



# Managing Omics Data at GLBRC

Yury V Bukhman

Bioinfo-Core Workshop, ISMB 2016

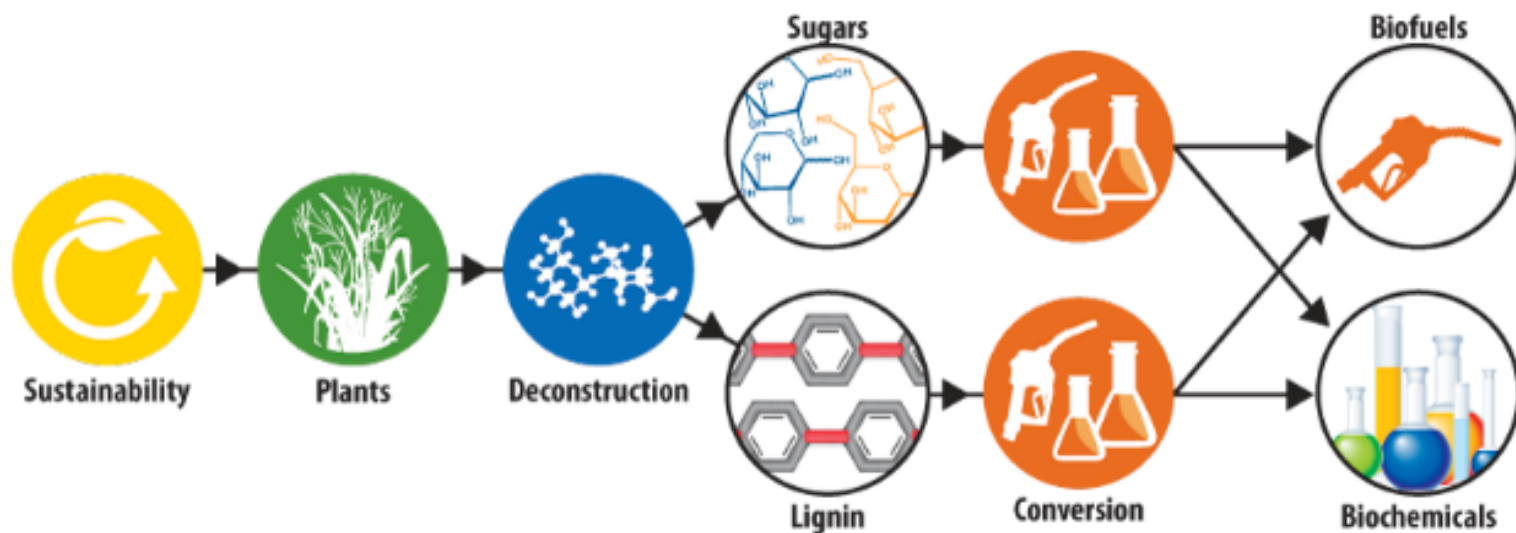
# GLBRC = Great Lakes Bioenergy Research Center

- ✧ The mission of the Great Lakes Bioenergy Research Center is grand, but simply stated: to perform the basic research that generates technology to convert cellulosic biomass to ethanol and other advanced biofuels.
- ✧ DOE-funded, university-based center
- ✧ UW – Madison, Michigan State and other partners

# GLBRC Structure

- ✧ Multiple projects done by academic labs led by professors
- ✧ Enabling technologies
  - Core facilities
  - Informatics and Information Technology (IIT)
    - IT
    - LIMS
    - Computational biology
      - Yury the scientist
    - MSU informatics group

# GLBRC Data Management Challenges

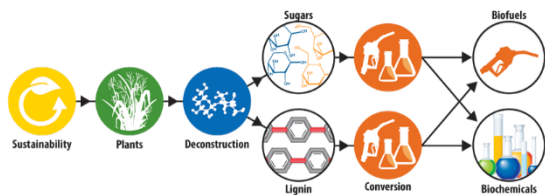


- ✧ A wide variety of research projects → many different data types, including omics
- ✧ University labs have a high degree of independence, different needs, different practices
- ✧ Geographic separation

