



# **Approaches to Bioinformatic Data Analysis**

**David Crossman, Ph.D.  
UAB Heflin Center for Genomic Science**

**Immersion Course  
Friday, April 5, 2013**

# Contents

- **Setting up your project**
- Analysis Challenges
- Tools available for analysis
- Microarrays
- NGS
- Bioinformatic Resources
- References and web links



# Factors to Consider in Genomic Studies

- What is my study design/hypothesis?
- How comprehensive does the data need to be?
- How much work can I afford?
- What is the quality and quantity of my sample
- How fast do I need to have results?
- Will my grant get funded if I don't use the latest technology?

# Study Design/Hypothesis

- I have an organism for which there are no off-the-shelf products for gene expression or sequence analysis.
- I need to comprehensively interrogate the entire genome of my model.
- I am studying a rare disease that I hypothesize to be attributed to private/rare/*de novo* mutation.

**Next Generation Sequencing is a good choice.**

# Study Design/Hypothesis

- I have an organism for which there are array-based products available.
- I want to expand my current work to identify new pathways involved in the physiology/organism I am studying.
- I have a large cohort of humans or animals and want to characterize all individuals.

**Microarray-based assays are a good choice.**

# How much work can I afford?

Microarray		Next Generation Sequencing	
Assay	Cost*	Assay	Cost*
Gene Expression	\$245-650	mRNA-Seq	\$650 <sup>§</sup>
Methylation	\$365	Methyl-Seq	\$650-1800 <sup>¥</sup>
ChIP-ChIP	\$550	ChIP-Seq	\$650-1800 <sup>¥</sup>
Genotyping	\$20-620	Whole Exome	\$2,000 <sup>§</sup>
		Whole Genome	\$5,000 <sup>€</sup>

\*Prices include labor and consumables and are subject to change.

§Price reflects running 28 samples per flowcell on HiSeq2000.

¥Prices reflect running several samples per lane v. one sample per lane.

€Price is for running 3 genomes per flowcell.

**NGS analysis is always going to provide more comprehensive data than an off-the-shelf microarray.**

# Contents



- Setting up your project
- **Analysis Challenges**
- Tools available for analysis
- Microarrays
- NGS
- Bioinformatic Resources
- References and web links

# NGS Analysis Challenges

- Advanced technologies require substantial computing resources.
- File sizes:

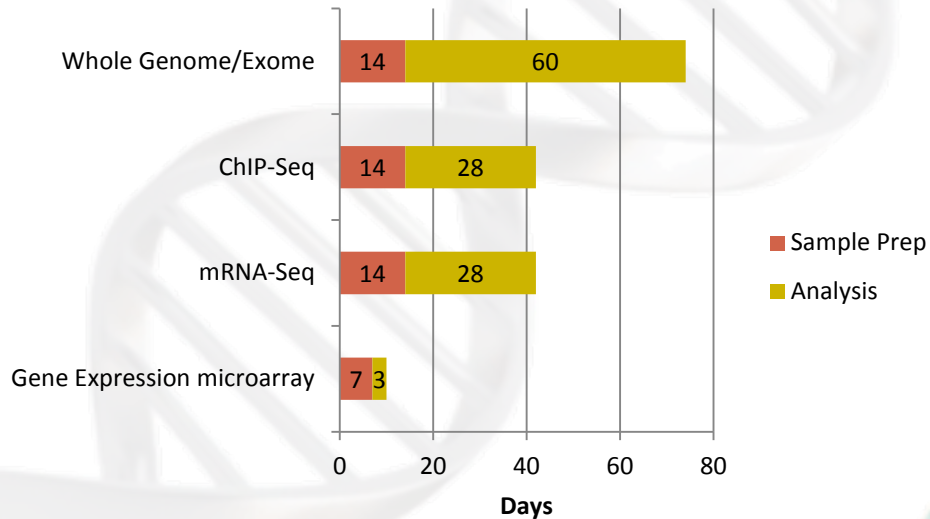
Per Sample (in GB)	Raw Data	Aligned	Spreadsheet	Total
Gene Expression microarray	0.01	NA	0.002-0.005	0.012-0.015
mRNA-Seq	10-26	10-26	0.01-0.03	20.01-52.03
Exome	10-30	10-30	0.01-0.05	20.01-60.05
Whole Genome	250-500	250-500	0.01-0.1	500.01-1000.1

- Processed NGS files can be several TB in size.
- Software for NGS analysis rapidly evolving.
- Need for realistic understanding of data complexity and timeline for analysis.



