

PRACTICE PAPER

A FAIR-Based Approach to Enhancing the Discovery and Re-Use of Transcriptomic Data Assets for Nuclear Receptor Signaling Pathways

Scott A. Ochsner^{1,2}, Yolanda F. Darlington^{1,3}, Apollo McOwiti^{1,3}, Wasula H. Kankanamge^{1,3}, Alexey Naumov^{1,3}, Lauren B. Becnel^{1,3} and Neil J. McKenna^{1,2}

¹ Nuclear Receptor Signaling Atlas (NURSA) Informatics, One Baylor Plaza, Houston, Texas, 77030, US

² Department of Molecular and Cellular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, Texas, 77030, US

³ Dan L. Duncan Cancer Center Biomedical Informatics Group, Baylor College of Medicine, One Baylor Plaza, Houston, Texas, 77030, US

Corresponding author: Neil J. McKenna (nmckenna@bcm.edu)

Public transcriptomic assets in the nuclear receptor (NR) signaling field hold considerable collective potential for exposing underappreciated aspects of NR regulation of gene expression. This potential is undermined however by a series of enduring informatic pain points that retard the routine re-use of these datasets. Here we describe a coordinated biocuration and web development approach to redress this situation that is closely aligned with ideals articulated in the FAIR (findable, accessible, interoperable, re-usable) principles on data stewardship. To improve findability, biocurators engage authors of studies in collaborating journals to secure datasets for deposition in public archives. Annotated derivatives of the archived datasets are assigned digital object identifiers and regulatory molecule identifiers that support persistent linkages between datasets and their associated research articles, integration in relevant records in gene and small molecule knowledgebases, and indexing by dataset search engines. To enhance their accessibility and interoperability, datasets are visualizable in responsively designed web pages, retrievable in machine-readable spreadsheets, or through an application programming interface. Re-use of the datasets is supported by their interrogation as a universe of data points through the Transcriptome search engine, highlighting transcriptional intersections between NR signaling pathways, physiological processes and disease states. We illustrate the value of our approach in connecting disparate research communities using a use case of persistent interoperability between the Nuclear Receptor Signaling Atlas and the Pharmacogenomics Knowledgebase. Our FAIR-aligned model demonstrates the enduring value of discovery-scale datasets that accrues from their systematic compilation, biocuration and distribution across the digital biomedical research enterprise.

Keywords: Transcriptomics; datasets; findability; accessibility; interoperability; re-use

Introduction

Signal transduction by members of the nuclear receptor (NR) superfamily of transcription factors encompasses interactions with small molecule ligands and coregulators that control cell- and tissue-specific transcriptomes in a wide variety of developmental and physiological contexts (McKenna and O'Malley, 2002, Mangelsdorf et al., 1995). Basic and clinical researchers in this field frequently pose many fundamental questions that directly relate to unappreciated or undeveloped aspects of NR signaling biology. What NR pathways regulate my gene of interest? What genes are most consistently regulated by a given NR pathway, and how do these targets differ between different tissues? What NR pathways impact my cellular process of interest in different tissues? Although the NR signaling field has generated a large number of expression

profiling datasets involving perturbations of NR signaling pathways, numerous factors combine to complicate re-use of these datasets to answer these and other biological questions. Datasets are not consistently archived (Ochsner et al., 2008) and those that are archived in repositories such as the US National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) (Barrett et al., 2009) and European Bioinformatics Institute (EBI) ArrayExpress (Kolesnikov et al., 2015) are frequently under-annotated and poorly exposed for discovery by researchers. The informatic isolation that results limits the reuse of these datasets as a biological continuum, preventing the routine generation or validation of research hypotheses by the NR signaling community. In many cases, researchers must invest time, effort and money in designing and carrying out an experiment for which relevant data points have already been published. Apart from the wasted effort, the current period of financial austerity in research funding makes a strong case for the development of tools that will provide for more effective and efficient use of already existing, but currently peripheral, data points.

The recently elaborated FAIR (findable, accessible, interoperable, re-usable) principles on data stewardship (Wilkinson et al., 2016) articulate an ideal biomedical research enterprise in which the major research roles, such as bench researchers, data repositories and publishers, interact in a way that promotes improved efficiency and re-use of research assets. The Nuclear Receptor Signaling Atlas (NURSA) was formed in 2002 with a mandate of improving the discoverability, accessibility and re-use of datasets for the NR signaling community and related research constituencies (Becnel et al., 2015). As a domain-specific biocuration and dataset repository, the FAIR principles have a high degree of relevance to our mission to enable researchers in this field to make greater use of the abundant 'omics-scale datasets that this field has generated. This paper describes our implementation of the FAIR principles as they relate to biomedical data stewardship, demonstrating the problems we encounter, and our FAIR-aligned solutions to those obstacles. We illustrate our approach with reference to specific examples and use cases, including an interoperability collaboration between NURSA and the Pharmacogenomics Knowledgebase (PharmGKB) that connects research communities with common interests in transcriptomic datasets relevant to nuclear receptor signaling.

Methodology

The methodologies supporting the data stewardship approaches described in this paper are discussed in detail in a recent publication (Darlington et al., 2016). Briefly, landing page fields for the biocurated datasets were aligned with recommendations of the Joint Declaration on Data Citation Principles (Martone, 2014) and DataCite (DataCite, 2015) organizations. Integration of the third party reference manager-integrated dataset citation widget was based upon the RIS file format standard. Automated integration between NURSA and the Pharmacogenomics Knowledgebase (PharmGKB) resource was achieved through the development of a RESTful application programming interface (API) (Darlington et al., 2016), permitting gene calls using Entrez Gene ID or approved gene symbol or, for small molecules, PubChem ID. Deposition of dataset metadata with dataset indexing services and search engines was supported by use of the Open Archives Initiative Protocol for Metadata Harvesting standard (Archives, 2016).

Results

At the inception of NURSA's third funding cycle in 2012, we embarked upon an extensive re-appraisal of NURSA website content and user interface. To tailor our modifications in scientific scope and web design principles to the needs of our end users, we carried out a survey of the NURSA user community. Although end users expressed satisfaction with the site overall, they requested an increased emphasis on 'omics dataset integration and analysis tooling. Based upon this response, and given that they represent the most abundant 'omics modality in the field of nuclear receptor signaling we set out on a systematic effort to enhance the re-use of transcriptomic datasets in the field. Here we highlight key aspects of our biocuration and web development approach as they relate to each of the four components of FAIR in turn, discussing for each the problems confronted and the solutions adopted.

1. Findability

1.A. The issue

Multiple points of failure in the existing model of academic scientific funding, research and publishing have contributed to poor findability of transcriptomic datasets. Reluctant to impose additional administrative burdens on authors, many publishers default to a *laissez-faire* disposition towards dataset archiving, such that only journals with the highest submission volume and rejection rates – the Nature Publishing Group

and Cell Press families, for example – can afford to mandate and actively enforce deposition of discovery scale datasets as a condition of publication. To compound matters, journal editorial staff lack the technical expertise to oversee dataset deposition, and manuscript reviewers are primarily preoccupied with the scientific content of the manuscript rather than the deposition status of the associated dataset. As a consequence, for the vast majority of journals we have encountered during our biocuration activities, deposition of a high quality, well annotated dataset, and inclusion of its accession number in the final published version of the article, are entirely at the author's discretion. Due to a lack of academic incentivization and unwillingness or inability on the part of many principal investigators to commit time and resources, many 'omics scale datasets are absent from public repositories (Ochsner et al., 2008, Witwer, 2013). Among the more ironic manifestations of this situation are those articles supported by unarchived transcriptomic datasets, but that validate their own findings using publically deposited transcriptomic datasets (Huber-Keener et al., 2012). Although the NIH Genomic Data Sharing Policy (Health, 2016) goes some way to addressing the problem of failure of researchers to archive datasets, it can mandate only that accession numbers be provided at the time of grant renewal (typically once every five years), rather than at the time of publication of the associated article.

