Identification and characterization of orphan genes in rice (Oryza sativa japonica) to understand novel traits driving evolutionary adaptation and crop improvement.

Rakesh Shah
rakeshks@iastate.edu

Follow this and additional works at: https://lib.dr.iastate.edu/creativecomponents

Part of the Computer Sciences Commons

Recommended Citation
Shah, Rakesh, "Identification and characterization of orphan genes in rice (Oryza sativa japonica) to understand novel traits driving evolutionary adaptation and crop improvement." (2018). Creative Components. 103.
https://lib.dr.iastate.edu/creativecomponents/103

This Creative Component is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Creative Components by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Identification and characterization of orphan genes in rice (*Oryza sativa japonica*) to understand novel traits driving evolutionary adaptation and crop improvement.

Abstract

Orphan genes are the genes that lack detectable homologs in other species. Orphan genes also known as TRS (Taxonomically Restricted Genes) are an elliptical component of the genome because their ancestry and functions are mostly unknown. Understanding the functional significance and the origin of such orphan genes may give us insights into novel traits that can be evolutionary and physiologically important. Using bioinformatics computation and statistical analysis of *Oryza sativa japonica*, the total of 1,507 orphan genes was found. Around 5,173,767 unannotated open reading frames (ORFs) in *Oryza sativa japonica* genome which could be potential coding for a novel protein specific to *Oryza sativa japonica* is obtained. The publicly available RNA-seq data from SRAdb were then used in order to explore the expression patterns of these orphan genes under various conditions. Most of the research disregards orphan genes as its importance and functions are not well known. Orphan genes are related to stress responses and specific traits which might be the key to improve crop yields and disease resistance in domestic plants.
Introduction

Orphan genes are the genes that lack sequence similarity identifiable to genes in other species. In other words, orphan genes are the genes with protein-coding regions which have arisen from scratch with no evidence of homology in other species (Arendsee et al., 2014). It was brought in light after the completion of the yeast genome since these sequences constituted about one-third size of the total yeast genome (Dujon, 1996). These genes are often termed as ORFans in the bacterial world and have been found to be present in archaea and phages as well. They have been reported to be in all genomes sequenced till date, typically making up 10% to 30% of all genes in a genome (Wissler et al., 2013). They are the universal feature of all genomes as shown by comparative genomics. These genes were originally believed to be more due to fewer sequencing coverage of genomes and were predicted to be in fewer numbers with the increase in genome sequencing (Cassari et al., 1996). But this has not been the case after extensive expansion on the sequencing of various genomes (Khalturin et al., 2009). This brings the question of its origin making its evolutionary nature still unknown and suggesting its species-specific roles rather than these orphan genes being a product of genome annotation anomaly.

In bacteria, as the number of complete genome sequence analyzed has been going up, the cumulative number of identified orphans doesn’t seem to be reducing (Wilson, 2005; Wilson et al., 2007). Even though the genome of the bacteria is increasing, the orphan genes that have been identified previously is still unchanged, hence it has no effect on the overall constituent of the orphan gene. This level of coverage of orphan genes suggests its taxon level significance. Hence, the concept of restricting these genes to a taxonomic group has redefined these orphans as species level-taxonomically restricted genes (TRGS) (Wilson, 2005). These genes due to its unexplained evolutionary and taxonomic significance have brought up many questions of whether they code for any functional protein if those proteins are unique or they show some structural similarity (Gauger, 2013). The emergence of such genes is found to be rather high but it is still not known if they are stabilized or further diverged along the evolution. These orphans or TRGs might be involved in lineage-specific adaptations as
a mechanism that favors the given lineage in the environment that is continuously evolving (Khalturin et al., 2009).

