Approaches to Bioinformatics 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
- NGS
- Bioinformatics Resources

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.

Contents

- Setting up your project
- Analysis Challenges
- Tools available for analysis
- NGS
- Bioinformatics Resources

NGS Analysis Challenges

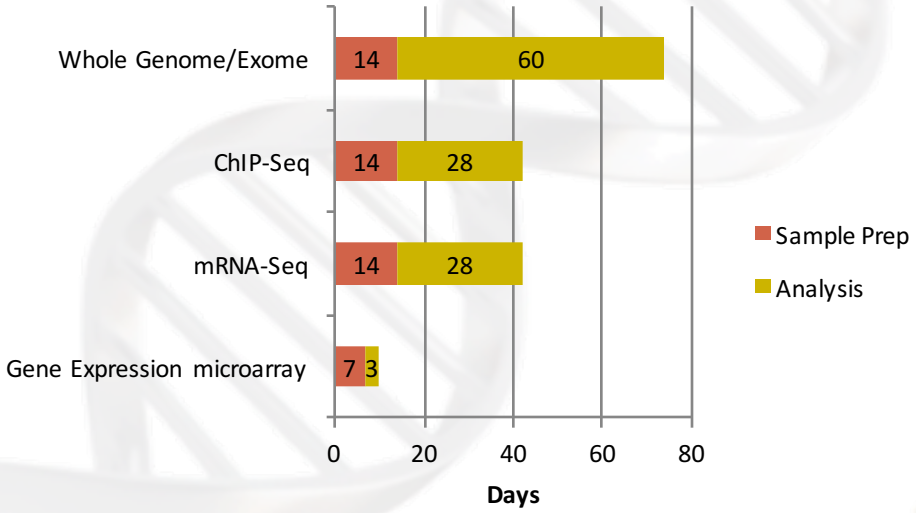
- Advanced technologies require substantial computing resources.
- File sizes:

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
- NGS
- Bioinformatics Resources

Tools available to UAB investigators

- Galaxy – NGS analysis for those afraid of the "blinking cursor."
- Command line tools to run on UAB's Cheaha compute cluster:
  - TopHat
  - Cufflinks
  - Bowtie
  - SAMTools
  - BWA
  - GATK
  - PicardTools
  - MACS2
  - Bismark
  - STAR
  - HISAT
  - Stringtie
  - Ballgown
  - Trinity
  - Freebayes
  - Gemini
  - Circos
  - Homer
  - FastQC
  - SNPEff
  - ANNOVAR
  - muTect
  - MutSig
  - Velvet
  - Abyss
  - vcftools
  - AND MANY MORE!

Contents

- Setting up your project
- Analysis Challenges
- Tools available for analysis
- NGS
  - What is Galaxy
  - What **isn't** Galaxy
  - FASTQ anatomy
  - Using Galaxy
- Bioinformatics Resources

