Supplementary Material

Suppl. Figure 1: Plot of the dN/dS (ω) and p-value

Depending on settings different number of the previously identified CPuORFs were identified compared with the total number of uSSRs identified using the pipeline.

Suppl. Figure 2: 5’ leader annotation is not limiting for classification of conservation. Boxplots with the number of species of which the most conserved gene of the CPuORFs possesses a 5’ leader sequence. All CPuORFs with a stop codon in the 5’ leader that were discovered in this study were used. The red diamonds indicate the averages and open circles indicate individual data points. Boxplots were made using the R package “ggplot2”.

Suppl. Figure 3: Sequence alignments of the newly discovered CCRs directly upstream of and in frame with the mORF.
Amino acid sequence alignments of predicted non-AUG initiating main ORFs. Area aligned is from last in frame stop codon till the annotated AUG start codon for each of the genes with discovered conserved 5’ extended mORF. Aligned using MAFFT v.7.307 (FFT-NS-2) and displayed using Jalview v. 2.10. Species abbreviations can be found in supplemental table 2.

Suppl. Figure 4: Sequence alignments of the newly discovered CPuORFs.
Amino acid sequence alignments of novel discovered CPuORFs. Area aligned is from last in frame stop codon to the next in frame stop codon for each of the genes with novel discovered CPuORFs. Aligned using MAFFT v.7.307 (FFT-NS-2) and displayed using Jalview v. 2.10. Species abbreviations can be found in supplemental table 2.

Suppl. Figure 5: Sequence alignments of the newly discovered CPuORFs with stop codon in main ORF.
Amino acid sequence alignments of novel discovered CPuORFs with stop codon in the main ORF. Area aligned is from last in frame stop codon to the next in frame stop codon for each of the genes with novel discovered CPuORFs. Aligned using MAFFT v.7.307 (FFT-NS-2) and displayed using Jalview v. 2.10. Species abbreviations can be found in supplemental table 2.

Suppl. Figure 6: Sequence alignment of AT1G01060 CPuORF.
Red arrows indicate position of the mORF start codon (other frame). Aligned using MAFFT v. 7.307 (FFT-NS-2) and displayed using Jalview v. 2.10. Species abbreviations can be found in supplemental data 1

Suppl. Figure 7: Boxplots of uCCR lengths. Lengths between the first conserved amino acid and the stop codon or mORF start codon (light blue plots) and lengths between the first conserved amino acid and the last conserved amino acid (light red plots) were determined for the three types of uCCRs. Conservation of amino acids were determined using the program “cons” from the EMBOSS package (version 6.6.6.0) with “-setcase” as 0.75 multiplied by the number of sequences in the alignment. The alignments show in suppl. figures 1-3 were used as input. The red diamonds indicate the averages and open circles indicate individual data points. Boxplots were made using the R package “ggplot2”.

Suppl. Figure 8: comparison of mRNA and peptide sequence of S6Kinase uCCR in different species.

Suppl. Table 1: Overview with number of genes transcripts and SSR remaining after each step of the pipeline.

Suppl. table 2: Overview of used species and their genome sources.

Suppl. Table 3: Identification of previously discovered CPuORFs.
Our pipeline was evaluated on CPuORFs discovered in four previous studies. Arabidopsis-rice (ATH-OSA) comparison and Arabidopsis-Arabidopsis comparison was performed by (Hayden and Jorgensen, 2007), BAIUCAS pipeline by (Takahashi *et al.*, 2012), MOTIF based search by (Vaughn *et al.*, 2012) and (Laing *et al.*, 2015) discovered two CPuORFs experimentally. Green checked boxes indicate that there was a SSR extracted in the 5' leader of the gene of the CPuORF during this step of the pipeline. Yellow "-" indicates that the CPuORF was not further analyzed, and a red "X" indicates that the CPuORF was not extracted after the indicated filtering step.

Suppl. Table 4: Predicted localization of proteins with (pep_morf) of without (morf) translation initiation from predicted alternative start site.
Three different programs were used to predict the localization of these proteins.

Suppl. Data 1: Overview of all CCRs discovered by our pipeline.
The SSR's location in the 5' leader of the mRNA in Arabidopsis, presence in the different clades, and p- and omega-values of the purifying selection method are shown.

Suppl. Data 2: Fasta file containing peptide sequences used for subcellular localization predictions.
Indicated in the fasta file is the transcript id followed by the SSR number and whether the sequence includes (pep_morf) or excludes (morf) the amino acid sequence from the non-AUG start codon to the annotated AUG start codon.