ABSTRACTS

ALPHY/PhyloSIB 2014: Swiss-French meeting on Bioinformatics and Evolutionary Genomics

February 4-5 2014 – Geneva

Editing duplicated nodes in vertebrate gene trees

Amélie Peres, Hugues Roest Crollius

Gene phylogenies are essential for many biological evolutionary studies. However, phylogenetic reconstructions are difficult to model, especially when they include gene duplications. In this study, we have developed a method to improve the positions of duplications in gene trees produced by TreeBest, a widely used method at the core of the "Ensembl compara" pipeline. We first investigated methods to automatically identify incorrectly positioned duplications with two independent criterions. The first relies on a "confidence score", a measure introduced by TreeBest and comprised between 0 and 1 and assigned to each duplicated node. The score reflects the ratio between the number of species with a duplicated genes and the total number of species under this node. A well-supported duplication will thus have a score close to 1. The second criteria relies on the topology of the species tree. Each node in the tree is ancestral to two subtrees that descend from it. If a duplication is well supported, we expect both subtrees to contain duplicated copies dating from this duplication. This criterion can be adjusted depending on the number or the ratio of species that must contain a duplicated gene in each subtree. Once a node is considered to be porrly supported using either one or both criterion, we edit the node by replacing it by a speciation node, and testing the nodes just below using the same criterion. If the nodes pass the test, the duplication is created at this new position in the tree. In order to assess the quality of the new edited trees, we tested both criterion on all phylogenetic trees available in the Ensembl compara database (20114 gene trees in Ensembl 71) and then used the new gene tree databases and the initial Ensembl gene tree database to reconstruct ancestral genomes using AGORA. AGORA is an algorithm developed in our laboratory to reconstruct ancestral gene orders, and its performances are very sensitive to the quality of the input gene trees. In particular, the length of the reconstructed ancestral chromosomal regions varies substantially depending on the quality of the input gene trees. We find that using the confidence score criteria significantly improves the positions of duplications within gene trees compared to the initial Ensembl gene tree database. The optimal value is a threshold score of 0.3, at which 39% of the 197 894 duplication nodes are edited, resulting in a 200% increase in the N50 length of ancestral Boreoeutheria reconstruction. In contrast, the second criteria (topological criteria) degrades the quality of gene trees compared to the original Ensembl gene trees, either used alone or in

combination with the confidence score. We will discuss the implications of these finding with respect to the nature of poorly and well-supported duplication nodes.

Benchmarking based on SwissTree

Brigitte Boeckmann, Adrian Altenhoff, Ioannis Xenarios, the SwissTree Consortium and Christophe Dessimoz

SwissTree is a collection of high confidence gene trees for the assessment of the quality status of phylogenomic databases. It currently contains 14 gene trees, comprising 1'309 genes and 76'547 resolved gene relationships. At the Quest for Orthologs meeting 2013 (QfO3) at the University of Lausanne, first benchmarking studies based on SwissTree were presented. Results are discussed

Model estimation using mixture of topologies - Application to the detection of conversion events

Laurent Guéguen, Eugénie Pessia, Gabriel Marais

Usually, to infer a molecular evolution process from a family of homologous sequences, we put the hypothesis that this process has been guided along a unique tree topology on all the sites. However, some evolution events - such as recombination or conversion – may contradict this hypothesis. In this case the inference procedure should take into account this event, which means that the likelihood of the model should be computed site-specifically on the correct topologies. But even if the changing topology event is known – or supposed - the sites where it happened may be ignored.

We propose to compute the likelihood of evolution process on several trees on the same time, assigning on the set of candidate topologies a priori probabilities (either independent or in a markovian framework). And a posteriori probabilities describe on which sites each topology fits the data. Using the Bio++ specific syntax, this can be done with heterogeneous models as well as with site-specific models. This modeling can also be used in the case when the exact topology is unknown and several candidate topologies are possible.

Using this type of modeling, we develop a procedure to detect conversion events between X and Y chromosomes on primate genes.

The Evolution of Base Composition in Monocots Genomes

Yves Clément, Margaux Alisson-Fustier, Benoit Nabholz & Sylvain Glémin

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 a unimodal GC-content distribution.

