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).

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 1^{st}, 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 families. The recent release of sorghum genome [16], the first C_4 plant to be completely sequenced, removed many obstacles on the road to C_4 comparative genomics. A recent work by Wang et al. [6] used a comparative analysis of rice and sorghum genomes to test for the importance of gene duplications for C_4 evolution and the action of adaptive evolution during the acquisition of C_4-specific enzymes. These authors demonstrated that gene duplication (e.g., via whole genome duplication, tandem duplication or single gene duplication) was indeed an important step allowing evolution of most C_4 genes, although it was not involved in the evolution of all enzymes of the C_4 pathway (e.g., *nadp-mdh*). A long time lag between the availability of duplicates and the appearance of first C_4 grasses, together with different genesis of C_4 genes, also suggested that the transition process was very long before the establishment of fully C_4 plants [6]. These results are key improvements of our understanding of C_4 evolution and are a first step toward understanding the genetic factors linked to the recurrent evolutions of C_4 photosynthesis in grasses, although their scope can be limited by the small number of species compared.

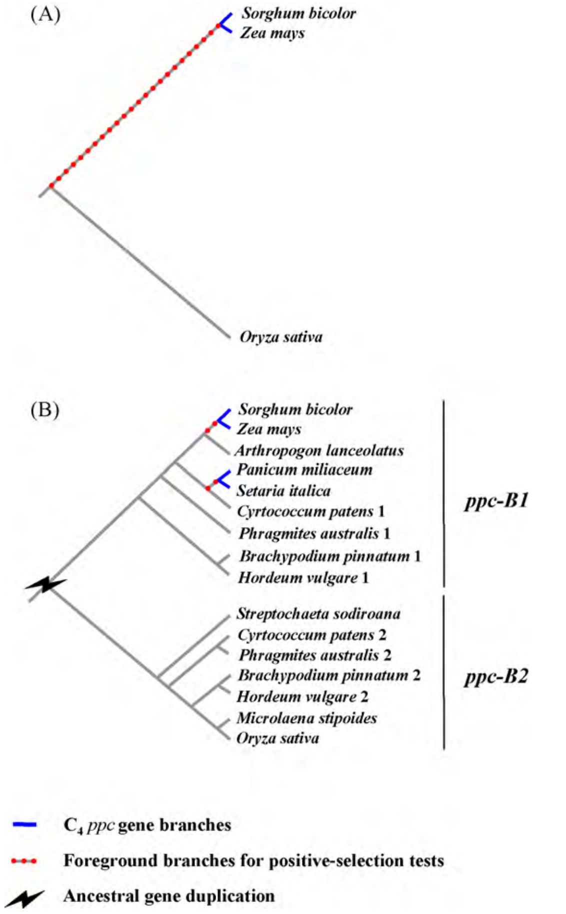

Fig. 1. Limited taxon sampling leading to erroneous statement of orthology and reducing the power of positive-selection tests. The case of the *ppc* gene family in grasses (i.e., gene duplicates *ppc-B1* and *ppc-B2*) is taken as an example [31,32]. **A.** First, when only complete genomes of rice, sorghum and maize are considered, one *ppc-B* copy is detected in each species. The gene of rice (Ehrhartoideae, C_3) is distantly related from the gene of sorghum and maize (Panicoideae, tribe Andropogoneae; C_4). As a consequence, the branch at the base of the C_4 specific gene, used for positive-selection tests, is long; **B.** Second, by analyzing *ppc-B* gene segments (generated by PCR) on a species sample covering different grass subfamilies, two gene clades (*ppc-B1* and *ppc-B2*) are identified in the phylogram. Gene duplication in the ancestor of these species is inferred from this topology (for a topology based on a more comprehensive species sampling see [31,32]). Gene *ppc-B2* was not isolated in some subfamilies such as Ehrhartoideae (e.g., *Oryza*), while gene *ppc-B1* was absent from numerous PACMAD species (in particular, *Sorghum* and *Zea*). This means that genes *ppc-B1* and *ppc-B2* were independently lost in different grass lineages. These observations were further confirmed by the presence of both genes in the complete genome of *Brachypodium distachyon* [22], as well as the distinct genomic locations (non-collinear regions) of *ppc-B1* and *ppc-B2* in rice and sorghum, respectively [16,32].