# Workflow Timeline

- Analysis time varies:



- Additional analyses will extend processing time

# Contents

- Setting up your project
- Analysis Challenges
- **Tools available for analysis**
- Microarrays
- NGS
- Bioinformatic Resources
- References and web links

# Tools available to UAB investigators

- [GeneSpring](#) – Statistical tools for microarray analysis to enable understanding of the data in a biological context.
- [Galaxy](#) – NGS analysis for those afraid of the “blinking cursor.”
- Command line tools to run on UAB’s Cheaha compute cluster:
  - TopHat
  - Cufflinks
  - Bowtie
  - BWA
  - GATK
  - Etc...

# Contents

- Setting up your project
- Analysis Challenges
- Tools available for analysis
- **Microarrays**
- NGS
- Bioinformatic Resources
- References and web links



Data generated is stored here.

# GeneSpring

Icons: heatmaps, tables, bar charts, etc...

Tools to use for analysis.

The screenshot displays the GeneSpring 12.1.1 software interface. At the top, a menu bar includes Project, Search, View, Tools, Annotations, Windows, and Help. Below the menu is a toolbar with various icons for data visualization and analysis. On the left, the Project Navigator shows a hierarchical tree structure for 'Example', including Samples, Interpretations, Analysis, and My Favorites. The main viewing window displays a profile plot of 'Normalized Intensity Values' (y-axis, ranging from -2 to 1) across 'All Samples' (x-axis, split into 'Control' and 'Treated'). The plot shows numerous lines, with red lines indicating up-regulated genes and green lines indicating down-regulated genes. A legend at the bottom right of the plot shows a color gradient from green (-0.9) to red (1.5) for 'Color By [Control]'. On the right side, a 'Workflow' panel lists various analysis tools such as 'Experiment Setup', 'Quality Control', 'Analysis', 'Class Prediction', 'Results Interpretations', 'Pathway Analysis', 'NLP Networks', and 'Utilities'. At the bottom, a status bar indicates 'Displaying 1070; 0 selected' and '319M of 431M'.

Viewing window

# Contents

- Setting up your project
  - Analysis Challenges
  - Tools available for analysis
  - Microarrays
  - **NGS**
  - Bioinformatic Resources
  - References and web links
- **What is Galaxy**
  - What isn't Galaxy
  - FASTQ anatomy
  - Using Galaxy
- 
- University of Alabama at Birmingham  
Heflin Center for Genomic Science

# What is Galaxy

- GUI for genomics
  - for complete analyses: analyze, visualize, share, publish
- A free (for everyone) web service integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage
- Open source software that makes integrating your own tools and data and customizing for your own site simple

For those afraid of the “blinking cursor!” |

# Datasources

- Upload file from your computer
  - FTP support for large datasets
- UCSC table browser
- UCSC Archaea table browser
- BX table browser
- EBI SRA
- BioMart
- Gramene Mart
- Flymine
- modENCODE fly server
- modENCODE modMine
- Ratmine
- YeastMine
- modENCODE worm server
- WormBase
- EuPathDB server
- EncodeDB at NHGRI
- EpiGRAPH server
- GenomeSpace import



# Tool Suites

- Generic Tools
  - Text Manipulation
  - Format Converters
  - FASTA Manipulation
  - Filtering and Sorting
  - Join, Subtract, Group
  - Sequence Tools
  - Multi-species Alignment Tools
  - Genomic Interval Operations
  - Summary Statistics
  - Graphing / Plotting
  - And More!
- NGS
  - QC and manipulation
  - Mapping
  - SAM Tools
  - GATK Tools (beta)
  - Variant Detection
  - Indel Analysis
  - Peak Calling
  - RNA Analysis
  - Picard (beta)
  - BEDTools
  - snpEff

# Create Workflows



# Sharing and Publishing

## Sharing and Publishing History 'Variant Analysis for Sample E18'

### Making History Accessible via Link and Publishing It

This history is currently restricted so that only you and the users listed below can access it. You can:

#### Make History Accessible via Link

Generates a web link that you can share with other people so that they can view and import the history.