Until recently, because of a lack of genomic sequence, the origin of the peculiar GC-content distribution in grasses remained unknown. Indeed, only grass genomes were available inside monocotyledons while the phylogenetic sampling outside this group was lacking. The recent publication 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 globally evolving very slowly and that that the gradient of GC-content observed along coding sequences is also conserved in grass plants. 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.

Long noncoding RNAs in mammalian development and evolution

Fabrice Darbellay and Anamaria Necsulea

It is now well established that mammalian genomes encode tens of thousands of long noncoding RNA (lncRNA) genes. However, only a small fraction of these genes are well characterized. To provide insights into lncRNA functionality, we have recently undertaken a large-scale evolutionary analysis of lncRNA repertoires and expression patterns, in eleven tetrapod species (Necsulea et al., 2014). In that study, we had identified approximately 2,500 highly conserved lncRNAs, including approximately 400 genes that likely originated more than 300 million years ago. Importantly, we had observed that the promoters of these highly-conserved lncRNAs were enriched in binding sites for homeobox transcription factors (which are key develpemental regulators), suggesting that ancient lncRNAs may function predominantly during embryonic development rather than in adult organs. Here, we test this hypothesis by analyzing lncRNA expression patterns during embryonic development and aging (including 2 embryonic stages, newborn, adult and aged individuals), in two model organisms (mouse and rat), across four major organs (brain, kidney, liver and testis). We show that lncRNAs are indeed more often subject to temporal regulation of gene expression than protein-coding genes. In addition, we found that developmental regulation is more often associated with ancient lncRNAs than with young lncRNAs, and that in somatic tissues ancient lncRNAs tend to have higher expression levels during embryonic development than during adulthood and aging, thus confirming our hypothesis. However, in the testes, both young and lncRNAs are preferentially expressed in the adult state, and are inactive before the onset of spermatogenesis. Ongoing analyses are focusing on further dissecting the relationships between lncRNA expression patterns and their evolutionary origin.

Patterns of positive selection in seven ant genomes

Julien Roux, Eyal Privman, Sebastien Moretti, Josephine T. Daub, Marc Robinson-Rechavi, Laurent Keller

The evolution of ant species is marked by remarkable adaptations that allowed the development of very complex social systems. To identify how ant-specific adaptations are associated with specific patterns of molecular evolution we searched for signs of positive selection on amino-acid changes in proteins during the evolution of the ant lineage. We identified 24 functional categories of genes which were enriched for positively selected genes in the ant lineage. We also reanalyzed genome-wide dataset in bees and flies with the same methodology to check if genes under positive selection in ants were also under positive selection in the other analyzed lineages. Notably, genes implicated in immunity were enriched for positively selected genes in the three lineages, ruling out the hypothesis that the evolution of hygienic behaviors in social insects caused a major relaxation of selective pressure on this set of genes. Our scan also indicated that genes implicated in neurogenesis and olfaction started to undergo increased positive selection before the evolution of sociality in Hymenoptera, although it is assumed that the main challenges of the olfactory and neural systems in this lineage occurred with the evolution of social living. Finally, the comparison between these three lineages allowed us to pinpoint molecular evolution patterns that were specific to the ant lineage. In particular, there was relaxed selective pressure for genes related to metabolism in ants but not in bees and flies, possibly reflecting the loss of flight in ant workers. By contrast, there was recurrent positive selection on genes with mitochondrial functions specifically in ants, suggesting that the activity of mitochondria was improved during ant evolution. This might have been an important step toward the evolution of extreme lifespan that is a hallmark of this lineage.

Can we predict genetic diversity from species biology ? An answer from 414 animal transcriptomes.