4. Toward an exhaustive taxon sampling

Nowadays, the low number of species completely sequenced limits the resolution of comparative genomics of C_4 photosynthesis. Rice, brachypodium and Andropogoneae (e.g., sorghum, maize) are only very distantly related and their most recent common ancestor dates back to more than 50 million years ago [29]. Sorghum and maize belong to the PACMAD clade and share a common C_4 ancestor, whereas rice and brachypodium belong to the sister BEP clade, which contains only C_3 species [29]. The recent genomic comparison of rice and Andropogoneae, two distantly related C_3 and C_4 taxa, can be problematic and is unlikely to accurately resolve the genetic mechanisms directly linked to C_4 evolution, since 50 million years of independent accumulation of genetic mutations can strongly blur any signal. For instance, the identification of orthologs between rice and sorghum-maize can be challenging, because independent losses of alternative homeologs could have occurred after gene duplication, as in the case of genes encoding the phospho*enol*pyruvate carboxylases [6,32]. Erroneous assessments of orthology can mislead interpretations regarding the number of gene duplications and their nature (Fig. 1). Moreover, the comparison of highly divergent genes can bias the estimations of past selective pressures [34]. This highlights the limits of comparative analyses based on a few whole genomes in reconstructing an accurate evolutionary history of genes responsible for the emergence of a novel adaptive trait. In the next years, the number of species to be compared will strongly increase, as tens of grass genomes should quickly become available [19]. Unfortunately, sampling of species to be sequenced was driven by economical interests and did not take into account grass diversity and evolutionary issues. In particular, all sequenced C_3 taxa belong to the exclusively C_3 BEP clade whereas the PACMAD clade is represented by C_4 species only [19], which will prevent a direct comparison of C_4 species with their C_3 sister taxa. Sequencing the whole genome of C_3 PACMAD species would definitively suppress problems associated with taxon sampling, but is not yet realistic due to the low economical and agronomical interests of such plants. An alternative is to set up dense comparative analyses of specific gene families, and full genome information of the model species are useful to design appropriate methodologies to sequence genes on non-model species and help understand the genomic context in which the studied genes lie.

In a recent study, such an approach was used to assess the genetic diversity of genes encoding NADP-malic enzyme (*nadpme*) in three model grasses (rice, sorghum and *Brachypodium distachyon*) [7]. Long fragments of *nadpme* were then sequenced from about 50 other grass species chosen to represent the different subfamilies and a variety of photosynthetic types. The joint analysis of genes extracted from full genomes and those isolated via PCR showed that four *nadpme* lineages appeared through recurrent gene duplications before grass diversification. The encoded enzyme of one of these lineages (*nadpme-IV*) acquired a plastid-specific localization through the acquisition of a first exon containing a transit peptide long before the different C_4 origins. Interestingly, this gene lineage became involved in C_4 photosynthesis at least five times independently, and it is strongly suggested that its plastid expression, which is necessary for the C_4 pathway, predisposed it for the C_4 function. On the other hand, the supposed absence of this *nadpme-IV* gene lineage in genomes of Chloridoideae may have prevented the evolution of the C_4 biochemical subtype based on NADP-malic enzyme in this large grass subfamily [7]. We are looking forward to the future release of additional C_4 grass genomes for exploring such hypotheses about C_4 evolutionary genetics.