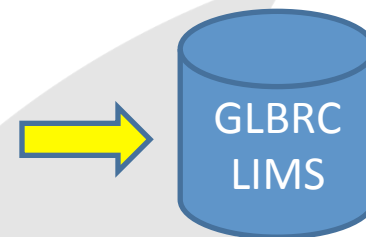
# GLBRC Data Management Solutions

- ✧ Each project has a data management plan, reviewed by IIT
  - Several projects have their own databases
- ✧ Centralized informatics resources
  - Shared file servers
  - SharePoint
  - LIMS
  - GLOW metadata database
  - GxSeq: Genome Browser, Expression Viewer etc.

# Omic Data Flow



Lab samples and materials: field samples, biomass, hydrolysates...



Sequencing Facilities



Bioinformatics workflows



# GLBRC LIMS

**STARLIMS** [About](#) [Add/Remove Content](#) [Logout](#)

[Dashboard](#) [GLBRC Sample Viewer](#)

**Folders List** [Search](#) [Export To Excel](#) Submitter's Lab:

Approval S	Approved	Approved	Status	Folder #	Exp #	Service	Test Plan	Submitted By	Project #	Service Lab	Notes
Releas...	Auto	06/11...	Done	UFS140054	0	Cell Wall Analyses	Cell Wall Compositio...	RRANATUN...	5.5.2	Cell Wall ...	
Releas...	Auto	06/11...	Done	UFS140054	0	Cell Wall Analyses	Cell Wall Compositio...	RRANATUN...	5.5.2	Cell Wall ...	
Releas...	Auto	06/11...	Done	UFS140053	0	Cell Wall Analyses	Cell Wall Compositio...	RRANATUN...	5.5.2	Cell Wall ...	
Releas...	Auto	06/11...	Done	UFS140053	0	Cell Wall Analyses	Cell Wall Compositio...	RRANATUN...	5.5.2	Cell Wall ...	
Releas...	Auto	06/11...	Done	UFS140052	0	Cell Wall Analyses	Cell Wall Compositio...	RRANATUN...	5.5.2	Cell Wall ...	
Releas...	Auto	06/11...	Done	UFS140052	0	Cell Wall Analyses	Cell Wall Compositio...	RRANATUN...	5.5.2	Cell Wall ...	
Releas...	Auto	06/02...	Done	UFS140051	0	Cell Wall Analyses	Cell Wall Compositio...	RRANATUN...	5.5.2	Cell Wall ...	
Releas...	Auto	06/02...	Done	UFS140051	0	Cell Wall Analyses	Cell Wall Compositio...	RRANATUN...	5.5.2	Cell Wall ...	
Releas...	Auto	06/02...	Done	UFS140050	0	Cell Wall Analyses	Cell Wall Compositio...	RRANATUN...	5.5.2	Cell Wall ...	

**Test Results** **Attachments**

Sample Info	Test	Analyte	Rep	Result	Final	Units	Status	Test Stat	Analyzed D:	Analyzed by	HIGH A	HIGH E	LC	LC
UFS140054001 - FSUWM000000...	Digestibility...	Glucose	1	12.957	12.957	%Yiel...	Done	Done	06/02/2...	CWLIMS				
UFS140054001 - FSUWM000000...	Digestibility...	Glucose	2	9.655	9.655	%Yiel...	Done	Done	06/02/2...	CWLIMS				

Ying Gao, Quansheng Yang, Feiran Yu, LiRong Tao, et al.

# GLOW: the File Metadata Database

Welcome, ybukhman | Logout | [Feedback](#)

GLOW

Experiments

Samples

Workflows

Files

Experimental Protocols

Workflow Templates

Reference Genomes

Reports

Help

File Details [New File](#)