Jonathan Romiguier

The population genomic literature has been so far dominated by a handful of model organisms in which the available genomic resources were concentrated. Next-generation sequencing (NGS) in principle offers the opportunity to investigate the molecular diversity of non-model species genome-wide in absence of any prior knowledge. The PopPhyl project takes a comparative approach to population genomics across animals. It aims at investigating the relationship between species biology and genome variation patterns based on RNA-seq data in >30 metazoan species (10 individuals each) and their outgroups, and de novo SNP and genotype calling. Here, we present a comprehensive view of the genome-wide genetic diversity of metazoan species. Surprinsingly, we found low levels of genetic diversity in both vertebrate and invertebrate, and show that life-history strategy is the major driver of genetic diversity disparities among animals.

Turtle population phylogenomics.

E. Loire, Y. Chiari, J. Romiguier, B. Nabholz, J. Lourenco, N. Galtier

Reptiles are a fascinating group of animals in which virtually no genomic data have been available until recently. We analysed the blood transcriptome of 27 individuals from 12 species of reptiles, with a specific focus on turtles. Based on this data set, we clarified the phylogenetic position of turtles within amniotes, calibrated the turtle molecular clock, estimated divergence times, and identified environmental determinants of the molecular substitution rate. Then we investigated the population genomics of Chelonoidis nigra, the giant Galapagos tortoise, an endangered species endemic to the Galapagos archipelago. We report an extremely low level of nuclear genetic diversity and an elevated mutation load in this insular species. We identify functional categories of genes undergoing an increased selective pressure in C. nigra, presumably as a response to its stressful environment. Finally, we do not detect any evidence for population structure in the giant Galapagos tortoise. These results have important implications in terms of species conservation and management.

Comparative study of the Ophioderma longicauda species complex: divergent patterns of connectivity and genetic diversity between brooding and broadcasting lineages

Alexandra Weber, Aurélien Bernard, Nicolas Galtier, Anne Chenuil

Closely related sympatric species with divergent life history traits represent excellent model species to infer the role of life history traits in connectivity because they display the same ecology, thus the differences in connectivity should only rely on the dispersal characteristics. The brittle star species complex Ophioderma longicauda is composed of three brooding (L2-L3-L4) and three broadcast spawning (L1-L5-L6) lineages. To infer connectivity between and within all six lineages, we analyzed 829 individuals for the COI marker. The lineage L1 displayed low genetic structure and high genetic diversity. In contrast, the genetic structure observed for the brooding lineages was high, the haplotype networks displaying high correlation with geography, and the genetic diversity was low. In addition, we sequenced the transcriptome of L1 and L3 individuals. We showed that the mean piN/piS ratios were higher in the brooding lineage and that the mean heterozygosity was higher in the broadcasting lineage. Our data showed strong differences in genetic structure and genetic diversity between closely related species even at the transcriptomic level, pointing out the strong influence of life history traits on connectivity. Further research is ongoing to compare the transcriptome characteristics of L1 and L3, and detect the different patterns of selection occurring in the brooding and broadcasting lineages.

Crossing the species barrier: genomic islands of introgression between two extremely divergent Ciona intestinalis species

Roux Camille, Tsagkogeorga Georgia, Bierne Nicolas, Galtier Nicolas

Inferring a realistic demographic model from genetic data is an important challenge to gain insights into the historical events during the speciation process and to detect the molecular signature of selection along genomes. Recent advances in divergence population genetics have reported that speciation in face of gene flow occurred more frequently than theoretically expected, but the approaches used rarely account for genome wide heterogeneity (GWH) in introgression rates. However, GWH is expected as a consequence of variation in effects of natural selection on migrant alleles. We investigated the impact of GWH on the inference of divergence with gene flow between two cryptic species of the marine model Ciona intestinalis. These morphologically indistinguishable entities are highly diverged molecular-wise, but evidence of hybridisation has been reported in both laboratory and field studies. We examined polymorphism and divergence patterns across 852 genes scattered throughout the C. intestinalis genome. We compared various speciation models and statistically tested for GWH under the ABC framework. Our results demonstrate the presence of significant extents of gene flow resulting from a recent secondary contact between the two gene pools, after more than 3My of divergence in isolation. The inferred rates of introgression are relatively low, highly variable across loci and mostly unidirectional, which is consistent with the idea that numerous genetic incompatibilities have accumulated over time throughout the genomes of these highly-diverged species and that introgression could be adaptive. A genomic map of the level of gene flow identified two islands of introgression, i.e. large genome regions of unidirectional introgression. This study clarifies the history and degree of isolation of two cryptic and partially sympatric model species, and provides a methodological framework for the study of GWH in introgression rates at various stages of the speciation process.

