ECCB 2014 Accepted Posters with Abstracts

F: Evolution and population genomics

F01: Xiaoyu Yu and Oleg Reva. Mathematical modeling of the genetic amelioration of horizontally transferred genomic islands in bacterial genomes

Abstract: Horizontal gene transfer (HGT) within bacteria studies has dated back several decades and has been well documented. Current studies are still undergoing to dwell deeper into its effects within phylogeny and evolution alongside improvement in new technology and techniques. Although HGT is well understood, amelioration the process where the base DNA composition of the transferred genes from a donor undergoes nucleotide substitutions over time and reflects similarly in DNA composition to the recipient genome, is understudied. The study of amelioration could enhance the understanding of many aspects such as the mutation process of transferred material (preference in composition mutation, directional mutation, mutation rate), and the effects of base composition of recipient on the amelioration process. Hence a logical yet practical mathematical model is needed to model amelioration.

As foreign inserts, 4 known genomic islands (GI) were used to model their amelioration process towards compositional profiles of genomes of organisms representing distant taxa and different GC content, i.e. Bacillus subtilis 168, Pseudomonas aeruginosa PA01, Escherichia coli K12, Xylellafastidiosa 9a5c and Streptomyces griseus NBRC 13350. Compositional methods were used on each combination of GI (tester) and recipient (target) whereby k-mer words (dimers, trimers, tetramers) were calculated and ranked based by their frequencies in descending order of oligonucleotide usage (OU). A logistic probability function is then used to convert the ranked frequencies into a probability which gives the likelihood that at any given position the nucleotide will be substituted into another. A program on Python to simulate the amelioration process of generations with a given mutation rate was designed and in turn simulates the amelioration process for the underlined GIs. An amelioration model was then derived and fitted to the standard Verhulst model, which in general used in population dynamics. The parameter within the model is well suited for the simulated data and represents a good fit for the sample simulations. The program predicts a graduate merging of the insert's OU profile with those of the host genomes that would stabilize at some level of pattern similarity. The dynamics of this process and the level of stabilization depend on the rate of mutations in the target organism as well as the composition of the target sequence. The developed algorithm is suitable for estimating of the time lapsed after GI acquisition by a bacterium and also for estimating of mutation rates in different organisms and genomic loci.

F02: Kay Prüfer, Janet Kelso and Svante Pääbo. Searching for Regions of Superarchaic Introgression in the Denisovan Genome

Abstract: Before the arrival of modern humans from Africa, Europe and Asia was inhabited by at least two distinct groups of archaic humans: Neandertals and Denisovans. High quality sequences of the genomes of these archaic humans have been generated recently and led to the discovery of a signal of admixture from a deeply divergent, unknown hominin in the Denisovan genome. However, the analysis did not yield candidate
regions in Denisova that may trace their ancestry to that unknown hominin.

We use a hidden markov model to predict regions where either Neandertal or Denisova fall outside of the variation of modern humans (external region). We observe an excess of external regions in Denisova compared to the Neandertal individual and a trend towards longer external regions in Denisova as compared to Neandertals. We observe that Denisova regions larger than 10,000 basepairs give a significantly increased divergence compared with Neandertal external regions. This observation is compatible with introgression of a deeply divergent hominin into Denisova and suggests that introgressed regions may be among the longer Denisova external regions.

F03: Corinne Rancurel, Martine Da Rocha and Etienne G J Danchin. Alienness : Rapid detection of horizontal gene transfers in metazoan genomes

Abstract: Horizontal gene transfer (HGT) is the transmission of genetic material between species by ways other than direct (vertical) inheritance from parents to the offspring. HGT is recognized as a major evolutionary force in prokaryotes as it is involved in acquisition of antibiotic resistance or pathogeny. HGT has long been overlooked and considered insignificant in eukaryotes. However, HGT have also played important roles in the evolutionary history and biology of these species, including animals. For example, HGT have contributed to the colonization of land by plants, in the emergence of plant parasitism in nematodes or in the development of capabilities like resistance to extreme temperatures or desiccation. Progress in genome sequencing technologies has allowed multiple animal genomes to be publicly released. Systematic searches for HGT events in the root-knot nematodes or in the bdeloid rotifer, have shown that between 3 and 9 % of protein-coding genes in these species were of foreign origin (1). However, in the absence of a user-friendly, rapid and publicly available tool to detect HGT events in metazoan genomes, we still lack a global view of the prevalence of HGT in animals.

Here, we propose a tool that allows automatic detection of putative HGT events, based on the predicted protein set from a genome or transcriptome. Our tool has been specifically designed to rapidly detect genes of non-metazoan origin (e.g. bacterial, fungal) in metazoan genomes (e.g. insects, mammals, nematodes).

Based on a blastP search against the NCBI's non-redundant library (nr), we retrieve putative homologs for each protein query in a whole proteome. Canonical metazoan proteins should return better blast hits to metazoans than to non-metazoans. Conversely, candidate HGT of non-metazoan origin are expected to return better blast hits to non-metazoans than metazoans. Our method uses the Alien Index metrics as described in (2) to detect a significant gap between the best metazoan and non-metazoan e-values as an indicator of putative HGT. An Alien Index (AI) > 0, indicates a better hit to non-metazoans than to metazoans. An AI > 30 corresponds to a difference of magnitude > e10 between the best non-metazoan and metazoan e-values and is estimated to be a clear indication of HGT. The main difficulty of this approach is to automatically retrieve taxonomic information and to assign a hit to either metazoan, or non-metazoan categories. Our methods not only allows classifying best blast hits in metazoans and non-metazoans but it further classifies non-metazoan best hits into virus, bacteria, archaea, fungi, plant and "other". Thus it allows both the detection of putative HGT acquisitions and identification of the putative donors.