<b>ID:</b>	539
<b>Name:</b>	1007510.GBK4b.03.FPKM.csv
<b>Description:</b>	Normalized counts FPKM
<b>File Path:</b>	/mnt/bigdata/processed_data/a3_rna-seq/Rlandick/1007510.GBK4b.03/1007510.GBK4b.03.counts/1007510.GBK4b.03.FPKM.csv
<b>File Type:</b>	Gene-centric Counts
<b>Subtype:</b>	FPKM
<b>Owner:</b>	omoskvin
<b>Submitted By:</b>	omoskvin
<b>Upload Date:</b>	01/15/2014
<b>File Size:</b>	1.619 MB (1,618,798 Bytes)
<b>Workflows:</b>	1007510.QC.RSEM.1007510.GBK4b.03
<b>Checksum:</b>	2060ed14ee426eae383b8c61beb72220
<b>Download File:</b>	<a href="#">Download</a>
<b>View File:</b>	<a href="#">View</a>

Darin Kalisak, Nathan DiPiazza, Adam Halstead, et al.



# Files are Associated With Samples

Welcome, ybukhman | [Logout](#) | [Feedback](#)

GLOW

Experiments

Samples

Workflows

Files

Experimental Protocols

Workflow Templates

Reference Genomes

Reports

Help

File Catalog

[Upload File](#)

[Adv. Search](#)

Search



By default, the File Catalog displays only files owned by the logged-in user. To view other files, use the Search feature located in the command bar above.

Show  entries

Filter table:

[Previous](#)

[Next](#)

ID ↑↓	Name	File type	Associated Samples	Owner ↑↓	Download	Update
4439	1020957.ReadMapping.unmapped.MERGED.fastq.gz	Sequence:FastQ	AARS, Cont.Corn, Block 1, 2010	dduncan		
4450	1020957.ReadMapping.unmapped.MERGED.fastq.gz	Sequence:FastQ	AARS, Cont.Corn, Block 1, 2010	dduncan		
4546	1020957.ReadMapping.unmapped.MERGED.fastq.gz	Sequence:FastQ	AARS, Prairie, Block 1, 2010	dduncan		
4642	1020957.ReadMapping.unmapped.MERGED.fastq.gz	Sequence:FastQ	AARS, Cont.Corn, Block 1, 2011	dduncan		
4813	1020957.ReadMapping.unmapped.MERGED.fastq.gz	Sequence:FastQ	AARS, Prairie, Block 1, 2011	dduncan		
5056	1020957.ReadMapping.unmapped.MERGED.fastq.gz	Sequence:FastQ	AARS, Cont.Corn, Block 1, 2012	dduncan		
5243	1020957.ReadMapping.unmapped.MERGED.fastq.gz	Sequence:FastQ	AARS, Prairie, Block 1, 2012	dduncan		
5458	1020957.ReadMapping.unmapped.MERGED.fastq.gz	Sequence:FastQ	AARS, Native Grass, Block 1, 2010	dduncan		
5502	1020957.ReadMapping.unmapped.MERGED.fastq.gz	Sequence:FastQ	AARS, Old Field, Block 1, 2010	dduncan		
5576	1020957.ReadMapping.unmapped.MERGED.fastq.gz	Sequence:FastQ	AARS, Miscanthus, Block 1 2010 replant, 2010	dduncan		

Showing 1 to 10 of 2,464 entries

[Previous](#)

[Next](#)

Darin Kalisak, Nathan DiPiazza, Adam Halstead, et al.

# Sample and Its Files

Welcome, ybukhman | Logout | [Feedback](#)

GLOW

Experiments

**Samples**

Workflows

Files

Experimental Protocols

Workflow Templates

Reference Genomes

Reports

Help

Sample Details

[New Sample](#)

[Next Sample](#)→

**ID:** 1

**Name:** 3533\_LIMS\_318\_V1-T2

**Description:** Sequencing Product Name: Microbial Transcriptome, rRNA Depletion, Annotation | Sow Item Type: RNA | Target Fragment Size (bp): 270 | rRNA Depletion: Y | Amplified: Y

