SRP Workshop: Bioinformatics Tools for Data Integration

SUSAN TILTON

Oregon State University
Corvallis, OR
September 13-14, 2017

Overview of omics data integration

Rationale

Data integration: use of multiple sources of information (or data) collectively within a system

Goal: to provide a better understanding of a system/situation/association to improve knowledge discovery

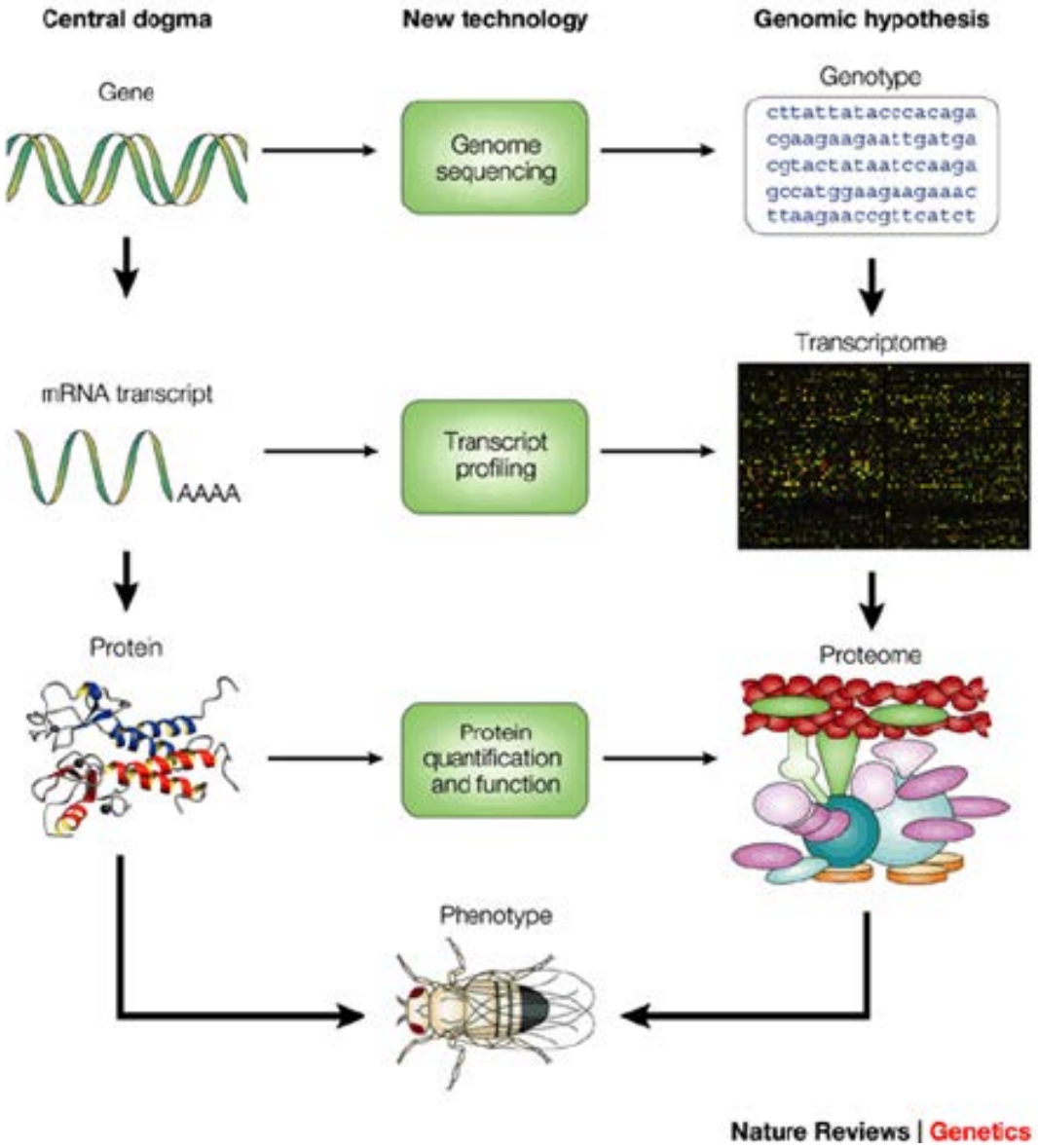
Challenges: data explosion associated with high content technologies

Rationale

Make comparisons among datasets:

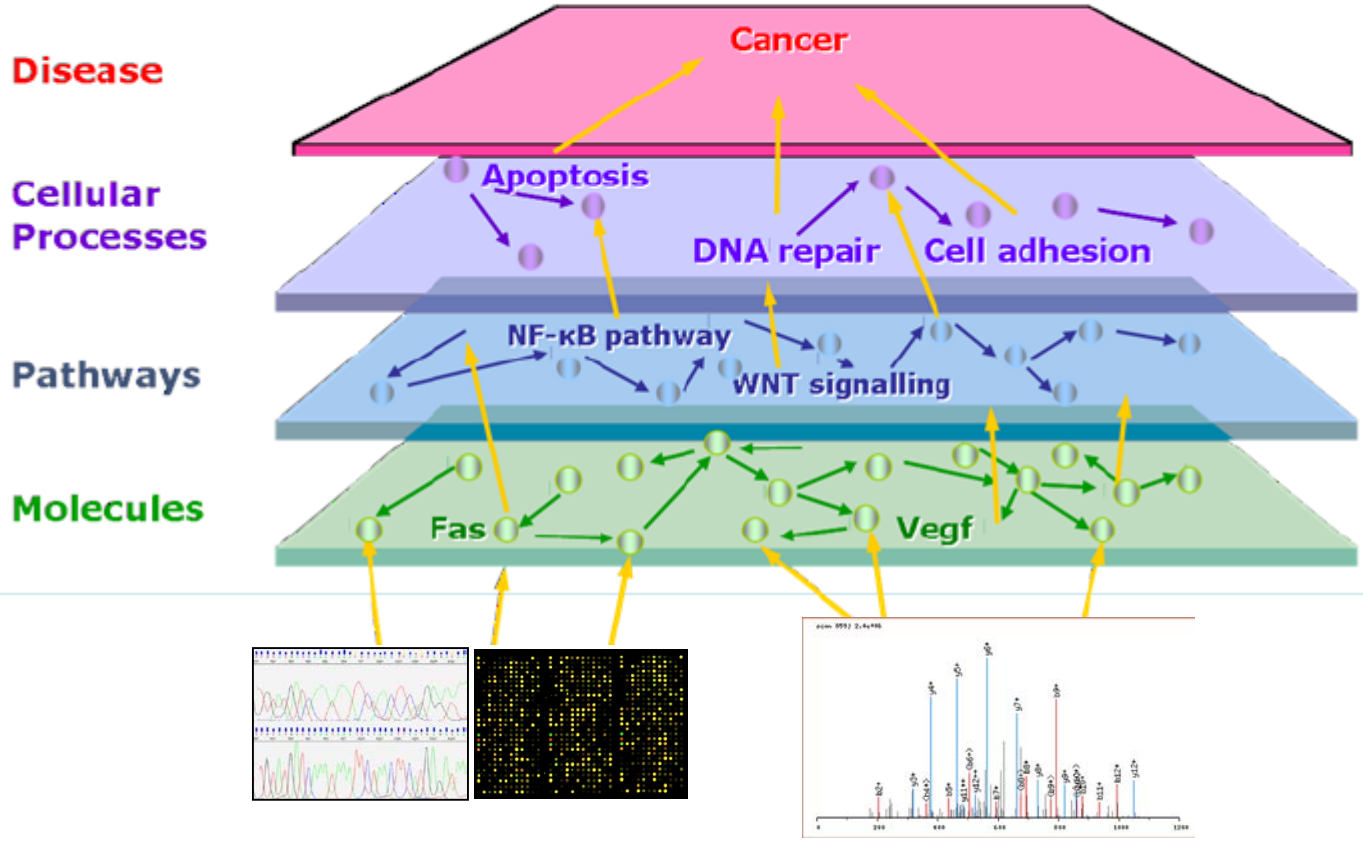
- Technologies (gene/proteins/metabolites)
- Species (mouse/human/zebrafish)
- Platforms (microarray/RNAseq)
- Phenotype (toxicity/disease)
- Experimental models (in vitro/in vivo)
- Dose or time
- Chemical class