References

1. J.-F. Flot et al., Genomic evidence for ameiotic evolution in the bdelloid rotifer Adineta vaga, Nature (2013).

2. E. A. Gladyshev, et al., Massive horizontal gene transfer in bdelloid rotifers, Science (2008).

F04: Anne Friedrich, Cyrielle Reisser, Paul Jung, Gilles Fischer and Joseph Schacherer. Population genomics reveals the evolutionary fate of a large-scale introgression in a protoploid yeast species

Abstract: With the advent of high-throughput technologies for sequencing, the complete description of genetic variation that occurs in populations is foreseeable but still far to be reached. Explaining the forces that govern the pattern of genetic variation is essential for elucidating the evolutionary history of species. In fact, genetic variation results from a wide assortment of evolutionary forces among which mutation and recombination play a major role in shaping the nucleotide composition of genomes.

In this context, yeast species are of particular interest, as they represent a unique resource for studying the evolution of intraspecific genetic diversity in a phylum spanning a broad evolutionary scale. To date, large-scale polymorphism surveys have only focused on two closely related yeast species: Saccharomyces cerevisiae and Saccharomyces paradoxus.

To obtain new insights into the evolutionary forces shaping natural populations, we sequenced the genomes of an expansive worldwide collection of isolates of a distant species relative to S. cerevisiae: Lachancea kluyveri. Among the 28 sequenced isolates, we identified 6.5 million SNPs, revealing a highly polymorphic population. The phylogeny and population structure of L. kluyveri, inferred from these polymorphic positions, provide clear evidence for well-defined geographically isolated lineages. Furthermore, we showed that a large introgression event of 1-Mb of GC-rich sequence in the chromosomal arm occurred in the last common ancestor of all L. kluyveri strains. Our population genomic data clearly revealed that the introgressed region underwent a molecular evolution pattern very different from the rest of the genome. It is characterized by a higher recombination rate, with a dramatically increased A:T→G:C substitution rate, which is the signature of a increased GC-biased gene conversion. The presence of two distinct recombination and substitution regimes within the same genome demonstrates that the chromosome-scale compositional heterogeneity will persist after the genome has reached mutational equilibrium. Altogether, the data presented herein suggest that large DNA introgressions resulting from interspecific hybridizations lead to different evolutionary patterns within a given genome.

F05: David Pflieger, Anastasie Sigwalt, Jing Hou and Joseph Schacherer. An automated R pipeline to analyze the genetic complexity of stress tolerance in Saccharomyces cerevisiae

Abstract: A fundamental challenge in genetics is to understand the molecular mechanisms governing the relationships between genotype and phenotype. In this context, the yeast Saccharomyces cerevisiae is an ideal model. Natural isolates of this species are found in various ecological and geographical niches, and present a broad range of phenotypic diversity. Dissecting the genetic architecture and complexity underlying the observed phenotypic variation is valuable to understand the patterns of evolution and natural selection.

To assess the genetic complexity of the phenotypic variations, we performed systematic crosses between the lab strain Sigma 1278b and 44 natural isolates of S. cerevisiae, and evaluated the phenotypic distribution in the offspring across a large panel of stress conditions. For each cross, quantitative measurements of the normalized growth ratio over 30 conditions (carbon sources, chemical compounds, etc) were obtained for 40 segregants from 10 tetrads, spanning 3 time points. This has generated a huge dataset of measurements that could only be efficiently analyzed using an automated method. To this end, we developed a complete pipeline in R, which allowed for automatic extraction of the overall phenotypic distribution for each cross under a given condition, as well as the pattern of segregation in the tetrads. For each case, we first tested the normality or bimodality of phenotypic distributions (Shapiro-Wilk Test and Dip Test of Unimodality) using the normalized growth rates, which gives an indication concerning the complexity of the genetic control. Then, for cases with a possible simple genetic architecture (cases with bimodal distributions), we automatically identified the segregation pattern in the tetrads using a clustering method (kmeans n=2). Modeling the segregation types allowed for systematic analysis of the type of genetical control (mendelian

or epistasic model).
Using this approach, we assessed the genetic architecture of multiple traits in a quantitative and high-throughput manner. We found that some resistance phenotypes such as copper sulfate and sodium chloride were mostly under monogenic control, whereas resistance to other condition like ethanol, potassium chloride or ribose was of complex origin. This pipeline also allowed for rapid identifications of crosses that might be useful for further linkage studies (Bulk Segregant Analysis, QTL mapping).

F06: Stefanie Mühlhausen and Martin Kollmar. Analysing the taxonomic distributions of conserved introns with GenePainter

Abstract: The conservation of intron positions comprises information useful for de novo gene prediction as well as for analyzing the origin of introns. Multiple sequence alignments can be improved by incorporating information about gene structures. GenePainter is a standalone tool that maps gene structures onto multiple sequence alignments and protein structures. Moreover, taxonomic information can also be incorporated and projected onto individual introns. In order to assess the taxonomic distribution of introns, lineages of the species represented in the multiple sequence alignment are collected from NCBI taxonomy. For every intron occurring in the provided gene structures, the number of occurrences is determined. Concerning the distribution of introns, their occurrence is resolved on species-level. The last common ancestor of all species harbouring the respective intron as well as the distribution of species onto direct descendants of that last common ancestor is provided. A preliminary analysis of the coronin gene family, a dataset comprising more than 500 coronin genes in all eukaryotic lineages, reveals introns with small taxonomic distribution as well as deeply conserved introns. For example, a large amount of introns present in Animalia and Fungi are conserved between these two groups, which diverged estimated a billion years ago. In addition, GenePainter is a valuable tool to study gene family evolution. The resolution of conserved exon-intron patterns within sub-families allows resolving sub-family memberships and is thus an alternative to phylogenetic analysis. In summary, the analysis of intron conservation reveals useful information about the evolution of gene families, such as intron conservation and sub-family resolution.

