• Quantification of the degree to which transcript abundance differs between C\textsubscript{3} and C\textsubscript{4} leaves
• Identification of novel components of C\textsubscript{4} metabolism
• Intersection with mathematical models to explain evolution of the complex C\textsubscript{4} phenotype
• Indication that C\textsubscript{4} photosynthesis is underpinned by both convergent and parallel evolution of structural genes and also regulators
Insights into C₄ metabolism from comparative deep sequencing

Steven J. Burgess and Julian M. Hibberd

Department of Plant Sciences, Downing Street, University of Cambridge, Cambridge, CB2 3EA, UK

Email: SJB: sjb287@cam.ac.uk, JMH: jmh65@cam.ac.uk

For correspondence; email - jmh65@cam.ac.uk
Abstract

C₄ photosynthesis suppresses the oxygenation activity of Ribulose Bisphosphate Carboxylase Oxygenase and so limits photorespiration. Although highly complex, it is estimated to have evolved in sixty-six plant lineages, with the vast majority lacking sequenced genomes. Transcriptomics has recently initiated assessments of the degree to which transcript abundance differs between C₃ and C₄ leaves, identified novel components of C₄ metabolism, and also led to mathematical models explaining the repeated evolution of this complex phenotype. Evidence is accumulating that this complex and convergent phenotype is partly underpinned by parallel evolution of structural genes, but also regulatory elements in both cis and trans. Furthermore, it appears that initial events associated with acquisition of C₄ traits likely represent evolutionary exaptations related to non-photosynthetic processes.
**Introduction**

$C_3$ plants inherited a carbon fixation system developed by the photosynthetic bacteria, with primary carbon fixation being catalysed by the enzyme Ribulose Bisphosphate Carboxylase Oxygenase (RuBisCO). The oxygenase activity of RuBisCO generates the toxic intermediate phosphoglycollate, and although this can be detoxified and carbon partially recovered by the photorespiratory pathway, energy is expended in the process. As the oxygenase function of RuBisCO increases with ambient temperature, it is thought that in tropical and sub-tropical habitats, significant selection pressure led to the convergent evolution of carbon concentrating mechanisms [1]. Phylogeny indicates that land plants have repeatedly evolved either temporal (Crassulacean Acid Metabolism) or spatial carbon concentrating mechanisms ($C_4$ photosynthesis) [2].

Although highly complex, the $C_4$ pathway is estimated to have evolved in at least sixty-six lineages of plants [3]. Initial analysis of clades that contain $C_3$ and $C_4$ species but also ‘$C_3$-$C_4$’ intermediates identified the most common early traits likely associated $C_4$ photosynthesis, and this led to the development of models that depict the evolution of this complex phenotype along a relatively linear path of trait acquisition [4]. More recently, probabilistic modelling within a Bayesian framework identified flexibility in when $C_4$ component traits evolve, but also found four major paths likely associated with acquisition of these traits [5]. Despite this flexibility in the acquisition of $C_4$ component traits, the core $C_4$ metabolic machinery has converged upon a similar architecture in all $C_4$ lineages. For example, in all $C_4$ species, $HCO_3^-$ is initially fixed by phosphoenolpyruvate carboxylase (PEPC) (Figure 1), which has a higher affinity for $HCO_3^-$ than RuBisCO does for $CO_2$ [6]. $C_4$ acids then diffuse down a concentration gradient into insulated cellular, or sub-cellular [7] compartments where $C_4$ acid decarboxylases increase the local concentration of $CO_2$ around RuBisCO, thereby reducing its oxygenation activity. In most $C_4$ species, an altered arrangement of cells within the leaf known as Kranz anatomy facilitates the compartmentation of carboxylation and decarboxylation (Figure 1A). There are three basic biochemical pathways defined by the predominant $C_4$ acid decarboxylase that releases $CO_2$ around RuBisCO, but there are also at least 25 forms of Kranz anatomy documented (Figure 1A and 1B).

