Maize Cytolines Unmask Key Nuclear Genes That Are under the Control of Retrograde Signaling Pathways in Plants

Mihai Miclaus[1,2,*], Ovidiu Balacescu[3,4], Ioan Has[5], Loredana Balacescu[3,4], Voichita Has[5], Dana Suteu[1], Samuel Neuenschwander[2,6], Irene Keller[2,7], and Rémy Bruggmann[2,*]

[1]National Institute of Research and Development for Biological Sciences, Cluj-Napoca, Romania

[2]Interfaculty Bioinformatics Unit and Swiss Institute of Bioinformatics, University of Bern, Bern, Switzerland

[3]The Oncology Institute "Prof Dr Ion Chiricuta", Cluj-Napoca, Romania

[4]Iuliu Hatieganu University of Medicine and Pharmacy, Cluj-Napoca, Romania

[5]Agricultural Research and Development Station, Turda, Romania

[6]Vital-IT, Swiss Institute of Bioinformatics, University of Lausanne, Lausanne, Switzerland

[7]Department of Clinical Research, University of Bern, Bern, Switzerland

*Corresponding authors: E-mails: mihai.miclaus@icbcluj.ro; remy.bruggmann@bioinformatics.unibe.ch.

Accepted: September 27, 2016

Data deposition: This project has been deposited at the European Nucleotide Archive (ENA) under the accession number PRJEB11473, and Gene Expression Omnibus (GEO) under the accession number GSE74096.

Abstract

The genomes of the two plant organelles encode for a relatively small number of proteins. Thus, nuclear genes encode the vast majority of their proteome. Organelle-to-nucleus communication takes place through retrograde signaling (RS) pathways. Signals relayed through RS pathways have an impact on nuclear gene expression but their target-genes remain elusive in a normal state of the cell (considering that only mutants and stress have been used so far). Here, we use maize cytolines as an alternative. The nucleus of a donor line was transferred into two other cytoplasmic environments through at least nine back-crosses, in a time-span of $>$ 10 years. The transcriptomes of the resulting cytolines were sequenced and compared. There are 96 differentially regulated nuclear genes in two cytoplasm-donor lines when compared with their nucleus-donor. They are expressed throughout plant development, in various tissues and organs. One-third of the 96 proteins have a human homolog, stressing their potential role in mitochondrial RS. We also identified syntenic orthologous genes in four other grasses and homologous genes in *Arabidopsis thaliana*. These findings contribute to the paradigm we use to describe the RS in plants. The 96 nuclear genes identified here are not differentially regulated as a result of mutation, or any kind of stress. They are rather key players of the organelle-to-nucleus communication in a normal state of the cell.

Key words: retrograde signaling, nuclear gene expression, cytolines, maize transcriptome, bioinformatics.

Introduction

Plants have an important energy-converting organelle besides the mitochondrion: the plastid. Both organelles contain a small number of genes in their genomes: 120 in the plastid and 57 in the mitochondria, respectively (Sugita and Sugiura 1996; Unseld et al. 1997). But the organellar genomes outnumber the nuclear genome by as much as 5,000 to 1 (Bendich 1987; Cavelier et al. 2000). Despite this ratio, it was originally thought that organellar DNA was highly conserved compared with its nuclear counterpart and therefore, that any phenotypic variation was mainly due to the latter (Wolfe et al. 1987). This view is currently changing with the aid of new technologies (e.g., next generation sequencing—NGS), which offer the possibility of transcriptome-wide gene expression and comparative analyses. In this respect, Moison et al. (2010) sequenced plastid and mitochondrial genomes in 95 accessions of *Arabidopsis* and concluded that there was considerable genetic polymorphism in both organelles. Furthermore, there is a significant body of evidence showing that cytoplasm–nucleus interaction is important in explaining phenotypic variation in many different species, like rice, mouse, yeast or *Drosophila* (Roubertoux et al. 2003;

Tao et al. 2004; Rand 2005; Dimitrov et al. 2009). Retrograde pathways have been defined in order to describe the existing cross-talk between the organelles and the nucleus, and to understand how nuclear gene expression (NGE) is modulated according to signals received from both organelles (Gray et al. 2003; Rhoads and Subbaiah 2007; Chi et al. 2013).

Mutants that are defective in the retrograde signaling (RS) pathways have been extensively used in trying to untangle how organelles control NGE (reviewed in Jung and Chory 2010). But mutants have limited potential to explain this phenomenon in a broader context. This is because the genes identified to respond to RS are linked to a single original stimulus, which is a mutation and all the expression changes downstream are the result of it. Therefore it still remains unclear if these genes are the only targets of the retrograde pathways or other adjustments come into play in a nonmutant cell environment. These other adjustments might change the paradigm we use to explain organelle-to-nucleus communication.

To circumvent the problem of using the less informative mutants, Joseph et al. (2013) recently took a metabolomics approach using the reciprocal Kas×Tsu *Arabidopsis* recombinant inbred lines (RILs) population to investigate the effect of the cytoplasm on cell metabolites. They concluded that 80% of the metabolites are controlled by the cytoplasmic genome. Thus, it is clear that the cytoplasmic genetic make-up of the cell plays an important role in the functioning of nuclear loci, but no key players responsible for the observed effect have been identified yet.

Cytolines represent a better model to study the effect of cytoplasm (including its organellar genomes) on NGE than RILs in *Arabidopsis*. By repeated backcrossing one can transfer the nucleus from a donor line (used as male/pollen donor) to several other cytoplasms, thus creating isonuclear lines, or cytolines. The plastids and mitochondria present in the resulting lines are only of maternal origin, as they are not transmitted by pollen in most angiosperms and all studied grasses (Conde et al. 1979; Soliman et al. 1987). Maize (*Zea mays* ssp. *mays*), like other cereal crops, is an ideal candidate for the creation of such cytolines. It has an easy pollination process and also has visible and easily measurable phenotypes, which may vary depending on the cytoplasm. For these reasons, Allen (2005) used one maize inbred line and back-crossed it repeatedly into various cytoplasms of teosinte (*Zea mays* ssp. *parviglumis*; maize's ancestor) and observed the phenotypes of the resulting cytolines. He concluded that cytoplasmic genomes have a significant effect on morphological, developmental, and functional characters. These results were based on empirical phenotypic observations. No molecular approach has been implemented to understand the changes taking place in the newly created lines at the gene expression level. Tang et al. (2013) took a step forward and analyzed the molecular background (not gene expression) of several maize cytolines that displayed significant phenotypic differences when compared with their donor lines. The authors were interested in two important traits for maize breeding: plant height and ear height. Using 154 microsatellite markers, 22 quantitative trait loci (QTLs) were identified through simple sequence repeat (SSR) mapping, which may contain genes directly under the control of the cytoplasmic genomes (chloroplast or mitochondria).

Interactions between cytoplasmic and nuclear genomes also influence agronomic traits of rice, but no specific loci have been identified either (Tao et al. 2004). In a recent study, Crosatti et al. (2013) transferred the wheat nucleus (*Triticum aestivum*) into two other species to examine NGE in depth using microarray technology. About 540 nuclear genes were found to have a significantly altered expression pattern when the wheat nucleus was transferred into *Hordeum chilense* cytoplasm, whereas only 11 and 28 genes significantly changed their expression in transfers to cytoplasm from *Aegilops uniaristata* and *Aegilops tauschii*, respectively.

To sum up, all studies above present significant evidence that organellar genomes are involved in controlling NGE, but mainly mutants have been used so far in untangling the RS pathways in various organisms. Maize is a model plant well suited for studies of NGE using cytolines, which circumvent the shortcomings of mutants. No study has used NGS technology to sequence the whole transcriptome in such lines. They have been subjected, instead, to phenotypic, microarray, and metabolomics analyses (Tao et al. 2004; Allen 2005; Crosatti et al. 2013; Joseph et al. 2013).

Here, we transferred the nucleus of a donor line into two other cytoplasmic environments of the same species (*Zea mays* ssp. *mays*), through at least nine back-crosses, thus creating three cytolines. Their transcriptome was sequenced using an Illumina HiSeq2500 instrument and the data validated using a custom-made microarray chip. We identified 96 nuclear genes that could potentially function as targets of the RS pathways. More importantly, these genes are not differentially regulated as a result of mutation or any kind of stress. They are rather key players of the organelles-to-nucleus communication in a normal state of the cell. We also identified syntenic genes in four other grasses and homologous genes in *Arabidopsis thaliana*, hinting towards a general mechanism in plants, where the RS pathways target these key nuclear genes.

Material and Methods

Plant Material and RNA Extraction

Inbred line TC208 was back-crossed ten times (as pollen donor) to inbred lines TC316 and W633, respectively, during 1992–2004. The resulting cytolines (TC208(cytTC316), TC208(cytW633), and TC208) have been maintained by self- or sib-cross ever since. The three original inbred lines had been created through at least ten self-crosses. All lines

used in the present study are fertile, including Cyt1 and Cyt2. There are no statistically significant phenotypic differences between the three cytolines. Seeds were sowed in the greenhouse and kept under natural light (9.5 h daylength). Nine-day-old seedlings were sampled at noon from all three lines being immediately frozen in liquid nitrogen. Total RNA was isolated with TriReagent (Sigma–Aldrich). Three biological replicates for each genotype were used. Total RNAs were further purified with RNeasy Mini Kit (Qiagen) and their quality was evaluated with Bioanalyzer 2100 (Agilent Technologies) based on RNA integrity. RIN (RNA integrity number) was ≥ 8 for all samples.