Data and technologies



Data and technologies

Ultimate Goal:
Combine data to increase knowledge

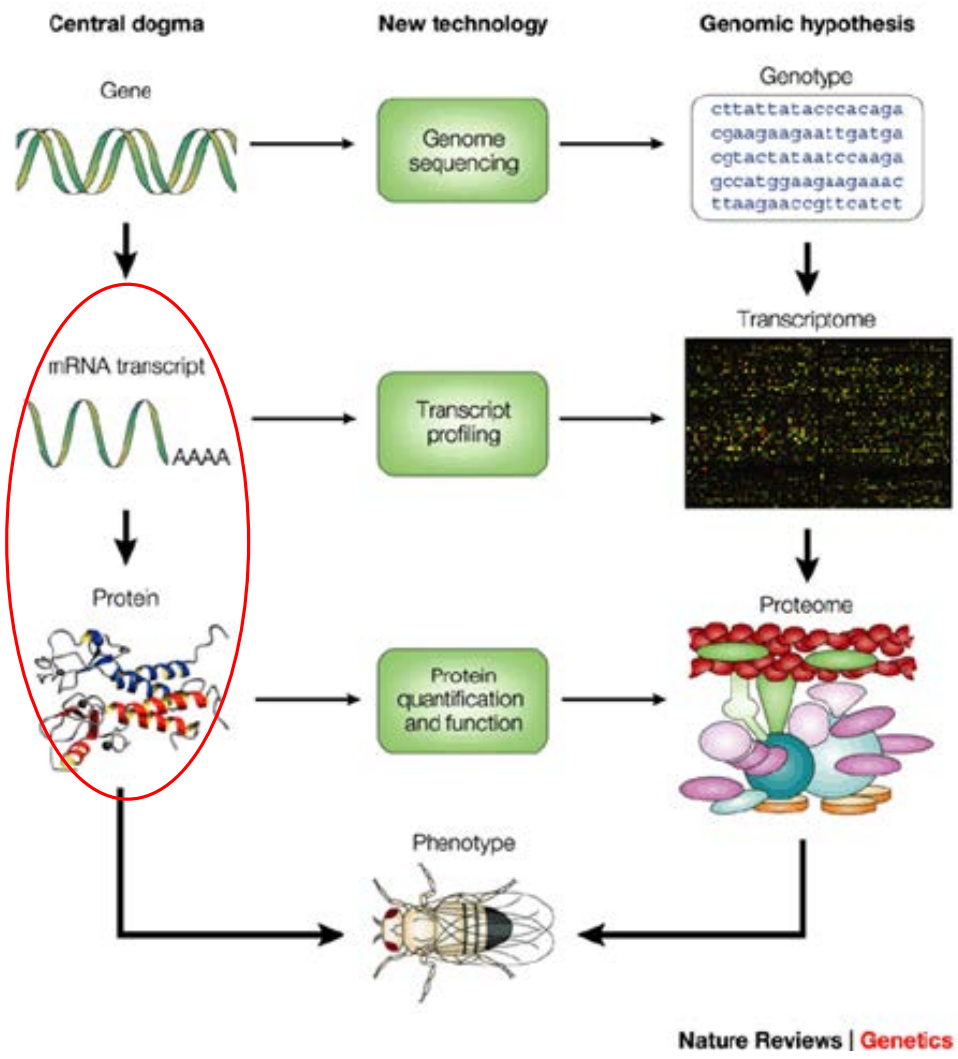


Approaches for data integration

- Direct integration
  - Genes/proteins
  - Cross-species
  - In vivo / in vitro
- Integration at functional or pathway level
- Integration based on common upstream regulators
  - Transcription factors
  - miRNAs
- Integration based on statistical and network-based approaches
  - Correlation
  - Clustering
  - Network structure

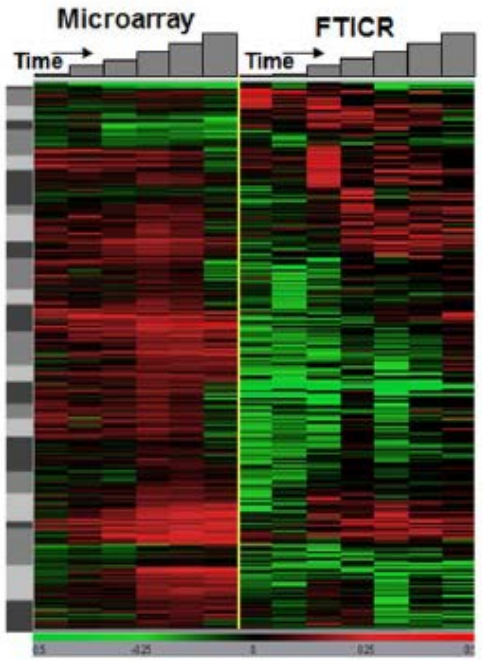
Direct integration of Transcriptomics and Proteomics

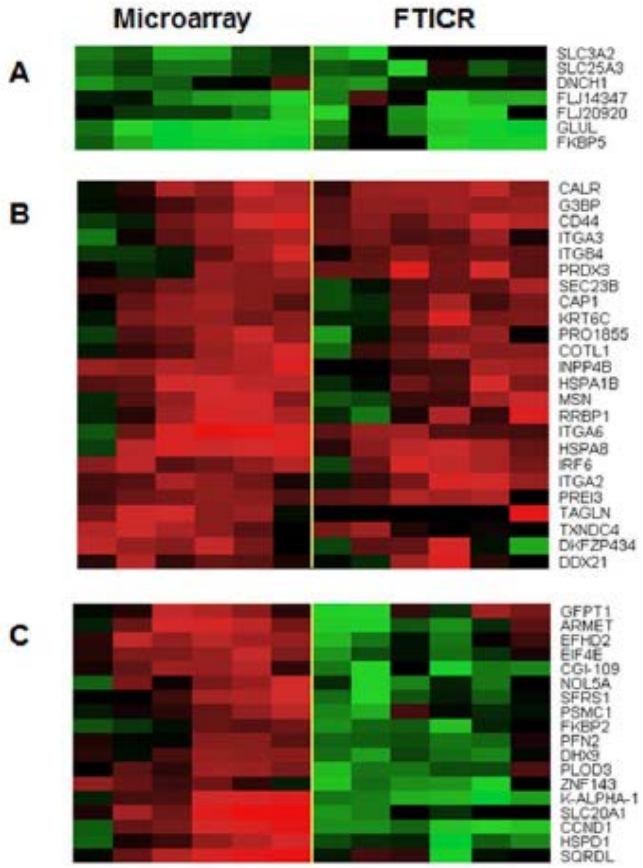
- Assumption that gene and protein regulation is directly correlated



Direct integration of Transcriptomics and Proteomics

- Assumption that gene and protein regulation is directly correlated
- Lack of direct correlation between genes & proteins due to following challenges:
  - Post-translational modifications
  - Gene/protein turnover and degradation
  - Statistical filtering / platform differences and limitations → missing data

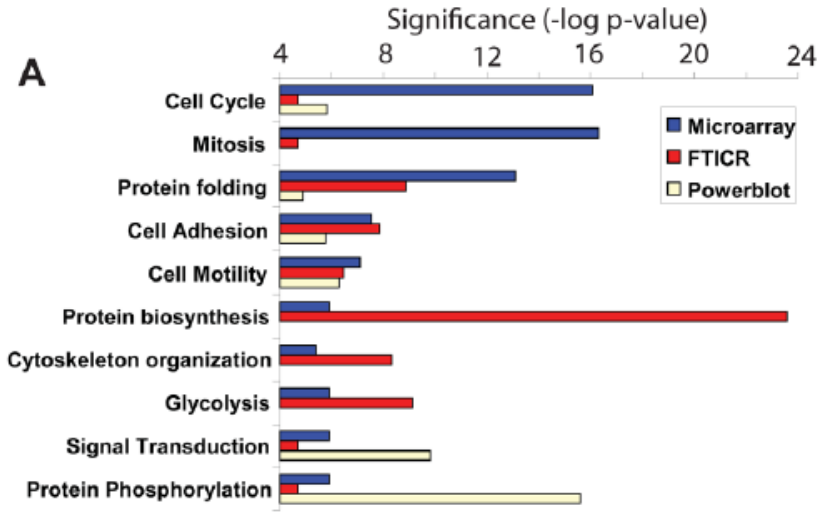


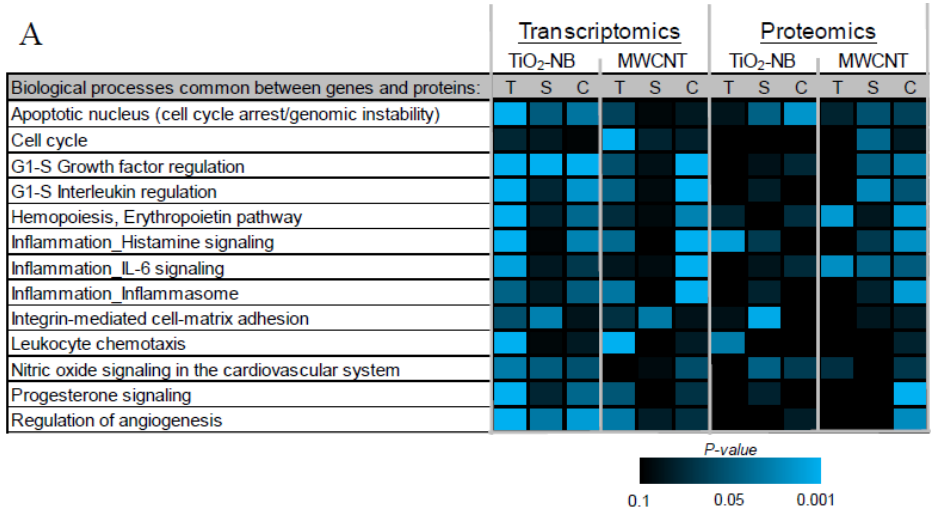


[1]WATERS ET AL. 2012. PLOS ONE, 7(3):E34515

Integration at functional or pathway level

- Correlation among datasets is observed at the functional or pathway level
- Even though the same genes and proteins are not regulated simultaneously within the system, genes and proteins within the same pathways are co-expressed
- Measured through pathway or functional enrichment analysis

