**Article**

**Copaifera langsdorffii** novel putative lncRNA: interspecies conservation analysis in adaptive response to different biomes

Monica F. Danilevicz1, Kanhu C. Moharana2, Thiago M. Venancio2, Luciana O. Franco3, Sérgio R.S. Cardoso1, Mónica Cardoso3, Flávia Thiebaut1, Adriana S. Hemerly1, Francisco Prosdocimi4, Paulo C.G. Ferreira1,*

1. Introduction

*Copaifera* is a genus of native trees from Latin America tropical regions and Western Africa. The species *C. langsdorffii*, *C. reticulata*, *C. cearensis*, *C. multijuga* among others are popularly known as “Copaiba” [69]. The *Copaifera sp.* oil is extracted through V-shaped cut in the stem bark and has been used by indigenous populations from the Brazilian Amazon as a powerful antimicrobial, anti-inflammatory and “overall healing purposes” [69]. Many of this alleged features have been investigated by pharmacological studies, confirming its anti-inflammatory capacity [7, 20, 70] and proving it to be an efficient alternative to treat dental infections [61], gastrointestinal disorders [50], endometriosis [59], skin ulcers [23, 41, 49], and to be applied to skin scaffold implants increasing tissue angiogenesis [43]. The Copaiba oil resin has been shown to exert larvicidal activity against *Aedes aegypti* [51, 58], and antibacterial activity *in vitro* [71]. The leaf extracts also presented leishmanicidal and antimalarial activities [36], also being effective as a biopesticide against lepidoptera [53].

The amount of Copaiba oil production is influenced by the climate and soil conditions, as more oil is produced at locations with clay soil, during the raining season [2]. In Brazil, *C. langsdorffii* is a
widely occurring species, included in a broad range of ecosystems like Cerrado, Atlantic Rain Forest and Caatinga, which are very distinct biomes [15], requiring diverse adaptive mechanisms. Such plasticity to adapt to different biomes is a complex regulation, involves several genetic, evolutionary and epigenetic fine tuning, which may also include long non-coding RNAs (lncRNAs).

LncRNAs are considered to be RNA transcripts, defined as longer than 200bp, with no apparent protein coding capacity [30, 31]. They hold many resemblances with mRNA, such as similar epigenetic marks to promote expression, binding sites for RNA polymerase II at their genome loci. They are often polyadenylated at 3' end and receive 5' CAPs, especially when acting outside the nucleus [30, 31]. Because they are (i) less conserved than mRNA at the sequence level across different species and (ii) regularly transcribed at low levels, lncRNAs were once considered transcriptional noise. LncRNAs tends to form secondary and tertiary structures, in which the molecule conformation is crucial to regulate their targets, leading to their functional domains and genomic positions to be more conserved than their sequence [31].

LncRNAs interspecies conservation may be influenced by the regulatory mechanism they play, whereas their sequence conservation depends on whether (a) the lncRNA molecule acts as a regulator or (b) the simply transcription of the lncRNA regulates its target gene [30]. Most functional lncRNAs undergo post-transcriptional processing and retain higher conservation of splice sites. This indicates that they most likely function in the mature form [24, 68]. In vertebrates, it is argued that transposable elements and bidirectional transcription may play an important role in the evolution and rapid turnover of lncRNAs [30].

There are several types of lncRNAs classified according to their genomic positions as: sense, natural antisense, bidirectional, intronic and intergenic; and they may also act through cis or trans regulation [5, 10, 55]. LncRNAs exhibit relatively low expression patterns compared to mRNAs, showing a specific profile depending on the (i) tissue or cell type observed [10], (ii) developmental stage and (iii) environmental stress response [9, 14].