Microarray Assay and Data Analysis

Cy-3 labeled microarray probes (cRNA-Cy3) were synthesized from 200 ng of total RNA using one-color Low Input Quick Amp Labeling Kit, according to Agilent manufacturer's protocol. The quality of synthesized cRNAs was checked with Nanodrop ND-1000 spectrophotometer (NanoDrop Technologies) considering a minimal yield of 1.65 μg and a specific activity >6 pmol/μl Cy3/μg cRNA. The probes were hybridized on Agilent maize custom arrays 4X180k containing 176,026 in situ synthesized 60-mer oligonucleotide features (without controls). Hybridization was carried out for 17 h at 65 °C, followed by washing and scanning on SureScan Microarray Scanner at 2 μm. In addition, to avoid the ozone effects on Cy-3 signal, a supplementary organic solvent containing an ozone scavenging compound dissolved in acetonitrile was used after washing step. Feature Extraction (FE) software v. 11.0 was used for image processing. Preprocessing and differential data analysis were performed on median signal from raw files generated by FE, using standard functions in R/Bioconductor (https://www.bioconductor.org/; last accessed October 6, 2016) and custom written routines. Control and flagged spots were removed and data were normalized between arrays using quantile method implemented in normalizeBetweenArrays function/*limma* package (Ritchie et al. 2015). Transcripts originating from the same gene were combined by taking the median value of the intensities. The differentially expressed sequences were selected with *limma* package by fitting a linear model for each sequence and using an empirical Bayes smoothing to moderate the standard errors. A gene was considered differentially expressed when the *P* value adjusted for multiple testing (Benjamini–Hochberg method) was <0.01.

Online Software and Tools Used

The patterns of gene expression in 25 maize tissues were compiled using data from Sekhon et al. (2011) curated in MaizeGDB (Lawrence et al. 2007) within the Gene Models tool: http://beta.maizegdb.org (last accessed October 6, 2016). Syntenic orthologs in the *Poaceae* family and *Arabidopsis* homologs for the 96 genes of interest were pulled out from the same data base, which curates data from Schnable et al. (2012), and annotations from the 284 *Zea mays* release of Phytozome 10 (Schnable et al. 2009). BLASTP searches for the 96 putative proteins were performed against the human genome on the NCBI webpage. No putative or predicted proteins were taken into account and only hits with e-value $\leq e^{-10}$ were considered.

TargetP 1.1 server (http://www.cbs.dtu.dk/services/TargetP/; last accessed October 6, 2016) (Emanuelsson et al. 2007) was used to predict the subcellular localization of the translated sequences corresponding to the 96 genes. There are four possible predictions (chloroplast, mitochondrion, secretory pathway and any other location), each with an associated reliability class (RC) from one (most reliable) to five (least reliable).

PLACE database (Higo et al. 1999) (http://www.dna.affrc.go.jp/PLACE/) was used for screening the 500 bp promoter region of the 96 genes of interest for GATA, G-box, and CCAC motifs. The 500-bp promoter sequences were retrieved from Gramene, using BioMart (http://ensembl.gramene.org/biomart; last accessed October 6, 2016).

Transcriptome Analysis

The transcriptome of each of the three cytolines was sequenced in triplicates using a HiSeq2500 Illumina sequencer. Each RNAseq library consisted of more than 35mio paired-end reads of 2×150 bp in length. Reads were first checked for sequence quality (http://www.bioinformatics.babraham.ac.uk/projects/fastqc; last accessed October 6, 2016) before mapping them to the reference maize genome (AGPv3.23) using the spliced mapping approach implemented in TopHat2 (Kim et al. 2013). For each gene of the corresponding annotation we counted the number of reads mapping to it using the program HTSeq-count (Anders and Huber 2010). Significant different expression levels between the three strains were assessed using the R-package DESeq2 (Love et al. 2014). *P* values were corrected for multiple testing following Benjamini and Hochberg (1995) and a significant threshold of 0.01 was applied.

Reads not mapping to the reference genome (AGPv3.23, which includes the two organellar genomes) were used to find novel transcripts of the two cytoplasm-donor lines (TC208(cytTC316) and TC208(cytW633). First, the unmapped reads were assembled individually for all three strains using Trinity version 2.1.1 (Grabherr et al. 2011). Second, the transcriptomes of the two cytoplasm-donor lines were filtered for transcripts not present in the reference transcriptome (AGPv3.25) and the donor strain TC208 using cd-hit-est (Li and Godzik 2006). Transcripts were defined to be novel if the sequence identity threshold or the length difference cutoff were <80%. Third, the transcriptomes of the two lines were filtered for common novel transcripts using cd-hit-est (Li and Godzik 2006). Transcripts were defined as identical if their sequence identity threshold and their length

difference cutoff were >99% and 80%, respectively. Transcripts with low complexity or repeat sections were filtered out using RepeatMasker (Smit et al. 2014). Finally, NCBI-BLAST version 2.2.29 (Camacho et al. 2009) and BLAST2GO (Conesa et al. 2005) were used to annotate the transcripts and to perform a GOterm enrichment analysis.

Results

Cytolines Are the Result of at Least Ten Crosses

Considering that maize has a life cycle of ~6 months, a breeder needs at least 10 years to create cytolines in temperate climate conditions, which allow for just one generation a year. We used the inbred line TC208 as nucleus donor (B in fig. 1; henceforth line "B") and the inbred lines TC316 and W633 as cytoplasm donors/nucleus acceptors (A1 and A2 in fig. 1). It is important to note that the three starting lines (TC208, TC316, W633) were also generated through at least ten self-crosses, resulting in inbred lines (i.e., all loci in the genome are homozygous). Based on their pedigree, TC208 and TC316 are part of the "Lancaster Sure Crop", whereas W633 is part of "Minnesota 13" heterotic groups, according to the maize germplasm classification by Troyer (1999). The extant genetic variability within the "Lancaster Sure Crop" heterotic group (Smith and Smith 1987), corroborated with the membership of W633 to a different heterotic group ("Minnesota 13") support our conclusion that the two cytoplasm donors/nucleus acceptors lines carry different cytoplasm.

Line B's nuclear material replaces half of the acceptor's nuclear material in the first cross (fig. 1). Consequently, the paternal line B will contribute 50% of the cell's proteome, i.e., line B's nuclear genes will code 50% of the proteins present in the cytoplasm of the F1 progenies. In subsequent crosses, line B's contribution towards the generation of the cytoplasm itself will constantly increase up to 99.95%. At the end of the tenth cross, the resulting cytoplasm will be composed of the organelles of line A1/A2, but in a cytoplasmic environment generated exclusively from translating line B's mRNA. Thus, the three cytolines (Cyt1, Cyt2, and B) have the same nucleus and cytoplasm, but different organelles (chloroplast and mitochondria) that were not transmitted through the pollen of line B (always used as paternal line). Therefore, the only influence that could trigger a change in NGE would be exerted through the RS pathway. The cytolines were later maintained by sibmating or self-mating. They are the result of approximately two decades of breeding efforts, considering the inbreeding process of the three lines (B, A1, and A2), which had preceded the cytolines creation.

The Organellar Genomes Differentially Regulate over 1,000 Nuclear Genes

Transcriptome sequencing, using an Illumina HiSeq 2500 instrument, revealed that 5,009 genes changed their expression

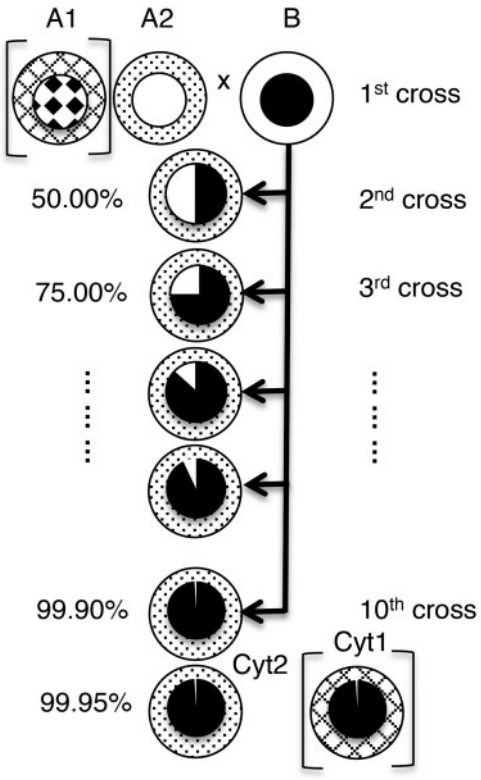

Fig. 1.—Cytolines are created by at least ten crosses. Three cytolines (B, Cyt1, and Cyt2) were created by crossing inbred line B (always as male: i.e., pollen donor) into inbred lines A1 and A2, respectively. After ten such crosses, 99.95% of the acceptor lines/cytoplasm donors' genetic material is replaced by that of inbred line B. Neither mitochondria, nor chloroplast are transmitted through pollen. The three cytolines share the same nucleus, which resides on different cytoplasmic environments, each characterized by its own organellar genomes.

pattern in Cyt1 and 1,914 in Cyt2, respectively, when compared with their nucleus donor line B (using a stringent $P_{adj} < 0.01$ cutoff). The comparison of the two cytolines alone resulted in 3,646 differentially regulated genes. These were filtered out when B versus Cyt1 and B versus Cyt2 gene sets were overlapped (supplementary table S1, Supplementary Material online) to help in the deconvolution of the genes of interest, i.e., common to the two cytolines when compared with the nucleus donor. Thus, Cyt1 and Cyt2 have 1,179 genes in common that share the same expression patterns when compared with the nucleus donor line B. Of these, 608 are up-regulated and 571 are down-regulated (fig. 2 and supplementary table S1, Supplementary Material online). These shared genes are potential end-receptors in a general mechanism used by the RS pathways to communicate with the nucleus in a normal state of the cell.

