Genomic context influence on gene evolution

Federico Abascal [1], María Victoria Aguilar [1], David de Juan [1], Daniel Rico [1], Alfonso Valencia [1]§

[1] Structural Biology and Biocomputing Program, Spanish National Cancer Research Centre (CNIO), C/ Melchor Fernández Almagro, 3, E-28029 Madrid, Spain
§ Corresponding author valencia@cnio.es
FA is the poster presenter fabascal@cnio.es

Patterns and rates of nucleotide substitutions, insertions and deletions vary largely across the human genome. GC content heterogeneity explains a substantial fraction of this variability, supporting the use of GC content as a descriptor of different genomic contexts. We recently found evidence for the influence of the genomic context on the evolution and subfunctionalization of the two ASF1 paralogs in vertebrates (Abascal et al, 2013). Whereas 3rd codon positions almost fully diverged in GC-content, most non-synonymous sites remained conserved in the two paralogs. Our results suggested a large impact of the genomic context on divergence at the functional level, including gene expression regulation, amino acid usage, and ultimately, interaction with other partners.

Here, we present results of a genome-wide assessment of the influence of the genomic context on the evolution of genes (Aguilar et al, in preparation). We looked for GC-content shifts along phylogenetic trees of vertebrate orthologs and paralogs. When the GC-content varies between orthologs, divergence is expected to be mostly related to the genomic context. In contrast, divergence between paralogs may also reflect changes in protein function. We found clear evidence of an acceleration in evolutionary rates and differential trends in amino acid replacements associated to GC-content shifts in orthologs. Interestingly, genes showing GC-content shifts accumulated in subtelomeric of the genome, which are known to be highly recombinogenic. Regarding paralogs, our results revealed a clear relationship between GC-content differences and the age of the duplication, as well as between the GC-content and the breadth of expression. Remarkably, we found evidence supporting the influence of the genomic context in the acquisition of either general or tissue-specific patterns of expression after gene duplication.

Abascal F, Corpet A, Gurard-Levin ZA, Juan D, Ochsenbein F, Rico D, Valencia A, Almouzni G (2013). Subfunctionalization via Adaptive Evolution Influenced by Genomic Context: The Case of Histone Chaperones ASF1a and ASF1b. *Mol Biol Evol*, **30**, 1853–1866.

Aguilar MV, Juan D, Rico D, Valencia A, Abascal F (2013). Influence of the Genomic Context on Gene Evolution. *In preparation*.

Rapid turnover of recombination hotspots since the divergence between Denisova and modern humans

Yann Lesecque [1], Dominique Mouchiroud [1], Laurent Duret [1] §.

[1] Laboratoire de Biométrie et Biologie Evolutive (UMR CNRS 5558) 43 bd du 11 novembre 1918 69622 Villeurbanne.
§ Corresponding author laurent.duret@univ-lyon1.fr
YL is the poster presenter yann.lesecque@univ-lyon1.fr

In mammalian genomes, meiotic recombination occurs predominantly in hotspots, whose position evolves very rapidly. Indeed, despite considerable similarity between human and chimpanzee sequences, their recombination hotspots are located at different sites. It is now clear that the PRDM9 protein plays a major role in determining hotspots location in the human genome, by binding a 13-bp consensus motif (HM-motif). It has been suggested that the turnover of recombination hotspots was driven by the evolution of PRDM9 DNA-binding domain, but the dynamics of this process remains unclear. The “hotspot paradox” model predicts that if a motif determines the position of a hotspot, then it is expected to rapidly fix mutations by biased gene conversion in favor of "cold" alleles. We used this specific signature of recombination hotspots activity to study their evolution since the divergence from Chimp, using Denisova sequences to date mutations affecting HM-motifs. We show that mutated HM-motifs are strong predictors of hotspot activity. The accumulation of mutations in different branches of the phylogeny (modern Humans, Denisova and their ancestor) indicates that the HM-motif was already a target of PRDM9 at least 1.5 Myrs ago, i.e. before the *Homo sapiens*/Denisova divergence. The observed substitution rate suggests that on average, mutations of HM-motifs decrease hotspots recombination activity by about 20%. It has been previously shown that the substitution pattern around human recombination hotspots is strongly biased towards GC, as a consequence of GC-biased gene conversion. Interestingly, we show that this GC-bias results specifically from substitutions in the modern human branch. Furthermore, we observed that HM motifs that were mutated in the Denisova lineage do not overlap with human hotspots. These observations indicate that the location of recombination hotspots differed between Denisova and modern humans, and hence reveal an extremely fast turnover of hotspots, despite sharing similar PRDM9 target motifs.

Is RAD-seq suitable for short-scale phylogenetic inference ? An in silico assesment

Marie Cariou, Laurent Duret, Sylvain Charlat

Laboratoire de Biométrie et Biologies Evolutive (UMR 5558, Université Lyon 1)

Abstract
Resolution of phylogenetic relationships between closely related species can be hindered by several problems. First, most nuclear markers lack informative variation at short evolutionary timescale. Second, because of incomplete lineage sorting and introgression, phylogenies infered from particular loci can differ from the average genome phylogeny. Finally, PCR-primer information can be lacking in poorly studied taxa. In this context Restriction site Associated DNA sequencing (RADseq) seems promissing (Davey *et al.* 2011). This technique can generate sequence data from DNA fragments flanking restriction sites, thus randomly distributed throughout the genome, from a large number of samples and without preliminary knowledge on the taxa under study.
RADseq was first developed for population genetics and quantitative trait mapping. The suitability of this method for phylogenetic inference, relying on the presence of conserved restriction sites in different species, thus remains to be evaluated. This need motivated the present study, where we simulated a RADseq experiment using the 12 *Drosophila* genomes. More than 100 restriction sites were conserved between the most distant species which diverged 40 million years ago. Inter-individual clustering of RAD sequences retrieved the majority of known orthologs. Using these data, we were able to recover the expected phylogenetic relationships between the 12 *Drosophila* species, with strong statistical support. This study therefore validates the suitability of the RADseq technique for phylogenetic inference between closely related species.

Références

Davey J. L. and M. W. Blaxter (2011) RADSeq: next-generation population genetics. *Briefings in functional genomics.* **9(5)** : 416-423

Hydrogen bonds are related to the thermal stability of 16S rRNA

Hiroshi Nakashima, Ai Fukuoka, Yuka Saitou

Department of Clinical Laboratory Science, School of Health Sciences,
Kanazawa University, Kanazawa, Japan
E-mail: naka@kenroku.kanazawa-u.ac.jp

Abstract

The number of base pairs in the 16S rRNA secondary structures of 51 bacterial sequences was counted, and the number of hydrogen bonds was estimated. The number of hydrogen bonds was highly correlated with the optimal growth temperature (OGT) rather than with the G+C content. Paired and unpaired nucleotides in mesophiles were compared to those in thermophiles. OGT exhibited a relationship with paired nucleotides but not with unpaired nucleotides. The total number of paired as well as unpaired nucleotides in mesophiles was very similar to that in thermophiles. However, the components in base pairs in mesophiles significantly differed from those in thermophiles. As compared with mesophiles, the number of G•C base pairs in thermophiles was high whereas that of A•U base pairs was low. In this study, we showed that hydrogen bonds are important for stabilizing 16S rRNAs at high temperatures.

Microorganisms can live in a wide temperature range from the freezing point of water to its boiling point. This indicates that the environment where water exists in the liquid state can be inhabited by microorganisms. At their living temperature, macromolecules such as protein, DNA and RNA are stable and can perform their biological functions. DNA is double stranded and RNA is single stranded. The method of DNA stabilization at high temperatures is different from that of RNA. The dinucleotide composition of DNA is related to the optimal growth temperature (OGT), and mononucleotide composition i.e., G+C content of RNA is proportional to their OGT. Hyperthermophiles have higher RNA G+C content. The G•C base pair has 3 hydrogen bonds and A•U base pair has 2 hydrogen bonds. Therefore, hydrogen bonds seem to play an important role for RNA thermal stability, however, the relationship between the number of hydrogen bonds and OGT has not reported yet. Ribosomes are the machinery necessary to produce proteins based on the mRNA, which is a blueprint of genetic information. There are 3 types of bacterial ribosomal RNAs—5S, 16S, and 23S named according to their molecular weights. 16S rRNA is the most conservative of the 3 rRNAs, and is used to identify bacterial species on the basis of the phylogenetic tree. The Gutell group predicted base pairs in 16S rRNA of bacteria, which are available through the web. Using these data, base pairs in the 16S rRNA structures were counted and the number of hydrogen bonds was estimated.

