Polytech network form for PhD Research Grants from the China Scholarship Council

This document describes the PhD subject and supervisor proposed by the French Polytech network of 14 university engineering schools. Please contact the PhD supervisor by email or Skype for further information regarding your application.

Supervisor information	
Family name	Berry
First name	Vincent
Email	Vincent.berry@umontpellier.fr
Web reference	http://www.lirmm.fr/~vberry/
Lab name	LIRMM (Montpellier Laboratory of Informatics, Robotics and Microelectronics)
Lab web site	http://www.lirmm.fr
Polytech name	Polytech Montpellier
University name	Université de Montpellier
Country	France

PhD information	
Title	Computational methods to characterize inter(sub)specific rice genomes
Main topics regards to CSC list (3 topics at maximum)	V.18 Biological diversification III.1 Functionnal genome and proteinomist

Required skills in science and engineering	Ideally the student would have a Computer Science background (in particular skills in designing and analyzing algorithms) together with applied mathematics or statistics skills. Curiosity or appeal towards multi-disciplinary work is neeed.

Subject description (two pages maximum)

This PhD project issues from recent success stories of the research community in Montpellier (France) in producing crop reference genome sequences. These landmark achievements together with the current possibility for massive resequencing data production have opened new opportunities to understand the organization and dynamics of these genomes and the related keys to a more efficient exploitation of their diversity in breeding programs. However, this also brought up new challenges since the full exploitation of these data requires development of new biomathematic / bioinformatic concepts, methods and tools. We want to tackle these challenges thanks to the strength of Montpellier in computer science, mathematics and agricultural sciences.

The PhD will focus on aspects related to the frequent inter(sub)specific events involved in the history of crops and more precisely rice varieties. The student will be supervised in the development of methodologies to decipher the mosaic structure of crop genomes (these genomes have accumulated portions of different origins).

The methods that have been used so far to characterize the mosaic structure in plant genomes consisted in analyzing the distribution of simple summary statistics along the genome (Wu et al. 2014, Curk et al. 2014). Although they have proved useful in some biological models such approaches do not make optimal use of the information contained in the data, and therefore lack statistical power in species where information from putative parental genomes is scarce or even missing. Moreover, these approaches will undoubtedly be affected by sequencing errors and/or uncertainties in variant calling.

The mosaic structure in plant genomes can be characterized by a phylogenomic approach. The PhD subject aims at extending phylogenetic methods and at applying them on genome-size data. More precisely, the student will couple the inference of hybridization networks, reconciliations and ancestral gene adjacencies.

Hybridization networks are often used to describe and explain how a few founders shaped current genomes through reticulate events such as rounds of hybridization or recombination. Such networks describe precisely the chain of major events that lead to current genomes. Their recovery is an important step to understand the broad evolutionary patterns that lead to the composition of current and ancestral genomes. The team at work shares some expertise in the field (Huson et al. 2009, Huson

et al. 2010, Gambette et al. 2012, Huson et al. 2012, Kelk et al. 2012, El Baidouri et al. 2013, van Iersel et al. 2014). Such networks are currently inferred either from sequence data (that is, ordered character data, in which case the network is called an Ancestral Recombination Graph — ARG) (Gusfield 2014), or from a set of unordered gene trees (Huson et al. 2011).
Species tree / gene tree reconciliations is a technique often used to track the evolutionary history of genome fragments or gene sequences. Such methods allow to tag nodes and branches of gene phylogenies so as to infer gene duplications, losses and transfersthat can be used to model recombinations due to hybridization (Doyon et al. 2010, Doyon et al. 2011, El Baidouri et al. 2013, Nguyen et al. 2013a, Nguyen et al. 2013b, Scornavacca el al. 2013,
Scornavacca el al. 2015). Such events allow to explain the origin of current genome parts having identified homologues in other species or subspecies.
Last, the inference of ancestral gene adjacencies is a common technique to obtain a realistic picture of the mosaic of ancestral genomes from which current genomes evolved. Many methods exist that propose estimates of the gene order of ancient genomes (Braga et al. 2008, Chauve and Tannier 2008), yet most of them operate by translation and permutation operators on genome fragments, without accounting for the specific history of genome fragments or gene families. With several founders involved, accounting for such histories is a prerequisite to obtain the mosaic of ancestral genomes. Methods along this track just start to appear, being elaborated in part in Montpellier (Bérard et al. 2012) and Lyon (Patterson et al. 2013). These methods track current gene adjacencies backward in time being guided by species / gene tree reconciliations. It's also plausible that there will be a connection with works inferring synteny blocks in ancient genomes (see Lucas et al. 2014) for a preliminary work resorting on gene phlyogenies for two species only) or those relying on phylogenies to assemble contigs of fragments ancient DNA (Luhmann et al. 2014). Though the phylogenomic methods mentioned above pave the way to explain the mosaic of current and ancient genomes in general, they need further development in order to be applied to plants studied in this project. Methods for hybridization networks inference are still too slow to scale up to handle genome- size data as that considered here (e.g., the 3000 rice genome project). Moreover, reconciliation methods and gene adjacencies inference methods are only for species phylogenies that can be modeled as trees. They need to be extended to handle the case of species with multiple founders, such as polyploid plants considered in this project. Adapting these three approaches to our data and using them together will permit us to conceive a powerful framework to unravel mosaic structures of plant genomes in general.

This subject requires mainly skiils in designing and analyzing algorihms. No prior knowledge in bilogy is necessary (the needed background will be aquired during the first semester and the biological expertise provided by partners of the supervisor. A collaboration on this subject is ongoing with C. Scornavacca (ISEM) and JC Glaszmann (CIRAD).