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!” |

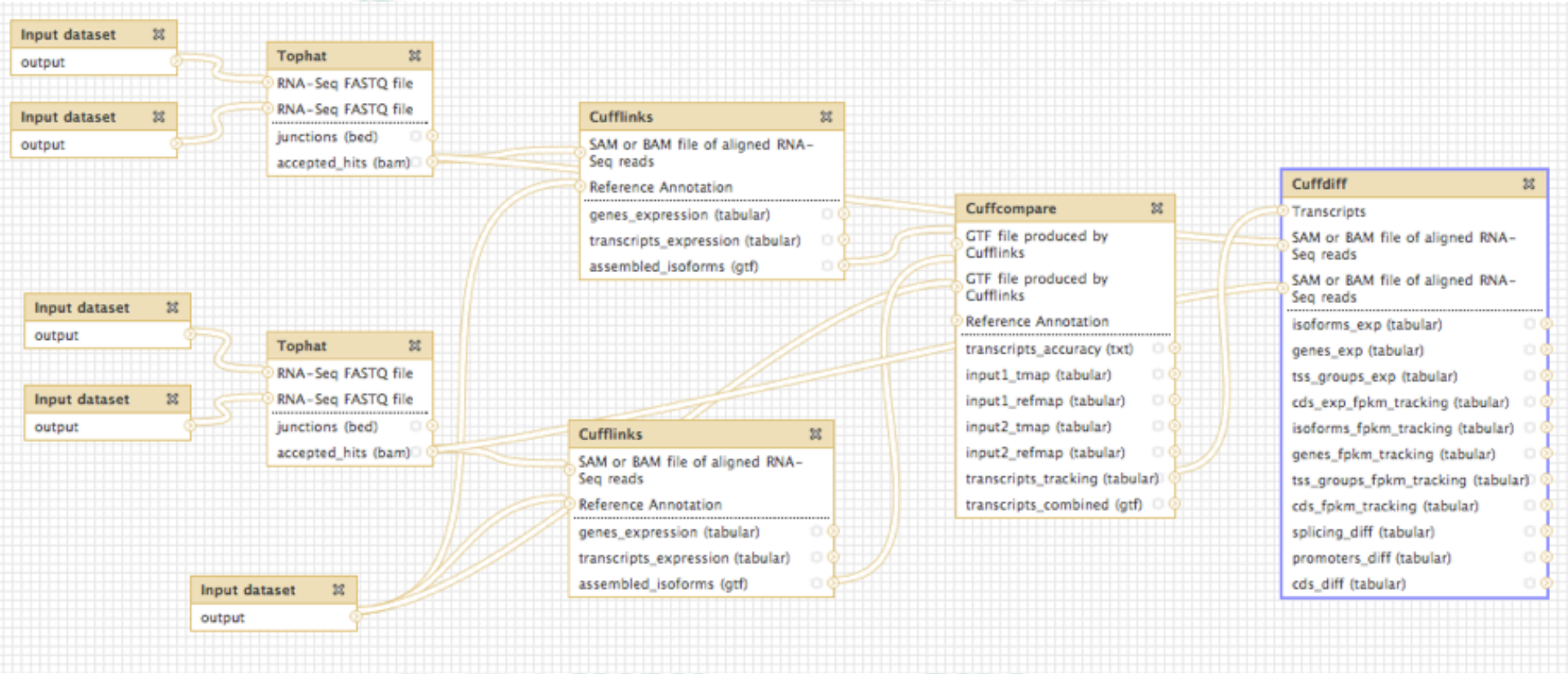
Datasources

- 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

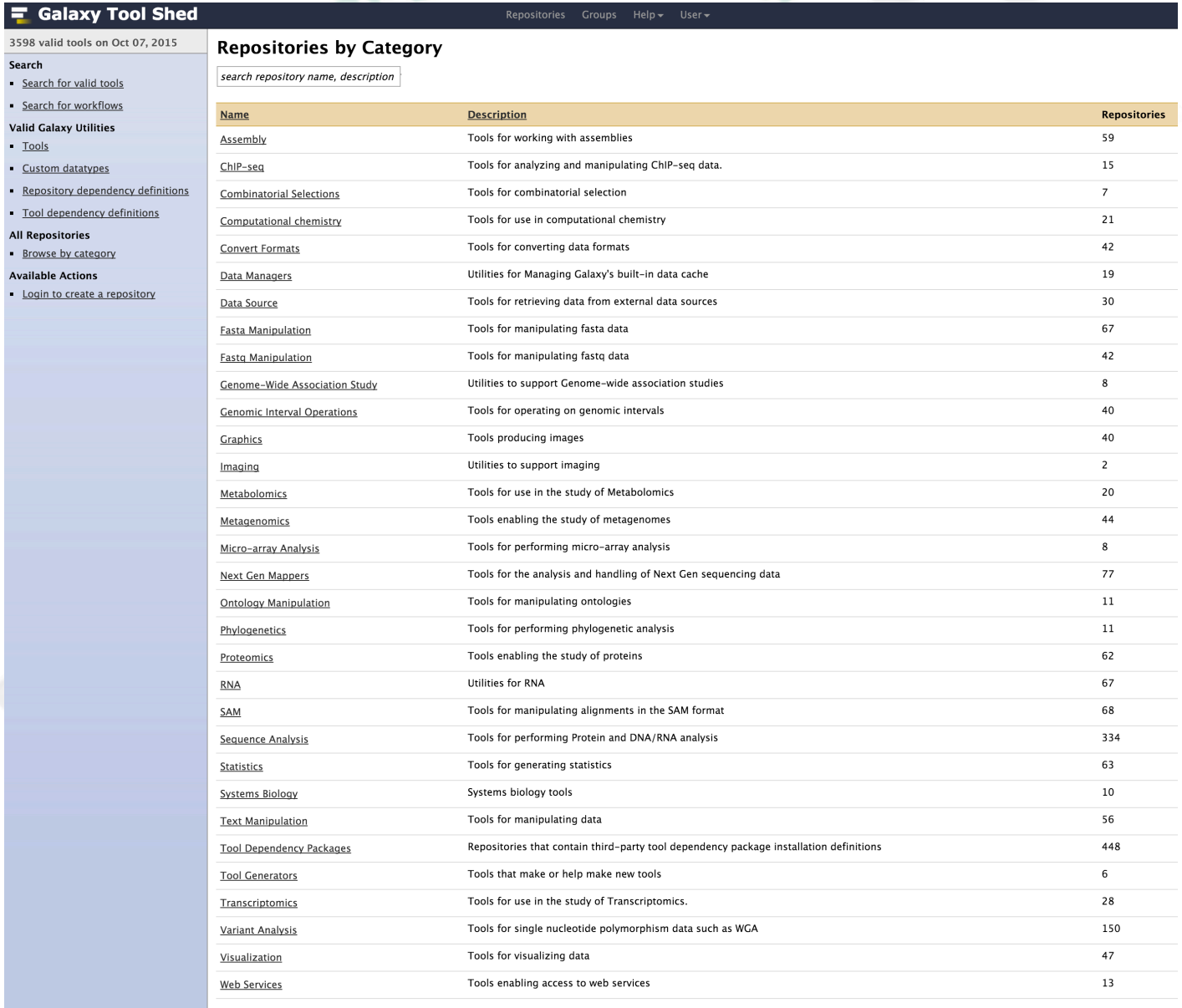
Back to Histories List

Where you can use and build Galaxy

- Penn State's instance: https://usegalaxy.org/
- UAB's instance: https://www.uab.edu/galaxy
- Local instance: http://getgalaxy.org
  - 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 Groups Help User

3598 valid tools on Oct 07, 2015

Search
- Search for valid tools
- Search for workflows

Valid Galaxy Utilities
- Tools
- Custom datatypes
- Repository dependency definitions
- Tool dependency definitions

All Repositories
- Browse by category

Available Actions
- Login to create a repository

Repositories by Category

search repository name, description

Name	Description	Repositories
Assembly	Tools for working with assemblies	59
ChIP-seq	Tools for analyzing and manipulating ChIP-seq data.	15
Combinatorial Selections	Tools for combinatorial selection	7
Computational chemistry	Tools for use in computational chemistry	21
Convert Formats	Tools for converting data formats	42
Data Managers	Utilities for Managing Galaxy's built-in data cache	19
Data Source	Tools for retrieving data from external data sources	30
Fasta Manipulation	Tools for manipulating fasta data	67
Fastq Manipulation	Tools for manipulating fastq data	42
