2017

Raising orphans from a metadata morass: a researcher’s guide to re-use of public ’omics data

Priyanka Bhandary  
*Iowa State University*, bhandary@iastate.edu

Arun S. Seetharam  
*Iowa State University*, arnstrm@iastate.edu

Zebulun Arendsee  
*Iowa State University*, arendsee@iastate.edu

Manhoid Hur  
*Iowa State University*, mhhur@iastate.edu

Eve Syrkin Wurtele  
*Iowa State University*, mash@iastate.edu

Follow this and additional works at: [http://lib.dr.iastate.edu/gdcb_las_pubs](http://lib.dr.iastate.edu/gdcb_las_pubs)

Part of the [Bioinformatics Commons](http://lib.dr.iastate.edu/gdcb_las_pubs), [Cell and Developmental Biology Commons](http://lib.dr.iastate.edu/gdcb_las_pubs), [Computational Biology Commons](http://lib.dr.iastate.edu/gdcb_las_pubs), and the [Molecular Genetics Commons](http://lib.dr.iastate.edu/gdcb_las_pubs)

The complete bibliographic information for this item can be found at [http://lib.dr.iastate.edu/gdcb_las_pubs/164](http://lib.dr.iastate.edu/gdcb_las_pubs/164). For information on how to cite this item, please visit [http://lib.dr.iastate.edu/howtocite.html](http://lib.dr.iastate.edu/howtocite.html).
Raising orphans from a metadata morass: a researcher's guide to re-use of public 'omics data

Abstract
More than 15 petabases of raw RNAseq data is now accessible through public repositories. Acquisition of other 'omics data types is expanding, though most lack a centralized archival repository. Data-reuse provides tremendous opportunity to extract new knowledge from existing experiments, and offers a unique opportunity for robust, multi-'omics analyses by merging metadata (information about experimental design, biological samples, protocols) and data from multiple experiments. We illustrate how predictive research can be accelerated by meta-analysis with a study of orphan (species-specific) genes. Computational predictions are critical to infer orphan function because their coding sequences provide very few clues. The metadata in public databases is often confusing; a test case with Zea mays mRNA seq data reveals a high proportion of missing, misleading or incomplete metadata. This metadata morass significantly diminishes the insight that can be extracted from these data. We provide tips for data submitters and users, including specific recommendations to improve metadata quality by more use of controlled vocabulary and by metadata reviews. Finally, we advocate for a unified, straightforward metadata submission and retrieval system.

Keywords
Orphan genes, Metadata, Meta-analysis, Transcriptomics, Metabolomics, 'Omics

Disciplines
Bioinformatics | Cell and Developmental Biology | Computational Biology | Genetics and Genomics | Molecular Genetics

Comments
This is a manuscript of an article published as Priyanka Bhandary, Arun S. Seetharam, Zebulun Arendsee, Manhoi Hur, Eve Syrkin Wurtele, Raising orphans from a metadata morass: a researcher’s guide to re-use of public 'omics data, Plant Science (2017), doi: 10.1016/j.plantsci.2017.10.014. Posted with permission.

Creative Commons License
This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License.
Accepted Manuscript

Title: Raising orphans from a metadata morass: a researcher’s guide to re-use of public ‘omics data

Author: Priyanka Bhandary Arun S. Seetharam Zebulun Arendsee Manhoi Hur Eve Syrkin Wurtele

PII: S0168-9452(17)30574-5
DOI: https://doi.org/doi:10.1016/j.plantsci.2017.10.014
Reference: PSL 9693

To appear in: Plant Science

Received date: 17-7-2017
Revised date: 7-10-2017
Accepted date: 15-10-2017

Please cite this article as: Priyanka Bhandary, Arun S. Seetharam, Zebulun Arendsee, Manhoi Hur, Eve Syrkin Wurtele, Raising orphans from a metadata morass: a researcher’s guide to re-use of public ‘omics data, <![CDATA[Plant Science]]> (2017), https://doi.org/10.1016/j.plantsci.2017.10.014

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
Raising orphans from a metadata morass: a researcher’s guide to re-use of public ’omics data

Priyanka Bhandary\textsuperscript{a,b}, Arun S. Seetharam\textsuperscript{c}, Zebulun Arendsee\textsuperscript{a,b}, Manhoi Hur\textsuperscript{a,b}, Eve Syrkin Wurtele\textsuperscript{a,b,}\ast

\textsuperscript{a}Dept. of Genetics Development and Cell Biology, Iowa State University, Ames IA, 50010, USA
\textsuperscript{b}Center for Metabolic Biology, Iowa State University, Ames, IA 50011, USA
\textsuperscript{c}Genome Informatics Facility, Office of Biotechnology, Iowa State University, Ames, IA 50011, USA

Abstract

More than 15 petabases of raw RNAseq data is now accessible through public repositories. Acquisition of other ’omics data types is expanding, though most lack a centralized archival repository. Data-reuse provides tremendous opportunity to extract new knowledge from existing experiments, and offers a unique opportunity for robust, multi-’omics analyses by merging metadata (information about experimental design, biological samples, protocols) and data from multiple experiments. We illustrate how predictive research can be accelerated by meta-analysis with a study of orphan (species-specific) genes. Computational predictions are critical to infer orphan function because their coding sequences provide very few clues. The metadata in public databases is often confusing; a test case with \textit{Zea mays} mRNA seq data reveals a high proportion of missing, misleading or incomplete metadata. This metadata morass significantly diminishes the insight that can be extracted from these data. We provide tips for data submitters and users, including specific recommendations to improve metadata quality by more use of controlled vocabulary and by metadata reviews. Finally, we advocate for a unified, straightforward metadata submission and retrieval system.

\ast Corresponding Author.

Email address: mash@iastate.edu (Eve Syrkin Wurtele)
Keywords: Orphan genes, Metadata, Meta-analysis, Transcriptomics, Metabolomics, 'Oomics

1. Highlights

Tips, tools, and pitfalls in accessing and harvesting knowledge from public datasets.

Case study using meta-analysis: Exploring the behavior and function of orphan genes.

Why metadata meta-matters.

Submitting high quality metadata to public archives.

Recommendations to improve the quality and usability of public metadata.

2. Meta-analysis of data with its metadata

2.1. The meta-importance of metadata

Meta-analysis, the systematic combining of data from multiple studies, aids in developing hypotheses that have greater statistical power than that from a single experiment alone. As such, it can provide a powerful high-throughput first-pass inference of gene function [1-7].

For some purposes, meta-analysis depend only on data without it’s metadata. Computer scientists can represent the patterns in the data without emphasis on the biology. Statisticians can evaluate data heterogeneity, determine value distributions, or compare normalization parameters. Bioinformaticians can implement high-throughput studies that test general ideas, for example about patterns and degrees of interactions among genes.

However, for much meta-analysis, the metadata is crucial. From a biologist’s vantage point, accurate and comprehensive metadata is nearly always essential for biological insight. We focus here on transcriptomics, because these data are most bountiful. However, the concepts apply generally to proteomics, metabolomics, and other omics data, which are rarely archived yet valuable in molecular function inference [8-10].
2.2. Using compiled sets of RNA-seq data and its metadata

Meta-analysis of multiple experiments can provide key information about the function of a particular gene. For example, if statistical analysis of a single RNA-seq dataset shows that expression of a gene is up-regulated 18-fold when a plant is under heat-stress, it may not be reasonable to design tests of knockout mutants of that gene under similar heat-stress conditions. However, if this gene is upregulated in ten independent studies of heat stress that include various organs, the gene might be prioritized for a test as to it’s potential involvement in a heat stress response. If the dataset shows the gene is downregulated in various mutants of ABA signal transduction, ABA might be mediating the effect; also a testable hypothesis. If this same gene is down-regulated by aphid infestation but not by nematodes, this provides yet another piece in the puzzle that can be addressed experimentally. In the primary publications arising from each independent study, that gene, one of perhaps many hundreds of genes with differential expression, may easily have been ignored.

Likewise, meta-analysis can be used to develop testable hypotheses about biological systems. For example, it can provide clues about the physiological significance of a biochemical pathway, or the interactions among members of genes in a gene family, or it can provide a unique gene expression fingerprint of a particular disease condition versus other related diseases.

