Hector Escriva

Observatoire Océanologique de Banyuls sur Mer
Université Pierre et Marie Curie (UPMC)
Centre Nationale de la Recherche Scientifique (CNRS)
Banyuls sur Mer, France

*The Mediterranean amphioxus (*Branchiostoma lanceolatum*) genome indicates a stepwise evolution of vertebrate Hox bimodal regulation*

Abstract:

Today two amphioxus species have their genome sequence published. Why should we sequence a new species? In this talk I will try to answer this question and give an overview of our first results in the study of the *Branchiostoma lanceolatum* genome. These results concern the study of the chromatin structure around the Hox cluster. The HoxA and HoxD gene clusters of jawed vertebrates are organized into bipartite three-dimensional chromatin structures that separate long-range regulatory inputs coming from the anterior and posterior Hox-neighboring regions. This architecture is instrumental in allowing vertebrate Hox genes to pattern disparate parts of the body, including limbs. Almost nothing is known about how these three-dimensional topologies originated. We perform extensive 4C-seq profiling of the Hox cluster in embryos of amphioxus, an invertebrate chordate. We find that, in contrast to the architecture in vertebrates, the amphioxus Hox cluster is organized into a single chromatin interaction domain that includes long-range contacts mostly from the anterior side, bringing distant *cis*-regulatory elements into contact with Hox genes. We infer that the vertebrate Hox bipartite regulatory system is an evolutionary novelty generated by combining ancient long-range regulatory contacts from DNA in the anterior Hox neighborhood with new regulatory inputs from the posterior side

Claudia Chica

Bioinformatics and Biostatistics Hub
Centre de Bioinformatique, Biostatistique et Biology Intégrative (C3BI)
Institut Pasteur
Paris, France

Comparative epigenomics in Brassicaceae reveals distinct modes of PRC2-mediated gene regulation

Abstract:

Polycomb group (PcG) proteins form chromatin-modifying complexes involved in maintaining transcriptional gene repression in most eucaryotes. Polycomb Repressive Complex 2 (PRC2), which is responsible for the trimethylation of histone H3 at lysine 27 (H3K27me3), represents a major determinant of plant development as it affects up to 30% of genes, throughout the *Arabidopsis thaliana* life cycle.

In order to determine the rate of evolutionary changes in PRC2-mediated regulation and its relation to genome sequence, gene expression and function, we performed ChIP-seq for H3K27me3 as well as RNA-seq in three Brassicaceae species, *Arabidopsis thaliana*, *Arabidopsis lyrata* and *Arabis alpina*, with divergence times ranging from 5 to 40 million years, respectively.

Comparison of the three epigenomes reveals an overall conservation of H3K27me3 marking with some lineage-specific variations. The epigenome stability across species correlates with different degrees of conservation of promoter information content and nucleosome occupancy; it also corresponds to distinct transcriptional regimes and biological processes. In addition, syntenic blocks enriched in ancestral H3K27me3 targets (i.e. genes marked in the three species) show preferential association with genomic regions involved in long-range, intrachromosomal interactions. These data suggest that PRC2-mediated regulation is subject to multiple levels of evolutionary constraints associated with distinct modes of transcriptional repression.

Yves Clément

Institut de Biologie de l'Ecole Normale Supérieure (IBENS)
Centre Nationale de la Recherche Scientifique (CNRS)
ENS Paris, France

Core vertebrate long range cis-regulatory interactions revealed by zebrafish – human comparative genomics

Abstract:

While finding long-range regulatory regions (or enhancers) is possible through a variety of techniques, finding which genes they regulate is more challenging. For example, simply considering the nearest gene as the target gene can be lead to errors as enhancers can regulate genes located several megabases away.

We previously developed a method to 1) identify putative enhancers in genomes by looking for conserved regions in multiple alignments and 2) identifying their target genes by looking for conserved enhancer - target genes association in multiple species by computing an association score. In this approach, the rationale is that the enhancer - target gene link, if functional, will be conserved by natural selection. We applied this method to human and to zebrafish in two separate analyses. Zebrafish is an ideal species as it is a model organism for vertebrate development, during which many enhancers are known to act in a time and tissue specific way. We found around 80,000 putative enhancers with at least one predicted target gene in zebrafish and more than 1,300,000 such putative enhancers in human. In these putative enhancers, coverage by histone marks, especially marks associated with development in zebrafish, increases with association score, showing that our method correctly predicts functional interactions.