Genome-Wide Association Study	Utilities to support Genome-wide association studies	8
Genomic Interval Operations	Tools for operating on genomic intervals	40
Graphics	Tools producing images	40
Imaging	Utilities to support imaging	2
Metabolomics	Tools for use in the study of Metabolomics	20
Metagenomics	Tools enabling the study of metagenomes	44
Micro-array Analysis	Tools for performing micro-array analysis	8
Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencing data	77
Ontology Manipulation	Tools for manipulating ontologies	11
Phylogenetics	Tools for performing phylogenetic analysis	11
Proteomics	Tools enabling the study of proteins	62
RNA	Utilities for RNA	67
SAM	Tools for manipulating alignments in the SAM format	68
Sequence Analysis	Tools for performing Protein and DNA/RNA analysis	334
Statistics	Tools for generating statistics	63
Systems Biology	Systems biology tools	10
Text Manipulation	Tools for manipulating data	56
Tool Dependency Packages	Repositories that contain third-party tool dependency package installation definitions	448
Tool Generators	Tools that make or help make new tools	6
Transcriptomics	Tools for use in the study of Transcriptomics.	28
Variant Analysis	Tools for single nucleotide polymorphism data such as WGA	150
Visualization	Tools for visualizing data	47
Web Services	Tools enabling access to web services	13