#### Make History Accessible and Publish

Makes the history accessible via link (see above) and publishes the history to Galaxy's Published Histories section, where it is publicly listed and searchable.

### Sharing History with Specific Users

You have not shared this history with any users.

#### Share with a user

[Back to Histories List](#)

# Where you can use and build Galaxy

- Public website
  - <https://main.g2.bx.psu.edu/>
- Local instance (<http://getgalaxy.org>)
  - <https://www.uab.edu/galaxy>
  - Galaxy is designed for local installation and customization
    - Just download and run, completely self-contained
    - Easily integrate new tools
    - Easy to deploy and manage on nearly any (unix) system
    - Run jobs on existing compute clusters
- On the cloud (<http://usegalaxy.org/cloud>)
- Tool shed/contributing tools (<http://toolshed.g2.bx.psu.edu/>)

# Tool Shed

(<http://toolshed.g2.bx.psu.edu/>)

## Galaxy Tool Shed

Repositories Help User

2203 valid tools on Dec 04, 2012

### Search

- Search for valid tools
- Search for workflows

### All Repositories

- Browse by category

### Available Actions

- [Login to create a repository](#)

## Categories

search repository name, description

Name	Description	Repositories
<a href="#">Assembly</a>	Tools for working with assemblies	21
<a href="#">Computational chemistry</a>	Tools for use in computational chemistry	4
<a href="#">Convert Formats</a>	Tools for converting data formats	29
<a href="#">Data Source</a>	Tools for retrieving data from external data sources	15
<a href="#">Fasta Manipulation</a>	Tools for manipulating fasta data	23
<a href="#">Genomic Interval Operations</a>	Tools for operating on genomic intervals	21
<a href="#">Graphics</a>	Tools producing images	11
<a href="#">Metagenomics</a>	Tools enabling the study of metagenomes	7
<a href="#">Micro-array Analysis</a>	Tools for performing micro-array analysis	3
<a href="#">Next Gen Mappers</a>	Tools for the analysis and handling of Next Gen sequencing data	47
<a href="#">Ontology Manipulation</a>	Tools for manipulating ontologies	6
<a href="#">Phylogenetics</a>	Tools for performing phylogenetic analysis	3
<a href="#">Proteomics</a>	Tools enabling the study of proteins	2
<a href="#">SAM</a>	Tools for manipulating alignments in the SAM format	20
<a href="#">Sequence Analysis</a>	Tools for performing Protein and DNA/RNA analysis	111
<a href="#">SNP Analysis</a>	Tools for single nucleotide polymorphism data such as WGA	19
<a href="#">Statistics</a>	Tools for generating statistics	26
<a href="#">Systems Biology</a>	Systems biology tools	2
<a href="#">Text Manipulation</a>	Tools for manipulating data	32
<a href="#">Tool Generators</a>	Tools that make or help make new tools	1
<a href="#">Visualization</a>	Tools for visualizing data	20
<a href="#">Web Services</a>	Tools enabling access to web services	3

# Contents

- Setting up your project
  - Analysis Challenges
  - Tools available for analysis
  - Microarrays
  - **NGS**
  - Bioinformatic Resources
  - References and web links
- What is Galaxy
  - **What isn't Galaxy**
  - FASTQ anatomy
  - Using Galaxy
- 
- University of Alabama at Birmingham  
Heflin Center for Genomic Science

# What isn't Galaxy

- Latest version of tools not always available (unless your willing to modify the wrapper for them)
- Not all options for tools are available
  - Examples:
    - TopHat unaligned reads file is not kept
    - Log files not kept
- Your favorite tool isn't there (need to write a wrapper to install it)
- Still buggy (although getting better with each new release!)
  - Example:
    - Job states is complete (by green colored box), but downstream tools can't use it because it didn't completely write all the file.
- Reproducible?