A considerable amount of lncRNAs may act as chromatin regulators [68, 52]. For example, APOLO, a lncRNA responsive to auxin, interacts with the chromatin, leading to a loop formation encompassing the PID gene (key regulator of polar auxin transport) and regulating its expression [4]. HOTAIR, ANRIL and KCNQ1OT1 also are known to bind to more than one histone-modifying complex, acting as regulators [34]. COLDAIR and COLDWRAP are lncRNA described to act regulating the FLOWERING LOCUS C (FLC), associate to Polycomb protein complex to stably repress FLC during the vernalization process, prompting adaptive fitness and development in Arabidopsis thaliana [33]. In rice, it was shown that photoperiodic-sensitive male sterility (PSMS) is carried out by a lncRNA called LDMAR, in which a SNP led to change in secondary structure and subsequently repression of its expression at long day conditions [19]. ENOD40 and ASCO are examples of conserved lncRNA among legumes, which are involved in the re-localization of Nuclear Speckles RNA Binding Protein (NSRs) in Medicago truncatula and A. thaliana, albeit ASCO interaction with AtNSRs induces alternative splicing activity in auxin signaling [6, 13]. Several other lncRNA have their functions experimentally tested, such as HDI promoting photomorphogenesis in red light [28] and NERDL association to wood formation in P. tormentosa [56].

In response to inorganic phosphorus starvation, the lncRNA cis-NAT PHO1;2 acts regulating Pi homeostasis as a translational enhancer of OsPHO1;2 increasing its uptake when overexpressed in a mutated rice lineage [27]. Evidence of lncRNA involved in stress response and adaptation was also observed in nitrogen deficient Populus [16], maize under drought stress [86] and A. thaliana submitted to salt stress [62]. Moreover, lncRNAs have also been found to act as sRNA precursors [9, 16, 40, 74, 82] or acting as bait to sRNA regulation by target mimicry [21, 57, 9]; some lncRNA are target mimics to the plant miRNA response to viral infection [74]. Collectively, lncRNA have been implicated in various cell and molecular processes, including post-transcriptional regulation, post-translational regulation of protein activity and protein re-localization, organization of protein complexes, cell–cell signaling and intrinsically connected to adaptive fitness and overall plant processes [5, 8, 55].
In this study, we identified 8,020 novel lncRNAs through bioinformatics analysis of which approximately 565 were shown to be up regulated above 5 times in CER or ARF populations, which might be related to their adaptation to such diverse environments. LncRNAs tend to have their primary sequence evolving under a relaxed constraint, being less likely to present high interspecies sequence similarity. Nevertheless, we were able to identify one single transcript that matched a *Glycine max* lncRNA from CANTATA database [63]. The high sequence similarity between this transcript from soybean and Copaiba indicates that it is potentially functional. Considering also that it was regulated under the conditions tested, it may be involved in the adaptive response. In addition, we were able to map 156 lncRNAs to seven *Fabaceae* genomes, of which ten transcripts exhibited more than 5-fold up regulation among the conditions. The emerging epigenetic studies involving lncRNAs function and conservation have shown their involvement to several types of biotic and abiotic stresses. Thus, the conservation of lncRNAs among *Fabaceae* species validates their transcription. The differential expression observed for hundreds of lncRNAs suggests that they take part in regulatory pathways that lead to adaptive responses in Copaiba.

2. Results

2.1. Identification of novel and differentially expressed lncRNA

In order to identify lncRNAs and elucidate their adaptive roles, we sequenced the transcriptomes of *C. langsdorffii* leaves obtained from trees growing at two different ecosystems, the (i) Atlantic Rain Forest (ARF, humid condition) and the (ii) Cerrado (CER, dry forest) in Brazil during the drought season for both locations. A total of 138,175 and 199,556 transcripts were assembled from ARF and CER, respectively. On these transcripts, a series of filters were applied. Initially we removed sequences shorter than 200bp and those containing ORFs encoding peptides of at least 100 aa. Next, we ran CPC (v0.9-r2) [35] and PLEK (v1.2) [37] to assess the protein coding potential of these transcripts. Transcripts classified as non-coding by both software were compared to establish one-to-one correspondence between ARF and CER transcripts. This correspondence was defined using bi-directional BLASTN (v2.2.31+) [12] and MCL (implemented in scps, v.0.9.8[45]). Reads were remapped against the assembled transcriptomes with Bowtie2 [64] and transcriptional levels estimated with cufflinks v2.2.1 [64]. Transcripts with at least 1 RPKM (Reads Per Kilobase of transcript per Million mapped reads) were kept for downstream analyses (8,020 transcripts). The majority of these (2,327 transcripts) presented over two fold regulation, while 566 transcripts were regulated above 5 fold (Figure 1, supplementary table 1).