Sufficient transcriptomic data for meaningful meta-analysis are available for many organisms. Thousands (eg. maize, yeast, *Arabidopsis*) to hundreds of thousands (humans) of individual RNA-seq runs have been deposited in public databases. Most are in NCBI/Sequence Read Archives (SRA contained more than 10 petabases of data as of April, 2017), NCBI/GEO, EMBL/ArrayExpress/ENA and DDBJ or JGI. These data come from many thousands of studies. Meta-analyses may be based on a pre-existing compiled data or new dataset.

Public transcriptomics datasets (RNAseq or microarray) composed of data from multiple experiments have been compiled for several model species and have led to valuable insights ([1], [11–22]). Some community crop databases have similar compilations ([25, 26]. Most compiled datasets are
limited in size (generally ranging from 50 to 1000 chips or runs and 4 to 50 experiments). For many species, including some major crops, compiled datasets are limited or non-existent. There is currently no central repository for large compiled biological datasets, although many published compiled datasets are accessible via dispersed project websites, and some links to large datasets are stored in diverse compilations, e.g., awesome-public-datasets. ([1, 11, 12, 16, 17, 23, 24, 27, 28].

2.3. Creating compiled sets of RNA-seq data and its metadata

A raw dataset can be compiled for meta-analysis. This involves a number of data normalization decisions. Creating a large dataset from multiple studies involves choosing which analytical approaches to use to combine the data and which statistical methods to use for normalization [1, 29–31]. These approaches have different strengths and weaknesses, and are likely to reveal different biological features. At one extreme, individual studies can be separately examined [32]. Information on expression levels is retained for each study, however, comparison of expression across experiments can be misleading and the statistical power of meta-analysis is lost.

Alternately, experiments can be co-normalized before meta-analysis. For example, the weighted mean of the correlations from individual studies could be calculated [32]. In another approach, data from individual studies could be pooled and then co-normalized using the same method for all runs, transforming the data to a common range [1, 33, 34]. Using the weighted mean approach reveals the most “true” correlation [32]; however, this method misses much about the biology that can be seen if the pooled method is used. A hypergeometric analysis of gene ontology (GO) terms [35] of the genes in the clusters obtained by weighted means method compared to the pooled method revealed both statistical approaches yield significant over-representation of GO terms [32, 33].

To implement meta-analysis, individual runs are grouped into samples, and the samples into studies. Each run in a study is identified by its experimental conditions. This is where accurate metadata describing the biological conditions,
design, and biological samples comes into the picture. A researcher’s flexibility and analysis options are greatly increased if s/he has access to a compendium of data with its comprehensive metadata. Conversely, research is hindered by an avoidable nightmare: incomplete, incorrect, and inaccessible metadata.

The ATH963 microarray dataset provides an example of how to build a high-quality dataset and the benefits that can be reaped. All microarray data available as of 2007 were retrieved from ArrayExpress \[1\]. After removal of samples with low replicate quality, this amounted to 71 experiments, 963 chips. A selection criterion for replicate reproducibility was imposed and samples with low quality replicates were removed. Data from each chip was normalized, after removal of outliers, such that the mean relative expression of a gene was a common value. The data were then pooled. The bottleneck in creating the dataset were the metadata, which were manually curated after downloading; this process took months and the end result was still lacking basic information about many samples. That said, analysis of the ATH963 dataset has led to hypotheses that have culminated significant discoveries \[1, 36, 37\]. For example, the surprising prediction that the chalcone isomerase-like FAP genes are involved in fatty acid metabolism \[36\]; a prediction that was later experimentally verified \[38\].

2.4. Meta-analysis and visualization

Existing software provides diverse analysis and visualization approaches to transcriptomics data \[10, 25, 36, 39–42\]. One example of visualization/analysis software, MetaOmGraph \[1, 43\], is a Java software that enables on-the-fly correlations of expression among all genes, as well as data sorting by, e.g., keywords and accumulation levels. MetaOmGraph empowers biologist’s explanation with use of metadata and analysis of large datasets. Researchers can analyze the pre-curated datasets for Arabidopsis that is available on MetaOmGraph, or can upload and analyze datasets of their own. The software is data-type agnostic, and can be used to analyze and visualize multiple data types (e.g., microarray, metabolomics, taxes, RNA-seq, grades, or epidemiology). Datasets can be large
(tens of thousands of samples by tens of thousands of genes). A researcher has direct access to all the data and metadata in the dataset. The researcher can calculate correlations on the fly for a selected gene against all other genes, display interactive results, and can analyze the entire dataset or a selected subset (e.g., all experiments representing reproductive structures and oxidative stress). One advantage of this approach is that the researcher drives the analysis, and can look at large datasets as a whole.

EfP browser [44, 45] is another example of a visualization software. It relies on selected sets of pre-curated, pre-analyzed experiments with high quality metadata. Rather than create a large combined dataset, experiments are analyzed independently. A biologist selects a gene and gets information about expression of that gene related to treatments, developmental stages, or organs. The expression of two genes can be compared, and the data can be displayed as a heat map of a plant. The datasets are smaller but accessible for download and metadata is well-curated. The variation among replicates and the details about the types of treatments are less easily available; the user must download each data set individually to examine it. This type of approach has the major advantage that little effort is required to get valuable information. However, if the experimental data from different studies is not combined, the user has less ability to determine the extent to which a change in gene expression is context-independent. Most importantly, if the data are not pooled or co-normalized, comparisons across experiments and experimental treatments can be misleading. A recent Shiny web application, ScanGEO enables the direct mining of all RNA-seq data in GEO. Like EfP, the analysis is on a gene-by-gene and experiment-by-experiment basis.

3. Orphans: New kids on the block

3.1. Something old, something new

The study of orphan genes provides a prime example of how metadata can be used to reveal gene function. New protein-coding protogenes are continuously
arising from a background of transcriptional and translational chatter. Most protogenes will not become fixed in evolution, however, some have selective value and may survive and evolve in structure, expression, and function. Thus, a given genome is comprised of genes that have emerged and become fixed in the genome at various times in evolution. Genes that originated early in evolution are far more likely to be well-characterized than genes of recent origin. Phylostratigraphic analysis, the classification of each gene in a species based on its emergence on a evolutionary time-scale, allows a researcher to identify and focus on genes of recent origin. One category of young genes, orphan genes (also called species-specific genes), is defined as genes unique to a given species, sharing no sequence similarity with genes from other taxa.

Both non-coding and protein coding orphan genes are likely present in every species. However, non-coding orphans are difficult to identify due to their usual low similarity to non-coding genes in other species. Furthermore, young but functional non-coding genes can disappear in the transcriptional noise; as a result, non-coding genes are seldom predicted or annotated. Protein-coding orphan genes are somewhat easier to identify than non-coding orphan genes because their ORF can be translated and used as part of the search criteria, and because ribosomal binding and proteomics can provide additional evidence that they are translated. Thus, protein coding orphans are partially annotated in most newly sequenced genomes.

Protein-coding orphan genes comprise ~1-10 percent of genes in most eukaryotic genomes. Using a conservative estimate of 15,000 genes/eukaryotic species, then there will be around 100 to 1000 orphan genes per species. Assuming there are around 10 million eukaryotic species on earth, then the total number of unique orphan protein sequences will be on the order of 1 to 10 billion.

3.2. What do they do?

Despite the multitude of orphans, only a handful of orphans or other highly lineage-specific genes have been studied from a functional perspective.
The antifreeze protein of the antarctic icefish and jellyfish toxin proteins were functionally characterized because of their dramatic properties; subsequently, these were determined to be orphans. The orphan gene dauerless, found in the zebrafish relative Pristionchus pacificus, controls the copy number of a pheromone-related gene, and thereby influences conversion of fish larvae to a stress resistant phenotype. The A. thaliana orphan gene QQS responds sensitively to biotic insults and interacts with the conserved NY-Y transcription factor complex to modulate carbon and nitrogen allocation and total protein content.

A study with E. coli showed that targeted expression of about 25 per cent of randomized ORF-containing sequences induced altered cell growth and response. This result reflects on a global level the potential for function of emergent, previously non-genic sequences even though they have not been subjected to selection pressure based on their transcription/translation.

Some orphans are recruited into unique stress responses, development-specific pathways, or defense networks - transient conditions where highly-optimized ancient proteins might be less dominant among the expressed genes. Such orphans may be expressed only under very limited developmental contexts or under specific environmental conditions. Several plant orphan genes from Arabidopsis thaliana, Oryza sativa, Solanum spp., and legumes have been shown to respond specifically to various abiotic and biotic stresses, or to express only in specific structures.