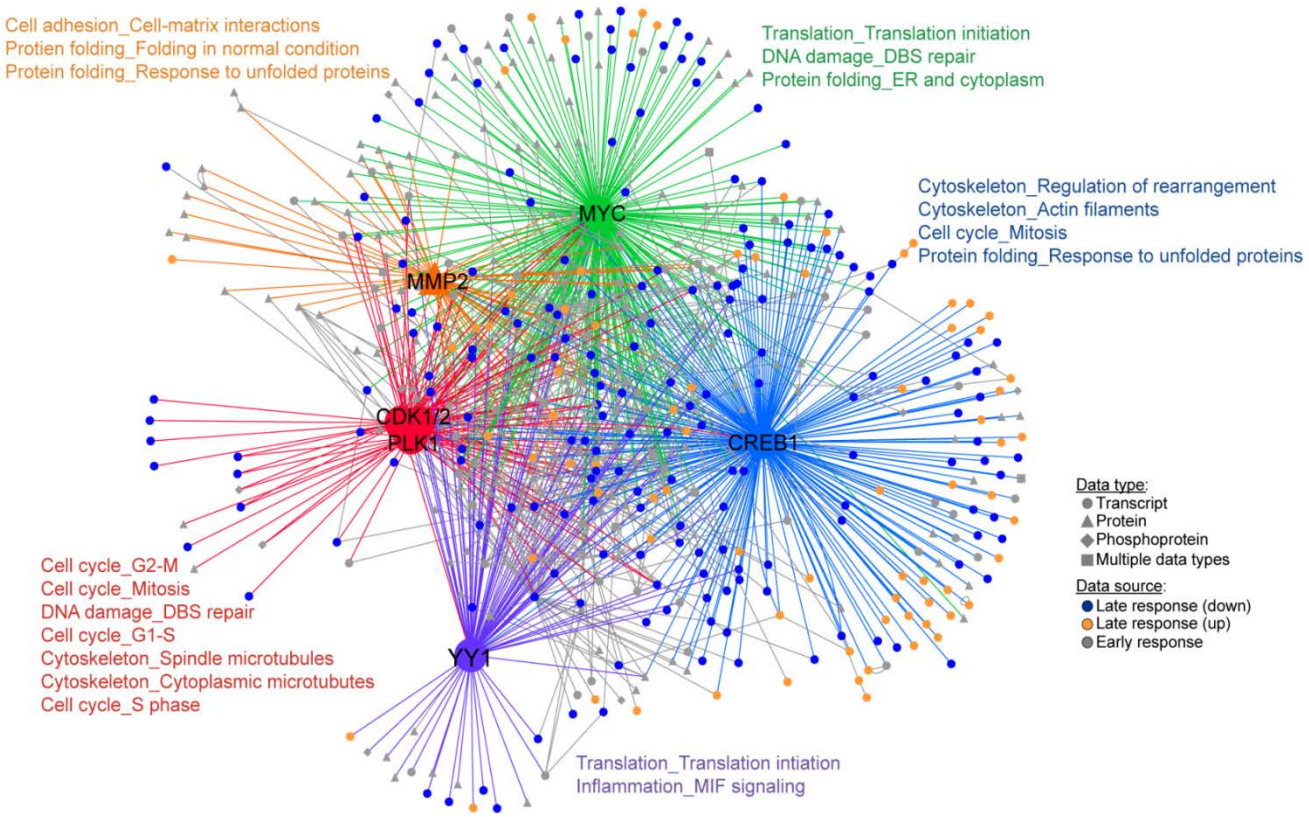




[1]WATERS ET AL. 2012. PLOS ONE, 7(3):E34515
[2]TILTON ET AL. 2014. NANOTOXICOLOGY, 8(5):533-48.

Integration based on common upstream regulators

- Similarly, overlap among datasets is observed through common upstream regulators (e.g. transcription factors)
- Measured through transcription factor enrichment, hub analysis



[3]TILTON ET AL. 2015. TOXICOL APPL PHARMACOL (*ACCEPTED*).

Practical challenges for data integration

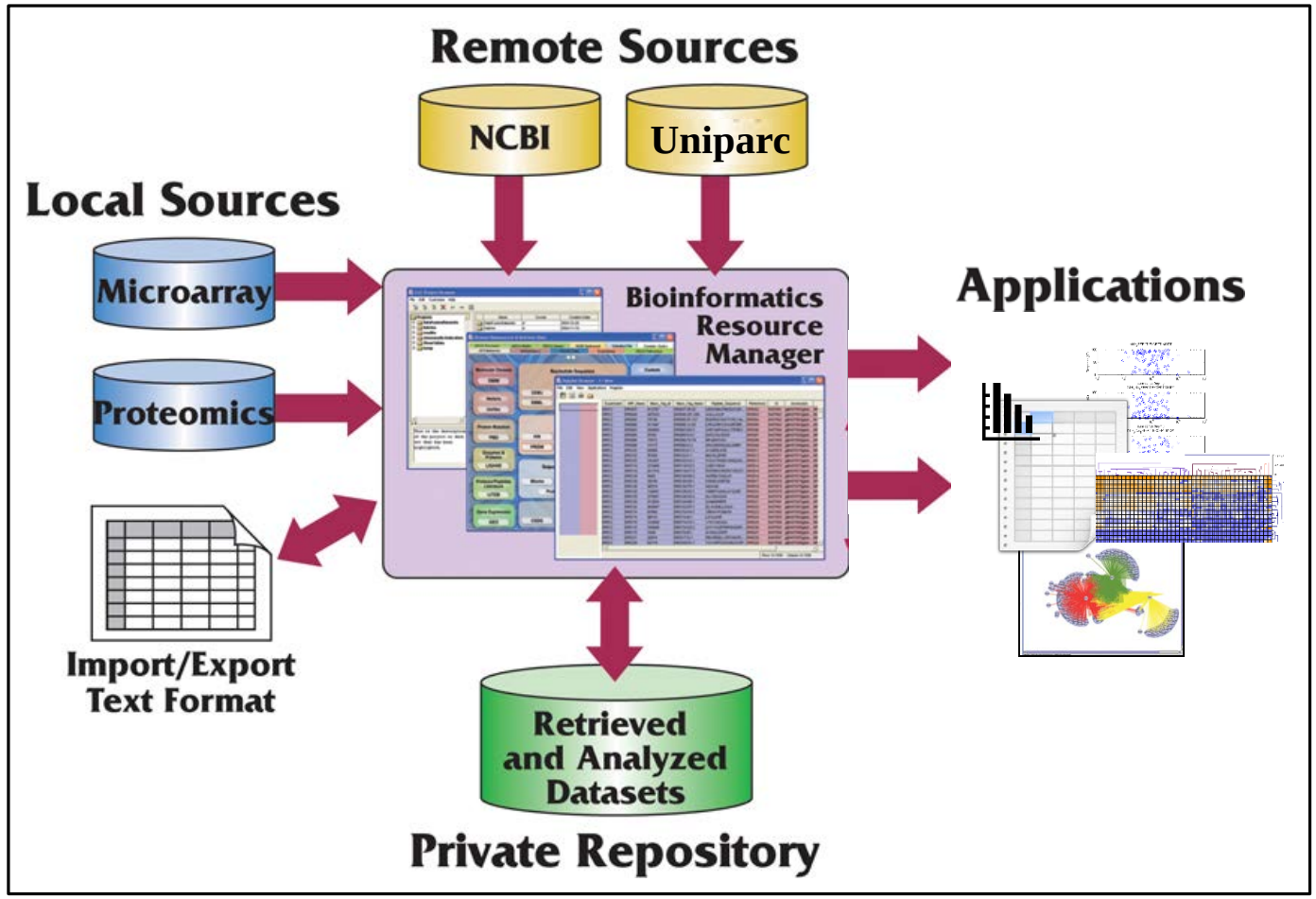
- Choosing appropriate approach to answer biological questions
  - Handle missing data
  - Applying statistical/filters to data

- Common bioinformatics problem of translating identifiers across data tables
  - Genes → proteins
  - Mouse → Human
  - Entrez Gene ID → Ensembl Gene ID

Bioinformatics Resource Manager (BRM): A systems biology web tool for miRNA and omics with data integration

BRM Overview

A web tool developed to provide biological scientists with computation tools for annotation retrieval, cross-referencing and integration of high-throughput (HTP) data.