![Figure 1: Putative lncRNA identified in ARF and CER samples and its fold change regulation in comparison to each other.](image-url)

There were 2893 differentially regulated transcripts identified from a total of 8020 copaiba lncRNAs. The majority of the transcripts are 2 to 5 times differently expressed on either sample, yet there are 581 transcripts regulated above 5 fold in either sample.
Table 1: Stem-loop secondary structure of highly regulated conserved copaiba lncRNA: The selected RNA have their MFE estimated by ViennaRNA package. The secondary structure is deemed to be stable when the MFE ≤ -80 kcal, thus from the 12 transcripts.

<table>
<thead>
<tr>
<th>ARF ID</th>
<th>CER ID</th>
<th>ARF RPKM</th>
<th>CER RPKM</th>
<th>ARF/CER</th>
<th>CER/ARF</th>
<th>MFE (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>JCF44_0000048876</td>
<td>JCF45_0000090780</td>
<td>1269.38</td>
<td>4.83663</td>
<td>262.4513349</td>
<td>0.00381023</td>
<td>-209.9</td>
</tr>
<tr>
<td>JCF44_0000039348</td>
<td>JCF45_0000129049</td>
<td>149.628</td>
<td>3.75544</td>
<td>39.84300109</td>
<td>0.01322936</td>
<td>-189.2</td>
</tr>
<tr>
<td>JCF44_0000061237</td>
<td>JCF45_000001210</td>
<td>286.172</td>
<td>9.83557</td>
<td>29.09561927</td>
<td>0.034369435</td>
<td>-220.5</td>
</tr>
<tr>
<td>JCF44_0000021616</td>
<td>JCF45_0000120089</td>
<td>80.4121</td>
<td>3.17854</td>
<td>25.29843891</td>
<td>0.039528131</td>
<td>-125.5</td>
</tr>
<tr>
<td>JCF44_0000041906</td>
<td>JCF45_0000057106</td>
<td>90.1082</td>
<td>3.75251</td>
<td>24.01278078</td>
<td>0.04164449</td>
<td>-52.5</td>
</tr>
<tr>
<td>JCF44_0000027393</td>
<td>JCF45_0000069991</td>
<td>55.4674</td>
<td>3.02796</td>
<td>18.3175141</td>
<td>0.05492561</td>
<td>-70.6</td>
</tr>
<tr>
<td>JCF44_0000021638</td>
<td>JCF45_0000007995</td>
<td>9.75664</td>
<td>108.186</td>
<td>0.090183942</td>
<td>11.0884848</td>
<td>-106.1</td>
</tr>
<tr>
<td>JCF44_0000094004</td>
<td>JCF45_0000091462</td>
<td>11.7198</td>
<td>131.109</td>
<td>0.089389744</td>
<td>11.1869565</td>
<td>-151.9</td>
</tr>
<tr>
<td>JCF44_0000075087</td>
<td>JCF45_0000057481</td>
<td>12.3338</td>
<td>138.6</td>
<td>0.088988456</td>
<td>11.23741264</td>
<td>-118.9</td>
</tr>
<tr>
<td>JCF44_0000042254</td>
<td>JCF45_0000109819</td>
<td>2.39024</td>
<td>28.4969</td>
<td>0.083877194</td>
<td>11.92219191</td>
<td>-162.3</td>
</tr>
<tr>
<td>JCF44_0000040403</td>
<td>JCF45_0000101583</td>
<td>80.5029</td>
<td>1132.28</td>
<td>0.07109805</td>
<td>14.06508337</td>
<td>-87.9</td>
</tr>
</tbody>
</table>

2.2. Interspecies lncRNA conservation analysis

2.2.1. Positional conservation and genome alignment analysis

Since there is no reference genome for *C. langsdorffii*, we performed a similarity search using Bowtie2 (v2.3.4.1) [64] against the genomes of the following related species: (i) *Vicia faba*, (ii) *Glycine max*, (iii) *Medicago truncatula*, (iv) *Phaseolus vulgaris*, (v) *Lotus japonica*, (vi) *Vigna unguiculata* and (vii) *Cicer reticulatum*. Genome sequences were downloaded from NCBI Genomes (ftp://ftp.ncbi.nlm.nih.gov/genomes/). From the alignments, we found 1,747 and 1,879 transcripts respectively from ARF and CER samples to match at least one of genomes used. Using DrawVenn (http://bioinformatics.psb.ugent.be/webtools/Venn) application we were able to identify 156 transcripts that aligned to all seven *Fabaceae* genomes in both ARF and CER samples (Figure 2).
represents the lncRNA gather from the ARF samples and the red line represents the lncRNA gather from the CER sample, their alignment profile is very similar as expected, around 1800 transcripts aligned to at least one of the genomes used.