New Universal Rules of Eukaryotic Translation Initiation Fidelity

Hadas Zur [1] and Tamir Tuller [2§]

[1] School of Computer Science, Department of Exact Sciences, Tel Aviv University
[2] Department of Biomedical Engineering, Faculty of Engineering, Tel Aviv University, Ramat Aviv 69978, Israel.
[§] Corresponding Author tamirtul@post.tau.ac.il
HZ is the poster presenter zurhadas@post.tau.ac.il

The accepted model of eukaryotic translation initiation begins with the scanning of the transcript by the pre-initiation complex from the 5'end until an ATG codon with a specific nucleotide (nt) context surrounding it is recognized (Kozak rule). A fundamental biological question in the field is related to the way the efficiency and fidelity of this stage is encoded in the transcripts, including the ORFs, and affects its evolution.
We perform for the first time, a genome-wide statistical analysis, uncovering a new, more comprehensive and quantitative, set of initiation rules for improving the cost of translation and its efficiency. Analyzing dozens of eukaryotic genomes, we find that in all frames there is a universal trend of selection for low numbers of ATG codons; specifically, 16--27 codons upstream, but also 5--11 codons *downstream* of the START ATG, include less ATG codons than expected. We further suggest that there is selection for *anti optimal* ATG contexts in the vicinity of the START ATG. Thus, the efficiency and fidelity of translation initiation is encoded in the 5'UTR as required by the scanning model, but also at the beginning of the ORF.
The observed nt patterns suggest that in all the analyzed organisms the pre-initiation complex often misses the START ATG of the ORF, and may start translation from an alternative initiation start-site. Thus, to prevent the translation of undesired proteins, there is selection for nucleotide sequences with low affinity to the pre-initiation complex near the beginning of the ORF. With the new suggested rules we were able to obtain a twice higher correlation with ribosomal density and protein levels in comparison to the Kozak rule alone (*e.g.* for protein levels $r = 0.7$ vs. $r = 0.31$; $p < 10^{-12}$).

SEX-DETector: a method for studying sex chromosomes and sex determination in non-model organisms using high-throughput sequencing data

Aline Muyle [1§], Franck Picard [1], Sylvain Mousset [1], Gabriel Marais [1]

[1] Laboratoire de Biométrie et Biologie Evolutive (UMR 5558), CNRS/Université Lyon 1, Villeurbanne, France
[§] Corresponding author aline.muyle@univ-lyon1.fr

Our current views on sex chromosome evolution mainly come from a few model systems such as humans and drosophila. The advance of next generation sequencing (NGS) technologies provides an unprecedented opportunity to study other "non-model" systems and test how general these views are. One very promising approach relies upon RNAseq data from male and female individuals which has been successfully applied to the dioecious plant *Silene latifolia* in three independent recent studies.
However, identifying sex-linked SNPs and contigs from RNAseq data reliably is not a simple task and this was done rather empirically in previous work. Here we propose a new pipeline that relies on genotypes from a new genotyper designed for species without reference genome. It also can deal with genes with differences in expression among alleles, which is key for sex-linked genes as the Y copy usually shows reduced expression. We then introduce a new method based on a model of allele transmission from parents to progeny to compute the probability of a contig to be sex-linked (both X and Y copies), X hemizygous (X-linked only) and autosomal, which uses an expectation maximization (EM) algorithm. This method can be run on RNAseq data of three types: (i) individually-tagged parents + individually-tagged progeny, (ii) individually-tagged parents + pooled progeny (one pool / sex) and (iii) individually-tagged brothers and sisters from an inbred line. Our pipeline was tested on a set of experimentally known autosomal and sex-linked genes in *Silene latifolia* and gave better results than previous work. It was also tested using simulations.
Our pipeline will be very useful to identify sex-linked genes in a number of non-model organisms where little or nothing is known on sex chromosomes and sex determination. Our pipeline will be available in a Galaxy workflow, a user-friendly platform widely used for analyzing NGS data.

Comparison of transposable elements insertions between *Latimeria chalumnae* and *Latimeria menadoensis*: is the "living fossil" genomically inert?

Magali Naville [1 * §], Domitille Chalopin [1 *] and Jean-Nicolas Volff [1]
[1] Institut de Génomique Fonctionnelle de Lyon, Ecole Normale Supérieure de Lyon.
[*] Equal authors contribution
[§] Corresponding author magali.naville@ens-lyon.fr

The coelacanth is a lobe-finned fish that has been considered extinct for 70 million years, until the discovery, in 1938, of a living specimen in South Africa. Nowadays, a total of 309 coelacanth individuals have been inventoried. From the evolutionary point of view, the coelacanth occupies a key phylogenetic position between ray-finned fish and tetrapods. With fossils dating back 300 mya that highly look like current specimens, coelacanths have been largely placed in the arguable class of "living fossils", and the hypothesis of such morphological stasis relying on an almost missing genomic evolution has been proposed.
The recent availability of coelacanth sequence data gives the opportunity to test this assumption. Since the genome of the African coelacanth *Latimeria chalumnae* is now sequenced, and since few BACs from its related *Latimeria menadoensis* are now available, an intra-genus comparison can be drawn that will provide information concerning the possible genome stasis following the separation of the two *Latimeria* species. One of the major and dynamic part of vertebrate genomes consists in transposable elements (TEs). TEs are DNA sequences that can change their position within the genome. They are sorted in several classes and families according to their transposition mode and to specific sequence features. Produced in the context of the coelacanth project, the coelacanth TE library show an intermediate diversity of families compared to mammals and fish, which reflects the intermediate position of *Latimeria* among vertebrates. The careful comparative analysis of TEs insertions between *L. chalumnae* and *L. menadoensis* we are developping is the first high-scale attempt applied in this clade that will possibly highlight direct TEs movements through an insertion polymorphism. These information allow to discuss the idea of an inactive genome as basis of coelacanths morphological stasis, and help to understand the particular evolution of this intriguing vertebrate species.

Identifying long range cis-regulation in the human genome using evolutionary co-segregation

Magali Naville [1], Alexandra Louis [2] and Hugues Roest Crollius [2 §]
[1] Institut de Génomique Fonctionnelle de Lyon, Ecole Normale Supérieure de Lyon, CNRS UMR 5242 - INRA USC 1370, 46 allée d'Italie, 69364 Lyon cedex 07, France
[2] Institut de Biologie de l'Ecole Normale Supérieure, UMR8197 CNRS, 46 rue d'Ulm, 75005 Paris, France
[§] Corresponding author hrc@ens.fr
poster presenter: magali.naville@ens-lyon.fr

Recent systematic screens of exonic sequences in patients cohorts suggest that only a small fraction of diseases is caused by coding variations [Tarpey2009], while a significant fraction could involve non-coding regulatory loci. Functional data now make it easier to find active regulatory elements, but the identification of their target genes, which still needs heavy experimental procedures, remains a challenging task. Due to the complex 3D structure of the vertebrate chromosome, enhancers often target genes distant from several kb, and simply hypothesizing the nearest gene as being the regulatory target is often incorrect. Computational prediction of such functional pairs is thus of great interest. Here we propose a bioinformatic method that identifies cases of evolutionary co-segregation between putative enhancers and specific genes, using the information on a large range of genomes. The method was developed on the human X chromosome.