A Genomicus tutorial: how to explore and study genomes using comparative genomics and phylogeny.

Alexandra Louis, Nga Thi Thuy Nguyen, Matthieu Muffato, Hugues Roest Crollius

Comparative genomics combined with phylogenetic reconstructions are powerful approaches to study the evolution of genes and genomes. However the current rapid expansion of the volume of genomic information makes it increasingly difficult to interrogate, integrate and synthetize comparative genome data while taking into account the maximum breadth of information available. Genomicus (http://www.genomicus.biologie.ens.fr/genomicus) is a database and a web server that addresses this issue by allowing users to explore genomes in an intuitive way, across the broadest evolutionary scales. Four main interfaces ("views") are available: (i) PhyloView combines phylogenetic trees with comparisons of genomic loci across any number of genomes (ii) AlignView aligns of locus of interest against all other genomes to visualise its topological conservation (iii) MatrixView compares two genomes in a classical dotplot representation (iv) Karyoview visualise chromosome karyotypes "painted" with colours of another genome of interest. All four views are interconnected and benefit from many customizable features. More than 150 eukaryote genomes can be analysed, broken down in five groups following Ensembl and Ensembl Genomes: Vertebrates, Plants, Metazoa, Fungi, Protists. For the vertebrate group, ancestral gene order provides a long-term chronological view of gene order evolution. Human conserved non-coding

element (CNEs) information, and their orthologs across vertebrate phylogenies, are also available in vertebrates. In this short tutorial we will present both classic and new functionalities of the server that enable evolutionary studies from ancestral genomes to extant species. 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.

MetaPIGA 3, current features and perspectives

Dorde Grbic & Michel C. Milinkovitch

MetaPIGA, a robust implementation of several stochastic heuristics for large phylogeny inference under maximum likelihood, has reached its third version with several new and exciting functionalities. The program can now be run as a client for XtremWeb-CH grid users, a feature that can provide >100 fold speed improvement for very large dataset analyses. MetaPIGA3 also supports analyses over multiple processing cores as well as with CUDA-compatible Nvidia graphics cards (GPU), a feature particularly useful for datasets run with protein or codon models. A user-friendly and intuitive graphical interface allows for automated selection of best-fit model and performing complex analyses (e.g, with data partitioning). The program can also be run with batch files directly on a local computer or on distant servers using the console mode. MetaPIGA3 provides tools for tree visualization and manipulation as well as ancestral state reconstruction based on Bayesian statistics. MetaPIGA 3 runs on Linux, Windows, and MacOSX, and is freely available from http://www.metapiga.org. In the last part of the talk I will discuss additional functionalities that we are currently implementing.

Toward a better insight into transcriptome changes in complex developing organs.

Marie Sémon, Manon Peltier, Coraline Petit, Vincent Laudet, Anne Lambert, Sophie Pantalacci

Transcriptomes are now extensively studied to understand adaptive evolution in many complex traits, as a key connection between the genome and the phenotype. Comparative transcriptomics in conjunction with models of gene expression evolution -- although less sophisticated than those used to understand changes in coding sequences-- have permitted to scan for loci that have been subjected to adaptive changes in expression. But meaningful expression changes for understanding the evolution of complex phenotypes are changes that occur in developing organs. Transcriptomes of developing organs are particularly difficult to interpret and model because i) they vary quickly and ii) estimated gene expression levels are a mix of cellular expression levels intermingled with spatial patterns. So we think that modelling the evolution of expression in such datasets will remain elusive while we do not apprehend the main