F07: Sophie Siguenza, Hélène Badouin, Stéphane De Mita, Jérôme Gouzy and Ludovic Cottret. WeggLib, a web interface for population genomics

Abstract: The new high-throughput technologies allow rapid sequencing of a large number of individuals within a given population. The differences between these genomes, taking into account their environments, can draw different evolutionary scenarios linking several populations. Thus, methods brought by population genomics afford interests far beyond the community of specialists of this field.
To facilitate the first steps of the neophyte in population genomics, we introduce here the WeggLib web interface.
The main features of WeggLib are :
-an easiest approach of population genomics methods by users without programming skills,
-an interpretation of results with online visualisation tools,
-a flexibility in development through quick integration of complex pipelines.
Easy to use : Today, WeggLib contains around twenty functions that apply on sequences, multiple alignments, phylogenetic trees or tables. Most of these are based on EggLib library, a python library concerning population genomics. WeggLib's friendly GUI eases the access to EggLib functions for any user by providing a helpfull graphical output.
A possible usage of WeggLib is the diversity analysis of a set of sequences. Depending on the

current object kind, WeggLib displays only the adapted tools, guiding the user in his choice for the next step. Thus, from an alignment, the user can launch tools for translation, extraction, character replacement, gaps filtering, data formating (Fasta, Nexus, Phylip, PhyML), ..., as well as polymorphism analysis. This last tool calculates the basic statistics in population genetics (Tajima's D, FST, Pi, estimators of theta, etc...) allowing the user to do comparisons thanks to interactive graphics. To complete his analysis, the user can compute and visualise the phylogenetic tree corresponding to the alignment studied.

The lightweight interface has been designed for running the same function on several objects at once and for making easier the analyses with interactive help, sorting, filtering and interactive result visualisation.

At last, being a web application, WeggLib does not need any software install excepted a modern web browser (Firefox or Chrome).

Implementation: Beyond friendly usage, WeggLib has also been designed to be maintainable. With the help of a formatted description file (xml), each new function is directly interfaced: it is not necessary to recode new input forms and results window, which saves lot of code and improves Wegglib's maintainability.

Furthermore, the command-line of a function is called from this xml descriptors, allowing us to use any programmation language.

WeggLib takes advantage of the ExtJs javascript framework for advanced charting and graphing capabilities, as for its use of " Model View Controller " (MVC) pattern that greatly facilitates development teams.

Link to WeggLib: http://symbiose.toulouse.inra.fr/wegglib

This work has been supported by the ERC GenomeFun and ANR Gandalf grants

F08: Hélène Badouin, Jérôme Gouzy, Alodie Snirc, Sophie Siguenza, Antoine Branca and Tatiana Giraud. Population genomics of the phytopathogenic fungi Microbotryum violaceum

Abstract: Understanding how species adapt to their environment is a key question in evolutionary biology. To achieve this goal, scientists want to identify what genes are responsible for local adaptation, how many there are and how they are distributed within genomes. Studying patterns of genetic variation within and between populations can reveal signatures of positive selection.

Pathogens are excellent models to study adaptation, because they tend to evolve rapidly in response to the coevolving host environment. The basidiomycete Microbotryum violaceum is a species complex that parasites more than a hundred species of Caryophyllaceae. The fungus hijacks the reproductive system of its host to produce teliospores instead of pollen, eventually causing the host sterility. Microbotryum species are highly specific to their host, as most species infect a single host plant species. Speciation often involves host jump between closely related hosts, and closely related fungus species can hybridize in nature.

The analysis of EST sequences of Microbotryum violaceum has revealed genes that potentially have experienced positive selection between species adapted to different hosts. Nevertheless, genes causing host adaptation remain largely unknown. The genome of Microbotryum violaceum has been recently sequenced. In order to identify candidates for local adaptation, we sequenced 50 individuals of two sister-species, M.violaceum ssp silenes-dioicae and M. violaceum ssp lychnidis-dioicae. After mapping the reads to the reference genome of M. violaceum ssp lychnidis-dioicae and filtering, we identified 197,929 autosomal single nucleotide polymorphisms for M. violaceum ssp lychnidis-dioicae and 30,000 for M. violaceum ssp lychnidis-dioicae.

We analyzed the genetic structure of both species with the Structure sofware and Discriminant Pluri-Component Analysis, and found no recent hybrids in our dataset. Furthermore, we confirmed that European Microbotryum lychnidis-dioicae populations are structured in 3

distinct genetic clusters. Those are more strongly differentiated than previously thought based on microsatellites, with pairwise Fst ranging from 0.5 to 0,7. Furthermore, we found that the Western cluster is divided into a southern and a northern cluster, occurring in sympatry at intermediate latitude.

Since strong genetic structure can bias tests for positive selection, we performed selection tests on each cluster separately. We ran genomic scans for positive selection using Tajima's D and Fay and Wu's H standardized by Zeng (Z). We found potential positively selected region, including two regions containing genes coding for small secreted proteins such as and Major Facilitator proteins. We also ran McDonald-Kreitman tests for positive selection and found several dozens of candidates, including putative virulence factors.

F09: Gabriel V Markov, Praveen Baskaran and Ralf J Sommer. The same or not the same: Lineage-specific gene expansions and homology relationships in multigene families in nematodes

Abstract: Homology is a fundamental concept in comparative biology and a crucial tool for the analysis of character distribution. Introduced by Owen in 1843 in a morphological context, homology can similarly be applied to protein-coding genes. However, in molecular biology the proper distinction between orthology and paralogy was long limited by the absence of whole-genome sequencing data. By now, genome-wide sequencing allows comprehensive analyses of the homology of genes and gene families at the level of an entire phylum. Here, we analyze a manually curated dataset of more than 2000 proteins from the genomes of 11 nematode species of seven different genera, including free-living and animal and plant parasites to study the principles of homology assignments in gene families. Using all sequenced species as an extensive outgroup, we specifically focus on the two model species Caenorhabditis elegans and Pristionchus pacificus and compare enzymes involved in detoxification of xenobiotics and synthesis of fatty acids. We find that only a small proportion of genes in these families are one-to-one orthologs and that their history is shaped by massive duplication events. Of a total of 348 and 543 genes from C. elegans and P. pacificus, respectively, only 37 are one-to-one orthologs. Thus, frequent amplifications and losses are a widespread phenomenon in nematode lineages, and we also report variation in birth and death rates depending on gene families and nematode lineages. We discuss the consequence of the near absence of one-to-one orthology in related organisms for the application of the homology concept to protein-coding genes in the era of whole-genome sequencing data.