To further refine the gene set and classify the genes according to their biological function we used their $\log_2$ fold change values as input for the MapMan software (Thimm

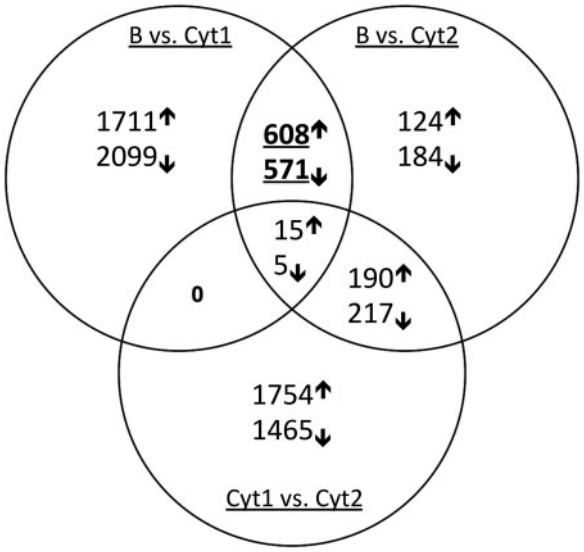

Fig. 2.—Up- (↑) and Down-regulated (↓) genes with the same expression pattern in the two cytolines (Cyt1 and Cyt2) when compared with the nucleus donor line B. Out of the 5,009 differentially regulated ($P_{adj} < 0.01$) genes in Cyt1 and 1,914 in Cyt2, respectively, 608 are up-regulated and 571 are down-regulated in both, when compared with B.

et al. 2004). MapMan is a widely used software in the Omics field (having collected more than 1,200 citations so far) due to its modular structure, which allows for the creation of nonredundant gene ontologies (Scavenger module), easy visualization of the genes of interest on schematic diagrams (Image Annotator module), and most importantly, due to its statistical power in evaluating the genes' responses in the context of metabolic pathways or biological processes (PageMan module). As described in Usadel et al. (2009; a follow-up article of Thimm et al. 2004), MapMan is superior to other tools that handle Omics data. The latter ones have a deficit precisely in the plant signaling pathway, which is of interest to us. We therefore used the latest mapping file version available for maize (Zm_B73_5b_FGS_cds_2012) in our analyses using MapMan.

Key Nuclear Receptors of the RS Pathways

Ninety-six nuclear genes have been identified by MapMan as differentially regulated in both Cyt1 and Cyt2, when compared with B (fig. 3 and supplementary table S2, Supplementary Material online). Down-regulated genes are more abundant than up-regulated ones, with 56 and 40 representatives, respectively. MapMan placed 91 of the genes in 14 of its 34 bins, which were used to categorize gene functions, whereas five could not be assigned (fig. 4). More than half of those assigned fit to just three of the bins: secondary metabolism (18), cell wall (16) or lipid metabolism (14). The remaining are involved in CHO metabolism (9), amino acid metabolism (8), glycolysis (6), photosystem (PS) (6), mitochondrial electron transport (4), fermentation (3), nucleotide metabolism (2), tetrapyrrole metabolism (2), and three more classes with just one representative (N-metabolism, RNA regulation of transcription, and TCA). On an average, two-thirds of the genes involved in secondary metabolism, cell wall, glycolysis, PS, and fermentation are down-regulated in the two cytolines. Only the genes involved in lipid metabolism and CHO metabolism have more up-regulated representatives (fig. 4). More details on the sub-bins used by MapMan and the distribution of the 96 genes among them can be found in columns 3–4 of supplementary table S3, Supplementary Material online.

To validate our findings, we analyzed the same three cytolines (B, Cyt1 and Cyt2) using the microarray technology. About 19 genes (highlighted in blue in fig. 3) had the same expression pattern in both Cyt1 and Cyt2 transcriptomes, whereas 19 more (highlighted in green) were validated in one of the two samples, when compared with B. As shown in figure 3 and supplementary table S2, Supplementary Material online, the 19 genes highlighted in blue tend to cluster towards the ends of the heat-map generated from $\log_2$ fold changes values.

Consequently, considering the stringent criteria used to define the set of 96 genes ($P_{adj} < 0.01$ and MapMan's internal filters) and the superior power of the NGS technique, our further analyses focused on the whole set of 96 genes defined above.

The Candidate Genes Are Ubiquitously Expressed throughout Plant Development

We used data from Sekhon et al. (2011), curated in MaizeGDB (Lawrence et al. 2007), to check for gene expression patterns in various tissues, organs, and developmental stages of the maize plant for all our candidate genes. Out of the 96, there were 82 for which data were available (supplementary table S4, Supplementary Material online). Leaf is the organ where most of the genes are expressed (79), followed by fruit, stem internode, shoot apical meristem, tassel and ear inflorescence, inflorescence bract, primary shoot system, seedling coleoptile, shoot apex, and central spike of the ear, all of which have over 70 genes expressed (fig. 5). There are 46 genes that are expressed in all 25 tissues analyzed, whereas 67 are expressed in at least 20 tissues. The remaining ones are expressed in more than ten tissues (nine representatives) and five more, whose expression is restricted to less than ten tissues (supplementary table S4, Supplementary Material online).

The 96 Genes Have Syntenic Orthologs in the *Poaceae* Family and Homologs in *A. thaliana* and Humans

Considering their ubiquitous expression above, we expected the genes to have orthologous copies at least in other grasses. Therefore, we took advantage of the exiting data on orthologous genes in the *Poaceae* family (Schnable et al. 2012), curated in MaizeGDB (Lawrence et al. 2007), and subtracted the

Gene ID	Heat-map	Loc.	MapMan annotation	GATA	G-Box	CCAC	A/Sb/Si/O/B orthologous genes	H.s.	Chr
GRMZM2G107076	5,54	_	Sec. met.	3	-	2	+ + + + +	+	5
GRMZM2G006290	3,94	_	Not assigned	2	-	2	+ x x x x		8
GRMZM2G076239	3,62	_	PS	10	-	4	+ + + + +	+	2
GRMZM2G046070	3,02	M	Sec. met.	4	-	3	+ + + + +	+	2
GRMZM2G401970	1,95	C	Lipid met	5	-	4	+ + + + +		6
GRMZM2G531230	1,93	_	AA met.	-	-	9	+ + + + +	+	2
GRMZM2G098346	1,89	_	Ferm.	4	-	2	+ + + + +	+	4
GRMZM2G135470	1,85	M	Sec. met.	6	2	3	+ + + + +	+	10
GRMZM2G116876	1,67	_	Lipid met	-	-	1	+ + + + +		5
GRMZM2G140996	1,65	_	Sec. met.	4	-	1	+ + + + +		6
GRMZM2G010348	1,40	_	Mito ET	2	-	3	+ + + + +	+	8
GRMZM2G156026	1,32	C	Sec. met.	5	-	2	+ + + + +		4
GRMZM2G147701	1,29	M	Lipid met	-	-	4	+ + + + +	+	4
GRMZM2G055667	1,26	C	Lipid met	3	-	4	+ + + + x		8
GRMZM2G174598	1,25	S	Cell wall	3	-	4	+ + + + +		4
GRMZM2G090980	1,18	_	Sec. met.	5	-	4	+ + x x x	+	9
GRMZM2G423137	1,14	M	PS	4	-	4	+ + + + +		4
GRMZM2G117357	0,84	_	Lipid met	4	-	3	+ + + + x	+	6
GRMZM2G443715	0,72	M	Cell wall	2	-	10	+ + + + +		9
GRMZM2G013357	0,71	S	Sec. met.	3	2	5	+ + x + x	+	2
GRMZM2G044775	0,68	C	Nucl. met.	5	-	-	+ + + + +	+	8
GRMZM2G157113	0,65	_	Lipid met	5	-	-	+ + x x x	+	6
GRMZM2G154124	0,61	_	Cell wall	1	-	6	+ + + + +		1
GRMZM2G042179	0,51	_	Cell wall	4	-	3	+ + + + +	+	4
GRMZM2G004534	-0,32	_	Glycolysis	1	-	6	+ + + + +	+	10
GRMZM2G074282	-0,33	C	AA met.	7	-	1	+ + + + +	+	5
GRMZM2G132898	-0,35	M	Lipid met	3	-	6	+ + + + +	+	1
GRMZM2G122715	-0,36	C	Not assigned	2	-	5	+ + + + +		4
GRMZM2G029566	-0,40	_	Cell wall	3	-	1	+ + + + +	+	4
GRMZM2G036759	-0,43	_	TCA	3	-	3	x x x x x	+	9
GRMZM2G112609	-0,43	_	Mito ET	3	-	7	+ + + + +	+	1
GRMZM2G106578	-0,44	C	Lipid met	-	2	2	+ + + + +	+	2
GRMZM2G454952	-0,47	C	Sec. met.	-	-	1	+ + + + +		7
GRMZM2G070199	-0,49	_	Mito ET	4	-	8	+ + + + +	+	6
GRMZM2G082007	-0,50	_	Sec. met.	4	-	2	+ + + + +	+	4
GRMZM2G345493	-0,51	_	Glycolysis	1	-	7	+ + + + +	+	9
GRMZM2G120724	-0,52	S	Cell wall	7	-	2	+ + + + +		2
GRMZM2G177631	-0,52	_	Cell wall	5	-	3	+ + + + +		7
GRMZM2G103281	-0,53	S	Lipid met	4	2	1	+ + + + +	+	4
GRMZM2G072091	-0,54	C	CHO met.	4	-	-	+ + + + +	+	9
GRMZM2G010555	-0,56	_	Mito ET	4	-	9	+ + + + +		2
GRMZM2G025171	-0,64	C	PS	1	-	1	+ + x + +	+	4
GRMZM2G139360	-0,70	M	Glycolysis	4	-	2	+ + + + +	+	1
GRMZM2G145029	-0,73	C	Sec. met.	5	-	-	+ + + + +	+	8
GRMZM2G060886	-0,74	_	Lipid met	8	-	3	+ + + + +		8
GRMZM2G027955	-0,87	C	CHO met.	1	-	5	+ + + + +		6
GRMZM2G155242	-0,91	_	CHO met.	2	-	-	+ + + + +	+	1
GRMZM2G140107	-0,92	_	CHO met.	-	-	7	+ + + + +		3
GRMZM2G093666	-0,94	C	AA met.	10	-	3	+ + + + +		1
GRMZM2G114127	-0,94	_	Cell wall	4	-	2	+ + + + +		5
GRMZM2G004528	-0,99	_	CHO met.	2	4	6	+ + + + +	+	9
GRMZM2G051185	-1,05	S	Cell wall	1	-	11	+ x x + x		4
GRMZM2G103197	-1,12	_	Tpyrl. synt.	2	-	3	+ + x x +		1
GRMZM2G110881	-1,23	_	Sec. met.	6	-	4	+ + + + +	+	5
GRMZM2G140994	-1,36	M	RNA r.trs.	3	-	1	+ + + + +	+	8
GRMZM2G306566	-1,38	_	Lipid met	-	-	2	+ + + + x		5
GRMZM2G044107	-1,43	S	Cell wall	3	-	3	+ x x x x		4
GRMZM2G021794	-1,46	S	Cell wall	7	-	5	+ + + + +		6
GRMZM2G119941	-1,51	S	CHO met.	3	-	2	+ + + + +		2
GRMZM2G079477	-1,61	C	Nucl. met.	3	-	7	x x x x x	+	4
GRMZM2G097297	-1,64	_	Sec. met.	1	-	8	+ x + x x		4