**Experiment Type:** RNA-seq

**Owner:** jgrass

**Submitted By:** jgrass

**LIMS References:** There are no references to external LIMS for this sample.

## General Sample Information

### Associated Files

Type	File: Name and Description
BAM:Forward	<a href="#">CBYS.1007510.GBK4b.03.genome.sorted.F.bam</a> Annotated transcriptome based binary alignment map sorted by coordinate and restricted to the forward strand of the genome
BAM:non-strand specific	<a href="#">CBYS.1007510.GBK4b.03.transcript.sorted.bam</a> Annotated transcriptome based binary alignment map based on the transcripts inferred by RSEM and sorted by coordinate <a href="#">CBYS.1007510.GBK4b.03.genome.sorted.bam</a> Annotated transcriptome based binary alignment map sorted by coordinate
BAM:Reverse	<a href="#">CBYS.1007510.GBK4b.03.genome.sorted.R.bam</a> Annotated transcriptome based binary alignment map sorted by coordinate and restricted to the reverse strand of the genome
<a href="https://glow.glbrc.org/datafiles/forward">https://glow.glbrc.org/datafiles/forward</a>	<a href="#">CBYS.1007510.GBK4b.03.genome.sorted.F.bam.bai</a>

Darin Kalisak, Nathan DiPiazza, Adam Halstead, et al.

# Samples are Grouped Into Experiment Sets

GLOW

- Experiments
- Samples
- Workflows
- Files
- Experimental Protocols
- Workflow Templates
- Reference Genomes
- Reports
- Help

Experiment Set Details

Add To... ▾

← Previous Experiment Set

**ID:** 1  
**Name:** Escherichia coli K-12 MG1655 3.1.4 Multiomics #1  
**Description:** Project 1007510 uploaded from JGI metadata  
**Owner:** jgrass  
**Submitted By:** jgrass

## Samples in this Experiment Set:

Show 10 ▾ entries

Filter table:

Previous Next

ID	Name	Description	Experiment Sets	Experiment Type	Category	Subcategory	Workflows	File Types	Owner	Upload Date	Action
1	3533_LIMS_318_V1-T2	Sequencing Product Name: Mi...	Escherichia coli K-12 MG1655 3.1.4 Multiomics #1;	RNA-seq			1007510.QC.RSEM.1007510.GBK4b.03;	BAM:Forward BAM:Reverse BAM:non-strand specific BAMI:Forward BAMI:Reverse BAMI:non-strand specific Sample-centric Cou...	jgrass	01/14/2014	<input type="checkbox"/> Add to set • Create workflow •
	3533_LIMS_318_V1-T3	Sequencing Product Name: Mi...	Escherichia coli K-12 MG1655 3.1.4 Multiomics #1;	RNA-seq			1007510.QC.RSEM.1007510.GBK4b.03;	BAM:Forward BAM:Reverse BAM:non-strand specific BAMI:Forward BAMI:Reverse BAMI:non-strand specific Sample-centric Cou...	jgrass	01/14/2014	<input type="checkbox"/> Add to set

Darin Kalisak, Nathan DiPiazza, Adam Halstead, et al.

# GLOW Experiments

Welcome, ybukhman | [Logout](#) | [Feedback](#)

GLOW

Experiments

Samples

Workflows

Files

Experimental Protocols

Workflow Templates

Reference Genomes

Reports

Help

Experiment Set Catalog

Adv. Search

Search



Show 10 entries

Filter table:

Previous

Next

ID	Name	Description	Category	Subcategory	Workflows	Owner	Upload Date	Actions
91	<a href="#">AARS</a>	Samples taken from the intensive cropping system experiment at AARS				dduncan	03/19/2015	
63	<a href="#">Aerobic_Anaerobic_RNA-seq_Comparison</a>	This experiment set was performed by Kevin Myers in the Gasch Lab to study and compare gene expression levels (via RNA-seq) for Trey Sato's evolved strains (Y22-3, Y127 and Y128) grown both aerobically and anaerobically at 30°C in YPD (2% Glucose), YPX (2% Xylose), YPGalactose (2% Galactose), YPS...	Strain comparison			kmyers	04/11/2014	
92	<a href="#">BCSEs</a>	Samples taken from the intensive cropping system experiments at AARS and KBS				dduncan	03/19/2015	
79	<a href="#">BSA Aneuploidy Tolerance Genomic-Seq</a>	BSA Pool F Genomic-Seq to determine W303 vs. YPS1009 allelic frequency in segregants from the hybrid of haploid W303 and YPS1009 ChrXII Non-Disomes (as a control). These segregants of Pool F were defined as having large colony	Strain comparison	Aneuploidy		jhose	07/17/2014	

<https://glow.glbrc.org/experiments#ASearchModal>

Darin Kalisak, Nathan DiPiazza, Adam Halstead, et al.

