C. R. Biologies 333 (2010) 577–581

Contents lists available at ScienceDirect

Comptes Rendus Biologies

www.sciencedirect.com

Concise review/Le point sur

Evolutionary genomics of C_4 photosynthesis in grasses requires a large species sampling

La génomique évolutive du caractère photosynthétique C_4 des graminées nécessite un large échantillonnage d'espèces

Guillaume Besnard [a,*], Pascal-Antoine Christin [b,c]

[a] *Imperial College London, Silwood Park Campus, Buckhurst Road, Ascot, Berkshire SL5 7PY, UK*
[b] *Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland*
[c] *Department of Ecology and Evolutionary Biology, Brown University, Providence, RI 02912 USA*

ARTICLE INFO

Article history:
Received 10 November 2009
Accepted after revision 29 March 2010
Available online 23 June 2010

Keywords:
Adaptive trait evolution
C_4 photosynthesis
Evolutionary genomics
Multigene family
Gene orthology
Poaceae

Mots clés :
Caractère adaptatif
Famille multigénique
Gène orthologie
Génomique évolutive
Photosynthèse C_4
Poaceae

ABSTRACT

Recent advances in genomics open promising opportunities to investigate adaptive trait evolution at the molecular level. However, the accuracy of comparative genomic studies strongly relies on the taxonomic coverage, which can be insufficient when based solely on a few completely sequenced genomes. In particular, when distantly-related genomes are compared, orthology of some genes can be misidentified and long branches of the phylogenetic reconstructions make inappropriate positive selection tests, as recently exemplified with investigations on the evolution of the C_4 photosynthetic pathway in grasses. Complementary studies addressing the diversification of multigene families in a broad taxonomic sample can help circumvent these issues.

RÉSUMÉ

Les récentes avancées dans le domaine de la génomique ont ouvert de nouvelles perspectives pour l'étude de l'évolution de caractères adaptatifs. Cependant, la précision des études de génomique comparative dépend de leur recouvrement taxonomique, qui peut être insuffisant lorsqu'il basé seulement sur quelques génomes complets. En particulier, lorsque des génomes phylogénétiquement distants sont comparés, l'orthologie de certains gènes peut être mal identifiée et les longues branches des reconstructions phylogénétiques sont peu appropriées pour des tests de sélection positive comme récemment illustré par des études sur les bases moléculaires de l'évolution de la photosynthèse C_4 chez les graminées. Dans ce cas de figure, des analyses complémentaires sur un échantillonnage approprié sont encore nécessaires pour mieux comprendre la diversification des familles multigéniques impliquées dans l'expression d'un caractère.

1. Introduction

Since the first sequencing of a whole eukaryote genome was completed (*Saccharomyces cerevisiae* in 1996 [1]), sequencing methods have greatly improved and costs have reduced significantly. This has opened the road to

---

* Corresponding author.
E-mail addresses: g.besnard@imperial.ac.uk (G. Besnard), Pascal-Antoine_Christin@brown.edu (P.-A. Christin).

1631-0691/$ – see front matter © 2010 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.crvi.2010.03.011

sequencing large genomes relatively quickly [2]. As a consequence, the number of organisms for which full genome information is available has vastly increased in the last decade. In plants, genomes of fourteen angiosperm species have already been completed (*Arabidopsis*, canola, grape vine, poplar, papaya, cucumber, rice, sorghum, maize, brachypodium, cassava, potato, soybean and African oil palm; on February 1st, 2010) and many additional genomes will be available soon (http://www.genomesonline.org). With predicted technological advances, the number of organisms completely sequenced is likely to grow exponentially [3]. This will open new avenues to comparative genomic analyses offering exceptional opportunities to better understand the mechanisms that shape genome organization [4], as well as shedding new light on the genetic mechanisms that led to the emergence of novel adaptations during evolutionary diversification [5]. In the following paragraphs, we detail the considerable value of full genome information for evolutionary studies in eucaryotes, and highlight the risks linked to the poor taxonomic coverage of full genomes, which will persist for a few years to come. Recent advances in the evolutionary genomics of the C_4 photosynthetic pathway in grasses [6,7] are discussed to illustrate the advantages and highlight some limits of molecular studies based solely on whole genomic data.