We defined orthologous putative enhancer – target gene interactions between human and zebrafish with a blast-based approach. We found a set of conserved interactions involving 150,000 enhancers in human and more than 50,000 enhancers in zebrafish targeting several thousands of orthologous genes in both species, a set that can be considered as ancestral to vertebrates. Using this set, we were able to show that orthologous genes with conserved regulation are enriched for development functions. Finally, we found that following whole genome duplication in zebrafish, enhancer retention on ohnologous genes pairs is mostly random or biased towards on one copy.

Together, these results bring exciting new insights to the function and evolution of long-range regulatory regions in vertebrates.

Delphine Larivière & David Couvin

UMR Amélioration Génétique et Adaptation des Plantes (AGAP)
Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD)
Montpellier, France

Comparative genomics of gene families in relation with metabolic pathway for gene candidates highlighting

Abstract:

The study of gene families is an important field of comparative genomics, allowing, by the analysis of evolutionary history, to identify homology relationships, gene losses and to help in annotation transfers. The addition of metabolic information improves the identification of candidate genes by addition of functional and gene network data. We propose new web systems to facilitate and improve the analysis of gene family for the search of candidate genes in plants. GenFam* is dedicated to the manual and precise analysis of gene families and includes specific workflows running under a Galaxy platform, allowing to gather several data sources, analysis and visualization tools, in order to (i) build custom families (ii) run analysis workflows dedicated to the analysis of gene families (iii) visualize analysis results and functional evidences through a dedicated visualization dashboard. In complement to the integration of data sources, tools and visualizations, we also suggest a new way to find evidences for the identification of evolutionary events through syntenic analyses. The IDEVEN algorithm is based on the study of syntenic blocks linked to a gene family to identify speciation, Whole Genome Duplication (WGD) events, and other duplications in a family history. The identified events will be reported on the phylogeny and aim to bring complementary evidences to have a clearer view of the evolutionary history of a gene family. To extend this tool to the analysis of multiple gene families and integrate metabolic pathways data, this tool has been integrated in genesPath, which will allow a deep identification and highlighting of candidate genes of interest for a specific project called “Biomass For the Future (BFF)”. This online tool will be soon available and could be notably used for searching candidate genes involved in biosynthesis of lignin and cellulose in various plant species (such as maize and sorghum).

* http://genfam.southgreen.fr/

Tristan Lefébure

Laboratoire d'Ecologie des Hydrosystèmes Naturels et Anthropisés (Lehna).
Université Claude Bernard Lyon1 (UCBL1)
Lyon, France

Reduced selection efficacy in larger genomes

Abstract:

The evolutionary origin of the striking genome size variations found in eukaryotes remains largely unexplained. The effective size of populations has been hypothesized as the key parameter controlling genome size evolution. We challenged this hypothesis using asellid isopods that have undergone multiple independent habitat shifts from surface water to low-energy groundwater. Using denovo transcriptomes, we show that these habitat shifts were associated with a long-term reduction in selection efficacy as evidenced by higher transcriptome-wide dN/dS. These independent population size reductions were paralleled by a massive increase in genome size (25% increase on average), an increase also confirmed in other groundwater taxa. Contrary to population size, life history traits such as body size and growth rate did not correlate with genome size. Genome sequencing showed that these genome size inflations were triggered by the invasion of the genome by repetitive elements. Our findings support the idea that variations in the strength of non-adaptive forces are at the root of genome size variations.

Anamaria Necsulea

Institut Suisse de Recherche Expérimentale sur le Cancer (ISREC)

Ecole Polytechnique Fédérale de Lausanne (EPFL)

Lausanne, Suisse.