Solution? Blinking Cursor! |

# Contents

- Setting up your project
  - Analysis Challenges
  - Tools available for analysis
  - Microarrays
  - **NGS**
  - Bioinformatic Resources
  - References and web links
- What is Galaxy
  - What isn't Galaxy
  - **FASTQ anatomy**
  - Using Galaxy
- 
- University of Alabama at Birmingham  
Heflin Center for Genomic Science





# Contents

- Setting up your project
  - Analysis Challenges
  - Tools available for analysis
  - Microarrays
  - **NGS**
  - Bioinformatic Resources
  - References and web links
- What is Galaxy
  - What isn't Galaxy
  - FASTQ anatomy
  - **Using Galaxy**
- 
- University of Alabama at Birmingham  
Heflin Center for Genomic Science

# Galaxy Splash Page

<https://www.uab.edu/galaxy>

<https://main.g2.bx.psu.edu/>

The screenshot shows the Galaxy UAB splash page. The browser window title is 'Galaxy' and the address bar contains 'https://galaxy.uabgrid.uab.edu'. The page has a dark blue header with navigation links: 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Admin', 'Help', and 'User'. A 'Using 1.2 Tb' indicator is visible in the top right. On the left, a 'Tools' sidebar lists various categories like 'Get Data', 'Send Data', 'ENCODE Tools', 'Text Manipulation', etc. The main content area features a green 'Welcome to UAB Galaxy!' message with a checkmark icon. Below the message is a workflow diagram titled 'WWFSMD? grow noodly appendages...' from 'usegalaxy.org'. The diagram shows a sequence of tools: 'Input dataset', 'Filter', 'Join', 'Strip', 'Sort', 'Join from Query', and 'Subset from'. The 'Join' tool is highlighted with a blue box. At the bottom, a text line reads: 'This project is supported in part by NSF, NHGRI, and the Huck Institutes of the Life Sciences.'

# Random Galaxy icons/colors

## Colors

Z: ControlR1FastQC data 4.html  
 Z: ControlR1FastQC data 4.html  
 Z: ControlR1FastQC data 4.html  
 48: MF2-3: Cuffmerge on data 42, data 15, and data 46: merged transcripts

Queued

Running

Completed

Failed

## Download/Save

**14: Control Tophat for Illumina on data 3, data 4, and data 2: accepted hits**  
 19.4 Mb  
 format: bam, database: ?  
 Info: Settings:  
 Output files:  
 "/mnt/galaxyData/tmp/15.1.all.q/t  
 mpv7F3\_v/dataset\_12.\*.ebwt"  
 Line rate: 6 (line is 64 bytes)  
 Lines per side: 1 (side is 64 bytes)  
 Offset rate: 5 (one in 32)  
 FTable chars: 10  
 Strings: unpacked  
 Max bucket size: d  
 display in IGB [Local](#) [Web](#)  
 Binary bam alignments file

**14: Control Tophat for Illumina on data 3, data 4, and data 2: accepted hits**  
 19.4 Mb  
 format: bam, database: ?  
 Info: Settings:  
 Output files:  
 "/mnt/galaxyData/tmp/15.1.all.q/t  
 mpv7F3\_v/dataset\_12.\*.ebwt"  
 Line rate: 6 (line is 64 bytes)  
 Lines per side: 1 (side is 64 bytes)  
 Offset rate: 5 (one in 32)  
 FTable chars: 10  
 Strings: unpacked  
 Max bucket size: d  
 Download Dataset  
 ADDITIONAL FILES  
 Download bam\_index

## Icons



Display data in browser



Edit attributes



Delete



Edit dataset annotation



View details



Run this job again



View in Trackster



Edit dataset tags

# Edit files in History

Edit Attributes

Name:  
Tophat for Illumina on data 3, data \*

Info:  
Settings:  
Output files:

Annotation / Notes:  
None

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build:  
Click to Search or Select

Number of comment lines:

Chrom column:  
1

Start column:  
2

End column:  
3

Strand column (click box & select):  
 1

Name/Identifier column (click box & select):  
 1

Score column for visualization:  
1  
2  
3

Save \*  
Auto-detect

This will inspect the dataset and attempt to correct the above column values if they are not accurate.