Putative enhancers are, in a first approximation, assimilated to Conserved Non-coding Elements (CNEs). As the simple conservation criteria does not stand exclusively for enhancers, we further use different annotations to characterize them: ENCODE data, co-activator p300 binding sites, or experimentally validated enhancers from literature. The target prediction consists in collecting immediate neighbor genes (within 1.5Mb) of each CNE in the human reference genome. A score of co-segregation then measures the frequency of co-retention of each CNE-gene pair during evolution, taking into account the rearrangement rate of the different genomes as well as their coverage. Genes showing the highest values of this score are considered as the most likely targets.

Strikingly, CNEs with high co-segregation scores are enriched in functional annotations, supporting the initial premises that elements segregating with genes are more likely to be true enhancers. Among the 1,704 “best” CNEs selected on their functional characteristics and associated with high confidence to a particular gene, 732 are associated to neurological disease genes. In collaboration with several human genetics groups, we are working on applying our procedure to detect pathological mutations in such non-coding regions. More globally, this study will allow the reconstruction of the evolution of chromosomal regulatory circuits, and a better understanding of the role of long range cis-regulation in negative selection of genomic rearrangements.

Classification of bacterial replicons based on the Genetic Information Transmission Systems. A comparative genomic approach

Olivier Poirion [1], Bénédicte Lafay [1 §]

[1] CNRS UMR5005-Laboratoire Ampère, École Centrale de Lyon, 36 avenue Guy de Collongue, F-69134 Écully
[§] Corresponding author benedicte.lafay@ec-lyon.fr
Poster presenter olivier.poirion@ec-lyon.fr

Abstract
The genome of bacteria is classically separated into the essential, stable and slow evolving chromosomes and the accessory, mobile and rapidly evolving plasmids. This distinction has been blurred in recent years by the characterization of multipartite genomes constituted of a primary "standard" chromosome and one or several additional and essential replicons adapted to the cell cycle. Depending on the authors, these genomic elements are either named "secondary chromosomes" or "megaplasmids". However, their true nature and evolution are yet to be determined. Here we investigate the relationships of these secondary essential replicons (SERs) to classical chromosomes and plasmids based on the key processes involved in the maintenance of genomes and replicons, i.e., their replication, partition and segregation, and perform a global comparative genomic analysis for all bacterial replicons available from public databases based on graph and data-mining methologies. Several classes of replicons could thus be characterized, and chromosomes, plasmids and SERs differentiated. This sets the basis to the investigation of the emergence of SER-like genomic structures.

Genomic context and distribution of effector genes in the plant pathogenic fungus *Fusarium oxysporum*

Sarah M. Schmidt [1], Peter van Dam[1], Petra M. Houterman [1], Ines Schreiver [2], Lisong Ma [1], Stephan Amyotte [3], Biju Chellappan [1], Sjef Boeren [4], Frank L. W. Takken [1], Martijn Rep [1§]

[1] Molecular Plant Pathology, Swammerdam Institute for Life Sciences, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, the Netherlands,
[2] Fachgebiet Medizinische Biotechnologie, Institut für Biotechnologie, Technische Universität Berlin, Gustav-Meyer-Allee 25, Germany
[3] Department of Plant Pathology, University of Kentucky, 201F Plant Science Building, 1405 Veterans Drive, Lexington, KY 40546-0312, USA
[4] Laboratory for Biochemistry, Wageningen University, Dreijenlaan 3, 6703HA, Wageningen, the Netherlands
[§]Corresponding Author m.rep@uva.nl
SMS is the poster presenter s.m.schmidt@uva.nl

Strains of the *Fusarium oxysporum* species complex (FOSC) are able to infect a wide range of mono- and dicotyledonous plants. Based on the host specificity of individual strains, the FOSC is divided into various *formae speciales*. All strains share a common core genome and possess additional lineage-specific (LS) chromosomes. The fungus secretes effector proteins into the host vascular system that presumably manipulate the host to promote infection. In the tomato pathogen *F. oxysporum* f. sp. *lycopersici* (*Fol*) these effectors are encoded by *SIX* (Secreted In Xylem) genes. Interestingly, all *SIX* genes are present on a single LS chromosome that can be transferred horizontally to a previously non-pathogenic *Fo* strain, resulting in gain of pathogenicity towards tomato. Upon close inspection of this tomato pathogenicity chromosome we discovered that a non-autonomous miniature transposable element (mite) is present in the promoters of all *SIX* genes. Promoter deletion analysis at two different *SIX* gene loci did not reveal a direct role of the mite for *SIX* gene expression. However, we were able to use this genomic signature to predict novel effector gene candidates in the *Fol* genome. Expression of several of these novel candidates during infection was confirmed by mass spectroscopic analysis of the xylem sap of *Fol*-infected tomato plants. We also discovered a small reservoir of 'silent' effector genes that are not expressed during infection. Next, we used our method to predict effector gene candidates in the genomes of several other *formae speciales* and developed a more global picture of the effector gene complement in the FOSC. Effector genes in *Fo* consistently reside in repeat-rich enviroments. Some strains contain 2 or 3 paralogs of an effector gene. Additionally, many genomes feature truncated effector gene homologs. Overall, the effector gene distribution among different *formae speciales* is patchy, and there is no unique set of effectors that is common to all plant pathogenic strains of the FOSC.

ProteINSIDE: a web service to computerize an in-depth analysis of functions, sequences, secretions and interactions for proteins.

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, F-63122 Saint-Genès-Champanelle, France
2 Clermont Université, VetAgro Sup, UMR1213 Herbivores, BP 10448, F-63000, Clermont-Ferrand, France
Corresponding author: nicolas.kaspric@clermont.inra.fr

Given the increasing quantity of genomic data produced, a strategy to perform a systemic and integrative analysis of protein biological meanings is to develop an online workflow with an interface devoted to reachable and fully customizable analysis and to an easy view of the results.
From a list of proteins/genes, ProteINSIDE collects and stores accession numbers, protein sequences and biological meanings from UniprotKB database and NCBI. ProteINSIDE annotates proteins/genes according to the Gene Ontology (GO) from UniprotKB database and calculates over- and under-represented terms. ProteINSIDE proposes a list of putative secreted proteins by both the prediction of signal peptides with SignalP software and the search for GO terms relevant to the secretion process. Protein-protein interactions (PPi) are identified with PsicQuick web service; our script checks 28 PPi databases and collects experimentally proven PPi. PPi are visualized by a cytoscape web implementation, with customizable view and analysis of network based on beetweeness and closeness centralities. A web interface for ProteINSIDE gives results by modules of analysis as dynamic tables and diagrams.
We validated this web service on a test list made of human proteins from glycolysis, citric acid cycle and hormones. ProteINSIDE collected sequences and biological meanings (functions, tissue specificity and cellular location) for 100% of these proteins. ProteINSIDE detected 93% of expected signal peptides and annotated at least 100% of the proteins involved in glycolysis and citric acid cycle with GOs related to these pathways.
In conclusion ProteINSIDE saves time and treatment information for researchers who get results from several softwares and database with a single query. ProteINSIDE imports accession numbers (usable in other web services) and biological meanings, identifies over- and under-represented biological functions and secreted proteins, and constructs a network from large lists of proteins/genes or accession numbers. ProteINSIDE is freely available at: http://147.99.129.193/i-analyse/index.php.