From the subset of 156 copaiba lncRNAs aligned to all seven Fabaceae genomes, 45 transcripts presented above 2-fold differential regulation (Figure 3). To understand if the copaiba lncRNA transcripts aligned to all seven Fabaceae genomes hold positional conservation, we selected 10 transcripts exhibiting at least 5-fold up-regulation to retrieve the information regarding their genomic locations. For this analysis, we used the reference genomes of P. vulgaris, G. max and M. truncatula, as they are better assembled and annotated. However, the genes closely located to the lncRNA loci are described only as “hypothetical protein-coding”, or “plant-like protein”. Further, none of the 6,133 lncRNA annotated in the G. max genome were located close to copaiba putative lncRNA alignment locations. The M. truncatula or P. vulgaris annotation files had no information on lncRNAs.

Figure 3: Subset of transcripts which aligned to all Fabaceae genomes presenting differential expression: From 156 transcripts which aligned to all genomes analyzed, there were 45 copaiba lncRNAs upregulated in either condition represented in this graph. Each transcript is represented by a single bar, in red are indicated the lncRNAs upregulated in CER samples in relation to ARF. In blue are indicated the lncRNAs upregulated in ARF samples in relation to CER. In the x axis is indicated the fold change regulation of the transcripts.

2.2.2. Identification of putative C. langsdorffii lncRNA in EST sequences of other Fabaceae species.

We downloaded EST and cDNA libraries from Phytozome and NCBI databases from C. reticulatum, P. vulgaris, M. truncatula, G. max, V. unguiculata and L. japonicus, to identify whether the copaiba lncRNA transcripts that aligned to the genomes could also be found in available transcriptome libraries, which comprise both mRNA and poly-A ncRNA. LncRNAs have a specific expression profile, considered to be lower than mRNAs. The likelihood of expression is highly influenced by tissue and specific condition [10]; therefore the identification of copaiba lncRNA in Fabaceae transcriptome is susceptible to be underestimated. This is due to the fact that transcripts specific profile of expression and relatively low amounts may not be observed even if present in the reference transcriptomes.

We used BLASTN (v2.2.31+)[12] to search putative copaiba lncRNAs against downloaded transcript sequences. Two identity thresholds, 50% and 90%, were used in this analysis, along with
e-value and coverage threshold of $10^{-15}$ and 50%, respectively. We found that 27% (2,194 lncRNA) aligned to at least one species’ transcriptome (using 50% identity filter on BLASTN) (Figure 4). While only 3.3% (264 lncRNA) aligned with 90% identity to at least one Fabaceae expressed transcripts library and a single transcript (JCF44_0000056614) aligned to five species. It is remarkable to notice that 227 transcripts aligned to *G. max* with more than 90% identity, which might due to *G. max* to be more intensively studied species than the others (Figure 5).

---

**Figure 4:** Count of copaiba lncRNA aligned to each Fabaceae transcriptome with Blastn (50% identity): In this diagram is shown the amount of copaiba lncRNA aligned to the transcriptome of each species, segregated by a color pattern indicated on the legend. In the colored overlapped area are the transcripts which were aligned to more than one species transcriptome, and its respective amount. There are only five species shown in this diagram, for illustration purposes we left out *P. vulgaris*.

**Figure 5:** Count of copaiba lncRNA aligned to each Fabaceae transcriptome with Blastn (90% identity): In this diagram is shown the amount of copaiba lncRNA aligned to the transcriptome of each species, segregated by a color pattern indicated on the legend. In the colored overlapped area
are the transcripts which were aligned to more than one species transcriptome, and its respective amount. There are only five species shown in this diagram, for illustration purposes we left out *P. vulgaris*.