5. Conclusion

While full genome sequencing projects bear great promises for evolutionary biology, we must keep in mind that the low taxonomic coverage they offer limits the scope of comparative genomics. In particular, the very long branches in the phylogenetic trees that include only genes from distantly related organisms can blur the signature of the past selection pressures. Similarly, the long evolutionary gap between completely sequenced organisms hampers causation between observed genetic differences and known phenotypic divergence. To maximize the impact of full genome projects, comparative analyses should be complemented by the sequencing of genes from non-model organisms of interest, to reduce the branch lengths in phylogenetic trees and obtain a taxon sampling suited for each research question. This can improve the accuracy of selection tests and, in the case of C_4 photosynthesis, already gave strong and novel insights into the genetic mechanisms linked to the recurrent origins of this complex and highly adaptive trait [6,7].

Acknowledgments

G.B. and P.A.C. were respectively funded by the Intra-European Fellowship PIEF-GA-2008-220813 and the Swiss National Science Foundation grant PBLAP3-129423.

References

[1] A. Goffeau, B.G. Barrell, H. Bussey, R.W. Davis, B. Dujon, H. Feldmann, et al., Life with 6000 genes, Science 274 (1996) 546–567.

[2] M. Margulies, M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, et al., Genome sequencing in microfabricated high-density picolitre reactors, Nature 437 (2005) 376–380.

[3] E.R. Mardis, Anticipating the $1,000 genome, Genome Biol. 7 (2006) 112.

[4] R.C. Hardison, Comparative genomics, PLoS Biol. 1 (2003) 156–160.

[5] D.H. Erwin, Early origin of the bilaterian developmental toolkit, Phil. Trans. Roy. Soc. B 364 (2009) 2253–2261.

[6] X. Wang, U. Gowik, H. Tang, J.E. Bowers, P. Westhoff, A.H. Paterson, Comparative genomic analysis of C_4 photosynthetic pathway evolution in grasses, Genome Biol. 10 (2009) R68.

[7] P.A. Christin, E. Samaritani, B. Petitpierre, N. Salamin, G. Besnard, Evolutionary insights on C_4 photosynthetic biochemical subtypes in grasses from genomics and phylogenetics, Genome Biol. Evol. 2009 (2009) 221–230.

[8] E.V. Koonin, Orthologs, paralogs, and evolutionary genomics, Annu. Rev. Genet. 39 (2005) 309–338.

[9] G.P. Wagner, V.J. Lynch, Evolutionary novelties, Curr. Biol. 20 (2010) R48–R52.

[10] H. Cai, R. Thompson, M.F. Budinich, J.R. Broadbent, J.L. Steele, Genome sequence and comparative genome analysis of *Lactobacillus casei*: insights into their niche-associated evolution, Genome Biol. Evol. 2009 (2009) 239–257.

[11] Y.H.E. Loh, L.S. Katz, M.C. Mims, T.D. Kocher, S.V. Yi, J.T. Streelman, Comparative analysis reveals signatures of differentiation amid genomic polymorphism in Lake Malawi cichlids, Genome Biol. 9 (2008) R113.

[12] L.E. Flagel, J.F. Wendel, Gene duplication and evolutionary novelty in plants, New Phytol. 183 (2009) 557–564.

[13] M. Zieman, M. Bhave, S. Zachgo, Origin and diversification of land plant CC-type glutaredoxins, Genome Biol. Evol. 2009 (2009) 265–277.

[14] A. Wagner, N. Blackstone, P. Cartwright, M. Dick, B. Misof, P. Snow, et al., Surveys of gene families using polymerase chain-reaction: PCR selection and PCR drift, Syst. Biol. 43 (1994) 250–261.

[15] R.L. Small, R.C. Cronn, J.F. Wendel, Use of nuclear genes for phylogeny reconstruction in plants, Austral. Syst. Bot. 17 (2004) 145–170.

[16] A.H. Paterson, J.E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood, H. Gundlach, et al., The *Sorghum bicolor* genome and the diversification of grasses, Nature 457 (2009) 551–556.

[17] J. Salse, M. Abrouck, S. Bolot, N. Guilhot, E. Courcelle, T. Faraut, et al., Reconstruction of monocotyledonous proto-chromosomes reveals faster evolution in plants than in animals, Proc. Natl. Acad. Sci. U S A 106 (2009) 14908–14913.