Fig. 3.—Nuclear genes that respond to retrograde signaling pathways. The complete form of this figure is presented as supplementary table S2, Supplementary Material online. Col. 1—gene IDs of those validated in both microarray analysis are highlighted in blue, whereas those validated by one of the

(continued)

syntenic orthologs in four other species: *Sorghum bicolor*, *Setaria italica* (foxtail millet), *Oryza sativa* ssp. *japonica* (rice), and *Brachypodium distachyon*. Indeed, the vast majority (88) of the 96 genes have such syntenic orthologs in the other grasses (fig. 3 and supplementary tables S2 and S3, Supplementary Material online). The remaining eight do have an orthologous copy in at least one other grass species but they are not syntenic.

When the *Arabidopsis thaliana* genome was queried, 82 of the genes had a homologous copy. The eight genes missing a syntenic ortholog in the *Poaceae* do not have a homologous copy in *A. thaliana*.

We also performed BLASTP searches of all 96 putative proteins against the human genome, taking into account only hits with an e-value $\leq e^{-10}$ and query coverage of at least 50%; no putative or predicted proteins in human were considered.

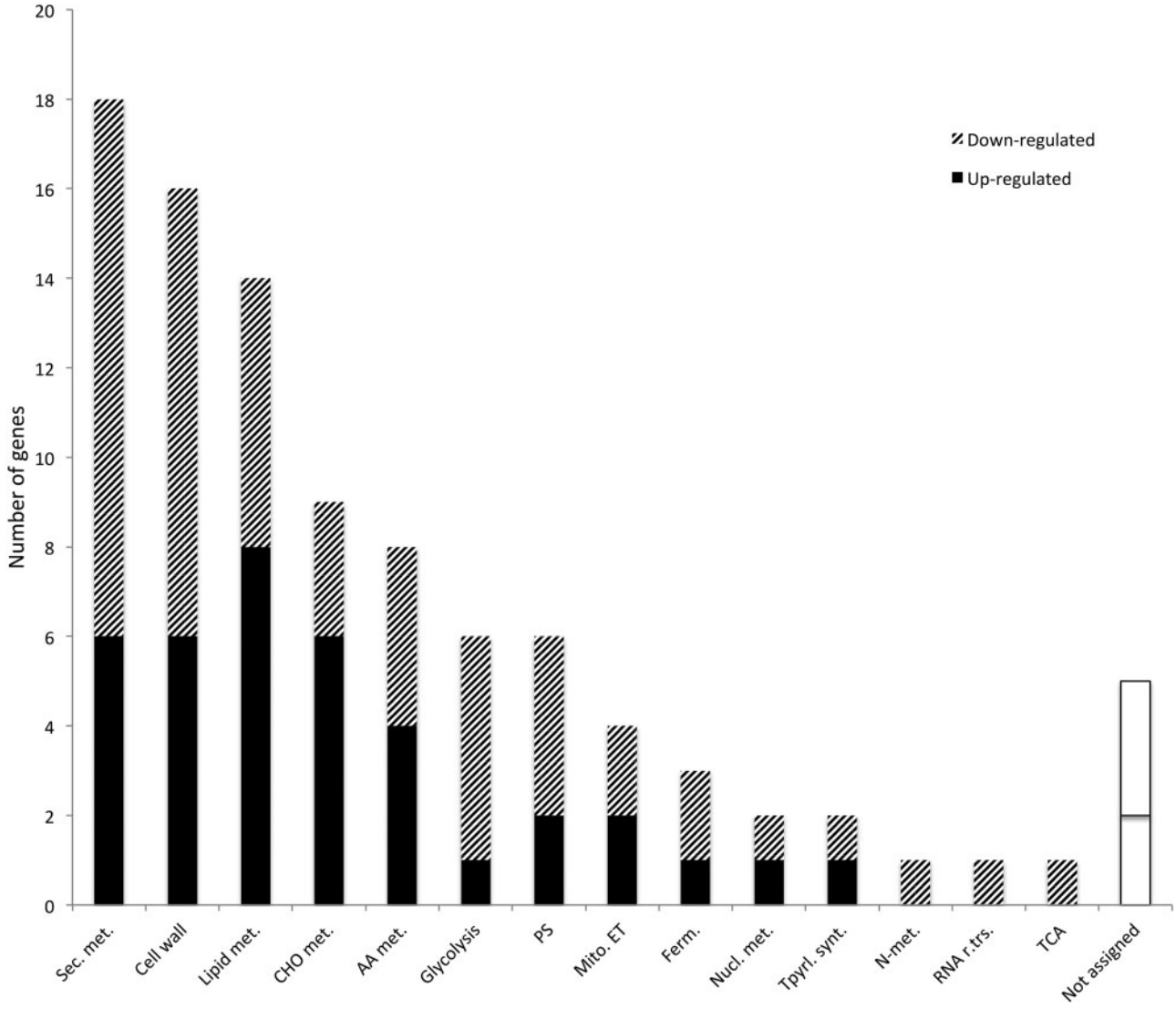

Fig. 4.—Gene functions of the putative proteins are sorted using MapMan into 14 of its 34 bins.

Fig. 3.—Continued

microarrays, in green; Col. 2—$\log_2$ fold changes in Cyt1 and Cyt2 were averaged an used for the heat-map; Col. 3—subcellular localization of the putative proteins, as predicted by TargetP 1.1 (C = chloroplast, M = mitochondrion, S = secretory pathway, any other location); Col. 4—gene annotation according to the classes used by MapMan; Col. 5–7—number of occurrences of three motifs in the 500 bp promoter region; Col. 8—Presence (+) or absence (x) of a homologous gene in the other five species (A = *Arabidopsis thaliana*, Sb = *Sorghum bicolor*, Si = *Setaria italica*, O = *Oryza sativa*, B = *Brachypodium distachyon*); Col. 9—Presence (+) of a homologous copy in human; Col. 10—Chromosomal location for each gene.

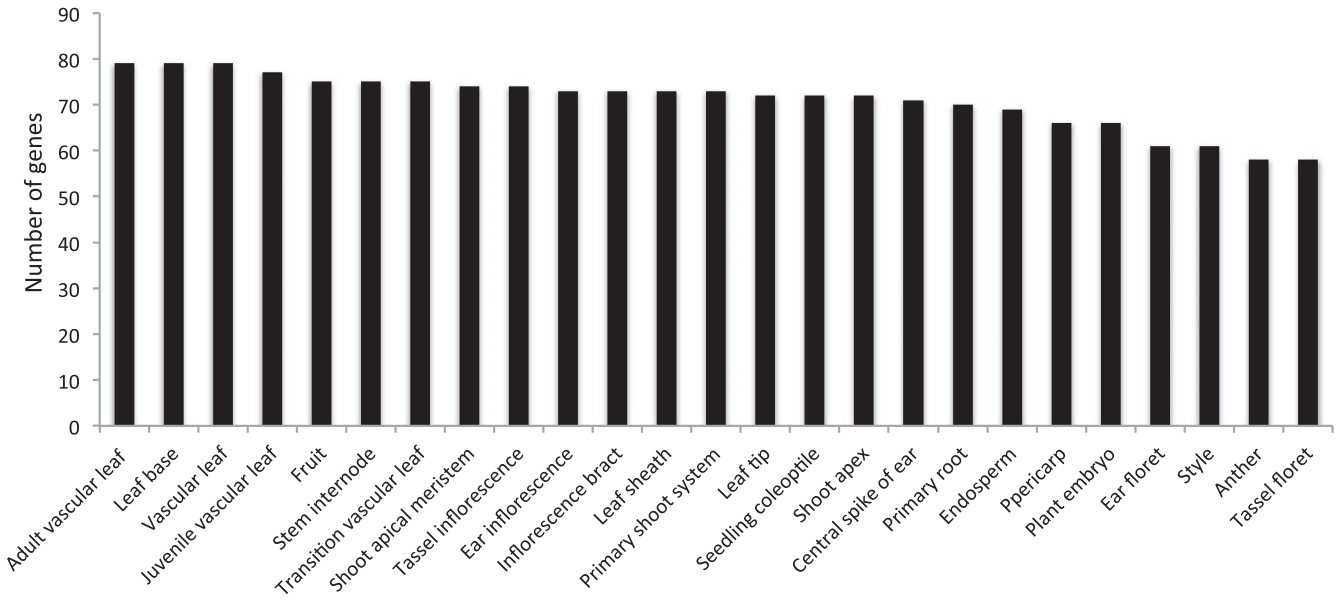

Fig. 5.—Expression pattern for 82 genes of interest across the 25 tissues investigated by Sekhon et al. (2011).

