre 1918, 69622 Villeurbanne Cedex, France

[2] Pôle Rhône-Alpes de Bioinformatique, Université Lyon 1, Bât. Grégor Mendel, 43 boulevard du 11 Novembre 1918, 69622 Villeurbanne Cedex

Contacting author: simon.penel@univ-lyon1.fr

La génomique comparative est une approche centrale dans l'analyse des séquences biologiques. Elle est utilisée depuis l'annotation et l'identification de régions fonctionnelles jusqu'à l'étude des processus évolutifs tels que la perte ou la duplication de gènes, voire la duplication complète de génomes. Dans ce contexte, la base de données HOGENOM permet d'accéder à un ensemble de familles de gènes homologues provenant d'organismes complètement séquencés. Ces familles sont associées à des alignements multiples ainsi que des arbres phylogénétiques. Du fait de l'intégration de données phylogénétiques, HOGENOM offre la possibilité d'utiliser des motifs d'arbre pour rechercher des ensembles de gènes orthologues et/ou paralogues dans ces familles. La précédente version d'HOGENOM (6) contenait environ 300 000 familles pour 1 470 organismes (bactéries, archées et eucaryotes). Cependant, l'explosion des données de séquençage a rendu l'approche de construction précédemment utilisée – fondée sur l'utilisation de l'intégralité des génomes complets disponibles – inexploitable. Une nouvelle stratégie de calcul des familles a donc été développée. Cette nouvelle stratégie repose sur la construction d'une base de données (HOGENOM-CORE) contenant des génomes représentatifs de différents phyla et de plusieurs bases contenant tous les génomes d'un phylum particulier (HOGENOM-PHYL). Les génomes représentatifs sont choisis de manière semi automatisée avec l'objectif de maximiser la représentation taxonomique. Il sera par ailleurs possible de placer un arbre de HOGENOM-PHYL à l'intérieur d'un arbre de HOGENOM-CORE afin d'étendre son éventail taxonomique. En outre une nouvelle méthode de partitionnement de données est utilisée pour le calcul des familles.

Link PDF

PockDrug-Server: a web server for predicting pocket druggability

Hiba ABI HUSSEIN[1], Alexandre BORREL[1,2], Colette GENEIX[1], Michel PETITJEAN[1], Leslie REGAD[1] and Anne-Claude CAMPROUX[1]

[1] Molécules Thérapeutiques *in silico* (Mti), University Paris Diderot, Sorbonne Paris Cité, Inserm, UMRS-973, Paris, France
[2] University of Helsinki, Division of Pharmaceutical Chemistry, Faculty of pharmacy, Finland

Corresponding authors: hiba.abihussein@univ-paris-diderot.fr & anne-claude.camprouxs@univ-paris-diderot.fr

Therapeutical molecules bind to preferred sites of action, which are mostly pockets located inside proteins or lying at their surface. Therefore, estimation and characterization of pockets are major issues for drug discovery projects. Drug-like molecules are small molecules with particular properties such as small size or the ability to cross the digestive tract. Predicting protein pocket's ability to bind drug-like molecules with high target affinity, i.e. druggability, is a key step of the modern drug design process. Existing druggability prediction models involve only one pocket estimation method at a time, despite pocket estimation uncertainties. Here we present PockDrug-Server, which predicts pocket druggability. It is efficient on both (i) estimated pockets guided by the ligand proximity (extracted by proximity to a ligand from a holo protein structure using several distance thresholds), and (ii) estimated pockets only based on protein 3D structure. In order to provide a mean druggability, PockDrug-Server involves a consensus model based on a combination of 7 linear discriminant analysis models using 9 pocket descriptors. It provides consistent and robust druggability results using different pocket estimation methods; thus it is efficient when using apo pockets difficult to estimate. It accurately distinguishes druggable from undruggable pockets, outperforming recent druggability models for apo pocket. It can be carried out from one or a set of apo/holo proteins. PockDrug-Server is publicly available at: http://pockdrug.rpbs.univ-paris-diderot.fr.

Link PDF

Workflows for comparative protein analysis using PipeAlign in Galaxy

Arnaud KRESS[1], Luc THOMES[1], Raymond RIPP[1] and Julie THOMPSON[1]

[1] LABORATOIRE ICUBE, UMR7357 CNRS, LBGI, 11 rue Humann, faculté de médecine, 67000 Strasbourg, Cedex, France

Corresponding author: akress@unistra.fr

In the context of the BISTRO project (Strasbourg bioinformatics platform: bioinfo-bistro.fr), the cascade of programs PipeAlign was ported to the Galaxy platform, available at toolshed.galaxeast.fr.

PipeAlign is a tool for comparative or evolutionary analysis of a protein family that allows the user, starting from a protein sequence, to run a pipeline of programs ranging from homolog searches in public sequence databases to construction of high quality multiple alignments and visualization of the results. First, we integrated into Galaxy the programs present in the original version of PipeAlign, namely BlastP, Ballast (BlastP post- processing), DBClustal (multiple alignment of detected sequences), NorMD (objective function), RASCAL (correction of local alignment errors), LEON (analysis of homologous regions), and Cluspack (classification of sequences). Then, we extended this workflow with additional programs, including MAFFT, MACSIMS (structural/functional annotation), and JalView to visualize associated annotations.

This work allows us to propose a modular assembly of the programs and allows the user to run all or part of the workflow either automatically, or step by step. The user can also customize the workflow by changing the execution parameters. PipeAlign thus provides biologists with an accessible and friendly environment for interactive comparative analysis, reuse of workflows, management / sharing of results, etc. The software suite has been adapted in Galaxy packages and integrated in a toolshed (depository) accessible to all, allowing easily installation on all instances and publishing of analysis protocols.

Link PDF

MyGeneFriends: a social network linking genes, diseases and researchers

Alexis ALLOT[1], Kirsley CHENNEN[1], Laetitia POIDEVIN[1], Alexia RHOMER[1], Arnaud KRESS[1], Raymond RIPP[1], Julie THOMPSON[1], Olivier POCH[1] and Odile LECOMPTE[1]

[1] LBGI, Computer Science Department, ICube, UMR 7357, University of Strasbourg, CNRS, Fédération de médecine translationnelle, 4 rue Kirschleger 67085 Strasbourg, France

Corresponding author: odile.lecompte@unistra.fr

With the constant and massive increase of biological information, efficient access to useful information and production of knowledge becomes a growing challenge for successful research processes. Here, we present MyGeneFriends, a social network that directly interconnects three types of actors: genes, diseases and humans. Networking between genes and diseases is ensured by a daily data mining process through public databases and by human interactions. Thus, MyGeneFriends transforms genes and diseases from passive entities into active actors of the research process interacting with clinicians and biologists. The goal is to optimize and speed up research processes by providing a user friendly web service allowing retrieval, annotation and interaction with information related to genes, diseases and researchers. MyGeneFriends allows users to become friends (publicly, collaboratively or privately) with humans, genes and diseases, manage these friendships, view new friendship suggestions, and follow the activity of their friends (a gene-friend befriends a new disease, a disease has its personal information updated...). The concept of "Topic" allows users to specify active research topics and organize genes, diseases, keywords and other information on which various analyses can be performed and visualized. It also allows a group of researchers to collaborate end efficiently communicate around a topic. Most importantly, these data allow MyGeneFriends to take into account user interests, and automatically adapt available structured or unstructured information such as publications or disease and gene textual descriptions by highlighting important data and suggesting new friends and pertinent publications.

Link PDF

BreedWheat genotyping and phenotyping data in GnpIS information system

Mathilde LAINE[1], Thomas LETELLIER[1], Raphaël FLORES[1], Loïc COUDERC[1], Nacer MOHELLIBI[1], Cyril POMMIER[1], Delphine STEINBACH[1], Hadi QUESNEVILLE[1], Frédéric SAPET[2], Nathalie RIVIERE[2], Etienne PAUX[3], Audrey DIDIER[3], François BALFOURIER[3], Gilles CHARMET[3], Jacques LE GOUIS[3] and Michael ALAUX[1]

[1] INRA, UR1164 URGI - Research Unit in Genomics-Info, INRA de Versailles, Route de Saint-Cyr, Versailles, 78026 France
[2] Biogemma, route Ennezat, Chappes, 63720 France
[3] INRA, UMR 1095 GDEC – Génétique Diversité Ecophysiologique des Céréales, INRA de Clermont-Ferrand, Domaine de Crouël, 5 chemin de Beaulieu, Clermont-Ferrand, 63039 France

Corresponding author: mathilde.laine@versailles.inra.fr

BreedWheat project aims to support the competitiveness of the French wheat breeding sector, answering to societal challenges for a sustainable and quality production. Moreover, the BreedWheat project will characterize yet poorly exploited genetic resources to expand the diversity of the elite germplasm. Finally, new breeding methods will be developed and evaluated for their socioeconomic impact.

In this frame, bioinformatics goals are (i) to establish and maintain a centralized repository of generated data,(ii) to support user requirements for data query and facilitate large scale statistical analyses through an integrated system.

URGI (Unité de Recherche en Génomique Info) is a genomics and bioinformatics research unit at INRA dedicated to plants and crop parasites. It develops and maintains a genomic and genetic Information System: GnpIS (Steinbach et al., Database 2013, doi:10.1093/database/bat058). In collaboration with the Biogemma biotechnology company, GnpIS was updated to manage and integrate genotyping and phenotyping data and new types of data such as GWAS and genomic selection.

BreedWheat data already available in GnpIS are BreedWheat germplasm (collection of 5113 accessions), SNP discovery (724020 SNPs from 10 sources), genotyping (Affymetrix Axiom TaBW420K array) and phenotyping data (48000 plots in 21 locations): wheat-urgi.versailles.inra.fr/Projects/BreedWheat. Association and genomic selection data will be generated and integrated in the coming years.

Link PDF

Post-105 (#59) - Florian PHILIPPE - Data integration in the agronomic domain: national and international data discovery system

Data integration in the agronomic domain: national and international data discovery system

Florian PHILIPPE[1], Aravind VENKATESAN[2], Nordine EL HASSOUNI[3], Cyril POMMIER[1], Manuel RUIZ[2,3], Pierre LARMANDE[2,4,5], Delphine STEINBACH[1] and Hadi QUESNEVILLE[1]

[1] INRA, UR1164 URGI – Research Unit in Genomics-Info, route de Saint-Cyr, 78026 Versailles, Cedex, France
[2] IBC, Université Montpellier 2, 860 rue St Priest Bâtiment 5 - CC05019, 34095 Montpellier, Cedex 5, France
[3] CIRAD, UMR AGAP – Amélioration Génétique et Adaptation des Plantes Méditerranéeennes et Tropicales, av. Agropolis – TA A-108/03, 34398 Montpellier, Cedex 5, France
[4] IRD, UMR 232 DIADE – Diversité Adaptation et Developpement des plantes, 911 Av. Agropolis, 34394 Montpellier, Cedex 5, France
[5] INRIA, ZENITH – Scientific Data Management, INFO/LIRMM – Département Informatique, 161 rue Ada, 34090 Montpellier, Cedex 5, France

Corresponding author: florian.philippe@versailles.inra.fr

Current research in Agronomy has produced a vast amount of genomic, genetic and phenomic data. To deal with the Volume, Variety and Velocity of those data, it is necessary to first refine to candidate datasets through data discovery then to integrate them through semantic web technologies. Data discovery is an approach that allows to easily search for datasets based on keywords and metadata. The plant bioinformatic node of the Institut Français de Bioinformatique (IFB) gives access to several public information systems hosting domain specific data. It is composed of five bioinformatics platforms : the South Green platform, the LIPM platform, the Roscoff platform ABiMS, the platform for Arthopods for Agroecosystems Arthropods and the URGI platform. The later one plays a key role in several national an international projects like the Whea Initiative. Those platforms integrate several plant genomic, genetic and phenomic data, which they need to expose in data discovery and integration systems. The distributed data discovery system need an ETL (Extraction, Transformation and Loading) based integration pipeline implemented on each platform. This ETL can either be developed from scratch or be based on existing technologies such as KarmaWeb, Talend or Open Refine. The pipeline is being developed at the URGI, and will be deployed on all IFB plant nodes. The data discovery system is based on SolR (implemented in the Transplant portal http://www.transplantdb.eu) which uses the Lucene search framework at its core for full-text indexing. Here, we will present the data discovery system architecture and the ETL solutions evaluation and comparison. Work funded by IFB investment for the future infrastructure project, IFB_Plant node.

Link PDF

Post-106 (#65) - Guillaume MERCERON - GnpAsso: a generic workbench for managing and exploiting genetic association studies results using high throughput genotyping and phenotyping data

GnpAsso: a generic workbench for managing and exploiting genetic association studies results using high throughput genotyping and phenotyping data

Guillaume MERCERON[1], Célia MICHOTEY[1], Raphael FLORES[1], Cyril POMMIER[1], Nacer MOHELLIBI[1], Christopher SAUVAGE[3], Mathilde CAUSSE[3], Stéphane NICOLAS[2], Alain CHARCOSSET[2], Vincent SEGURA[9], Frédéric SAPET[4], Pierre DUBREUIL[4], Jacques LEGOUIS[5], Alexis DEREEPER[6], Patrice THIS[7], Yannick DE OLIVIERA[2], Catherine DELAITRE[8], Guy-Ross ASSOUMOU-ELLA[2], Félix HOMA[7], Hadi QUESNEVILLE[1] and Delphine STEINBACH[1]

[1] LABORATOIRE, URGI, INRA, RD 10 Route de St Cyr 78000 Versailles, France
[2] LABORATOIRE UMR GQE-Le Moulon, INRA, 91190 Gif-sur-Yvette, France
[3] LABORATOIRE UMR GAFL, INRA, 67 allée des chênes 84143 MONTFAVET, France
[4] Biogemma compagny, Route d'Ennezat 63720 Chappes, France
[5] LABORATOIRE UMR GDEC, INRA, Site de Crouel 63000 Clermont-Ferrand, France
[6] LABORATOIRE UMR RPD, IRD 911 avenue Agropolis BP 64501 Montpellier France
[7] LABORATOIRE UMR AGAP, INRA Campus Supagro Montpellier, 2 place Viala 34060 Montpellier, France
[8] LABORATOIRE UMR Agroécologie, INRA, 17 rue de Sully 21065 Dijon, France
[9] LABORATOIRE, UR AGPF, INRA 2163 avenue de la pomme de pin 45075 Orléans Cedex 2, France

Corresponding author: delphine.steinbach@versailles.inra.fr

GnpIS, (D. Steinbach et al. Databases Journal 2013, http://dx.doi.org/10.1093/database/bat058) is an information system for genomic, genetic and phenomic data for plants and its bioagressors, http://urgi.versailles.inra.fr/gnpis. We will present here the work done in the frame of th ANR bioinformatics project GnpAsso (2011-2014), whose aim was to set up in collaboration with scientists experts in GWAS on plants (wheat, maize, tomato, grape, pea, poplar), a complete environment for hosting and mining into GWAS association data and in its associated highthroughput data (genotyping and phenotyping). GnpAsso is composed of 1) an information system GnpIS extended to manage GWAS data (GnpIS-Asso), accessible for query at: https://urgi.versailles.inra.fr/association, 2) a tool for diversity analysis, Sniplay http://sniplay.cirad.fr extended to run GWAS analysis tools and view results, 3) Thaliadb, a genotyping and phenotyping laboratory database extended to provide data to the 2 other components. GnpIS-Asso database contains today 2 datasets published by partners, one on tomato and one on maize. The datawarehouse is used to manage data for 4 running projects funded by ANR 'Investment for the Future' on wheat, maize, pea and rapeseed.