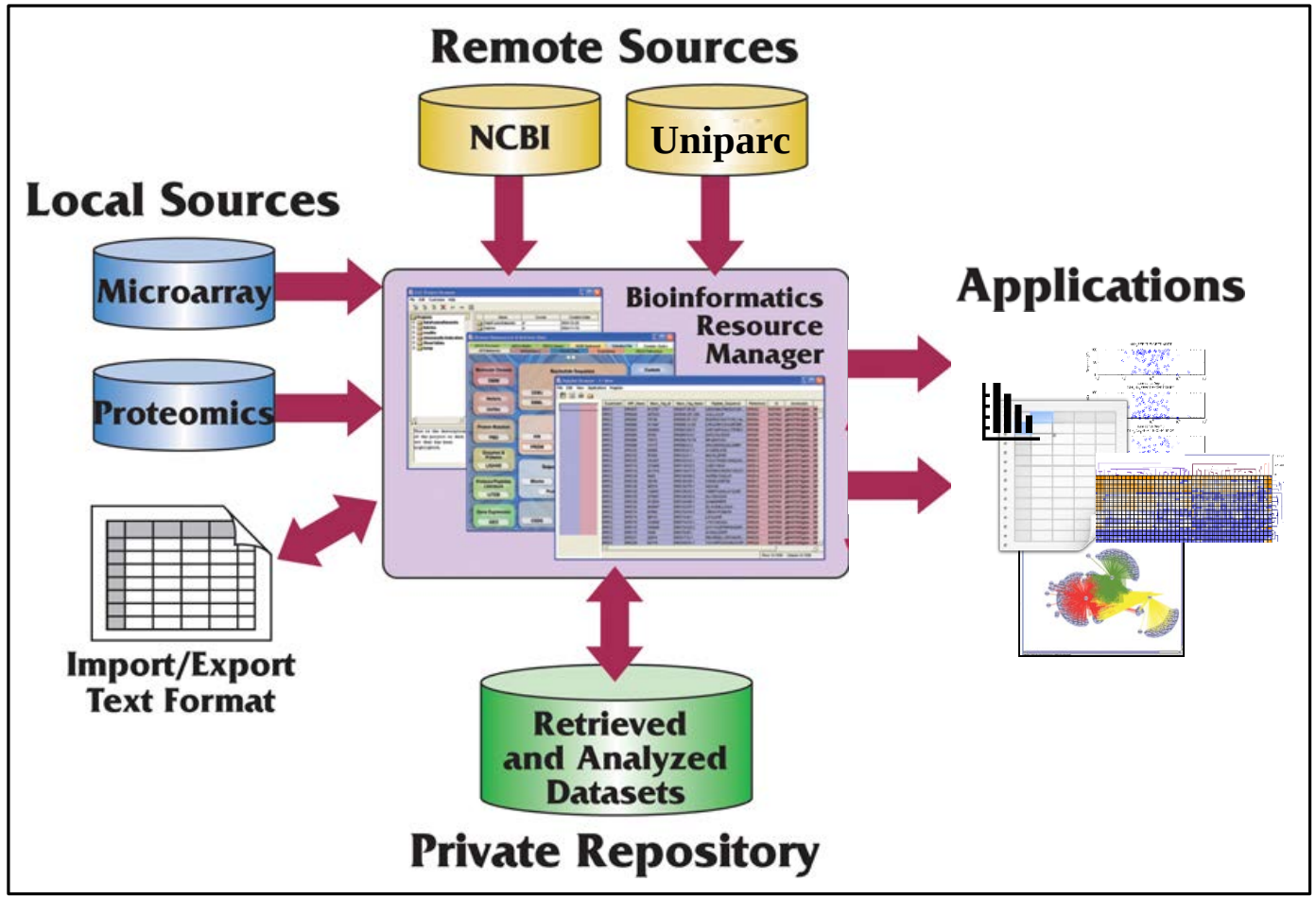
TILTON ET AL. 2019. BMC BIOINFORMATICS. 20(1):255

BRM Overview

Provide a platform for integration across HTP datasets

Examples include:

- miRNA/mRNA
- Transcriptomics/ Proteomics
- Mouse/Human/Rat/ Zebrafish/Macaque
- Cross-Identifier



TILTON ET AL. 2019. BMC BIOINFORMATICS. 20(1):255

BRM Overview

Brown *et al. BMC Bioinformatics* (2019) 20:255
https://doi.org/10.1186/s12859-019-2805-6

BMC Bioinformatics

DATABASE Open Access

Bioinformatics Resource Manager: a systems biology web tool for microRNA and omics data integration

Joseph Brown[1,5], Aaron R. Phillips[2], David A. Lewis[2], Michael-Andres Mans[3], Yvonne Chang[3], Robert L. Tanguay[3,4], Elena S. Peterson[2], Katrina M. Waters[1,3,4*] and Susan C. Tilton[3,4*]

BRM Capabilities

Bath Annotation Retrieval

- Gene Annotation (NCBI)
- Protein Annotation (Uniparc)
- miRNA Annotation (MiRBase/microCosm)

XREF Identifier Retrieval

- Gene Identifiers (NCBI)
- Protein Identifiers (Uniparc)
- Cross Species Orthologs (Ensembl)

Integration of Data Tables

- Includes all data in output

BRM Capabilities

miRNA Datasource Retrieval

- miRNA Predicted Targets (TargetScan, microcosm, microRNA)
- miRNA Identifiers
- miRNA Metadata

Data Workflows *(performs bioinformatics tasks in background)*

- miRNA Predicted Target Retrieval
- miRNA Target Prediction and Integration with Co-expressed Data
- Data Integration Workflow
  - Transcriptomic and Proteomic Data
  - Cross-Species Data

BRM Web Access[5]
(http://cbb.pnnl.gov/brm/)

Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

U.S. DEPARTMENT OF ENERGY

PNNL Home | About | Research | Publications | Jobs | News | Contacts Search PNNL

Bioinformatics Resource Manager

- Add Identifiers
- Merge Tables
- miRNA Targets
- miRNA Convert

Cite BRM Software

Tilton SC, Tal TL, Scroggins SM, Franzosa JA, Peterson ES, Tanguay RL and Waters KM. 2012. Bioinformatics Resource Manager v2.3: An integrated software environment for systems biology with microRNA and cross-species analysis tools. BMC Bioinformatics. 13:311. doi: 10.1186/1471-2105-13-311. PMCID: PMC3534564. PMID: 23174015.

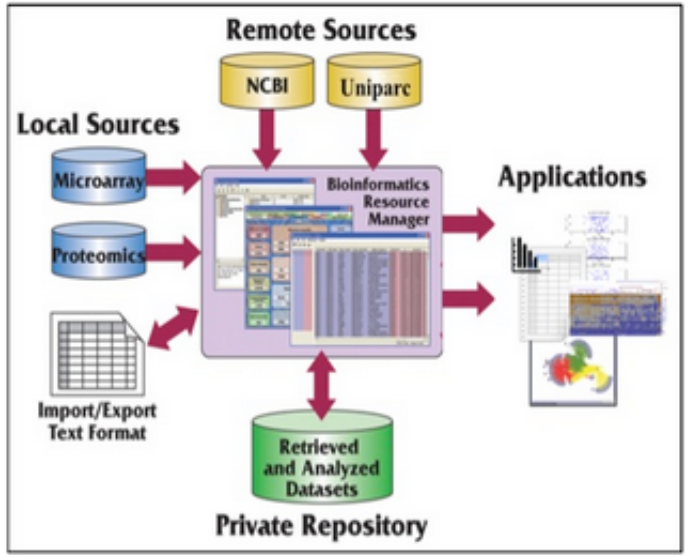
Shah AR, M Singhal, KR Klicker, EG Stephan, HS Wiley, KM Waters. 2007. "Enabling high-throughput data management for systems biology: The Bioinformatics Resource Manager." Bioinformatics 2007 23(7):906-909. doi:10.1093/bioinformatics/btm031. PMID: 17324940.

Bioinformatics Resource Manager Info

The Bioinformatics Resource Manager (BRM) is a software environment developed for data retrieval, integration and analysis of high-throughput (HTP; transcriptomic, proteomic or sequencing) biological data. BRM provides computational tools for biologists to merge datasets, cross-reference gene or protein identifiers, map identifiers across species and add functional annotation from NCBI, UniProt, Ensembl or microRNA databases, including predicted miRNA targets from multiple sources. BRM utilizes easy to navigate workflows for identification of predicted miRNA gene targets and integration with experimental mRNA or protein datasets. BRM further provides generic workflows for integrating cross-platform, cross-technology or cross-species. These tools provide biological researchers with a platform for straightforward integration and analysis of heterogeneous HTP datasets critical for biomedical research.