F10: Dong Seon Kim, Hye Ji Oh, Dongjin Choi and Yoonsoo Hahn. MOXD2 Gene Inactivation in Apes and Whales

Abstract: The MOXD2 gene encodes a membrane-bound monooxygenase similar to dopamine-beta-hydroxylase, and has been proposed to be associated with olfaction. In this study, we analyzed MOXD2 genes from 64 mammalian species, and identified loss-of-function mutations in apes (humans, Sumatran and Bornean orangutans, and five gibbon species from the four major gibbon genera), toothed whales (killer whales, bottlenose dolphins, finless porpoises, baijis, and sperm whales), and baleen whales (minke whales and fin whales). We also identified a shared 13-nt deletion in the last exon of Old World cercopithecine monkeys that results in conversion of a membrane-bound protein to a soluble form. We hypothesize that the frequent inactivation of MOXD2 genes in apes and whales may be associated with the evolution of olfaction in these clades.

F11: Hye Ji Oh, Dongjin Choi and Yoonsoo Hahn. Evolution of Intronless Genes in Ciona genus

Abstract: Tunicates, the sister clade of vertebrates, have miniature genomes and numerous intronless genes compared to other animals. It is still unclear how the tunicates acquired such a large number of intronless genes and achieved a high degree of genome compaction. Here, we analyzed sequences and intron-exon organizations of homologous genes from two closely related tunicates, Ciona intestinalis and Ciona savignyi. We found seven cases in which ancestral introns of a gene were completely lost in a species after their divergence. In four cases, both the intronless copy and the intron-containing copy were present in the genome, indicating that the intronless copy was generated by retroduplication. In the other three cases, the intron-containing copy was absent, implying it was lost after retroduplication. This result suggests that retroduplication and loss of parental genes is a major mechanism for the accumulation of intronless genes and genome compaction in tunicates.

F12: Dong Seon Kim, Hye Ji Oh, Dongjin Choi and Yoonsoo Hahn. Gains and Losses of N-glycosylation Sites during Human Evolution

Abstract: N-linked protein glycosylation plays important roles in various biological processes including cell-cell adhesion, protein excretion, and regulation of protein stability and activity. We developed a bioinformatics method and identified gains and losses of N-glycosylation sites in human proteins during evolution by analyzing human and mouse glycoproteome data and orthologous mammalian protein sequences. We examined 2534 N-glycosylation sites of 1027 human proteins and found that 124 sites of 99 proteins newly appeared in the human lineage. We also examined 5531 N-glycosylation sites of 1659 mouse proteins and found that 42 sites of 38 proteins disappeared in the human lineage. We identified 5 cases of human-specific gains (Asn518 of ALBU, Asn196 of APMAP, Asn91 of CD166, Asn453 of FIBA, and Asn76 of THYG proteins) and 3 cases of human-specific losses (Ser2155 of CELR1, Lys279 of SIAT9, and Ile121 of VSI10 proteins). We propose that gains and losses of N-glycosylation sites may result in evolution of protein function and novel phenotypes during human evolution.

F13: Jackson Peter, Anne Friedrich, Agnès Llored, Anders Bergstrom, Anastasie Sigwalt, Kelle Freel, Gianni Liti and Joseph Schacherer. The 1002 yeast genomes project: a framework for genome-wide association studies

Abstract: In all species, genetic diversity is the raw material for phenotypic diversity. Genome-wide investigation of the patterns of polymorphism in a large sample of individuals is the first step to assess the relationship between genotype and phenotype within a species. Large-scale polymorphism surveys and analyses were initiated for a small number of species including Arabidopsis thaliana and Homo sapiens. Today, projects with the goal of describing the whole-genome sequence variation in more than 1000 human genomes and 1001 accessions of A. thaliana are underway. To date, yeast population genomics only focused on a limited number of isolates (less than 100 strains) and this stands in contrast with these projects. Because of their small and compact genome, yeasts and more precisely S. cerevisiae represent a powerful model for population genomics. Here, we present our project: "The 1002 yeast genomes project: a framework for genome-wide association studies". We sequenced approximately 1000 genomes of S. cerevisiae using the Illumina HiSeq 2000 technology, with a mean coverage of 250X.

Due to the broad diversity of strains selected for sequencing, this population genomics dataset reveals an accurate picture of the genomic variation within S. cerevisiae. Indeed, the next generation sequencing data are suitable to reveal the entire repertoire of single nucleotide polymorphisms (SNPs) as well as the degree of copy number variation (CNV). We therefore expand the current catalogue of SNPs described in S. cerevisiae so far. Furthermore, our

strategies to sequence diploid strains (as they are mostly isolated in nature) allow us to fully characterize the spectrum of CNV (such as aneuploidies) and the degree of heterozygosity. The distribution of the genetic diversity across the genome offered an opportunity to understand patterns of variation and recombination. We also analyzed the SNPs to reconstruct the population history and structure.

Furthermore, we performed extensive phenotyping experiments with our strain collection, based on quantitative measurements of the normalized growth ratio over 53 conditions (carbon sources, diverse chemical compounds, for example). The high SNPs density allowed us to perform genome-wide association studies. This dataset leads to the identification of a large set of functional polymorphisms that underlie phenotypic variation.