3.3. Meta-analysis of gene expression data is a best bet for first-pass delineation of orphan function

Ultimately, experiments are required to prove the function of a gene. One goal of functional annotation is to provide a basis upon which more detailed experimental studies can be designed. However, a genome’s orphan complement is usually left with default labels such as “Hypothetical protein” or “Unknown function”. This is because the standard homology-based inference, so useful for more established genes, cannot be used for orphan genes, that by definition have
no homologs. Thus, an alternative strategy is needed.

One approach to narrow in on gene function would be to use high throughput screens on orphan mutants [69]. However, it is difficult to design the screen or know what to screen for without prior information.

An alternative would be to infer function from computational analysis of motifs within the sequence. For example, the presence of signals targeting the protein for extra-cellular transport [70]. However, such motifs are rare in orphan genes.

Identification of an interacting protein or other molecule could provide an idea of what biological function an orphan might participate in. *Ab initio* structural prediction and subsequent docking simulations could be applied to predict proteins that an orphan gene might interact with. However, docking simulations have not yet been applied for any orphans, which are predominantly composed of “unknown” secondary structure [47], and which have only rarely been purified for analytical structural determination. Two hybrid screens have been used in at least one case to determine an interacting protein of an orphan protein [62, 71].

Another alternative to postulate function is to use high-throughput, homology independent data, e.g. ‘omics data. Diverse statistical and computational methods of meta-analysis can emphasize various features of gene expression, for example, Pearson’s or Spearman’s correlation [1, 72], Markov Clustering [1], Machine Learning methods [29, 73] or Mutual Information [30, 31]. A. thaliana orphan genes and other genes of unknown function have been placed into groups according to co-expression patterns [72] and by co-expression followed by Markov Clustering [1]. More comprehensive RNAseq datasets, combined with metabolomics and other ‘omics datasets, will enable yet more powerful analyses. Orphans are a wildly deviant group often with a highly variable pattern of expression [1]; thus, even more than for ancient genes, it is important to use a wide variety of experimental conditions in any meta-analysis to evaluate orphan gene function.
3.4. Meta-analysis with exploration to postulate A. thaliana orphan function

A. thaliana has around 1,000 orphans genes [47], which are varied in expression patterns and levels. Contrary to popular perception, some orphan genes have high expression under many conditions. Fig. 1 shows 20 of the orphans from A. thaliana whose transcripts accumulate to high levels under multiple conditions. The biological samples (X axis) have been sorted by level of QQS expression (a feature of the MetaOmGraph software [43] that was used for visualization). Of these 20 orphan genes, only QQS has been functionally characterized [62, 71]; the other 19 are annotated as, e.g., “unknown protein”.

Figure 1: Expression pattern of 20 promiscuously-expressed orphan genes across multiple genetic, environmental, and developmental conditions. Samples ordered by expression level of orphan gene AT3G30720(QQS). Each line represents the expression pattern of a single orphan gene. X axis is 483 samples. Y axis is gene expression level (mean of replicates). Genes expression values on each chip are normalized to an average value of 100. ATH963 dataset [1] is visualized in MetaOmGraph [43]. Mean expression value of genes is normalized to 100 (green arrow).

Fig. 2 depicts the patterns of expression of nine A. thaliana orphan genes that have sparse expression. Without data from samples of seeds at 8-10 days after fertilization, the conditions under which the orphan AT1G05450 is highly expressed, this gene may not have been considered as a gene nor annotated in the genome (Fig. 2A). Nor would researchers know which tissues to harvest to search for the presence of a AT1G05450 protein, or how to devise a test for the biological role for this gene. Clearly, to postulate functions for sparsely expressed orphans, the biological samples in a meta-analysis must encompass many environmental, developmental and genetic conditions.

Samples with high accumulation of AT1G31520 transcript are either derived from pollen or from shoot apical meristem (SAM). It is only because the dataset being visualized, ATH963 [1], contains samples of apical meristems of rosette shoots, that a researcher can understand that AT1G31520, though barely expressed under most conditions, is highly expressed in SAM (Fig. 2B), and that it might function in that complex progenitor of leaf and stem morphology and
<table>
<thead>
<tr>
<th>Locus ID</th>
<th>TAIR Annotation</th>
<th>Phylostratum</th>
<th>Pearson Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT1G31520</td>
<td>Unknown protein</td>
<td><em>Arabidopsis thaliana</em></td>
<td>1</td>
</tr>
<tr>
<td>AT4G19480</td>
<td>Unknown protein</td>
<td><em>Arabidopsis thaliana</em></td>
<td>0.95</td>
</tr>
<tr>
<td>AT3G26920</td>
<td>FBD/Leucine Rich Repeat domains containing protein</td>
<td>Magnoliophyta</td>
<td>0.93</td>
</tr>
<tr>
<td>AT3G11490</td>
<td>rac GTPase activating protein</td>
<td>Eukaryota</td>
<td>0.93</td>
</tr>
<tr>
<td>AT3G26120</td>
<td>terminal EAR1-like 1(TEl1)</td>
<td>Eukaryota</td>
<td>0.93</td>
</tr>
<tr>
<td>AT1G31430</td>
<td>Pentatricopeptide repeat, PPR-like) superfamily</td>
<td>Eukaryota</td>
<td>0.9</td>
</tr>
<tr>
<td>AT1G03445</td>
<td>BRI1 SUPPRESSOR 1(BSU1)</td>
<td>Cellular organisms</td>
<td>0.89</td>
</tr>
<tr>
<td>AT5G59900</td>
<td>Pentatricopeptide repeat, PPR-like) superfamily</td>
<td>Eukaryota</td>
<td>0.89</td>
</tr>
<tr>
<td>AT5G37620</td>
<td>Cys/His-rich C1 domain family-zinc ion binding</td>
<td>Embryophyta</td>
<td>0.88</td>
</tr>
<tr>
<td>AT4G09455</td>
<td>copia-like retrotransposon family</td>
<td>NA</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Table 1: The nine genes with expression patterns most highly correlated with *A. thaliana* orphan gene AT1G31520 across 483 conditions. The phylostratum classification is from [47]; the Pearson correlation is calculated using the MetaOmGraph software [43] and is based on the ATH963 dataset [1].
metabolism. Likely, using data from only a limited number of conditions led NCBI AceView [74] to report that the “AT1G31520 gene is expressed at a low level, only 3.2 percent of the average gene in this release”.

Figure 2: Expression patterns of 9 sparsely-expressed orphan genes across multiple genetic, environmental, and developmental conditions. The need for multiple samples obtained under a wide range of conditions is clear from the restricted expression patterns of these genes. Many datasets would not contain the specific conditions in which a given orphan is expressed. A) Samples ordered by expression level of orphan gene AT1G05450. AT1G05450 expression is detected in seeds at 8-10 days after pollination. B) Samples ordered by expression level of orphan gene AT1G31520. All samples that have detectable AT1G3152 expression are either meristems or pollen. Each line represents the expression pattern of a single orphan gene. X axis is 483 samples. Y axis is gene expression (mean of replicates). ATH963 dataset [4] is analyzed, sorted and visualized in MetaOmGraph [43]. Mean expression value of genes is normalized to 100 (green arrow).

To determine which, if any, genes might be expressed under the same conditions as AT1G31520, the MetaOmGraph software [43] was used for a correlation analysis of the accumulation of the AT1G31520 transcript as compared to the other 22,000 genes represented on the Affymetrix chip. The expression of AT1G31520 is highly correlated with nine other genes in the Arabidopsis genome across the varied experiments and conditions (Table 1). Those among the correlated genes that have any annotated function appear to be involved in regulatory processes. One such gene, BRI1 suppressor BSU1, is a member of the 4-gene Protein Phosphatase with Kelch-Like repeat domains (PPKL) family; the proteins from this family form homo- and hetero-oligomers, complexes that are needed for brassinosteroid signaling [75]. BSU1 inhibits BR signaling by dephosphorylating BR-signaling kinase BIN2, and hence targeting it for degradation [76–78]. Interestingly, until functional studies indicated otherwise, BSU1 was annotated as a “Hypothetical protein” with minimal evidence for its expression [79]. These results might lead a researcher to investigate a potential role of AT1G31520 in relation to BR signaling in SAM.