sources of transcriptome differences between stages, between organs, and between species. To do the spadework in preparation for this modelling, we gathered and analysed transcriptome timeserie data in a developing organ, the upper and lower molar tooth germ in mouse. The molar germ is a good example of a complex organ, where the three main cell types (mesenchyme, epithelium and enamel knot) are well identified and quantifiable. Upper and lower molars are also serial organs, ie similar organs produced at different locations in the body, making them a good model for studying differences of developmental trajectories (including heterochronies) within a species. Transcriptomes of developing molars capture many sources of variation. In particular, the comparison of small-scale and large-scale temporal variations shows that there is no such thing as replicates for developmental transcriptomes, because developmental differences are extremely quick and pervasive. On the positive side, even small-scale differences are meaningful and document morphogenesis. In line with these observations, in a blind multivariate analysis, transcriptomes are ordered in accordance with developmental stage. This record of development timing, including heterochronies between upper and lower molar, is carried by a steady increase of enamel knot cells, corresponding to the progress of differentiation. The other sources of variation also reflect changes in cellular composition (changes in the proportion of the two other cell populations, and colonisation by neuronal and blood cells). Whole genome differences between tooth transcriptomes therefore largely reflect patterns --cell proportions-- more than cellular levels. With this in mind, we reanalysed published data on many organs and we propose that this is a general effect that could also be a major signal in interspecific comparisons. The contribution of cell proportions in transcriptome differences and species divergences is particularly important because it makes it necessary to redefine a null hypothesis in expression evolution.

Comprehensive metagenomic analysis of glioblastoma reveals absence of known virus despite antiviral-like type I interferon gene response

Érika Cosset, Tom J. Petty, Valérie Dutoit, Samuel Cordey, Ismael Padioleau, Patricia Otten-Hernandez, Laurent Farinelli, Laurent Kaiser, Pascale Bruyère-Cerdan, Diderik Tirefort, Soraya Amar El-Dusouqui, Zeynab Nayernia, Karl-Heinz Krause, Evgeny M. Zdobnov, Pierre-Yves Dietrich, Emmanuel Rigal, Olivier Preynat-Seauve

Glioblastoma is a deadly malignant brain tumor, and one of the most incurable forms of cancer in need of new therapeutic targets. As some cancers are known to be caused by a virus, the discovery of viruses could open the possibility to treat, and perhaps prevent, such a disease. Although an association with viruses such as cytomegalovirus or Simian virus 40 has been strongly suggested, involvement of these and other viruses in the initiation and/or propagation of glioblastoma remains vague, controversial, and warrants elucidation. To exhaustively address the association of virus and glioblastoma, we developed and validated a robust metagenomic approach to analyze patient biopsies via high-throughput sequencing, a sensitive tool for virus screening. In addition to traditional clinical diagnostics, glioblastoma biopsies were deep-sequenced and analyzed via a multi-stage computational pipeline to identify known or potentially discover unknown viruses. In contrast to the studies reporting the presence of viral

signatures in glioblastoma, no common or recurring active viruses were detected, despite finding an anti-viral-like type I interferon response in some specimens. Our findings highlight a discrete and non-specific viral signature and uncharacterized short RNA sequences in glioblastoma. This study provides new insights into glioblastoma pathogenesis and defines a general methodology that can be used for high-resolution virus screening and discovery in human cancers.

A quality control system for profiles obtained by ChIP sequencing

Marco-Antonio Mendoza-Parra, Wouter Van Gool, Mohamed Ashick Mohamed Saleem, Danilo Guillermo Ceschin and Hinrich Gronemeyer

The absence of a quality control (QC) system is a major weakness for the comparative analysis of genome-wide profiles generated by next-generation sequencing (NGS). This concerns particularly genome binding/occupancy profiling assays like chromatin immunoprecipitation (ChIP-seq) but also related enrichment-based studies like methylated DNA immunoprecipitation/methylated DNA binding domain sequencing, global run on sequencing or RNA-seq. Importantly, QC assessment may significantly improve multidimensional comparisons that have great promise for extracting information from combinatorial analyses of the global profiles established for chromatin modifications, the bindings of epigenetic and chromatin-modifying enzymes/machineries, RNA polymerases and transcription factors and total, nascent or ribosome-bound RNAs. Here we present an approach that associates global and local QC indicators to ChIP-seq data sets as well as to a variety of enrichment-based studies by NGS. This QC system was used to certify >5600 publicly available data sets, hosted in a database for data mining and comparative QC analyses.