F14: Francesc Peris-Bondia and Laurence Van Melderen. Comprehensive analysis of the genomic localization of bacterial toxin-antitoxin systems

Abstract: Bacterial toxin-antitoxin (TA) systems are composed of a stable toxin and an unstable antitoxin. These modules were originally detected on plasmids in the 80's. In this context, TA systems are involved in plasmid addiction. This phenomenon relies on differential stability of the toxin and antitoxin components. When plasmid-free daughter-cells are produced, antitoxin molecules are degraded. Toxins are then free to act on their targets and this lead to growth inhibition and eventually cell death. Therefore, the cells are addicted to the presence of the TA systems. These systems are widely spread in bacterial and archeal genomes. Although a variety of functions have been assigned to chromosomally-encoded systems, their biological roles are still unclear or limited to particular cases, making the 'selfish' behaviour an attractive hypothesis to explain the evolutionary success of TA systems. Type II TA systems are thought to invade bacterial genomes through horizontal gene transfer but it still remains unclear. Are these systems mainly present in mobile elements? Are the TA genes always part of TA systems or some of them can exist without the presence of its partner?

We have performed an exhaustive search for proteins belonging to TA systems (type II to V) and a comprehensive analysis of their distribution, genomic organization and context in order to answer these questions. Using the known and experimentally tested TA systems we have search for similar proteins in all the bacterial, viral and plasmidic genomes available in the NCBI database using hmm profiles. We have determined if toxin and antitoxin ORFs co-localize and look for the presence of signature of mobile elements in the neighbouring genes. The primary results of this work are that each TA family has a particular distribution, that different antitoxins can be associated to the same toxin, as it has been seen in other studies, and in some cases toxin or antitoxins can appear without known antitoxin or toxin, respectively.

Genomic distribution of some families is coherent with 'selfish' behaviour while others are more easily explained by acquisition of different roles in bacterial physiology throughout evolution.

This might thus reveal novel putative functions and/or TA integrations in host regulatory networks. In addition, this work highlighted potential novel toxins or antitoxins that will be experimentally validated in our lab.

F15: Julien Fumey, Céline Noirot, Hélène Hinaux, Sylvie Rétaux and Didier Casane. Evo Devo of Astyanax mexicanus cavefish: A new time frame and its consequence on the underlying evolutionary mechanisms.

Abstract: Populations of blind cavefish belonging to the Mexican tetra species Astyanax mexicanus are outstanding models to study the evolution of vertebrates at a small time scale.

In particular, the phenotypic convergence of independently-evolved, cave-adapted populations allows questioning whether the evolution of similar cave phenotypes involved the fixation of standing genetic variation or the apparition of de novo mutations. To get an estimation of the time frame of the evolution of the Astyanax Pachón cave population which is considered as one of the “oldest” and most isolated population in the Sierra del Abra, we applied a population genomics approach. We compared transcriptome-wide the polymorphism and substitution rates of the Pachón population and a surface fish population (San Solomon Spring, Texas), using the Buenos Aires tetra (Hyphessobrycon anisitsi) as a close outgroup to identify ancestral and derived alleles. These data were compared to simulations of population evolution in which various parameters varied, such as the size and age of the populations and the gene flow between populations, to determine parameters that are compatible with the differentiation observed. The polymorphism was higher in the surface population than in the cave population, suggesting, as expected, a higher effective population size for the river-dwelling fish population. We also observed higher substitution rates in cavefish than in surface fish, also in accordance with a lower cavefish population size allowing a more rapid fixation of derived alleles, but implying that the Pachón cave population is much “younger” than previously estimated and may have spent less than 100.000 years underground. We will discuss the consequences of this new time frame on the underlying evolutionary mechanisms responsible for the morphological changes observed in cavefish populations.

Work supported by ANR BLINDTEST.

F16: Fanny Pouyet, Marc Bailly-Bechet and Laurent Guéguen. Evolution of Codon Usage in E. coli

Abstract: Codon substitution models describe how a codon sequence evolves along a phylogenetic tree. We implement a new codon model that separate mutational bias from selection on codon usage. It is available in Bio++ suite [Guéguen L.,et al. Mol. Biol. Evol 30(8):1745-50, Aug (2013)].

The genetic code is redundant, and some codons, called synonymous, are translated in the same amino acid. Degeneracy of the code does not however lead to a uniform or a random usage of those synonymous codons which is called codon usage bias – CUB. CUB may vary between species and between genes. These differences are easily observable and measurable on extant sequences. However, to fully understand how this bias arose and is maintained in a set of genes, we need a model that studies codon usage in an evolutionary approach. We develop such a model, inspired from Yang and Nielsen [Yang, Z. and Nielsen, R. Mol. Biol. Evol 25(3):568-579, Mar (2008)]. Our model distinguishes evolution of coding sequences at nucleotidic, codons and amino acids level. Our model explicitely describes separately the mutational bias of nucleotides which applies on all nucleotides independently of their position in the codon, the selection between synonymous codons and the preferences among amino acids.

We apply this model in an homogeneous and non-stationary context, in a maximum likelihood framework. We study the evolution of codon usage bias in the core genome of thirty-five strains of E. coli. We show that nucleotidic mutational bias and codon bias have different equilibrium points: mutational bias tends to increase AT composition mostly by an increase of T content whereas codon bias favors GC enrichment by an increase of C and a decrease of A. Codon bias impact on nucleotidic composition is homogeneous across the genome. We measure the selection on codon usage. We dismiss the role of mutational bias and corroborate the one of codon bias on codon usage evolution.