Few of the mechanisms to explain the origin of these orphan genes includes gene rearrangement or duplication. Any paralogues will still have a traceable founder gene taking those genes out of the orphan category. One mechanism might be divergence after duplication of some functional genes that result in loss of the similarity below the detection level (Wissler et al., 2013; Schlötterer, 2015). It is the fact that de novo evolution results in the gene that shares no similarity with other species (Tautz et al., 2011). These may arise from the non-coding region and gain some functions during the evolution. This might be the most prominent cause to the origin of orphan genes than previously assumed. Besides, transposons are also a likely way to rearrange the genes that eventually develops into an orphan (Vloff, 2006). These orphan genes might have a very fast evolving rate in a short lifespan making it impossible to trace the origin. In *Drosophila*, around 30% genes were found to be fast evolving while all those were not the orphan genes as there are some known regulatory genes with the traceable origin that are evolving fast (Schmid et al., 1997). But few studies have reported that the non-conserved genes are shorter and hence suggesting they might have weak selection constraint resulting in fast evolution (Lipman et al., 2002). These genes shorter in length might be the easier target for deletion or other mutation resulting in the loss of similarity with any other genes in the species.

Many orphan genes are not studied and hence their functionality still remains unknown, although few other studies have established the crucial functionality of orphan genes and in some case its essentiality (Prabh et al., 2016; Arendsee et al., 2014; Chenet al., 2010). Sometimes, these genes are governed by the environmental stresses that include both biotic and abiotic stress. As reported in Hydra, these novel genes may play a role in the species-specific morphological trait. An analysis on ants and Hymenoptera shows the role of new genes in the social evolution (Wissler et al., 2013). In humans, they are shown to be responsible for neurological defects when disrupted (Li et al., 2010) while these novel genes were shown to be the essential genes in
Drosophila (Chen et al., 2010). The first orphan gene of A. thaliana with defined function is QQS that has a metabolic role in which it is involved in Carbon and nitrogen allocation. It has been shown to increase protein content and lower the starch level when overexpressed (Arendsee et al., 2014; Li et al., 2009).

There has been an obvious bias in representing orphan genes in the study since most of the studies focus on the genes that are identified as significant in one species, and is conserved among many others that shares the sequence and function similarity and hence represents a good candidate gene to be studied (Prabh et al., 2016; Aubourg et al., 2000). This transfer of annotations between species has decreased our focus on other non-conserved genes (Calabrese et al., 2009). To focus more on those genes found to be unique for a taxon group, our study is trying to identify new likely orphan genes in Oryza sativa japonica and it’s functionality in terms of its quantitative expression using published transcriptomic data. As a general agreement or rule, in BLASTp proteins which do not show any sequence similarity with cut-off values E < 10^{5} or E < 10^{10} are taken as possible orphans that further goes down for filtering with a more stringent condition.(Khalturin et al., 2009).

RNA-Seq (RNA Sequence) uses deep-sequencing technologies to profile, quantify and discover transcriptome (Wang et al., 2009). The transcriptome is the complete set of all RNA molecules in a cell and their quantity according to the environmental or physiological conditions.

SRA (Sequence Read Archive) is the largest bioinformatics database which makes sequencing data available to researchers to allow for new research. The datasets are generally raw sequencing data and alignment information and generated by High-throughput sequencing platforms. The platform includes Roche 454 GS System, Illumina Genome Analyzer, Applied Biosystems SOLiD System, Helicos Heliscope, and others. It is difficult to extract the data of interest using available tools. SRAdb makes easy to access the data which is achieved by parsing all the NCBI SRA metadata (associated with submission, study, sample, experiment and run) into a
SQLite database. The sequences were downloaded and queried like a simple SQL query to get all the RNA-seq data for the *Oryza sativa japonica*.

**ORF (Open Reading Frame)** is a part of the reading frame that includes the start codons and stop codons. Codons are the sequences of contiguous nucleotide triplets equating to amino acids or start and stop signals during translation. A reading frame is a sequence of codons that contains a start codons like AUG, a stop codon like UAA, UAG or UGA and others codons specifying amino acids. One strand of DNA sequence has three possible reading frames. The common use of ORF is as a piece of evidence to help in the prediction of genes. Long ORFs are used to identify the candidate protein-coding regions or functional RNA-coding regions is a DNA sequence.

**Salmon** (Patro, Rob, et al. 2017) is a software used to quantify RNA-seq data. Salmon uses RNA-seq reads in FASTA format to map to a reference transcriptome in our case *Oryza sativa japonica*. This tool uses a dual-phase inference algorithm in which the first phase is to indexing the initial expression level and model parameters, which is independent of reads. The second phase comprises of quantifying those estimates that give the transcript abundance in TPM (count or transcript per million) value (Patro, Rob, et al, 2017).