Progress in understanding $C_4$ leaf anatomy has recently been critically assessed [8]. Here we focus on how deep sequencing is influencing our understanding of $C_4$ biochemistry and argue that combined with allied technologies it is opening up a new era of $C_4$ research. These approaches are helpful for at least three reasons. First, many years of mutant screens, biochemistry and molecular biology have so far failed to unlock many of the molecular components that regulate or induce the $C_4$ system [9,10], sequencing offers the opportunity to identify candidate genes for these traits. Second, the $C_4$ pathway should correctly be viewed as a system. Deep sequencing now makes it possible to move from analysis of individual genes and their gene products, to assessing the simultaneous behaviours of both the system and its components. Third, computational advances that have been driven by deep sequencing datasets provide the opportunity to study the natural diversity of all $C_4$ lineages, rather than being limited to well-studied ‘model’ species for
which genome sequence is available. With this as background, we now assess how deep sequencing has influenced the understanding of core components of the C₄ cycle, the \textit{trans}-factors likely responsible for their compartmentation between mesophyll (M) and bundle sheath (BS) cells of the C₄ leaf, and also the evolutionary processes that have governed the transition from the ancestral C₃ photosynthetic system to the derived C₄ metabolic pathway.
Defining mRNAs associated with C₄ photosynthesis

Approximately forty genes encoding core C₄ cycle enzymes and components of the Calvin-Benson-Bassham cycle (CBB) have long been known to be involved in C₄ metabolism. RNA-seq has been used to report mRNA signatures associated with the ‘NAD-ME’, ‘NADP-ME’ or ‘PEPCK’ biochemical sub-types [11–13], and along with theoretical and modelling approaches, has provided clear evidence that often two of the decarboxylases operate in parallel, with their relative contributions varying depending on conditions [14–17]. However, our understanding of what changes leaf anatomy such that contact between tissues involved in carbon assimilation and reduction is increased (Figure 1A), and also what sets up and then maintains the patterns of gene expression required for the C₄ cycle are rudimentary. These factors are important, as an understanding of C₄ genetics has implications for strategies being adopted to engineer the pathway into C₃ crop species, dictating whether efforts should be focused on alterations to individual genes, transcriptional regulators or hormone metabolism and signalling. Deep sequencing has allowed estimates of the extent to which global patterns of mRNA abundance differ between C₃ and C₄ leaves. This approach was initiated in the Cleomaceae, which in addition to containing C₃ and C₄ species, is phylogenetically the closest-C₄-containing clade to C₃ Arabidopsis thaliana [18]. 603 genes showed differential mRNA abundance in C₄ compared with C₃ leaves [12]. Furthermore, in addition to confirmation that mRNAs encoding core C₄ and CBB cycles were up and down-regulated respectively, previously unidentified characteristics of the C₄ leaf as well as new components of the C₄ cycle were reported. For example, reduced abundance of mRNAs encoding ribosomal sub-units in C₄ compared with C₃ leaves was reported [12], while BASS2, which was subsequently shown to encode the long-sought-after pyruvate transporter associated with C₄ photosynthesis was up-regulated [19]. Subsequent analysis has led to increased numbers of genes being linked to the C₄ cycle [13] and Table 1. The highest reported differences in transcript abundance between C₃ and C₄ tissues are derived from Eleocharis, a species that is able to switch from C₃ to C₄ depending on whether it is aquatic or terrestrial (Table 1). However, a proportion of the mRNAs reported to be differentially abundant in C₄ compared with C₃ Eleocharis are likely associated with the different light and temperature conditions caused by the aquatic to terrestrial switch [20].

Comparison of estimates of the number of changes associated with each of the three biochemical sub-types (Figure 1) led to suggestions that establishment of the PEPCK C₄ sub-type requires the fewest changes, in part because of reduced requirements for alterations in photosystem accumulation between mesophyll and bundle sheath cells [11]. An overview of statistics from these studies (Table 1) shows that as sequencing depths have increased there has been an increase in the predicted number of differentially expressed genes, likely due to better quantification of low abundance transcripts. However, as no annotated genomes were available for these species, the data are based either on cross-species mapping of reads, or gene models created by de novo transcriptome assembly [21–23]. Both of these approaches introduce inaccuracy compared with direct read mapping to a well-annotated genome. It is important to note that the absolute number of differentially expressed genes detected through congeneric comparisons...
is clearly dependent on the phylogenetic distance, statistical cut-offs, quality of transcriptome assemblies and number of species sampled (Table 1). As the number of independent C₄ lineages that are assessed with RNA-seq increase, estimates of the conserved alterations to mRNA abundance will become more reliable. However, it is clear from the current estimates which range from hundreds to thousands of genes showing differential expression in C₄ compared with C₃ leaves, research needs to focus on identification of key transcription factors and signalling events that underlie these patterns of gene expression.