The lack of active engagement on the part of many publishers with respect to dataset deposition has negative repercussions for the integration between GEO, ArrayExpress and NCBI's public bibliographic database, PubMed. Given that many GEO accession numbers are absent from final accepted versions of articles, and consequently unavailable to PubMed curators, integration of GEO records with the corresponding PubMed record is patchy and inconsistent (Neveol et al., 2012). As a result, GEO identifiers are inconsistently annotated in the full indexed PubMed record for a given article and, to compound matters, neither GEO nor MEDLINE support searching using each other's unique identifiers (GSE accession numbers and PubMed IDs (PMIDs), respectively). To further complicate the situation, PMID is not a required field when the datasets are submitted to GEO, neither are authors required by journals to add PMIDs to their GEO records upon manuscript acceptance. A direct consequence of this situation is that a considerable percentage of the GEO records encountered during our biocuration activities, which we refer to as "orphaned" datasets (**Figure 1**),

NCBI > GEO > Accession Display [?](#) Not logged in | [Login](#) [?](#)

Scope: Format: Amount: GEO accession:

Series GSE56778 [Query DataSets for GSE56778](#)

Status	Public on Aug 01, 2014
Title	Expression data from Panc1 pancreatic epithelial cells
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	Krüppel-like factors (KLFs) are a group of master regulators of gene expression conserved from flies to human. However, scant information is available on either the mechanisms or functional impact of the coupling of KLF proteins to chromatin remodeling machines, a deterministic step in transcriptional regulation. In the current study, we use genome-wide analyses of chromatin immunoprecipitation (ChIP-on-Chip) and Affymetrix-based expression profiling to gain insight into how the KLF11, a human transcription factor involved in tumor suppression and metabolic diseases, works by coupling to three co-factor groups: the Sin3-histone deacetylase system, WD40-domain containing proteins, and the HP1-histone methyltransferase system. We utilized genome-wide expression analysis of wild type KLF11 and three mutants that disrupt KLF11-chromatin machinery interactions to examine the relationship of the transcription factor and chromatin systems in the regulation of gene networks.
Overall design	Panc1 epithelial cells were plated at a density of 1x10 ⁶ cells/100mm dish and transduced with empty vector, KLF11, KLF11-A347S, KLF11-486, or KLF11-EAPP adenovirus at an MOI of 150. RNA was prepared as previously described from pooled biological triplicates.
Contributor(s)	Raul U, Ezequiel C
Citation missing	Has this study been published? Please login to update or notify GEO .

Figure 1: "Orphaned" dataset in GEO. GEO relies upon users for retrospective annotation of records that are not mapped to a corresponding published study.

are not mapped to PMIDs. As a result, updating GEO records with their corresponding PMIDs has become a routine component of our biocuration standard operating procedure.

A final problem with regard to findability of transcriptomic datasets in the current academic research model relates to their visibility in searches. Whether in archived CEL files or supplementary PDFs, transcriptomic data points are largely opaque to popular search engines such as Google. Such search engines perform excellently in text retrieval, but are comparatively inadequate in the retrieval of experimental gene regulation data points, generating large amounts of noisy search results that require considerable parsing on the part of the user with little guarantee of gleaning meaningful information.

1.B. Our solution

The cornerstone of our strategy to improve data findability is the creation of secondary versions of the primary archived datasets and the establishment of persistent linkages between their landing pages and key relevant nodes in the digital biomedical research ecosystem (Darlington et al., 2016). Given its well-established infrastructure, broad community adoption and familiarity to researchers, the DOI standard was selected to support the creation of these linkages. NURSA DOIs support bidirectional links between NURSA dataset landing pages and publisher partner articles (**Figure 2A**) (Darlington et al., 2016) and, since few publishers support retroactive journal-database record linkages, PubMed records (**Figure 2B**). The cultivation of alliances with strong publishing brands such as Elsevier and Public Library of Science has the added benefit of strengthening NURSA's own brand in the community and has resulted in their facilitating proactive contact of our biocuration team with authors of accepted manuscripts. This direct contact with authors addresses the largest obstacle to dataset biocuration archiving – lack of access to the original authors of the datasets – and results in more rapid deposition of more complete, comprehensively annotated datasets. Findability is also enhanced by providing for indexing of dataset metadata by dataset search engines such as bioCADDIE DataMed (**Figure 2C**) and Thomson Reuters Web of Science Data Citation Index (Darlington et al., 2016). To increase the visibility of NURSA transcriptomic data assets to researchers in disparate communities with no explicit connection to NR signaling pathways, we have established access points to the NURSA datasets and/or the Transcriptomine search engine from databases that curate information on small molecules (ChEBI, **Figure 2D**) and their genomic targets (NCBI Entrez Gene LinkOut). The benefit of NURSA's more intensive biocuration is demonstrated in **Figure 2D**, which shows Gene Expression studies mapped to the ChEBI record for the GR agonist dexamethasone. Not only are NURSA records more numerous than ArrayExpress records (28 vs 21), they are more accurately mapped – note the number of *Arabidopsis* studies inappropriately mapped to dexamethasone. Depositors often assign names for primary dataset depositions using the title of the associated article, which frequently gives no indication as to the nature or design of the underlying dataset (Swindell et al., 2014). To mitigate against this, the NURSA biocuration team assigns names to curated datasets that unambiguously declares the essential regulatory parameters and the design of the dataset: compare the name for the GEO dataset cited above and that of its NURSA-curated derivative (Darlington et al., 2016).

Journal research articles and their associated reference lists represent rich environments for the discovery of unfamiliar science. By providing for citation of their datasets as scholarly works alongside citations of research articles in article reference lists, repositories can tap into this infrastructure to greatly enhance the findability of their dataset records. Indeed, the importance of citing datasets in articles and research proposals to their status as first class research objects is widely accepted (Borgman, 2011, Goodman et al., 2012, Margolis et al., 2014, Martone, 2014). Accordingly, we have made provision for one-click downloading of a JDDCP/FAIR-compliant dataset metadata record in a format compatible with the four major reference managers of choice. In addition to supporting their discovery by article readers, citation of datasets in this way ensures accreditation to original authors, which in turn incentivizes their deposition of future datasets to complete the cycle of discovery and re-use (Darlington et al., 2016).

2. Accessibility

2.A. The issue

The quality of archived datasets and their associated metadata is a major determinant of their accessibility. The often superficial or non-existent peer-review of datasets during the manuscript review process however, and the lack of rigorous oversight of dataset deposition, give rise to recurring problems that at best complicate, and at worst completely prevent, their downstream re-use. Problems at the dataset level that we have encountered during our biocuration activities include incomplete information on data

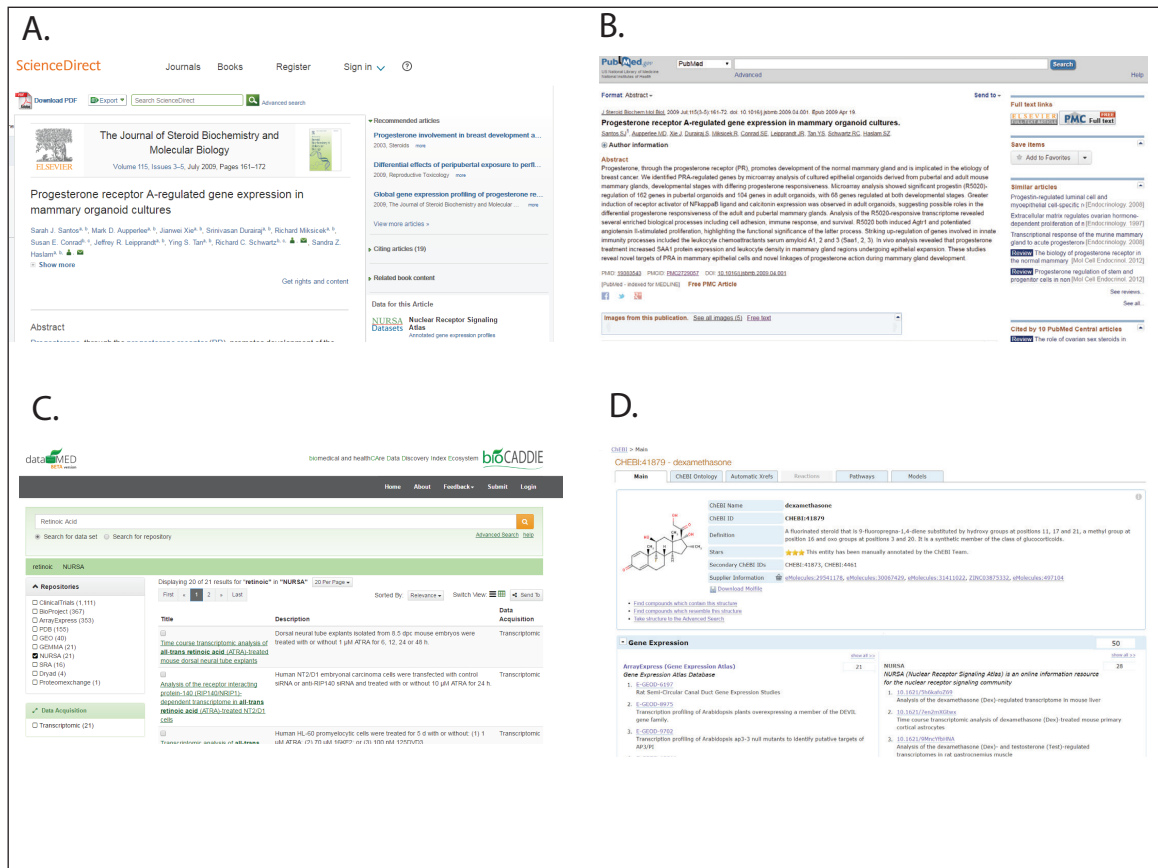


Figure 2: Findability: enhanced exposure of NURSA datasets across the biomedical research enterprise. Automated processed embed links to NURSA datasets from (A) journal articles of publisher collaborators who support journal-database linkage, (B) PubMed records, (C) dataset search engines such as DataMed and (D) records in small molecule knowledgebases such as ChEBI.