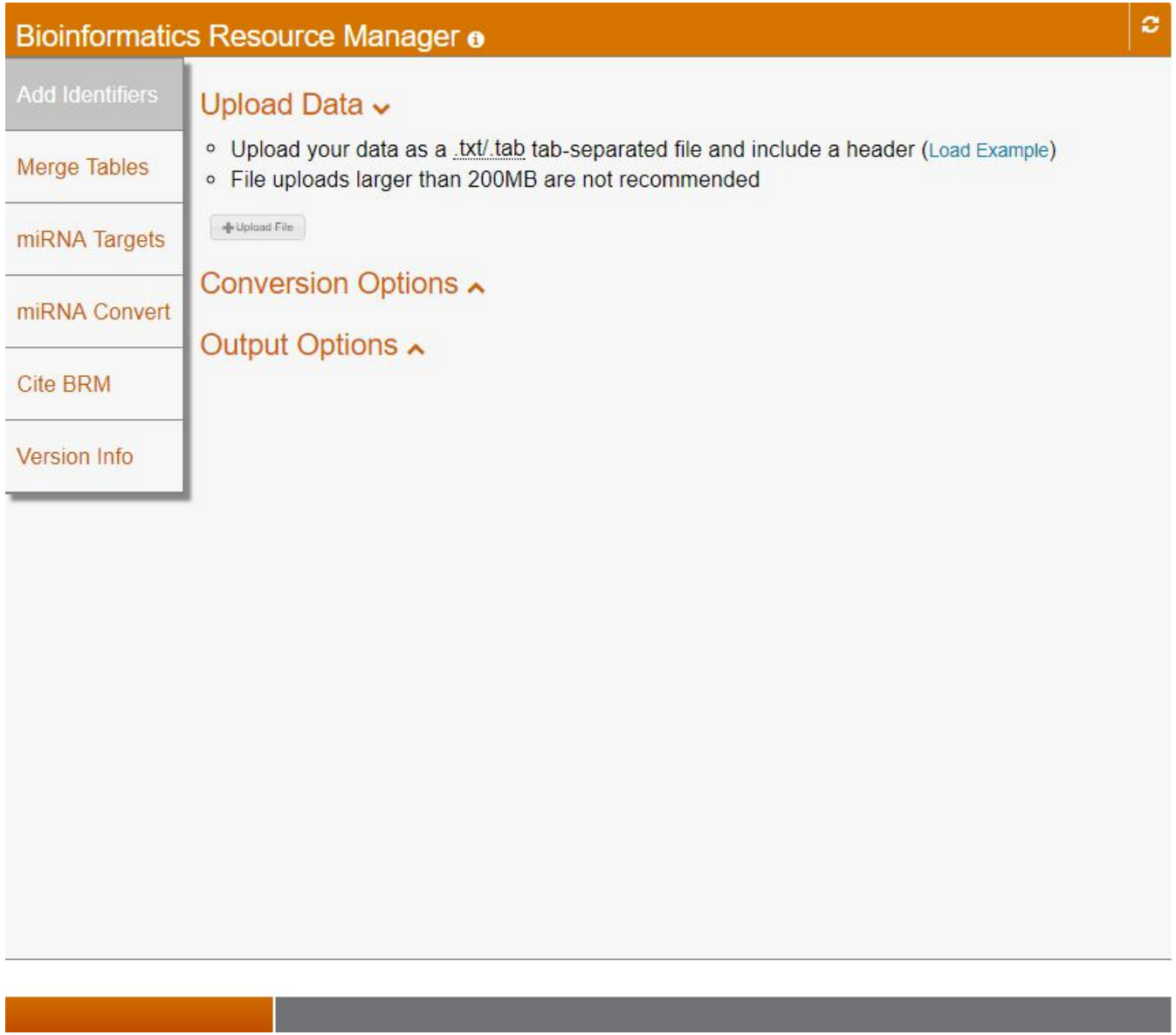
Use BRM for integrating across HTP experimental datasets. Examples include:

- miRNA/mRNA
- Transcriptomics/Proteomics
- RNAseq/Microarray
- Mouse/Human/Rat/Zebrafish/Macaque
- Cross-Identifier (e.g. Entrez gene ID/Ensembl gene ID)



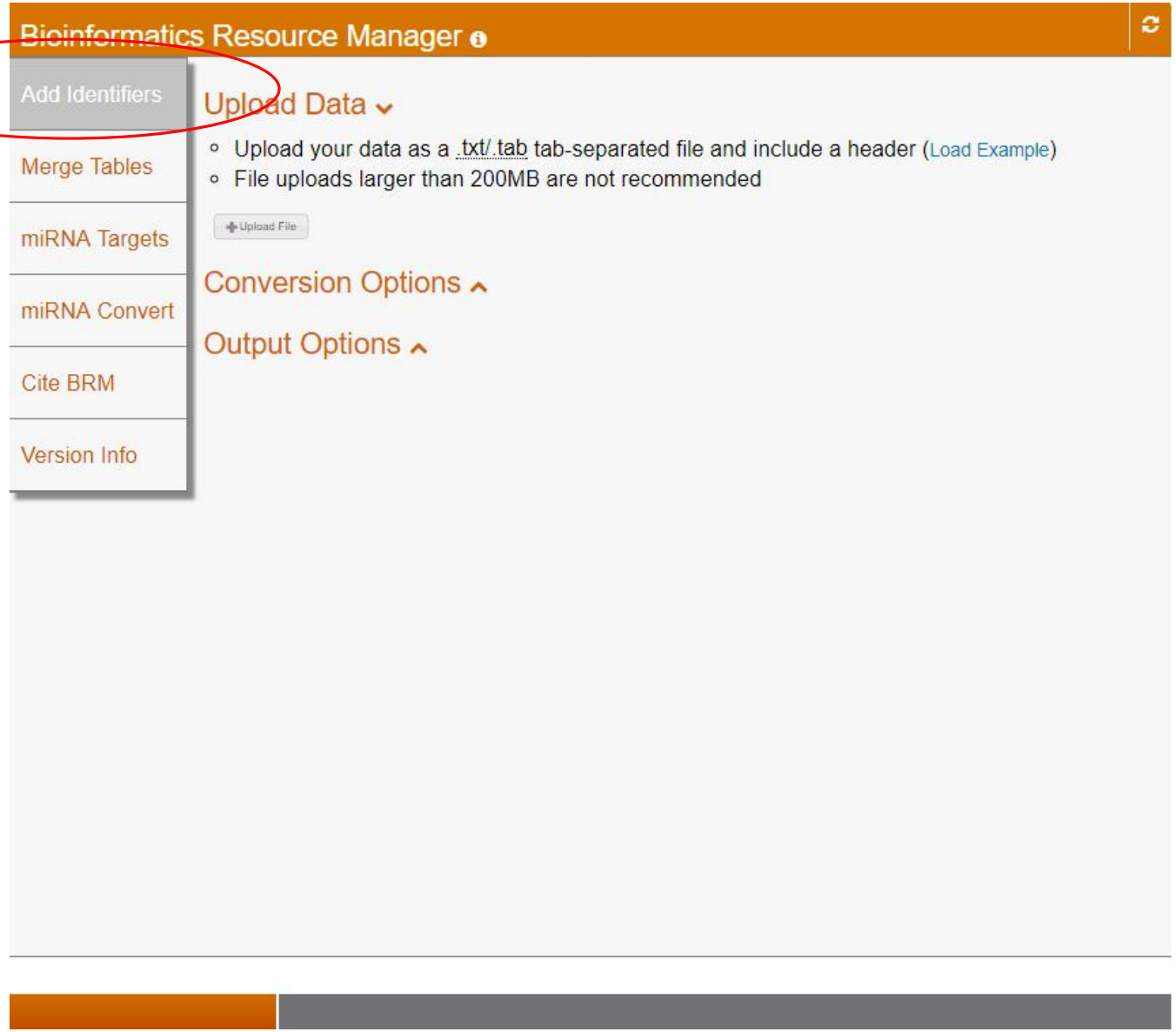
BRM Organization: Homepage

Workflows:



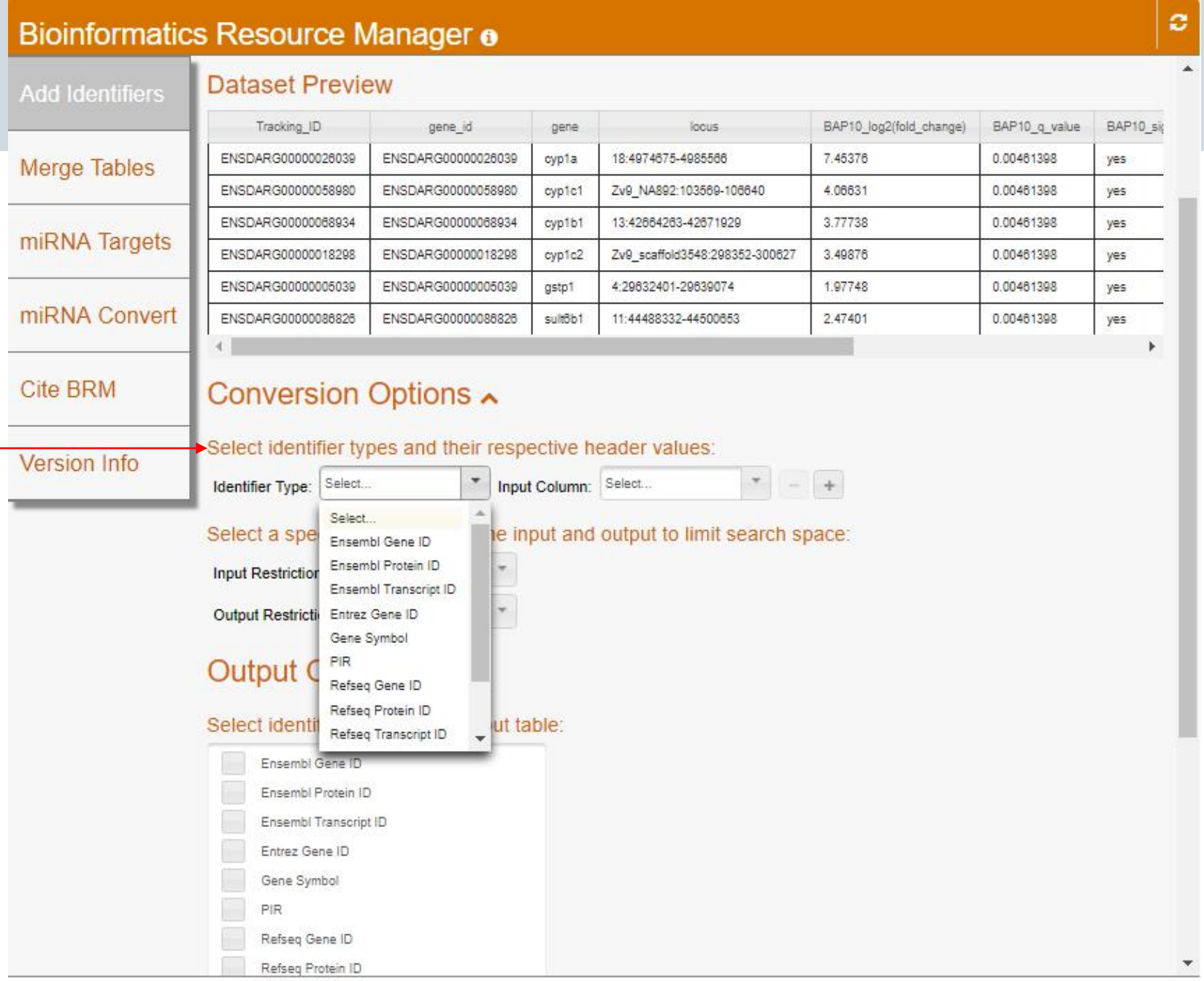
BRM Organization: Homepage

Workflows:



BRM: Batch Annotation Retrieval

Cross-identifier/platform



BRM: Batch Annotation Retrieval

Cross-species

Bioinformatics Resource Manager

- Add Identifiers
- Merge Tables
- miRNA Targets
- miRNA Convert
- Cite BRM
- Version Info

Dataset Preview

Tracking_ID	gene_id	gene	locus	BAP10_log2(fold_change)	BAP10_q_value	BAP10_si
ENSDARG00000026039	ENSDARG00000026039	cyp1a	18:4974675-4985566	7.45376	0.00461398	yes
ENSDARG00000058980	ENSDARG00000058980	cyp1c1	Zv9_NA892:103569-106640	4.06631	0.00461398	yes
ENSDARG00000068934	ENSDARG00000068934	cyp1b1	13:42664263-42671929	3.77738	0.00461398	yes
ENSDARG00000018298	ENSDARG00000018298	cyp1c2	Zv9_scaffold3548:298352-300627	3.49876	0.00461398	yes
ENSDARG00000005039	ENSDARG00000005039	gstp1	4:29632401-29639074	1.97748	0.00461398	yes
ENSDARG00000086826	ENSDARG00000086826	sult6b1	11:44488332-44500653	2.47401	0.00461398	yes

Conversion Options

Select identifier types and their respective header values:

Identifier Type: Ensembl Gene ID Input Column: gene_id

Select a species restriction on the input and output to limit search space:

Input Restriction: None

Output Restriction:

- None
- Rat
- Human
- Macaque
- Mouse
- Zebrafish

Output Op

Select identifier table:

- Ensembl Gene ID
- Ensembl Protein ID
- Ensembl Transcript ID
- Entrez Gene ID
- Gene Symbol
- PIR
- Refseq Gene ID
- Refseq Protein ID

BRM: Batch Annotation Retrieval

Bioinformatics Resource Manager

- Add Identifiers
- Merge Tables
- miRNA Targets
- miRNA Convert
- Cite BRM
- Version Info

Identifier Type: Ensembl Gene ID Input Column: gene_id

Select a species restriction on the input and output to limit search space:

Input Restriction: Zebrafish

Output Restriction: Zebrafish

Output Options

Select identifiers to add onto input table:

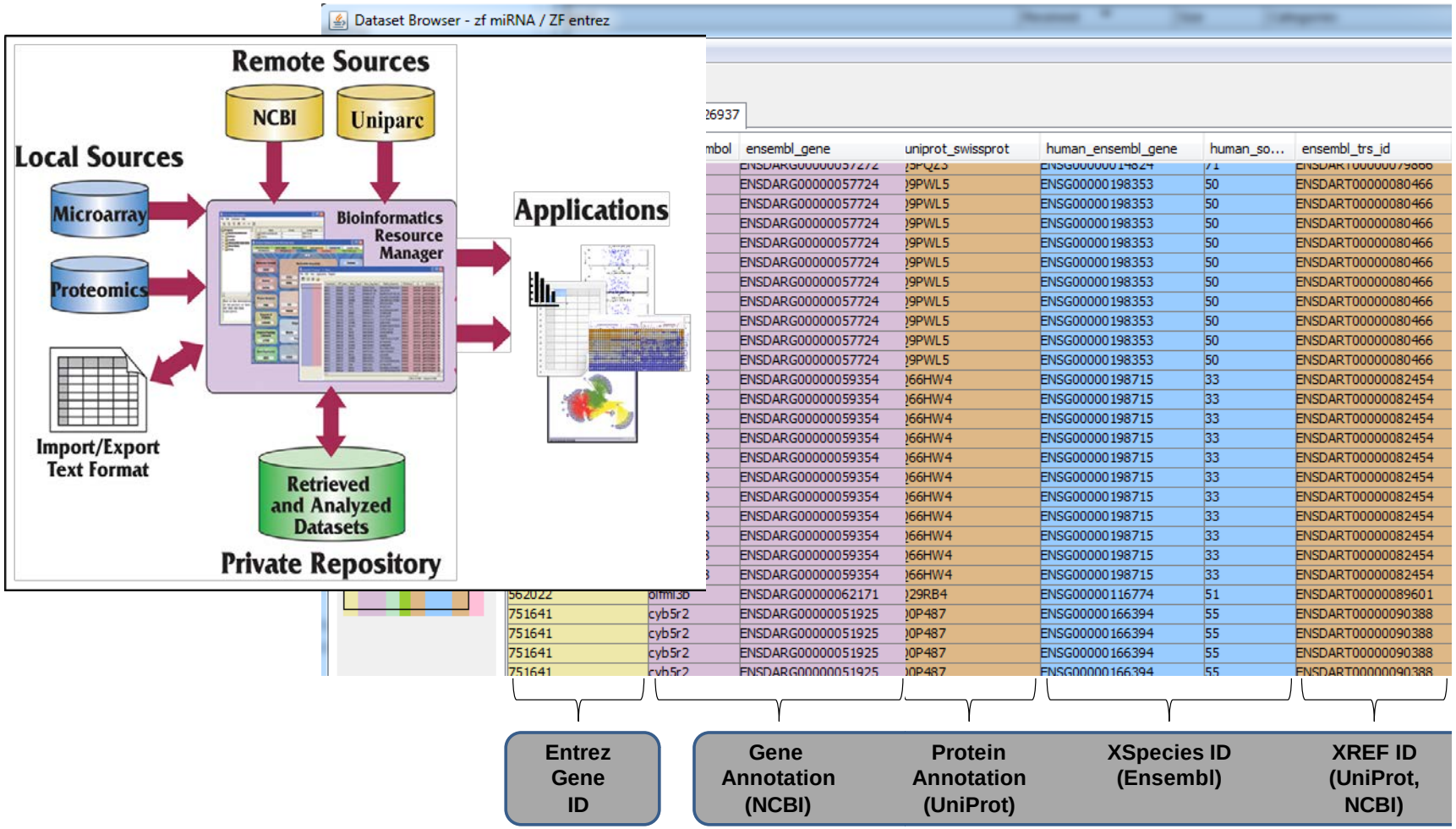
- [ ] Ensembl Gene ID
- [ ] Ensembl Protein ID
- [ ] Ensembl Transcript ID
- [x] Entrez Gene ID
- [ ] Gene Symbol
- [ ] PIR
- [ ] Refseq Gene ID
- [ ] Refseq Protein ID
- [ ] Refseq Transcript ID
- [ ] Swissprot Accession
- [ ] Swissprot ID
- [ ] Tax ID

Select how to handle hits with multiple results:

- (•) Return 1 result
  - (•) The first result
  - ( ) The alphabetically last result
- ( ) Multiple entries per row (separated by semi-colons)
- ( ) Generate extra rows for multiple results

