Midsummer Phylogenetics at UEA

Tuesday, July 17, 2018, School Meeting Room (SCI 2.29)

Program

14:00-14:20 Jo Dicks, Quadram Institute, Challenges and opportunities of ribosomal DNA micro-heterogeneity detection and analysis in yeast NGS datasets

14:20-14:35 Kara Martin, UEA, Bioinformatics approaches for assessing microbial communities in the surface ocean

14:35-14:50 Chris Pyatt, Quadram Institute, Cytochrome P450 diversity in the NCYC collection and their use in biosynthetic gene cluster discovery

14:50-15:20 coffee break

15:20-15:40 Will Nash, The Earlham Institute, Expansion of gene families and signatures of selection in the Australian marsupials

15:40-15:55 Anne-Marie Keane, Quadram Institute, Computing a yeast tree of life

15:55-16:10 Guillaume Scholz, UEA, From phylogenetic networks to MUL-trees (and back)

16:10-16:25 Andris Thompson, UEA, Distributions of cherries and pitchforks in random phylogenetic trees

16:25 End (Pub)

Abstracts

Jo Dicks, Quadram Institute, Challenges and opportunities of ribosomal DNA micro-heterogeneity detection and analysis in yeast NGS datasets

The ribosomal DNA (rDNA) is essential for life, underpinning protein synthesis in all living organisms. Furthermore, the rDNA is a rich source of evolutionary information and has recently been implicated in many diverse biological phenomena ranging from genome integrity and sensitivity to DNA damage to ageing and acting as a target for nutritional signals.

Despite the ubiquity of the rDNA, it exhibits remarkable compositional variation between the kingdoms of life, ranging from a single operon in some bacteria and archaea to hundreds of tandemly repeated unit sequences, organised across several genomic loci, in some eukaryotes. In yeast, rDNA unit sequences of ~9kb are repeated in tandem arrays of tens or hundreds of copies, at one or more genomic loci. While yeast rDNA unit sequences evolve in concert due to homogenising processes such
as gene conversion and unequal sister chromatid exchange, disruptive processes such as point mutation may still occur frequently enough to cause significant levels of sequence heterogeneity within rDNA arrays, including variants within the rRNA genes.

Here, we examine the pros and cons of using different next generation sequencing (NGS) technologies and software tools to detect rDNA micro-heterogeneity. We investigate patterns of particular rDNA sequence variants in two contrasting yeast species – *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* – and link them to key characteristics of genomic evolution. Finally, we look forward to opportunities offered by technologies such as RNA-Seq to understand the expression, and ultimately the function, of rDNA sequence and copy number variation.

Kara Martin, UEA, *Bioinformatics approaches for assessing microbial communities in the surface ocean*

Microbes are vital for life on Earth. For example, they are the major primary producers in the ocean and contribute greatly to the biogeochemical cycling of the elements and which in turn influence the global climate. Microbes inhabit the oceans throughout the world and these cover over 71% of the surface of the earth. Thus, marine microbes can be thought of as the forests of the sea.

Microbes have evolved in different environments in the oceans and in different ways. To gain an understanding of the microbe communities in the Atlantic and Artic Oceans environmental scientists based at University of East Anglia collected ocean samples from 68 stations along a transect of the Arctic Ocean, North Atlantic Ocean and South Atlantic Ocean. In addition, they recorded environmental data at the time of sampling, such as temperature and salinity. They then filtered out the microbes, extracted their DNA and sequenced this using high-throughput sequencing.

The purpose of this work is to analyse this sequencing data so as to shed light on the composition and distribution of microbe communities in the surface of the ocean. To this end, we designed bioinformatic pipelines in order to analyse 18S and 16S rDNA datasets from the various stations. In addition, we developed a novel methodology for normalising 18S and 16S rDNA copy numbers. This enabled us to perform additional analyses such as, biodiversity, co-occurrence and breakpoint analyses.

In our results, we observed a greater diversity of microbes the south Atlantic and tropical regions, versus the artic polar regions of the Atlantic ocean. Moreover, in the co-occurrence analysis on the 18S and 16S rDNA datasets, we found two community networks, one positively correlated to temperature and the other negatively. We also performed a breakpoint analysis on our 18S and 16S rDNA datasets and found a shift in diversity occurring in the North Atlantic Ocean. In particular, the shift occurs in the temperate region of the Ocean, between the polar Arctic and tropical South Atlantic Ocean.