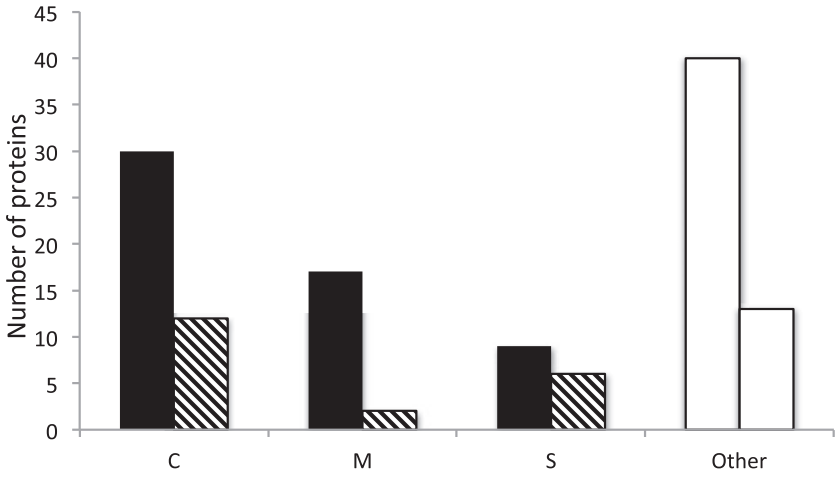

Fig. 6.—Subcellular localization according to TargetP 1.1. Black columns—all predicted results; grey columns—only RC (reliability class) ≤ 2 (i.e., strongest prediction). C, chloroplast transit peptide; M, mitochondrial targeting peptide; S, secretory pathway signal peptide.

Thus, we identified 35 human proteins with significant homology to our maize sequences. The full list, containing maize genes and their corresponding human proteins, query coverage (%), e-value, and identity (%), is included in supplementary table S3, Supplementary Material online.

There Are Twice as Many Proteins Being Targeted to the Chloroplast than Mitochondria and Secretory Pathway Taken Together

TargetP 1.1 server (Emanuelsson et al. 2007) was used to predict the subcellular localization of the translated sequences corresponding to the 96 genes. Among those, 30 contain a chloroplast transit sequence, 17 a mitochondrial targeting peptide, and nine are directed to the secretory pathway (fig. 6). However, the software differentiates five reliability classes (RC) when predicting the localization, an RC = 1 indicating the strongest prediction. When considering only proteins with RC = 1 or RC = 2, the results have a similar pattern, with chloroplast-directed proteins accounting for almost double the ones targeted to mitochondria and secretory pathway (fig. 6 and supplementary table S2, Supplementary Material online).

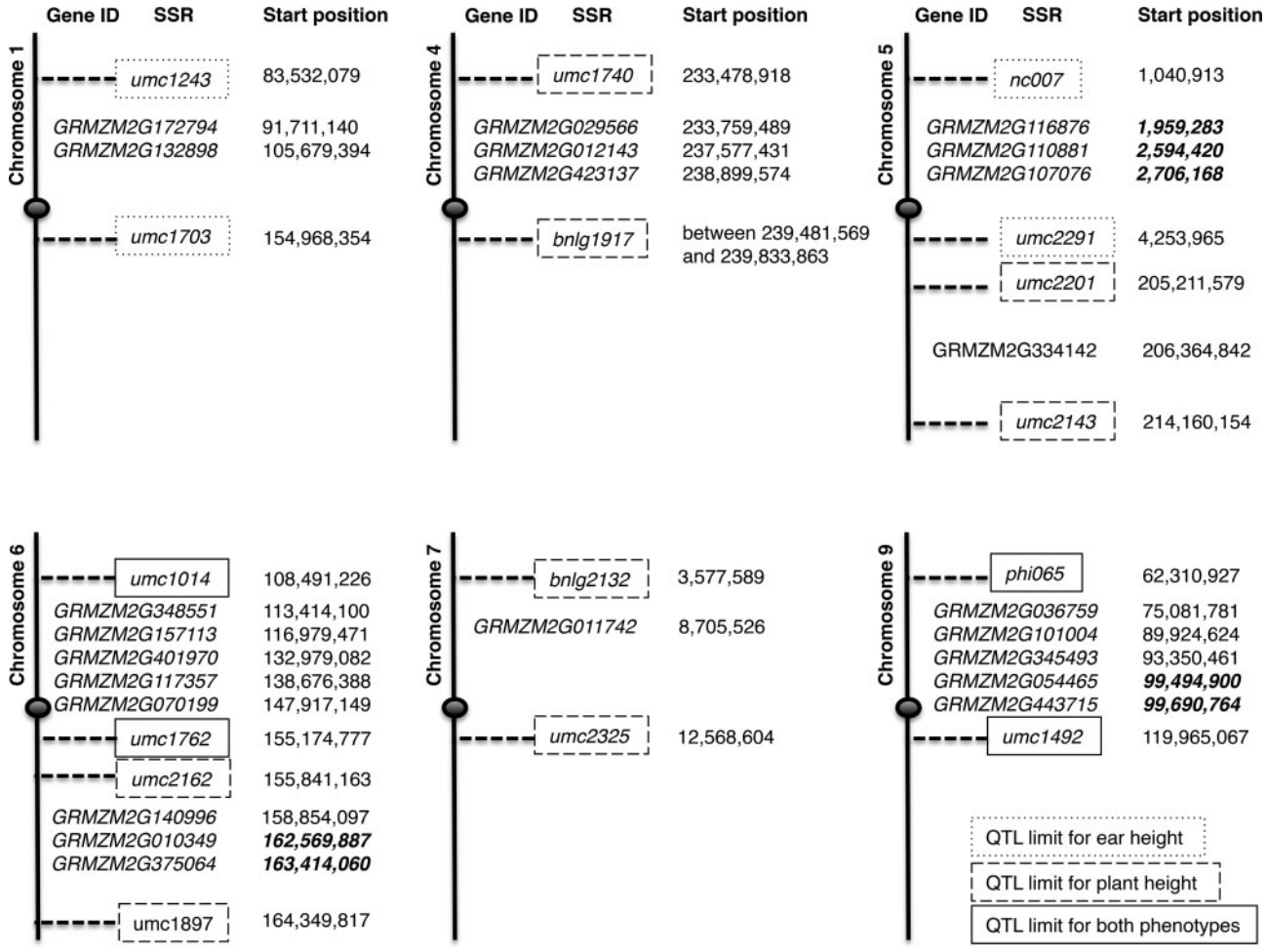

Fig. 7.—Chromosomal positions of the 23 genes that fit within previously defined QTLs for two agronomically important traits. Start positions are given for the SSR markers (boxed) and the genes (in italic). Start positions of the genes that cluster together are bolded.

Of special interest were the 35 genes identified above as having a human homologous copy. We expected none of those to be targeted to the chloroplast. Indeed, 29 genes translate into proteins predicted to target the mitochondria, secretory pathway or another location. However, six were predicted to have chloroplast localization (highlighted in yellow in supplementary table S2, Supplementary Material online), based on their chloroplast transit sequence. Among those, five had RC $\geq$ 3 (i.e., not a reliable prediction) whereas one had an RC = 2. Furthermore, all six have an identity of <40% when their amino acid sequence was used as query in BLASTP against human data (supplementary table S3, Supplementary Material online).

Promoter Analysis (500 bp)

The promoter regions of genes that are under the control of retrograde pathways have been shown to contain several motifs, like GATA, G-box, or CCAC. The first two are present in the promoters of genes responsive to light and plastid retrograde signals (Chi et al. 2013), whereas the CCAC motif is bound by ABI4, a transcription factor that modulates NGE and integrates signals coming from three RS pathways (Koussevitzky et al. 2007). We screened the 500-bp promoter region of our candidate genes for such motifs using the PLACE database (Higo et al. 1999).

The G-box occurred at least twice in the promoters of 14 genes, with two of those having the motif present 4 times. In addition, nine genes have been identified as having more than ten GATA or CCAC motifs, a strong indication that they are under the control of RS pathways (bolded in red in fig. 3). Another set of 18 genes have 5–10 GATA motifs in their promoters, whereas 23 have 5–10 CCAC motifs (supplementary table S2, Supplementary Material online).

The number of occurrences of each of the three motifs investigated represents a significant enrichment. By chance alone, one would expect the G-box (which is a six-bases motif: CACGTG) to occur 0.195 times in a 500-bp window, the CCAC motif to occur 3.153 times, and the GATA motif to occur 4.659 times.

Twenty-Three Genes from This Study Fit within QTLs Previously Defined in Maize Cytolines, Responsible for Plant Height and Ear Height

The 96 genes of interest are evenly spread across the ten maize chromosomes (fig. 3 and supplementary table S3, Supplementary Material online). We used their chromosomal positions in a comparison with data available from Tang et al. (2013). The authors' main objectives were to use maize cytolines to assess the impact of a new cytoplasmic environment on two phenotypes of interest (namely, plant height and ear height) and to identify the QTLs responsible for them. Our hypothesis was that part of the 96 genes likely fit within such QTLs, as proof of their role in RS. We retrieved the chromosomal positions of the SSR markers that had been used to define those QTLs and anchored our 96 genes to the map (fig. 7). There are 23 genes that co-localize with eight QTLs defined by Tang et al. (2013). Among these, 15 fit within QTLs defined for ear height, and 18 within QTLs defined for plant height, respectively. Ten are part of two QTLs responsible for both characters, on chromosomes 6 and 9. Some are clustered together, with two of the QTLs (delimited by *umc1014-umc1762* and *phi065-umc1492*) harboring five genes each on chromosome 6 and 9, respectively. Furthermore, there are two instances where the genes are <200 kb apart: *GRMZM2G110881* is 102,279 bp away from *GRMZM2G107076* (QTL defined by *nc007* and *umc2291*, on chromosome five) and *GRMZM2G054465* is 186,489 bp away from *GRMZM2G443715* (QTL defined by *phi065* and *umc1492*, on chromosome nine). Three others are <1 Mb apart (fig. 7).