[18] O. Zhaxybayeva, C.L. Nesbø, W.F. Doolittle, Systematic overestimation of gene gain through false diagnosis of gene absence, Genome Biol. 8 (2007) 402.

[19] C.R. Buell, Poaceae genomes: Going from unattainable to becoming a model clade for comparative plant genomics, Plant Physiol. 149 (2009) 111–116.

[20] J. Yu, S.N. Hu, J. Wang, G.K.S. Wong, S.G. Li, B. Liu, et al., A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*), Science 296 (2002) 79–92.

[21] P.S. Schnable, D. Ware, R.S. Fulton, J.C. Stein, F. Wei, S. Pasternak, et al., The B73 maize genome: complexity, diversity, and dynamics, Science 326 (2009) 1112–1115.

[22] The International Brachypodium Initiative, Genome sequencing and analysis of the model grass *Brachypodium distachyon*, Nature 463 (2010) 763–768.

[23] J.C. Preston, E.A. Kellogg, Reconstructing the evolutionary history of paralogous *APETALA1/FRUITFULL*-like genes in grasses (Poaceae), Genetics 174 (2006) 421–437.

[24] D.E. Soltis, H. Ma, M.W. Frohlich, P.S. Soltis, V.A. Albert, D.A. Oppenheimer, et al., The floral genome: an evolutionary history of gene duplication and shifting patterns of gene expression, Trends Plant Sci. 12 (2007) 358–367.

[25] M. Charles, H.B. Tang, H. Belcram, A. Paterson, P. Gornicki, B. Chalhoub, Sixty million years in evolution of soft grain trait in grasses: emergence of the softness locus in the common ancestor of Pooideae and Ehrhartoideae, after the divergence from Panicoideae, Mol. Biol. Evol. 26 (2009) 1651–1661.

[26] A. Zamora, Q. Sun, M.T. Hamblin, C.F. Aquadro, S. Kresovich, Positively selected disease response orthologous gene sets in the cereals identified using *Sorghum bicolor* L. Moench expression profiles and comparative genomics, Mol. Biol. Evol. 26 (2009) 2015–2030.

[27] R.K. Monson, Gene duplication, neofunctionalization, and the evolution of C_4 photosynthesis, Int. J. Plant Sci. 164 (2003) S43–S54.

[28] C.P. Osborne, D.J. Beerling, Nature's green revolution: the remarkable evolutionary rise of C_4 plants, Phil. Trans. Roy. Soc. B 361 (2006) 173–194.

[29] P.A. Christin, G. Besnard, E. Samaritani, M.R. Duvall, T.R. Hodkinson, V. Savolainen, N. Salamin, Oligocene CO_2 decline promoted C_4 photosynthesis in grasses, Curr. Biol. 18 (2008) 37–43.

[30] R.F. Sage, Environmental and evolutionary preconditions for the origin and diversification of the C_4 photosynthetic syndrome, Plant Biol. 3 (2001) 202–213.

[31] P.A. Christin, N. Salamin, V. Savolainen, M.R. Duvall, G. Besnard, C_4 photosynthesis evolved in grasses via parallel adaptive genetic changes, Curr. Biol. 17 (2007) 1241–1247.

[32] P.A. Christin, G. Besnard, Two independent C_4 origins in Aristidoideae (Poaceae) revealed by the recruitment of distinct phosphoenolpyruvate carboxylase genes, Am. J. Bot. 96 (2009) 2234–2239.

[33] G. Besnard, A.M. Muasya, F. Russier, E.H. Roalson, N. Salamin, P.A. Christin, Phylogenomics of C_4 photosynthesis in sedges (Cyperaceae): multiple appearances and genetic convergence, Mol. Biol. Evol. 26 (2009) 1909–1919.

[34] M. Anisimova, Z.H. Yang, Multiple hypothesis testing to detect lineages under positive selection that affects only a few site, Mol. Biol. Evol. 24 (2007) 1219–1228.