### 2.2.3. Comparison analysis of conserved putative lncRNA

To assess whether the copaiba lncRNAs aligned to *Fabaceae* transcribed libraries were also aligned to *Fabaceae* genomes, we used the Bowtie2 [64] aforementioned result, aiming to determine the amount of transcripts that are present in the genome and also are actively transcribed in the cDNA, EST libraries used (Figure 6, supplementary table 2). Thus, we identified 1,141 transcripts that aligned to both genome and transcriptome comparisons of at least one species. From this total, there were 36 copaiba lncRNA transcripts which aligned to both genome and transcriptome of six *Fabaceae* species (Figure 7).

![Figure 6: Blastn comparison of copaiba lncRNA aligned to each species genome and cDNA](image_url)

In the venn diagram is compared the amount of copaiba transcripts aligned to the genome (blue circle), to the transcriptome (red square) or to both (overlapped area). A is the *V. unguiculata* comparison, B is *M. truncatula*, C is *P. vulgaris*, D is *L. japonicus*, E is *G. max* and F is *C. reticulatum*. It is possible to observe that *G. max*, *L japonicus* and *M. truncatula* presented a higher number of overall aligned transcripts and also overlapped ones.
Figure 7: Total transcripts aligned to Fabaceae genomes and transcriptome: This diagram shows the amount of copaiba IncRNA aligned to the both genome and transcriptome of each species segregated by a color pattern indicated on the legend. In the colored overlapped area are the transcripts which were aligned to more than one species genome and transcriptome, and its respective amount.

2.2.4. Expression analysis of the conserved putative IncRNA among Fabaceae.

Most of the IncRNA transcripts that aligned to both genome and transcriptomes of Fabaceae species displayed similar RPKM values for ARF and CER. However, several IncRNA (254 out of 1,141) were regulated at either condition above 2-fold, indicating that they might be involved with the adaptation to different environments. (Figure 8). From the total transcripts which aligned to both genome and transcriptome of Fabaceae species, it is possible to identify the formation of two clear groups of putative IncRNA formed in each condition. The values and identification of the putative IncRNA are available at supplementary table 3.
Figure 8: Fold change comparison of copaiba lncRNA conserved in Fabaceae species: From 1141 lncRNA which aligned to multiple Fabaceae genome and transcriptome, there was 254 lncRNA regulated above 2 fold in either ARF or CER samples. A displays the regulated transcripts between 2 and 10 fold, the blue bars are up regulated in ARF samples, and red bars are up regulated in CER samples. Each bar corresponds to a single transcript. B displays the regulated transcripts above 10 fold, the blue bars are up regulated in ARF samples, and red bars are up regulated in CER samples. Each bar corresponds to a single transcript.

2.2.5. Identification of Known lncRNA

In order to investigate conserved lncRNAs in closely related species, we used the putative transcripts obtained and compared with BLASTN against CANTATA and GREEnc lncRNA databases for S. bicolor, G. max, M. truncatula and P. vulgaris, which summed up approximately 32,000 sequences. This comparison found a single transcript (JCF45_0000011974/ JCF44_0000015840) that matched a G. max lncRNA from CANTATA database (CNT2032069) with 90% identity and e-value of 4e^{-97}. This single transcript is 6.5 fold up-regulated in ARF samples. This transcript was also conserved in G. max, L. japonicus and P. vulgaris genome analysis.

The high sequence identity of the JCF45_0000011974/JCF44_0000015840 transcript with other legume genomes and with a G. max lncRNA transcript indicates that it plays important biological roles, which could be related to the adaptation to different niches, as it is differentially regulated between ARF and CER populations.

2.3. Stem-loop secondary structure of regulated putative lncRNA.

Recent studies have suggested that the secondary and even tertiary structure of lncRNAs are conserved and critical for the transcript to be functionally active [42, 44]. In vertebrates it was observed that the conservation of secondary structure of the lncRNA MEG3 regulates a tumor suppression function, regardless of its primary sequence [85]. The secondary structure of lncRNA is regarded as one of the multidimensional conservation pressure that long no-coding transcripts can
suffer [32], and as a result of their sequence length, most RNA transcripts are prone to form secondary structure [66]. However ncRNA present some distinguishable features, they show higher thermostability than coding transcripts, (i) their temperature melting (Tm) is significantly higher and (ii) with greater negative free energy values (Minimum Free Energy - MFE) [32, 42]. Another study observed that functional transcripts tend to present higher in silico second structure stability (with greater negative MFE), suggesting a link between secondary structure stability and functionality [53, 73].