F17: Darius Kazlauskas and Česlovas Venclovas. Viral DNA replication: new insights and discoveries from large scale computational analysis

Abstract: Ability to replicate is essential for all living entities. Duplication of genetic information is carried out by replication proteins. DNA replication is well studied in T7, T4 phages and herpes viruses; however, the information about replication mechanisms from other groups of viruses is either scarce or missing altogether. Double-stranded (ds) DNA viruses infect cells from all domains of life, evolve fast and are very diverse. Their genome size varies from 5 to 2500 kbp. To better understand viral DNA replication we identified replication proteins in dsDNA viruses using current state-of-the-art homology detection methods. Over 150000 proteins from 1574 genomes were analyzed. We found out that the composition of replication machinery depends on virus genome size. Small viruses (<40 kbp) use protein-primed DNA replication or rely on replication proteins from the host. Large viruses (>140 kbp) have their own RNA-primed replication apparatus often supplemented with processivity factors and DNA topoisomerases to increase replication speed and efficiency. Latter insight led us to a search for „missing“ replication components in large genomes. This has resulted in a discovery of single-stranded DNA binding (SSB) proteins in largest eukaryotic viruses. Surprisingly these proteins turned out to be homologs of SSB proteins previously thought to be specific for T7-like phages. Another surprise came from the analysis of the herpesviral helicase-primase complex. We found that its component (UL8) is a highly diverged inactivated B-family DNA polymerase.

F18: Héloïse Philippon, Céline Brochier-Armanet, Christine Brun, Evelyne Goillot and Guy Perrière. Origin and Evolution of Cellular Signaling Pathways: PI3K as a case study

Abstract: Phosphatidylinositol-3-kinases (PI3K) are a family of enzymes modifying phosphoinositides in phosphatidylinositols-3-phosphate. It is divided into three main classes, including 14 human homologs. While class II enzymes are composed of a single catalytic subunit, class I and III contain also a regulatory subunit. Located upstream of the AKT/mTOR signaling pathway, PI3K are second messengers of extracellular signals, and involved in many critical cellular processes such as survival, angiogenesis or autophagy (Graupera and Potente, Exp. Cell Res., 2013).

Moreover, it was shown that mutations affecting PI3K coding genes or their expression are associated to many cancers, including breast, colon, and prostate (Engelman, Nat. Rev. Cancer, 2009). Despite their biological and medical importance, PI3K have not been studied in detail from an evolutionary point of view, and only two very incomplete phylogenies have been published to date (Kawashima et al., Dev. Genes Evol., 2003; Brown and Auger, BMC Evol. Biol, 2011). These studies concluded that given their crucial role as second messenger signaling, the increasing complexity linked to multi-cellularity was accompanied by the evolution of more diverse and specialized PIKs signaling cascades.

Here we present an in-depth phylogenetic analysis of the PI3K. We confirmed that PI3K catalytic subunits form a monophyletic group, whereas regulatory subunits form three distinct groups. The phylogeny of the catalytic subunit indicates that two major duplications events occurred during the evolutionary history of Eucarya: the most ancient arose in their Last Common Ancestor (LECA) and led to the emergence of class III and class I/II. The second, which led to the separation between classes I and II, took place in the ancestor of Unikonta (i.e. Amoebozoa, Fungi, and Metazoa). These two major events were followed by many duplications in particular in vertebrates (including Human), but also in various protist lineages. This asks the question of the function of the resulting genes, especially in unicellular organisms. Regarding the regulatory subunits, we identified homologs of class III in all eukaryotic groups indicating that for this class both the catalytic and the regulatory subunits

were presents in LECA. In contrast, homologs of class I formed two subgroups restricted to Metazoa and Amoebozoa, and to Chordata, respectively, suggesting a relatively recent origin. This result and the absence of class II regulatory subunit is puzzling given that the class I/II catalytic subunit was present in LECA and has been conserved in most present-day eukaryotic lineages. This suggests different enzyme action in these lineages.
Our analyses shed a new light on the evolutionary history of PI3K enzymes and open the door for broader studies of the AKT/mTOR signaling pathway. By this way, we would like to provide a better understanding of their functioning in present-day eukaryotes.

F19: Annalisa Fierro, Sergio Cocozza, Antonella Monticelli, Giovanni Scala and Gennaro Miele. Continuos and Discontinuos Phase Transitions in Quantitative Genetics: the role of stabilizing selective pressure

Abstract: By using the language of statistical mechanics we have analyzed the evolution of a population of N diploid hermaphrodites in random mating regime. The population evolves under the effect of drift, selective pressure in form of viability on an additive polygenictrait, and mutation. The analogy allows to determine a phase-diagram in the plane of mutation rate and strength of selection. The involved pattern of phase transitions is characterized by a line of critical points, corresponding to a selective pressure smaller than a critical value, whereas we observe discontinuous phase transitions for a stronger selective pressure. In this framework, the mutation rate, which allows the system to explore the accessible microscopic states, is the parameter controlling the transition from a disordered to an ordered state, and plays a role, being in some sense analogous to that of magnetic field in ferromagnetic systems. As it is expected, we observe a metastable hysteresis in correspondence of the discontinuous phase transitions. Interestingly we can describe such hysteresis phenomena in terms of population heterozygosity, properly averaged over the loci.

F20: Giovanni Scala, Ornella Affinito, Gennaro Miele, Antonella Monticelli and Sergio Cocozza. Evidence for Evolutionary and non Evolutionary phenomena shaping the genetic variant distribution near Transcription Start Sites

Abstract: Transcription Start Sites (TSS) represent hot spots of the genome particularly stressed by transcription related phenomena. At the same time these regions are very susceptible to variations due to their critical role in the initiation and regulation of the transcriptional activity. In this framework, the study of genetic variability can shed light on the evolutionary and non evolutionary mechanisms acting near TSS.
In this work we performed a genome-wide analysis of the distribution of SNPs inside a 10 kb region flanking human Transcription Start Sites (TSS) by dividing SNPs in four classes according their frequency (rare, two intermediate classes, and common).
We found that, in this region, the variants' distribution depends on their frequency class, and on their localization relative to TSS . Splitting TSSs in those located inside a CpG island (CGI-TSS), and not located inside a CpG island (NCGI-TSS), we found that variant distribution is generally different for these two subsets. We found a significant relationship of the rare variants distribution with nucleosome occupancy scores. Furthermore, the analysis seems to suggest that evolutionary (purifying selection) and non evolutionary (Biased Gene Conversion) forces play a relevant role in determining the relative SNP frequency around TSSs. As a last step, we analyzed the potential pathogenicity of each class of variants using the Combined Annotation Dependent Depletion score.
This study provides a novel and detailed picture of the distribution of genomic variant around TSSs, giving insights on the forces that play a role in the insurgence and maintenance of variability in such critical regions.