Validation of Pooled Whole-Genome re-Sequencing by Individual-Based Genotyping By Sequencing

M. Fracassetti, P. Griffin and Y. Willi Validation of Pooled Whole-Genome re-Sequencing by Individual-Based Genotyping By Sequencing

M. Fracassettia, P. Griffinb and Y. Willia a Evolutionary Botany, Institute of Biology, University of Neuchâtel, Neuchâtel, Switzerland. b Department of Genetics, University of Melbourne, Parkville, Australia.

Sequencing pooled DNA of multiple individuals of a population instead of individual-specific sequencing has become a very popular method, due to cost-effectiveness and simple wet-lab protocol. The goal of this study was to validate SNP frequencies obtained with Pooled Whole-Genome re-Sequencing (PWGS) with those obtained by individual-based Genotyping By Sequencing (GBS). With PWGS, DNA of all individuals of one population are pooled and sequenced together. Population SNP frequencies are estimated from whole-genome sequence alignments for which coverage is sufficient. With GBS, DNA is digested with a restriction enzyme to obtain a reduced representation

of the genome. Then DNA is sequenced by use of individual-specific labels and SNP are called for each individual. To compare the two methods, we prepared libraries with PWGS and GBS using the individuals of one population of Arabidopsis lyrata. Libraries were sequenced with Illumina Hiseq, 100bp paired-end. Sequences were mapped against the A. lyrata genome v1.0 (Hu et al. 2011 Nat Genet) with BWA-MEM (Li 2013 arXiv). SNP frequencies of GBS reads were calculated with the two software pipelines GATK (McKenna et al. 2010 Genome Res) and Stacks (Stacks 2013 Mol ecol). For comparison, SNP frequencies of PWGS reads were calculated with GATK and PoPoolation (Kofler et al. 2011 PLoS ONE); SNP calling was tested with different minimum read-depth thresholds, to choose a suitable threshold to get reliable frequency data. Comparison of SNP frequencies showed that PWGS is a valid method for acquiring population-level SNP frequency data. Therefore, we will use the method in a Genome-Wide Association Study (GWAS) involving 40 populations of the whole North American distribution of A. lyrata. SNP variation will be investigated for association with climatic variables to investigate the genetic basis of climate adaptation.

The importance of newly sequenced genomes and functional annotations for phylogenetic profiling

Nives Skunca and Christophe Dessimoz

Phylogenetic profiling methods use patterns of presence and absence of genes in different species to predict protein-protein interactions and functional annotations. Since their introduction by Pellegrini et al. in 1999 [1], numerous methodological refinements have been proposed [2]. But a much greater difference lies in the amount of available genomic and functional data. In my talk, I will explore the extent to which new data improves the performance of phylogenetic profiling. Using a state-of-the-art phylogenetic profiling method [3], we quantified the improvement in prediction accuracy afforded by additional sequence and function information. Firstly, I will discuss an impressive difference in performance between phylogenetic profiles that use only the data available in 2005 and phylogenetic profiles that use the most recently available data. Further, I will discuss the difference in performance when having more organisms in phylogenetic profiles, compared to having more comprehensive functional annotations. I will briefly reflect on the difference in the performance of phylogenetic profiling in the three kingdoms of life. Finally, I will discuss one avenue of reducing the computational costs related to phylogenetic profiling: a careful selection of organisms that provides similar performance as when using the full set of sequenced organisms.

References

1. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96: 4285-4288.

2. Kensche PR, van Noort V, Dutilh BE, Huynen MA (2008) Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. J R Soc Interface 5: 151-170. doi:10.1098/rsif.2007.1047.

3. Skunca N, Bosnjak M, Krisko A, Panov P, Dzeroski S, et al. (2013) Phyletic profiling with cliques of orthologs is enhanced by signatures of paralogy relationships. PLoS Comput Biol 9: e1002852. doi:10.1371/journal.pcbi.1002852.