2. Complete genomes to study evolutionary novelties

Identifying the genetic changes linked to the emergence of adaptive novelties is an important challenge contributing to our deep understanding of the evolutionary processes at the molecular level [8,9]. Comparison between related organisms that exhibit different phenotypes can help identify the genetic changes responsible for a novel adaptive trait as well as some genetic features promoting its evolution [10,11]. For example, the impact of gene duplication and polyploidy on phenotypic diversification is an attractive topic that is better addressed by comparison of genome portions between related species [12]. Moreover, when genes involved in a trait have been previously identified, comparative approaches can give strong insights into the constraints on the recruitment of particular genes for the new function [7,13]. The quality and spectrum of the data on genes and genomes is a key factor determining the accuracy of comparative evolutionary approaches and the high amount of information provided by full genomes projects will transform comparative genetics into comparative genomics, a step that is necessary for an integrative understanding of evolutionary biology.

Comparative analyses of multigene gene families are strongly facilitated when full genomes are available [8]. First, complete gene sequences are directly accessible, including introns and non-coding flanking regions that often contain promoters, while the sequencing of complete genes on a large panel of species is often time-consuming using PCR-based cloning and can be challenging [14,15]. In addition, full genomes provide information that would be almost unattainable with other techniques. For instance, the exact genomic location of the studied genes can reveal that two paralogues lie on duplicated chromosomes or are tandemly repeated and thus helps reconstruct the genomic mechanisms linked to genetic diversity [16,17]. Finally, a precise knowledge of the number of genes that compose any multigene family almost necessarily requires complete genomes, since demonstrating that one gene lineage is absent from a non-model organism is difficult [18], particularly with PCR-based methods [14]. When merging the genomic information with functional and evolutionary approaches, an exhaustive picture can emerge, bringing our understanding of evolution to a level that was never reached before.

3. The case of C_4 photosynthesis in grasses

In plants, several of the most economical crops belong to Poaceae (or grass family) promoting intensive genetic and genomic studies in this family [19]. Poaceae is a worldly dominant family distributed in various environments from wet or dry tropical conditions to extremely cold habitats. The complete genome of four grass species, rice, sorghum, maize and brachypodium [16,20–22], is now available and others should be released in the next months (e.g., foxtail millet [19]), with a predicted burst of complete grass genomes in the coming years [19]. This high quantity of genomic data will be exceptional for a plant clade offering wonderful opportunities to understand evolution of traits at the molecular level. In particular, the multiplication of genomes will allow comparative analyses, shedding new lights on the molecular changes that gave rise to adaptive novelties, such as for developmental transitions to modulate flowering time or modify floral organ morphology [23,24], to change grain morphology [25], to develop new disease resistance [26] or photosynthetic adaptation, such as the C_4 trait in tropical conditions [27].

Sixty percents of C_4 species belong to the grass family (Poaceae), with several major crops, such as maize, sorghum or sugarcane [28]. The C_4 pathway consists of a set of morphological and biochemical modifications that together allow concentrating CO_2 around Rubisco and thus reducing photorespiration. The emergence of the C_4 traits is an evolutionary puzzle since the establishment of such a CO_2-pump has involved a high number of changes but occurred up to 18 times independently in grasses [29]. A key point to understand the evolution of this trait is that all enzymes involved in the C_4 pathway already exist in the C_3 ancestors, but are responsible for other functions [27]. In addition, the clustering of C_4 origins in some plant clades strongly suggests that these groups of organisms possess attributes that increase the probability of C_4 evolution [30]. C_4 facilitators should be searched for in genomic properties, such as the propensity of some C_3 lineages to create gene duplicates (particularly via polyploidisation) [27]. Besides theoretical works, genetic promoters of C_4 evolution remained out of reach until recently. While comparative analyses of multigene families encoding C_4 enzymes identified some changes in the protein sequences that are likely linked to C_4 evolution [31–33], the lack of genomic information hampered our understanding of the genome dynamics that led to genetic diversity of these gene