EMBnet: The Global Bioinformatics Network

Domenica D'Elia [1§], Erik Bongcam-Rudloff [2], Etienne de Villiers [3], Pedro L. Fernandes [4], Andreas Gisel [1], George Magklaras [5], Teresa K. Attwood [6]
[1] CNR – Institute for Biomedical Technologies, Bari, Italy
[2] Department of Animal Breeding and Genetics, SLU-Global Bioinformatics Centre, Swedish University of Agricultural Sciences, Uppsala, Sweden
[3] Centre of Geographical Medicine Research Coast (CGMRC), Kenya Medical Research Institute, Africa
[4] Instituto Gulbenkian de Ciência, Oeiras, Portugal
[5] Biotechnology Center of Oslo, University of Oslo, Norway
[6] Faculty of Life Sciences & School of Computer Science, The University of Manchester, United Kingdom
[§] Corresponding author domenica.delia@ba.itb.cnr.it

Abstract

EMBnet is a global network of bioinformatics communities committed to providing expertise and services to support and advance research in the fields of bioinformatics, the life sciences and biotechnology. EMBnet's mission is to provide bioinformatics education and training, to exploit network infrastructures to investigate, develop and deploy state-of-the-art public software, to assist biotechnology- and bioinformatics-related research, bridging commercial and academic sectors. EMBnet has thus been, and continues to be, both a pioneer and provider of bioinformatics services, expertise and training, and an incubator of multidisciplinary projects across Europe and beyond (*e.g.*, members of EMBnet are playing leading roles in the NGS data-analysis network, SeqAhead COST Action and AllBio Coordination Action). EMBnet has supported the creation and growth of bioinformatics and computational biology communities in many countries, aiding both the implementation of bioinformatics infrastructures and the development and use of the most advanced bioinformatics tools and resources. As a global bioinformatics network, EMBnet shares common interests and works closely with other international organisations, networks and societies, including the ISCB, APBioNet, SoIBio, ASBCB, ISB and GOBLET. In particular, GOBLET (Global Organisation for Bioinformatics Learning, Education & Training), an initiative spear-headed by EMBnet, currently includes 26 participating organisations and several individuals, working together to provide a global, sustainable support and networking infrastructure for bioinformatics trainers and students. EMBnet also provides links and support to many smaller collaborative and thematic networks, and produces a variety of widely read online publications, such as *EMBnet.journal*, EMBnet.digest and EMBnet QuickGuides. Since January 2013, EMBnet has enabled and encouraged individual membership, allowing academics, postdoctoral researchers, students, teachers and industrialists to join the oldest bioinformatics network in the world, to benefit from participation in its Special Interest Groups, its Project Committees and International Alliances, and to gain visibility for their activities via EMBnet's website and publications.

Links: www.embnet.org www.embnet.org/joinus journal.embnet.org/index.php/embnetjournal

Horizontal Gene Transfer and Nitrogen Fixation Genes in Gamma- and Beta-proteobacteria

Luiz T Rangel [12], João C Setubal [1§]

[1]Instituto de Química, Universidade de São Paulo, São Paulo, Brasil
[2]Interunidades em Bioinformática, Universidade de São Paulo, São Paulo, Brasil
[§] Corresponding author setubal@iq.usp.br
LTR is the poster presenter lthiberiol@gmail.com

We report results of phylogenetic and phylogenomic analyses of genes related to nitrogen fixation in Gamma- and Beta-proteobacteria. Based on our results we hypothesize that 25 genes related to nitrogen fixation have been horizontally transferred from a diazotroph ancestor of Gamma-proteobacteria to the ancestor of *Candidatus Accumulibacter phosphatis*, *Dechloromonas aromatica* and *Azoarcus sp.* (and a possible secondary transference to *Sideroxydans lithotrophicus*), all of which are Beta-proteobacteria (henceforward named 4B). We clustered homologous proteins of 463 Proteobacteria organisms, only one strain per species, using OrthoMCL. Using as a filter the required presence of *nifHDKENB* genes we identified 83 organisms as real or putative diazotrophs. We then built phylogenetic trees for each of these six genes. The genes of the four Beta-recipients were more closely related to genes from *Pseudomonas stutzeri*, *Azotobacter vinelandii*, *Allochromatium vinosum*, *Thiocystis violascens*, *Teredinibacter turnerae* and *Methylomonas methanica* than to genes of the additional nine Beta-proteobacteria present in the set of 83 organisms. We then built individual trees for genes shared by all 4B genomes and their homologs in the other 459 organisms, when present (1298 homolog groups); we detected a similar pattern in the trees of another 19 gene families. All of them are present in the diazotroph island identified in *Pseudomonas stutzeri* A1501, and are thus presumably related to nitrogen fixation. The other 9 Beta-proteobacteria organisms tend to cluster together, separately from the 4B organisms, confirming a distinct evolutionary history. The conservation of the diazotrophy island of *Pseudomonas stutzeri* A1501 was detected in the genomes of the 4B genomes, while no significant alignment was observed in the genomes of additional Beta-proteobacteria. The island fragments identified in the genomes has suffered some rearrangements. We did not detect any evidence of HGT by compositional methods, suggesting that the hypothesized transfer is ancient. We hypothesize that the 4B genomes are recipients of nitrogen-fixing genes in an ancient horizontal gene transfer from gamma-proteobacteria.

The Evolution of Base Composition in Grass Plants

Yves Clément [§], Margaux-Alisson Fustier, Benoit Nabholz, Sylvain Glémin
Institut des Sciences de l'Evolution de Montpellier, Montpellier, France
Poster presenter: yves.clement@univ-montp2.fr

Abstract:
In grass plants such as rice or maize, the distribution of GC-content of third codon positions is well known to be bimodal. This feature is thought to be specific to grass plants as closely related species like banana have an unimodal GC-content distribution. Until recently, because of a lack of genomic sequence, the origin of the peculiar GC-content distribution in grasses remains unknown. Indeed, only grass genomes were available inside monocotyledons and while the phylogenetic sampling outside this group was lacking. The recent publications of several complete genomes and transcriptomes of non grass monocots allows us to study with details the evolution of GC-content within monocots.
We studied more than 1,000 groups of one to one orthologous genes in seven grass plants and two outgroup species (banana and palm tree). Using a maximum likelihood-based method, we reconstructed for each group the GC-content at third codon positions at several ancestral nodes. We found that the bimodal GC-content distribution observed in grass plants is ancestral to both grasses and outgroup species, and that other species have lost this peculiar structure. We also found that GC-content in grass lineages is evolving very slowly. While in the two outgroup species GC-content at third codon positions is globally decreasing, in the ancestor of grass plants first exons of genes saw their base composition increasing while the rest of genes did not. Such observations are consistent with an influence of GC-biased gene conversion on GC-content evolution in plants. Globally, these findings have implications for plant genome evolution as well as phylogenetic reconstructions in plants.

Modeling the evolution of gene relationships

Magali Semeria [1§] , Laurent Guéguen [1] , Eric Tannier [1 2]

[1] Laboratoire de Biométrie et Biologie Évolutive, UMR CNRS 5558, Université Lyon 1, 69622 Villeurbanne, France
[2] INRIA Grenoble Rhône-Alpes, 38334 Montbonnot, France
[§] Corresponding author and poster presenter: magali.semeria@univ-lyon1.fr

Abstract
A genome is not only a set of independent genes. It can be seen as an organized and functioning set of interactions between genes that are embedded in their environment. To study the evolution of genomes, some integrative methods have been developed that reconstruct species and gene history simultaneously, and take into account sequence evolution, gene birth, duplication, transfer and loss. These methods give a first hint at the gene content of ancestral species, but since they assume that every gene evolves independently from each other, they do not give information about the relationships between these genes.
We propose a method to model the evolution of these relationships. Our method aims to be general but, at first, we apply it to modeling the evolution of gene adjacencies on chromosomes. Given a species tree, a set of gene trees, a set of present gene adjacencies, and a model of adjacencies evolution, we calculate the probability of adjacencies along the branches of the gene trees. As genes undergo evolutionary events such as duplication and rearrangement, adjacencies can be kept, gained or lost. The pseudo-likelihood of observed adjacencies can be computed using the usual dynamic algorithm proposed by Felsenstein.
We thus establish the methodological basis that will enable us to integrate information about gene relationships into models of genome evolution. This information will allow us to improve the reconstruction of gene and species histories. It will also be helpful to reconstruct the genomes of ancestral species.