Phylogeny of the Hessian Fly, Mayetiola destructor

Panagiotis Ioannidis, Robert M Waterhouse, Jeffrey Stuart, Stephen Richards and The Hessian Fly Genome Consortium

The great Dipteran diversity is traditionally partitioned into two principal suborders: the Nematocera comprise mosquito-like flies with long antennae, while the Brachycera contain stout and fast-moving flies with short antennae. The Hessian fly, Mayetiola destructor, along with march flies, gnats and other midges make up the Nematoceran infraorder, Bibionomorpha. The placement of the Bibionomorpha as a sister group to Brachycera rather than with mosquitoes makes the Nematocera a paraphyletic group within the Dipteran phylogeny. The sequencing of the Hessian fly genome now provides an opportunity to investigate these relationships using whole-genome data. Phylogenies built from the concatenated alignments of single-copy orthologs are widely used to resolve uncertain branches of a phylogenetic tree; however, the incongruence between the individual gene trees is usually masked in the concatenation tree. Filtering of the input data to select genes with the strongest phylogenetic signal therefore reduces incongruences and provides a more confident resolution. Specifically, filtering individual gene trees using bootstrap support values and a recently described measure of "tree certainty", reliably placed of the Hessian Fly as basal to Brachycera.

Evolution of the mitochondrial genomes of vampire bats.

F. Botero-Castro, M. Tilak, F. Justy, F. Catzeflis, F. Delsuc, E.J.P Douzery

Vampire bats are the only group of mammals feeding exclusively on blood. This diet shift implies strong functional constraints needing of particular physiological modifications, namely to deal with the volume of ingested liquid and the high levels of potentially toxic iron in blood. This physiological changes are energetically costly and can thus an impact on the mitochondrial activity in order to assure energy supply. We explore, in a comparative phylogentic contex, some aspects of the nucleotide and amino acid composition of mitochondrially encoded genes that seem to be impacted by the shift to hematophagy.

Estimating past life-history traits in mammals : the contribution of mitochondrial DNA

Emeric Figuet, Nicolas Galtier, Jonathan Romiguier, Julien Dutheil

Reconstructing the ancestral characteristics of species is a major goal in evolutionary and comparative biology. Unfortunately, fossils are not always available or sufficiently informative, and phylogenetic methods based on models of character evolution can be unsatisfactory. Genomic data offer a new opportunity to estimate ancestral character states, through: (i) the correlation between DNA evolutionary processes and species life-history traits, and (ii) available reliable methods for ancestral sequence inference. We assess the relevance of mitochondrial DNA - the most popular molecular marker in animals - as a predictor of ancestral life-history traits in three mammalian orders, namely Cetartiodactyla, Carnivora and Primates. Using the complete set of 13 mitochondrial protein-coding genes, we show that the lineage-specific nonsynonymous over synonymous substitution rate ratio (dN/dS) is closely correlated with the species body mass, longevity and age of sexual maturity in Cetartiodactyla, and can be used as a marker of ancestral traits provided that the noise introduced by short branches is appropriately dealt with.

PhylDiag: identifying complex synteny blocks that include tandem duplications using phylogenetic gene trees

Joseph Lucas, Matthieu Muffato and Hugues Roest Crollius

Extant genomes share regions where genes have the same order and orientation. Such regions are often called synteny blocks and are conventionally thought to arise from the same ancestral block of genes. Precisely identifying these ancestral blocks is a prerequisite to better understand the evolutionary history of genomes. Here we describe PhylDiag, a software that finds statistically significant synteny blocks in pairwise comparisons of eukaryote genomes. Contrary to previous methods, PhylDiag uses gene trees to define gene homologies, thus allowing gene deletions to be considered as events that may break the synteny. PhylDiag also accounts for genes orientations, blocks of tandem duplicates and lineage specific gene apparitions. Starting from two genomes and the corresponding gene trees PhylDiag returns synteny blocks with gaps lower or equal to the maximum gap parameter gapmax. This parameter is theoretically estimated, and together with a utility to graphically display results, contributes to making PhylDiag a user friendly method. In addition, synteny blocks are subject to a statistical test to verify that they cannot be due to a random combination of genes. PhylDiag correctly identifies small synteny blocks even with insertions, deletions, incorrect annotations or micro-inversions. In this study we also benchmark several metrics to measure the distance in a matrix of homologies and we compare PhylDiag to ADHoRe on real and simulated data.