**11: Tophat for Illumina on data 3, data 4, and data 2: insertions**

**11: Control Tophat for Illumina on data 3, data 4, and data 2: insertions**

# Contents

- Setting up your project
- Analysis Challenges
- Tools available for analysis
- Microarrays
- NGS
- **Bioinformatic Resources**
- References and web links

Bioinformatics Resources



Heflin Center for Genomic Sciences



UAB  
CCTS

Facilities to help



Heflin Center for Genomic Sciences

Department of Pathology  
Division of Informatics

Other Genomic Centers

You are here

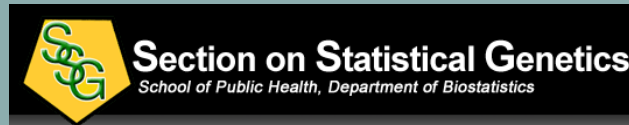
CCC Biostatistics & Bioinformatics Shared Facility (BBSF)



HUDSONALPHA  
INSTITUTE FOR BIOTECHNOLOGY



HUDSONALPHA  
INSTITUTE FOR BIOTECHNOLOGY



Section on Statistical Genetics  
School of Public Health, Department of Biostatistics

# Bioinformatics Resources

- **Heflin Center**
  - David Crossman, Ph.D.
    - [dkcrossm@uab.edu](mailto:dkcrossm@uab.edu)
    - (205) 996-4045
- **CCTS-BMI**
  - Elliot Lefkowitz, Ph.D.
    - [ElliotL@uab.edu](mailto:ElliotL@uab.edu)
    - (205) 934-1946
- **Section on Statistical Genetics** (School of Public Health)
  - Hemant Tiwari, Ph.D.
    - [Htiwari@soph.uab.edu](mailto:Htiwari@soph.uab.edu)
    - (205) 934-4907
- **Department of Pathology Division of Informatics**
  - Jonas Almeida, Ph.D.
    - [jalmeida@uab.edu](mailto:jalmeida@uab.edu)
    - (205) 975-3286
- **Comprehensive Cancer Center (CCC) Biostatistics and Bioinformatics Shared Facility (BBSF)**
  - Karan Singh, Ph.D.
    - [kpsingh@uab.edu](mailto:kpsingh@uab.edu)
    - (205) 996-6122



# Contents

- Setting up your project
- Analysis Challenges
- Tools available for analysis
- Microarrays
- NGS
- Bioinformatic Resources
- **References and web links**

# References and web links

- Galaxy
  - Public website: <https://main.g2.bx.psu.edu/>
  - UAB: <https://www.uab.edu/galaxy>
- GeneSpring
  - <http://genespring-support.com/>
- TopHat
  - Trapnell C, Pachter L, Salzberg SL. [TopHat: discovering splice junctions with RNA-Seq](#). *Bioinformatics* doi:10.1093/bioinformatics/btp120
  - <http://tophat.cbcb.umd.edu/>
- Bowtie
  - Langmead B, Trapnell C, Pop M, Salzberg SL. [Ultrafast and memory-efficient alignment of short DNA sequences to the human genome](#). *Genome Biol* 10:R25.
  - <http://bowtie-bio.sourceforge.net/index.shtml>
- Cufflinks
  - Trapnell C, Williams BA, Pertea G, Mortazavi AM, Kwan G, van Baren MJ, Salzberg SL, Wold B, Pachter L. [Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation](#) *Nature Biotechnology* doi:10.1038/nbt.1621
  - Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. [Improving RNA-Seq expression estimates by correcting for fragment bias](#) *Genome Biology* doi:10.1186/gb-2011-12-3-r22
  - Roberts A, Pimentel H, Trapnell C, Pachter L. [Identification of novel transcripts in annotated genomes using RNA-Seq](#) *Bioinformatics* doi:10.1093/bioinformatics/btr355
  - <http://cufflinks.cbcb.umd.edu/>
- TopHat and Cufflinks protocol
  - Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. [Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks](#) *Nature Protocols* 7, 562-578 (2012) doi:10.1038/nprot.2012.016
- GATK
  - BWA: <http://bio-bwa.sourceforge.net/>
  - GATK: <http://www.broadinstitute.org/gatk/>
- IGV
  - <http://www.broadinstitute.org/igv/>

# Thanks! Questions?

## Contact info:

David K. Crossman, Ph.D.

Bioinformatics Director

Heflin Center for Genomic Science

University of Alabama at Birmingham

<http://www.heflingenetics.uab.edu>

[dkcrossm@uab.edu](mailto:dkcrossm@uab.edu)