The Algebraic Theory for Genome Rearrangements

Pedro Feijao [1 §] , João Meidanis [2]

[1] Genome Informatics Group, Faculty of Technology, Bielefeld University, Germany.
[2] Institute of Computing, University of Campinas, Brazil.
[§] Corresponding author and poster presenter: pfeijao@cebitec.uni-bielefeld.de

Abstract

Genome rearrangements are evolutionary events where large, continuous pieces of the genome shuffle around, changing the order of genes in the genome of a species. Gene order data can be very useful in estimating the evolutionary distance between genomes, and also in reconstructing the gene order of ancestral genomes.

In 2000, Meidanis and Dias proposed a framework for studying rearrangement problems, called "Algebraic Formalism", based on permutation groups to model genomes and rearrangement operations. In its original formulation, it focuses on representing the order in which genes appear in chromosomes, and applies to circular chromosomes only.

Recently, Feijão and Meidanis introduced an extension of this formalism, called the "Adjacency Algebraic Theory", where permutations represent the adjacencies between genes in a genome. This allowed the algebraic theory to model linear chromosomes and the use of the original algebraic distance formula in the general multichromosomal case, with both linear and circular chromosomes. It was also shown that there is a direct relationship between the original model "chromosomal" and the adjacency model.

This resulting algebraic rearrangement distance is very similar, but not quite the same, to the Double-Cut-and-Join distance, a well known comprehensive model of genome rearrangents.

This poster is intended as an introduction to algebraic concepts used in genome rearrangement problems, where we will present the main ideas of the Algebraic Theory and some of its most recent developments.

Toward the identification of feminizing genes of the bacterial endosymbiont Wolbachia

Myriam Badawi, Isabelle Giraud, Pierre Grève, Richard Cordaux [§]

Laboratoire Ecologie et Biologie des interactions UMR CNRS 7267, Equipe Ecologie Evolution Symbiose, Université de Poitiers, Bât. B8, rue Albert Turpin, 86022 Poitiers Cedex, France
[§] Corresponding author richard.cordaux@univpoitiers.fr
Poster presenter myriam.badawi@univpoitiers.fr

Symbiotic interactions are a major driver of evolution. The symbiont genotype is able to alter the host phenotype, and the other way round : it is called “extended phenotype”. In this respect, Wolbachia endosymbiosis is remarkable. Indeed, this intracellular bacterium is a wellknown reproductive parasite able to induce feminization of genetic males or cytoplasmic incompatibility (which cause infertile crossings) in its terrestrial isopod crustacean hosts. Currently, no molecular genetic basis of these reproductive manipulations has been described. In order to identify genes involved in the feminizing process, we used a comparative genomics approach using the genomes of two closely related strains of Wolbachia : one inducing feminization of its host *Armadillidium vulgare* (wVulC) and the other one inducing cytoplasmic incompatibility in *Cylisticus convexus* (wCon). The effect induced by these strains is considered strainspecific as it is conserved when Wolbachia is crosstransfected to the other host. As the wVulC genome has previously been sequenced, we first sequenced the genome of wCon using pyrosequencing 454 (Roche). We assembled the wCon genome in ~200 contigs for a total length of ~2Mb and ~2000 predicted genes. We then elaborated a bioinformatics pipeline that compared all the wVulC coding sequences with those of wCon. Thus, we eliminated all the genes supposed not to be feminizing, as the repeats, the genes of the Wolbachia core genome and all the genes with an aminoacid sequence similar to wCon. Therefore, we determined ~300 candidate genes for feminization. Finally, we investigated the expression of the candidate genes during host sexual differentiation in order to pinpoint genes involved in the manipulation of host sex determination. This work is funded by an ERC grant (EndoSexDet) to RC, which aims to identify the genetic factors implicated in the sex determination of *A. vulgare*, thereby contributing to evaluate the evolutionary impact of endosymbionts on sex determination of their hosts.

Degeneration of mating type chromosomes in the anther smut Microbotryum fungi

Eric Fontanillas[1§], Michael E. Hood[2], Elsa Petit[2], Valérie Barbe[3], Gabriela Aguileta[4], Julie Poulain[3], Patrick Wincker[3], Christina Cuomo[5], Michael H. Perlin[6], Tatiana Giraud[1]

[1] Paris-Sud, France.
[2] Amherst College, Amherst, Massachusetts
[3] Genoscope, Evry, France
[4] CRG Barcelona, Spain
[5] Broad Institute of MIT and Harvard, Cambridge, Massachusetts
[6] University of Louisville, Kentucky
[§] Corresponding author and poster presenter: eric.fonta@gmail.com

Currently several hypotheses about the evolution of sexual chromosomes remain difficult to disentangle. The pathogen *Microbotryum* fungi represents a good model to explore sex chromosome evolution as their haplo-diploïd cycle (mating compatibitlity is determined in the haploid stage) may represent a model of evolution simpler to investigate than typical diploid-matings (e.g. XX/XY) as they display no asymmetry in the sheltering of mutation, or in effective population size. This may simplify the test and interpretation of expected consequences of recombination suppression in mating type chromosome.

We reconstructed from low-depth sequencing data the evolutionary relationships between 12 species of *Microbotryum* specific to different host plants. We applied stringent filtering in order to reconstruct putative ortholog groups, build phylogenomic relationships between fungal species and ultimately assess effect and efficiency of purifying selection on coding sequences comparatively between mating type chromosomes and autosomes. Three other potential consequences of reduction of recombination rate on mating chromosome were additionally investigated, including content in transposable elements (TEs), degeneration of codon usage, and alteration of the negative relationship observed between the GC content in the first two codon positions (GC12) and the third one (GC3).

As theoretically expected, suppression of recombination on mating type chromosomes led to significantly increased rate of non-synonymous mutations relatively to synonymous mutations (dN/dS) when compared to autosomes. This significant increase in protein sequence evolutionary rate relatively to silent substitution is interpreted as a decreased efficiency of purifying selection. This observation remains true when controlling for GC3 content and codon bias. Additionally, two other footprints of suppressed recombination are observed on mating type chromosomes: they accumulated more TEs and displayed a weaker relationship between GC12 and GC3. Finally, no significant difference was observed in codon usage between mating type chromosomes and autosomes.

Comparative Transcriptomics (RNA-seq): Applications and Methods Based on the Yeast *Yarrowia lipolytica*

Hugo Devillers [1,2§], François Brunel [1,2], Caroline Proux [3], Claude Gaillardin [1,2], Jean-Y. Coppée [3], Cécile Neuvéglise [1,2]

[1] INRA, UMR 1319 MICALIS, F-78352 Jouy-en-Josas, France
[2] AgroParisTech, UMR 1319 MICALIS, F-78352 Jouy-en-Josas, France
[3] Département Génomes et Génétique, Institut Pasteur, Plate-forme Transcriptome et Epigénome, 25 Rue du Docteur Roux, F-75015 Paris, France
[§] Corresponding author and poster presenter hugo.devillers@jouy.inra.fr

The advent of next generation sequencing (NGS) now allows high-throughput and high-resolution transcriptome analyses. Whole transcriptome shotgun sequencing, also called RNA-seq, offers invaluable opportunities to better understand the functioning and the structure of genomes. In this work, we investigated RNA-seq data from *Yarrowia lipolytica* strain E150, an oleaginous yeast of prime interest in biotechnology, under 6 different culture conditions allowing variations in pH, temperature, oxidative stress and carbon source. We first showed how this data helped us to improve the structural annotation of *Y. lipolytica* genome. Indeed, RNA-seq can be used to identify new genes or non-coding transcripts, to study mRNA splicing and more especially alternative splicing, and to determine 5' and 3' untranslated regions (UTR). *Y. lipolytica* is known as a relatively intron-poor species, but here, about 40% more introns were identified with RNA-seq, especially in 5'UTR, leading to 20% the number of intron-containing genes. Additionally, 95 putative non-coding RNA were detected whose expression differs according to the culture conditions.

In a second step, based on these RNA-seq data, the pros and cons of existing methods for the comparison of gene expression level were discussed. More precisely, we focused on tools rooted on multiple testing such as DESeq and edgeR (Bioconductor version 2.12). A particular attempt was made to illustrate the constraints and the limits of these approaches as well as the statistical issues they raised. Thus, for instance, we showed a relatively high heterogeneity of the outputs provided by these different methods as well as a critical impact of the number and the quality of replicates on the results.