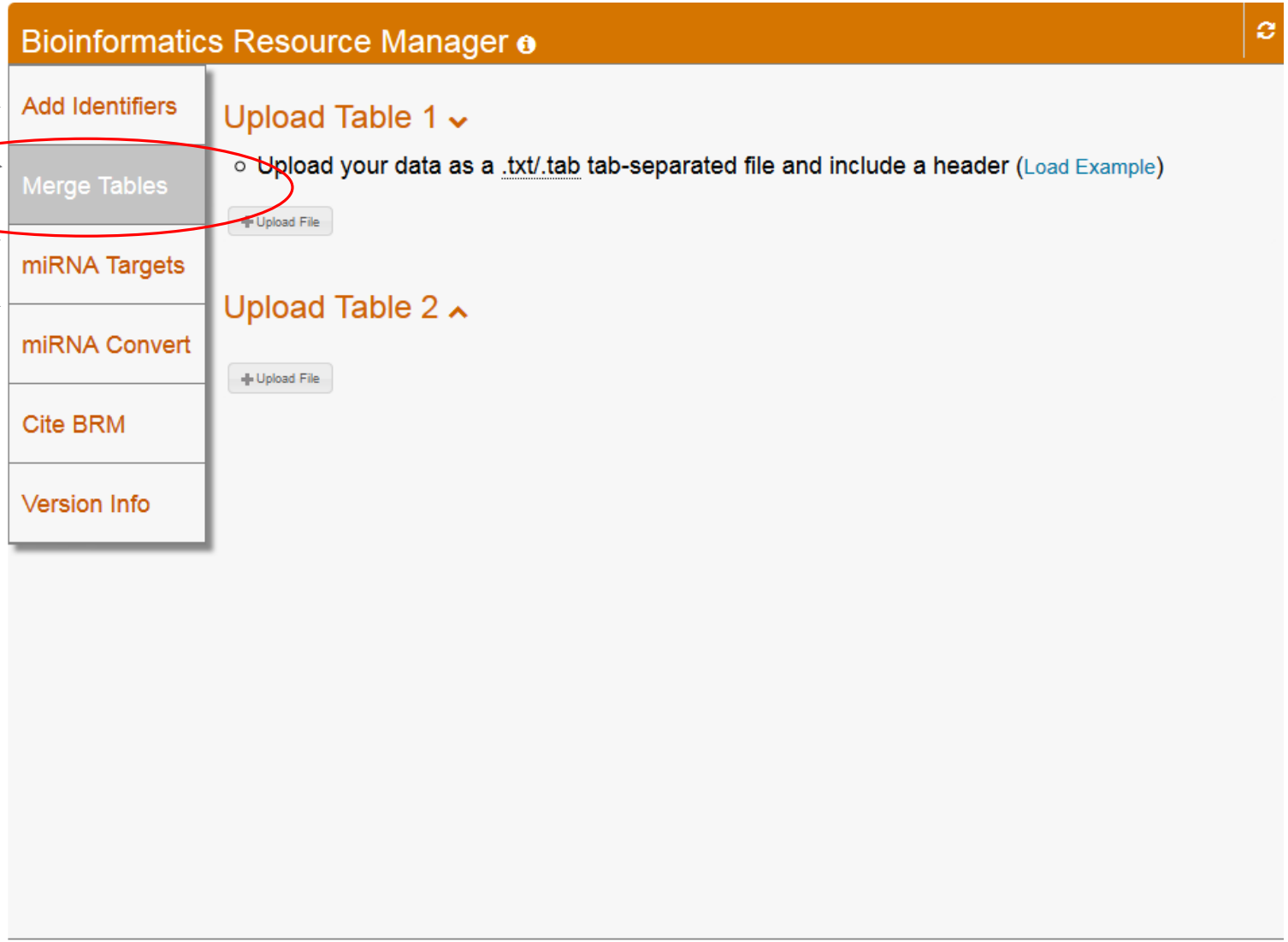
Run Query

BRM: Batch Annotation Retrieval



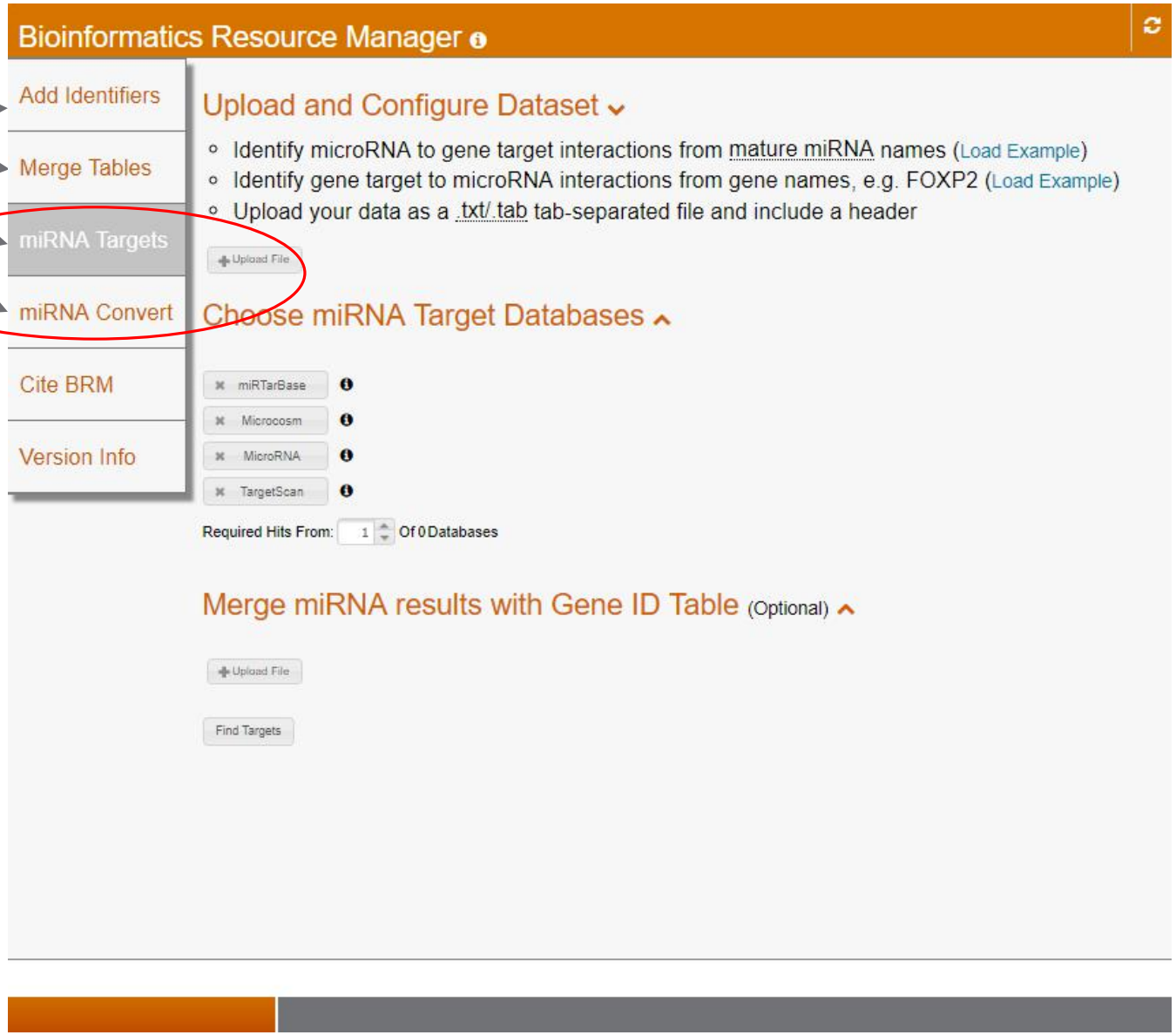
BRM Organization: Homepage

Workflows:



BRM Organization: Homepage

Workflows:



BRM: microRNA tools

- Query predicted and experimentally validated miRNA targets from multiple data sources
  - TargetScan (http://www.targetscan.org/)
  - MicroCosm/miRBase (http://mirbase.org/)
  - miRNA (http://www.microrna.org/)
  - miRTarBase (http://mirtarbase.mbc.nctu.edu.tw/)
- Retrieve potential regulatory miRNAs for known genes
- Retrieve updated miRNA annotation
- Integrate experimentally derived miRNA and mRNA datasets for
  - *Homo sapiens* (human)
  - *Mus musculus* (mouse)
  - *Danio rerio* (zebrafish)
- Use output with other bioinformatics tools:
  - Visualize miRNA regulated gene signatures as heatmap or network
  - Determine functional consequences of miRNA regulation

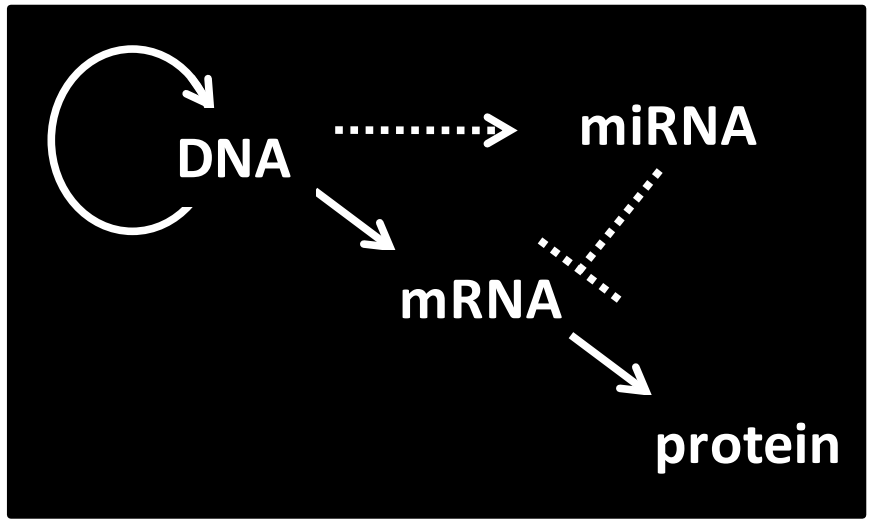
Overview

miRNAs – small non-coding RNAs that regulate gene expression in a sequence specific manner

Importance in biology:

- Developmental timing
- Cell differentiation, proliferation and apoptosis
- Energy metabolism
- Antiviral defense

→ Cancer biology, neurobiology, infectious disease, etc.