Transcriptome dynamics in evolution and development

Abstract:

The development of high-throughput transcriptome profiling technologies (e.g., RNA-seq) has considerably changed our view of mammalian genome complexity. It has thus become evident that mammalian genomes encode tens of thousands of non-coding RNA genes, the most recent estimates indicating that the long non-coding RNA (lncRNAs) repertoire far surpasses the number of protein-coding genes in the human genome (Iyer et al., Nature Genetics, 2015). Moreover, numerous alternative isoforms have been discovered for protein-coding genes, adding a further layer of transcriptome complexity. However, the functional relevance of these newly identified genes and isoforms remains uncertain and rigorous tests of their functions lag behind the rapid discovery of new transcripts. Evolutionary approaches can provide important insights into their functionality, by identifying transcripts that are subject to purifying natural selection to maintain their functional properties, or to positive selection following the acquisition of new functions.

Aiming to explore the evolutionary and functional properties of protein-coding and non-coding transcripts, we generated and analyzed RNA-seq profiles in two mammalian species (mouse and rat), for four major organs (cerebral cortex, kidney, liver and testis). To obtain an overview of the temporal dynamics of gene expression throughout development, we analyzed a series of five developmental stages from mid-gestation embryo to aged individuals. We were thus able to explore the temporal and developmental transcriptional dynamics of coding and non-coding genes, including more than 15,000 previously unknown lncRNA genes in each of the two species.

We found that in both species, the majority of coding and non-coding transcripts were highly expressed in the adult testes and in isolated spermatogenic cell types, consistent with our previous observation that spermatogenesis is defined by a highly permissive chromatin environment that allows transcription of potentially non-functional genomic elements. Confirming previous findings that lncRNA repertoires evolve rapidly, we found that only approximately 60% of lncRNA genes are shared between these two closely related mammalian species. Interestingly, lncRNAs expressed in somatic organs and early in development showed increased levels of sequence and expression conservation compared to those detected exclusively in the testes and in adult and aged individuals. By contrasting expression level divergence between and within species, we show that there is significantly more selective constraint on lncRNA expression in early developmental stages than in adult and aged individuals, for all surveyed organs. Taken together, our results suggest that the majority of non-coding transcripts may be of little functional relevance, but they also indicate that functional lncRNAs may be preferentially associated with developmental processes, outlining an important direction for further studies of non-coding transcript functionality.

Carine Rey[1,2], Bastien Boussau[2] and Marie Sémon[1]

1. Laboratoire de Biologie et Modélisation de la Cellule (LBMC), Ecole Normale Supérieure de Lyon, Lyon, France.
2. Laboratoire de Biométrie et de Biologie Evolutive (LBBE), Université Claude Bernard, Lyon1, Lyon, France.

Genome-scale detection of interspecies convergent evolution: problems and prospects

Abstract:

Convergent phenotypes are widespread in nature, yet we only have a fragmented grasp of the genetic mechanisms underlying their evolution. Recently a controversy erupted around studies aiming at finding genome scale convergent evolution in species with convergent phenotypes: some studies find vast amounts of genomic convergence, while others do not.

Differences between those studies (which used the same data) may come from the design of the dataset, in which two independent events of convergence are considered. This narrow sampling may lead to a lack of statistical power and a high rate of false positives. Indeed, some convergent genomic changes can be due to chance and not correlated with a convergent phenotype.

We simulated evolution of sequences along trees and within them we changed the evolution model at randomly chosen nodes to simulate a convergent change in the substitution pattern. The number of nodes is thus the number of considered independent events of convergence in the dataset. From these simulated sequences, we tested our ability to detect this substitution pattern change in function of the number of considered nodes. Although our simulation settings, in which lots of parameters are known, should make it easy to detect convergent events, the sensibility of our test is paltry under 5 events of convergence. Using several events of convergent evolution is thus primordial to detect genomic convergence with statistical confidence.

Even if there is a lot of available genomic data, it is still hard to find enough data for species with a convergent phenotype to detect convergent genomic changes. RNA-seq data are today a good way to study transcriptome of non-model species, but getting coding sequences (assembly and annotations) and merging them with genomic data is complex especially if there are no phylogenetically close reference genomes.

To rise to this challenge, we built a robust and accurate pipeline (which is available on github) to merge data from transcriptomes and genomes in order to get datasets with numerous species. We used methods to assemble high quality transcripts from RNA-seq reads, took special care to build accurate sequence alignments, avoided potential artifacts by using reconciliation methods to build accurate gene trees.

Datasets built using this pipeline can then be used to study interspecies convergent genomic evolution.