These results are important as from our co-occurrence analysis we can hypothesise that different microbial communities have different preferences for temperature. Moreover, as global warming is predicted to raise the temperatures in the ocean, our results could potentially enable us to forecast how climate change will affect these microbial communities using climate models underpinned by genetic information.
Chris Pyatt, Quadram Institute, *Cytochrome P450 diversity in the NCYC collection and their use in biosynthetic gene cluster discovery*

The National Collection of Yeast Cultures (NCYC; http://www.ncyc.co.uk) contains around 4,000 strains from over 500 species, representing a large part of global yeast diversity. The genomes of nearly one quarter of these strains have now been sequenced, forming the foundation of a project aiming to uncover the metabolic potential of the collection. One important aspect of this project is the computational discovery of biosynthetic gene clusters. Existing gene cluster-finding tools are mostly limited to identifying known ‘core genes’ of particular types (e.g. PKS, NRPS), which can be found easily and used as anchors around which gene cluster searching can be focussed. Cytochrome P450 (CYP) genes are also frequent components of gene clusters and are easily identifiable by four very highly conserved protein motifs. They therefore have the potential to serve as anchor genes.

In this study, the genomes of 792 NCYC strains were searched for CYP genes. 10,740 putative sequences were identified, of which 3,395 were of suitable length. These were classified according to sequence similarity with known CYPs; 2,388 and 538 were found to be members of known subfamilies or families respectively, while the remaining 469 did not match any known fungal CYP. Intriguingly, the novel CYP genes are taxonomically clustered in the basidiomycetes, in which the overall number of CYP genes per genome is also highest, suggesting that these organisms are understudied and may harbour useful diversity. Future work within this project will involve assessing the regions surrounding the CYP genes identified here for novel gene clusters.

Will Nash, The Earlham Institute, *Expansion of gene families and signatures of selection in the Australian marsupials*

The marsupials are an ancient and unique branch of the mammalian evolutionary tree. We share a common ancestor with them, existing around 160 million years ago. In the time since, this enigmatic group of species has followed an evolutionary path with huge differences to that of our own. Using a unique genomic resource partly developed by The Earlham Institute, we can finally start to explore the genetic basis of some of the specialised traits seen in the marsupials. With particular focus on the Koala, I will discuss recent innovations in genome sequencing, how the high-quality data we gain from them can be used in biology, and what can be learnt about evolution as a result.

Anne-Marie Keane, Quadram Institute, *Computing a yeast tree of life*

Phylogenetic analysis both informs our view of the divergence of species and develops a framework on which to view and exploit phenotypic information. The UK National Collection of Yeast Cultures (NCYC) consists of over 4,000 diverse strains, ideal for the construction of such a framework for yeast. Recent genome sequencing of ~1,000 NCYC strains has provided raw material for next-generation sequencing (NGS)-based tree estimation. Several NGS-based approaches to phylogenetic analysis have emerged in the past few years. One highly popular approach uses feature frequency profiles (Sims et al., 2009), essentially comparing frequency distributions of k-mers between genome pairs as a proxy for evolutionary divergence. This approach is simple to use but has shown some problems with computational efficiency and in taking biological features into account. Here, we present a comparison of multi-locus sequence, whole genome SNP, and NGS-based phylogenetic approaches, focussing on a well-studied set of 40 Saccharomyces complex strains. The success of the various approaches was
assessed by computational measures (e.g. Robinson Foulds distances, Mantel tests). Simulation studies were also used to assess the accuracy of the different phylogenetic methods. The results will inform future work aiming to develop new NGS-based approaches that incorporate additional biological knowledge.

Guillaume Scholz, UEA, *From phylogenetic networks to MUL-trees (and back)*

Phylogenetic networks are becoming of increasing interest to evolutionary biologists due to their ability to capture complex non-treelike evolutionary processes. From a combinatorial point of view, such networks are certain types of rooted directed acyclic graphs whose leaves are labelled by, for example, species. A number of mathematically interesting classes of phylogenetic networks are known. We present here the biologically relevant class of stable phylogenetic networks whose members are defined via certain “fold-up” and “un-fold” operations, that link them with certain type of rooted trees, known as MUL-tree. We then highlight some interesting properties of such networks with regard to the attractive notion of tree-displaying.

Andris Thompson, UEA, *Distributions of cherries and pitchforks in random phylogenetic trees*

Cherries and pitchforks are two of the simplest subtree structures found in phylogenetic trees. Using a recursive approach, we investigate the statistical distributions of these subtrees in random phylogenetic trees generated under two different models. We extend this work to unrooted trees, discuss applications to actual phylogenetic trees and beyond, and briefly explore possible extensions and generalisations.

Organisers: Katharina Huber and Andris Thompson, UEA