There are lncRNA in which the secondary structure dictates their functionality, in humans the lncRNA MEG3 acts as a tumor suppressor based rather in the structure than primary sequence conservation [42], also the steroid receptor RNA activator (SRA), one of the few lncRNA that has its secondary structure experimentally defined, is reported to interact with many proteins and be related to breast cancer development. Although SRA primary sequence is mutated, the secondary structure, and the nucleotides involved in the stabilization of the structure are highly conserved, suggesting their direct involvement in the lncRNA functionality [47]. Experimental analysis comparing the folding energy of lncRNAs and mRNAs were capable to differentiate lncRNA based on their higher MFE [32].

Therefore, we performed folding analysis using ViennaRNA (v2.4.8) [39]. In which the stability of secondary structure of RNA can be inferred by MFE, regarding any values below -80 kcal/mol to be structurally stable [47]. We selected ten differentially regulated lncRNAs (table 1), from the set of conserved lncRNA in both Fabaceae genome and transcriptome analysis, the majority of them presented MFE below -100 kcal/mol. Once these structures were shown to be stable, it is expected that they may have a functional role in which a secondary structure formation is likely to be meaningful. At the supplementary figure 1, is possible to observe the expected second structure of the given lncRNAs analyzed.

### 3. Discussion

Forest trees are unique group to study adaptability traits, based on their life span and endurance to biotic and abiotic stress [79]. Thus, in the present study we identified 8,020 putative lncRNAs, some of which regulated above 2- and 5-fold in either copaiba population originated from different biomes. Also a thorough analysis of copaiba coding transcripts is being conducted by our group (Franco et al., in preparation), leading to a deeper understanding about their adaptability to these different environments at mRNA and epigenetics level, which is critical to copaiba management and conservation [1]. Similarly, Xu et al. [83] performed a study with Miscanthus lutarioparrius populations from two different environments, and observed an expression profile on which the lncRNA presented higher fold change expression than usually observed with mRNA, suggesting them to be more sensitive and responsive to environmental changes than coding transcripts.

Positional conservation in the genome analysis of lncRNA loci neighboring specific orthologous genes plays an interesting role, particularly when their primary and secondary structures are not completely conserved through species. For example, in humans, a small segment of AIRN lncRNA overlaps with the IGF2R promoter (insulin-like growth factor 2 receptor) and is sufficient to cause IGF2R silencing. This phenomenon was also identified in other lncRNAs [25, 44]. Through previous interspecies comparison [66], it was observed that the positional conservation may act by regulating the target gene solely by the transcription of the lncRNA.

In the present copaiba analyses, a total of 22% of the lncRNAs aligned to at least one of the genomes analyzed. The observed amount lncRNA conservation is coherent with the literature, since the conservation of lncRNA among the same family is much smaller than that of protein-coding genes [29, 65]. A conservation analysis between human and other placental mammals estimates only 44% of the lncRNA is conserved, while also their promoter seem to be under conservative pressure [17]. In plants, between Z. mays and S. bicolor there are approximately 25% lncRNA conserved, and Z. mays compared to A. thaliana presented only 2% conservation [37]. Thus C. langsdrorffii’s 22% conservation along the Fabaceae family lies in the expected conservation threshold.
Within these conserved putative lncRNAs, some were regulated in either condition and is possible to observe that the majority of them are upregulated in ARF plants, with a high fold change. The high conservation and regulation among these few transcripts suggests that they are under evolutionary constraints, possibly involved in the plant regulatory machinery [44].

Often the lncRNA primary sequence can be degenerated while the position is maintained and the act of transcription itself is enough to function as epigenetic regulators to closely related genes [29, 42, 44, 65, 66]. A study performed comparing putative lncRNA among the Brassicaceae, Aethionemae and Cleomaceae families uncovered several transcripts that were thought to be lineage-specific, instead they were in fact positionally conserved, although sequence divergent [44].