# Workflows are Run on Experiment Sets, Consume and Produce Files

Welcome, ybukhman | Logout | Feedback

GLOW

Experiments

Samples

Workflows

Files

Experimental Protocols

Workflow Templates

Reference Genomes

Reports

Help

Workflow Details

File Path List

**Template:** [Low-level RNA-Seq data processing: Bowtie-RSEM \(Standard RNA-seq processing workflow template #1\)](#)

**Owner:** jgrass

**Submitted By:** jgrass

**Parameter Signature:** Rscript GLSeq.top.R updateFromDb dataPrepare exprRun resCollect expID protID

**Actual Parameters:** /mnt/bigdata/processed\_data/a3\_rna-seq/Rlandick/1007510.GBK4b.03/1007510.GBK4b.03.stat/1007510.GBK4b.03.runParam.txt

**Software Version:**

**Command Line Invocation:**

**Status:** Complete

**Start Time:**

**Completion Time:**

**Comment:**

**Experiment Set:** [Escherichia coli K-12 MG1655 3.1.4 Multiomics #1](#)

**Reference Genome:** [E.coli MT203\\_pPET double feature version](#)

## Files Associated with this Workflow:

Workflow Name	Filetype	File Name (check box to select)
1007510.QC.RSEM.1007510.GBK4b.03	Gene-centric Counts:FPKM	<input type="checkbox"/> 1007510.GBK4b.03.FPKM.csv
		<input type="checkbox"/> m.1007510.GBK4b.03.FPKM.csv
	Gene-centric Counts:FPKM_lower	<input type="checkbox"/> 1007510.GBK4b.03.FPKM_lower.csv

Darin Kalisak, Nathan DiPiazza, Adam Halstead, et al.

# GxSeq Gene Expression Viewer <sup>14</sup>

GLBRC GxSeq

ybukhman | My Account | Log Out

Help Tools Samples Features Sequence

## Expression Viewer - Matrix for: *Brachypodium distachyon* ( Bd21 )

Update Selection Download

Filter Favorites: None Search definition and locus:  Go

(Advanced Options)