F21: Cécile Pereira, Alain Denise and Olivier Lespinet. A new method for improving the prediction and the functional annotation of ortholog groups

Abstract: Motivation: In comparative genomics, orthologs are used to transfer annotation from genes already characterized to new sequenced genomes. Many methods have been developed for finding orthologs in sets of genomes (Altenhoff and Dessimoz, 2009)□. However, the application of different methods on the same proteomes can lead to distinct orthology predictions (Chen et al., 2007; Dalquen et al., 2013)□. In this work we propose to combine results obtained by several of these methods by developing a new method based on a meta-approach. The purpose of this method is to produce better quality results by using the overlapping results obtained by several individual orthologous gene prediction methods.
Results: We developed a new method, based on a meta-approach that is able to combine the results of several methods for orhologous genes prediction.
Our new method runs in two steps. The first step aims to construct seeds for groups of orthologous genes; these seeds correspond to the exact overlaps between the results of all or several methods. In the second step, these seed groups are expanded by using HMM profiles. We evaluated our method on two standard reference benchmarks, OrthoBench (Trachana et al., 2011)□ and Orthology Benchmark Service (http://orthology.benchmarkservice.org/) . Oure method presents a higher level of accurate predicted groups than the selected gene ortholog prediction input methods. Moreover, our method presents the largest GO term annotation similarity and the largest number of predicted pairs compared to nine state-of-the-art methods.
Conclusion: The meta-approach based method presented here appears to be a reliable method of prediction of ortholog groups. Based on the combination of existing methods, it allows to find a consensus of higher quality. Both ortholog group quality and consistence of group annotation have been positively tested. As a large number of methods for predicting groups of orthologous genes exist, it is quite conceivable to apply this meta-approach either to different methods or to more methods.
Availability: The software is freely available at http://bim.igmors.u-psud.fr/mario/

F23: Alexandra Vatsiou, Eric Bazin and Oscar Gaggiotti. Pathways enriched for selection

Abstract: Motivation: The identification of genetic variations that contribute to a genetic disease remains a major challenge in the research in human genetics. One of the factors that could have shaped the genetic diversity in a population is natural selection. Many studies investigate the effects of mutations and genes in the phenotype independently and therefore do not consider for large functional and gemonic effects that could originate from multiple small ones. Assessing and analyzing large-scale genomic data based on the gene sets level or in the network of gene-sets could provide further biological insight. Our objective in this work is to detect signals enriched for positive selection in the biological pathway level that could improve our knowledge in the biological interpretation of these effects.
Methods & Results: We used the HapMap data phase II for genetic data http://hapmap.ncbi.nlm.nih.gov/downloads/phasing/2007-08_rel22/, the human genes from coding regions from the Entrez NCBI database (data downloaded on 5/2014) and the human pathways from the Biosystems database http://www.ncbi.nlm.nih.gov/biosystems (data downloaded on 5/2014). The analysis consists of the following steps:
1. Analyze the haplotype and SNP data using a composite likelihood method to test for positive selection between populations. The method XPCLR was chosen as it was proved to detect the wider range of signals over other methods (Vatsiou et al. in prep). Three comparisons were conducted: a. CEU-YRI, b. CEU-CHB/JTP and c. YRI-CHB/JTP.

2. Match the SNPs to 27081 genes according to their start and end position that were extracted from the Entrez database. We also included 50kb upstream or downstream away from the gene to account for intergenic regions.
3. We considered the highest XPCLR normalized score as the representative of the whole gene. A further normalization though was made to account for genes with large number of SNPs.
4. We tested for enrichment signals for positive selection calculating the sum of all the scores in each of the gene sets (total number of gene sets: 2362) and we estimated the significance taking into account the different gene set sizes.
Conclusions: It is about to be shown the exact signals we observe from the human genome pathways and the way that they could possibly be involved in adaptation events. Even though many enrichment analysis have been conducted before (Daud et al. 2013), our study will possibly reveal more biological pathways enriched for positive selection as we are mainly based on pairwise comparisons between populations using XPCLR, and such an analysis has not yet been considered.

F24: Rafael Piergiorge, Ana Carolina Ramos Guimarães and Marcos Catanho. Evolution and functional genomics of analogous enzymes in the human genome

Abstract: Since enzymes catalyze almost all chemical reactions occurring in living organisms, it is extremely important that the genes encoding such activities are properly identified and functionally characterized. Several studies suggest that the fraction of enzymatic activities in which multiple events of independent origin have taken place during evolution is substantial. However, this topic is still poorly explored, and a comprehensive investigation of the occurrence, distribution and implications of these events, involving organisms whose genomes have been completely sequenced, has not been done so far. Fundamental questions, such as how analogous enzymes originate, why so many events of independent origin have apparently occurred during evolution, and what is the reason for the coexistence in the same organism of distinct enzymatic forms remain to be answered. In this context, the purpose of this project is to investigate the biological importance and the evolutionary role of functional analogous enzymes identified in metabolic pathways annotated in the human genome. We intend to achieve an overview of the processes of de novo origin (convergence) and extinction (pseudonization) of enzymatic forms in the human species occurred during evolution, as well as to acquire a functional profile of the genes encoding these analogous enzymes, analyzing the structure and organization of these genes in the human genome, their transcriptional activity in different conditions, and their essentiality.