Towards synteny-aware gene tree reconstruction

Magali Semeria, Laurent Guéguen, Eric Tannier

Phylogenetic reconstruction methods are traditionally based on models of nucleotide or amino-acid sequence evolution. But the evolution of genomes can be studied at different scales: the gene level, accounting for gains and losses, and the genome level, accounting

for rearrangements of chromosome organization. Integrative methods have been developed that reconstruct species and gene histories simultaneously and take into account sequence evolution, gene birth, duplication, transfer and loss [1]. Because they account for gene level and sequence level signal, these methods provide far more reliable species trees and gene trees for the inference of evolutionary histories. Using such reconciled gene trees, and available software, it is possible to model the evolution of adjacent genes and to reconstruct ancestral synteny [2]. We show that, by a backward loop, ancestral synteny can be used as a control for reconciled gene trees. We detect inconsistencies in the ancestral syntenies reconstructed with [2] and propose an algorithm that modifies gene trees to solve these inconsistencies [3]. We show that, in the majority of cases, adding the gene-order information to guide the construction of gene trees does not lower the statistical support for these trees. We argue that our method is a first step towards the integration of synteny evolution in gene tree construction methods.

[1] B. Boussau, G. J. Szöllosi, L. Duret, M. Gouy, E. Tannier, and V. Daubin, "Genome-scale coestimation of species and gene trees.," Genome Res., vol. 23, no. 2, pp. 323-30, Feb. 2013.

[2] S. Bérard, C. Gallien, B. Boussau, G. J. Szöllesi, V. Daubin, and E. Tannier, "Evolution of gene neighborhoods within reconciled phylogenies.," Bioinformatics, vol. 28, no. 18, pp. i382-i388, Sep. 2012.

[3] C. Chauve, N. El-Mabrouk, L. Guéguen, M. Semeria, and E. Tannier, "Duplication, Rearrangement and Reconciliation: A Follow-Up 13 Years Later," in in Models and Algorithms for Genome Evolution, C. Chauve, N. El-Mabrouk, and E. Tannier, Eds. Springer, 2013, pp. 47-62.

Evaluating Synteny for Improved Comparative Studies

Cristina G. Ghiurcuta, Bernard M.E. Moret

Comparative genomics aims to understand the structure and function of genomes by translating knowledge gained about some genomes to other genomes that constitute the object of study. Early approaches used pairwise comparisons, but today researchers are attempting to leverage the larger potential of multiway comparisons. Comparative genomics relies on the structuring of genomes into syntenic blocks: blocks of sequence that exhibit conserved features across the genomes. Syntenic blocks are required for complex computations to scale to the billions of nucleotides present in many genomes; they enable comparisons across broad ranges of genomes because they filter out much of the individual variability; they highlight candidate regions for in-depth studies; and they facilitate whole-genome comparisons through visualization tools. However, the concept of syntenic block remains loosely defined. Tools for the identification of syntenic blocks yield quite different results, thereby preventing a systematic assessment of the next steps in an analysis. Current tools do not include measurable quality objectives and thus cannot be benchmarked against themselves. Comparisons among tools have also been neglected - what few results are given use superficial measures unrelated to quality or consistency. As a first step towards tacking the issues of syntenic block usage in comparative studies, we propose a theoretical model as well as an

experimental basis for comparing syntenic blocks. Our formal approach introduces an evolutionary principled quality criterion; it provides a solid basis for improving the construction of syntenic blocks and the design principles for developing tools for syntenic block mining. We apply the model and the measures to syntenic blocks produced by 3 different contemporary tools (DRIMM-Synteny, i-ADHoRe and Cyntenator) on a dataset of 8 yeast genomes. Our findings highlight the need for a well founded, systematic approach to the decomposition of genomes into syntenic blocks, while our experiments demonstrate widely divergent results among these tools, throwing into question the robustness of the basic approach in comparative genomics.