← Previous 1 2 3 4 5 6 7 8 9 ... 521 522 Next →

26,091 Matching Results

	↕Bd_192	↕tair10	↕Sbicolor 1.4	↕CTL-RNAi-pal1-1	↕CTL-RNAi-pal1-2	↕RNAi-pal1-1-1	↕RNAi-pal1-1-2	↕Sum	Options
GeneID:100839285				9435.72	6826.8	6553.75	11019.86	33836.13	Details   Graph   Browser
onein-like protein 2C-like	Bradi2g38120.1		Sb05g022840	6039.65	3105.31	2871.86	9070.81	21087.63	Details   Graph   Browser
se small chainPW9,	Bradi3g26391.1	AT1G67090.1	Sb05g003480	4264.48	4556.47	4398.18	3662.85	16881.98	Details   Graph   Browser
se small chainPWS4.3,	Bradi1g65564.1	AT1G67090.1	Sb02g001725	4172.7	4684.72	3873.82	3445.35	16176.59	Details   Graph   Browser
se small chainPWS4.3,	Bradi4g08500.1	AT1G67090.1	Sb02g042765	4193.09	4534.38	3965.85	3044.95	15738.27	Details   Graph   Browser
s; PF03953.7.fs;	Bradi1g10150.1	AT4G14960.2	Sb01g009560	2935.65	3541.07	3257.52	3087.65	12821.89	Details   Graph   Browser
osylmethionine synthase	Bradi2g12160.1	AT4G01850.2	Sb03g013170	2243.15	2260.26	2906.21	2533.62	9943.24	Details   Graph   Browser
842305	Bradi1g12787.1	AT4G39260.1	Sb01g012300	2431.77	2192.41	1965.51	2842.94	9432.63	Details   Graph   Browser
s; PF00891.8.ls;	Bradi3g16530.1	AT5G54160.1	Sb07g003860	2129.64	1782.27	1959.39	3405.29	9276.59	Details   Graph   Browser
ethyltransferase-like	Bradi1g13290.1	AT5G17920.2	Sb01g012960	1835.59	2214.82	2885.17	2284.07	9219.65	Details   Graph   Browser
	Bradi4g19457.1	AT4G13940.1	Sb05g014470	1818.18	1888.53	2270.99	2540.1	8517.8	Details   Graph   Browser
02496.6.ls	Bradi4g24650.1		Sb08g004190	2084.03	1813.89	1706.95	2717.46	8322.33	Details   Graph   Browser
439.8.ls	Bradi2g06220.1	AT3G15353.1		1911.09	2016.92	2088.74	1668.07	7684.82	Details   Graph   Browser
polyubiquitin-like	Bradi3g04737.1	AT5G03240.3	Sb04g004270	2065.79	1757.84	1875.13	1923.75	7622.51	Details   Graph   Browser
ethyltransferase-like	Bradi4g01200.2	AT3G03780.3	Sb08g022210	1502.49	1755.13	2052.45	1845.14	7155.21	Details   Graph   Browser
LOC100846426	Bradi3g01470.1	AT4G35090.1	Sb04g001130	1434.33	1369.15	1790.25	2377.0	6970.73	Details   Graph   Browser
rm 1: chlorophyll a-b	Bradi2g16290.1	AT2G34420.1	Sb03g027030	2201.31	1062.69	1445.91	1066.33	6676.24	Details   Graph   Browser

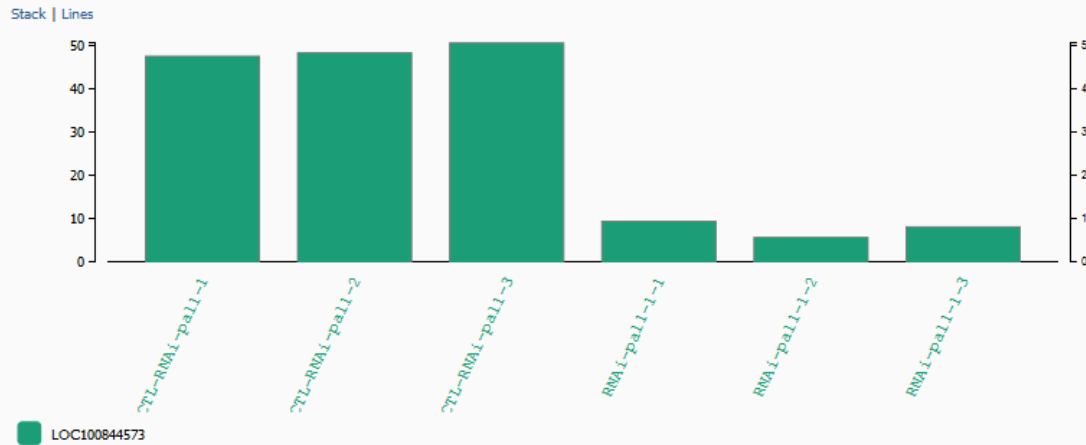
Nick Thrower, Matt Larson, Curt Wilkerson