F25: Fabricia Nascimento and Allen Rodrigo. The evolutionary dynamics of endogenous retroviruses: testing the strict master and transposon models by computer simulations

Abstract: Endogenous retroviruses (ERVs) are viewed as ancient retroviral infections in vertebrate genomes and are commonly referred to as viral “fossils”, accounting for approximately 8% of the human genome. They are retroviruses that have infected and colonized germ cells and are passed to the offspring. Because proviruses have the potential to disrupt genes and to change the host gene expression, which might lead to diseases, they are negatively selected and end up losing their viral function, becoming unable to re-infect. However, some ERVs are still able to re-infect because they still have intact viral genes. They are therefore theoretically able to form viral particles, even after millions of years within the host genome, yet the reason for this remains elusive. In this study we developed algorithms to understand the evolutionary dynamics of full-length ERVs (genome size = 10,000bp) in host genomes by expanding described transposable element (TE) models known as the “strict

master" and "transposon" models. The first model assumes that only one element of a given family is capable of producing a copy, while the second model assumes that all elements of a given family are equally like to produce a new copy in the host genome. Algorithms in Python were written to simulate phylogenetic trees under variations of the "strict master" and "transposon" models, by testing how viral inactivity and ongoing activity for retrotransposition or reinfection of an ERV family influences phylogeny shape. We also included additional parameters such as ERV mutation rate and the rate of ERV replication in the host genome. The program Seq-Gen was used to simulate DNA sequence alignments (10,000bp) under the simulated trees. These alignments were used to reconstructed phylogenetic trees using the program RAxML. Beta-splitting statistics were calculated for tree shape and results show that maximum likelihood analysis do not always recover the "true" tree topology when DNA sequences are used for phylogenetic reconstructions. These results show promise with respect to our understanding of activity and inactivity of ERVs in vertebrate host genomes. It suggests that although ERV tree topologies are commonly referred to as following a "strict master" or "transposon" models by visual inspection, we should be careful in doing so. Results suggest that there are regions in the tree topology space where we cannot distinguish between the different models.

F26: Nadav Rappoprt and Michal Linial. Comparative Analysis of Insects' Complete Proteomes

Abstract: Insects are the most diverse class of animals known, with million species that represent the majority of animals on earth. At present, there are tens of fully sequenced insects' genomes that cover a range of habitats, social behavior, sizes and life cycles. In view of such immense collection of genomes, the evolution and functional relationships of their proteins remain a challenging task.

We analyzed the relatedness of the complete proteome of 17 genomes from insects as well as an outlier from the subphylum Crustacea. We have applied an unsupervised hierarchical protein clustering method to extract biological insights from these proteomes. The resulting hierarchical tree was used to chart the proteins into protein families. The hierarchical tree, called ProtoBug covers 287,405 protein sequences that are partitioned to 20,134 families (excluding singletons). The represented proteomes include lice, ants, bee, beetle, flies, mosquitoes and water flea. We are able to highlight groups of putative gene-loss events and to quantify their dynamics. We characterize protein families that were expanded or reduced in specific species and in distinct evolution branches. We marked novel proteins with unique functionality and identified specie specific paralogs. The results from the comprehensive comparative analysis show cases of adaptation and evolutionary novelty that satisfy the unique social or physiological needs.

We conclude that the ProtoBug and the analytical methods that are developed provide a rich resource towards automatic annotation of additional insects. We propose ProtoBug as a prototype for inferring the relatedness and function of newly sequenced proteomes.

F27: Nicolas Rodrigue and Nicolas Lartillot. Phylogenetic measurements of departures from the mutation-selection equilibrium

Abstract: Codon substitution models have traditionally been used to measure selective pressures in protein-coding genes by evaluating the ratio of rates of nonsynonymous to synonymous substitutions. Recently, we have proposed a mutation-selection framework in which site-specific purifying selection at the amino acid level is explicitly modeled (Rodrigue et al., PNAS, 2010). Loosely speaking, under this model, substitutions at a given position occur at the neutral or near-neutral rate when they are either synonymous, or when they correspond to replacements within a sub-set of suitable amino acids---substitutions to ill-suited amino acids have much lower rates. As an alternative to traditional methods, we explore the idea of using our recent model as the null against which to test for deviations from the neutral/nearly-neutral regime. We present applications of this approach on a few data set of protein-coding genes, and discuss how the null model can be extended so as to test for different reasons for measured deviations, such selection on codon usage.

F28: Thies Gehrmann and Marcel Reinders. Proteny Discovering and visualizing statistically significant syntenic clusters at the proteome level between divergent genomes

Abstract: Background: With more and more genomes being sequenced, detecting synteny between genomes becomes more important, amongst other things, to derive functional interpretations of orthologous genes.
However, for microorganisms the genomic divergence quickly becomes large, resulting in (for example) different codon usage and shuffling of gene order and gene elements such as exons.
Approach: We present Proteny, a tool to detect synteny between diverged genomes. It operates on the amino acid sequence level to be insensitive to codon usage adaptations.
Furthermore, Proteny assigns significance levels to the syntenic clusters such that clusters can be selected on statistical grounds, including appropriate techniques to perform multiple testing correction.
Finally, Proteny provides novel ways to visualize results at different scales, facilitating the exploration and interpretation of syntenic regions.
Results: We illustrate the performance of Proteny on two closely related genomes (two different strains of Aspergillus niger) and on two distant genomes (two species of Basidiomycota).
In comparison to other tools, we find that Proteny finds larger and generally tighter clusters, encompassing more genes.
Further, we show how genome rearrangements, assembly errors, gene duplications and the conservation of specific genes can be easily studied with Proteny.
Availability : Proteny (together with the examples shown) is available from http://bioinformatics.tudelft.nl/dbl/software.