**Compartmentation of gene expression between cell-types of the C₄ leaf**

As with analysis of any organ or tissue, the C₄ leaf is composed of multiple distinct cell types, and the specialisation of M and BS cells in C₄ leaves (Figure 1) is considered a hallmark of the C₄ pathway. The first publications on global mRNA populations of M and BS cells of C₄ leaves were conducted on maize and supported existing knowledge of genes known to be differentially expressed between these cell types [24,25]. Analysis of two independent C₄ lineages from the grasses indicated that the absolute abundance of mRNAs in M and BS cells of grasses that evolved C₄ photosynthesis independently was statistically more convergent than other differentially expressed genes [26]. This implies that strong selection pressures acted on genes associated with the C₄ pathway to generate very similar expression in separate C₄ lineages. As the M and BS transcriptomes of more C₄ species become available this quantitative convergence could be used to generate a predictive framework that allows unknown components of C₄ photosynthesis to be identified. Although it has long been clear that transcriptional, post-transcriptional and post-translational processes all play a part in generating the C₄ metabolic system [9], omics approaches are now initiating non-biased and systems level quantification of their importance. For example, quantitative proteomics and transcriptomics indicated that the ratio of each cognate protein to its mRNA varies during C₄ leaf development, and that the ratio is often highest where protein function is most relevant [27]. Taken together, these findings start to provide an oversight of the extent of post-transcriptional and post-translational regulation in the C₄ leaf.

Transcriptomic datasets derived from M and BS cells of C₄ leaves highlight an area of ignorance, namely the mRNA populations associated with these two cell types in leaves of ancestral C₃ plants. Without this information it has not been possible to define how much patterns of gene expression have altered in M and BS cells of C₄ compared with those cells in C₃ leaves. A major hurdle was our inability to isolate M and BS cells from C₃ leaves, however immunopurification of ribosomes from specific cell types [28] has initiated our understanding of the BS in C₃ *Arabidopsis thaliana*. Although it was previously known that veinal cells of C₃ plants possessed characteristics of C₄ photosynthesis [30,31], ribosome tagging and deep sequencing of associated mRNAs indicated that components of the C₄ cycle are also preferentially expressed in the C₃ BS [29]. This work also highlighted a role for the C₃ BS in sulphur metabolism, a characteristic that had previously been reported of the C₄ BS [32]. Thus, as more C₃ lineages are sampled, we will develop a much clearer understanding of the extent to which metabolic characteristics currently associated with C₄
photosynthesis are actually ancestral and present in either M or BS cells of C₃ leaves. We therefore conclude that technologies are in place to significantly improve our understanding of M and BS cells in both C₃ and C₄ plants. Data from these approaches are being used to formulate models that relate to the molecular drivers associated with the repeated evolution of this complex trait, and it is this that will be explored in the next section.

Insights into the molecular drivers of C₄ metabolism

It has been clear for some time that prior to their recruitment into C₄ photosynthesis, the major proteins of C₄ photosynthesis typically accumulate at relatively low levels in a constitutive manner in C₃ leaves [33]. Through comparison with a gene expression atlas of closely related species, it is now proposed that expression of orthologues to C₄ genes show a variety of expression patterns, and peak in various tissues, in the C₃ ancestral system [34]. Deep sequencing data has also now provided the insight into the extent to which genes of the C₄ cycle become co-regulated with photosynthesis genes in leaves of both C₄ monocotyledons and dicotyledons [23,35]. Overall, these data imply that during the evolution of C₄ photosynthesis, genes of the C₄ cycle are co-opted into the gene regulatory networks that govern photosynthesis gene expression in the ancestral C₃ state [23,34].