Acknowledgments

This work was supported by the French National Agency for Research (ANR), project YeastIntron 2010 BLAN 1620. Thanks are expressed to Marie-Agnès Dillies for helpful advices on statistics.

Automated Filtering of Multiple Sequence Alignment Worsens Phylogenetic Inference

Ge Tan[1], Matthieu Muffato[2*], Christian Ledergerber[1], Javier Herrero[3], Nick Goldman[2], Manuel Gil[1], Christophe Dessimoz[4,1,2§]

[1] ETH Zurich, Computer Science, Switzerland and Swiss Institute of Bioinformatics,
[2] European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, UK.
[3] The Genome Analysis Centre, Norwich Research Park, Norwich, NR4 7UH, UK
[4] University College London, Gower Street, London, WC1E 6BT, UK
[§] Corresponding author: c.dessimoz@ucl.ac.uk
MM is the poster presenter: muffato@ebi.ac.uk

Abstract:

In the context of phylogenetic analysis, a multiple sequence alignment (MSA) is the general way of comparing simultaneously several sequences. However, computing a MSA appears to be difficult because of errors in the initial sequences (due to sequencing, the assembly or the annotation) and the divergence of the sequences. As this could seriously impact the tree estimation, some automated algorithms are thus used to filter out the less-supported columns to increase the signal *vs* noise ratio.
Here, we test the assumption that filtering a MSA will increase its overall quality for the purpose of single gene phylogeny reconstruction, in the context of whole-genome analysis. We show that the trees obtained from filtered MSAs are on average worse than those obtained from unfiltered MSAs. We confirm that this result holds for a wide range of parameters and methods. It appears that filtering the MSAs can only significantly improve the reconstruction for optimized combinations of the filtering algorithm, its parameters, and the dataset, which require some prior information about the expected result, and make it unsuitable for general analysis.
Thus, contrary to widespread practice, we do not generally recommend the use of current alignment filtering methods for phylogenetic inference (except perhaps light filtering to speed computations). We hope that the methodology described in this study will guide the development of better filtering algorithms, by providing a common way of measuring the impact of filtering the alignments.

An Orthology Quality Measure and its Applications

Maribel Hernandez Rosales [1 §], Gabriel Moreno Hagelsieb [2], Sarah Berkemer [1], Peter F. Stadler [1]

[1] Department of Computer Science, Univ. Leipzig, Haertelstr. 16-18, D-04107 Leipzig, Germany
[2] Department of Biology, Wilfrid Laurier University, 75 University Ave. W. Waterloo, Ontario, Canada N2L 3C5
[§] Corresponding author and author presenter maribel@bioinf.uni-leipzig.de

Abstract

Here we present a graph-theory based analysis of the quality of predicted orthologous relationships. We have applied this method to sets of genes from the *Escherichia coli* pangenome (defined as the complete set of genes present within a set of genomes belonging to a particular taxa, such as a species). The set of genes requires the clustering of genes into groups wherein the genes in the group would be deemed as a "single" gene. Therefore, the genes in the same group have to be orthologs, homologs arising after speciation events. Orthologs are expected to have a higher probability of keeping their functions across organisms. However, orthology is not transitive. Duplication events after lineage separation (outparalogy), for example, complicates this simple definition, since it produces a paralogy down the lineage, but the set of paralogs remain co-orthologs to genes out of such lineage. Other events might complicate the picture both for natural reasons and for technical ones. However, it is expected that groups of orthologous genes within a pangenome defined at the species level would suffer less from problems arising from natural events and technical issues. In this work we represent genes as vertices of a graph and place an edge between two vertices if they are predicted orthologs. It has been proven that a graph representing orthology relationships must be a cograph if the orthology relationship is valid. A cograph is a graph where no induced subgraph in four vertices forms a path, called a P4. Forbidden subgraphs are induced subgraphs that contain more than one P4.

In this work we are interested in quantifying forbidden subgraphs in ortholog relationships and in random graphs, in such a way that, this quantification would give us evidence of how good an estimated orthology relationship is. In other words, we would like to differentiate between graphs that can be edited and converted to cographs (which in turn would give us a valid orthology) and graphs that look more like random graphs. We have simulated graphs that represent valid ortholog relationships, and therefore are cographs, then we have added noise to them. We have also simulated random graphs with same vertex degree as in our noisy cographs. We have then quantified the number of forbidden subgraphs found in each graph. As a result we found out that noisy graphs with more than 20% of noise converge to a certain range and therefore become very similar to those of random graphs, while graphs with less than 20% of noise contain many less forbidden subgraphs and therefore can be edited to be converted to cographs.

We therefore explored the clusters of genes representing the *Escherichia coli*

pangenome and measured their consistency. We found many clusters that are highly connected and contain no forbidden subgraphs or few of them, while only a few of them vary in terms of cleanliness. We suggest that the existence of many forbidden subgraphs in some clusters might be mostly due to horizontal gene transfer.

The Ancestrome project, an effort towards better diffusion of phylogenetic developments

Thomas Bigot [1*§], Rémi Planel [1,2*]

[1] Université de Lyon; Université Lyon 1; CNRS; UMR 5558, Laboratoire de Biométrie et Biologie Evolutive
[2] Pôle Rhône-Alpes de Bioinformatique
[*] Equal author contribution and both poster presenters
[§] Corresponding author thomas.bigot@univ-lyon1.fr

Introduction

All over the world, scientists work hard to generate data and develop sophisticate methods to analyze them. Yet, only few of them are used outside of their original lab. This is the case for many phylogenetic tools and databases. Many of them are ignored by the scientific community because they do not comply with minimal quality standards, i.e. portability of the code but also ways of diffusion.

The Ancestrome project develops many original algorithms for phylogeny, ancestral state reconstruction and comparative genomics and proposes original databases for phylogenetic analyses. A central concern of the project is the spread of our developments. We present some of the tools developed within the Ancestrome project as well as a set of good practices and reflexions for a better diffusion of bioinformatics tools.

Diffusion of Programs

In the scientific community, a lot of developments are made as FLOSS (Free/Libre/OpenSource Software) which is a nice starting point. But to make it compilable and installable on a lambda user machine, it must meet some extra requirements. In the Ancestrome project, we try to: i) use standard programming languages, ii) use cross-plaform compilation tools such as CMake. We also bring efforts to keep an up-to-date central repository, with bug tracking which allows us to be very reactive when user report some issues.
We think that designing good methods is unfortunately only half of the job, the other half being packaging tasks. We will describe the progress we have done with our lead softwares: Phyldog, ALE, DeCo, and Prunier.

Data visualization through web services

Visualization plays an important role when it comes to interpret and understand large and complex datasets, which is typically the case in comparative genomics. We humans are intensely visual creatures and a synthetic data visualization tool can allow the identification of patterns within complex data. It is also the fastest way to share it

with the community. It is hence necessary to develop schematic representations, which can be adapted to users needs.

The Ancestrome project is developing web applications for appealing visualization and interactive experience with genomic data in an evolutionary framework. We propose these tools for visualizing and querying our databases, such as HOGENOM and HOMOLENS.

References

Sèverine Bérard, Coralie Gallien, Bastien Boussau, Gergely J. Szöllősi, Vincent Daubin and Eric Tannier. Evolution of gene neighborhoods within reconciled phylogenies, Bioinformatics, 2013.

B Boussau, GJ Szollosi, L Duret, M Gouy, E Tannier, V Daubin. Genome-scale coestimation of species and gene trees. Genome research 23 (2), 323-330.