Fig. 3 shows the pattern of expression of AT1G31520 and these nine co-
expressed genes. The data is sorted according to terms in the metadata. (These sort terms were chosen based on the metadata from the Fig. 2B samples that showed high AT1G31520 accumulation.)

Fig. 2A indicates that some of the samples with significant accumulation of the AT1G31520 transcript are pollen; however, it is also possible that other samples of pollen did not accumulate AT1G31520 transcript. Perhaps AT1G31520 accumulation is specific only to mature pollen, or pollen developing under some particular environmental condition. Fig. 3A shows the 483 Arabidopsis samples after sorting to bring near the Y axis any study with the word “pollen” in it. Three studies, each with samples from multiple flowers parts, including pollen, petal, carpel, stamen and sepal, were identified in this sort. Stamen (which contain pollen) showed slight peaks of AT1G31520 accumulation; the samples composed exclusively of pollen show much higher peaks. The other flower parts did not have detectable AT1G31520 expression. Based on this somewhat limited set of pollen samples, AT1G31520 was expressed in pollen under all conditions tested but not in the rest of the flower.

In Fig. 3B, the terms “apical” OR “apex” were used to identify samples containing meristems; these would include SAM as well as samples from root meristem, inflorescence meristem, etc. Of the 483 conditions in this dataset, the highest expression of AT1G31520 was in SAMs of Col-0 seedlings grown under long days or continuous light. All SAM samples showed high AT1G31520 accumulation. Neither inflorescence meristems nor root meristems of Col-0 plants showed detectable expression of AT1G31520. SAMs from seedlings in ler, CS175 or CS184 showed little detectable expression of AT1G31520. Furthermore, all SAM samples from Col-0 plants showed high AT1G31520 accumulation.

The conditions under which a given orphan gene is expressed, and the genes of known function with which it is co-expressed, have enabled functional predictions in at least one case. Meta-analysis of the ATH963 dataset provided a key to elucidating the function of the QQS orphan gene in carbon and nitrogen allocation and stress response [54, 62], predictions that were experimentally verified [47, 62, 71, 80]. Meta-analysis also led to the hypothesis that the SAQR
Brassicaceae-specific gene is involved in the senescence response [55].

Figure 3: Genes with expression most highly correlated to the orphan gene AT1G31520. Annotations and Pearson correlations of the mostly highly co-expressed genes are shown in Table 1. A) Samples sorted according to key term “pollen” (grayed region). B) Samples sorted according to key terms “apical” OR “apex” (grayed region). The 9 genes shown are those among the 22,000 represented on the Affymetrix chip with the highest Pearson correlation to AT1G31520 across all samples. Each line represents the expression pattern of a single gene. X axis is 483 samples. Y axis is gene expression. ATH963 dataset [1] is analyzed, sorted and visualized in MetaOmGraph [43]. Mean expression value of genes is normalized to 100 (green arrow).

Thus, though orphans lack phylogenetic context, meta-analysis of transcriptomics data and metadata can provide a powerful high-throughput approach to place them in a physiological context. Genes can be clustered based on expression profile and these clusters of co-expressed genes (also referred to as regulons) can be linked to a potential physiological or developmental phenomenon based on the metadata of the samples with high expression of the gene of interest and the identity of other genes in the co-expression regulon. Orphans in such clusters can be considered as candidates for targeted experimental analysis. The more comprehensive the data and metadata, the more potential for an accurate functional hypotheses.

4. Standardization and simplification of metadata, its submission and retrieval

As exemplified for the orphan genes of Figs. 2 and 3, an ‘omics-based approach to predict gene function is dependent on high quality and complete metadata. Community efforts have led to development of standards for biological and analytical metadata [12, 81, 82]. MIAME [81], the seminal effort, describes microarray metadata standards, stating that “metadata should include the type of the experiment (such as normal-versus-diseased comparison, time course, dose response, and so on) and the experimental variables (... time, dose, genetic variation or response to a treatment or compound)” [81]. These
standards do not reflect the quality of much of the real-world metadata submissions in major databases. Metadata quality could be greatly improved by 1) more thoughtful use of the extant database submission systems by the submitter, and 2) improvement in submission system infrastructures and explanations by the database. Section 4 provides some tips for first-time metadata submitters and retrievers to facilitate data reuse.

4.1. Submission

The data/metadata submission process can be complex. There are currently a variety of ways by which RNAseq data and metadata can be submitted to a public repository, and they are not all created equal. GEO, SRA and ArrayExpress house the major submission systems, and the metadata and data are theoretically synchronized across them. Data is assigned IDs and stored in GEO, SRA, BioProject and BioSample for NCBI submissions and in ArrayExpress and ENA for EMBL database submissions. Fig. 4 compares the naming hierarchy across each of these databases, as defined in the user interface of that database. Multiple terms are used to refer to the same concept across the various databases; in this paper we use consistent terminology to minimize confusion. The various IDs assigned by NCBI/GEO, SRA, BioProject, BioSample and EMBL databases (ArrayExpress, ENA) sometimes appear to refer to the same information. Most notably confounding, the term “experiment” is used to refer to a “study” in ArrayExpress, but elsewhere “experiment” is used to refers to a subcategory of a sample: “unique sequencing results for a specific sample” (SRA) or “An experiment contains information about the sequencing experiments” (ENA).
4.1.1. Submitting via GEO: Highly Recommended.

RNA-seq short read data, processed data, and metadata can be submitted via GEO [88] or directly to SRA. The GEO submission format uses a single spreadsheet (excel or tsv) for each analytical Platform. Thus, all metadata about a study and its samples are in one place and easily compared (Supplementary file 1 has the template file along with two example sheets). The GEO portal prompts the submitter to select a template spreadsheet file based on the type of Platform used (e.g., Affymetrix, Agilent, Illumina, Nimblegen). The template spreadsheet varies according to the Platform. GEO has clear explanations of each metadata field in the spreadsheet (Series, Samples and Protocols) [89].

Each GEO spreadsheet contains columns for the submitter to enter all metadata concerning the study and its samples. The Summary field details the study goals and provides a description of the study. The Overall Design field enables submitters to describe the numbers and arrangements of samples, replicates, controls; Overall Design is important to understand the experiment and assess its quality. The Protocol field (total mRNA-seq, ribosomal footprinting, small RNA-seq, etc.), is a critical field for any future meta-analysis.

One downside to the GEO submission system is that there is no field in which to enter a Secondary Accession. A Secondary Accession would provide links to the same data in other databases, such as DDBJ or EBI. When contacted by email, GEO staff indicated that a secondary accession “could be added to the Overall Design or Summary field”, and indicated that it was “not their priority” to provide metadata for meta-analysis. (Although, the submitter could be encouraged to add any applicable Secondary Accession by a mouse-over prompt, which would take maybe 30 minutes of time for a GEO programmer).

GEO’s template file has fields for user-designated characteristics:tag that are specific to each sample (Supplementary file 1 is a template file for GEO submission). Some recommended example for the field characteristics:tag are provided. Because the GEO submission is via spreadsheet, a submitter knows at one glance what s/he has entered, and can easily change errors before sub-
mission. The final step is submitting the spreadsheet along with the raw and processed data files in a zip file to the GEO portal. Raw data for runs are automatically ported to the SRA database and assigned SRR-IDs (Fig. 4).

4.1.2. Submitting via SRA: not recommended

Submitting sequence read data and metadata via SRA [90] is complex, and involves the submitter filling out a series of web pages and spreadsheets, sometimes with multiple choices as to which should be completed next. Information about the submission process is dispersed across multiple web forms[91]; some of these explanations are incomplete or unclear, while others are useful (SRA Metadata Overview and Submission [90]; SRA Overview [92]; Search in SRA [93]; SRA Factsheet (Supplementary file 2). One example, the SRA Overview, does not mention that there are three possible nomenclatures for Study Accession IDs, depending on the source database to which the data was initially submitted: SRA designates studies as: SRP-ID if they are from SRA; ERP-ID, from EBI; or DRP-ID, from DDBJ. The multiple IDs are alluded to elsewhere in SRA [93], but no information is provided as to how these nomenclatures are assigned or accessed in the database.

SRA encourages SRA data submission via the SRA Submission Portal system [94] (Supplementary File 3). Three separate submissions are required. Specifically, the submitter registers the study via BioProject Submission Portal [95], the sample(s) via BioSample Submission Portal [96], and the platform metadata via SRA Submission Portal. BioProject Submission Portal users submit data about the study through a series of web forms and a template file (Supplementary file 4). Each BioProject Submission requires a Project Type, Organism Name, Description of the study, and provides an option of adding a single BioSample ID. A BioProject ID is assigned once the completed web form is submitted. Samples are submitted via BioSample [96]. BioSample Submission Portal provides a series of web forms and a spreadsheet to which the user submits metadata about the sample/samples (Supplementary file 5 shows the template file for entering Biosamples in case of Plants). To link BioSam-
ples to a BioProject, the submitter provides the BioProject ID. Then, the user goes to the SRA Wizard web page of the SRA Submission Portal. The SRA Submission Portal requires the submitter to enter his/her name and other contact information. The subsequent web form prompts the submitter to enter the BioProject ID, BioSample IDs and a release date for the submission. The next web form enables the submitter to download a spreadsheet template file and populate it with information about the analytical Platforms (eg. Helicos, ABI Solid, PacBio SMRT) and about the (processing) Strategy (eg. RNAseq, WGS, miRNAseq) which is similar to the Protocols entry in a GEO submission. Each field is assigned a column within the spreadsheet. The field entries have a drop-down menu, creating a controlled vocabulary. Some fields have pop-up comments or definitions, which give the submitter guidance as to how to fill in the fields. This spreadsheet can be edited by the submitter before submission. The spreadsheet provided no field to allow a user to enter a Secondary Accession (such as a GEO accession or an ArrayExpress accession). Because the submission process is more complex than submitting to GEO, there are many more opportunities for errors to be introduced.

Submission of data/metadata via the SRA User Interface is insufferable and should be avoided if possible (Supplementary file 6 shows required fields vs. not required fields). The SRA interface itself no longer recommends using this method to submit metadata. (Initially, this was the only way to submit data to SRA).

4.1.3. Submitting via ArrayExpress

The new ArrayExpress submission system, Annotare, provides a simple eight-step data/metadata submission system with clear submission guidelines. Like GEO, it accepts microarray and Next Generation Sequencing studies. For each data type, pre-submission checklists provide the submitter an idea of what s/he should have in hand. Submission is via a series of webpages. Comprehensive metadata fields include Sample Name, Material Type, Organism. Like GEO, variable samples attributes that are specific to the study are added.
by the user. The submissions can be saved as spreadsheets. A validation step
ensures that all fields are filled in, but not whether they make any logical sense.
Raw data and some metadata is printed to the ENA database.

GEO and ArrayExpress have equally intuitive submissions systems; the very
major downside to ArrayExpress is that the data and metadata in ArrayExpress
is much less easily accessible for meta-analysis (discussed in Section 4.3).

4.1.4. Unique IDs at NCBI and ArrayExpress

During the submission process, each database generates unique IDs (Fig.
4). Understanding to what these IDs refer is critical to being able to design
effective queries for data retrieval. In total, six unique IDs are created for
an RNAseq submission to NCBI: BioProject ID (PRJNA-ID), BioSample ID
(SAMN-ID), SRA Study ID (SRP-ID), SRA Sample ID (SRS-ID), SRA Exper-
iment ID (SRX-ID), and SRA Run ID (SRR-ID). If the entry was submitted
initially via GEO, three more unique identifiers are generated: a Series ID
(GSE-ID), a Platform ID (GPL-ID), and a GEO Sample ID (GSM-ID); only
the GSE-ID is displayed in the corresponding SRA entry, and thus the GSE-ID
provides a critical unique identifier and link between the metadata in GEO and
the run data in SRA. The BioProject ID (PRJNA-ID), SRA Study ID (SRP-ID)
and GEO Series ID (GSE-ID) seem to point to identical data.

A submission to GEO enables users to easily access the corresponding raw
data files, as they are automatically stored in SRA via the GEO Submission.
So, for any GEO entry describing an RNAseq experiment, the corresponding
raw file(s) are present in the SRA database. ArrayExpress assigns a single
ID to each study, the ID varies according to the database to which the data and
metadata was initially submitted. Four additional IDs (SAMEA-ID, ERX-ID,
ERR-ID and PRJEB-ID) are assigned to ArrayExpress submissions for the ENA
database. The ENA database is rarely used in retrieving data and metadata for
re-use.
4.2. Submission errors and revisions

Submission in GEO is based almost entirely on a single editable spreadsheet; thus, all information is handy until the point of submission and errors can be easily corrected. In contrast, because the SRA User Interface involves three separate submissions (to BioProject, BioSample and the SRA User Interface), as well as a series of web submission pages that need to be sequentially understood and completed, errors are easily introduced. Furthermore, it is quite difficult to correct errors that may have occurred during an SRA submission. Changes to most parts of a submission require the submitter to send an email to the SRA staff, with a request to update/edit. This process is complex and can be confusing. (For an actual sample of an email exchange between an SRA staff member and a submitter, with identifying information removed, see Supplementary File 7).

4.3. Tools for retrieving metadata: A test case with maize RNAseq data and metadata

A prerequisite to data/metadata reuse is being able to efficiently search, retrieve and understand the biological and technical parameters of that data/metadata. Several methods for data/metadata retrieval are possible for each database. We illustrate these methods using Zea mays coding RNAseq data/metadata. We then assess the quality of the retrieved metadata.

Data and metadata can be retrieved from SRA using the SRA online search engine [100] or the SRAdb R toolkit [101, 102] (Fig. 5). The SRAdb R toolkit includes dbGetQuery and getSRA, two APIs that are useful for metadata retrieval. We compared the functionality of these standard data retrieval approaches in retrieving all Zea mays, paired-end, Illumina, coding RNA transcriptomic runs (Fig. 5). The SRA online search engine identified 125 studies (3519 runs). However, only limited metadata could be retrieved, and some runs were not relevant to the search query (eg. genomic source, metagenomic source, etc.). The two SRAdb R toolkit software gave similar results— dbGetQuery API retrieved 96 studies (3251 runs) while the getSRA API retrieved 102 studies (3591 runs)
(Supplementary file 8 contains R scripts used for retrieval of study, sample, experiment and run metadata from SRA). For each study, the metadata was complete and all samples and runs were retrieved in a single spreadsheet. Some of the samples or runs were irrelevant to the search query (e.g., single-end, meta-transcriptomic, genomic, or metagenomic reads), and these runs were filtered out manually (Fig. 5).

We used the secondary accession IDs (GSE-IDs and E-MTAB-IDs) obtained from SRA Study Alias and Study Attribute fields (search of Fig. 5) to retrieve the corresponding metadata from the GEO and ArrayExpress databases (Fig. 6). The GSE-IDs were queried via the GEOmetadb R toolkit dbgetQuery API [103, 104] to retrieve the GEO run and study metadata; this metadata was complete and didn’t require filtering out of irrelevant runs. A flowchart detailing the search parameters obtained from GEO and ArrayExpress is shown in Fig. 6.

A SQL query of the Zea mays data in GEO using the GEOmetadb R toolkit, which is designed based on the structure of that database, retrieved 6901 runs. This exceeded the number of the GEO runs obtained through secondary accessions provided by SRA metadata with many runs that did not meet the search criteria. This is because GEOmetadb R toolkit, the query structure doesn’t allow searching for fields containing the keywords 'TRANSCRIPTOMIC', 'ILLUMINA' or 'PAIRED' while the SRA query structure allowed use of these specific keywords. It allows the search, it just doesn’t retrieve the data. Thus, many irrelevant runs are obtained, and these would need to be filtered out before a meta-analysis of the data.

ArrayExpress was problematic for search and retrieval of metadata corresponding to multiple experiments. Like GEO and SRA, ArrayExpress has an R package, ArrayExpress R package [105, 106], and a search engine [107] for data retrieval. However, retrieving metadata using the R package required that three different kinds of files (IDF, SDRF, and processed files) are present in the database. Because most studies in ArrayExpress do not include all three file types, retrieving metadata using the 199 E-MTAB-IDs from SRA did not
provide any result. Searching the same 199 E-MTAB-IDs via the ArrayExpress search engine retrieved metadata from four studies; however, all of these studies had been originally submitted via SRA.

Retrieving metadata using the ArrayExpress search engine directly, using the query 'Zea mays AND “coding RNA” NOT “non coding RNA”' and then filtering out those studies that were not from Zea mays, retrieved 106 studies.

Neither the ArrayExpress search engine nor the ArrayExpress R Package provided a pipeline to retrieve metadata for the combined runs from multiple studies; rather, runs must be accessed study by study. This requires the user to navigate through each study to each sample. Then, the researcher must filter out from each study the runs that are irrelevant to the retriever’s purpose (e.g., in this test case, single end read runs, ribosomal profiling, or DNA seq runs); this process is tedious. We did not develop code for JSON queries, Rest-style queries, (the former two being additional approaches suggested to query the ArrayExpress database [108]) nor did we study the ArrayExpress and ENA database schema to develop SQL queries. Thus, we cannot comment on the efficacy of these methods in retrieving studies nor evaluate the quality of the ArrayExpress metadata. The ArrayExpress database has links to the ENA database. This database stores sequence reads data and metadata. A search of ENA (’“Zea mays” AND LibrarySource=“TRANSCRIPTOMIC” AND LibraryLayout=“PAIRED” AND InstrumentModel=“ILLUMINA”) retrieved a total of 114 studies, with limited metadata. The studies in the ENA database do not appear to reference an external database.

GEO stores all submitted metadata while SRA stores all the corresponding raw data files. Since SRA has the most intuitive query structure as stated above, the simplest method to access comprehensive metadata submitted to GEO would be to query SRA, obtain corresponding GEO accessions and ultimately retrieve metadata from GEO using these accessions.
Figure 5: Workflow of search and retrieval in the SRA database. The SRA database engine allows an advanced search but doesn’t allow metadata to be downloaded. The SRAdb R toolkit [101] has two APIs, dbGetQuery and getSRA. dbGetQuery allows SQL queries while getSRA allows a text search to retrieve metadata. Both of the queries retrieved varying number of runs and both of them contained unwanted entries which required filtering it out. *The metadata retrieved by getSRA API was used as a test case for missing annotations. GSE-IDs, GSM-IDs and E-MTAB-IDs were used to retrieve metadata from GEO and ArrayExpress, respectively (see Fig. 6).

4.4. Missing and confusing annotations: a test case

Datasets with missing or incorrect metadata are much harder to reuse and more open to misinterpretation. Information is lost if related terms are not merged into a common categorical variable. We illustrate the extent of this annotation challenge by analysis of all metadata from the 2341 paired end Illumina coding RNAseq runs from Zea mays (i.e, the set of all runs available in SRA as of 03/01/2017 and retrieved via the getSRA API and the subset of those runs retrieved via GEO) (Fig. 5 and Supplementary file 9). Because of the more limited search and retrieve capability of ArrayExpress, we did not include this metadata in our analysis.

Figure 6: Workflow of search in GEO and ArrayExpress. The metadata from the SRA search (Fig. 5) provided GEO Series and Sample IDs which were used to retrieve GEO metadata using the GEOmetadb R toolkit. The same SRA metadata also contained ArrayExpress E-MTAB-IDs, which were used to retrieve ArrayExpress metadata. The ArrayExpress R package didn’t retrieve any metadata (explained in text).

Figure 7: Metadata entries for coding RNA-seq Zea mays runs from the SRA database. Metadata for 2341 paired-end Illumina runs were retrieved using the SRA R toolkit (Fig. 5). For each field, the number of blank entries, the number actually filled in by the submitter (helpful metadata) or the number filled in automatically filled by SRAdb (Unique IDs assigned) are shown. *metadata imported from BioSample entries (required to be filled out by submitters prior to an SRA submission); **metadata imported from the BioProject (required to be filled out by submitters prior to an SRA submission). Seventeen fields were left blank by all submitters.
4.4.1. SRA

In total, 61 metadata fields were retrieved from the SRA search (Fig. 7). Twenty seven percent of the metadata fields were left blank by every submitter; these included fields conveying important information, such as Sample URL link, Primary Study, Study Attribute and Study Description. Thirty-seven percent of the fields were filled in by some of the submitters. For example, 54 percent of submitters left the Experiment attribute field blank; 48 percent of the submitters left the Design Description field empty (Fig. 8). Twenty-nine per cent of the fields were always filled in by submitters: Library Layout, Taxon ID, Updated Date and Instrument Model were among these. Ten percent of the fields were generated automatically by SRA, including the various IDs.

Confusing or non-helpful annotations are extensive, and data fields are often filled in with inappropriate information (Fig. 9 and Supplementary file 10 part a and b). For example, one entry for the field Experiment Name is “BGl-CAUO5-CORLJSD3AE”; an entry in the Experiment Title field is “LL100”, another “lane2.bc8”; an entry in the Sample Alias field is “FGMG-O2”; a Sample entry, “71d23972-1dc9-11e6-bfeb-000e1e0af2dc”; the submitter may understand these entries, but they are not useful to the broader community.

4.4.2. GEO

One thousand seven hundred and forty seven of the 2341 maize RNAseq runs in SRA (Fig. 5) also had a GEO sample entry (meaning that the study was submitted via GEO)(Fig. 6). These corresponded to sixty five studies. GEO entries tend to have much more comprehensive metadata than SRA. This more comprehensive metadata is likely because: the GEO submission process is clear and relatively simple; the GEO field names are intuitive; and the GEO submitter is encouraged to use controlled vocabulary. GEO has fewer required metadata fields relative to SRA, yet it covers the important information about the experimental conditions. The various fields provided by the submitter to GEO in the format characteristics:tag are merged into one field, characteristics, by the database before submission, which makes metadata retrieval and reuse
easy.

Figure 8: Metadata entries for coding RNA-seq *Zea mays* runs from the GEO database. The metadata fields for Studies/Series and Samples for each GEO submission (1747 paired-end Illumina runs) were retrieved using the GEOmetadb R toolkit (Fig. 6). (Supplementary file 8 contains R scripts used for retrieval of series and sample metadata from GEO.) For each field, the number of blank entries, the number filled in by the submitter, or the number filled in automatically filled by GEO (mostly IDs) are shown. The metadata entered in GEO was more completely annotated than that entered in SRA. Only 1 field out of the total 34 fields were left blank by all the submitters. 13 fields were always filled in by all the submitters. The field *characteristics* is a submitter-designated compilation of the characteristics for each sample, such as line, tissue and developmental stage.

Figure 9: Confusing and counterintuitive metadata entries from SRA, GEO and ArrayExpress. A few examples, of many, from the metadata for coding RNA from *Zea mays*. Bold headings are fields.

GEO submissions include metadata for studies and samples (Fig. 8). Of the 32 fields retrieved from the GEO entries, 21 were filled in by all submitters, and four were predominantly blank. A qualitative assessment indicated the fields *Overall Design*, *Summary* and *Type* were generally quite well written and informative from a data-reuse perspective. Six unique ID fields were automatically populated by GEO. Sample metadata was generally good quality. Less informative entries included, for example, one entry in the field *Title* was “10-OPEA-1”, another was “10R” (Fig. 9). See [109] for a GEO submission with exemplary metadata, submitted by Ruairidh Sawers.[110]

4.4.3. ArrayExpress

Due to the time required for merging and manually analyzing the ArrayExpress metadata study by study (see Section 4.3), we did not systematically evaluate metadata quality for ArrayExpress. However, we did examine metadata from 10 studies. ArrayExpress, like GEO, has metadata fields called *characteristics*, which include tags. The submitter generates the *characteristics* for each experiment that best describe her/his experimental variables. However,
ArrayExpress compartmentalizes metadata rather than merging all characteristics into one field. Thus, extracting and combining ArrayExpress metadata from runs from multiple studies may require a researcher to merge all the characteristics columns into a single field, as well as to merge the run metadata from the experiments (Merging characteristics can be done using the “find and merge” function in Excel).

4.5. Incorrect annotations

Some RNA-seq data is wrongly annotated. One example is a study that annotated the *Saccharomyces cerevisiae* strain incorrectly for all but the first sample, presumably as a consequence of “dragging” the strain number through the column in the spreadsheet, the strain number increasing by one at each row (See Supplementary file 10).

Another example is provided in a clever study of metadata from 4,010 microarrays chips based on humans derived from 70 experiments in GEO, which investigated the number of mis-annotated samples by evaluating the expression of sex-specific genes. Forty-six percent of these experiments had one or more samples that were mislabeled with respect to gender (male or female) [111].

Identifying mis-annotated samples is challenging unless there are clear cut markers (such as the sex-specific genes used in [111]) that identify particular conditions. However, if only some of the run replicates comprising a given sample are mislabeled, cross-replicate data comparisons at the time of data submission could help to eliminate some of the mislabeling. A useful tool for cross-replicate comparisons might be NGSCheckMate [112]. Potentially misannotated replicates could then be flagged for the researcher/PI to review.

4.6. Retrieving common submissions from SRA, GEO and ArrayExpress

A promising approach to improve the metadata for a meta-analysis would be to pool the metadata submitted to the three databases. The sequence databases share data on a weekly basis [87], therefore information in GEO, SRA and ArrayExpress is theoretically synchronized; however, this does not appear to always
be the case. Some experiments in SRA are not in ArrayExpress. ArrayExpress does not provide easily apparent accessions/IDs that link to other databases. It does not appear that data flows from ArrayExpress to GEO. None of this is stated on the respective websites. In practice, retrieving submissions that were common to GEO, SRA and ArrayExpress was non-trivial.

In our case study of maize RNAseq data, not a single Sample ID was retrieved that was common to all three databases using any of the R-toolkits or search engines. It should not take weeks (or months) for a researcher to retrieve and curate sets of metadata and data. A single intuitive portal and database (worldwide) with a consistent clearly defined structure and terminologies, would save time and reduce the number of databases that a researcher would need to understand and search in order to retrieve comprehensive data and metadata.

5. Controlled vocabulary

To easily compare data and metadata across submissions, it is imperative that the metadata contains standard terms, which submitters can choose from pre-designed libraries/ontologies. For example, if the term “iron stress” is used in the title of every sequencing run in which high or low iron is a variable, it is simple to identify those experiments in which “iron stress” is a variable. A researcher could leverage data from these experiments to gain new information, such as a fingerprint of those genes up-regulated under high iron, regardless of other environmental and developmental contexts. Extrapolating to the orphan theme, a researcher could determine whether any orphan gene was part of this fingerprint. If so, a reasonable postulate might be that the orphan plays a role in iron metabolism or stress. The researcher could then design an experiment to test her/his hypothesis. However, if one submitter uses the term “iron stress” for a plant grown under unusually high iron, while others describe a similar treatment as “Fe stress”, “FTH”, “high iron”, “metal”, “treated” or “iron-treated”, this type of analysis would be more difficult and require much more human time in searching and classifying the metadata.
Likewise, using a controlled vocabulary is important for gleaning insight from other types of annotations, such as characteristics designations. Suppose an experiment involved biotic stress (e.g., fungal disease), one submitter might create a metadata field column headed `characteristics:fungal type` with entries describing the fungal species used in each treatment. Another user might create a column headed `characteristics:pathogen type`, also with entries describing the fungal species used in each treatment. A third submitter might create a metadata field column headed `characteristics:treatment` with entries describing the fungal species used in each treatment.

RNA isolation and sequencing methods provide examples of a particular controlled vocabulary challenge, because of the rapidly changing nature of the technologies, methodologies, sequencing apparatus and resolutions. Yet, it is critical to for a researcher to be aware of what data types and analytical platforms comprise a given meta-analysis.

In summary, inconsistencies in naming can hinder high-throughput meta-analyses and vastly reduce reusability of data. Conversely, metadata that are well-annotated with controlled vocabulary (ontological terms, tabs, mesh terms, keywords) enable a researcher to extract and analyze sets of experiments and samples using flexible criteria, enriching the value and understanding of the analysis.

5.1. Ontologies

The 2001 MIAME standards state “A list of qualifiers initially left at a submitter’s discretion may progressively be made standard when applicable ontologies are made public.... Provided in a format of controlled vocabularies, these will enable accurate queries and more formal data analysis than free-text descriptions” [81]. Indeed, ontologies can provide excellent avenues to standardize metadata [16, 39, 113, 114]. Unfortunately, there is still a paucity of comprehensive ontologies: biology may be far too complex for comprehensive ontologies.
5.2. Tabs/MeSH terms/keywords

Tabs/MeSH terms/keywords can alleviate the challenge of developing ontologies, yet still enable the terms used for conditions, tissues, organisms and strains to be compatible across studies, creating a uniformity that is maintained across all the metadata, thus rendering the metadata easier to process and search. GEO and ArrayExpress prompt the users to submit these types of information. For example, in GEO the submitter is instructed to fill in “as much metadata as is needed” into the characteristic:tag fields “to describe the variable and key characteristics” of her/his biological samples (e.g., characteristic:plant part, characteristic:time of day, and characteristic:mutant). Although this approach still involves subjectivity as to which characteristics are important and how they should be defined, from a data-reuse perspective these fields can be extremely beneficial. A greater emphasis and integration of use of MeSH terms into the characteristics would further standardize metadata.

5.3. Reasons for inadequate metadata and metadata retrieval.

Major challenges on the database side include: the lack of a comprehensive validation system for submission; the difficulty of connecting within and across the varied SQL-based relational databases used by SRA, GEO, and ArrayExpress, the context of big data with ever-changing data types and requirements; and the inadequate funding or appreciation for curation. These challenges are exacerbated in SRA by the complex process of metadata submission, metadata fields that are not well-defined, and metadata fields that are (or appear to be) repetitive. For ArrayExpress, big-(meta)data retrieval is problematic because the current structure of the database supports querying and retrieving individual experiments, and then going within each experiment to examine the runs, but does not appear to support retrieving samples/runs from multiple experiments en masse.

Challenges on the submission side are that many individual submitters are not trained in the submission process or in its relevance. Metadata/data submitters are often students who may not realize the extent to which providing
comprehensive metadata can increase the impact of their own research, nor how they themselves can reuse metadata/data to develop hypotheses, nor how a well-annotated public repository for biological data can benefit the scientific community. Individual PIs are ultimately responsible for training their students in data submission, however, not all PIs understand the extent to which metadata enables re-purposing of data, most are busy, and submission itself can be cumbersome and confusing. There is little incentive for excellence in metadata.

6. Recommendations

The petabases of RNAseq data being generated via elegant, well-executed experiments are deposited in data archives. The data is often used only one time by it's creators for one purpose. Yet, within this data is valuable information that could be re-used to generate new insight into complex biological challenges.

We present a series of recommendations, each of which on it's own could increase the information content in the metadata and/or efficiency of its retrieval, and thus provide powerful data re-use capabilities.

Recommendations for submitters: Submit comprehensive, accurate metadata. To enhance the reusability of omics data, we advocate that data submitters expand the metadata in their submissions by including controlled vocabulary-tags/keywords/MeSH terms and ontology terms to describe their experiment. This would enable a (computational) parser to more efficiently extract the terms and apply them to the analysis. Well-annotated datasets enable a researcher to analyze multiple experiments using a flexible variety of criteria. We also recommend that PIs check all metadata submissions and take responsibility for them. Metadata submission provides a great opportunity to train students about the importance of consistent, controlled-vocabulary and comprehensive data in a public database. Finally, because the large-scale retrieval of quality data/metadata is the most straightforward for GEO-submitted metadata, we recommend submission of data/metadata via GEO, in order to maximize potential impact of data with its metadata.
Recommendations for journals and funding agencies: Implement policies of metadata/data review. Reviews could take a variety of forms. Journals could establish a policy in which the metadata/data submissions described in the manuscript are included in the review process. This would put more burden on reviewers, however, it would improve the utility of public data. Federal funding agencies could implement a random spot-check of the metadata and data submissions included in annual reports. There is little so frightening to PIs as to lose their funding. Such a spot-check would ensure much higher quality metadata. Journals could require the approval of each data/metadata submission by the PI before the data is entered into the database. A positive incentive to increase the quality of metadata would be an award for best metadata submissions by students.

Recommendations for biologically-challenged meta-analyzers (read, computational geeks): Include biologists in any effort. When analyzing metadata and data, it is imperative that biologists verify that the metadata is appropriately designated. A biologist can often identify major flaws in the metadata classifications, and/or can become drawn into the analysis to create case studies or suggest hypothesis.

Recommendations for using meta-analysis of RNAseq data to explore gene function. Bioinformatics can provide many clues as to a gene’s function or a biological process. The same data can, and should be, analyzed in multiple ways. Going to one website and pulling out an expression result for a gene is rarely sufficient. A researcher should be skeptical, and take into account the methods by which a meta-analysis was conducted. Looking at multiple datasets or huge datasets can help to derive valid hypotheses. And, if a gene encoding an “hypothetical protein” comes your way, take the chance.