Contents

- Setting up your project
- Analysis Challenges
- Tools available for analysis
- NGS
  - What is Galaxy
  - What **isn't** Galaxy
  - FASTQ anatomy
  - Using Galaxy
- Bioinformatics Resources

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
- NGS
  - What is Galaxy
  - What **isn't** Galaxy
  - FASTQ anatomy
  - Using Galaxy
- Bioinformatics Resources

NGS FASTQ file format

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ..............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
33                        59   64       73                            104                   126

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
   with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
   (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

Line1: Begins with '@' and followed by a sequence identifier and optional description
Line2: Raw sequence letters
Line3: '+'
Line4: Encodes the quality values for the sequence in Line2 (see above figure)
Repeat Lines1-4 format again and again and again…

```
1 @D5VG2KN1:116:C0NTMACXX:5:1101:1606:2077 2:N:0:GTGAAA
2 CTTNNCTTCATGTNCCTTTCCTCTCATGTCTTCCCTGAGGTCCTCGTAATC
3 +
4 B@@##2=2AFDHH#2<CDHHGIII9HHIIEFF:CEHB@DGHGIIIDGEIEH
  @D5VG2KN1:116:C0NTMACXX:5:1101:1584:2079 2:N:0:GTGAAA
  GGGNNTTCATGATNAAGATGAGAGTGCACGGCTTCTCCTCTGAGAAGGACT
  +
  @?;##22=AD84D#2<<;CDH@HG<C>FHGDBFGEH??DBFGEBB<9CEFC
  @D5VG2KN1:116:C0NTMACXX:5:1101:1526:2088 2:N:0:GTGAAA
  TTTNGCAGCACGGCTTTGTCCTCTGGGGTGAGGGCTGGTGTGGGTAGGGCA
  +
  BBB#4=DDBHHHFIJIJIJJGHEGGIJJIJIJJJGIJJIJHIHJGGJGHFE
  @D5VG2KN1:116:C0NTMACXX:5:1101:1730:2093 2:N:0:GTGAAA
  CCCCCAGGCCAGGTAGCCCAAGCCAAGTGTCCAGAGGTTGACCCTGTGCGT
  +
  CC@FFFFFHHHHHIJJIIJIJIJJJJJGIIGIJJHJJHHJIJJJJJJJJIG
```

Contents

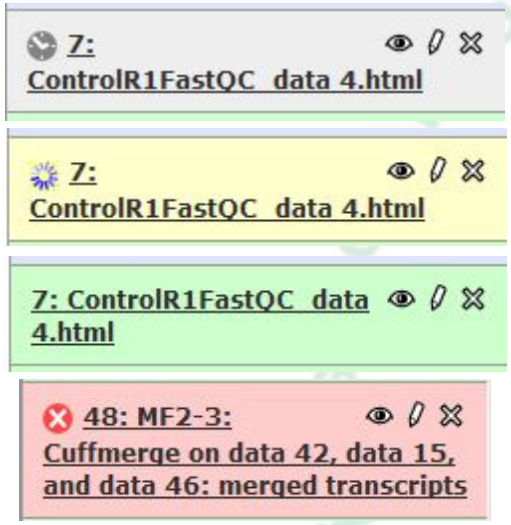
- Setting up your project
- Analysis Challenges
- Tools available for analysis
- NGS
  - What is Galaxy
  - What **isn't** Galaxy
  - FASTQ anatomy
  - Using Galaxy
- Bioinformatics Resources