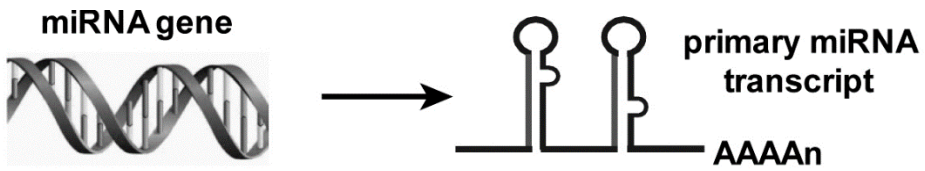


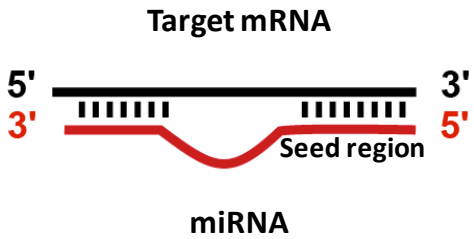
→ Important to identify miRNAs, targets and functional consequences

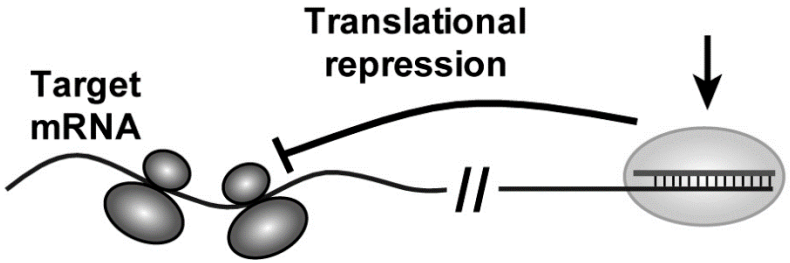
Important Computational Considerations

For miRNA identification and target predictions:

1. Sequence-based
2. Reliance on secondary structure/interactions
3. Evolutionary species conservation







[6]ADAPTED FROM EXP MOL MED (2010) 42:233-244.

BRM: microRNA tools

- Query predicted miRNA targets from multiple data sources
  - TargetScan (http://www.targetscan.org/)
  - MicroCosm/miRBase (http://mirbase.org/)
  - miRNA (http://www.microrna.org/)
  - miRTarBase (http://mirtarbase.mbc.nctu.edu.tw/)

TargetScan
MiRanda

- Retrieve potential regulatory miRNAs for known genes
- Retrieve updated miRNA annotation
- Integrate experimentally derived miRNA and mRNA datasets for
  - *Homo sapiens* (human)
  - *Mus musculus* (mouse)
  - *Danio rerio* (zebrafish)
- Use output with other bioinformatics tools:
  - Visualize miRNA regulated gene signatures as heatmap or network
  - Determine functional consequences of miRNA regulation

miRNA Target Prediction Resources[7-9]

Resource	Type	Website
DIANA-microT-CDS	miRNA target prediction	http://www.microrna.gr/microT-CDS
EvoFold	miRNA target prediction	http://users.soe.ucsc.edu/~jsp/EvoFold/
MicroCosm	miRNA target prediction	http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/
microRNA.org	miRNA target prediction	http://microrna.org
miRDB	miRNA target prediction	http://www.mirdb.org
miRiam	miRNA binding	http://ferrolab.dmi.unict.it/miriam.html
MiRscan	miRNA target prediction	http://genes.mit.edu/mirscan/
PicTar	miRNA target prediction	http://pictar.mdc-berlin.de/
PITA	miRNA target prediction	http://genie.weizmann.ac.il/pubs/mir07/
RNA22	miRNA target prediction	http://cbcsrv.watson.ibm.com/rna22.html
RNAz	miRNA target prediction	http://www.tbi.univie.ac.at/~wash/RNAz/
TargetMiner	miRNA target prediction	http://www.isical.ac.in/~bioinfo_miu/targetminer20.htm
Targetscan	miRNA target prediction	http://www.targetscan.org/
DIANA-TarBase	Manually curated validated mirna target database	http://www.microrna.gr/tarbase
miRecords	Manually curated validated mirna target database	http://mirecords.biolead.org
miRTarBase	Manually curated validated mirna target database	http://mirtarbase.mbc.nctu.edu.tw

Balancing specificity and sensitivity in target prediction

TargetScan

- TargetScanS algorithm
- higher specificity, lower sensitivity
- Considers many parameters (seed matching, 3’ complementarity, local AU content, cross-species conservation)
- Omits sites with poor seed pairing

microRNA.org

- miRanda algorithm
- lower specificity, higher sensitivity
- Considers complementarity and free energy binding, species conservation
- More false positives

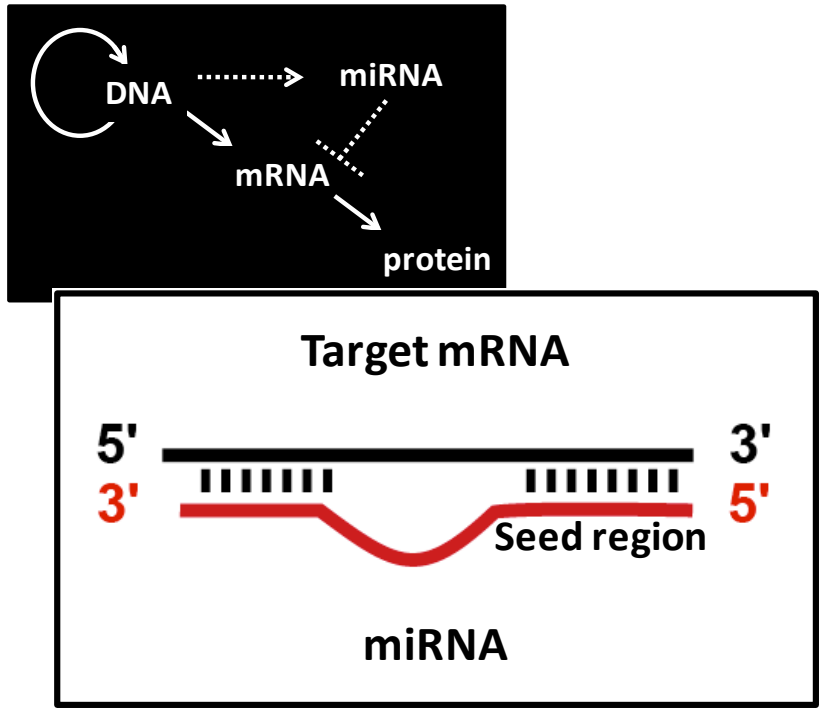
Challenges and Considerations

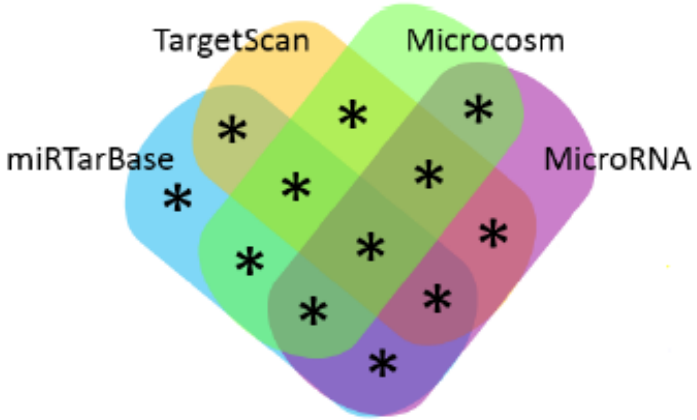
<u>False positives in miRNA prediction</u>:

- Some targets with perfect seed pairing are not repressed experimentally

- Continued discussion about which attributes are most important (sequence v. structure)
  - Low seed pairing stability
  - High target site abundance

- Methods to reduce false positives:
  - Species conservation
  - Combinatorial approaches (multiple tools for prediction)
  - mRNA coexpression





BRM: microRNA tools

- Query predicted and experimentally validated miRNA targets from multiple data sources
  - TargetScan (http://www.targetscan.org/)
  - MicroCosm/miRBase (http://mirbase.org/)
  - miRNA (http://www.microrna.org/)
  - miRTarBase (http://mirtarbase.mbc.nctu.edu.tw/)
- Retrieve potential regulatory miRNAs for known genes
- Retrieve updated miRNA annotation
- Integrate experimentally derived miRNA and mRNA datasets for
  - *Homo sapiens* (human)
  - *Mus musculus* (mouse)
  - *Danio rerio* (zebrafish)
- Use output with other bioinformatics tools:
  - Visualize miRNA regulated gene signatures as heatmap or network
  - Determine functional consequences of miRNA regulation

BRM as a tool for data integration

BRM is a web-based software platform that allows biological researchers the necessary computational and bioinformatics tools for integration and comparison of multiple HTP omics datasets through easy-to-navigate workflows.

- Incorporates cross-reference IDs and annotation directly into your dataset
- Integrates data tables that have no common identifiers using NCBI, Uniparc and Ensembl
- Integrates data tables using multiple IDs and data sources for optimal matching
- Integrates data tables with any common information (string match)
- Works independent of data type and format
- Output includes user’s full dataset(s) (all rows and columns) after data integration or annotation retrieval

Approaches for data integration

- Direct integration
  - Genes/proteins
  - Cross-species
  - In vivo / in vitro
- Integration at functional or pathway level
- Integration based on common upstream regulators
  - Transcription factors
  - miRNAs
- Integration based on statistical and network-based approaches
  - Correlation
  - Clustering
  - Network structure

Hands-on session

BRM: Acknowledgements, publications and funding

BRM Acknowledgements and Funding:

- NIEHS Superfund Research Program P42_ES016465
- BRM was developed through funding from the Pacific Northwest National Laboratory, a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy under Contract DE- AC05-76RL01830.

BRM Web Tool Development Team

Katrina Waters
Elena Peterson
Aaron Phillips
Joe Brown
Susan Tilton

Pacific Northwest NATIONAL LABORATORY

OSU Oregon State UNIVERSITY

Srp Superfund Research Program