normalization and absence of replicates (Notas et al., 2010, Sharova et al., 2007) and failure to deposit all the required data files (Boyer et al., 2005). In some more extreme cases (Ajj et al., 2013), technical details on the transcriptomic datasets are entirely absent from the materials and methods and supplementary material sections, even though they contribute in large part to the entire basis for a study, and are discussed in the results. In other cases, articles do not contain any fold changes from the arrays and as a consequence our curators have been unable to validate the processed data files against author-reported data points (Liu et al., 2015). Even many properly archived datasets are only nominally accessible to researchers lacking informatics expertise: many of the file types involved are unfamiliar to most bench researchers, and the computational investment involved in generating relative abundances and associated measures of statistical significance across multiple datasets is prohibitive. Although the GEO2R feature in GEO alleviates some of this burden (NCBI, 2016), it is limited to a single dataset, and assumes that the user is sufficiently conversant with the experimental design to set up meaningful experimental contrasts.

Transcriptomic and ChIP-Seq datasets reflect highly specific spatiotemporal contexts and accordingly, if they are to be meaningfully interpreted and compared with each other, they must be associated with a substantial amount of detailed, accurately mapped metadata. Unfortunately, the lack of familiarity with metadata standards of many of the laboratory personnel tasked with their deposition, and the limited domain expertise on the part of primary data repository curators, has given rise in many cases to cursory and uneven annotation. Rather than being mapped to community-endorsed universal identifiers for regulatory small molecules and genes or biosamples, these critical experimental parameters are often encoded in a cryptic and ad hoc shorthand. The consequences of this is that related datasets are not organized into biologically meaningful categories in public repository user interfaces, requiring users to identify and retrieve “like” datasets using free text queries with often noisy or unpredictable search results. An additional consequence

is that our own biocurators are required to expend considerable time and effort in retrospectively parsing metadata from the related publications.

2.B. Our solution

2.B.1. Biocuration

As the number of datasets in our resource grows, so the need increases to organize biologically related datasets for routine retrieval by specific research constituencies: some researchers are interested in all datasets related to a specific organ, for example, whereas others might be focused on datasets related to a given signaling pathway across all organs. The two primary descriptors that define a transcriptomic dataset are the biosample from which the RNA was generated, and the regulatory molecule that defines the primary experimental variable (Becnel et al., 2016). To enhance the accessibility of datasets to users of our resource, our biocuration approach incorporates a step that maps all datasets to hierarchical “catch-all” terms that group datasets related by pathway and biosource. In this way, users are not required to run multiple iterative queries to retrieve all datasets related to a given signaling pathway, or physiological system and organ (Becnel et al., 2016). Instead, they can filter datasets in the directory according to their pathway or biosample of interest using intuitive drop-down menus (**Figure 3**). An additional benefit of this approach is that when researchers arrive at a dataset landing page from an external site, they can identify datasets related by regulatory molecule or by biosource (Darlington et al., 2016).

2.B.2. Quality Control

Once an accession number is issued by a primary data repository and the associated paper is published, opportunities to correct the data record are greatly limited. There is little enthusiasm either on the part of authors to correct problems with archived datasets, or on the part of publishers of articles associated with such datasets to mandate such retrospective action from the authors. Given that the time and effort required on the part of our biocuration group to troubleshoot flawed depositions would exclude timely curation of properly-archived datasets, our role with respect to depositions with irretrievable deficits is necessarily limited to flagging them for exclusion from our database. One notable success story, and a validation of our model of retrospective biocuration of primary datasets, is the instance of a mismatch between relative abundance values generated during our biocuration and those reported by the authors in the article, which resulted in a correction to the published article (Mamrosh et al., 2015).

2.B.3. Web Development

Our web development approach places a strong emphasis on usability factors and of the amenability of the datasets to re-use by researchers: put another way, users are unlikely to make use of a resource that is not intuitively accessible. When a researcher lands on a dataset page, experimental contrasts and their associated transcript relative abundances and measures of statistical significance are all pre-defined and accessible using a drop-down menu (**Figure 3B**). Detailed metadata are available at both the experiment and dataset levels, and detailed regulation reports on individual transcripts across the entire universe of data points can be visualized in visually engaging scatterplots with the click of a mouse (Becnel et al., 2016). Given the widespread adoption of mobile devices, in particular cellular handsets, to access Internet content, we adhere to responsive principles in web design, ensuring as far as possible that the accessibility requirements of mobile devices users are addressed. Finally, data and metadata can be downloaded in spreadsheet format for downstream analysis in a user’s third party software of choice.

3. Interoperability

3.A. The issue

The same biocuration deficits that beset findability of datasets in GEO and ArrayExpress impacts their automated interoperability with external entities wishing to leverage the underlying data points to add value to their own resource. The problems posed are exemplified by the fact that despite being components of the same organization, biologically meaningful connections between GEO and other NCBI entities are yet to be established – the PubChem record for 17 β -estradiol, for example, does not contain a listing of GEO datasets in which this small molecule is an experimental variable. This lack of integration represents a substantial missed opportunity to leverage transcriptomic datasets as knowledge mines to connect disparate research communities.

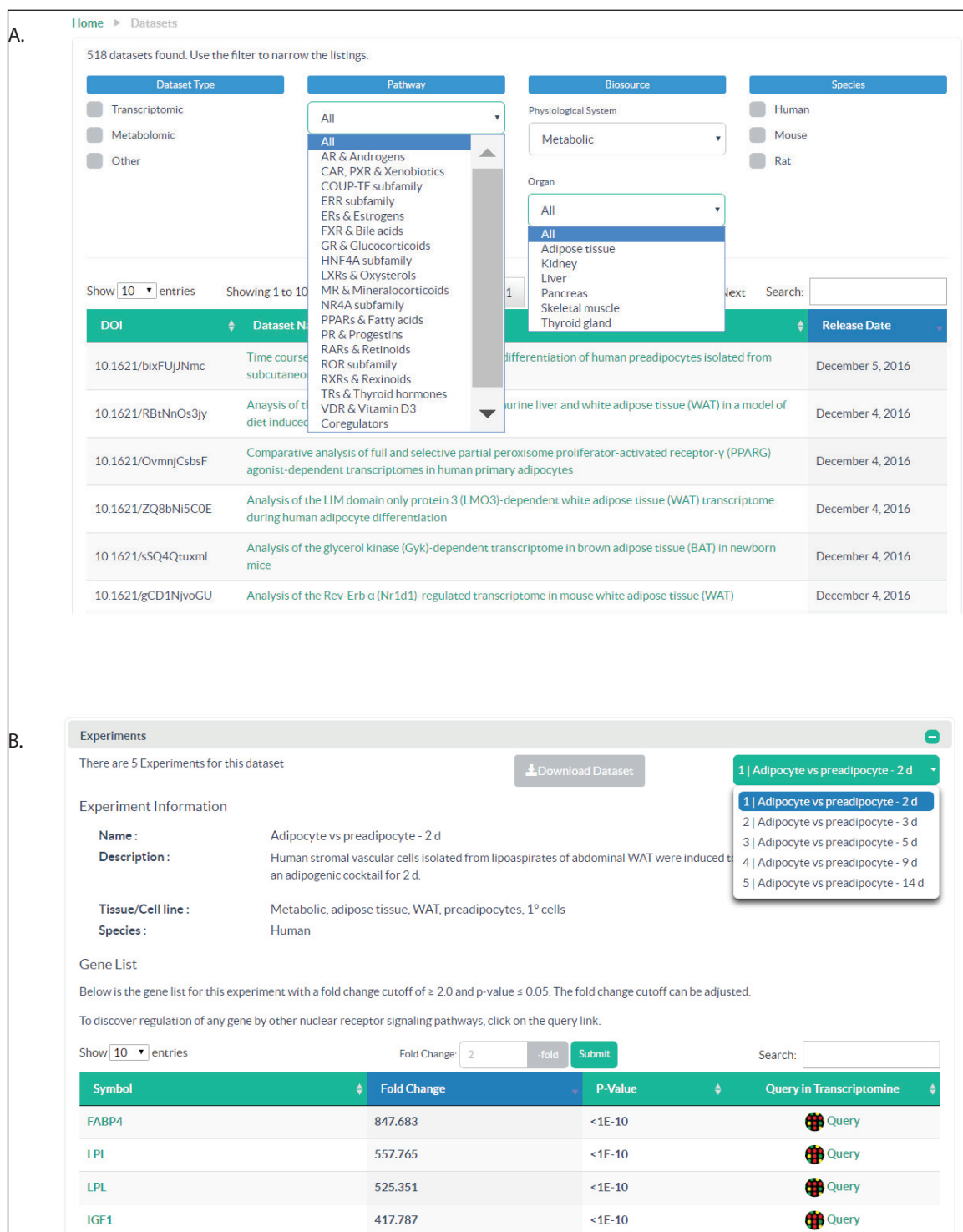


Figure 3: Accessibility: web development usability principles support dataset filtering and knowledge extraction. (A) The dataset directory is filterable by intuitive drop-down menus for nuclear receptor signaling pathway or biosample physiological system class and organ subclass. **(B)** Drop down menus on dataset landing pages provide for toggling between gene lists to identify significant data points in each experimental contrast.