Recommendation for public archives: a single intuitive metadata submission system. Ideally, metadata should be submitted, stored and retrieved via a single system, world-wide. It is likely that a NoSQL database platform [116] for metadata/data storage, such as MongoDB [117], would be preferable to the current
table-based relational database storage. NoSQL platforms, used currently by most giant companies \[118\] as well as by the 'omics databases such as PMR \[10, 119\], would provide the greater flexibility and more rapid data retrieval that is needed for a rapidly growing and changing data storage system.

In the near term, we recommend submission of RNAseq data via GEO, because of the relative advantages of that system in enabling efficient retrieval of large-scale high quality metadata and data. A specific corollary: SRA should redirect potential submitters to GEO. This would require only a simple prompt to any would-be SRA submitter, for example, “Please submit all RNA-Seq metadata via GEO”, with a link to the GEO submission site. It would take twenty minutes for an SRA programmer and ultimately save many millions of research dollars.

To ensure the best product, federally-funded metadata/data submission systems should be subject to public review starting in the planning stages and through beta releases, with comments/ratings from reviewers posted online (similar to reviews at Amazon or Yelp). Reviews would provide important guidance to the developers of the metadata/data submission-system, making them more aware of community needs and more responsive to the end-users.

Recommendation for a public site for metadata compilations. A public site should be created to store large, well-curated sets of metadata that have been generated by researchers, with links to the data. Upon publication of a meta-analysis, the relevant metadata set could be deposited in this site, with a link to the metadata set in the publication. Such metadata sets would provide a valuable resource for the biological and computational communities.

7. Funding

This research was funded in part by National Science Foundation IOS 1546858(to ESW) and by the Center for Metabolic Biology at Iowa State University (in partial support of PB, AS, ZA, MH).
7.1. Conflict of interest statement.
None declared.

References


URL https://github.com/caesar0301/awesome-public-datasets


[34] In RNA-Seq, 2 != 2: Between-sample normalization (Dec. Date accessed 2017-05-09).


[43] MetaOmGraph (Date accessed 2017-05-09).
URL http://metnetweb.gdcb.iastate.edu/MetNet_MetaOmGraph.htm


URL http://bar.utoronto.ca/efp/cgi-bin/efpWeb.cgi


URL http://dx.doi.org/10.1038/s41559-017-0127


[84] SRA Metadata Overview and SRA submission overview (Date accessed 2017-05-11).


[86] Submit project and samples for SRA (Date accessed 2017-05-11).
[87] GEO Data < Help < ArrayExpress < EMBL-EBI (Date accessed 2017-05-10).
URL https://www.ebi.ac.uk/arrayexpress/help/GEO_data.html

[88] GEO Submitter Login - GEO - NCBI (Date accessed 2017-05-09).

[89] GEOarchive submission instructions - GEO - NCBI (Date accessed 2017-05-09).

[90] Submission Portal (Date accessed 2017-05-09).

[91] SRA Submission Quick Start (Date accessed 2017-05-09).

[92] The Sequence Read Archive (SRA) Overview (Date accessed 2017-05-09).

[93] Searching and Browsing in SRA (Date accessed 2017-05-09).

[94] Submissions | Sequence Read Archive (SRA) | Submission Portal (Date accessed 2017-05-09).

[95] Submissions | BioProject | Submission Portal (Date accessed 2017-05-09).


[98] Submissions Overview < ArrayExpress < EMBL-EBI (Date accessed 2017-05-09).
URL https://www.ebi.ac.uk/arrayexpress/submit/overview.html


[100] Home - SRA - NCBI (Date accessed 2017-05-09).

[101] Bioconductor - SRAdb (Date accessed 2017-05-09).


[103] Bioconductor - GEOmetadb (Date accessed 2017-05-09).


[105] Bioconductor - ArrayExpress (Date accessed 2017-05-09).
URL http://www.bioconductor.org/packages//2.7/bioc/html/ArrayExpress.html


[107] ArrayExpress < EMBL-EBI (Date accessed 2017-05-09).
URL https://www.ebi.ac.uk/arrayexpress/
URL https://www.ebi.ac.uk/arrayexpress/help/programmatic_access.html#Protocols

[109] GEO Accession viewer (Date accessed 2017-05-10).


[115] Medical Subject Headings - Home Page (Date accessed 2017-06-12).

[116] Best NoSQL Databases Software in 2017 (Date accessed 2017-05-21).
URL https://www.g2crowd.com/categories/nosql-databases

URL https://www.mongodb.com/index
[118] Announcing Google Cloud Bigtable: The same database that powers Google Search, Gmail and Analytics
(Date accessed 2017-05-21).

URL http://metnetweb.gdcb.iastate.edu/PMR/
<table>
<thead>
<tr>
<th>Performer</th>
<th>Source Name</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>group 1</td>
<td>Sample 2</td>
<td>1L</td>
</tr>
<tr>
<td>group 1</td>
<td>Sample 2</td>
<td>2L</td>
</tr>
<tr>
<td>group 2</td>
<td>Sample 3</td>
<td>3L</td>
</tr>
<tr>
<td>group 2</td>
<td>Sample 3</td>
<td>4L</td>
</tr>
<tr>
<td>group 2</td>
<td>Sample 3</td>
<td>5L</td>
</tr>
<tr>
<td>group 3</td>
<td>Sample 4</td>
<td>6L</td>
</tr>
<tr>
<td>group 3</td>
<td>Sample 4</td>
<td>7L</td>
</tr>
<tr>
<td>group 3</td>
<td>Sample 4</td>
<td>8L</td>
</tr>
<tr>
<td>group 4</td>
<td>Sample 5</td>
<td>9L</td>
</tr>
<tr>
<td>group 4</td>
<td>Sample 5</td>
<td>10R</td>
</tr>
<tr>
<td>group 4</td>
<td>Sample 6</td>
<td>11R</td>
</tr>
<tr>
<td>group 5</td>
<td>Sample 7</td>
<td>12R</td>
</tr>
<tr>
<td>group 5</td>
<td>Sample 7</td>
<td>13R</td>
</tr>
<tr>
<td>group 5</td>
<td>Sample 8</td>
<td>14R</td>
</tr>
</tbody>
</table>

### Database

<table>
<thead>
<tr>
<th>experiment_name</th>
<th>experiment_title</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>BGI-CAU05-CORLJSD3AE</td>
<td>LL100</td>
<td>10-OPEA-1</td>
</tr>
<tr>
<td>BGI-CAU05-CORLJSD3AE</td>
<td>LL150</td>
<td>10-OPEA-2</td>
</tr>
<tr>
<td>BGI-CAU05-CORLJSD3AE</td>
<td>LL200</td>
<td>10-OPEA-3</td>
</tr>
<tr>
<td>BGI-CAU05-CORLJSD3AE</td>
<td>LR100</td>
<td>10-OPEA-4</td>
</tr>
<tr>
<td>BGI-CAU04-CORLJSD3AD</td>
<td>LR150</td>
<td>12-OPDA-1</td>
</tr>
<tr>
<td>GSM1379803</td>
<td>LR200</td>
<td>12-OPDA-2</td>
</tr>
<tr>
<td>GSM1652858</td>
<td></td>
<td>12-OPDA-3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12-OPDA-4</td>
</tr>
</tbody>
</table>
SRA Database Search Engine

- SRA DataBase Search Engine
  - SRAdb
    - Retrieved 3519 hits, 125 studies
    - Filter not needed
    - 3519 Runs, 125 Studies with limited metadata

SRAdb R Toolkit

- dbGetQuery API
  - Retrieved 3274 runs, 110 studies
  - Filtered out 910 runs (genomic, metagenomic [source]) and single end [layout]
  - 2945 Runs, 92 Studies with complete SRA metadata

- getSRA API
  - Retrieved 3591 Runs, 265 Studies
  - Filtered out 361 runs (genomic, metatranscriptomic, metagenomic [library source]) and single end [library layout]
  - 2341 Runs, 96 Studies with complete SRA metadata (Fig. 7)

- rs <- getSRA(search_terms = "Zea mays transcriptomic paired Illumina RNA-seq", out_types = c('submission', 'study', 'experiment', 'sample', 'run'), sra_con)
  - 1747 GSE-IDs & GSM-IDs for GEO metadata (Fig. 6)
  - 199 E-MTAB-IDs for ArrayExpress metadata (Fig. 6)