Novel Transcripts Are Being Produced in Cyt1 and Cyt2

We used Trinity (Grabherr et al. 2011) to perform a *de novo* assembly of the RNA-seq data in the three cytolines. Only the transcripts that did not map to the reference genome sequence of inbred line B73 (Schnable et al. 2009) were used. Next, the transcripts flagged by RepeatMasker (Smit et al. 2014) as having a transposable or retro-transposable element origin were discarded. Thus, approximately 600 transcripts were identified as common in the two cytoplasm-donor lines (Cyt1 and Cyt2) when compared with the nucleus donor line B (supplementary table S5, Supplementary Material online). The genes coding for them are silent (or expressed at undetectable levels) in the donor line and become expressed as a result of the novel cytoplasmic environment of Cyt1 and Cyt2, respectively. To get more insights into the functions of these putative proteins we used the software BLAST2GO (Conesa et al. 2005). Interestingly, more than 100 transcripts are categorized as GO:0006355 ("regulation of transcription, DNA-templated"), GO:0006351 ("transcription, DNA-templated"), GO:1903506 ("regulation of nucleic acid-templated transcription") or GO:0010468 ("regulation of gene expression"). Many more are involved in signal transduction (e.g., GO:0007165), protein phosphorylation (e.g., GO:0006468) and other processes that could potentially alter NGE as a result of retrograde signals received from the new organellar genomes of the two cytolines (supplementary figs. S1 and S2, Supplementary Material online).

Discussion

Maize Cytolines Provide an Alternative to Studying RS

Retrograde signaling pathways are still the subject of much debate, despite many of their components being identified and integrated into complex networks (Chi et al. 2013). Research on mitochondria-to-nucleus communication is less advanced. It is however clear that a certain overlap does exist between chloroplast and mitochondrial RS, with ABI4 transcription factor providing a strong case in this regard (Giraud et al. 2009). Studies on the plastid RS were first reported in barley mutants (Bradbeer et al. 1979) but quickly shifted to *Arabidopsis*, as a more amenable model organism, whereas research on mitochondrial RS mainly focused on yeast (*Saccharomyces cerevisiae*) (Liu and Butow 2006). In a recent review, Ng et al. (2014) show that *Arabidopsis* was almost exclusively used for understanding the impact of mitochondrial RS on nuclear genes and acknowledge the need to move to other plant models. The same is true in the field of chloroplast signaling, where *Arabidopsis* has been extensively used (Chi et al. 2013), mainly because of the availability of mutants for almost all the nuclear genes.

Indeed, mutants that are defective in the RS pathways and stress, have both been center points in trying to understand how NGE is regulated by signals coming from the organelles. But studying how nuclear genes expression changes as a result of a (single) stimulus may not provide a holistic view of the mechanism. Therefore, the use of mutants and stress does not have the potential to identify key nuclear genes that respond to retrograde signals in a normal state of the cell, but cytolines do. This is because the same nucleus is under the influence of different cytoplasms, each having its own set of organellar genomes. By identifying genes that are differentially regulated in such cytolines, we can argue that these are molecular switches that the RS pathways use to relay signals to the nucleus and control NGE according to the cell's needs. However, our experimental design cannot capture the entire set of nuclear genes that respond to retrograde signals. To do so, one can envision a similar approach to defining the pan-genome/pan-transcriptome of a species by sequencing the

transcriptomes of a larger cytolines pool. The 96 genes identified here are a valuable addition to the scant extant knowledge on the nuclear targets of RS. We labeled them as "key nuclear genes" that respond to RS pathways, having been filtered from the larger set of 1,179 candidates (fig. 2 and supplementary table S1, Supplementary Material online) and based on their functional analysis, as we show below.

Filter Criteria and Microarray Validation

Here, we identified 96 nuclear genes that are differentially regulated in the two cytoplasm-donor lines (Cyt1 and Cyt2 in fig. 1) compared with the nucleus-donor line (B in fig. 1). However, depending on how stringent we set the selection criteria, there are up to 1,179 candidate genes (fig. 2 and supplementary table S1, Supplementary Material online). This is almost twice the number of genes identified by Crosatti et al. (2013) as being differentially regulated when wheat nucleus was transferred to *Hordeum chilense* cytoplasm. The authors identified 540 genes as differentially regulated in an interspecies comparison—where one would expect significant changes to occur—whereas our experimental setup is intraspecific. The filtering criteria used in the study above were a 2-fold change in expression and a P value ≤ 0.05 for a gene to be considered as having altered expression. In contrast, we used a much more stringent P value (≤ 0.01) that was even adjusted for multiple testing, but did not set a fold-change threshold. Using these criteria we avoided missing important genes from the analyses, like transcription factors, whose impact on NGE might be substantial even with a slight change in their transcript level. Working with the entire dataset is also important when analyzing gene expression using MapMan, as the software sets the results in the context of metabolic pathways or biological processes (Usadel et al. 2009). As an example of genes that would have been missed when using a fold-change threshold is *GRMZM2G010349*, which has a $\log_2$ fold change of 0.56 (supplementary table S2, Supplementary Material online). Its homologous copy in *Arabidopsis*, *AT1G68830* (supplementary table S3, Supplementary Material online), codes for STN7, a protein that is proposed to act as a signaling or a sensing component for the redox signals within the plastid (Pesaresi et al. 2009; Dietzel et al. 2015), and is part of one of the four classic chloroplast RS pathways (plastid redox state).

GRMZM2G010349 is also a good example when the two platforms for transcriptome analysis are compared: NGS versus microarray. Using just the microarray technique this gene would have been missed from the analysis (supplementary table S2, Supplementary Material online), whereas NGS provides a higher resolution. As we show above, when the experiments carried out using the NGS platform were replicated using a custom-made microarray there was a core set of 19 genes that could be validated in both Cyt1 and Cyt2 and 19 more genes in just one (i.e., 38 out of the 96). The first 19, highlighted in blue in figure 3, tend to cluster towards the ends of the heat-map generated from $\log_2$ fold change values. This can be attributed to the microarray platform, which has a lower dynamic range compared with an NGS instrument being able to identify only those genes that are either strongly up- or down-regulated. It is also worth mentioning that not all of the 96 genes identified by NGS-transcriptome analysis were spotted on the custom-made microarray chip. These are: *GRMZM2G071226, GRMZM2G066865* and *GRMZM2G066791*.

When comparing the number of differentially regulated genes from B versus Cyt1 and B versus Cyt2, respectively, they differ by 2.6-fold (fig. 2 and supplementary table S1, Supplementary Material online). We speculate that the extant difference relates to the pedigree of the three inbred lines used to create the three cytolines. This translates into different polymorphisms of the organellar genomes of Cyt1 and Cyt2 having a differentiated impact (through RS) on the gene expression of B line nucleus. However, in order to properly assess this impact, one would have to sequence a representative pool of organellar genomes in maize, similar to the work of Moison *et al.* (2010) in *Arabidopsis.* The number of inbred lines selected for this analysis would have to be proportional to the maize genetic diversity worldwide. To the best of our knowledge such a comprehensive study has not been undertaken yet.

Ubiquitous Expression Pattern for the Genes of Interest

We argue that the 96 genes identified here are molecular switches targeted by the RS pathways. Their ubiquitous expression pattern throughout plant development strengthens our hypothesis. There are 46 genes (among the 82 for which data was available) that are expressed in all 25 tissues analyzed in Sekhon et al. (2011) (gene expression pattern in all 25 tissues is detailed in supplementary table S4, Supplementary Material online). The fact that only 82 of the 96 genes of interest are found within the data reported by Sekhon et al. (2011) can be explained by: (1) the platform used by the authors in their transcriptome analysis (i.e., microarray vs. NGS, in the present study)—it is worth noting that one gene is reported to be expressed only in the flower (*GRMZM2G306566*), whereas we confirm here that it is expressed in nine-day-old seedling, too, (2) the arbitrary threshold used by the authors, which excluded data coming from 2,647 probes (8.6% of total), and (3) the high intraspecific variability of maize (Hirsch et al. 2014), which leads to significant differences when different inbred lines are analyzed, including gene copy number variation (Springer et al. 2009).

Figure 5 summarizes the expression patterns for the 82 genes for which data were available. Based on these observations, we speculate that the genes we identified here are housekeeping genes, playing important roles in all tissues and throughout plant development. Our study complements

the "Genome-wide atlas of transcription during maize development" by Sekhon et al. (2011) adding new data at whole seedling stage (nine days old). It also provides evidence for eight more genes (*GRMZM2G114486*, *GRMZM2G054465*, *GRMZM2G058675*, *GRMZM2G101004*, *GRMZM2G029856*, *GRMZM2G038821*, *GRMZM2G172794*, and *GRMZM2G079477*) being expressed in maize seedlings, which are not mentioned in the atlas above. These eight genes seem to be restricted to maize, when searching for homologous copies in *Arabidopsis*, members of the *Poaceae* family, or human (supplementary table S2, Supplementary Material online).

Majority of Genes of Interest Are Involved in Secondary Metabolism, Cell Wall, or Lipid Metabolism

The fact that *secondary metabolism* genes are well represented (fig. 4) may be an indication that cytolines' nuclei are still exposed to a certain level of stress in their new cytoplasmic environment. This is because various stress factors have been shown to trigger the production of secondary metabolites (Ramakrishna and Ravishankar 2011). Consequently, one could hypothesize that cytolines act as a mutation per se, thus resembling the widely used mutations and stress factors as tools for studying RS. The high number of genes mapped to the "*cell wall*" bin of MapMan may be explained by the high complexity of the cell wall itself, a structure that is only present in plants and whose maintenance requires an intensive traffic through the secretory pathways (Kim and Brandizzi 2014). Indeed, several of our candidate genes code for proteins that are targeted to the secretory pathway (fig. 6). *Lipid metabolism* is the third most abundant bin, grouping 14 of the 96 genes. Lipids play important roles for instance in cell signaling (Hannun and Obeid 2008), thus conferring potential roles as relays to some of those 14 proteins. For example, GRMZM2G060886 and GRMZM2G093666 putative proteins are part of the adenosyl-L-methionine-dependent methyltransferases superfamily, whose members are involved in signal transduction (Schubert et al. 2003). Several other genes, including STN7-ortholog, which are not part of the three most abundant classes mentioned here, may play a role in relaying signals received from the retrograde pathways: (1) GRMZM2G122715 is a putative calcium sensing receptor, with a role in sensing and signaling environmental stimuli in *Arabidopsis* (Bordo and Bork 2002); (2) GRMZM2G454952, a ZDS (zeta carotene desaturase) putative protein, has been shown to act in a signaling pathway that controls chloroplast and leaf development. Furthermore, in cases of ZDS deficiency numerous nuclear and chloroplast genes are differentially regulated, causing abnormal leaf development (Avendaño-Vázquez et al. 2014); (3) GRMZM2G106578 is a putative DGK5 (diacylglycerol kinase). These kinases catalyze the conversion of DAG (diacylglycerol), a known second messenger, to PA (phosphatidic acid), playing a central role in cell signaling (Mérida et al. 2008).