3.B. Our solution

To support interoperability with external resources, data points across all NURSA datasets are interoperably exposed via RESTful web services for retrieval using controlled vocabularies and regulatory molecule unique identifiers (Darlington et al., 2016). An example of the use of APIs to connect NURSA with resources

curating content complementary to its own is an ongoing BD2K-funded interoperability project between NURSA and the Pharmacogenomics KnowledgeBase (PharmGKB, www.pharmgkb.org). PharmGKB is a comprehensive compendium documenting the effect of variations in the sequences of human genes on the response to drugs and is widely used by both clinicians and basic researchers (Thorn et al., 2013). Given that many drugs are small molecule regulators of NR function, and NURSA curates datasets in which these molecules are regulatory variables, the establishment of automated, programmatic connections between NURSA and PharmGKB would enhance the other's website content and the research experience of their respective user bases. The NURSA and PharmGKB biocuration standard operating procedures involve mapping records to universal molecule identifiers such as PubChem ID and Entrez Gene IDs, which are exposed through APIs that both groups have developed. Accordingly, API-supported interoperability was established between the two sites such that PharmGKB records for drugs that are NR ligands linked to NURSA datasets in which these molecules were perturbants (**Figure 4A**) and, reciprocally, available PharmGKB-curated drug metabolism pathway and pharmacogenomic drug-gene variant interaction modules were displayed in appropriate NURSA ligand Molecule Pages (**Figure 4B**). The net result of this interoperability is to expose essential information on aspects of drug action that might not otherwise be readily accessible to, respectively, the pharmacogenomic and nuclear receptor signaling research communities.

4. Reusability

4.A. The issue

The format of the journal research article, predicated upon lengthy exposition, interpretation and attribution, has remained substantially unchanged since its introduction in the mid-19th century. The structural and dimensional constraints of the research article dictate however that it is neither practical nor feasible to convey all the findings from global expression studies in detail. As a result, authors of articles containing 'omics-scale datasets typically validate and interpret only those data points most relevant to their experimental hypothesis, consigning hundreds or thousands of potentially useful expression fold changes to unwieldy, patchily annotated spreadsheets or PDFs. Due to the problems with dataset findability and accessibility outlined above, pointing to primary dataset records from journal articles does little to advance the potential of these datasets for re-use.

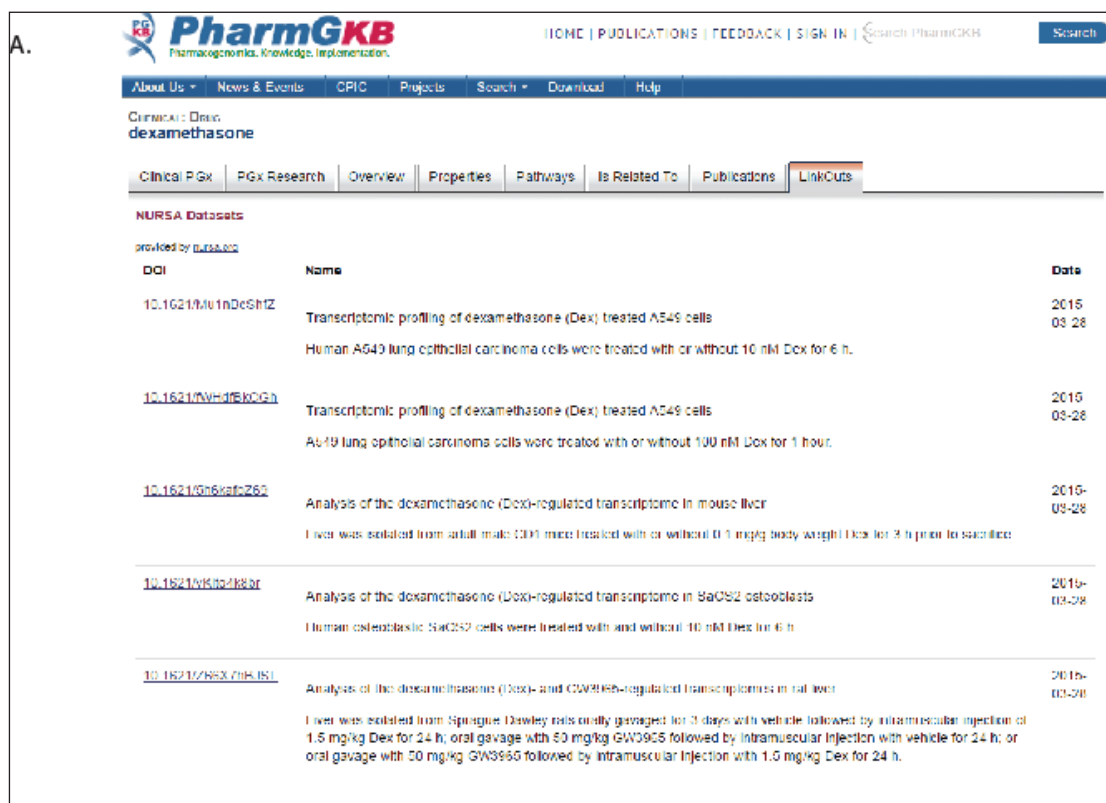
4.B. Our solution

As described in the Findability section above, NURSA places a strong emphasis on cultivating relationships with publishers of scientific research journals. In addition to facilitating our biocuration efforts, the development of these relationships gives NURSA the opportunity to work with publisher development teams to embed links from research articles to their respective dataset landing pages on the NURSA website. By providing journal readers with one-click access to a universe of contextual data points that enrich and add value to the original research article, NURSA datasets make a significant contribution to the re-use of these data points for the discovery of unfamiliar or unappreciated biology. An example of the striking visual impact of the Transcriptome Regulation Report is shown in **Figure 5**, which summarizes the NR signaling pathways impacting expression of the gene encoding fatty acid binding protein 4 (*FABP4*), a lipid transport protein present primarily in adipose tissue and macrophages. Firstly, numerous animal and cell model data points place expression of the gene in context, redundantly documenting induction of *FABP4* in adipogenesis (Ochsner et al., 2009, Christian et al., 2005) and the correlation between its expression and fat whitening (Wu et al., 2012). Next, data points from numerous datasets illustrate the close regulation of *FABP4* expression by the PPARA and PPARG signaling pathways, two very well characterized adipose regulatory paradigms (Szatmari et al., 2007, Zacharewski et al., 2013, Ochsner et al., 2009, Finck et al., 2005). Other signaling modalities that are convincingly conveyed by the *FABP4* Regulation Report are: repression by the AR/androgen pathway (Lin et al., 2009, Kazmin et al., 2006), consistent with androgenic suppression of adipogenesis (Chazenbalk et al., 2013, Singh et al., 2006); and induction by the GR/glucocorticoid pathway (Hoffman et al., 2005, James et al., 2007, lab et al., 2010), reflective of the widespread use of the GR agonist dexamethasone as an adipogenic stimulant in cultured cells (Scott et al., 2011).

Discussion

The existing NURSA infrastructure, developed over 14 years of funding with support from six different NIH institutes, has to date has focused on transcriptomics of nuclear receptor (NR) signaling pathways (Becnel et al., 2015). NURSA has developed an international userbase with a nearly 3,000 person-strong

A.



B.