www.glbrc.org



# GxSeq Gene View



## Density of Aligned Reads

Toggle Stacking

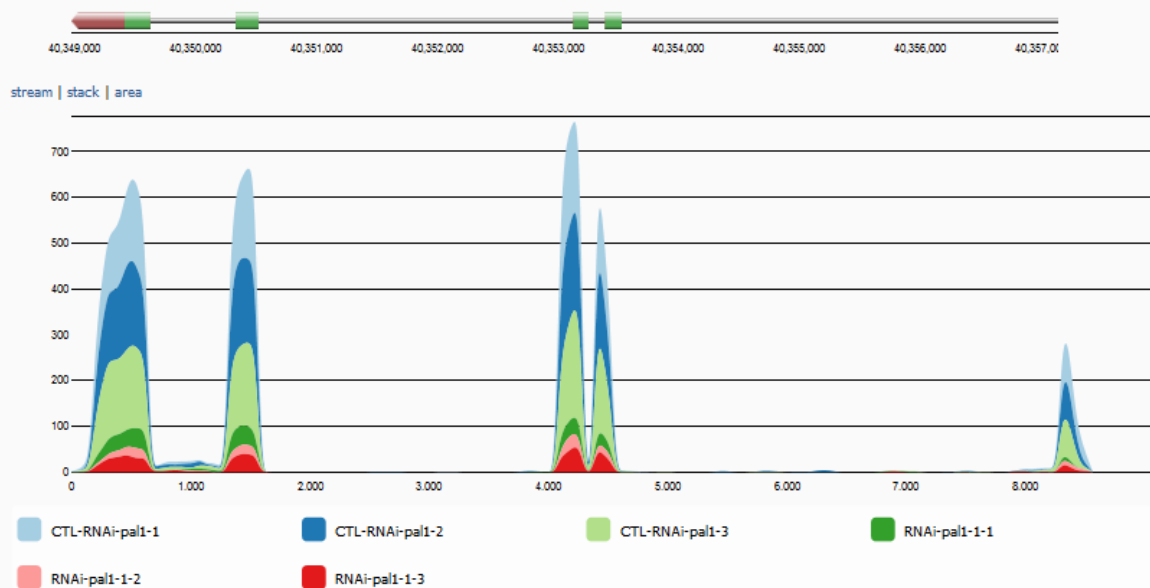


Chart Design Courtesy: Jim Vallandigham - <http://flowingdata.com/>

Nick Thrower,  
Matt Larson,  
Curt Wilkerson

# Open Source

- ✧ GxSeq:  
<https://github.com/Michigan-State-University/gxseq> (version 1)
- ✧ GLOW: coming soon



# Acknowledgements



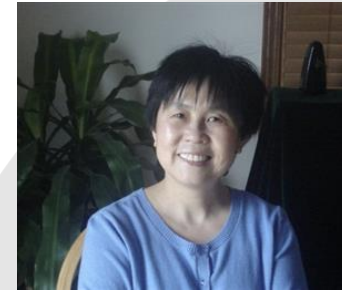
David Benton,  
retiring IIT Director



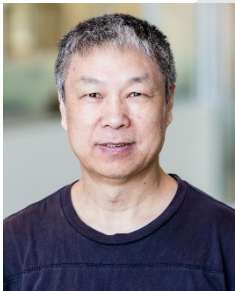
Nihar Sheth,  
new IIT Director



WEI IT: Dirk Norman  
Kevin Leigeb, Branden  
Timm, ...



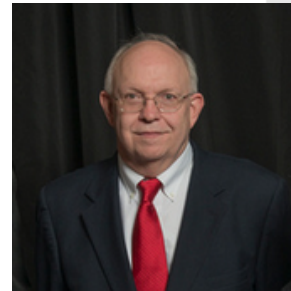
LIMS: Ying Gao  
Quansheng Yang, Feiran Yu,  
LiRong Tao, ...



QA/QC:  
Enhai Xie



GLOW:  
Darin Kalisak,  
Spencer Thiel, Nathan  
DiPiazza, Adam  
Halstead, ...



Curt Wilkerson,  
MSU Informatics  
Group Lead



GxSeq: Nick Thrower,  
Matt Larson

A large, light gray smiley face is centered on the page. The eyes are represented by two large, teardrop-shaped leaves. The mouth is a simple upward-curving arc. The nose is a four-pointed starburst shape. The background is white with a green gradient at the top and bottom.

**THANK YOU**

# Growing Core Data Management Systems

- ✧ Thoughtful design → ability to extend systems and create customized solutions
  - Needs database and software design expertise
  - Needs scientific domain expertise
- ✧ Sound software development practices: documentation, version control, testing etc. → maintainable and extensible systems, robust to staff turnover
- ✧ Agile: release early, release often → get user input as frequently as possible
- ✧ Do well by early adopters → gradually extend reach