The identification of transcription factors responsible for these alterations in expression of genes encoding components of the C₄ cycle is an area where significant progress still needs to be made. However, comparative transcriptomics has now identified candidate regulators for the C₄ cycle in maize [24,25,35–37], Setaria [26,38], Flaveria [13] and Gynandropsis gynandra (formerly known as Cleome gynandra) [23,34]. Interestingly, independent lineages of C₄ plants appear to have up-regulated homologous transcriptional regulators in either M or BS cells. This has been reported for two independent lineages of C₄ grasses [26] but also for the C₄ dicotyledon G. gynandra and the C₄ monocotyledon maize [23]. These data indicated that M or BS preferential expression is not only associated with parallel evolution of regulatory DNA [39] and histone marks [40], but also the recruitment of transcription factors [23,26].

Another striking finding facilitated by deep sequencing has been quantification of the extent to which specific members of multi-gene families are recruited into the C₄ pathway. This was initially reported after phylogenetic reconstructions of individual genes such as PEPC [41], but the extent of this process was not clear. Transcriptomics has now quantified this phenomenon in Alloteropsis, which contains C₃ and C₄ subspecies [42]. In maize and Setaria, which represent two independent lineages of C₄ grass, 87% of C₄ cycle proteins that are up-regulated in C₄ leaves are syntenic orthologues, indicating that the same ancestral gene has repeatedly been recruited into the pathway [26]. Again, the mechanism behind this phenomenon is not clear, but it is possible that these orthologues are repeatedly used into the C₄ pathway because they are part of pre-existing gene regulatory networks that are recruited into C₄ photosynthesis. These data further emphasize that the highly complex C₄ photosynthesis trait is underpinned by a mixture of both convergent and parallel evolution [39,42].
The combination of deep sequencing and metabolic flux modelling has demonstrated the power of an integrated approach, and lead to an enticing hypothesis concerning the repeated evolution of C₄ photosynthesis. Comparing C₃, C₃-C₄ and C₄ species in *Flaveria*, RNA-seq data coupled to metabolic modelling predicted that loss of the full photorespiratory pathway in the M cells of C₃ plants, which is the most common biochemical alteration thought to initiate C₄ evolution [2], leads to a nitrogen imbalance between M and BS cells [43] (Figure 2). The most parsimonious alterations to central metabolism that corrects this imbalance in the leaf is to induce, and compartment, the key components of the C₄ cycle into either M or BS cells (Figure 2). These data strongly imply that the metabolic remodelling during these early stages of C₄ evolution represent an evolutionary exaptation that was initially not related to photosynthetic efficiency *per se*. Thus, it now appears that metabolic and also morphological alterations to C₃ leaves were both unrelated to photosynthesis [5,42,44]. Later in the evolutionary process it is thought that each alteration to the C₄ cycle leads to a steady increase in photosynthetic performance [45], and this is then followed by evolutionary fine-tuning mediated by amino acid substitutions that modify allosteric regulation of these proteins for the C₄ leaf [46]. In the future, deep sequencing will also allow us to determine whether parallel changes to amino acids are associated with parallel or convergent evolution to the nucleotides encoding them. Moving ahead, perhaps a similar combined modelling, sequencing and hormone approach is required to make progress in understanding the molecular basis of Kranz anatomy.

**Summary**

The use of deep sequencing in C₄ research is in its infancy, and so far is mostly limited to RNA-seq. It is also true that the initial phase has identified many genes that could be important for C₄ photosynthesis, but for which functional analysis has not yet been undertaken. However, it is clear that use of deep sequencing has initiated an unbiased and objective study of C₄ photosynthesis in species that previously lacked any transcriptomic or genomic resources. As outlined above, deep sequencing and improved computational pipelines for data analyses have started to provide significant new insight. This includes defining core components of the C₄ cycle, identifying variations in C₄ metabolism both within and between species, and also providing inference into evolutionary mechanisms associated with the polyphyletic appearance of this highly complex system.
Acknowledgements

We thank the BBSRC for grant BB/1002243/1 and the EU 3to4 program for financial support.


++ The initial use of deep sequencing of closely related C₃ and C₄ plants to provide quantitative insight into the degree that their leaf transcriptomes differ, and identifying new candidate proteins important for the C₄ cycle.


++Functional analysis showing that a gene identified after deep sequencing of closely related C₃ and C₄ species encoded an elusive pyruvate transporter.


++ Detailed analysis of samples from closely related C₃ and C₄ plants providing key insights into genes recruited into C₄ photosynthesis.


++ Deep sequencing of C₃ and C₄ grasses combined with an algorithm to normalise between leaves of different growth habits.