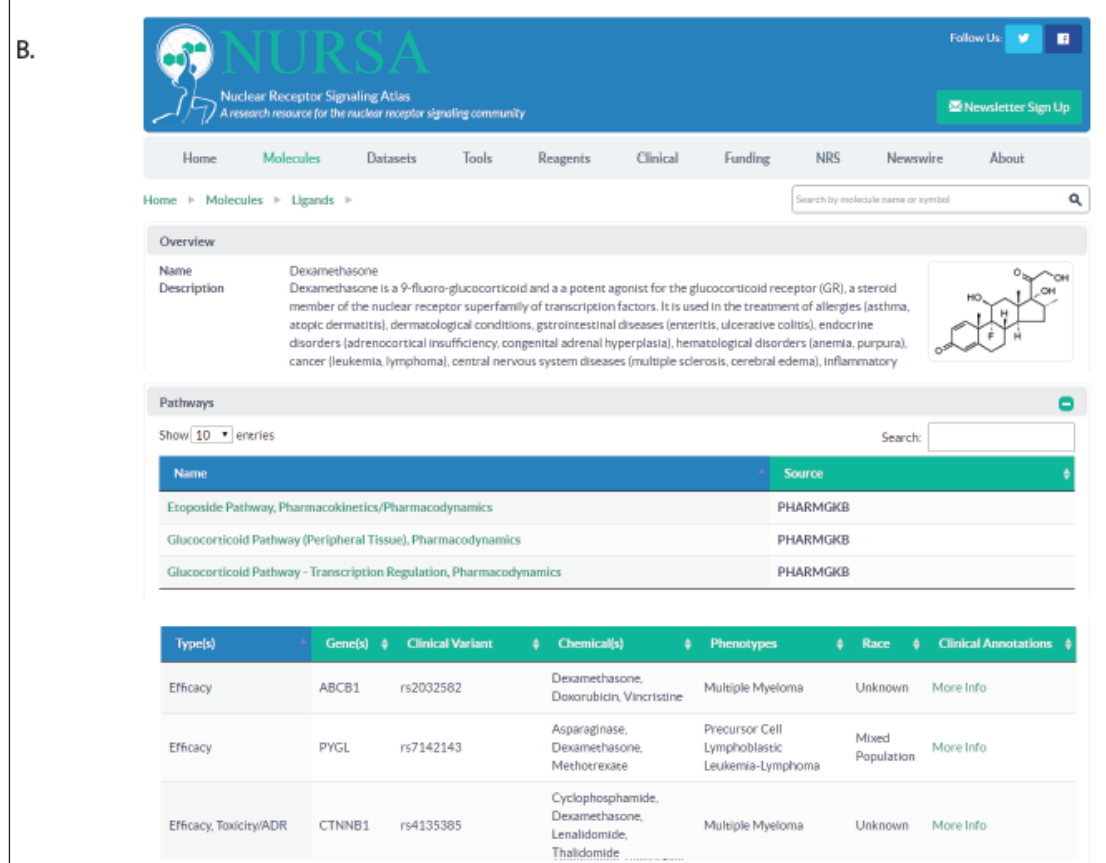


Table 1: NURSA Datasets for Dexamethasone

DOI	Name	Date
10.1021/MU1nDeSHZ	Transcriptomic profiling of dexamethasone (Dex) treated A549 cells Human A549 lung epithelial carcinoma cells were treated with or without 10 nM Dex for 6 h.	2015-03-28
10.1021/NV1nDFK0Gh	Transcriptomic profiling of dexamethasone (Dex) treated A549 cells A549 lung epithelial carcinoma cells were treated with or without 100 nM Dex for 1 hour.	2015-03-28
10.1021/Gn6kafz6Z	Analysis of the dexamethasone (Dex)-regulated transcriptome in mouse liver Liver was isolated from adult male C57BL/6 mice treated with or without 0.1 mg/kg body weight Dex for 3 h prior to sacrifice.	2015-09-28
10.1021/NK6kafz6Z	Analysis of the dexamethasone (Dex)-regulated transcriptome in SaOS2 osteoblasts Human osteoblastic SaOS2 cells were treated with and without 10 nM Dex for 6 h.	2015-03-28
10.1021/NK6kafz6Z	Analysis of the dexamethasone (Dex)- and GW3965-regulated transcriptomes in rat liver Liver was isolated from Sprague-Dawley rats orally gavaged for 3 days with vehicle followed by intramuscular injection of 1.5 mg/kg Dex for 24 h; oral gavage with 50 mg/kg GW3965 followed by intramuscular injection with vehicle for 24 h; or oral gavage with 50 mg/kg GW3965 followed by intramuscular injection with 1.5 mg/kg Dex for 24 h.	2015-03-28

Table 2: Drug-Variant Interactions

Type(s)	Gene(s)	Clinical Variant	Chemical(s)	Phenotypes	Race	Clinical Annotations
Efficacy	ABCB1	rs2032582	Dexamethasone, Doxorubicin, Vincristine	Multiple Myeloma	Unknown	More Info
Efficacy	PYGL	rs7142143	Asparaginase, Dexamethasone, Methotrexate	Precursor Cell Lymphoblastic Leukemia-Lymphoma	Mixed Population	More Info
Efficacy, Toxicity/ADR	CTNIB1	rs4135385	Cyclophosphamide, Dexamethasone, Lenalidomide, Thalidomide	Multiple Myeloma	Unknown	More Info

Figure 4: Interoperability and re-usability: persistent API-supported, DOI-driven connections between NURSA and the Pharmacogenomics Knowledgebase connect disparate research communities. (A) PharmGKB drug reports list NURSA datasets in which these drugs are regulatory molecules. (B) NURSA Molecules Pages for ligands that are prescription drugs display links to PharmGKB-curated pathway and pharmacogenomic drug-variant interactions.

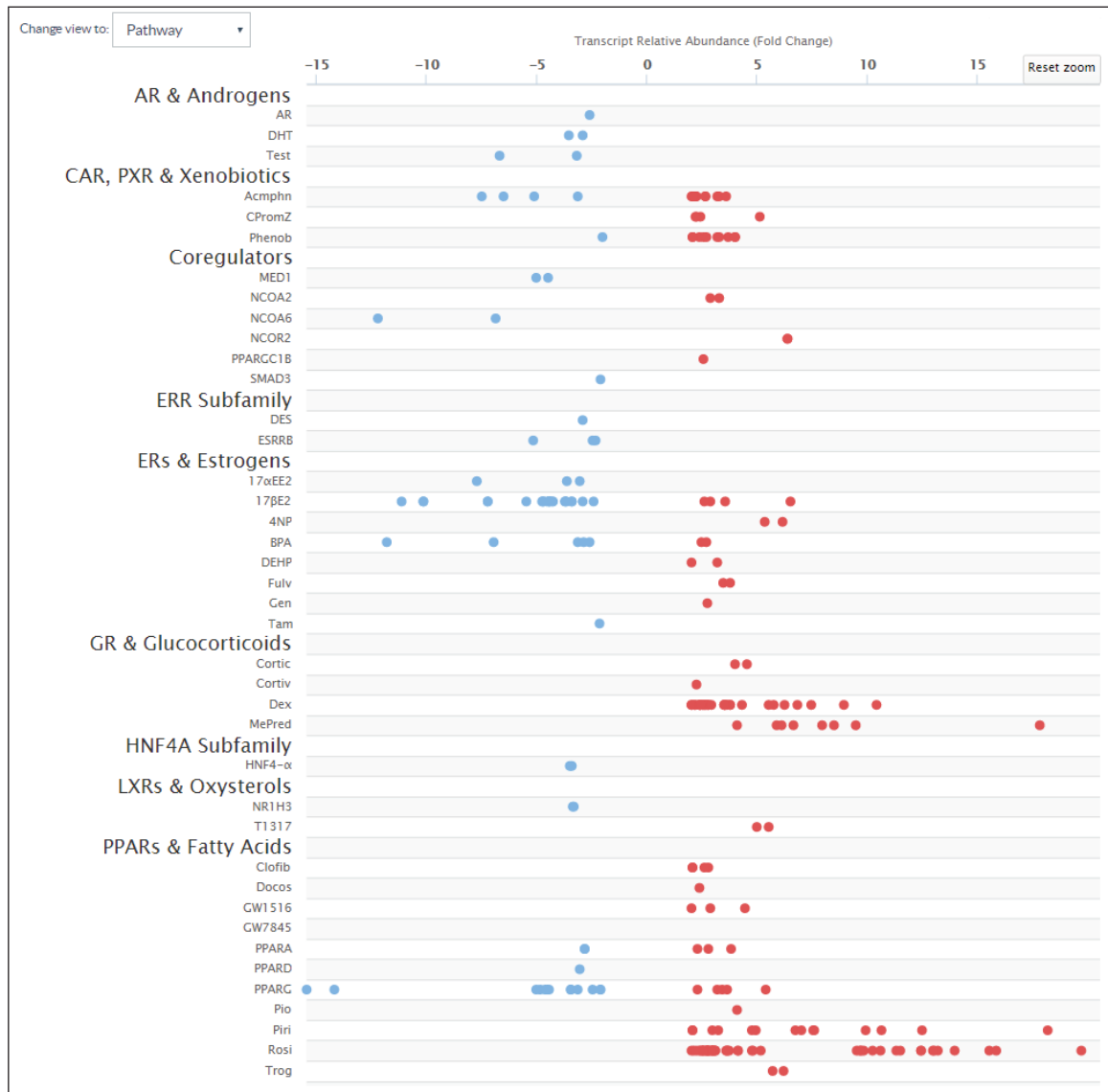


Figure 5: Re-usability: biocuration mapping provides for intuitive visualization of both known and previously uncharacterized gene regulatory paradigms. The grouping of experiments into hierarchical pathway categories in the Pathway View Transcriptome *FABP4* Regulation Report (Pathway View) makes immediately evident the key signaling pathways that modulate expression levels of this gene.