Identifying and understanding these orphan genes can be helpful in eventually deciphering its genesis and evolution, which could be explored for further study of genetic variation in crops. Owing to the recently discovered essential functions of orphan genes, their study will in long run help study the environmental stress, its effect, adaptability, evolution of novel genes and proteins along with optimum productivity of crops. This study is focused on identifying and characterizing orphan genes in one of the major crops, *Oryza sativa Japonica* using BLAST and look at the expression and protein profile of each of these identified orphans. This will eventually will be taken forward to understand and predict the possible function of these orphan genes. This study also focus on predicting and identifying the unannotated ORF which could be potential coding for the novel proteins.
Data and Methods

Data Collection:

The recently published 13 wild and domesticated rice genome which contains protein and cDNA sequences on different species of wild and domesticated rice were downloaded from Data Commons (CyVerse) (Stein J. C. et. al, 2018). The whole genome sequence of *Oryza Sativa Japonica* was downloaded from NCBI RefSeq (GCA_000005425.2 Build 4.0). The *Oryza sativa japonica* whole genome sequences were used to extract unannotated ORFs in order to compile a list of unannotated expressed ORFs which have no hits in the ‘non-redundant’ (nr) database.

<table>
<thead>
<tr>
<th>Species (Genome type)</th>
<th>Assembly size (Mb)</th>
<th>TAX ID</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Oryza_sativa_vg_japonica</em></td>
<td>374</td>
<td>39947</td>
</tr>
<tr>
<td><em>Oryza_sativa_vg_indica</em></td>
<td>375</td>
<td>39946</td>
</tr>
<tr>
<td><em>Oryza_sativa_N22</em></td>
<td>372</td>
<td>39946</td>
</tr>
<tr>
<td><em>Oryza_sativa_IR8</em></td>
<td>389</td>
<td>39946</td>
</tr>
<tr>
<td><em>Oryza_rufipogon</em></td>
<td>338</td>
<td>4529</td>
</tr>
<tr>
<td><em>Oryza_punctata</em></td>
<td>394</td>
<td>4537</td>
</tr>
<tr>
<td><em>Oryza_nivara</em></td>
<td>338</td>
<td>4536</td>
</tr>
<tr>
<td><em>Oryza_meridionalis</em></td>
<td>336</td>
<td>40149</td>
</tr>
<tr>
<td><em>Oryza_glumaepatula</em></td>
<td>373</td>
<td>40148</td>
</tr>
<tr>
<td><em>Oryza_glaberrima</em></td>
<td>285</td>
<td>4538</td>
</tr>
<tr>
<td><em>Oryza_brachyantha</em></td>
<td>261</td>
<td>4533</td>
</tr>
<tr>
<td><em>Oryza_barthii</em></td>
<td>308</td>
<td>65489</td>
</tr>
<tr>
<td><em>Leersia_perrieri</em></td>
<td>267</td>
<td>77586</td>
</tr>
</tbody>
</table>

**Table 1.** Assembly and annotation statistics of 13 Oryzeae reference genomes. The taxonomy database is maintained by NCBI/GenBank. These are the data downloaded from (Stein J. C. et. al, 2018).
Methods:

Figure 1. Workflow diagram to find the orphans genes in *Oryza sativa japonica*. 13 protein sequences from the paper (Stein J. C. et. al, 2018). BLASTp was run against *Oryza sativa japonica* with other rice species from the paper. Total of 925 orphans genes in OSJ against 13 other rice species from the paper was found with threshold e-value $10^{-2}$ cutoff. We also ran phylostratigraphy to classify genes of *Oryza sativa japonica*. We found the total of 1508 orphan genes in OSJ against all other species and all the protein sequences of OSJ were stratified according to their age.
1. Identify potential orphans genes in *Oryza sativa japonica* using BLAST:

BLAST stands for “Basic Local Alignment Search Tool”. BLAST is a tool for comparing coding sequences and finding the region of similarity between them. This tool compares peptide (known as BLASTp) or nucleotide (known as BLASTx) sequences to sequence databases and calculates the statistical significance (NCBI). The typical way to identify the phylloclade of genes in a genome is to compare the sequences of the genome of interest with identified genes from other species and try to locate local similarity between sequences in eukaryotes. BLAST has been shown to be effective to classify orphan genes in *Drosophila*, *Arabidopsis*, and viral genome (Arendsee, et al., 2014, Zhang and Long, 2010, Domazet-Loso et al., 2003) and the filtering criteria can be manipulated as the study demands. It enables a researcher to identify sequences that resemble the query sequence above a certain threshold also known as expectation (E) - value. We used $1.0 \times 10^{-2}$ as the expectation value for defining the number of orphan genes.

2. Age stratification of genes using phylostratigraphy:

Phylostratigraphy is a method that traces the evolution of a gene from its most recent ancestor based on the identification of homologs obtained by BLAST search (Moyers et al., 2015; Domazet-Loso et al., 2007). It is a statistical technique that stratifies the genes to trace modern genes back to their founder. The phylostratigrapic analysis uses protein-BLAST as a basic simple search algorithm. For example, Figure 1. Shows a phylostratigraph of the protein-coding genes of *Arabidopsis thaliana*. This methodology assumes that each query gene or protein is the results of the evolution of a founder that has been scattered throughout time. Therefore, these recent ancestors or the founder represent the evolutionary novelties (Domazet-Lošo et al. 2007; Domazet-Lošo and Tautz 2010; Domazet-Lošo and Tautz 2008). This enables us to determine the minimum age of that gene. The approach is to select the
hierarchical taxonomic groups (*Oryza sativa japonica* in our case) that ascends from the focal species and for each candidate gene trace back the oldest taxon in which it has a homolog. (Arendsee et al., 2014, Domazet-Lošo et al. 2007).

**Figure 2.** Example of phylostratigraphic analyses of the *Arabidopsis thaliana* genomes. Genes are stratified according to age and each gene is assigned to the oldest clade as inferred by a BLASTp for each *A. thaliana* genes against a selected set of genomes with an E-value of $1 \times 10^{-5}$. The age is estimated time since the diversification of the clade from its most recent ancestor and the age is in Ma (Millions of years ago). This figure is from (Arendsee et al., 2014).
3. Workflow:

Using the GETORF program (EMBOSS: Getorf) the open reading frames (ORF) was extracted from *Oryza sativa japonica* genome. These are used to identify the candidate protein-coding regions or functional RNA (coding regions in a DNA sequence) i.e potential coding for novel proteins. The downloaded ORFs were mapped and quantified against the RNA-Seq reads from 14 diverse datasets using Salmon (Patro, et al. 2017). The 14 datasets that contain RNA-Seq runs were selected from SRAdb that represented the more diverse and deeply sequenced set of sequencing data for the *Oryza sativa japonica*. The quant files which contained fields like Name, Length, Effective Length, TPM (Transcript per million) and Number of Reads after mapping and quantification using the Salmon tool was obtained. The ORFs were furthered filtered based on TPM value from the quant output files. The target transcript that had value more than 15 TPM in any RNA-seq runs were selected to run BLASTx against the database to find any homologs among different species. The ORFs that lacked any other homologs (i.e. orphans) were further filtered. The potential orphan genes and cDNA data from the recently published paper (Stein J. C. et. al, 2018) were mapped against the whole RNA-Seq data taken from NCBI SRAdb using Salmon to study the expression pattern of all ORFs and transcripts. These were further analyzed using our in-house tool called MetaOmGraph (MOG). MOG is an application to display and analyze large expression datasets.
Figure 3: Flowchart showing the workflow diagram for finding the expression pattern of *Oryza sativa japonica*. Using getorf program, the ORFs were downloaded and then mapped to 14 deeply sequenced RNA-Seq runs. The ORFs which had transcript per million of 15 or more in any runs was further taken. The ORFs with homologous proteins was removed by using BLASTx against the SwissProt database. We then combined the cDNA sequences from (Stein J. C. et. al, 2018) and novel unannotated expressed ORF sequences to map to all RNA-seq runs to get the expression.
Results and Discussion