Genes of Interest Have Syntenic Orthologs in the Grasses and Homology to Human Proteins

To further build upon the importance of our genes of interest, we identified their syntenic orthologs (i.e., genes that occur in the same order on a chromosome) in four more plant species from the *Poaceae* family, using data from Schnable et al. (2012). We identified syntenic orthologs for the vast majority of the 96 genes (fig. 3 and supplementary table S2, Supplementary Material online) and conclude that they play important roles in the cell, which are conserved throughout the *Poaceae* lineage. Aside from the grasses, we identified 82 *Arabidopsis* homologs, hinting towards a general mechanism for the retrograde control of NGE in plants, where the genes identified here play key roles.

In addition, there were 35 human proteins with significant homology to our maize sequences. About 7 of the 35 proteins (GRMZM2G004528, GRMZM2G010348, GRMZM2G029566, GRMZM2G070199, GRMZM2G345493, GRMZM2G098346, GRMZM2G155242) exhibit >50% identity with the human homologous protein sequences and a query coverage >90% (fig. 3). Based on their conservation within the animal lineage we hypothesize that they are potential targets of the mitochondrial RS, rather than chloroplast. The first five have not been functionally characterized in maize, but their *Arabidopsis* homologs retain the function described in human (supplementary table S3, Supplementary Material online). The last two are part of the "classical" gene set of maize (Schnable and Freeling 2011), which includes approximately 500 genes that are supported by at least three publications and have mutant phenotype data available. Their function diverged from that in humans. Interestingly, *GRMZM2G098346*, which codes for alcohol dehydrogenase-2 (*Adh2*), originated from a duplication event of *Adh1* 65 million years ago, before the radiation of grasses (Gaut et al. 1999). Therefore, all grasses have this second copy, whose role remains elusive, but could be involved in RS, as we show here.

Previous studies have proven that the process of mitochondrial RS is generally conserved in the animal lineage, including humans (reviewed in Liu and Butow 2006). Thus, *Saccharomyces cerevisiae* has mainly been used to study mitochondrial RS in a eukaryotic cell. Based on the ancient endosymbiosis of the mitochondrion with the eukaryotic cell we can speculate that parts of these pathways are conserved in the plant lineage, too; the five *Arabidopsis* genes that retain the same function as in human support this hypothesis.

Most of the Proteins Are Targeted to the Chloroplast

In terms of cellular localization, most of the proteins that carry a signal peptide are targeted to the chloroplast, whereas the mitochondria and the secretory pathways are less

represented, in accordance with their reduced proteomes. The chloroplast proteome is composed of 3,500–4,000 polypeptides (Soll and Schleiff 2004), almost twice that of the mitochondria (2,000–3,000) (Millar et al. 2005) and 3 times larger than that of the secretory pathway (1,400) (Gilchrist et al. 2006; Rojo and Denecke 2008). The estimate for the secretory pathway may not be accurate for plants, where gene families involved in trafficking have expanded considerably to cope with the characteristics of the endomembrane system, which differs from that of animals (Rojo and Denecke 2008). Thus, when only RC values ≤ 2 are considered (i.e., strongest prediction by TargetP), there are more proteins targeted to the secretory pathway, compared with the mitochondria, but the majority is still composed of chloroplast-targeted proteins (fig. 6).

Potential Roles of Genes Identified Here in the Already Described RS Pathways

There are four widely accepted plastid RS pathways: tetrapyrrole intermediate biosynthesis, plastid gene expression (PGE), plastid redox state, and reactive oxygen species (ROS). ABI4 is a transcription factor that integrates signals from the first three (Chi et al. 2013) but also those received through the mitochondrial retrograde pathway (Giraud et al. 2009). It binds the CCAC motif found in the promoter region of a plethora of genes, including other transcription factors (Koussevitzky et al. 2007; León et al. 2012). Among the 96 gene identified here there are six that have more than ten binding sites for ABI4 in their promoter region (fig. 3 and supplementary table S2, Supplementary Material online), hinting towards their role in the nuclear response to signals coming from the organelles. Three others have more than ten GATA motifs, which are part of light-regulation of transcription (Reyes et al. 2004), thus linked to the tetrapyrrole intermediate biosynthesis pathway of RS. Also, signals from the plastid redox state are initiated in the plastid with the participation of STN7 (Pesaresi et al. 2009; Dietzel et al. 2015). Its maize ortholog, *GRMZM2G010349*, is among the 96 genes we have identified as differentially regulated in the three cytolines. A number of other members from our gene set have *Arabidopsis* homologs that act under the control of the plastoquinone redox state (Jung et al. 2013). These are: *GRMZM2G155242, GRMZM2G140994, GRMZM2G401970, GRMZM2G147701, GRMZM2G013357, GRMZM2G044107, GRMZM2G103197, GRMZM2G122715, GRMZM2G174598*. The *Arabidopsis* homologs for these genes are listed in supplementary table S3, Supplementary Material online. Furthermore, *GRMZM2G004534* is a putative pyruvate kinase, which could be involved in the MAPK cascades that are characteristic to the fourth RS pathway, i.e., reactive oxygen species (Chi et al. 2013). All of the above indicate that many of the genes we have identified here are potentially involved at different levels in the four RS pathways currently defined for plastid to nucleus communication. By means of ABI4 transcriptional activity, which integrates signals coming from the mitochondria, they may also act as nuclear receptors for this organelle's retrograde signals.

The signals relayed through the four plastid RS pathways are grouped into two categories according to their function: biogenic control and operational control (reviewed in Pogson *et al.* 2008). The first category includes signals related to organelle biogenesis, which are mainly generated during early plant development, whereas the latter responds to developmental and environmental cues that command adjustments of the energy metabolism. Our experimental setup captures both modes of RS, biogenic and operational, probing NGE in 9-day-old seedlings. However, it does not differentiate between the two.

Co-Localization of Genes and QTLs Defined for Two Important Agronomic Traits

Plant height and ear height are two of the most important agronomic traits for maize, directly linked to biomass production and yield, respectively. Tang et al. (2013) demonstrated that 39.91% of the phenotypic variation observed for ear height and 8.75% for plant height was due to the influence of the cytoplasmic environment on the nucleus. We hypothesize that the 23 genes identified here to co-localize with the QTLs defined by Tang et al. (2013) are nuclear targets of the cytoplasmic signals that cause the observed phenotypic variation. Nine of those genes are part of either lipid metabolism (5) or cell wall (4) bins defined by MapMan (supplementary tables S2 and S3, Supplementary Material online), two classes of importance for the phenotypic traits of interest. *GRMZM2G443715*, e.g., is involved in cellulose synthesis, whereas *GRMZM2G157113* is involved in fatty acid synthesis and elongation. *GRMZM2G348551* is one of the "classical" genes of maize (*Sugary2*) and functions as a starch synthase. However, other genes are involved in lipid or cell wall degradation, an indication that there is a complex interplay among them leading to the observed phenotypes (supplementary table S3, Supplementary Material online).

Transcript De Novo Assembly of the Reads Not Matching the Reference Genome

Because of the incomplete status of the B73 genomic sequence (Schnable et al. 2009), one would expect that not all of the transcripts identified in an RNA-seq experiment of another inbred line would match the reference. Plus, the average rate of polymorphism in two maize inbred lines is ten times higher than that in humans and higher than that observed between humans and chimpanzees (Buckler et al. 2006). Copy number variation (CNV) and presence/absence variation (PAV) are also high when two maize inbred lines are compared (Springer et al. 2009). In this context it is conceivable that new transcripts are identified every time a new

inbred line is investigated. Here, we used Trinity to perform a *de novo* assembly of the reads that did not map to the reference genome and then subjected them to BLAST2GO analysis (Conesa et al. 2005). Next, we compared those *de novo* transcripts that are common to the two cytoplasm-donor lines versus the nucleus-donor line (i.e., common to Cyt1 and Cyt2 compared with B). We hypothesize that the ~ 600 transcripts identified (supplementary table S5, Supplementary Material online) are the result of a change in the methylation status of the genes encoding them. They only become active in the new cytoplasmic environments of Cyt 1 and Cyt2. Future work will need to validate the presence of these predicted transcripts before probing the methylation status of the respective gene bodies and promoter sequences in the nucleus-donor line and the two cytoplasm-donor lines. The approximately 100 transcripts included in GO terms related to gene expression (GO:0006355, GO:0006351, GO:1903506, and GO:0010468) (supplementary figs. S1 and S2, Supplementary Material online) are of particular interest, as they could further impact NGE in a cascade effect (e.g., the transcripts that are included in GO:0006355 regulate those in GO:0006351).

Conclusions

Through a laborious breeding process that took >10 years to complete we have created three cytolines, sharing the same nucleus but different organellar genomes in their corresponding cytoplasmic environments. We used an Illumina HiSeq 2500 instrument to sequence their transcriptome and identified 96 key nuclear genes, which integrate signals coming through the retrograde pathways. Our approach differs from previous studies through the use of cytolines, rather than the use of mutants that are defective in the RS pathways or cells that are under some sort of stress. This allowed us to investigate RS in a normal state of the cell. In total, 96 genes are differentially regulated in both Cyt1 and Cyt2 compared with the nucleus donor line B. They have a ubiquitous expression pattern and the vast majority of them have a syntenic ortholog in the four other grass species investigated, as well as an orthologous copy in *A. thaliana*. Therefore, these findings contribute to the paradigm we use to describe the RS in plants. Concurrently, we present strong evidence that at least 7 of the 96 genes are well conserved in the animal lineage, representing potential targets in a mitochondria-to-nucleus communication, for which no distinct pathway has been described so far.