e-mail newsletter following. It is a J2EE web application with a D3.JS-based interactive graphics user interface and a platform aware design, as well as controlled terminologies for semantic interoperability and RESTful web services for data exchange and syntactic interoperability (Becnel et al., 2015). Our resource combines a number of innovative strategies and features that both distinguish it from, and complement existing pathway resources such as Reactome (Fabregat et al., 2016) and Pathway Commons (Cerami et al., 2011). These include publisher alliances to link articles to biocurated datasets (Darlington et al., 2016), novel approaches to biocuration and mining of ‘omics datasets (Becnel et al., 2017), and providing for routine citation of datasets as first-class research objects (Darlington et al., 2016).

Public transcriptomic archives such as GEO and ArrayExpress are outstanding resources for long term deposition of raw files for transcriptomic datasets, and are deserving of continued financial support. Unfortunately, datasets in these archives have not been effectively integrated with an expanding volume of scientific research output, a state of affairs that substantially undermines their collective informatic potential. Such deficits relate primarily to the fact that these repositories operate under a passive model with respect to public dataset deposition. As such, the onus is on investigators to approach the repository, rather than the repository actively seeking out publically-funded datasets associated with accepted manuscripts. In response to this we

have combined proactive engagement of authors and publishers by our biocuration team with data analysis tool development to place data points from transcriptomic studies relevant to NR signaling pathways at the fingertips of research biologists (Darlington et al., 2016) (Becnel et al., 2016). Future efforts will extend our model to other 'omics modalities and pathway paradigms to provide basic and translational researchers with broader insights to the relationship between cellular signaling and human disease.

Although such direct feedback from the community is encouraging, our model does have a number of limitations. Principle among these is the uncertain extent to which the time and effort invested in biocuration is repaid in the form of added value for data re-users. Given the popularity of research resources and tools that do not incorporate a biocuration element, such as Galaxy (Afgan et al., 2016), it is far from clear that the value of biocuration is universally appreciated. Another problematic aspect of our approach is that by its very nature, interoperability between two resources establishes a dependence of each upon the other for full functionality of their respective resources. As more and more nodes are connected interoperably via web services, so the need for each resource to maintain and update their web services and documentation grows proportionately, which in turn adds to the workload for software architects and web developers. A third limitation is one that applies to the field of biocuration more generally, and that relates to the existing prevailing model of academic reward and recognition. Although their work is just as intellectually demanding as bench research, biocurators – at least in the field of cell signaling – do not enjoy the same potential for personal recognition and reward that is afforded hypothesis-driven bench scientists. The change in biomedical research culture heralded by the BD2K initiative (Margolis et al., 2014) suggests however that on the part of NIH at least, there is a commitment to making biocuration attractive as a viable career option for a broader population of trained scientists.

The FAIR Principles were formally published in 2016 as the synthesis of extensive efforts and discussions across the data science and scholarly research stewardship communities (Wilkinson et al., 2016). Our intent in this paper was to convey the experiences of a biocuration and web development group in a specific area of research – nuclear receptor signaling – that has incorporated the FAIR principles into its standard operating procedures. This is not to imply that issues around accessibility and interoperability are limited to basic biology: poor deposition rates to clinical trials repositories such as ClinicalTrials.gov have also been reported, for example (Piller, 2015). This issue is compounded by partial reporting and bias toward the selective publication of positive findings (Rifai et al., 2014), despite the intrinsic value of data from failed trials in helping avoid duplication of expensive human and co-clinical trials. Metadata and data exchange standards exist from the Clinical Data Interchange Standards Consortium (CDISC), whose tabulation and analysis datasets standard structures are now required for regulated clinical trial submissions to the Food and Drug Administration and Japan's Pharmaceuticals and Medical Devices Agency (FDA, 2014). These initiatives notwithstanding however, no broad requirement for data standardization for non-regulated clinical protocols currently exists.

It can be quite reasonably pointed out that we have not provided here any objective, quantitative metrics on the efficacy of our approach. Given the fact that our FAIR biocuration strategy has only recently reached maturity however, we submit that any current analysis of the efficacy or impact of our efforts, or the FAIR principles in general, would be previous. That is not to diminish in any way the importance of such an appraisal taking place in the future however, and a reliable perspective on the true value of the FAIR principles can emerge only from a series of comprehensive and objective retrospective appraisals of their impact across a variety of scholarly fields. A number of models have been proposed for such evaluations, most notably the case study-based approach espoused by Darke, (Darke et al., 1998) Yin (Yin, 2013) and colleagues. The experiences of biocuration groups such as our own, as well as of data re-users in the research community, will be of considerable value in furnishing data for such case studies, so that the lasting impact of the FAIR principles can be accurately – and fairly – gauged.

Acknowledgements

We thank Mike Wise for assistance preparing the manuscript. We acknowledge the assistance of key personnel at the Pharmacogenomics Knowledgebase (5R24GM061374) in Stanford University, namely, Dr Teri Klein, Dr Michelle Whirl-Carrillo and Ryan Whaley. This work was supported by National Institutes of Health Big Data To Knowledge (BD2K) program supplements (DK097748-S1 and DK097748-S3) and a Cancer Center Support Grant to BCM (CA125123). The NURSA Consortium is funded by an award from NIDDK and NICHD (DK097748). We thank the previous & current NURSA Program Officers, Drs. Ron Margolis and Corinne Silva (National Institute of Diabetes Digestive and Kidney Diseases, NIDDK) and Dr. Koji Yoshinaga (Eunice Kennedy Shriver National Institute of Child Health and Development, NICHD).

Competing Interests

The authors have no competing interests to declare.

Authors Information

Lauren B. Becnel is currently at Clinical Data Interchange Standards Consortium, 401 W. 15th Street Suite 800, Austin, TX, 78701, USA, Austin, TX.

Alexey Naumov is currently at Apple, Inc., Sunnyvale, CA.

Scott A. Ochsner, Yolanda F. Darlington and Lauren B. Becnel had made equal contributions to this work.