Szöllősi, G. J.; Boussau, B.; Abby, S. S.; Tannier, E. & Daubin, V. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations Proceedings of the National Academy of Sciences, National Acad Sciences, 2012, 109, 17513-17518

Abby, S. S.; Tannier, É.; Gouy, M. & Daubin, V. Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests BMC Bioinformatics, 2010, 11, 324

PhylDiag : identifying complex cases of conserved synteny that include tandem duplications

Joseph Lucas [1], Matthieu Muffato [1,2] and Hugues Roest Crollius [1§]

[1] IBENS Intitute of Biology of the Ecole Normale Superieure, Paris, 45 rue d'Ulm, France
[2] new address : EBI, Welcome Trust Genome Campus, Cambridge, BC10 1SD, Hinxton, United Kingdom
[§] Corresponding author : hrc@ens.fr
JL is the poster presenter jlucas@biologie.ens.fr

Abstract:
Many methods have been developed to identify synteny blocks, with varying degrees of sophistication to deal with specific cases while allowing some flexibility. Here, we define a synteny block from the point of view of the last common ancestor of two modern genomes. A synteny block is a successive sequence of ancestral genes where the order and orientation has not been modified during evolution, and is thus exactly the same in two modern descendants.
We developed an algorithm called PhylDiag to identify such conserved synteny blocks. The strengths of our algorithm is to use phylogenetic trees to identify and allow species-specific tandem duplications within the synteny block and to allow species-specific genic insertions or deletions. The algorithm includes a utility to visually represent synteny and segmental duplications can also be studied with our method by comparing a genome with itself.
PhylDiag takes as input two modern genomes and the corresponding forest of gene trees and returns accurate synteny blocks that can easily be merged if the distance that separates their extremities is less than a customizable gap. The user can choose many different distances among which the Diagonal Pseudo Distance (DPD) used in ADHoRe and diagHunter. Synteny blocks eventually undergo a statistical test to verify that they cannot be due to a random combination of genes. The statistical test takes into account the density in homologies, gene order and gene orientations. In this way, it is possible to identify real but small synteny blocks with insertions, deletions, incorrect annotations or even micro-rearrangements. PhylDiag has been compared to i-ADHoRe 3.0 to identify synteny blocks conserved between human and mouse. The performances are very similar in terms of coverage (~91% of genes are included in synteny blocks) and in terms of the distribution of synteny blocks lengths. However, PhylDiag accounts more rigorously for the orientation of genes duplicated in tandem, defines more rigorously the p-value reporting the significance of synteny blocks, defines an optimal inter-block gap size, resulting in both a smaller number of arbitrary user-defined parameters and a more predictable behaviour of the algorithm.

Half of the sox genes remained duplicated since the teleost specific whole genome duplication

Frédéric Brunet, Emilien Voldoire, Jean-Nicolas Volff , Delphine Galiana

Institut de Génomique Fonctionnelle de Lyon, Ecole Normale Supérieure de Lyon, 46, allée d'Italie, F-69364 LYON Cedex 07, France
Corresponding author and poster presenter: frederic.brunet@ens-lyon.fr

Abstract:
Two successive events of whole genome duplications (WGD) occurred at the base of the vertebrate lineage, coined 1R and 2R for Rounds of WGD. An additional third round of WGD (3R) occurred at the base of the teleostean fish. Sox genes encode a family of transcription factors that has experienced a phase of expansion leading to 20 sox genes well described in human and mouse. This gene expansion preceeded the vertebrates lineage and enhanced even more through both the 1R+2R WGDs and tandem duplication occurrences. In fish, additional sox genes have been described with orthologous relationship assessed by phylogenetical analyses. We were interested to know how have evolved such a group of genes since the 3R event. To this end, we carried out a bioinformatic analysis, searching exhaustively the public releases of fish genomes as well as other public databases. We combined both the phylogenetic information and synteny analyses to assess the evolutionary history of fish sox genes. We found evidence that in fish, 10 of the 20 mammalian orthologues of these sox gene family come from and remain duplicated since their 3R origination. This 50% ratio is way above the estimated global average of 12% of genes that remained duplicated since the 3R event, a value in favor of the idea that transcription factors have played a key role in the diversification of the teleost lineage. In addition, we performed expression analyses of these sox genes in three fishes (zebrafish, medaka and platyfish) and observed species specific expression for some of them, in agreement with this hypothesis.

An insight into the century timescale evolution of transposable element populations from the sequencing of the recent allotetraploid oilseed rape (*Brassica napus*)

Jérémy Just [1,4,§], Mathieu Charles [2], France Denoeud [3], Patrick Wincker [3], Boulos Chalhoub [1]

[1] Organization and Evolution of Plant Genomes, URGV – INRA 2, rue Gaston Crémieux – 91000 Évry – France
[2] EPGV – INRA 2, rue Gaston Crémieux – 91000 Évry – France
[3] Génoscope / CNS – IG, CEA 2, rue Gaston Crémieux – 91000 Évry – France
[4] RDP – INRA, CNRS ENS Lyon – 46, allée d'Italie – 69364 Lyon – France
§ Corresponding author: jeremy.just@ens-lyon.fr

Abstract:
Polyploidization and differential proliferation of transposable elements (TEs) have played a major role in the dynamics of plant genome evolution. Within the family of *Brassicaceae*, the genus *Brassica* includes species which genomes have been recurrently duplicated during their evolution, by frequent polyploidization events. Among these, oilseed rape (*B. napus*) has been formed through an allotetraploidization event between *B. rapa* (A genome) and *B. oleracea* (C genome), 5-10 centuries ago.
It is now established that TEs account for a large part of the divergence between the diploid A and C genomes, both by their distribution along chromosomes and the relative composition of their populations. And the fusion of those two diverging TE population into the *B. napus* genome is a rare opportunity to understand the diploidization process starting right after any polyploidization event, at a century timescale.
In the course of sequencing the genome of oilseed rape, we have put a special effort to annotate transposable elements in an unbiased way, to accurately compare TE populations in genomes A and C both from the tetraploid and the diploid species. We mostly relied on the identification of structural features and *de novo* prediction of repeats. We have structurally identified ~32% of TEs, while the remaining ~68% has been annotated by similarity against public databases.
As a whole, *B. napus* genome assembly contains 26.7% of TEs (39.3% in the unassembled reads), but more than half of the families are more abundant in one diploid genome than in the other, and more than 250 families exhibit a statistically significant difference in their representation between the addition of the diploid genomes and the tetraploid one. We have identified transpositions that could have occurred between A and C genomes, after the allopolyploidization event. Based on RNA-seq data analysis, we have also identified TE insertions which alter expression of homoeologous genes and could play a role in the on-going diploidization process.

Keywords : *Brassica*, oilseed rape, polyploidy, genome evolution, proliferation of transposable elements, comparative annotation

The combinatorics of Tandem Duplications

Luca Penso-Dolfin [1], Christopher Greenman [1 §]

[1] University of East Anglia, Norwich, UK; The Genome Analysis Center, Norwich, UK
[§] Corresponding author c.greenman@uea.ac.uk
LPD is the poster presenter l.penso-dolfin@uea.ac.uk

Abstract:
Tandem Duplications represent a common source of genomic variation, often crucial for a clear understanding of common phenotypic variation, genetic disorders, and the evolution of gene families.
We have developed an efficient algebraic representation of a Tandem Duplication (TD) evolution, representing chromosomes as row vectors (*words*) where each TD-like connection is uniquely represented by a number (*letter*). Based on such a representation, we have faced the following challenges:

i. *Describing the full space of TD evolutionary paths* in terms of number of words of size *m* after *k* TD events
ii. *Ordering TD events through a recursive deconstruction process of a word*. The total orders of events are obtained by recursively eliminating symmetries and unique letters from the word, going backward in time until all letters are eliminated
iii. *Using Hasse diagrams, describing the combinatorial properties of the TD space* in terms of
   a. Number of different breakpoint orders which are consistent with a specific word
   b. Number of different chronological orders of events which are consistent with a specific word

This study represents an in-depth description of both the geometric and combinatorial properties of Tandem Duplications. Moreover, the newly introduced algebraic representation fits particularly well to the interpretation of coupled Copy Number Variation and Paired-End Sequencing data, providing a tool to infer both the structure and the evolutionary history of a TD-rearranged chromosome.