++RNAseq used in conjunction with flux modeling to provide amazing insight into the evolution of C₄ photosynthesis.


Figure Legends

Figure 1: Schematics illustrating variation in leaf anatomy and C₄ biochemical cycles of C₄ leaves. A. Diagrams representing transverse sections through a C₃ leaf, and four anatomical variations in Kranz anatomy. Images are based on those reported by [47]. B. The three main cycles that have classically been used to define the three biochemical sub-types of C₄ photosynthesis. AlaAT = Alanine aminotransferase, AspAT = Aspartate aminotransferase, CA = Carbonic anhydrase, PEPC = Phosphoenolpyruvate carboxylase, PEPCK = Phosphoenolpyruvate carboxykinase, NADP-MDH = NADP-dependent malate dehydrogenase, NADP-ME = NADP-dependent malic enzyme, NAD-ME = NAD-dependent malic enzyme, PPDK = Pyruvate,orthophosphate dikinase, CBB = Calvin Benson Bassham cycle, Ala = alanine, Asp = aspartate, Mal = malate, OAA = oxaloacetic acid, Pyr = Pyruvate, PEP = phosphoenolpyruvate.

Figure 2: Impacts of deep sequencing on understanding C₄ metabolism. Representation of model predicting initial events associated with the evolution of C₃-C₄ intermediacy (based on [43]). Loss of photorespiration in the mesophyll cells would lead to an imbalance in nitrogen metabolism between mesophyll and bundle sheath cells, and accumulation of ammonia (yellow circle) in the bundle sheath. Upregulation of a C₄-like pathway rebalances this nitrogen imbalance. The three panels represent photorespiration (C₂ cycle) operating in both mesophyll and bundle sheath cells of a C₃ leaf (A), the C₂ cycle being lost in the mesophyll cells of C₃-C₄ intermediate species, and the subsequent development of a C₄-like cycle (B), and finally complete implementation of the C₄ cycle (C). Abbreviations as in Figure 1, as well as Glu = glutamate, Gly = glycine, 2-OG = 2-oxoglutarate, Ser = serine. Dashed lines indicate low metabolic flux. Red circles represent carbon atoms while yellow circles represent amine groups.
**C3**

**C4 atriplicoid**

**C4 kochioid**

**C4 salsoloid**

**C4 suaedoid**

*Note: In C3 species both carbon assimilation and reduction predominately occur in the mesophyll.*
<table>
<thead>
<tr>
<th></th>
<th>Bräutigam et al. (2011) 12</th>
<th>Gowik et al. (2011) 13</th>
<th>Bräutigam et al. (2014) 11</th>
<th>Chen et al. (2014) 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of DE transcripts</td>
<td>603</td>
<td>3582</td>
<td>1168</td>
<td>8848</td>
</tr>
<tr>
<td>Transcripts more abundant in C₃</td>
<td>258</td>
<td>1418</td>
<td>792</td>
<td>4184</td>
</tr>
<tr>
<td>Transcripts more abundant in C₄</td>
<td>345</td>
<td>2164</td>
<td>376</td>
<td>4664</td>
</tr>
<tr>
<td>% Transcriptome DE</td>
<td>1.4</td>
<td>NA*</td>
<td>6.1</td>
<td>13.5</td>
</tr>
</tbody>
</table>

**Table 1: Comparisons of transcript abundance in closely related C₃ versus C₄ photosynthetic tissues.** The total number of transcripts annotated as being differentially expressed (DE) in each study is listed, along with the numbers up or down regulated. Data expressed as percent of the total transcriptome are also reported for each study. Bräutigam et al. 2011 assessed C₄ Gynandropsis gynandra versus C₃ Tareneya hassleriana. Gowik et al. 2011 assessed C₄ Flaveria bidentis and Flaveria trinervia as well as C₃-C₄ Flaveria ramosissima and C₃ Flaveria pringlei and Flaveria robusta. Bräutigam et al. 2014 assessed Panicum maximum and Diancanthelium clandestinum. Chen et al. 2014 assessed C₄ and C₃ culms of Eleocharis baldwinii. *NA: the values for DE transcripts were based on multispecies comparisons which prohibits expressing the number of DE transcripts as a percentage of transcriptome.