References

- Afgan, E, Baker, D, Van Den Beek, M, Blankenberg, D, Bouvier, D, Cech, M, Chilton, J, Clements, D, Coraor, N, Eberhard, C, Gruning, B, Guerler, A, Hillman-Jackson, J, Von Kuster, G, Rasche, E, Soranzo, N, Turaga, N, Taylor, J, Nekrutenko, A and Goecks, J** 2016 The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res*, 44(W1): W3–W10. DOI: <https://doi.org/10.1093/nar/gkw343>
- Ajj, H, Chesnel, A, Pinel, S, Plenat, F, Flament, S and Dumond, H** 2013 An alkylphenol mix promotes seminoma derived cell proliferation through an ERalpha36-mediated mechanism. *PLoS One*, 8(4): e61758. DOI: <https://doi.org/10.1371/journal.pone.0061758>
- Barrett, T, Troup, D B, Wilhite, S E, Ledoux, P, Rudnev, D, Evangelista, C, Kim, I F, Soboleva, A, Tomashevsky, M, Marshall, K A, Phillippy, K H, Sherman, P M, Muetter, R N and Edgar, R** 2009 NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res*, 37(Database issue): D885–90. DOI: <https://doi.org/10.1093/nar/gkn764>
- Becnel, L B, Darlington, Y F, Ochsner, S A, Easton-Marks, J R, Watkins, C M, Mcowiti, A, Kankanamge, W H, Wise, M W, Dehart, M, Margolis, R N and McKenna, N J** 2015 Nuclear Receptor Signaling Atlas: Opening Access to the Biology of Nuclear Receptor Signaling Pathways. *PLoS One*, 10(9): e0135615. DOI: <https://doi.org/10.1371/journal.pone.0135615>
- Becnel, L B, Ochsner, S A, Darlington, Y F, Mcowiti, A, Kankanamge, W, Dehart, M D, Naumov, A and McKenna, N J** 2016 (In Press) Hierarchical pathway and biosource mapping support visualization of nuclear receptor signaling transcriptional networks in Transcriptomine. *Science Signaling*.
- Becnel, L B, Ochsner, S A, Darlington, Y F, Mcowiti, A, Kankanamge, W, Dehart, M D, Naumov, A and McKenna, N J** 2017 (In Press) Pathway and biosample mapping support hypothesis generation through visualization of nuclear receptor signaling networks in Transcriptomine. *Science Signaling*.
- Borgman, C L** 2011 *Why are the attribution and citation of scientific data important?* [Online]. Available at: http://sites.nationalacademies.org/PGA/brdi/PGA_064019 (Accessed May 21, 2015).
- Boyer, L A, Lee, T I, Cole, M F, Johnstone, S E, Levine, S S, Zucker, J P, Guenther, M G, Kumar, R M, Murray, H L, Jenner, R G, Gifford, D K, Melton, D A, Jaenisch, R and Young, R A** 2005 Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, 122(6): 947–56. DOI: <https://doi.org/10.1016/j.cell.2005.08.020>
- Cerami, E G, Gross, B E, Demir, E, Rodchenkov, I, Babur, O, Anwar, N, Sschultz, N, Bader, G D and Sander, C** 2011 Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res*, 39(Database issue): D685–90. DOI: <https://doi.org/10.1093/nar/gkq1039>
- Chazenbalk, G, Singh, P, Irge, D, Shah, A, Abbott, D H and Dumesic, D A** 2013 Androgens inhibit adipogenesis during human adipose stem cell commitment to preadipocyte formation. *Steroids*, 78(9): 920–6. DOI: <https://doi.org/10.1016/j.steroids.2013.05.001>
- Christian, M, Kiskinis, E, Debevec, D, Leonardsson, G, White, R and Parker, M G** 2005 Analysis of the receptor-interacting protein-140 (RIP140/Nrip1)-dependent transcriptome in the mouse adipogenic program. *Nuclear Receptor Signaling Atlas Datasets* (Jul 11, 2016). DOI: <https://doi.org/10.1621/WugGwn4VIV>
- Darke, P, Shanks, G and Broadbent, M** 1998 Successfully completing case study research: combining rigour, relevance and pragmatism. *Information Systems Journal*, 8: 273–289. DOI: <https://doi.org/10.1046/j.1365-2575.1998.00040.x>
- Darlington, Y F, Naumov, A, Mcowiti, A, Kankanamge, W H, Becnel, L B and McKenna, N J** 2016 Improving the discoverability, accessibility, and citability of omics datasets: a case report. *J Am Med Inform Assoc*. DOI: <https://doi.org/10.1093/jamia/ocw096>

- DataCite** 2015 *Cite Your Data* [Online]. Available at: <https://www.datacite.org/services/cite-your-data.html> (Accessed 21 May, 2015).
- Fabregat, A, Sidiropoulos, K, Garapati, P, Gillespie, M, Hausmann, K, Haw, R, Jassal, B, Jupe, S, Korninger, F, McKay, S, Matthews, L, May, B, Milacic, M, Rothfels, K, Shamovsky, V, Webber, M, Weiser, J, Williams, M, Wu, G, Stein, L, Hermjakob, H and D'Eustachio, P** 2016 The Reactome pathway Knowledgebase. *Nucleic Acids Res*, 44(D1): D481–7. DOI: <https://doi.org/10.1093/nar/gkv1351>
- FDA** 2014 *Providing Regulatory Submissions in Electronic Format – Submissions Under Section 745A(a) of the Federal Food, Drug, and Cosmetic Act* [Online]. Available at: <http://www.fda.gov/downloads/Drugs/Guidances/UCM384686.pdf>.
- Finck, B, Finck, B N, Kelly, D P, Finck, B N, Bernal-Mizrachi, C, Han, D H, Coleman, T, Sambandam, N, Lariviere, L L, Holloszy, J O, Semenkovich, C F and Kelly, D P** 2005 Analysis of the peroxisome proliferator activated receptor- α (PPAR α /Ppara)-regulated transcriptome in mouse skeletal muscle. *Nuclear Receptor Signaling Atlas Datasets* (Sep 15, 2016). DOI: <https://doi.org/10.1621/4rQYtJ4ZdB>
- Goodman, L, Lawrence, R and Ashley, K** 2012 Data-set visibility: Cite links to data in reference lists. *Nature*, 492(7429): 356. DOI: <https://doi.org/10.1038/492356d>
- Hoffman, E, Almon, R, Almon, R R, Lai, W, Dubois, D C and Jusko, W J** 2005 Time course analysis of the methylprednisolone (MePred)-regulated transcriptome in rat kidney. *Nuclear Receptor Signaling Atlas Datasets* (Sep 15, 2016). DOI: <https://doi.org/10.1621/h9hgIA2vKz>
- Huber-Keener, K J, Liu, X, Wang, Z, Wang, Y, Freeman, W, Wu, S, Planas-Silva, M D, Ren, X, Cheng, Y, Zhang, Y, Vrana, K, Liu, C G, Yang, J M and Wu, R** 2012 Differential gene expression in tamoxifen-resistant breast cancer cells revealed by a new analytical model of RNA-Seq data. *PLoS One*, 7(7): e41333. DOI: <https://doi.org/10.1371/journal.pone.0041333>
- James, C, James, C G, Ulici, V, Tuckermann, J, Underhill, T M, Beier, F, James, C G, Ulici, V, Tuckermann, J, Underhill, T M and Beier, F** 2007 Time course transcriptomic analysis of dexamethasone (Dex)-treated mouse chondrocytes. *Nuclear Receptor Signaling Atlas Datasets* (Sep 15, 2016). DOI: <https://doi.org/10.1621/ZyPkWLa8Sl>
- Kazmin, D A, McDonnell, D P, Kazmin, D, Prytkova, T, Cook, C E, Wolfinger, R, Chu, T M, Beratan, D, Norris, J D, Chang, C Y and McDonnell, D P** 2006 Analysis of the dihydrotestosterone (DHT)- and RTI 6413-018-dependent transcriptomes in LNCaP prostate cancer cells. *Nuclear Receptor Signaling Atlas Datasets* (Sep 15, 2016). DOI: <https://doi.org/10.1621/PG6srxDzNt>
- Kolesnikov, N, Hastings, E, Keays, M, Melnichuk, O, Tang, Y A, Williams, E, Dylag, M, Kurbatova, N, Brandizi, M, Burdett, T, Megy, K, Pilicheva, E, Rustici, G, Tikhonov, A, Parkinson, H, Petryszak, R, Sarkans, U and Brazma, A** 2015 ArrayExpress update—simplifying data submissions. *Nucleic Acids Res*, 43(Database issue): D1113–6. DOI: <https://doi.org/10.1093/nar/gku1057>
- Lin, B, Lin, B, Wang, J, Hong, X, Yan, X, Hwang, D, Cho, J H, Yi, D, Utleg, A G, Fang, X, Schones, D E, Zhao, K, Omenn, G S and Hood, L** 2009 Analysis of the androgen receptor (AR)-dependent transcriptome in PC3 prostatic carcinoma cells. *Nuclear Receptor Signaling Atlas Datasets* (Sep 14, 2016). DOI: <https://doi.org/10.1621/wGxM43bqxz>
- Liu, J, Lee, J, Salazar Hernandez, M A, Mazitschek, R and Ozcan, U** 2015 Treatment of obesity with celestrol. *Cell*, 161(5): 999–1011. DOI: <https://doi.org/10.1016/j.cell.2015.05.011>
- Mamrosh, J L, Lee, J M, Wagner, M, Stambrook, P J, Whitby, R J, Sifers, R N, Wu, S P, Tsai, M J, DeMayo, F J and Moore, D D** 2015 Correction: Nuclear receptor LRH-1/NR5A2 is required and targetable for liver endoplasmic reticulum stress resolution. *Elife*, 4: e10084. DOI: <https://doi.org/10.7554/eLife.10084>
- Mangelsdorf, D J, Thummel, C, Beato, M, Herrlich, P, Schutz, G, Umesono, K, Blumberg, B, Kastner, P, Mark, M, Chambon, P and Evans, R M** 1995 The nuclear receptor superfamily: the second decade. *Cell*, 83(6): 835–9. DOI: [https://doi.org/10.1016/0092-8674\(95\)90199-X](https://doi.org/10.1016/0092-8674(95)90199-X)
- Margolis, R, Derr, L, Dunn, M, Huerta, M, Larkin, J, Sheehan, J, Guyer, M and Green, E D** 2014 The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *J Am Med Inform Assoc*, 21(6): 957–8. DOI: <https://doi.org/10.1136/amiainjnl-2014-002974>
- Martone, M E** 2014 *Data Citation Synthesis Group: Joint Declaration of Data Citation Principles*. *San Diego, FORCE11* [Online]. Available at: <https://www.force11.org/datacitation>.
- McKenna, N J and O'Malley, B W** 2002 Combinatorial control of gene expression by nuclear receptors and coregulators. *Cell*, 108(4): 465–74. DOI: [https://doi.org/10.1016/S0092-8674\(02\)00641-4](https://doi.org/10.1016/S0092-8674(02)00641-4)
- NCBI** 2016 *GEO2R* [Online]. Available at: <https://www.ncbi.nlm.nih.gov/geo/geo2r/>.