Link PDF

Post-107 (#76) - Vincent J. HENRY - OMIC-Onto: a resource for search and indexation of databases and analytical software for omic data

OMIC-Onto: a resource for search and indexation of databases and analytical software for omic data

Vincent J. HENRY[1,2,3], Julien GROSJEAN[2], Lina SOUALMIA[2,4], Anita BANDROWSKI[5], Bruno J. GONZALEZ[6], Stéfan DARMONI[2,4] and Arnaud DESFEUX[1]

[1] STATSARRAY, Sotteville-lès-Rouen, France
[2] CiSMeF, LITIS-TIBS EA 4108, Rouen University Hospital, Normandie Université, France
[3] Master de bioinformatique de Rouen, Normandie Université, Mont-Saint Aignan, France
[4] LIMICS, U1142, INSERM, Paris, France
[5] Center for Research in Biological Systems, University of California, San Diego, La Jolla, CA 92093
[6] Haute-Normandie-INSERM ERI-28, Rouen, France

Corresponding author: vincent.henry@omictools.com

Recent advances in ‘omic’ technologies have created unprecedented opportunities for biological research. Now, the main challenge consists in omic data analysis and interpretation. The metadatabase OMICtools was developed to index and classify analytical resources that are fragmented. During its development, the lack of controlled vocabularies and the lack of minimal information about these tools were pointed out. To fill it, we build the terminology "OMIC-termino" which is composed by more than 650 terms gathered into 3 generic concepts: “raw data analysis”, “biological interpretation” and “minimal information associated to the tools”. Manual alignment of OMIC-termino with other resources such as EDAM ontology or NCI Thesaurus represents 55% of specific terms of OMIC-termino, which justifies its development. An automatic alignment will be performed after its integration into the multiple- terminologies portal HeTOP. The use of a common vocabulary will allow semantic and syntactic interoperability between the index tools and other domain-related resources. Moreover, the objective of the concept “minimal information associated to the tools” is to contribute to the good practice of release software tools standardization effort. As a next step, OMIC-termino will be formalized into the ontology OMIC-onto to automatically verify its consistency and validating it.

Link PDF

Post-108 (#80) - Emeric SEVIN - *de novo* Assembly of large and complex genomes: a cloud-oriented hybrid solution

de novo Assembly of large and complex genomes: a cloud-oriented hybrid solution

Emeric SEVIN[1], Alexis GROPPI[1], Pierre PETERLONGO[2], Macha NIKOLSKI[1,3] and Raluca URICARU[1,3]

[1] CENTRE DE BIOINFORMATIQUE DE BORDEAUX (CBiB), Université de Bordeaux, 146 rue Léo Saignat, case 68, 33076 Bordeaux Cedex, France
[2] GENSCALE, INRIA/IRISA CENTRE, Campus de Beaulieu, 35042 Rennes Cedex, France
[3] CNRS/Laboratoire Bordelais de Recherche en Informatique (LABRI), 33405 Talence, France

Corresponding author: emeric.sevin@u-bordeaux.fr

Thanks to great progress in DNA sequencing technologies these last few years, the number of whole genome sequencing projects has increased exponentially over that timeframe. However, releasing a high-quality finished genome remains a difficult and time-consuming task, because of the complex underlying assembly stage of DNA fragments. This is especially true in the case of certain plants, whose genome may have undergone several duplication events throughout evolution, and thus display very large sizes as well as higher frequencies of repetitions.

To tackle these challenges, we developed a parallel assembly pipeline to reduce the computation time. Our pipeline combines both long PacBio reads and short Illumina reads. This hybrid approach takes advantage of the low error rate in Illumina reads, so as to insure optimal accuracy in the reconstructed sequences, while still being able to rely on the length of PacBio reads to better guide the scaffolding stage and improve sequence contiguity. This combination of data is also at the root of our assembly distribution method. Assembly is indeed usually considered hard to parallelize because of the highly connected nature of the data structure commonly used to solve this problem, *i.e.* graphs. Finally, the pipeline is presently being adapted to fit a cloud-oriented implementation, based on the Hadoop and MapReduce frameworks, in order not to restrain its utilization to sole laboratories equipped with High Performance Computing clusters.

Link PDF

Mise à jour du site web spécialisé à l'analyse des CRISPRs

Juraj MICHALIK[1], Christine DREVET[2], Christine POURCEL[2], Claire TOFFANO-NIOCHE[2], Daniel GAUTHERET[2] and Maude PUPIN[1]

1 Bonsai/bilille, CRIStAL, UMR CNRS 9189, Univ Lille, 59650 Villeneuve d'Ascq, France
2 Institute for Integrative Biology of The Cell, UMR CNRS 9198, Univ Paris-Sud, 91405 Orsay, Cedex, France

Corresponding author: juraj.michalik@univ-lille1.fr

Certains génomes de Bactéries et d'Archées présentent des éléments appelés CRISPR pour « *Clustered Regularly Interspaced Short Palindromic Repeats* ». Ces éléments sont composés des segments répétés espacés par des régions uniques. Leur fonction principale est la protection contre l'ADN viral. Sous l'impulsion de l'Institut de Biologie Intégrative de la Cellule à Orsay, la première base de données publique dédiée aux CRISPR (CRISPRdb) a été développée, associée à un logiciel spécialisé permettant leur identification et leur extraction (CRISPRFinder), tous deux accessibles via internet (http://crispr.u-psud.fr/). Le maintien et le développement du site, ainsi que la mise à jour des informations contenues dans la base de données ont été repris par la plate-forme Bioinformatique de l'Université Paris-Sud (eBio), en collaboration avec la plate-forme de services en bioinformatique de Lille, bilille.

Nous présenterons le processus de mise à jour de la base de données qui a été modernisé. Notamment, le téléchargement des séquences à partir du serveur ftp du NCBI a été remplacé par l'utilisation des E-Utilities, les services web assurant l'interface avec le système de gestion des bases de données du NCBI. Nous ferons également un bilan des prédictions de CRISPR réalisées sur l'ensemble des génomes complets disponibles.

Link PDF

The transcript isoform quantification conundrum: an overview

Thibault DAYRIS[1], Oussema SOUIAI[1], Coline BILLEREY[1], Claire TOFFANO-NIOCHE[1] and Daniel GAUTHERET[1]

1 Institute for Integrative Biology of The Cell, UMR CNRS 9198, University Paris-Sud, 91405 Orsay cedex, France

Corresponding author: daniel.gautheret@u-psud.fr

In all domains of life, genes are transcribed into multiple transcript isoforms. RNA-seq is currently the preferred technology for the identification and quantification of these isoforms. However, correctly inferring isoforms is very challenging due to the small read size in RNA-seq data and the diversity of isoforms that result from multiple events, including variations in transcription start, termination, processing and degradation. This limitation strongly impacts our ability to properly quantify isoforms based on a given RNA-seq library.

In this study, we review about 30 isoform quantification tools. We highlight their diversity in terms of methods for representing and quantifying "events". In turn this defines different ranges of application such as detecting specific event classes, detecting novel events, or performing differential expression analysis. We show that those tools are neither equivalent nor interchangeable in their use.

We then focus on software that measure differential isoform expression between pairs of conditions. With the aim of benchmarking and rating these tools through an objective value (such as an f-measure), we are developing simulated RNA-seq datasets with fully controlled proportions of alternative events. We will present these ongoing developments, as well as results of our initial comparisons on "real" RNA-seq datasets.

Link PDF

How UniProtKB/TrEMBL Tackles High Redundant Proteomes

Benoit BELY[1], Ramona BRITTO[1], Borisas BURSTEINAS[1], Maria MARTIN[1,2,3] and UNIPROT CONSORTIUM[1,2,3]

[1] EMBL - European Bioinformatics Institute (EMBL-EBI), Welcome Trust Genome Campus, CB10 1SD Hinxton, Cambridge, United Kingdom
[2] SIB Swiss Institute of Bioinformatics (SIB), Centre Medical Universitaire, 1, rue Michel Servet, Geneva, Switzerland
[3] Protein Information Resource (PIR), Georgetown University Medical Center, 3300 Whitehaven Street NW, Washington, United States of America

Corresponding authors: benoit.bely@ebi.ac.uk & help@uniprot.org

The UniProt Knowledgebase (UniProtKB) has witnessed exponential growth in the last few years with a two-fold increase in the number of entries in 2014. This follows the vastly increased submission of multiple genomes for the same or closely related organisms. This increase was accompanied by a high level of redundancy in unreviewed UniProtKB (TrEMBL), and many sequences were over-represented in the database. This was especially true for bacterial species where different strains of the same species have been sequenced and submitted (e.g. 1,692 strains of Mycobacterium tuberculosis, corresponding to 5.97 million entries). High redundancy led to an increase in the size of UniProtKB, and thus to the amount of data to be processed internally and by our users, but also to repetitive results in BLAST searches for over-represented sequences.

To reduce this redundancy, we have developed a procedure to identify highly redundant proteomes within species groups using a combination of manual and automatic methods. We applied this procedure to bacterial proteomes (which constituted 82% of UniProtKB/TrEMBL as of release 2015_03) beginning in the 2015_04 release. Sequences corresponding to redundant proteomes (47.0 million entries) were removed from UniProtKB. From release 2015_04 on, we no longer create new UniProtKB/TrEMBL records for proteomes identified as redundant. The redundant sequences removed from UniProtKB (in 2015_04) or never added to UniProtKB (from 2015_05) are still available in the UniParc sequence archive dataset. All proteomes remain searchable through the Proteomes pages. The history (i.e. previous versions) of redundant UniProtKB records remains available.

Link PDF

Post-112 (#116) - Sandrine PERRIN - An automated and modular output quality control pipeline for Illumina sequencers

An automated and modular output quality control pipeline for Illumina sequencers

Sandrine PERRIN[1], Sophie LEMOINE[1], Stéphane LE CROM[1,*] and Laurent JOURDREN[1,*]

[1] Plateforme Génomique, École normale supérieure, Institut de biologie de l'ENS, UMR8197 CNRS, U1024 INSERM, 46 rue d'Ulm, 75005 Paris, France

[*] Co-auteurs

Contacting author: aozan@biologie.ens.fr

Aozan [1] is an automated post sequencing data processing pipeline, it automatically handles run detection, data transfer, demultiplexing conversion with bcl2fastq2 [2] and quality control with FastQC [3] and FastqScreen [4] once a run is completed. Moreover it checks if critical low disk space appears.

Aozan can be deployed using the Docker technology [5], it allows an easier installation of the software without any requirements. After its installation and configuration, Aozan runs without any manpower, it will inform users about all summaries at the end of each step by email.

Aozan is very flexible; it suits small and large sequencing platforms that have one or more Illumina sequencers (HiSeq or NextSeq). In order to reduce runtime in multi-servers context, it is possible, as an example, to dedicate a or more servers at each step.

The quality control report is customizable with a single configuration file. Data from sequencing control and output files like InterOp Illumina, FastQC report or XML files are compiled in several optional sections : runs, lanes, projects and samples. The plugin architecture allows easier integration of new functionalities in the pipeline like a complementary analysis on reads.

To conclude, Aozan is autonomous and modular. Moreover, our versatile tool should assist genomic platforms in tracking their run information and data quality results required for quality labels such as ISO 9001 certification.

[1] http://transcriptome.ens.fr/aozan/
[2] http://support.illumina.com/content/dam/illumina- support/documents/documentation/software_documentation/bcl2fastq/
[3] Andrews S: FastQC 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
[4] Andrews S: FastQ Screen 2011. http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen
[5] https://registry.hub.docker.com/u/genomicpariscentre/aozan/

Link PDF

Validate and keep up-to-date the annotations dedicated to RNA-Seq

Sophie LEMOINE[1,2], Laurent JOURDREN[1,2] and Stéphane LE CROM[1,2]

[1] Ecole normale supérieure, Institut de Biologie de l'ENS, IBENS, Plateforme Génomique, Paris, 75005 France;
[2] Inserm, U1024, Paris, 75005 France; CNRS, UMR 8197, Paris, 75005 France

Corresponding author: slemoine@biologie.ens.f

The genomics platform of the École normale supérieure is an open infrastructure dedicated to functional genomics. We cover sequencing projects whatever the specie is since it is a eukaryote and since a sequence file and a consistent GFF3 [1] annotation can be provided. These files are needed to run Eoulsan [2], our analysis pipeline.

Annotation file formats are very diverse what implies to validate properly each genome-annotation couple before it is used in Eoulsan for a real analysis. This is a quite simple step but what comes before needs a to be well defined.

As we are an open platform, we retrieve data from more than one repository to satisfy our needs and we cannot change that. But, if we promote Ensembl [3] when possible, it becomes feasible to: (i) query directly Ensembl using its API to build a native gff3, (ii) control the genome and annotation versions through the API, (iii) use BioMart [4] for additional annotations.

Each time Ensembl is updated, our pipeline will: (i) download genomes, GFF3 files and BioMart annotations in an automatic way, (ii) simulates reads on these downloaded genome files, (iii) run Eoulsan to validate genomes, GFF3 annotations and additional annotations.

This validation pipeline allows us to complete consistently this crucial task for any functional genomic project and avoid its tedious side.

[1] K. Eilbeck, S.E. Lewis, C.J. Mungall, M. Yandell, L . Stein, R. Durbin, M. Ashburner. The Sequence Ontology: A tool for the unification of genome annotations, *Genome Biology*. 6:R44, 2005
[2] L. Jourdren, M. Bernard, M-A. Dillies, S. Le Crom. *Bioinformatics*. 28: 1542-1543, 2012
[3] P. Flicek *et al*, Ensembl 2014. *Nucleic Acids Research*. 42:D749-D755, 2014
[4] J.M. Guberman *et al*. BioMart Central Portal: an open database network for the biological community. *Database*. bar041, 2011.

Link PDF

Post-114 (#123) - Nicolas FRANCILLONNE - A computational architecture designed for genome annotation: oak genome sequencing project as a use case

A computational architecture designed for genome annotation: oak genome sequencing project as a use case

Nicolas FRANCILLONNE[1], Tina ALAEITABAR[1], Françoise ALFAMA[1], Loïc COUDERC[1], Claire GUERCHE[1], Thomas LETELLIER[1], Mikael LOAEC[1], Isabelle LUYTEN[1], Célia MICHOTEY[1], Jean-Marc AURY[2], Hadi QUESNEVILLE[1], Christophe PLOMION[3], Joëlle AMSELEM[1] and the Oak genome sequencing consortium

[1] INRA, UR1164 URGI - Research Unit in Genomics-Info, INRA de Versailles, Route de Saint- Cyr, Versailles, 78026 France
[2] CEA, Institut de Génomique, Genoscope, BP5706, 91057 Evry, France
[3] INRA, UMR1202 – Biodiversity, Genes and Communities - 69, route d'Arcachon, 33610 Cestas, France

Corresponding author: nicolas.francillonne@versailles.inra.fr

The ANR Genoak project aims to study the two key evolutionary processes that explain the remarkable diversity found within the oak genus. We performed an automated structural annotation (transposable elements (TEs) and genes) and functional annotation of predicted genes using robust pipelines i/ REPET for TEs ii/ Eugene for gene prediction iii/ FunAnnotPipe (in-house pipeline) mainly based on InterproScan for functional annotation. Further objectives were to: i/ integrate the whole genome with all the features annotated into a Genome Browser, ii/ provide an interface for gene prediction curation/validation, and iii/ provide an information system pointing towards accessibility and interoperability. We set up a fast and flexible genome browser WebApollo_oak allowing the edition of genes structure. This tool based on JBrowse is used to visualize all the annotations, identify and curate gene predictions. We set up a data warehouse QuercusRoburMine based on Intermine technology to integrate functional annotation. Its user friendly interface allows cross queries between different data sources. These combined tools constitute powerful resources for whole genome annotation. Next step will be to automatically update all functional information in QuercusRoburMine for gene curated through WebApollo_oak. We will present some case studies to illustrate questions raised by the users.

Link PDF

TASKMANAGER: Massively parallelizable workflow management

Artem KOURLAIEV[1], Carole DOSSAT[1], Stefan ENGELEN[1] and Jean-Marc AURY[1]

[1] CEA/Institut de Génomique/Genoscope/LIS/RDBIOSEQ, 2 rue Gaston Crémieux, 91000 Evry, Cedex France

Corresponding author: akourlai@genoscope.cns.fr

TaskManager allows to develop generic workflows in Perl, potentially executable in various environments and with different degrees of parallelization. TaskManager can execute thousands tasks in parallel, depending on the available resources, using different "batch manager" (Slurm or Lsf). The divide and conquer approach is particularly well adapted to TaskManager to reduce the computing time of NGS data analysis. TaksManager generates logs with the same structure in all execution environments that facilitate analysis and error recovery. TaskManager is developed using object-oriented programming that facilitates future evolution, adding functionality and new runtime environments. Thanks to TaskManager, the developer will always have the same way of implementing workflows on different plateforms and different level of parallelization. TaskManager has been used to export mapping workflows from Genoscope to TGCC (HPC center) in order to treat big data projects (TARA oceans).

Link PDF

HUB Bioinformatique et Biostatistique à l'Institut Pasteur

Rachel TORCHET[1], Damien MORNICO[1], Marie-Agnès DILLIES[1,2] and Christophe MALABAT[1,3]

[1] Centre de Bioinformatique, Biostatistiques et Biologie Intégrative (C3BI), Institut Pasteur, 25-28 Rue du Docteur Roux, 75015 Paris France
[2] Plate-forme Transcriptome et Epigénome, Institut Pasteur, 25-28 Rue du Docteur Roux, 75015 Paris France
[3] Unité de Génétique des Interactions Macromoléculaires, Institut Pasteur, 25-28 Rue du Docteur Roux, 75015 Paris France

Corresponding author: damien.mornico@pasteur.fr

The bioinformatics and biostatistics HUB was created at the Pasteur Institute on November 1, 2014. It is a new entity whose mission is to provide Bioinformatics and Biostatistics support to research units and platforms of the campus. This support ranges from simple consultations to development and data analysis for research projects. The HUB members also play an important role in developing and conducting numerous training courses.

The HUB currently has 7 members and is projected to grow to 40 by 2017. The request for support is very straightforward, and achieved through an online submission form. In the first five months of existence, the Hub has received over 40 project applications from 10 of the 11 scientific departments of the Institute and on various thematics such as NGS data analysis, proteomics, metagenomics and CRISPR technology.

Its position within the Center for Bioinformatics, Biostatistics and Integrative Biology (C3BI) allows it to benefit from the support of eight research units, the Center of Informatics for Biology (CIB) and the International Group for Data Analysis (IGDA). The CIB ensures the development and maintenance of Bioinformatics tools and IGDA focuses on coordinating resources in bioinformatics and statistics between the international network of 32 institutes in 25 countries around the world. These relationships allow the Hub to bridge, build and support strong collaborations between experimental and computational biology.

Link PDF

Le Cloud Académique IFB pour les Sciences du Vivant

Christophe BLANCHET[1], Mohamed BEDRI[1], Stéphane DELMOTTE[2], Marie GROSJEAN[1], Yvan LE BRAS[3], Cyril MONJEAUD[3], Olivier SALLOU[3], Bruno SPATARO[2], Jean-François GIBRAT[1] and Olivier COLLIN[3]

[1] IFB-Core, UMS3601 CNRS, Avenue de la Terrasse, Bât 21, 91190 Gif-sur-yvette, France
[2] PRABI / LBBE UMR5558 CNRS, 43 bd du 11 novembre 1918, 69622 VILLEURBANNE cedex, France
[3] Plate-forme GenOuest, IRISA-INRIA, Campus de Beaulieu, 35000 Rennes, France

Corresponding author: christophe.blanchet@france-bioinformatique.fr

Ces dernières années, beaucoup d'Infrastructures de Recherche dédiées aux sciences du vivant, mises en place à l'échelle nationale et européenne, produisent d'énormes quantités de données expérimentales hétérogènes à analyser. Leur analyse demande un grand nombre d'outils logiciels et souvent la comparaison avec des collections de données de références (en sciences du vivant, plus d'un millier de telles collections, génériques ou spécifiques, sont communément utilisées). De la même manière, les interfaces requises couvrent un spectre large : de la ligne de commande au portail web, en passant par des logiciels graphiques.

L'objectif de l'Institut Français de Bioinformatique (IFB) est de fournir aux scientifiques et ingénieurs, utilisateurs comme développeurs, les capacités de stockage et de calculs requises dans une solution flexible, simple d'utilisation et facilement adaptable. Pour simplifier l'utilisation des logiciels bioinformatiques pertinents, nous les avons intégrés dans des machines virtuelles préconfigurées (les *appliances*), prêtes à l'emploi sur le *cloud* de l'IFB. Le modèle d'organisation envisagé vise un équilibre entre les mouvements de données, qui peuvent être très couteux et dans certains cas impossible pour les données confidentielles, et le déploiement automatique des *appliances* sur des *clouds* locaux, au plus près des dépôts de données mais nécessitant une infrastructure locale compatible.

Pour répondre aux besoins les plus courants, nous avons installé une sélection de plusieurs logiciels scientifiques dans des *appliances* pour différents domaines des sciences du vivant. L'IFB référence les outils des machines virtuelles dans un catalogue, disponible en ligne, et contenant déjà vingt *appliances* pour différents sujets comme l'analyse de séquence, l'écologie et la dynamique de populations, la génomique avec le portail Galaxy et des outils comme *Stacks* pour le *RAD-seq*, la bio-imagerie, les statistiques avec la suite R, le portail R-studio et l'API web Shiny, *etc.*

Link PDF

A RAINBio over the Life Sciences Cloud

Marie GROSJEAN[1], Claudia HERIVEAU[2], Renaud JULLIEN[2], Olivier COLLIN[2], Jean-François GIBRAT[1] and Christophe BLANCHET[1]

[1] IFB-Core, UMS3601 CNRS, Avenue de la Terrasse, Bât 21, 91190 Gif-sur-yvette, France
[2] Plate-forme GenOuest, IRISA-INRIA, Campus de Beaulieu, 35000 Rennes, France

Corresponding author: marie.grosjean@france-bioinformatique.fr

The life sciences community has heterogeneous needs in terms of bioinformatics resources and services. However, the cloud technology, thanks to its flexibility, can provide a solution that fits this panel of requirements. Therefore, the French Institute of Bioinformatics (IFB) set up a cloud in order to fulfill the community needs [1]. Users can access the tools and data available on this cloud through template of virtual machines called appliances that are referenced and described within a bioinformatics marketplace. The amount of services available rose quickly to meet the increasing demand of scientists so it became necessary to set up a registry which aim is to help users to choose the most relevant appliances. The RAINBio catalogue, *Répertoire des Applications Intégrées au Nuage Bioinformatique*, combines the descriptions of services (tools and databases) and appliances available on the IFB cloud. The descriptions of services follow the ELIXIR registry model [2]. This model uses the EDAM ontology (*Ison*, 2013) in order to annotate various features of a service: for example their topic, operation, data type and format. The compatibility of RAINBio with a European registry aims to ease the annotation effort. RAINBio is a graph database built with Neo4j and queried in Cypher [3]. Every EDAM term is a database node, the services are linked with their describing terms and the appliances carrying them. Therefore, an appliance can be found through the tools available on it. Moreover, recursive queries are easier to build which allows to browse EDAM quickly. A program written in Python language, using the py2neo library, processes the files describing the services and appliances in JSON format in order to feed RAINBio. And a user web interface is also under development with this library. The flexibility of the graph model allows a quick evolution of RAINBio and helps maintaining its compatibility with other registries.

[1] http://www.france-bioinformatique.fr/?q=fr/core/cellule-infrastructure/le-cloud-ifb
[2] https://elixir-registry.cbs.dtu.dk
[3] http://neo4j.com

Link PDF

AFFICHES / POSTERS

DEUXIEME APPEL

SECOND CALL

Les résumés présentés dans cette session ont été reçus lors de la seconde vague d'appel à communication par affiches/posters. Ils n'ont pas fait l'objet d'un processus de relecture par le Comité de Programme, mais d'une relecture par un comité restreint.

BIOCHIMIE, BIOLOGIE STRUCTURALE & BIOINFORMATIQUE STRUCTURALE

BIOCHEMISTRY, STRUCTURAL BIOLOGY & STRUCTURAL BIOINFORMATICS

Thématique - Topic 02

Post-119 (#188) - Manel ZOGHLAMI - A study of multiple instance prediction methods for bacterial ionizing radiation

A study of multiple instance prediction methods for bacterial ionizing radiation

Manel ZOGHLAMI[1,2,3], Sabeur ARIDHI[1,2,3,4], Haitham SGHAIER[5,6], Mondher MADDOURI[3] and Engelbert MEPHU NGUIFO[1,2]

[1] LIMOS, UBP, Clermont University, BP 10125, 63173 Clermont Ferrand, France
[2] CNRS, UMR 6158, LIMOS, 63173 Aubiere, France
[3] LIPAH, University of Tunis El Manar, Faculty of sciences of Tunis, 1060, Tunis, Tunisia
[4] University of Trento, Trento, Italy
[5] National Center for Nuclear Sciences and Technology (CNSTN), Sidi Thabet Technopark, Ariana 2020, Tunisia
[6] Laboratory BVBGR, ISBST, University of Manouba, La Manouba, 2010, Tunisia

Corresponding author: manel.zoghlami@gmail.com

Ionizing radiation resistant bacteria (IRRB) are important in biotechnology. In this work, we deal with the problem of phenotype prediction of bacterial ionizing radiation resistance (IRR). We analyzed basal DNA repair proteins of most known proteome sequences of IRRB and ionizing radiation sensitive bacteria (IRSB) in order to learn a classifier that correctly predicts this bacterial phenotype. We formulated the problem as a multiple instance learning (MIL) problem for sequential data in which bacteria represent bags and proteins of each bacterium represent instances.

In order to solve this prediction problem, we proposed two approaches. The first approach uses local alignment technique to measure the similarity between proteins in order to predict bacterial IRR. The second approach focuses on discriminating bacteria using discriminative classifiers. For each set of orthologous proteins, we learn a classifier that predicts IRR given one protein sequence of the query bag. We evaluated the proposed approaches in a real dataset and we demonstrated that both of them are efficient. The proposed prediction system is available online at http://www.isima.fr/~mephu/IRR/.

Link PDF

EVOLUTION, PHYLOGÉNIE & PALÉOGÉNOMIQUE

EVOLUTION, PHYLOGENY & PALEOGENOMICS

Thématique - Topic 05

Hitchhiking DNA in *Magnaporthe oryzae*

Ludovic MALET[1,2], Cyprien GUÉRIN[1], Joelle AMSELEM[3,4], Elisabeth FOURNIER[5] and Hélène CHIAPELLO[1,6]

[1] INRA, UR MaIAGE, 78352 Jouy-en-Josas, France
[2] Evolutionary Bioinformatics, University of Münster, Germany
[3] INRA, URGI, 78026 Versailles, France
[4] INRA, UMR BIOGER, 78850 Thiverval-Grignon, France
[5] INRA, UMR BGPI, TA 54K, 34398 Montpellier, France
[6] INRA, UR MIA-T, 31326 Castanet-Tolosan, France

Corresponding authors: helene.chiapello@toulouse.inra.fr & ludovic.mallet@uni-muenster.de

Magnaporthe oryzae is a successful pathogen of crop plants and a threat for food production world- wide. This species gathers pathogens of different Poaceae including rice and wheat and causes the main fungal disease of rice worldwide. The Evolutionary Genomics of *Magnaporthe oryzae* (GEMO) project is an attempt to identify the genomic determinants and evolutionary events involved in pathogenicity, host specificity and adaptation. We have analyzed and compared a dataset of ten closely related genomes of the *Magnaporthe oryzae*/grisea species complex selected for their different host range. We put emphasis on the horizontally acquired genetic material that we predicted with a parametric detection method based on tetranucleotide signature. We analyzed the general content of the predicted transferred regions and proposed for some candidates the likely taxonomy of their potential donors. First results pointed out a few large transferred regions potentially acquired from distant species identified by alignment-free methods, including several plant-pathogen fungi. These candidates called for further research around their potential contribution to phenotype, in a step to open the general study of the yet unresolved evolutionary tangram of the *Magnaporthe oryzae* pathogenicity origin.

Link PDF

Post-121 (#177) - Noé PONTOIZEAU - Comparative Analysis of Marine and Freshwater Brown Algae: Insights Into the Biology and Evolution of an Extremophile Ectocarpus

Comparative Analysis of Marine and Freshwater Brown Algae: Insights Into the Biology and Evolution of an Extremophile Ectocarpus

Noé PONTOIZEAU[1], Simon M. DITTAMI[1], Irène GONZALEZ[2], André E. MINOCHE[2,3], Alexandre CORMIER[1], Pierre PERICARD[4], Erwan CORRE[4], Catherine BOYEN[1], Heinz HIMMELBAUER[2,3,5] and Thierry TONON[1]

[1] Sorbonne Universités, UPMC-CNRS, UMR 8227, Station Biologique, 29680 Roscoff, France
[2] Centre for Genomic Regulation (CRG), 08003 Barcelona, Spain
[3] Max Planck Institute for molecular Genetics, 14195 Berlin, Germany
[4] ABiMS plateform, FR 2424 CNRS UPMC, Station Biologique, 29680 Roscoff, France
[5] Universität für Bodenkultur (BOKU), Department of Biotechnology, 1190 Vienna, Austria

Corresponding author: noe.pontoizeau@sb-roscoff.fr

Ectocarpus is a cosmopolitan brown algal genus with high capacity to acclimate to a wide range of salinities. Indeed, while most *Ectocarpus* isolates are marine, at least one strain of *E. subulatus* (named hereafter FWS) was isolated from a freshwater habitat, and can be grown in seawater. To elucidate the evolution of the FWS and its phenotypic plasticity, the genomes of this strain (diploid) and of one of its close relatives (haploid) were sequenced, and comparative structural and functional analysis comprising also a previously sequenced marine strain (MS) of a distinct *Ectocarpus* species is underway. Genome assembly metrics were computed and gene prediction performed using Eugene. The three *Ectocarpus* genomes exhibit long 3'-UTRs with 643 bp average size. Transposable elements (TEs) being important drivers for genome evolution, two TEs identification approaches were tested: *ab initio* and library-based. The latter, based on RepeatMasker, detected approximately 5% of TEs vs. 22.7% for *ab initio* predictions (pipeline REPET) in the MS genome. Furthermore, functional annotation and proteome comparison were carried out. OrthoMCL was used to determine which proteins belong to core proteome and which are specific to certain *Ectocarpus* strains. This resulted in 163 clusters of orthologous proteins specific to the MS, and 459 and 313 specific clusters in the two other proteomes. Then, all predicted proteins were annotated with Interproscan and Blast2GO, and these information have been integrated into a genome browser. Subsequently, comparison of Gene Ontology analysis were performed to assess possible enrichment or loss of pathways among the three genomes. Genome organization and synteny will also be analyzed across the three genomes based on the new genetic map available for the MS.

Link PDF

Looking for long range cis-regulatory interactions in the zebrafish genome

Yves CLEMENT[1] and Hugues ROEST CROLLIUS[1]

[1] Institut de Biologie de l'ENS, UMR8197 CNRS, 46 rue d'Ulm, 75005 Paris, France

Corresponding author: yclement@biologie.ens.fr

While finding long-range regulatory regions (or enhancers) in a genome is becoming easier, finding which genes these regions regulate is more challenging. For example, considering the gene nearest an enhancer as the target gene can lead to many false positives as enhancers can regulate genes located several megabases away, often bypassing intervening genes. We previously developed a method to 1) identify putative enhancers in a given genome by looking for conserved regions in multiple alignments and 2) identify their target genes by looking for conserved physical linkage between the putative enhancer and neighboring genes in multiple species, under the hypothesis that the enhancer - target gene link, if functional, will be conserved by natural selection. We now apply this method to the genome of zebrafish, a model organism for vertebrate development whose lineage underwent a whole genome duplication about 300 million years ago. We built genome-wide multiple alignments of 7 well-sequenced and assembled fish genomes. We found about 130 000 putative enhancers that are evolutionary linked with at least one gene. Out of the 25 000 putative enhancers with only one target gene, at least 14 000 of them are located more than one gene away from their target genes, which demonstrates the value of our target prediction method. We found that putative enhancers overlap histone marks associated with development and that this overlap increases with the linkage score of putative enhancer - target gene, which suggests that these putative enhancers play an important role in fish development. Finally, comparing these results with others obtained in human can help us identify conserved gene regulatory circuits in vertebrates and thus provide a unique opportunity to study the evolution of gene regulation in vertebrates.

Link PDF

ORGANISATION & EXPRESSION DES GENOMES

ORGANIZATION & GENOME EXPRESSION

Thématique - Topic 01

Post-123 (#176) - Tristan DUBOS - Toward the de novo assembly of tandem repeats of 5S rDNA in *Arabidopsis thaliana*

Toward the *de novo* assembly of tandem repeats of 5S rDNA in *Arabidopsis thaliana*

Tristan DUBOS[1], Lauriane SIMON[1], Hervé DUBORJAL[2], Axel POULET[1], Jean-Philippe PICHON[2], Christophe TATOUT[1] and Aline V. PROBST[1]

[1] Reproduction Génétique et Développement, UMR CNRS 6743, INSERM 1103, Clermont Université, 63000 Clermont Ferrand, France
[2] BIOGEMMA, route d'Ennezat, Centre de Recherche de Chappes, CS 90126, 63720 Chappes, France

Corresponding authors: christophe.tatout@univ-bpclermont.fr & aline.probst@univ-bpclermont.fr

The 5S rRNA is one of the RNA components of the large subunit of the ribosome. *Arabidopsis thaliana* contains about 1 000 5S rDNA copies organised in clusters of tandem repeats and located in the pericentric regions. In the Columbia ecotype the 5S rDNA clusters are localized on chromosome 3, 4 and 5 and can be differentiated by a short DNA signature specific for each locus. The 5S rRNA gene consists of a transcribed sequence of 120bp and of an intergenic sequence of 380pb, which contain only few polymorphisms along a given cluster or between the three major clusters. Despite the fact that the Arabidopsis genome was sequenced 15 years ago, the sequence of the 5S rDNA clusters is only partially assembled and missing for chromosome 4. Our project proposes to assemble the three 5S rDNA clusters from a MiSeq genome sequencing generating 300bp reads. We will present the evaluation of different *de novo* assemblers tested in the course of our analysis as well as preliminary results of the assembly of the different 5S rDNA clusters. This work will complete the sequence of the Arabidopsis genome in highly repeated regions of low genetic complexity and provide the reference sequences required for further studies of the epigenetic regulation of 5S rRNA gene expression.

Link PDF

FLAGdb^{++}: a new functionality to compare genome structures in plants

Sandra PLUMAS[1], Nathalie BOUDET[1], Cécile GUICHARD[1], Philippe GREVET[1], Marie-Laure MARTIN-MAGNIETTE[1], Véronique BRUNAUD[1] and Jean Philippe TAMBY[1]

[1] Institut de Sciences des Plantes Paris-Saclay (IPS2), UMR9213/UMR1403, CNRS, INRA, Université Paris-Sud, Université d'Evry, Université Paris-Diderot, Sorbonne Paris-Cité, Saclay Plant Sciences, Bâtiment 630, 91405 0rsay, France.

Corresponding author: Jean-Philippe.Tamby@evry.inra.fr

The FLAGdb^{++} information system, *i.e.* a core database coupled to a distributed Java interface, gathers complete sequences of selected plant genomes and heterogeneous data coming from experimental approaches or computer predictions. It allows considering genes in large contexts like the chromosome, gene family or group of orthologs. The FLAGdb^{++} version 6.1 gives access to six complete genomes, *i.e.* *Arabidopsis thaliana*, rice, poplar, grape, melon and tomato. Along with structural and functional annotations, a range of new features are available including 8,787 SNPs for the tomato genome, as well as 171,900 RNAseq contigs used to improve gene discovery in melon. FLAGdb^{++} includes the alternative gene models predicted for 30% and 27% of the *A.thaliana* and melon loci respectively. FLAGdb^{++} provides an integrated environment to explore various structural and functional features. While hosting different genomes, FLAGdb^{++} provides comparisons tools at different scales. First, the orthology interface shows the reciprocal blast hits between proteomes, and can perform multiple alignments of proteins, a particularly powerful feature when inferring functions to genes. Second, a new functionality has been developed, extending comparisons to genome scale. Distributions of structural traits such as CDS length, 5' and 3' UTR length, number of exons per gene and percent of coding nucleotides in genes were computed for each genome. Non-parametric Mann-Whitney-Wilcoxon statistical tests were performed at a level of 0.05 to compare the distributions and results are displayed in a graphical interface using box-plots. As a convenience, structural traits distributions of a species-specific gene list can be visualized on the same figures, while performing the statistical test on-the-fly. Thereafter, visualizing the distributions of corresponding lists of orthologs in every organism is easily done within the interface. Combining orthology and genome comparison appears complementary when carrying comparative analyses and transferring knowledge from model species to crops, and can help to formulate new hypotheses for translational research in plant biology.

Link PDF

Adopted strategy for functional genomic analysis of a non-model perennial species *Hevea brasiliensis* using NGS technology

Stéphanie POINTET[1], KUSWANHADI[2], Julie LECLERCQ[1] and Pascal MONTORO[1]

[1] CIRAD, UMR AGAP, Montpellier, 34398, France
[2] Sembawa Research Centre, Indonesian Rubber Research Institute, Palembang 30001, Indonesia

Corresponding author: stephanie.pointet@cirad.fr

Hevea brasiliensis is the sole commercial source of natural rubber. The rubber production is affected by a physiological syndrome called Tapping Panel Dryness (TPD). The haploid rubber tree genome was assessed at 2Gb, it is highly heterozygous and includes a high rate repeat sequences. Hevea genome sequencing programmes have been initiated since 2009. The only sequence published resulted in the assembly of 1,2Gb. Our goal is to produce a reference transcriptome to study functional genomics of the TPD. The transcriptional and post-transcriptional regulations playing a major role, particularly in stress response, we completed the analysis of the transcriptome by a microtranscriptome. The reference transcriptome has been assembled using reads from 454-sequencing. To cover all the transcriptional levels, we have developed a comprehensive transcriptome using several tissues from plants growing under different environmental conditions. Several gene families involved in redox system and hormone biosynthetic and signalling pathways (ethylene and jasmonate) were identified from this transcriptome. Most recently, it served as a basis for RNAseq analysis. Similarly, microRNAs have been identified after developing a small RNA database. Experimental validation by PCR of miRNA target cleavages is laborious. Therefore, a degradome analysis strategy has been implemented. The reference transcriptome quality has been validated by characterization of some gene families. It allowed a better understanding of the transcriptional regulation of these families and to initiate a high-throughput RNAseq analysis. The degradome analysis will allow us to identify mRNA cleaved by small RNA. An international initiative is currently set-up to improve the rubber genome sequence in order to re-sequence genomes of commercial rubber clones at a lower cost for further development.

Link PDF

IMAGERIE & TRAITEMENT DE L'IMAGE

IMAGING & IMAGE PROCESSING

Thématique - Topic 10

Post-126 (#174) - Axel POULET - *Nucleus J*, a new ImageJ plugin dedicated to nuclear shape and chromatin organization

Nucleus J, a new ImageJ plugin dedicated to nuclear shape and chromatin organization

Axel POULET[1,2], Philippe ANDREY[3], David E. EVANS[2], Aline PROBST[1] and Christophe TATOUT[1]

[1] Reproduction Génétique et Développement, UMR CNRS 6743, INSERM 1103, Clermont Université, 63000 Clermont Ferrand France
[2] Plant Cell Biology, Oxford Brookes University, Oxford, United Kingdom,
[3] Imagerie Numérique et Modélisation, Institut Jean-Pierre Bourgin, UMR1318 INRA-AgroParisTech INRA, Versailles, France

Corresponding author: christophe.tatout@univ-bpclermont.fr

Gene regulation by nuclear architecture involves a specific organization of chromatin within the nucleus; however, the mechanisms underneath are still largely unknown. Recent studies in yeast and humans indicate that positioning at the nuclear periphery contributes to the assembly and/or maintenance of repressed chromatin domains known as heterochromatin. We address this question in the model species *Arabidopsis thaliana*, in which heterochromatin is grouped in dense structures called chromocentres easily detected at the cytological level by staining with DNA dyes. In order to better characterize heterochromatin organization and to investigate its possible interaction with the nuclear envelope, we developed a set of bio-imaging tools to analyse images from structured light illumination microscopy, which retain all 3D spatial information. We will describe *NucleusJ* (Poulet *et al.* 2015) a new imaging analysis tools, which have been released as ImageJ plugins (download at [2]) *NucleusJ* calculates nuclear shape and chromatin/heterochromatin parameters through two distinct 3D segmentation processes. Results obtained using this pipeline for a mutant known to modify nuclear shape and heterochromatin in *Arabidopsis thaliana* will be presented. We believe that our imaging pipeline will be of general interest as it can be extrapolated to other model systems or in the future applied to additional methods such as Fluorescent in situ hybridization (FISH).

[1] A. Poulet, *et al.* NucleusJ: an ImageJ plugin for quantifying 3D images of interphase nuclei. *Bioinformatics* 31:1144–1146, 2015.
[2] http://imagejdocu.tudor.lu/doku.php?id=plugin:stacks:nuclear_analysis_plugin:start

Link PDF

RESEAUX, REGULATION & MODELISATION

NETWORK, REGULATION & MODELING

Thématique - Topic 06

Post-127 (#168) - Marc CHAKIACHVILI - Software for computer-assisted artificial biological networks modeling for synthetic biology: biosensors design

Software for computer-assisted artificial biological networks modeling for synthetic biology: biosensors design

Marc CHAKIACHVILI[1,2], Franck MOLINA[1] and Patrick AMAR[1,3]

[1] SYS2DIAG, FRE3690 CNRS/ALCEDIAG, Cap Delta, 1682 Rue de la Valsière, 34184 Montpellier, France

[2] Master STIC & Santé, Spécialité BCD Faculté des Sciences de Montpellier 2 Place Eugène Bataillon, 34090 Montpellier, France

[3] Équipe LRI Bioinfo / Équipe Projet INRIA AMIB LRI, Bâtiment 650 Université Paris-Sud 91405 ORSAY Cedex, France

Corresponding author: marc.chakiachvili@gmail.com

The new knowledge on life and its mechanisms, current NGS sequencing development and recent progress in bioinformatics, open now huge perspectives on the evolution of our relationship to life and understanding of its operation. More and more biologist and researcher plan to user such knowledge, to modify or create new systems from life based component for human applications, in areas such as energy, ecology and medicine. These new basic research fields are referred to as synthetic biology. SYS2DIAG, research laboratory operated by CNRS and ALCEDIAG, led by Franck Molina, is particularly interested in synthetic biochemical systems and their potential application as tools for early diagnosis in human health. The laboratory has developed a specific database called "CompubioTicDB", which references biological basic protein components, partnering in devices (sensors, switches, timers ...) to allow the modeling of complete synthetic systems. These devices are abstractly described in the database but can be "implemented" with real proteins and metabolites issued from the base. In parallel, the laboratory has developed a computer- aided modelling tool, BioNetCad, to set up the three steps of Engineering: design / simulation / experimental validation. It comes as a CellDesigner© plugin and allows access to CompubioTicDB, the selection of components in order to design biosensors. This poster presents the status and changes to be made to the system to allow more intuitive decision in BioNetCad tool and improved completeness of the associated database.

Link PDF

Post-128 (#173) - Vasundra TOURÉ - STON translator: applying graph database to standard knowledge representation in Systems Biology

STON translator: applying graph database to standard knowledge representation in Systems Biology

Vasundra TOURÉ[1,2], Alexander MAZEIN[1], Irina ROZNOVAT[1], Mansoor SAQI[1], Johann PELLET[1] and Charles AUFFRAY[1]

[1] European Institute for Systems Biology and Medicine (EISBM), CNRS-ENS-UCBL, Université de Lyon - 50 Avenue Tony Garnier, 69007 Lyon, France
[2] Université Paris-Sud, UFR Sciences Bât. 301, 91405 Orsay cedex, France

Corresponding authors: vasundra.toure@gmail.com & amazein@eisbm.org

Graph databases enable to work with extensive and complex heterogeneous information with a better response time than in relational databases. In Systems Biology, graphs are a natural way of representing biological networks. Thus applying graph database technologies to biological network data management can offer new advanced opportunities in the field. Here we present STON (**S**BGN **TO** **N**eo4j), a Java-based framework developed for translation between SBGN (Systems Biology Graphical Notation) and Neo4j graph database environment. It imports and translates SBGN maps into a Neo4j graph-oriented format. Possible applications include: 1) exploiting the power of querying languages, 2) combining multiple pathways for developing even more complex networks, and 3) combining different layers of granularity for addressing difficulties related to incomplete knowledge representation.

Link PDF

Post-129 (#183) - Swann FLOC'HLAY - Logical modelling of the regulatory network governing dorsal-ventral axis specification in the sea urchin *P. lividus*

Logical modelling of the regulatory network governing dorsal-ventral axis specification in the sea urchin *P. lividus*

Swann FLOC'HLAY[1], Mathurin DOREL[1], Morgane THOMAS-CHOLLIER[1], Thierry LEPAGE[2] and Denis THIEFFRY[1]

[1] IBENS, UMR ENS - CNRS 8197 - INSERM 1024, Paris, France.
[2] Institut Biologie Valrose (iBV), UMR UNS - CNRS 7277, Nice, France.

Corresponding author: thieffry@ens.fr

Located at the basis of the deuterostome branch, echinoderms occupy a unique position to study the regulatory networks governing embryo morphogenesis. The dorsal-ventral (D-V) axis specification in the sea urchin *Paracentrotus lividus* is controlled by various transcriptional factors and signalling molecules, including two members of the TGF-β family: Nodal and BMP2/4. However, the signalling network downstream of these key morphogens is not yet fully understood. To identify Nodal and BMP2/4 target genes, we have performed a systematic functional analysis using RNA sequencing and in situ hybridization screens. The analysis of these data enabled us to delineate novel interactions.

To gain further insights into this developmental process, we are currently developing a predictive dynamical model of the corresponding signalling/regulatory network. More specifically, using a logical modelling framework, we aim to account for the specification of three main ectodermal regions along the D-V axis (ventral, ciliary and dorsal ectoderm) in terms of specific marker gene expression patterns. Built with the software *GINsim* (http://ginsim.org), our current logical model encompasses 21 components and 29 regulatory interactions. In our model analysis, we first focused on the computation of stable states and on their reachability in single representative cells, depending on signalling inputs. Next, taking advantage of the software *EpiLog* (http://ginsim.org/epilog), we have simulated grids of cells connected through signalling molecules. These simulations correctly reproduce the wild-type pattern, as well as various embryo mutant phenotypes, including double Nodal mRNA injections.

Link PDF

Topological and semantic Web based method for analyzing TGF-*β* signaling pathways

Jean COQUET[1,2], Geoffroy ANDRIEUX[1,2], Jacques NICOLAS[1], Olivier DAMERON[1] and Nathalie THERET[2]

[1] Université de Rennes 1 - IRISA/INRIA, UMR6074, 263 avenue du Général Leclerc, 35042 Rennes Cedex, France
[2] Université de Rennes 1 - IRSET EA 4427 SeRAIC, IFR140, 2 avenue Pr Léon Bernard, Rennes Cedex, France

Corresponding author: jean.coquet@inria.fr

Targeting the deleterious effects of Transforming Growth Factor TGF-*β* without affecting its physiological role is the common goal of therapeutic strategies aiming at curing fibrosis, the final outcome of all chronic liver disease. The pleiotropic effects of TGF-*β* are linked to the complex nature of its activation and signaling net- works which understanding requires modeling approaches. Our group recently developed a model of TGF-beta signal propagation based on guarded transitions (ref, Andrieux et al, 2014). In this initial work, we explored the combinatorial complexity of cell signaling, developing a discrete formalism based on guarded transitions. We imported the whole database Pathway Interaction Database into a single unified model of signal transduction. We detected 16,000 chains of reactions linking TGF-*β* to at least one of 159 target genes in the nucleus. The size and complexity of this model place it beyond current understanding. Its analysis requires automated tools for identifying general patterns.

Currently, we focus on designing one reasoning method based on Semantic Web technologies for the analysis of signaling pathways. Our method aims at leveraging external domain knowledge represented in biomedical ontologies and linked databases to rank these candidates. We consider a signaling pathway as a set of proteins involved in the respons of a cell to an external stimulus and influencing at least one gene. The underlying reasoning methods are based on graph topological analysis, formal concepts analysis (FCA) and semantic similarity and particularity measures. First, we determine the formal concepts, maximal bi-cliques, between proteins sets and genes. Then, to determine the biological relevance of theses gene clusters, we calculate a similarity score for each cluster based on Wang semantic similarity. Using such approaches, we identify groups of genes sharing signaling networks.

Link PDF

GENETIQUE & GENETIQUE DES POPULATIONS

GENETICS & POPULATION GENETICS

Thématique - Topic 03

Post-131 (#172) - Yvan HERENGER - dbSTAR: a reference DataBase of human STructural vARiation from sequencing data

dbSTAR: a reference DataBase of human STructural vARiation from sequencing data

Yvan HERENGER[1], Arnaud KRESS[2], Hélène DOLLFUS[1,3], Jean MULLER[1,4] and Véronique GEOFFROY[1]

[1] LABORATOIRE DE GENETIQUE MEDICALE, UMR_S INSERM U1112, IGMA, Faculté de Médecine FMTS, 11,rue Humann, 67085 Strasbourg, Cedex, France
[2] BISTRO platform, LABORATOIRE ICUBE, Faculté de Médecine, 11, rue Humann, 67085 Strasbourg, France
[3] SERVICE DE GENETIQUE MEDICAL, CARGO, Hôpitaux Universitaires de Strasbourg, NHC, 1 place de l'hôpital, 67000 Strasbourg, Cedex, France
[4] LABORATOIRE DE DIAGNOSTIC GENETIQUE, Hôpitaux Universitaires de Strasbourg, NHC, 1 place de l'hôpital, 67000 Strasbourg, Cedex, France

Corresponding author: veronique.geoffroy@inserm.fr

For the last several years, Next Generation Sequencing (NGS) has been widely used to highlight the molecular causes for inherited diseases. Among those causes Structural Variation (SV) play a key role in human diseases. Different types of SV exist including deletions, duplications, insertions, inversions and translocations. When unbalanced and resulting in a gain or loss of material, they are called Copy Number Variation (CNV).

Many tools have been developed with the aim of detecting these SV, based on NGS data from Targeted Exome Sequencing (TES) through Whole Exome Sequencing (WES) and Whole Genome Sequencing (WGS). Different algorithms have been implemented, each with their own strengths and weaknesses, but none of them are currently able to detect all types of SV. Thus, despite the obvious utility of detecting SV in human health, no global guidelines have been discussed yet.

Our project entitled dbSTAR (http://dbstar.lbgi.fr/dbstar/) aims at delivering a public reference human SV set in order to, first evaluate objectively all software available and then provide the opportunity to enhance those tools. This public benchmark is based i) on real data of human SV identified from WGS, WES and TES data and all validated by additional molecular biology techniques, ii) on artificial data. Each SV is categorized according to its type, its size, etc. The systematic evaluation of SV detection software will also aim to write guidelines in detecting SV from NGS data.

Link PDF

METHODOLOGIES POUR L'ANALYSE DES SEQUENCES ET DES DONNEES OMIQUES

METHODOLOGIES FOR SEQUENCE ANALYSIS & OMICS DATA

Thématique - Topic 11

VCF_creator: Mapping and VCF Creation features in DiscoSnp++

Chloé RIOU[1], Claire LEMAITRE[1] and Pierre PETERLONGO[1]

[1] INRIA Rennes-Bretagne Atlantique/IRISA, EPI GenScale, Campus Beaulieu, 263 Avenue Général Leclerc, 35042 Rennes, France

Corresponding author: chloe.riou@inria.fr

The software DiscoSnp++ is designed to detect genomic variants such as Single Nucleotide Polymorphism (SNPs) and insertion/deletion (INDELs) from raw read set(s) using a reference genome. This *de novo* method enables to find variations, among or between individuals, in particular for non-model organism, for which there is often no reference genome available or poor quality one. These markers, because of their number and their distribution on the genome are used in many biological areas: agronomy, health, medicine, or environment.

To facilitate downstream analyses and selection of these variants, we propose VCF_creator, a new feature of DiscoSnp++. Starting from the DiscoSnp++ predictions and a reference genome, VCF_creator performs an alignment of the predictions on the reference. It outputs the variants in the VCF format (Variant Calling Format), which is the text file format commonly used to report variants. The pipeline is as follows: the first step consists in aligning all the predictions on the reference using a mapping tool (BWA). VCF_creator analyses the information obtained as results of the mapping, in order to extract the mapping position and to distinguish unique from multiple ones. The validation algorithm is able to ascertain if both alleles of a variant have a unique mapping position on the reference genome. For each variant, the VCF output file provides the genomic position and the name of the sequence where it is aligned, the reference and the alternative allele, and the DiscoSnp++ information (coverage for each dataset, genotyping, rank ...). VCF_creator was applied on simulated data on the human chromosome 1. Results show that the majority of false positives predicted by DiscoSnp++ corresponds to nonaligned or multiply aligned predictions. Therefore, this tool not only makes downstream analyses easier but also improves the precision of DiscoSnp++ predictions when a reference genome is availiable.

Link PDF

Read Mapping on de Bruijn graph

Antoine LIMASSET[1] and Pierre PETERLONGO[1]

[1] INRIA/IRISA/GenScale, Campus de Beaulieu, 35042 Rennes Cedex, France

Corresponding author: antoine.limasset@irirsa.fr

Mapping reads on references is a central task in numerous genomic studies.

Since references are mainly extracted from assembly graphs, it is of high interest to map efficiently on such structures.

The problem of mapping sequences on a De Bruijn graph has been shown NP-complete1 and no scalable generic tool exists yet.

We motivate here the problem of mapping reads on a de Bruijn graph and we present a practical solution and its implementation called BGREAT.

BGREAT handles real world instances of billions reads with moderat resources. Results show that BGREAT is able to map million reads per CPU hour, even on a complex human de Bruin graph.

This tool has several major applications in various treatments of sequencing data as de novo assembly, correction, compression and quantification.

[1] arxiv.org/pdf/1505.04911v1.pdf

Link PDF

Post-134 (#184) - Jocelyn DE GOËR DE HERVE - Indexation et comparaison de séquences ADN à partir d'une fonction de hachage perceptuel

Indexation et comparaison de séquences ADN à partir d'une fonction de hachage perceptuel

Jocelyn DE GOËR DE HERVE[1,2,3], Hayfa AZIBI[1,2,4], Myoung-Ah KANG[1,2], Sami ZGHAL[4], Xavier BAILLY[3] and Engelbert MEPHU NGUIFO[1,2]

[1] Clermont Université, Université Blaise Pascal, LIMOS, BP 10448, 63000 CLERMONT-FD
[2] CNRS, UMR 6158, LIMOS, 63173 AUBIERE
[3] INRA, UR346 Épidémiologie Animale, 63122 ST GENÈS CHAMPANELLE
[4] University FSJEG Jendouba Tunisia

Corresponding author: jgoer@clermont.inra.fr

La recherche de similarité entre des séquences ADN est une problématique commune à toute étude en génomique. Elle intervient notamment durant les phases d'assemblage, pour détecter des mutations, ou pour déterminer la diversité biologique d'un échantillon lors d'études en métagénomique. Afin de répondre à cette problématique, de nombreux algorithmes ont été développés. Nous pouvons citer comme références, des algorithmes d'alignement global (Needleman-Wusch ou Smith-Waterman). Par la suite, sont apparus des algorithmes d'alignement local (BLAST). La littérature décrit aussi de nombreux algorithmes utilisant des tables de hachage (SSAHA). Enfin, des méthodes basées sur des arbres des suffixes permettent d'indexer les k-mots d'une séquence. Étant donnée la production de plus en plus massive des données séquencées, un nombre non négligeable de ces algorithmes pourraient difficilement supporter le changement d'échelle qui s'opère.

Nous proposons une nouvelle méthode de comparaison de séquences ADN et son processus de validation. L'objectif est d'accélérer la recherche de similarité entre une séquence ADN et une base de données de séquences de référence. Cette méthode est basée sur une fonction de hachage perceptuel utilisant une Transformée en Cosinus Discrète à Coefficients Signés (TCD-CS), pour le calcul des clés de hachage à partir des séquences à comparer, et exploite les propriétés d'inter-corrélation de la TCD- CS pour déterminer leur similarité. Le calcul des clés de hachage perceptuel à partir de séquences ADN permet d'opérer une diminution importante de l'espace d'exploration des données, lors des comparaisons de similitude. Cette méthode a été validée à partir d'un jeu de données constitué à partir de génomes de références. Les résultats de l'évaluation montrent des temps d'exécution qui ouvrent des perspectives quant à son utilisation à grande échelle.

Link PDF

Post-135 (#185) - Tiffany DELHOMME - A sensitive statistical model for the detection of ctDNA variants using multi-sample deep next generation sequencing data

A sensitive statistical model for the detection of ctDNA variants using multi-sample deep next generation sequencing data

Tiffany DELHOMME[1], Graham BYRNES[2], Patrice AVOGBE[1], Noémie LEBLAY[1], Lynnette FERNANDEZ-CUESTA[1], James D. MCKAY[1] and Matthieu FOLL[1]

[1] GENETIC CANCER SUSCEPTIBILTY GROUP, SECTION OF GENETICS, International Agency for Research on Cancer (IARC-WHO), 69008 Lyon, France
[2] BIOSTATISTICS GROUP, SECTION OF GENETICS, International Agency for Research on Cancer (IARC-WHO), 69008 Lyon, France

Corresponding authors: delhommet@students.iarc.fr & follm@iarc.fr

Next generation sequencing (NGS) now allows the identification of genetic alterations in large number of cancer tumor samples. However, detecting somatic mutations observed in circulating free DNA raises challenging statistical questions and most currently available variant callers are not suited for this task. Indeed, a component of sequencing errors tend to occur in a systematic way, but also to vary among sites, with error rates potentially reaching the very low allele fractions found in circulating tumor DNA (ctDNA). Therefore, one needs to precisely quantify the level of sequencing error at each site considered. Here we introduce a new variant caller specifically designed for studies containing a large number of samples. We estimate the distribution of sequencing errors from all samples together using a recently published robust negative binomial regression that we further optimized to suit our purpose, and outliers from this regression are considered as somatic variants. We study the sensitivity and specificity of our method using simulated samples mimicking real ctDNA NGS data. We compare its performance with a published variant caller (DeepSNV/Shearwater) also dedicated to the calling of mutations from large cohorts. We show that our method is much more sensitive and is able to detect variants with allelic fractions as low as 0.01% while retaining a high specificity. We illustrate using real data examples that we can correctly identify single nucleotide substitutions as well as, importantly, small insertions and deletions, while controlling for false calls in problematic regions, such as homopolymers. We finally discuss the results of our sensitivity and power analysis that should help researchers designing their studies (number of samples, coverage).

Link PDF

Identifying and analysing microRNAs in plant genomes with miRkwood

Isabelle GUIGON[1,2], Sylvain LEGRAND[3], Jean-Frédéric BERTHELOT[1,2], Mohcen BENMOUNAH[1] and Hélène TOUZET[1,2]

[1] CRIStAL, UMR CNRS 9189 University of Lille, Cité Scientifique, 59655 Villeneuve d'Ascq cedex, France
[2] Inria Lille Nord Europe, Parc scientifique de la Haute-Borne, 40 avenue Halley, 59650 Villeneuve d'Ascq, France
[3] Evolution, Ecologie et Paléontologie, UMR CNRS 8198 University of Lille, Cité Scientifique, 59655 Villeneuve d'Ascq cedex, France

Corresponding author: helene.touzet@univ-lille1.fr

MicroRNAs (miRNAs) play a crucial role in the post-transcriptional regulation of eukaryotic gene expression, in plants and animals. Many aspects of the biogenesis and evolution of miRNAs in animals and plants differ. For example, unlike miRNAs of animals, which are mainly found in introns or exons from protein coding genes, most plant miRNAs are encoded by discrete genes. Moreover, miRNAs are released from their precursors using distinct pathways in the two kingdoms. Also, miRNA precursors are more heterogeneous in plants than in animals, varying greatly in size and structure. These differences have justified dedicated approaches for miRNA gene finding. However although several prediction tools are available for metazoan genomes, the number of tools dedicated to plants is relatively limited. Considering this gap, we have developed miRkwood, a user-friendly web server specifically designed for plant miRNAs. miRkwood is able to face the diversity of plant pre-miRNAs and allows the prediction of precursors of both conserved and non-conserved miRNAs. miRkwood can deal with both full small RNA sequencing reads and short genomic sequences (up to 100 000 nt). Moreover, it offers an intuitive and comprehensive user interface to navigate in the data, as well as many export options (GFF, CSV, FASTA, ODT) to allow the user to conduct further analyses on a local computer. It is accessible at http://bioinfo.lifl.fr/mirkwood.

Link PDF

META-OMIQUES & GENOMIQUE ENVIRONNEMENTALE

META-OMICS & ENVIRONNEMENTAL GENOMICS

Thématique - Topic 07

Post-137 (#179) - Gaëtan BENOIT - Fast kmer-based method for estimating the similarity between numerous metagenomic datasets

Fast kmer-based method for estimating the similarity between numerous metagenomic datasets

Gaëtan BENOIT[1], Pierre PETERLONGO[1], Dominique LAVENIER[1] and Claire LEMAITRE[1]

[1] Inria Rennes Bretagne Atlantique, UMR 6074 Irisa, Genscale, 263 Avenue Général Leclerc, Campus de Beaulieu, 35042 Rennes, Cedex, France

Corresponding author: gaetan.benoit@inria.fr

Comparative metagenomics aims to provide high-level information based on DNA material sequenced from different environments. The purpose is mainly to estimate proximity between two or more environmental sites at the genomic level. One way to estimate similarity is to count the number of similar DNA fragments. From a computational point of view, the problem is thus to calculate the intersections between datasets of reads. Resorting to traditional methods such as all-versus-all sequence alignment is not possible on current metagenomic projects. For instance, the Tara Oceans project involves hundreds of datasets of more than 100M reads each.

To tackle this issue, we introduce a new similarity function between two datasets, called Simka, based on their amount of shared kmers. To scale on large metagenomic projects, we use a new technique which is able to count the kmers of N datasets simultaneously. This method also offers new possibilities such as filtering low frequency kmers which potentially contain sequencing errors.

Simka was tested and compared to the state of the art on 21 Tara Oceans samples. This shows that our kmer- based similarity function is very close to the read-based ones. Regarding sample proximity, different methods identify the same clusters of datasets. The fastest method of the state of the art required a few weeks to compute all the intersections whereas Simka took only 4 hours.

Link PDF

BIOLOGIE TRANSLATIONNELLE & PHARMACOGENOMIQUE

TRANSLATIONAL BIOLOGY & PHARMACOGENOMICS

Thématique - Topic 08

Post-139 (#175) - Romain GUIDOUX - eMouveRecherche: a scientifically valid application to promote physical activity and well-being

eMouveRecherche: a scientifically valid application to promote physical activity and well-being

Romain GUIDOUX[1], Yves BOIRIE[1,2], Magalie DEBOUIT-MIOLANE[1,2], Nicolas FARIGON[1,2], Martine DUCLOS[1,2], Gérard FLEURY[3], Philippe LACOMME[4], Nicolas LAMAUDIERE[2], Libo REN[4], Pierre-Henri MANENQ[5], Ludivine PARIS[1] and Sylvie ROUSSET[1]

[1] Unité de Nutrition Humaine, UMR 1019 INRA, Theix, 63009 Clermont-Ferrand Cedex, France
[2] Service Médecine du Sport et des Explorations fonctionnelles, CHU, 63003 Clermont-Ferrand, France
[3] Laboratoire de Mathématiques, UMR 6620 CNRS, Campus des Cézeaux, 63177 Aubière, Cedex, France
[4] Laboratoire d'Informatique, UMR 6158 CNRS, Campus des Cézeaux, 63177 Aubière, Cedex, France
[5] Almerys, 46 rue du Ressort, 63967 Clermont-Ferrand

Corresponding author: Sylvie.Rousset@clermont.inra.fr

Physical activity is an essential part of life style to take into account in human health. Thus the increase in sedentary behavior is associated with a deleterious body composition and growing chronic diseases. Nowadays clinical research teams *lack* of valid and cheap devices to track usually volunteer or patient progress in *free-living conditions*. On the other hand *interactivity*, i.e. behavioral feedback related to a goal, should be helpful to change a behavior from a distance and in real time. *Smartphones* are the ideal vector for collecting and receiving information. They sense the movement intensity with the accelerometers. This study aims to develop and test algorithms to estimate physical activity intensity from *accelerometry* data collected by smartphones and to deduce *energy expenditure*. Four activity intensities were recognized: immobile, light-, moderate- and vigorous-intensity in two independent groups of volunteers studied in controlled (n=25) and free-living conditions (n=60). The adult male and female volunteers were either *normal weight* or *obese*. They wore reference devices (Armband and/or FitmatePro) and had a smartphone in a trouser front pocket. The validated algorithms were implemented on *ActivCollector* server (https://activcollector.clermont.inra.fr). *eMouveRecherche* application was developed to send accelerometry data and to receive feedbacks in terms of activity duration and energy expenditure. The algorithms differed widely between normal-weight and obese people. Thus the thresholds which discriminated the four activity categories fitted the body mass of the user. The algorithm errors in absolute value compared to the references were 5.6 ± 4.5 % and 8.5 ± 6.6 % for normal-weight and obese subjects. In free-living conditions, our application showed that obese people stayed *motionless 17%* more time than normal-weight people. *eMouveRecherche* is available for research projects on Google Play : http://goo.gl/SHQ0D7.

Link PDF

Post-140 (#182) - Aurélie AUGUSTE - Small cell carcinoma of the ovary, hypercalcemic type (SCCOHT) beyond SMARCA4 mutations: a comprehensive genomic analysis

Small cell carcinoma of the ovary, hypercalcemic type (SCCOHT) beyond SMARCA4 mutations: a comprehensive genomic analysis

Aurélie AUGUSTE[1], Romy CHEN-MING-TAO[2], Marc DELOGER[2], Joanna CYRTA[3], Audrey LE FORMAL[1], Catherine RICHON[4], Nathalie DROIN[4], Beatrice BRUNN[3], Olivier CARON[5], Mojgan DEVOUASSOUX[6], Catherine LHOMME[3], Patricia PAUTIER[1], Catherine GENESTIE[7] and Alexandra LEARY[1,3]

[1] INSERM U981, Gustave Roussy, 94805 Villejuif, France
[2] Bioinformatics core facility UMS CNRS 3655/INSERM 23 AMMICA, Gustave Roussy cancer center, 94805 Villejuif, France
[3] Gynecological Cancer Unit, Departement of Medecine, Gustave Roussy, 94805 Villejuif, France
[4] Genomic Plateform, Gustave Roussy Cancer Institute, 94805 Villejuif, France
[5] Departement of Genetics, Gustave Roussy, 94805 Villejuif, France
[6] Pathology Departement, Hôpital de la Croix Rousse, Lyon, France
[7] Pathology Departement, Gustave Roussy, 94805 Villejuif, France

Corresponding author: romy.chen-min-tao@gustaveroussy.fr

Small cell ovarian carcinoma of the hypercalcemic type (SCCOHT) is an extremely rare and aggressive tumor that affects mainly young women (median age = 24 years). Prognosis is poor as most patients die within 2 years of diagnosis. Until recently, the literature describing the genomic profile of SCCOHT was scarce with common SMARCA4 loss of expression in 85% of patients. We performed an integrated genomic analysis using WES, RNA-seq and CGH approaches to identify other recurrent genomic alteration and potential therapeutic targets. Candidate Single Nucleotide Variants identified by WES were validated by Sanger and RNA-Seq. SCCOHT demonstrates a remarkably homogeneous genomic profile and potentially actionable alterations. More interestingly, SMARCA4 mutated SCCOHT demonstrated complete loss of both SWI/SNF catalytic subunit (SMARCA4 and SMARCA2), suggesting that SCCOHTs are characterized by a catalytically inactive SWI/SNF complex. This dual null SMARCA2/SMARCA4 SCCOHTs could be selected for clinical trials with EZH2, HDAC or MT inhibitors.

Link PDF

EPIGÉNÉTIQUE & EPIGÉNOMIQUE

EPIGENETICS & EPIGENOMICS

Thématique - Topic 04

Post-141 (#186) - Corinne GREY - PRDM9 a genome wide and site specific chromatin modifier for meiotic recombination

PRDM9 a genome wide and site specific chromatin modifier for meiotic recombination

Corinne GREY[1*], Julie CLEMENT[1*], Jérôme BUARD[1], Benjamin LEBLANC[2], Yvo GUT[3], Marta GUT[3] and Bernard DE MASSY[1]

[1] Institut de Génétique Humaine (IGH), Montpellier, France,
[2] BRIC and Centre for Epigenetics, University of Copenhagen, Copenhagen, Denmark
[3] Centre Nacional d'Anàlisi Genòmica (CNAG), Barcelona, Spain
[*] Authors contributed equally to this work

Corresponding author: bernard.de-massy@igh.cnrs.fr

During early meiosis, the programmed induction of DNA double strand breaks (DSB) initiates the process of homologous recombination. In mice and humans, the locations of these DSBs are determined by the sequence specific DNA binding domain of PRDM9 [1-4]. Here we demonstrate that PRDM9 binds *in vivo* (ChIPseq method) in mouse spermatocytes to the sites where meiotic DSB can be detected, highlighting the high correlation between PRDM9 binding and DSB activity. Meiotic DSBs sites are also known to be enriched for H3K4me3, a histone post translational modification dependent on PRDM9 potentially through its methyltransferase activity. We discover that these sites are also enriched for another histone mark, potentially mediated by PRDM9. This combination of chromatin modifications raises novel perspectives for the control of meiotic DSB formation and repair.

[1] F. Baudat, *et al*. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science*, 327:836-40, 2010.
[2] S. Myers, *et al*., Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science*, 327:876-879, 2010.
[3] E.D. Parvanov, P.M. Petkov and K. Paigen. Prdm9 controls activation of mammalian recombination hotspots. Science, 327:835, 2010.
[4] C. Grey, *et al*. Mouse PRDM9 DNA-Binding Specificity Determines Sites of Histone H3 Lysine 4 Trimethylation for Initiation of Meiotic Recombination. PLoS Biol. 9:e1001176, 2011.

Link PDF

SERVICES, RESSOURCES & INFRASTRUCTURES POUR LA BIOINFORMATIQUE

SERVICES, RESOURCES & INFRASTRUCTURE FOR BIOINFORMATICS

Thématique - Topic 09

Post-142 (#169) - Loraine BRILLET-GUÉGUEN - A French Galaxy Tool Shed to federate the national infrastructures and offering quality assessed tools

A French Galaxy Tool Shed to federate the national infrastructures and offering quality assessed tools

Loraine BRILLET-GUÉGUEN[1], Christophe CARON[1], Valentin LOUX[2]
and the French Galaxy Working Group[3]

[1] ABIMS, FR2424 CNRS-UPMC, Station Biologique, Place Georges Teissier, 29680, Roscoff, France
[2] UR1404 Mathématiques et Informatique Appliquées du Génome à l'Environnement, INRA, 78352 Jouy-en-Josas, France
[3] Institut Français de Bioinformatique [ANR-11-INBS-0013], France Génomique [ANR-10-INBS-0009] and MetaboHUB [ANR-11-INBS-0010]

Corresponding author: gtgalaxy@groupes.france-bioinformatique.fr

The *Galaxy* environment, notably dedicated to bio-analyses, is finding a growing success within bioinformatics and biology communities. The "*Institut Français de Bioinformatique*" (IFB) commissioned in 2013 a Working Group around the Galaxy platform. This group gathers several national platforms, and manages animation actions (Galaxy Day, thematic schools, etc.) and actions to structure (training, good practices guides, etc.) users and developers communities.

Besides, as part of the bioinformatics work packages funded by the "*France-Génomique*" project, the community has developed or evaluated many tools and set up analysis workflows. Exploitation and diffusion of these pipelines dedicated to people unfamiliar with the command line instructions now lies on using a common platform (Galaxy) and on creating a common repository (Tool Shed). From this perspective and in the Working Group dynamic, the IFB offers a *reference repository* to centralize and promote the bio-analyses tools of the French community. The scope of this repository, initially dedicated to "France-Génomique" NGS pipelines, is now extending to other national infrastructures (MetaboHUB, etc.) and to training actions (e.g. "Ecole NGS AVIESAN").

The *IFB Tool Shed* is part of a strategy to federate the community around *good practices* for integrating tools into Galaxy and *training* of engineers from concerned platforms. A special effort is made on the *quality* of tools and workflows integration, with functional tests and validation procedures.

Link PDF

Multi-omics data integration platform in public-private partnership

Nicolas SAPAY[1], Audrey CAUCHARD[1], Jacques GARNIER[1], Tanguy LE GARROUDI[1] and Pierre VEYRE[1]

[1] Institut de Recherche Technologique BIOASTER, 321 avenue Jean Jaurès, 69007 Lyon

Corresponding author: Nicolas.sapay@bioaster.org

BIOASTER is a novel French Technology Research Institute in the fields of infectious diseases and microbiology. The Institute aims at fostering the transition between the proof of concept and the mature technology in the fields of diagnostics, vaccines, antimicrobials, or microbiota-based technologies. Currently, BIOASTER is involved in more than 16 projects involving academics, small and medium enterprises, and pharmaceuticals companies.

Since its creation in 2012, BIOASTER develops the capacity to grant the access to biological collections of infectious diseases, and to analyze using several omics technologies: next generation sequencing, mass spectrometry, or mass cytometry. Hence, for each biological sample, BIOASTER can generate a large set of preclinical, clinical and molecular data. This capacity requires the conception of a scientific information system able to collect, to track, to analyze and to share the raw data and their associated information at the scale of the Institute and the BIOASTER's partners.

BIOASTER and the CNRS/IN2P3 computing center (CC-IN2P3) have initiated in 2013 a partnership to deploy the capacity to host the data storage and computing resources for the BIOASTER's projects. BIOASTER has thus the access to a robust infrastructure to host its computing capacity. The Institute now focuses its efforts on the deployment or the developments of the different services that link the data production to the result visualization: sample tracking, browsing of clinical studies, etc. Those components constitute the foundation of a knowledge management system, where large data set will be explored by the bioinformatics workflows.

Link PDF

Liste des Conférences Invitées
List of Keynote lectures

Table des matières

Liste des articles originaux (Proceedings)

Table des matières

Liste des résumés étendus (Highlight)

Table des matières

Liste des Démos

Table des matières

Liste des résumés courts (Affiches/Posters) – *Premier appel*

Table des matières

Liste des résumés courts (Affiches/Posters) – *Deuxième appel*

Table des matières

Index des contributeurs

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z

Liste des participants JOBIM 2015 – Clermont-Ferrand

Liste disponible au 9 Juillet 2015

AGRET	Clément	clement.agret@gmail.com	UNIVERSITÉ DE MONTPELLIER	France
AIT BRAHAM	Lydia	lydia.ait-braham@moulon.inra.fr	INRA	France
ALAEITABAR	Tina	talaeitabar@versailles.inra.fr	INRA	France
ALATERRE	Elina	elina.alaterre@horiba.com	HORIBA ABX SAS	France
ALAUX	Michael	michael.alaux@versailles.inra.fr	INRA	France
ALBORZI	Seyed Ziaeddin	seyed-ziaeddin.alborzi@inria.fr	INRIA	France
ALLARIO	Thierry	thierry.allario@univ-bpclermont.fr	UNIVERSITÉ D'AUVERGNE	France
ALLOT	Alexis	allot@unistra.fr	ICUBE	France
AMATO	Pierre	pierre.amato@univ-bpclermont.fr	CNRS - FÉDÉRATION DE RECHERCHE EN ENVIRONNEMENT	France
ANDONOV	Rumen	rumen.andonov@irisa.fr	INRIA	France
ANTOINE-LORQUIN	Aymeric	aymeric.antoine-lorquin@irisa.fr	CNRS - IRISA	France
AROUMOUGOM	Jean-Michel	jean-michel.aroumougom@sb-roscoff.fr	UMPC	France
ASSOUMOU-ELLA	Guy-Ross	guy-ross.assoumou-ella@moulon.inra.fr	INRA	France
AUER	Lucas	auer@insa-toulouse.fr	INRA	France
AYACHINE	Houssam	houssambtv@yahoo.fr	UNIVERSITÉ BLAISE PASCAL	France
BAA-PUYOULET	Patrice	Patrice.Baa-Puyoulet@lyon.inra.fr	INRA	France
BADEL	Anne	anne.badel@univ-paris-diderot.fr	UNIVERSITÉ PARIS DIDEROT - INSERM	France
BAHIN	Mathieu	mathieu.bahin@ens.fr	CNRS	France
BARD	Emilie	ebard@clermont.inra.fr	INRA	France
BARRE	Aurélien	aurelien.barre@u-bordeaux.fr	UNIVERSITÉ DE BORDEAUX	France
BARTOLO	François	f.bartolo11@gmail.com	INSTITUT DE MATHÉMATIQUES DE TOULOUSE	France
BATUT	Bérénice	berenice.batut@udamail.fr	UNIVERSITÉ D'AUVERGNE	France
BEAUMEUNIER	Sacha	sacha.beaumeunier@live.fr	UNIVERSITÉ DE MONTPELLIER	France
BEDRI	Mohamed	mohamed.bedri@france-bioinformatique.fr	IFB-CORE	France
BELLENANNEE	Catherine	catherine.belleannee@irisa.fr	CNRS - IRISA	France
BELY	Benoit	bbely@ebi.ac.uk	EMBL-EBI	UK
BENOIT	Gaëtan	gaetan.benoit@inria.fr	INRIA	France
BENTO	Pascal	pascal.bento@igh.cnrs.fr	CNRS	France
BERLAND	Magali	magali.berland@jouy.inra.fr	INRA	France
BERNARD	Aurélien	aurelien.bernard@clermont.inra.fr	INRA	France
BESSIERE	Chloé	chloe.bessiere@gmail.com	UNIVERSITÉ DE MONTPELLIER	France
BESSOLTANE BENTAHAR	Nadia	nadia.bessoltane-bentahar@curie.fr	INSTITUT CURIE	France
BEUGNOT	Réjane	rejane.beugnot@udamail.fr	UNIVERSITÉ D'AUVERGNE	France
BILLEREY	Coline	coline.billerey@u-psud.fr	CNRS	France
BISSEUX	Maxime	m.bisseux@laposte.net	UNIVERSITÉ D'AUVERGNE	France
BLANCHET	Christophe	christophe.blanchet@france-bioinformatique.fr	CNRS	France
BONNAFOUS	Fanny	fanny.bonnafous@toulouse.inra.fr	INRA	France
BONNOT	Titouan	titouan.bonnot@clermont.inra.fr	INRA	France
BORNES	Stéphanie	stephanie.bornes@udamail.fr	UNIVERSITÉ D'AUVERGNE	France

BORREL	Alexandre	alexandre.borrel@univ-paris-diderot.fr	UNIVERSITÉ PARIS DIDEROT	France
BOUCHER	Delphine	delphine.boucher@udamail.fr	UNIVERSITÉ D'AUVERGNE	France
BOUDELLIOUA	Imene	imene.boudellioua@kaust.edu.sa	KING ABDULLAH UNIVERSITY OF SCIENCE AND TECHNOLOGY	Saudi Arabia
BOUDET	Nathalie	boudet@evry.inra.fr	INRA	France
BOUILHOL	Emmanuel	emmanuel.bouilhol@u-bordeaux.fr	UNIVERSITÉ DE BORDEAUX	France
BOURGEADE	Laetitia	l.bourgeade@bordeaux.unicancer.fr	INSTITUT BERGONIÉ	France
BOYER	Roxanne	roxane.boyer18@gmail.com	IUT Aurillac	France
BOYER	Delphine	delphine.boyer@clermont.inra.fr	INRA	France
BRANCOTTE	Bryan	brancotte@lri.fr	CNRS	France
BRASSET	Emilie	emilie.brasset@udamail.fr	GReD CNRS	France
BRAYET	Jocelyn	diana.keremian@curie.fr	INSERM	France
BRETON	Catherine	catherine.breton02@univ-montp2.fr	IFREMER	France
BRISARD	Pierre	pierre.brizard@angers.inra.fr	INRA	France
BRONNER	Gisèle	gisele.bronner@univ-bpclermont.fr	UNIVERSITÉ BLAISE PASCAL - CNRS	France
BRULARD	Céline	celine.brulard@inserm.fr	INSERM	France
BULTEAU	Laurent	laurent.bulteau@inria.fr	CNRS - INRIA	France
CABANETTES	Floréal	floreal.cabanettes@toulouse.inra.fr	INRA	France
CABAU	Cédric	cedric.cabau@toulouse.inra.fr	INRA	France
CALLERI	Antoine	acalleri@totalinux.fr	TotaLinuX	France
CAMBON	Mélissa	mc.1989@live.fr	IAE TOULOUSE	France
CAMPAN-FOURNIER	Amandine	amandine.fournier@univ-lyon1.fr	PRABI	France
CARON	Christophe	christophe.caron@sb-roscoff.fr	CNRS	France
CARRERE	Sébastien	sebastien.carrere@toulouse.inra.fr	INRA	France
CARVALHO	Gabriel	gabriel.carvalho@irstea.fr	IRSTEA - LISC	France
CASSAN	Elodie	elodie.cassan@cpbs.cnrs.fr	UNIVERSITE DE MONTPELLIER	France
CAUMES	Géraldine	caumes.geraldine@gmail.com	UPMC	France
CERUTTI	Franck	fcerutti@toulouse.inra.fr	INRA	France
CHAKIACHVILI	Marc	marc.chakiachvili@gmail.com	UNIVERSITÉ DE MONTPELLIER	France
CHARMET	Gilles	gilles.charmet@clermont.inra.fr	INRA	France
CHAUVOT DE BEAUCHENE	Isaure	isaure@debeauchene.fr	TECHNISCHE UNIVERSITAET MUENCHEN	Germany
CHAZALVIEL	Maxime	maxime.chazalviel@toulouse.inra.fr	INRA - TOXALIM	France
CHEN-MIN-TAO	Romy	romy.chen-min-tao@gustaveroussy.fr	GUSTAVE ROUSSY	France
CHENTLI	Imène	chentli.imene@hotmail.fr	UNIVERSITÉ DE MONTPELLIER	France
CHIAPELLO	Hélène	helene.chiapello@toulouse.inra.fr	INRA	France
CHOLET	Pierre	pierre.cholet@etu.udamail.fr	IUT Aurillac	France
CHOULET	Frédéric	frederic.choulet@clermont.inra.fr	INRA	France
CHUFFART	Florent	florent.chuffart@ens-lyon.fr	CNRS	France
CISSOKO	Jules	jules.cissoko@clermont.inra.fr	INRA	France
CLEMENT	Yves	yclement@biologie.ens.fr	INSTITUT DE BIOLOGIE DE L'ENS	France
CLEMENT	Julie	julie.clement@igh.cnrs.fr	CNRS	France
CLERC	Olivier	olivier.clerc@cea.fr	CEA / CNRS	France
COLINGE	Jacques	jacques.colinge@inserm.fr	IRCM	France
COLLANGE	Guillaume	Guillaume.collange@limagrain.com	LIMAGRAIN	France
COMTE	Blandine	blandine.comte@clermont.inra.fr	INRA	France
CONESA	Ana	aconesa@cipf.es	CENTRO DE INVESTIGACIONES PRINCIPE FELIPE	Spain

CONRADI	Marin	marin.conrady@etu.udamail.fr	IUT Aurillac	France
COQUEL	Anne-Sophie	anne-sophie.coquel@sanofi.com	SANOFI CHIMIE	France
COQUET	Jean	jean.coquet@irisa.fr	CNRS - IRISA	France
CORMIER	Alexandre	acormier@sb-roscoff.fr	UPMC	France
CORRE	Erwan	erwan.corre@sb-roscoff.fr	CNRS	France
CORREA	Margot	margot.correa@genopole.cnrs.fr	CNRS	France
COSQUER	Robin	robincosquerisdd@gmail.com	IUT Aurillac	France
COURT	Franck	franck.court1@udamail.fr	GRED	France
COUVIN	David	david.couvin@cirad.fr	CIRAD	France
COX	Anthony	ACox@illumina.com	ILLUMINA CAMBRIDGE LTD	UK
DA COSTA	Quentin	quentin.da-costa@inserm.fr	CENTRE DE RECHERCHE EN CANCÉROLOGIE DE MARSEILLE	France
DA ROCHA	Martine	martine.darocha@sophia.inra.fr	INRA	France
DACCORD	Nicolas	nicolas.daccord@free.fr	UNIVERSITÉ EVRY	France
DASSIEU	Cécile	cecile.dassieu@orange.fr	IUT Aurillac	France
DAUSSET	Caroline	c.dausset@probionov.com	PROBIONOV	France
DAYRIS	Thibault	thibault.dayris@igmors.u-psud.fr	CNRS	France
DE GOER DE HERVE	Jocelyn	jgoer@clermont.inra.fr	INRA / LIMOS	France
DE OLIVEIRA	Yannick	ydeoliveira@moulon.inra.fr	INRA	France
DE OLIVEIRA	Romain	romain_de_oliveira@yahoo.fr	UNIVERSITÉ BLAISE PASCAL	France
DEBROAS	Didier	didier.debroas@univ-bpclermont.fr	UNIVERSITÉ BLAISE PASCAL - CNRS	France
DELHOMME	Tiffany	tiffany.delhomme@ens-lyon.fr	IARC/WHO	France
DENONFOUX	Jeremie	jeremie.denonfoux@genoscreen.fr	GENOSCREEN	France
DESCORPS-DECLERE	Stéphane	stephane.descorps-declere@pasteur.fr	INSTITUT PASTEUR	France
DESPLAT	Nelly	n.desplat@bordeaux.unicancer.fr	INSTITUT BERGONIÉ	France
DILLIES	Marie-Agnès	marie-agnes.dillies@pasteur.fr	INSTITUT PASTEUR	France
DJEBALI	Sarah	sarah.djebali@crg.eu	CENTRE FOR GENOMIC REGULATION	Spain
DO SOUTO	Laura	dosouto.laura@gmail.com	UNIVERSITE DE ROUEN / GENOSCOPE	France
DRU	Philippe	dru@obs-vlfr.fr	CNRS	France
DUARTE	Jorge	jorge.duarte@biogemma.com	BIOGEMMA	France
DUBOS	Tristan	tristan.duos33@gmail.com	CNRS / INSERM	France
DUFAYARD	Jean-François	jean-francois.dufayard@cirad.fr	CIRAD	France
DUFRESNE	Yoann	yoann.dufresne@etudiant.univ-lille1.fr	UNIVERSITÉ LILLE 1 / INRIA LILLE	France
DUHNEN	Alexandra	alexandra.duhnen@toulouse.inra.fr	INRA	France
DUPOUY	Marion	dup.marion@gmail.com	UNIVERSITÉ DE ROUEN / GÉNOSCOPE	France
DUPRAT	Elodie	elodie.duprat@upmc.fr	UPMC	France
DUPUIS	Pierre	pierre.dupuis@univ-lyon1.fr	PRABI	France
DUROUX	Patrice	Patrice.Duroux@igh.cnrs.fr	CNRS	France
DUVERGEY	Hugo	hugo.duvergey@igh.cnrs.fr	CNRS	France
DUVERNOIS-BERTHET	Evelyne	evelyne.duvernois@mnhn.fr	MUSEUM NATIONAL D'HISTOIRE NATURELLE	France
DWIVEDI	Ankit	aankitddwivedi@gmail.com	CNRS	France
DYNOMANT	Emeric	emeric.dynomant@gmail.com	INRA	France
EL HASSOUNI	Nordine	nordine.el_hassouni@cirad.fr	INRA	France
EL KAOUTARI	Abdessamad	abdessamad.el-kaoutari@inserm.fr	INSERM	France
ENAULT	François	Francois.ENAULT@univ-	CNRS	France

		bpclermont.fr		
ESCUDIÉ	Frédéric	frederic.escudie@toulouse.inra.fr	INRA	France
FALENTIN	Cyril	cyril.falentin@rennes.inra.fr	INRA	France
FALENTIN	Hélène	helene.falentin@rennes.inra.fr	INRA	France
FALLAH	Manel	manel.fallah@clermont.inra.fr	INRA	France
FARCY	Cédric	cedric.farcy@cirad.fr	CIRAD	France
FARHAT	Sarah	sfarhat@genoscope.cns.fr	CEA/GENOSCOPE	France
FAUBLADIER	Michael	michael.faubladier@limagrain.com	LIMAGRAIN	France
FAURE	Adrien	afaure@yamaguchi-u.ac.jp	Université de Yamaguchi	Japon
FAVIER	Virginie	vfavier@totalinux.fr	TotaLinuX	France
FELTEN	Arnaud	arnaud.felten@anses.fr	LABORATOIRES DE MAISONS-ALFORT	France
FEY	Lonathan	jonathan.fey@univ-lyon1.fr	PRABI	France
FINN	Rob	rdf@ebi.ac.uk	EMBL-EBI	UK
FISTON-LAVIER	Anna-Sophie	asfiston@univ-montp2.fr	UNIVERSITE DE MONTPELLIER	France
FLOC'HLAY	Swann	flochlay@biologie.ens.fr	INSTITUT DE BIOLOGIE DE L'ÉCOLE NORMALE SUPÉRIEURE	France
FRANCHINARD	Marie-Laure	marie-laure.franchinard@jouy.inra.fr	INRA	France
FRANCILLONNE	Nicolas	nicolas.francillonne@versailles.inra.fr	INRA	France
FUMEY	Julien	julien.fumey@egce.cnrs-gif.fr	CNRS / UNIVERSITÉ PARIS SUD	France
GARAVILLON-TOURNAYRE	Marie	Marie.GARAVILLON@univ-bpclermont.fr	INRA	France
GASC	Cyrielle	cyrielle.gasc@sfr.fr	UNIVERSITÉ D'AUVERGNE	France
GAUTREAU	Guillaume	guillaume.gautreau@free.fr	Genoscope	France
GENCE	Guillaume	guillaume.gence@gmail.com	UNIVERSITÉ DE MONTPELLIER	France
GENEIX	Colette	colette.geneix@univ-paris-diderot.fr	UNIVERSITÉ PARIS DIDEROT	France
GIACOMONI	Franck	franck.giacomoni@clermont.inra.fr	INRA	France
GIBRAT	Jean-François	jean-francois.gibrat@france-bioinformatique.fr	CNRS	France
GOLLAIN	Matthis	matthis.gollain@orange.fr	IUT Aurillac	France
GOUE	Nadia	Nadia.Goue1@etudiant.univ-bpclermont.fr	MASTER AMD - UNIVERSITÉ BLAISE PASCAL	France
GRAVOUIL	Kévin	k.gravouil@gmail.com	INDIVIDUEL	France
GRÉGOIRE	Laura	laura.gregoire@cea.fr	IRTSV/PCV CEA GRENOBLE	France
GRENIER-BOLEY	Benjamin	benjamin.grenier-boley@pasteur-lille.fr	INSTITUT PASTEUR DE LILLE	France
GRISSA	Dhouha	dhouha.grissa@clermont.inra.fr	INRA	France
GROSJEAN	Marie	marie.grosjean@france-bioinformatique.fr	CNRS	France
GRUNAU	Christoph	christoph.grunau@univ-perp.fr	UNIVERSITE DE PERPIGNAN	France
GUERIN	Cyprien	cyprien.guerin@jouy.inra.fr	INRA	France
GUIGON	Isabelle	isabelle.guigon@univ-lille1.fr	CNRS	France
GUILHOT	Nicolas	nicolas.guilhot@clermont.inra.fr	INRA	France
GUYOT	Dominique	dominique.guyot@univ-lyon1.fr	PRABI	France
HARDY	Sara	soune.hardy@gmail.com	INRA AUZEVILLE	France
HATTE	Victor	vihatte1@gmail.com	UNIVERSITÉ BLAISE PASCAL	France
HENRY	Vincent	vincent643@gmail.com	UNIVERSITE DE ROUEN / LITIS - TIBS/CISMEF	France
HERENGER	Yvan	yvanherenger@yahoo.fr	INSERM	France
HILPERT	Cécile	cecile.hilpert@etu.udamail.fr	IUT Aurillac	France
HIRCHAUD	Edouard	edouard.hirchaud@anses.fr	ANSES	France
HOCHART	Corentin	corentin.hochart.pro@gmail.com	UNIVERSITÉ BLAISE PASCAL	France

HOEDE	Claire	claire.hoede@toulouse.inra.fr	INRA	France
HOURDEL	Véronique	veronique.hourdel@pasteur.fr	INSTITUT PASTEUR	France
HUBER	Katharina	k.huber@uea.ac.uk	UNIVERSITY OF EAST ANGLIA	UK
JAZIRI	Faouzi	faouzi.jaziri@udamail.fr	UNIVERSITÉ D'AUVERGNE	France
JOB	Bastien	bastien.job@gmail.com	INSERM	France
JOLY	Dominique	Dominique.JOLY@cnrs-dir.fr	CNRS	France
JOSSELIN	Ambre-Aurore	am.josselin@laposte.net	INRA	France
JOSSO	Adrien	ajosso@genoscope.cns.fr	CEA / GENOSCOPE	France
JOST	Daniel	daniel.jost@imag.fr	CNRS - UNIVERSITÉ GRENOBLE ALPES	France
JOURDREN	Laurent	jourdren@biologie.ens.fr	ENS	France
JUST	Jérémy	jeremy.just@ens-lyon.fr	LABORATOIRE RDP	France
KAHN	Daniel	Daniel.Kahn@univ-lyon1.fr	INRA	France
KARAMI	Yasaman	yasaman.karami@upmc.fr	CNRS	France
KASPRIC	Nicolas	nicolas.kaspric@clermont.inra.fr	INRA	France
KERGROHEN	Thomas	kergrohen.thomas@gmail.com	Institut Pasteur	France
KERVELLEC	Joseph	joseph.kervellec@sb-roscoff.fr	CNRS	France
KHANNA	Varun	varun.khanna@pasteur.fr	INSTITUT PASTEUR	France
KOURLAIEV	Artem	akourlai@genoscope.cns.fr	CEA/IG/GENOSCOPE	France
KRESS	Arnaud	akress@unistra.fr	CNRS	France
LABEL	Philippe	philippe.label@clermont.inra.fr	INRA	France
LAHURE	Alan	alan.lahure@univ-lille1.fr	UNIVERSITÉ LILLE 1 / INRIA LILLE	France
LAINE	Mathilde	mathilde.laine@versailles.inra.fr	INRA	France
LANDES-DEVAUCHELLE	Claudine	claudine.devauchelle@genopole.cnrs.fr	UNIV EVRY	France
LANDI	Marion	marion.landi@clermont.inra.fr	INRA	France
LAPREVOTTE	Ivan	ivan.laprevotte@genopole.cnrs.fr	CNRS	France
LARIVIERE	Delphine	delphine.lariviere@cirad.fr	CIRAD	France
LARSONNEUR	Elise	elise.larsonneur@pasteur.fr	CNRS	France
LAUBER	Damien	damien.lauber@gmail.com	UNIVERSITÉ BLAISE PASCAL	France
LAURENT	Yannick	yannick.laurent@genoscreen.fr	GENOSCREEN	France
LAVOIE	Arthur	arthur.lavoie@igh.cnrs.fr	CNRS	France
LE	Luyen Ngoc	luyenlengoc@gmail.com	CNRS	France
LE	Quang Hien	hien.le@limagrain.com	LIMAGRAIN VEGETABLE SEEDS	France
LE BOUDIC-JAMIN	Mathilde	mathilde.le_boudic-jamin@inria.fr	INRIA	France
LEBEURRIER	Manuel	lebeurrier.manuel@hotmail.fr	UNIVERSITE DE ROUEN / CEA GENOSCOPE	France
LEBLANC-FOURNIER	Nathalie	nathalie.leblanc@univ-bpclermont.fr	INRA	France
LECHAT	Pierre	plechat@pasteur.fr	INSTITUT PASTEUR	France
LEFORT	Vincent	lefort@lirmm.fr	CNRS	France
LEGEAY	Marc	legeay@info.univ-angers.fr	LERIA	France
LEGENDRE	Matthieu	legendre@igs.cnrs-mrs.fr	CNRS	France
LEGENDRE	Rachel	rachel.legendre@i2bc.paris-saclay.fr	UNIVERSITÉ PARIS-SUD	France
LEGRAND	Sylvain	sylvain.legrand@univ-lille1.fr	UNIVERSITÉ LILLE1	France
LEMOINE	Sophie	slemoine@biologie.ens.fr	ENS	France
LEROY	Philippe	leroy@clermont.inra.fr	INRA	France
LESLUYES	Tom	t.lesluyes@bordeaux.unicancer.fr	INSTITUT BERGONIÉ	France

LESUR KUPIN	Isabelle	ilesur@pierroton.inra.fr	INRA	France
LEVEUGLE	Magalie	Magalie.leveugle@biogemma.com	BIOGEMMA	France
LIMASSET	Antoine	antoine.limasset@irisa.fr	INRIA	France
LIONNET	Clément	clement.lionnet@etu.univ-rouen.fr	UNIVERSITÉ DE ROUEN / GENOSTAR	France
LIU	Xi	xi.liu@sb-roscoff.fr	CNRS	France
LOISEAU	Camille	loiseauc48@gmail.com	UNIVERSITÉ BLAISE PASCAL	France
LOMBARD	Vincent	vincent.lombard@afmb.univ-mrs.fr	CNRS	France
LOPEZ	David	lopez.dav@icloud.com	UNIVERSITÉ BLAISE PASCAL	France
LORENZO	Jonathan	jonath.lorenzo@gmail.com	UNIVERSITÉ DE MONTPELLIER	France
LOUIS	Alexandra	alouis@biologie.ens.fr	CNRS	France
LOUX	Valentin	valentin.loux@jouy.inra.fr	INRA	France
LY	Delphine	delphine.ly@clermont.inra.fr	INRA	France
MADELAINE	Guillaume	guillaume.madelaine@ed.univ-lille1.fr	UNIVERSITÉ. LILLE 1	France
MAHUL	Antoine	antoine.mahul@lifegrid.fr	CRRI UBP	France
MALABAT	Christophe	christophe.malabat@pasteur.fr	INSTITUT PASTEUR	France
MALLET	Ludovic	ludovic.mallet@uni-muenster.de	Université Muenster	Germany
MANCHERON	Alban	alban.mancheron@lirmm.fr	UNIVERSITÉ DE MONTPELLIER	France
MARIADASSOU	Mahendra	mahendra.mariadassou@jouy.inra.fr	INRA	France
MARKOV	Gabriel	gabriel.markov@tuebingen.mpg.de	MAX-PLANCK INSTITUT	Germany
MARQUAND	Elodie	emarquand@versailles.inra.fr	INRA	France
MARTIN	Aurélie	aurelie.martin@ipsen.com	IPSEN INNOVATION	France
MARTINANT	Jean-Pierre	jean-pierre.martinant@limagrain.com	LIMAGRAIN	France
MAUPETIT	Julien	julien@tailordev.fr	TAILORDEV	France
MENG	Arnaud	christelle.cagniart@upmc.fr	UPMC	France
MERCE	Clémentine	clementine.merce@gmail.com	DKFZ	Germany
MERCIER	Jonathan	jmercier@genoscope.cns.fr	CEA / GENOSCOPE	France
MESTIVIER	Denis	denis.mestivier@u-pec.fr	INSERM	France
MICHALIK	Juraj	juraj.michalik@univ-lille1.fr	CNRS	France
MOLINIER	Mélanie	melanie.molinier@gmail.com	UNIVERSITÉ BLAISE PASCAL	France
MONAT	Cécile	cecile.monat@ird.fr	IRD	France
MONNERIE	Stephanie	ste.monnerie@gmail.com	UNIVERSITÉ BLAISE PASCAL	France
MONOT	Marc	marc.monot@pasteur.fr	INSTITUT PASTEUR	France
MONTAGNE	Thibault	thibaut@montagne.pw	BIOMERIEUX	France
MOREAU	Laurence	moreau@moulon.inra.fr	INRA	France
MORNICO	Damien	damien.mornico@pasteur.fr	INSTITUT PASTEUR	France
MORVAN	Micks	micks.morvan@gmail.com	UPMC	France
MOTTIER	Stéphanie	stephanie.mottier@univ-rennes1.fr	CNRS	France
MOUSCAZ	Yoann	yoann.mouscaz@bioaster.org	BIOASTER	France
MULLER	Hugo	mullerhugo01@gmail.com	IUT Aurillac	France
MURAT	Florent	florent.murat@clermont.inra.fr	INRA	France
NANDY	Sivasangari	nandy@cng.fr	INRA	France
NAVARRO	Julien	julien.navarro@rennes.inra.fr	INRA	France
N'GOU	Kissi	kissi.ngou@live.fr	UNIVERSITÉ BLAISE PASCAL	France
NICOLAS	Aurélie	aurelie.nicolas@rennes.inra.fr	INRA	France
OGER	Christine	christine.oger@univ-lyon1.fr	UNIVERSITÉ CLAUDE BERNARD LYON1	France
ORENGO	Christine	c.orengo@ucl.ac.uk	UNIVERSITY COLLEGE LONDON	UK
PAILLOUX	Marie	marie.pailloux@isima.fr	ISIMA / LIMOS	France

PARISOT	Nicolas	nico.parisot@gmail.com	INSA LYON	France
PASCAL	Géraldine	geraldine.pascal@toulouse.inra.fr	INRA	France
PASI	Marco	marco.pasi@ibcp.fr	CNRS	France
PAULET	Damien	damien.paulet@lirmm.fr	CNRS	France
PAULHE	Nils	nils.paulhe@clermont.inra.fr	INRA	France
PAUVERT	Charlie	charlie.pauvert@jouy.inra.fr	UNIVERSITE DE ROUEN / INRA	France
PAVOT	Balthazar	balthazar.pavot@hotmail.fr	IUT Aurillac	France
PAYSANT-LE-ROUX	Chritsine	cpaysant@evry.inra.fr	INRA	France
PELEGRIN	Romain	romain.pelegrin@etu.udamail.fr	IUT Aurillac	France
PELLETIER	Sandra	sandra.pelletier@angers.inra.fr	INRA	France
PELLETIER	Eric	Eric.Pelletier@genoscope.org	CEA / GENOSCOPE	France
PENAUD	Benjamin	benjamin.penaud@hotmail.fr	UNIVERSITÉ BLAISE PASCAL	France
PENEL	Simon	simon.penel@univ-lyon1.fr	CNRS	France
PERA	Berengère	berengere.pera@etu.udamail.fr	IUT Aurillac	France
PERALTA	Marine	marine.peralta@igh.cnrs.fr	CNRS	France
PEREIRA	Hugo	hugo.pereira@angers.inra.fr	INRA	France
PERICARD	Pierre	pierre.pericard@ed.univ-lille1.fr	UNIVERSITÉ LILLE 1 & INRIA LILLE	France
PEROUX	Yannick	yannick@tailordev.fr	0	France
PERRIERE	Guy	guy.perriere@univ-lyon1.fr	CNRS	France
PERRIERE	Marion	marion.perrier73@gmail.com	IUT Aurillac	France
PERRIN	Sandrine	sperrin@biologie.ens.fr	ENS	France
PETERA	Mélanie	melanie.petera@clermont.inra.fr	INRA	France
PETERLONGO	Pierre	pierre.peterlongo@inria.fr	INRIA	France
PETIT	Corinne	corinne.petit@univ-bpclermont.fr	LABORATOIRE MICROORGANISMES	France
PEYRET	Pierre	pierre.peyret@udamail.fr	UNIVERSITÉ D'AUVERGNE	France
PEYRETAILLADE	Eric	eric.peyretaillade@udamail.fr	UNIVERSITÉ D'AUVERGNE	France
PHILIPPE	Florian	florian.philippe@versailles.inra.fr	INRA	France
PHILIPPE	Romain	romain.philippe@clermont.inra.fr	INRA	France
PICARD	Lea	lea.picard@gmail.com	UNIVERSITÉ BLAISE PASCAL	France
PICARD	Franck	franck.picard@univ-lyon1.fr	CNRS	France
PLANEL	Rémi	remi.planel@univ-lyon1.fr	PRABI	France
POINTET	Stéphanie	stephanie.pointet@cirad.fr	CIRAD	France
POLONAIS	Valérie	valerie.polonais@udamail.fr	CNRS	France
POMIES	Lise	lise.pomies@gmail.com	UMR 547 PIAF - INRA - UBP	France
PONELLE	Flora	flora.ponelle@cjp.fr	CENTRE JEAN PERRIN	France
PONGER	Loic	ponger@mnhn.fr	MUSÉUM NATIONAL D'HISTOIRE NATURELLE	France
PONTOIZEAU	Noé	noe.pontoizeau@sb-roscoff.fr	CNRS	France
PUJADE-RENAUD	Valérie	valerie.pujade-renaud@cirad.fr	CIRAD	France
PUJOS-GUILLOT	Estelle	estelle.pujos@clermont.inra.fr	INRA	France
QUILLET	Aurélien	aurelien.quillet@etu.univ-rouen.fr	INSERM	France
RAFFEL	Raoul	raoul.raffel@igh.cnrs.fr	CNRS	France
RANCUREL	Corinne	corinne.rancurel@sophia.inra.fr	CNRS	France
RAVEL	Sébastien	sebastien.ravel@cirad.fr	CIRAD	France
REBOUL	Guillaume	reboul_guillaume@yahoo.fr	IUT Aurillac	France
REGAD	Leslie	leslie.regad@univ-paris-diderot.fr	UNIVERSITÉ PARIS DIDEROT / INSERM	France

REICHSTADT	Matthieu	matthieu.reichstadt@clermont.inra.fr	INRA	France
RIBIERE	Céline	celine.ribiere1@udamail.fr	UNIVERSITÉ D'AUVERGNE	France
RICHARD	Magali	magali.richard@ens-lyon.fr	CNRS	France
RIFA	Etienne	etienne.rifa@clermont.inra.fr	INRA	France
RIMBERT	Hélène	helene.rimbert@clermont.inra.fr	INRA	France
RIOU	Chloé	chloe.riou@inria.fr	INRIA	France
RIVIERE	Nathalie	nathalie.riviere@biogemma.com	BIOGEMMA	France
ROCHE	David	droche@genoscope.cns.fr	CEA / GENOSCOPE	France
ROCHER	Tatiana	tatiana.rocher@gmail.com	UNIVERSITÉ LILLE 1 / INRIA LILLE	France
ROHMER	Marine	marine.rohmer@mgx.cnrs.fr	CNRS	France
ROMBAUTS	Stéphane	strom@psb.vib-ugent.be	VIB	Belgium
ROUSSET	Sylvie	rousset@clermont.inra.fr	INRA	France
ROUSSIGNE	Jason	jason.roussigne@etu.udamail.fr	IUT Aurillac	France
RUDEWICZ	Justine	j.rudewicz@bordeaux.unicancer.fr	UNIVERSITÉ DE BORDEAUX	France
RUÉ	Olivier	olivier.rue@toulouse.inra.fr	INRA	France
SAAD	Chadi	chadi.saad@univ-lille1.fr	UNIV LILLE 1 & INRIA LILLE	France
SABOT	Francois	francois.sabot@ird.fr	IRD	France
SAFFARIAN	Azadeh	azadeh.saffarian@pasteur.fr	INSTITUT PASTEUR DE PARIS	France
SAGOT	Marie-France	Marie-France.Sagot@inria.fr	INRIA	France
SAIDI	Rabie	rsaidi@ebi.ac.uk	EMBL-EBI	UK
SALSE	Jérôme	jsalse@clermont.inra.fr	INRA	France
SANDRON	Florian	sandron@cng.fr	CEA	France
SAPAY	Nicolas	Nicolas.SAPAY@bioaster.org	BIOASTER	France
SAPET	Frédéric	frederic.sapet@biogemma.com	BIOGEMMA	France
SAULNIER	Emma	emma.saulnier@ird.fr	MIVEGEC - IBC - UM	France
SCHBATH	Sophie	sophie.Schbath@jouy.inra.fr	INRA	France
SCHERMA	Frédéric	frederic.scherma@clermont.inra.fr	INRA	France
SEGARD	Camille	camille.segard@limagrain.com	LIMAGRAIN	France
SEGRE	Daniel	dsegre@bu.edu	UNIVERSITY of BOSTON	France
SEPTIER	Amandine	septier.amandine@gmail.com	UNIVERSITÉ BLAISE PASCAL	France
SEVIN	Emeric	emeric.sevin@u-bordeaux.fr	UNIVERSITÉ DE BORDEAUX	France
SICARD	Emilie	milie.sicard@hotmail.fr	IUT Aurillac	France
SIEKANIEC	Grégoire	gregoire.siekaniec@etu.udamail.fr	IUT Aurillac	France
SINAIMERI	Blerina	blerina.sinaimeri@gmail.com	INRIA	France
SOROKINA	Maria	msorokina@genoscope.cns.fr	CEA / GENOSCOPE	France
SOURDEIX	Manon	manon.sourdeix@udamail.fr	CENTRE JEAN PERRIN	France
SPATARO	Bruno	bruno.spataro@univ-lyon1.fr	CNRS	France
STEINBACH	Delphine	delphine.steinbach@versailles.inra.fr	INRA	France
TAGNY N'GOMPE	Gildas	tagnyngompe@gmail.com	CNRS	France
TAMBY	Jean Philippe	tamby@evry.inra.fr	INRA	France
TANDO	Ndomassi	ndomassi.tando@ird.fr	IRD	France
TATOUT	Christophe	christophe.tatout@univ-bpclermont.fr	CNRS / INSERM	France
TELETCHEA	Stéphane	stephane.teletchea@univ-nantes.fr	UNIVERSITÉ DE NANTES	France
TEMPEL	Sébastien	sebastien.tempel@univ-amu.fr	UNIVERSITE DE MARSEILLE	France
TERRAPON	Nicolas	terrapon@afmb.univ-mrs.fr	UNIVERSITÉ AIX-MARSEILLE	France
THOMAS-CHOLLIER	Morgane	mthomas@biologie.ens.fr	INSTITUT DE BIOLOGIE DE L'ECOLE NORMALE SUPÉRIEURE	France

THONIER	Florian	florian.thonier@inserm.fr	INSERM	France
TOFFANO-NIOCHE	Claire	claire.toffano-nioche@u-psud.fr	CNRS	France
TORCHET	Rachel	rachel.torchet@pasteur.fr	INSTITUT PASTEUR	France
TOURE	Vasundra	vtoure@eisbm.org	CNRS	France
TOURNAYRE	Jérémy	jeremy.tournayre@clermont.inra.fr	INRA	France
TOUZAIN	Fabrice	fabrice.touzain@anses.fr	ANSES	France
TRAN	Van Giang	van-giang.tran@clermont.inra.fr	INRA	France
TRAN CONG	Kevin	k.tran@bordeaux.unicancer.fr	INSTITUT BERGONIÉ	France
URBINI	Laura	laura.urbini@inria.fr	LABORATOIRE BBE	France
VALLENET	David	dvallenet@genoscope.cns.fr	CEA / GENOSCOPE	France
VAN DILLEN	Sabine	sabine.van.dillen@dupont.com	DUPONT	France
VAN HELDEN	Jacques	Jacques.van-Helden@univ-amu.fr	AIX-MARSEILLE UNIVERSITÉ	France
VARET	Hugo	hugo.varet@pasteur.fr	INSTITUT PASTEUR	France
VERNETTE	Caroline	caroline.vernette@sb-roscoff.fr	CNRS	France
VEYSSIERE	Maëva	maeva.veyssiere@clermont.inra.fr	INRA	France
VIGNE	Patricia	patricia.vigne@etu.udamail.fr	IUT Aurillac	France
VILCHEZ	Gaelle	gaelle.vilchez@toulouse.inra.fr	INRA	France
VINCENT	Jonathan	Jonathan.vincent@biogemma.com	BIOGEMMA	France
VOLANT	Stevenn	stevenn.volant@pasteur.fr	INSTITUT PASTEUR	France
WALTER	Thomas	thomas.walter@mines-paristech.fr	INSTITUT CURIE	France
WUCHER	Valentin	v.wucher@gmail.com	CNRS	France
YANG	Ziheng	z.yang@ucl.ac.uk	UNIVERSITY COLLEGE LONDON	UK
YUNFEI	Liu	liuyunfei2010@gmail.com	UNIVERSITÉ BLAISE PASCAL	France
ZANZONI	Andreas	zanzoni@tagc.univ-mrs.fr	AIX-MARSEILLE UNIVERSITÉ	France
ZINE EL AABIDINE	Amal	makrini@lirmm.fr	CNRS	France
ZOGHLAMI	manel	manel.zoghlami@gmail.com	UNIVERSITÉ DE TUNIS EL MANAR	Tunisie
ZOUAOUI	Anissa	a-zouaoui@chu-montpellier.fr	CENTRE HOSPITALIER RÉGIONAL UNIVERSITAIRE DE MONTPELLIER	France

JOBIM est le rendez-vous annuel de la communauté bioinformatique francophone, ouvert à toutes les personnes travaillant aux frontières de la biologie, de l'informatique, des mathématiques et de la physique.
La seizième édition de ces rencontres, JOBIM 2015, est organisée à Clermont-Ferrand du 6 au 9 Juillet 2015 par l'Université d'Auvergne en partenariat avec l''INRA et l'Université Blaise-Pascal, et sous l'égide de la Société Française de Bioinformatique (SFBI)
Cet ouvrage numérique rassemble les abstracts des 37 communications orales, dont 15 démonstrations académiques, et 140 affiches exposées, ainsi que les résumés des 11 conférences invitées.

Avec le soutien de :

Editeur INRA & Université d'Auvregne
ISBN : 2-7380-1377-5
Code EAN : 978 273 801 3774