The rice annotated proteome i.e. *Oryza sativa japonica* which has 50,567 protein sequences, was aligned to the NCBI nr database using the NCBI BLAST tool. For defining the number of orphan genes, we used $1 \times 10^{-2}$ as the threshold E-value cutoff. We compared the number of orphan genes to that threshold E-value and found 925 orphan genes in *Oryza sativa japonica*. Figure 5 is a bar chart which depicts the number of orphan genes across all thirteen rice species at a BLASTp E-Value cutoff of $1 \times 10^{-2}$. The x-axis specifies the number of orphan genes and y-axis specifies the respective *Oryza* and related species.
Figure 4: Pair-wise comparison of the number of orphan genes in *Oryza sativa japonica* against all 13 species. BLASTp was run against the *Oryza sativa japonica* with other 13 rice species. The number 4,629 indicates the number of orphans in *Oryza sativa japonica* against *Oryza sativa vg Indica*. The x-axis indicates the number of orphan genes in *Oryza sativa japonica* with an E-value cutoff of $10^{-2}$ and the y-axis indicates the respective *Oryza* and related species. At threshold E-value of $10^{-2}$, we obtained the total of 925 orphan genes out of 50,567 protein sequences in *Oryza sativa japonica*. 
To stratify the *Oryza sativa japonica* genes according to the age, we used phylotestratr (an in-house R framework for phylostratigraphy developed by Zeb in our lab) to find around 1,500 genes that lacked any founder genes when traced for its ancestry and hence were classified under orphan genes category. The table below signifies each gene that is assigned to the oldest clade that contains a homolog. Here the clades currently unnamed are given designators of a (the eldest of the unnamed clade) through h (the clade of the most recent origin).

\[
\begin{array}{l}
O. \text{brachyantha} \\
O. \text{punctata} \\
O. \text{meridionalis} \\
O. \text{glumaepatula} \\
O. \text{glaberrima} \\
O. \text{barthii} \\
O. \text{nivara} \\
O. \text{sativa vg. indica} \\
O. \text{rufipogon} \\
O. \text{sativa vg. japonica}
\end{array}
\]
b. Most frequently estimated species phylogeny in rice. The clades currently unnamed are given designators of a (the eldest of the unnamed clade) through h (the clade of the most recent origin). 4b. Phylostratigraphy of all genes in *Oryza sativa japonica*. Genes are stratified according to age and each gene is assigned to the oldest clade as inferred by a BLASTp for each *Oryza sativa japonica* genes against a selected set of genomes. The age is estimated time since the diversification of the clade from its most recent ancestor and the age is in Millions of years ago.

<table>
<thead>
<tr>
<th>Clades</th>
<th>Clade-specific (Numbers)</th>
<th>Genes (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cellular organisms</td>
<td>16190</td>
<td>32.0%</td>
</tr>
<tr>
<td>Eukaryota</td>
<td>8470</td>
<td>16.8%</td>
</tr>
<tr>
<td>Viridiplantae (Green plants)</td>
<td>2366</td>
<td>4.7%</td>
</tr>
<tr>
<td>Streptophyta (Plants)</td>
<td>2118</td>
<td>4.2%</td>
</tr>
<tr>
<td>Embryophyta (Land)</td>
<td>3042</td>
<td>6.0%</td>
</tr>
<tr>
<td>Tracheophyta (Vascular)</td>
<td>337</td>
<td>0.7%</td>
</tr>
<tr>
<td>Magnoliophyta (flowering)</td>
<td>1861</td>
<td>3.7%</td>
</tr>
<tr>
<td>Mesangiospermae</td>
<td>1420</td>
<td>2.8%</td>
</tr>
<tr>
<td>Liliopsida</td>
<td>64</td>
<td>0.1%</td>
</tr>
<tr>
<td>commelinids</td>
<td>408</td>
<td>0.8%</td>
</tr>
<tr>
<td>Poales</td>
<td>132</td>
<td>0.3%</td>
</tr>
<tr>
<td>Poaceae</td>
<td>2593</td>
<td>5.1%</td>
</tr>
<tr>
<td>BOP clade</td>
<td>371</td>
<td>0.7%</td>
</tr>
<tr>
<td>a</td>
<td>845</td>
<td>1.7%</td>
</tr>
<tr>
<td>b</td>
<td>355</td>
<td>0.7%</td>
</tr>
<tr>
<td>c</td>
<td>1602</td>
<td>3.2%</td>
</tr>
<tr>
<td>d</td>
<td>2556</td>
<td>5.1%</td>
</tr>
<tr>
<td>e</td>
<td>2216</td>
<td>4.4%</td>
</tr>
<tr>
<td>f</td>
<td>1048</td>
<td>2.1%</td>
</tr>
<tr>
<td>g</td>
<td>747</td>
<td>1.5%</td>
</tr>
<tr>
<td>h</td>
<td>305</td>
<td>0.6%</td>
</tr>
</tbody>
</table>