- Neveol, A, Wilbur, W J and Lu, Z** 2012 Improving links between literature and biological data with text mining: a case study with GEO, PDB and MEDLINE. *Database (Oxford)* (bas026). DOI: <https://doi.org/10.1093/database/bas026>
- NIH** 2016 *NIH Genomic Data Sharing Policy* [Online]. Available at: <https://gds.nih.gov/03policy2.html>
- Notas, G, Pelekanou, V, Castanas, E and Kampa, M** 2010 Conjugated and non-conjugated androgens differentially modulate specific early gene transcription in breast cancer in a cell-specific manner. *Steroids*, 75(8–9): 611–8. DOI: <https://doi.org/10.1016/j.steroids.2009.10.004>
- OAI** 2016 *The Open Archives Initiative Protocol for Metadata Harvesting* [Online]. Available at: <http://www.openarchives.org/OAI/openarchivesprotocol.html>
- Ochsner, S A, Schupp, M, Cristancho, A G, Lefterova, M I, Hanniman, E A, Briggs, E R, Steger, D J, Qatanani, M, Curtin, J C, Schug, J, Ochsner, S A, McKenna, N J, Lazar, M A, Schupp, M, Cristancho, A G, Lefterova, M I, Hanniman, E A, Briggs, E R, Steger, D J, Qatanani, M, Curtin, J C, Schug, J, Ochsner, S A, McKenna, N J and Lazar, M A** 2009 Analysis of the peroxisome proliferator-activated receptor- γ (PPAR γ /Pparg)-regulated transcriptome during adipogenesis in mouse 3T3-L1 adipocytes. *Nuclear Receptor Signaling Atlas Datasets* (Sep 15, 2016). DOI: <https://doi.org/10.1621/nWjv1UhrEE>
- Ochsner, S A, Steffen, D L, Stoeckert, C J, Jr. and McKenna, N J** 2008 Much room for improvement in deposition rates of expression microarray datasets. *Nat Methods*, 5(12): 991. DOI: <https://doi.org/10.1038/nmeth1208-991>
- Piller, C** 2015 *Law ignored, patients at risk* [Online]. Available at: <https://www.statnews.com/2015/12/13/clinical-trials-investigation/>
- Rifai, N, Bossuyt, P M, Ioannidis, J P, Bray, K R, McShane, L M, Golub, R M and Hooft, L** 2014 Registering diagnostic and prognostic trials of tests: is it the right thing to do? *Clin Chem*, 60(9): 1146–52. DOI: <https://doi.org/10.1373/clinchem.2014.226100>
- Scott, M A, Nguyen, V T, Levi, B and James, A W** 2011 Current methods of adipogenic differentiation of mesenchymal stem cells. *Stem Cells Dev*, 20(10): 1793–804. DOI: <https://doi.org/10.1089/scd.2011.0040>
- Sharova, L V, Sharov, A A, Piao, Y, Shaik, N, Sullivan, T, Stewart, C L, Hogan, B L and Ko, M S** 2007 Global gene expression profiling reveals similarities and differences among mouse pluripotent stem cells of different origins and strains. *Dev Biol*, 307(2): 446–59. DOI: <https://doi.org/10.1016/j.ydbio.2007.05.004>
- Singh, R, Artaza, J N, Taylor, W E, Braga, M, Yuan, X, Gonzalez-Cadavid, N F and Bhasin, S** 2006 Testosterone inhibits adipogenic differentiation in 3T3-L1 cells: nuclear translocation of androgen receptor complex with beta-catenin and T-cell factor 4 may bypass canonical Wnt signaling to down-regulate adipogenic transcription factors. *Endocrinology*, 147(1): 141–54. DOI: <https://doi.org/10.1210/en.2004-1649>
- Swindell, W R, Johnston, A and Gudjonsson, J E** 2014 Early tissue responses to etanercept in psoriasis lesions. *Gene Expression Omnibus*.
- Szatmari, I, Nagy, T, Agostini, M, Chatterjee, K, Nagy, L, Szatmari, I, Töröcsik, D, Agostini, M, Nagy, T, Gurnell, M, Barta, E, Chatterjee, K and Nagy, L** 2007 Time course- and dose-dependent analysis of the rosiglitazone (Rosi)-dependent and peroxisome proliferator-activated receptor- γ (PPAR γ /Pparg) DNA-binding domain-dependent transcriptomes in dendritic cells. *Nuclear Receptor Signaling Atlas Datasets* (Sep 15, 2016). DOI: <https://doi.org/10.1621/p3ZSUojETc>
- Thorn, C F, Klein, T E and Altman, R B** 2013 PharmGKB: the Pharmacogenomics Knowledge Base. *Methods Mol Biol*, 1015: 311–20. DOI: https://doi.org/10.1007/978-1-62703-435-7_20
- Wilkinson, M D, Dumontier, M, Aalbersberg, I J, Appleton, G, Axton, M, Baak, A, Blomberg, N, Boiten, J W, Da Silva Santos, L B, Bourne, P E, Bouwman, J, Brookes, A J, Clark, T, Crosas, M, Dillo, I, Dumon, O, Edmunds, S, Evelo, C T, Finkers, R, Gonzalez-Beltran, A, Gray, A J, Groth, P, Goble, C, Grethe, J S, Heringa, J, T Hoen, P A, Hooft, R, Kuhn, T, Kok, R, Kok, J, Lusher, S J, Martone, M E, Mons, A, Packer, A L, Persson, B, Rocca-Serra, P, Roos, M, Van Schaik, R, Sansone, S A, Schultes, E, Sengstag, T, Slater, T, Strawn, G, Swertz, M A, Thompson, M, Van Der Lei, J, Van Mulligen, E, Velterop, J, Waagmeester, A, Wittenburg, P, Wolstencroft, K, Zhao, J and Mons, B** 2016 The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, 3: 160018. DOI: <https://doi.org/10.1038/sdata.2016.18>
- Witwer, K W** 2013 Data submission and quality in microarray-based microRNA profiling. *Clin Chem*, 59(2): 392–400. DOI: <https://doi.org/10.1373/clinchem.2012.193813>
- Wu, J, Boström, P, Sparks, L M, Ye, L, Choi, J H, Giang, A, Khandekar, M, Virtanen, K A, Nuutila, P, Schaart, G, Huang, K, Tu, H, Van Marken Lichtenbelt, W D, Hoeks, J, Enerbäck, S, Schrauwen, P, Spiegelman, B M, Wu, J, Boström, P, Sparks, L M, Ye, L, Choi, J H, Giang, A H, Khandekar, M,**

Virtanen, K A, Nuutila, P, Schaart, G, Huang, K, Tu, H, Van Marken Lichtenbelt, W D, Hoeks, J, Enerbäck, S, Schrauwen, P and Spiegelman, B M 2012 Comparative transcriptomic analysis of white, brown and beige fat cell lines. *Nuclear Receptor Signaling Atlas Datasets* (Sep 15, 2016). DOI: <https://doi.org/10.1621/Hn4dqri4Pd>


Yin, R K 2013 *Case Study Research: Design and Methods*. SAGE Publications, Inc.

Zacharewski, T R, Kim, S, Kiyosawa, N, Burgoon, L D, Chang, C, Kim, S, Kiyosawa, N, Burgoon, L D, Chang, C C and Zacharewski, T R 2013 Acute and chronic time course analysis of the WY14643-regulated transcriptome in Female ovariectomized C57BL/6 mouse liver. *Nuclear Receptor Signaling Atlas Datasets* (Sep 15, 2016). DOI: <https://doi.org/10.1621/bhKR1Y5xgj>

How to cite this article: Ochsner, S A, Darlington, Y F, McOwiti, A, Kankanamge, W H, Naumov, A, Becnel, L B and McKenna, N J 2017 A FAIR-Based Approach to Enhancing the Discovery and Re-Use of Transcriptomic Data Assets for Nuclear Receptor Signaling Pathways. *Data Science Journal*, 16: 11, pp. 1–15, DOI: <https://doi.org/10.5334/dsj-2017-011>

Submitted: 16 September 2016 **Accepted:** 16 February 2017 **Published:** 23 March 2017

Copyright: © 2017 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS 