Approaches to Bioinformatic Data Analysis

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

Immersion Course
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?

<table>
<thead>
<tr>
<th>Microarray</th>
<th>Next Generation Sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assay</td>
<td>Assay</td>
</tr>
<tr>
<td>Gene Expression</td>
<td>mRNA-Seq</td>
</tr>
<tr>
<td>$245-650</td>
<td>$650§</td>
</tr>
<tr>
<td>Methylation</td>
<td>Methyl-Seq</td>
</tr>
<tr>
<td>$365</td>
<td>$650-1800¥</td>
</tr>
<tr>
<td>ChIP-ChIP</td>
<td>ChIP-Seq</td>
</tr>
<tr>
<td>$550</td>
<td>$650-1800¥</td>
</tr>
<tr>
<td>Genotyping</td>
<td>Whole Exome</td>
</tr>
<tr>
<td>$20-620</td>
<td>$2,000§</td>
</tr>
<tr>
<td>Whole Genome</td>
<td>$5,000€</td>
</tr>
</tbody>
</table>

*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:

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

- 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.
• Analysis time varies:

- **Gene Expression microarray**: 7/3/60 days
- **mRNA-Seq**: 14/28/60 days
- **ChIP-Seq**: 14/28/60 days
- **Whole Genome/Exome**: 14/60/60 days

• 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.

Viewing window

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

Tools to use for analysis.
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
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 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/)

## Categories

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Repositories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly</td>
<td>Tools for working with assemblies</td>
<td>21</td>
</tr>
<tr>
<td>Computational chemistry</td>
<td>Tools for use in computational chemistry</td>
<td>4</td>
</tr>
<tr>
<td>Convert Formats</td>
<td>Tools for converting data formats</td>
<td>29</td>
</tr>
<tr>
<td>Data Source</td>
<td>Tools for retrieving data from external data sources</td>
<td>15</td>
</tr>
<tr>
<td>Fasta Manipulation</td>
<td>Tools for manipulating fasta data</td>
<td>23</td>
</tr>
<tr>
<td>Genomic Interval Operations</td>
<td>Tools for operating on genomic intervals</td>
<td>21</td>
</tr>
<tr>
<td>Graphics</td>
<td>Tools producing images</td>
<td>11</td>
</tr>
<tr>
<td>Metagenomics</td>
<td>Tools enabling the study of metagenomes</td>
<td>7</td>
</tr>
<tr>
<td>Micro-array Analysis</td>
<td>Tools for performing micro-array analysis</td>
<td>3</td>
</tr>
<tr>
<td>Next Gen Mappers</td>
<td>Tools for the analysis and handling of Next Gen sequencing data</td>
<td>47</td>
</tr>
<tr>
<td>Ontology Manipulation</td>
<td>Tools for manipulating ontologies</td>
<td>6</td>
</tr>
<tr>
<td>Phylogenetics</td>
<td>Tools for performing phylogenetic analysis</td>
<td>3</td>
</tr>
<tr>
<td>Proteomics</td>
<td>Tools enabling the study of proteins</td>
<td>2</td>
</tr>
<tr>
<td>SAM</td>
<td>Tools for manipulating alignments in the SAM format</td>
<td>20</td>
</tr>
<tr>
<td>Sequence Analysis</td>
<td>Tools for performing Protein and DNA/RNA analysis</td>
<td>111</td>
</tr>
<tr>
<td>SNP Analysis</td>
<td>Tools for single nucleotide polymorphism data such as WGA</td>
<td>19</td>
</tr>
<tr>
<td>Statistics</td>
<td>Tools for generating statistics</td>
<td>26</td>
</tr>
<tr>
<td>Systems Biology</td>
<td>Systems biology tools</td>
<td>2</td>
</tr>
<tr>
<td>Text Manipulation</td>
<td>Tools for manipulating data</td>
<td>32</td>
</tr>
<tr>
<td>Tool Generators</td>
<td>Tools that make or help make new tools</td>
<td>1</td>
</tr>
<tr>
<td>Visualization</td>
<td>Tools for visualizing data</td>
<td>20</td>
</tr>
<tr>
<td>Web Services</td>
<td>Tools enabling access to web services</td>
<td>3</td>
</tr>
</tbody>
</table>
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
What *isn’t* Galaxy

- Latest version of tools not always available (unless you're 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
NGS FASTQ file format

<table>
<thead>
<tr>
<th>Line</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>@Sanger</td>
</tr>
<tr>
<td>2</td>
<td>Phred+33, raw reads typically (0, 40)</td>
</tr>
<tr>
<td>3</td>
<td>+</td>
</tr>
<tr>
<td>4</td>
<td>@D5VG2KN1:116:CONTAMACXX:5:1101:1606:2077 2:N:0:GTGAAA CTNNCTTCATGTNCCTTCTCTCATGCTTCCCTGAGGCTCTCGTAATC</td>
</tr>
</tbody>
</table>

- Line 2 encodes the quality values for the sequence in Line 1 (see above figure).
- Repeat Lines 1-4 format again and again…
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**
Welcome to UAB Galaxy!

Welcome to the UAB Galaxy platform for experimental biology and comparative genomics designed to help you analyze multiple alignments, compare genomic annotations, profile metagenomic samples and more from your web browser. This platform is built on Galaxy, backed by the Chinese computer cluster, and powered by UABgrid. Documentation on the UAB installation can be found on the UAB Galaxy wiki.

The UAB instance of Galaxy is live as of May 27th, 2011. Please be aware, however, that not all tools or data sets are currently available. Additional tools and data sets are planned, and more can be requested.

Galaxy User Support: In order to facilitate interaction among UAB Galaxy users, share experience, and provide peer-support we have established a galaxy-users group. To join this group and participate in email discussions please subscribe to the galaxy-user group. On-line archives of these discussions are available here. Please note, the email discussions are a public forum. You are advised to only post information you are authorized to share and comfortable with being public.

Galaxy is developed by Penn State and Emory University. The UAB Galaxy platform is a collaborative project between the Biomedical Informatics group of the Center for Clinical and Translational Science and UAB IT Research Computing. This project is supported in part by the UAB Center for Clinical and Translational Science under grant U54 RR025777 from the NIH National Center for Research Resources and by the Office of the Vice President for Information Technology at UAB. Please reference these in any publications resulting from your use of this platform.

WWFSMD?
grow noddy appendages...

usegalaxy.org

This project is supported in part by NSF, NIH, and the Huck Institutes of the Life Sciences.

https://www.uab.edu/galaxy
https://main.g2.bx.psu.edu/
Random Galaxy icons/colors

**Colors**
- Queued
- Running
- Completed
- Failed

**Download/Save**
- 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

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

Facilities to help

You are here

CCC Biostatistics & Bioinformatics Shared Facility (BBSF)

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

Department of Pathology
Division of Informatics
Bioinformatics Resources

- **Heflin Center**
  - David Crossman, Ph.D.
    - dkcrossm@uab.edu
    - (205) 996-4045

- **CCTS-BMI**
  - Elliot Lefkowitz, Ph.D.
    - ElliotL@uab.edu
    - (205) 934-1946

- **Section on Statistical Genetics** (School of Public Health)
  - Hemant Tiwari, Ph.D.
    - Htiwari@soph.uab.edu
    - (205) 934-4907

- **Department of Pathology Division of Informatics**
  - Jonas Almeida, Ph.D.
    - jalmeida@uab.edu
    - (205) 975-3286

- **Comprehensive Cancer Center (CCC) Biostatistics and Bioinformatics Shared Facility (BBSF)**
  - Karan Singh, Ph.D.
    - 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/](https://main.g2.bx.psu.edu/)
  - UAB: [https://www.uab.edu/galaxy](https://www.uab.edu/galaxy)

- **GeneSpring**
  - [http://genespring-support.com/](http://genespring-support.com/)

- **TopHat**
  - [http://tophat.cbcb.umd.edu/](http://tophat.cbcb.umd.edu/)

- **Bowtie**

- **Cufflinks**
  - [http://cufflinks.cbcb.umd.edu/](http://cufflinks.cbcb.umd.edu/)

- **TopHat and Cufflinks protocol**

- **GATK**

- **IGV**
  - [http://www.broadinstitute.org/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