While comparing the copaiba lncRNA to several Fabaceae genomes has many advantages, such as to understand their position, classify the lncRNA as sense, natural antisense, bidirectional, intronic and intergenic, and possibly infer its function according to gene proximity [44]. Mapping the lncRNA transcripts to other species cDNA and EST sequences is a way to confirm that it is actively transcribed. Some previous studies that performed this analysis using vertebrates’ lncRNA noticed that some transcripts may hold sequence similarities to genomic untranscribed regions [25]. In our study, a higher percentage of lncRNA was paired with EST and cDNA libraries than to the genomic alignment analysis; it might be due to lncRNA post transcriptional processing which prevents the lncRNA to be correctly mapped to the genome. The transcripts alignment analysis suggests more orthologous lncRNAs are being actively transcribed throughout the Fabaceae family than was estimated by genomic analysis.

Regarding that lncRNAs frequently present low expression profile and tend to be expressed in a tissue or condition specific manner [29, 47, 65], the transcripts that didn’t aligned to any of the expressed Fabaceae libraries analyzed may still be conserved, although not expressed in the particular condition in which the library samples were taken from. It can be illustrated by the fact that, when comparing copaiba lncRNAs to two databases of putative plants lncRNAs, there was a single transcript that aligned to a known G. max lncRNA, and mapped to three Fabaceae genomes but didn’t aligned to G. max EST and cDNA libraries. Hence, it is essential to understand if the comparison analysis is insufficient to state whether the transcript is expressed or not in a given species.

Moreover, we observed that 14.2% of copaiba lncRNA overlapped the genomic and transcription comparative analysis, while the majority presented similar expression levels in both ARF and CER samples, there were 254 transcripts regulated in either sample above 2-fold. In figure 8, we can detect a tendency to form two distinct expression profiles among the samples, in which 17 lncRNA were up regulated above 10-fold. Considering the samples are from the same type of tissue, and similar developmental stage, it is reasonable to infer that if the lncRNA identified is functional, the RPKM differential expression between the samples is associated to the plants response to environmental stimuli. It should be mentioned also that, although there are no replicates in the experiment, RNA samples were prepared from a pool of 10 plants from each local.

The in silico folding prediction of 12 most regulated lncRNAs which were conserved at genome and transcriptome level, most of the transcripts were regarded to have a stable secondary structure. Thus, a further investigation of these transcripts regarding their functionality is needed, since their structural stability and post-transcriptional processing indicates they are likely to regulate their target expression as molecule, not solely by being transcribed. Thus, a model of their putative second structure and possible targets identification are the natural research directions to further understand the role of lncRNAs in the adaptive response of copaiba and possibly other closely related woody plants.

5. Materials and Methods

5.1. Plant Material Collection

The Copaifera langsdorffii leaf samples were collected from two different biomes, ten individual samples were taken from Atlantic Rainforest, at Área de Proteção Ambiental da Bacia do Rio São
João – Mico Leão, Silva Jardim, RJ, Brazil. And another ten individual samples were collected from Cerrado ecosystem at Estação Ecológica do Jardim Botânico de Brasília – EEJBB, Distrito Federal, Brazil, in August, during the conspicuous annual dry season. All biological material harvested was placed in RNA later-like buffer, kept at -80°C until extraction.

5.2. RNA extraction to Sequencing

RNA extraction was performed individually from samples collected from Atlantic Rainforest and Cerrado. After, they were pooled together for the library construction, and named JCF-44 and JCF-45, respectively. It was then sent to Fasteris Life Sciences SA (Plan-les-Ouates, Switzerland) for the multiplex sequencing performed on Illumina HiSeq 2000I using the single-end 76 cycle protocol.