Supplementary Material

Supplementary figures S1 and S2 and tables S1–S5 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We thank Dr Hugo Dooner (Rutgers University) for critically reviewing the article. Help from Michèle Ackermann and Muriel Fragnière is also acknowledged for their excellent technical support, and the Next Generation Sequencing Platform of the University of Bern for performing the high-throughput sequencing experiments. This work was supported by two grants of the Romanian National Authority for Scientific Research, CNDI-UEFISCDI, numbers PN-II-PT-PCCA-2011-3.1-0511-103/2012 and PN-II-RU-TE-2014-4-1767-41/2015 to MM, who was also partially supported by a Sciex-NMSch grant, number 13.334/2014 to RB and PN 16 19 BIODIVERS, institutional funding. We also thank the canton of Berne for financial support.

Literature Cited

Allen JO. 2005. Effect of teosinte cytoplasmic genomes on maize phenotype. Genetics 169:863–880.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. Genome Biol. 11:R106.

Avendaño-Vázquez A-O, et al. 2014. An uncharacterized apocarotenoid-derived signal generated in ζ-carotene desaturase mutants regulates leaf development and the expression of chloroplast and nuclear genes in Arabidopsis. Plant Cell 26:2524–2537.

Bendich AJ. 1987. Why do chloroplasts and mitochondria contain so many copies of their genome?. BioEssays 6:279–282.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach tomultiple testing. J R Statist Soc Ser B. 57:289–300.

Bordo D, Bork P. 2002. The rhodanese/Cdc25 phosphatase superfamily. Sequence-structure-function relations. EMBO Rep. 3:741–746.

Bradbeer W, Atkinson Y, Börner T, Hagemann R. 1979. Cytoplasmic synthesis of plastid polypeptides may be controlled by plastid-synthesised RNA. Nature 279:816–817.

Buckler ES, Gaut BS, McMullen MD. 2006. Molecular and functional diversity of maize. Curr Opin Plant Biol. 9:172–176.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Cavelier L, Johannisson A, Gyllensten U. 2000. Analysis of mtDNA copy number and composition of single mitochondrial particles using flow cytometry and PCR. Exp Cell Res. 259:79–85.

Chi W, Sun X, Zhang L. 2013. Intracellular signaling from plastid to nucleus. Annu Rev Plant Biol. 64:559–582.

Conde MF, Pring DR, Levings CS. 1979. Maternal inheritance of organelle DNA's in Zea mays-Zea perennis reciprocal crosses. J Hered. 70:2–4.

Conesa A, et al. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–3676.

Crosatti C, et al. 2013. Cytoplasmic genome substitution in wheat affects the nuclear-cytoplasmic cross-talk leading to transcript and metabolite alterations. BMC Genomics 14:1–22.

Dietzel L, et al. 2015. Identification of early nuclear target genes of plastidial redox signals that trigger the long-term response of arabidopsis to light quality shifts. Mol Plant. 8:1237–1252.

Dimitrov LN, Brem RB, Kruglyak L, Gottschling DE. 2009. Polymorphisms in multiple genes contribute to the spontaneous mitochondrial genome instability of *Saccharomyces cerevisiae* S288C strains. Genetics 183:365–383.

Emanuelsson O, Brunak S, Heijne von G, Nielsen H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc. 2:953–971.
Gaut BS, Peek AS, Morton BR, Clegg MT. 1999. Patterns of genetic diversification within the Adh gene family in the grasses (*Poaceae*). Mol Biol Evol. 16:1086–1097.
Gilchrist A, et al. 2006. Quantitative proteomics analysis of the secretory pathway. Cell 127:1265–1281.
Giraud E, Van Aken O, Ho L, Whelan J. 2009. The transcription factor ABI4 is a regulator of mitochondrial retrograde expression of alternative oxidase1a. Plant Physiol. 150:1286–1296.
Grabherr MG, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29:644–652.
Gray JC, Sullivan JA, Wang J-H, Jerome CA, MacLean D. 2003. Coordination of plastid and nuclear gene expression. Philos Trans R Soc B Biol Sci. 358:135–144.
Hannun YA, Obeid LM. 2008. Principles of bioactive lipid signalling: lessons from sphingolipids. Nat Rev Mol Cell Biol. 9:139–150.
Higo K, Ugawa Y, Iwamoto M, Korenaga T. 1999. Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Res. 27:297–300.
Hirsch CN, et al. 2014. Insights into the maize pan-genome and pan-transcriptome. Plant Cell 26:121–135.
Joseph B, Corwin AJ, Baohua L, Atwell S, Kliebenstein DJ. 2013. Cytoplasmic genetic variation and extensive cytonuclear interactions influence natural variation in the metabolome. eLife 2:e00776:1–21.
Jung H-S, Chory J. 2010. Signaling between chloroplasts and the nucleus: can a systems biology approach bring clarity to a complex and highly regulated pathway?. Plant Physiol. 152:453–459.
Jung H-S, et al. 2013. Subset of heat-shock transcription factors required for the early response of Arabidopsis to excess light. PNAS 110:14474–14479.
Kim D, et al. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14:R36.
Kim S-J, Brandizzi F. 2014. The plant secretory pathway: an essential factory for building the plant cell wall. Plant Cell Physiol. 55:687–693.
Koussevitzky S, et al. 2007. Signals from chloroplasts converge to regulate nuclear gene expression. Science 316:715–719.
Lawrence CJ, Schaeffer ML, Seigfried TE, Campbell DA, Harper LC. 2007. MaizeGDB's new data types, resources and activities. Nucleic Acids Res. 35:D895–D900.
León P, Gregorio J, Cordoba E. 2012. ABI4 and its role in chloroplast retrograde communication. Front Plant Sci. 3:1–13.
Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659.
Liu Z, Butow RA. 2006. Mitochondrial retrograde signaling. Annu Rev Genet. 40:159–185.
Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15:550.
Mérida I, Ávila-Flores A, Merino E. 2008. Diacylglycerol kinases: at the hub of cell signalling. Biochem J. 409:1–18.
Millar AH, Heazlewood JL, Kristensen BK, Braun H-P, Møller IM. 2005. The plant mitochondrial proteome. Trends Plant Sci. 10:36–43.
Moison M, et al. 2010. Cytoplasmic phylogeny and evidence of cyto-nuclear co-adaptation in Arabidopsis thaliana. Plant J. 63:728–738.
Ng S, et al. 2014. Anterograde and retrograde regulation of nuclear genes encoding mitochondrial proteins during growth, development, and stress. Mol Plant. 7:1075–1093.
Pesaresi P, et al. 2009. Arabidopsis STN7 kinase provides a link between short- and long-term photosynthetic acclimation. Plant Cell 21:2402–2423.
Pogson BJ, Woo NS, Förster B, Small ID. 2008. Plastid signalling to the nucleus and beyond. Trends Plant Sci. 13:602–609.
Ramakrishna A, Ravishankar GA. 2011. Influence of abiotic stress signals on secondary metabolites in plants. Plant Signal Behav. 6:1720.
Rand DM. 2005. Nuclear-mitochondrial epistasis and drosophila aging: introgression of *Drosophila simulans* mtDNA modifies longevity in *D. melanogaster* nuclear backgrounds. Genetics 172:329–341.
Reyes JC, Muro-Pastor MI, Florencio FJ. 2004. The GATA family of transcription factors in Arabidopsis and rice. Plant Physiol. 134:1718–1732.
Rhoads DM, Subbaiah CC. 2007. Mitochondrial retrograde regulation in plants. Mitochondrion 7:177–194.
Ritchie ME, et al. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43:e47:1–13.
Rojo E, Denecke J. 2008. What is moving in the secretory pathway of plants?. Plant Physiol. 147:1493–1503.
Roubertoux PL, et al. 2003. Mitochondrial DNA modifies cognition in interaction with the nuclear genome and age in mice. Nat Genet. 35:65–69.
Schnable JC, Freeling M. 2011. Genes identified by visible mutant phenotypes show increased bias toward one of two subgenomes of maize. PLoS One 6:e17855.
Schnable JC, Freeling M, Lyons E. 2012. Genome-wide analysis of syntenic gene deletion in the grasses. Genome Biol Evol. 4:265–277.
Schnable PS, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115.
Schubert HL, Blumenthal RM, Cheng X. 2003. Many paths to methyltransfer: a chronicle of convergence. Trends Biochem Sci. 28:329–335.
Sekhon RS, et al. 2011. Genome-wide atlas of transcription during maize development. Plant J. 66:553–563.
Smit AFA, Hubley R, Green P, eds. 2014. *RepeatMasker*. 4th ed. http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker.
Smith JSC, Smith OS. 1987. Associations among inbred lines of maize using electrophoretic, chromatographic, and pedigree data. Theor Appl Genet. 73:654–664.
Soliman K, Fedak G, Allard RW. 1987. Inheritance of organelle DNA in barley and *Hordeum* × *Secale* intergeneric hybrids. Genome 29:867–872.
Soll J, Schleiff E. 2004. Plant cell biology: protein import into chloroplasts. Nat Rev Mol Cell Biol. 5:198–208.
Springer NM, et al. 2009. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet. 5:e1000734.
Sugita M, Sugiura M. 1996. Regulation of gene expression in chloroplasts of higher plants. Plant Mol Biol. 32:315–326.
Tang Z, et al. 2013. Cytonuclear epistatic quantitative trait locus mapping for plant height and ear height in maize. Mol Breed. 31:1–14.
Tao D, et al. 2004. Cytoplasm and cytoplasm-nucleus interactions affect agronomic traits in Japonica rice. Euphytica 135:129–134.
Thimm O, et al. 2004. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J. 37:914–939.
Troyer AF. 1999. Background of US hybrid corn. Crop Sci. 39:601–626.
Unseld M, Marienfeld JR, Brandt P, Brennicke A. 1997. The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. Nat Genet. 15:57–61.
Usadel B, et al. 2009. A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, maize. Plant Cell Environ. 32:1211–1229.
Wolfe KH, Li WH, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. PNAS 84:9054–9058.

Associate editor: Bill Martin