Detecting positive selection using the full spectra of genetic variationfrom whole-genome sequencing data

Maud Fagny [1,2,3], Etienne Patin [1,2], David Enard [4], Luis B. Barreiro [5], Lluis Quintana-Murci [1,2,§] and Guillaume Laval [1,2,§]

[1] Institut Pasteur, Human Evolutionary Genetics, Department of Genomes and Genetics, F-75015 Paris, France
[2] Centre National de la Recherche Scientifique, URA3012, F-75015 Paris, France
[3] Univ. Pierre et Marie Curie, Cellule Pasteur UPMC, F-75015 Paris, France
[4] Department of Biology, Stanford University, Stanford, CA 94305-5020, USA
[5] Sainte-Justine Hospital Research Center, Department of Pediatrics, University of Montreal, Montreal, Quebec H3T 1C5, Canada
[§] Corresponding authors: quintana@pasteur.fr, glaval@pasteur.fr

Abstract:
Multiple genome-wide scans for selection have identified hundreds of regions of the human genome as being targeted by positive selection. However, only a small proportion of these regions have been replicated across studies, and the genuine prevalence of positive selection as a mechanism of adaptive change in humans remains controversial. Here we explore the power of two haplotype-based statistical methods – iHS and DIND – in the context of next-generation sequencing data, and evaluate their robustness to human demography and other selection modes. We show that these statistics are both powerful for the detection of recent positive selection, regardless of population history, and robust to variation in coverage, with DIND being insensitive to very low coverage. We apply these statistics to whole-genome sequence datasets from the 1000 Genomes Project (Pilot and Phase 1) and Complete Genomics. We found that putative targets of selection were highly significantly enriched in genic and non-synonymous SNPs and that DIND was more powerful than iHS in the context of small sample sizes, low-quality genotype calling or poor coverage. As we excluded genomic confounders and alternative selection models, such as background selection, the observed enrichment attests to the action of recent, strong positive selection. Our analyses identified not only well-known selection targets, but also new regions, for some of which we were able to determine the phenotypic directionality of selection events. Overall, our results indicate that hard sweeps targeting low-frequency standing variation have played a significant, albeit moderate, role in the last 60,000 years of human evolution.

Comparative And Functional Genomics Of The *Wolbachia* Bacteria: Towards The Insight Of Endosymbiosis

Sandrine Geniez [1,2 §], Bouziane Moumen[2], Jeremy Foster[1], Sanjay Kumar[1], Barton E. Slatko[1] and Pierre Grève[2]
[1] New England Biolabs, Inc. Ipswich MA 01938 USA
[2] University of Poitiers, UMR CNRS 7267 Ecologie et Biologie des Interactions, Equipe Ecologie, Evolution, Symbiose. 86022 Poitiers Cedex, France
[§] Corresponding author: sandrine.geniez@gmail.com

Abstract:
Cross-strain genome studies are based on orthologs, genes present in different species that have only evolved through speciation events. Orthologs are known to have the same biological functions whereas duplication allows development of new functions. Identification of orthologs is a powerful tool of understanding the genealogy of genes and to investigate the mechanism of evolutionary process. These groups of genes with the same biological function are also an inestimable pool of information for functional analyses such as pathway comparison across species.
Using Next Generation Sequencing technologies (Illumina HiSeq and MiSeq), we successfully sequenced 6 new strains of the bacterial endosymbiont *Wolbachia* that induce either feminization or cytoplasmic incompatibility in their arthropod hosts. These genomes as well as the 16 already published genomes were used for comparative genomics. On the basis of the orthologous database we generated using OrthoMCL, we divided these bacterial genomes into (1) the "core-genome", containing genes shared by all strains, (2) the "dispensable-genome" composed by genes present in two or more but not all strains and finally (3) genes unique to single strains, all these categories representing the "pan-genome". We investigated the composition of this pan-genome in order to look at the insight of the basis of *Wolbachia* symbiosis. Based on all the single-copy core-genes, a phylogenomy of the *Wolbachia* strains was inferred. To complement the database of putative bacterial effectors established from pan-genomic analyses, a conversed motif recognition approach using HMMER3 allow us to identify eukaryote-like proteins such as ankyrin repeats containing proteins that are known to be involved in host-symbiont interactions.

Functional gene groups are concentrated within chromosomes, among chromosomes and in the nuclear space of the human genome

Annelyse Thévenin [1 2 *], Liat Ein-Dor [3 *], Michal Ozery-Flato [3], Ron Shamir [2 §]

[1] Genome Informatics, Faculty of Technology and Institute for Bioinformatics, Center for Biotechnology (CeBiTec), Bielefeld University, Germany
[2] Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, 69978 Israel
[3] Machine Learning and Data Mining Group, IBM Haifa Research Lab, Mount Carmel, Haifa, 31905, Israel

[*] Equal author contribution
[§] Corresponding author rshamir@tau.ac.il
AT is the poster presenter atheven@cebitec.uni-bielefeld.de

Genomes undergo changes in organization as a result of gene duplications, chromosomal rearrangements and local mutations, among other mechanisms. In contrast to prokaryotes, in which genes of a common function are often organized in operons [1] and reside contiguously along the genome, most eukaryotes show much weaker clustering of genes by function, except for few concrete functional groups [2]. We set out to check systematically if there is a relation between gene function and their organization in the human genome. We tested this question for three types of functional groups: pairs of interacting proteins, complexes and pathways. We found a significant concentration of functional groups both in terms of their distance within the same chromosome and in terms of their dispersal over several chromosomes. Moreover, using Hi-C based contact map of the tendency of chromosomal segments to appear close in the 3D space of the nucleus [3], we show that members of all three types of functional groups that reside on distinct chromosomes tend to concentrate in space. Hence, the human genome shows substantial functional organization on all the tested levels, most likely due to co-regulation effect. Our statistical testing methodology, based on permutation, applies equally to all types of data.

[1] Malke H. J. H. Miller and W. S. Reznikoff (Editors), The Operon (2nd Edition). VII, 469 S., 128 Abb., 36 Tab. Cold Spring Harbor 1980. Cold Spring Harbor Laboratory. Zeitschrift für allgemeine Mikrobiologie. 1981;21(9):697-697.
[2] Blumenthal T, Evans D, Link CD, Guffanti A, Lawson D, Thierry-Mieg J, et al. A global analysis of Caenorhabditis elegans operons. Nature. 2002;417(6891):851-4.

[3] Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326(5950):289-93.

Consimilar Intervals

Daniel Dörr [1], Jens Stoye [1], Sebastian Böcker [2], and Katharina Jahn [1]

[1] Institute for Bioinformatics, Center for Biotechnology (CeBiTec), Bielefeld University
[2] Lehrstuhl für Bioinformatik, Friedrich-Schiller-Universität Jena, Germany

Poster presenter: Daniel Dörr <ddoerr@cebitec.uni-bielefeld.de>

Abstract:
Comparative analyses of chromosomal gene orders are successfully used to predict gene clusters in bacterial and fungal genomes. Present models for detecting sets of co-localized genes in chromosomal sequences require prior knowledge of gene family assignments between genes in the dataset of interest. These families are often computationally predicted on the basis of sequence similarity or higher order features of gene products. Erroneous gene family assignments emerging in this process are amplified in subsequent gene order analyses and thus may deteriorate gene cluster prediction.

In this work we introduce a gene family-free approach of gene cluster prediction based on a novel graph-theoretical concept called consimilar intervals. We further present an efficient algorithm to compute all consimilar intervals between two chromosomal sequences. Moreover, we introduce a ranking scheme which orders consimilar intervals by their degree of local preservation and similarities between the contained genes. Finally we compare our model and algorithm to common intervals-based methods on a dataset of 93 bacterial genomes.

Our method is able to detect gene clusters that would be also detected with well-established gene-family based approaches. Moreover, we show that it is able to detect conserved regions which are missed by gene family-based methods due to wrong or deficient gene family assignments.