5.3. De novo transcriptome assembly and gene expression profiling

Sequencing reads were checked using FastQC and searched against NCBI nr database using BLASTN (megablast, e-value 1e-5; alignment length ≥90 and identity ≥80%)[3]. Reads aligning to non-plant organisms were removed. Raw reads were pre-processed using Trimmomatic (v0.36) [11] for trimming adapters, trailing 15 bases and tailing 5 bases for each read and other quality cleaning. De novo transcriptome assembly was performed using Trinity (v2.3.2; default settings)[72]. Fasta headers were renamed according to the sample ids (e.g. JCF44_xxx and JCF45_xxx). One-to-one correspondence between the transcriptomes was detected by bi-directional BLASTN (v.2.2.31+)[3]. Stringent criteria were used for this analysis and only top hits with e-value 1e-10, query coverage of ≥50%, identity ≥90%, bit-score ≥50 were considered. Filtered reads were aligned to one-to-one transcripts using Bowtie2 (v. 2.2.9)[64] and RPKM values estimated using Cufflinks (v2.2.1)[64].

5.4. IncRNA identification

The one-to-one transcripts were screened for IncRNA on several parameters. Transcripts longer than 200bp with maximum ORF size of less than 100 amino acids by Getorf (EMBOSS:6.6.0.0) were used to predict putative IncRNA. These transcripts were analyzed using two software CPC (v.0.9-r2) [35] and PLEK (v1.2) [37]. Only transcripts classified as non-coding in both CPC and PLEK analysis were considered as putative IncRNA. Further we kept only those putative IncRNAs with RPKM ≥1.

5.5. Interspecies IncRNA conservation analysis

To identify other IncRNA which holds sequence conservation among Fabaceae species, we used BlastN 2.7.1+ (evalue 10^-15, identity 90% - qcov 50%) [12] against the downloaded GREENC (http://greenc.science designers.com/) and CANTATA database (http://cantata.amu.edu.pl/), both databases present putative IncRNA obtained through their own bioinformatics pipeline, which is similar to the one we used to filter our own. Using DrawVenn application (http://bioinformatics.psb.ugent.be/webtools/Venn/) we identified the ones mapped to more than one genome. The EST and cDNA from Fabaceae species was obtained from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/) and Phytozome databases (https://phytozome.jgi.doe.gov/), we identified similar transcripts using BlastN 2.7.1+ (evalue 10^-15, identity 50% and 90%, query cov 50%) [12], followed by DrawVenn analysis (http://bioinformatics.psb.ugent.be/webtools/Venn/).

5.6. Second structure modeling

The second structure modeling of IncRNA was made using Vienna RNAfold (v2.4.8) (Lorenz, 2011) [39] package, at 25º C, default parameters.

Supplementary Materials: Table 1: Presents the novel putative copaiba IncRNA identified to be regulated above 2-fold in the comparison of CER and ARF samples. In this table are the identification of the transcripts in each sample, separated by columns and its respective RPKM expression. Table 2: Presents the novel copaiba IncRNA which were expressed in both genome and transcriptome of other Fabaceae species at the conservation analysis. The table presents their identification in each sample, and the respective RPKM
expression. Table 3: Presents novel copaiba lncRNA that were expressed in both genome and transcriptome of other Fabaceae species at the conservation analysis, and were regulated above 2-fold between CER and ARF samples, indicating they are potentially involved in the plant adaptive response to the different biomes. This table presents the identification in each sample, and the respective RPKM expression. Figure 1: Presents the secondary structures predicted by Vienna RNAfold for the 12 most highly regulated copaiba lncRNA conserved in both genome and transcriptome analysis. The image displays the MFE secondary structure with the base-pair probability in color code (blue indicates zero and red indicates maximum probability).

Author Contributions: Conceptualization, PF, MC, AH, SC, F.T. and LF.; Methodology, LF, SC, MD, TV and KM; Software, TV, KM, FP; Validation, MD, LF, and SC Formal Analysis, MD.; Investigation, MD.; Resources, SC, LF; Data Curation, MD.; Writing-Original Draft Preparation, MD.; Writing-Review & Editing, MD., FP, KM, TV; Visualization, MD.; Supervision, PF.; Project Administration, PF.; Funding Acquisition, PF.

Funding: This research was funded by FAPERJ grant number E12/200.134/2014

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References


43. Millás ALGM., Siveira JVW., Barbosa R., Puzzi M., Koh IHJ., Bittencort E., Implantes de sáculos de plga incorporados com óleo-resina do gênero copaíba. 14° Congresso da Sociedade Latino Americana de Biomateriais, Órgãos Artificiais e Engenharia de Tecidos 2017 - SLABO.