*Oryza sativa Japonica*  1508  3.0%
Finding ORFs using whole genome data

The total of 5,173,767 open reading frames in the *Oryza sativa japonica* genome was found using the getorf program (EMBOSS: Getorf). 14 diverse and deeply sequenced RNA-Seq runs were used from SRA to map the ORFs using salmon. As criteria for selecting expressed ORFs, we kept all 440,206 ORFs which had TPM of 15 or more in any of the 14 RNA-seq runs. We then used BLASTx against swissprot database to remove any ORFs with homologous proteins. We were left with 9,794 unannotated ORFs which could be potential coding for the novel protein.

Mapping 606 RNA-Seq runs to *Oryza sativa japonica* transcriptome

The cDNA sequences (Stein J. C. et. al, 2018) and the novel unannotated expressed ORF sequence was used to make the final *Oryza sativa japonica* transcriptome. Total of 60,361 *Oryza sativa japonica* transcriptome was obtained. It was used to map 606 RNA-seq runs using salmon and find the expression patterns of all the transcripts in all the different runs to compile a comprehensive dataset of expression patterns of *Oryza sativa japonica* transcripts.
Analysis of genes using MetaOmGraph

Expression pattern of 1508 orphan genes:

**Figure 6.** Expression pattern of *Oryza sativa japonica* orphan genes across 606 RNA-seq runs. The data is grouped by gsm_source_name metadata. Many orphan genes are expressed under different conditions. Correlation values can be computed using MetaOmGraph to find co-expressed genes which can give further insight about the genes.
Expression pattern of *Oryza sativa japonica* ORFs:

Figure 7. Expression pattern of *Oryza sativa japonica* ORFs across 600 RNA-seq runs. The ORFs are expressed under different conditions. The data is grouped by gsm_source_name metadata. Correlation values can be computed using MetaOmGraph to find co-expressed genes which can give further insight about the genes.
Expression pattern of an orphan gene (*Osjap03g13490.1*). This orphan gene shows high expression pattern in SRR1288363 RNA-Seq run. This SRA studies suggests that they have performed mRNA-Seq on the shoots of root rot nematode (*Hirschmanniella oryzae*) infected rice plants.

**Figure 8.** *Osjap03g13490.1* was identified as orphan genes in *Oryza sativa japonica* using our pipeline. This orphan gene shows high expression patterns in RNA-seq from studies related to shoots of ho-infected plants during post infection. Orphan genes may be expressed highly in only certain condition this making their identification a difficult task. Condition-specific expression also includes the gene may have a biological role. The analysis was done using MetaOmGraph.
Conclusion

The functional significance and the origin of the orphan genes provides an opportunity for evolutionary innovation. The careful analysis of the orphan genes can reveal the novel traits which can be evolutionary and physiologically important and can be involved in the biological functions. Also, there is the presence of the biological markers in the genes that are totally related to the crops yields and resistance of the disease.

We analyzed OSJ data and found 1,508 orphan genes and 9,794 unannotated ORFs expressed in 600 different runs. High expression of these orphans genes and unannotated ORFs suggest that they have the functional role and they are sources of novel proteins. The more rigorous analysis is required to find their biological functions and their evolutionary origins in *Oryza sativa japonica*. 
References