Galaxy Splash Page

https://www.uab.edu/galaxy
https://usegalaxy.org/

Random Galaxy icons/colors

Colors



Queued

Running

Completed

Failed

Download/Save

Icons

Display data in browser

Edit attributes

Delete

Edit dataset annotation

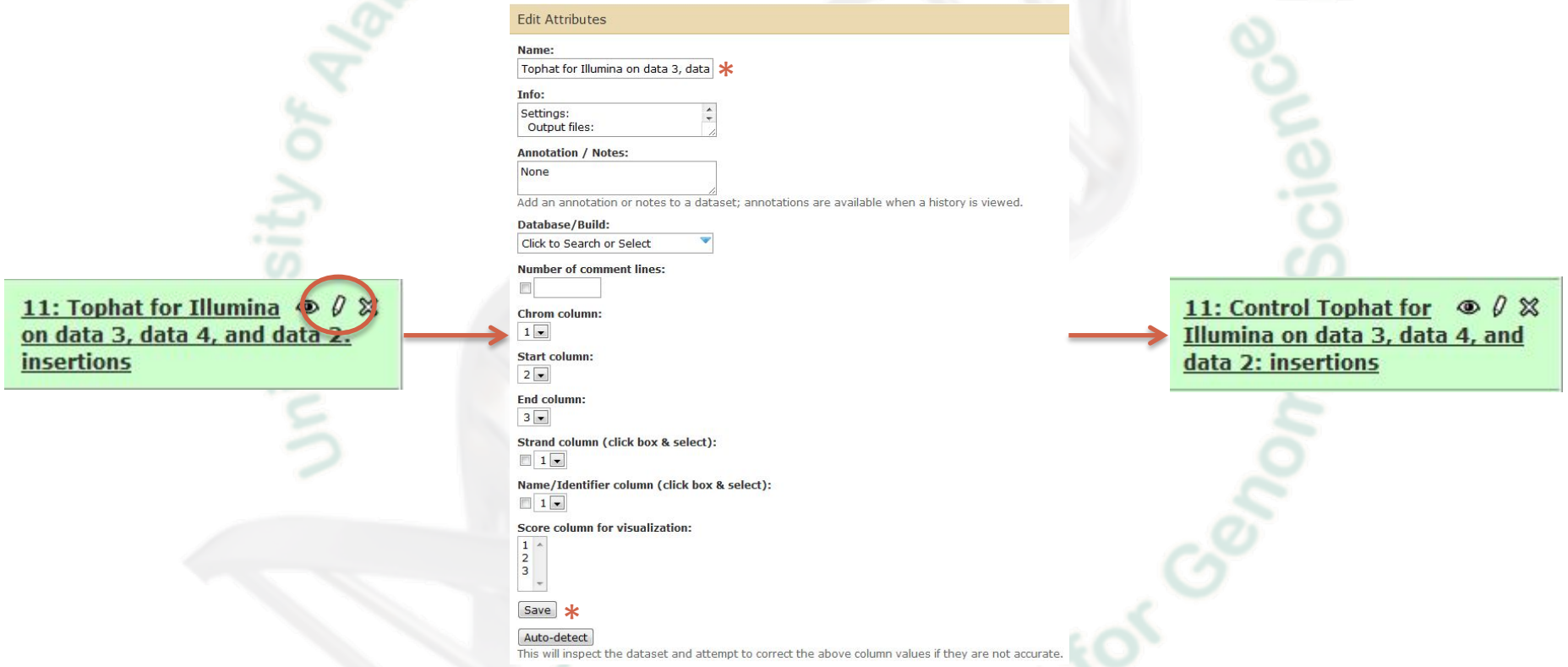
View details

Run this job again

View in Trackster

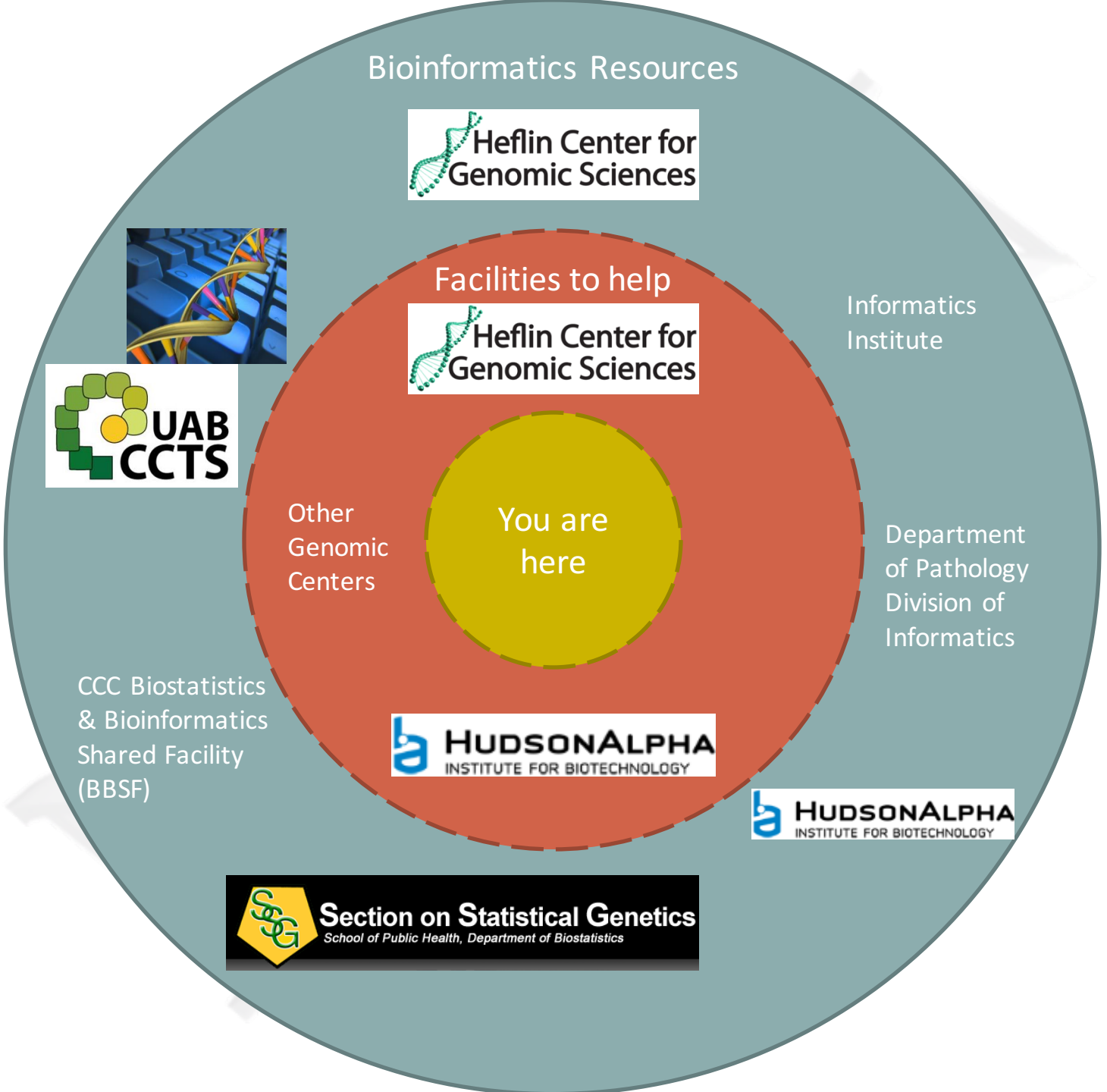
Edit dataset tags

Edit files in History



Contents

- Setting up your project
- Analysis Challenges
- Tools available for analysis
- NGS
- Bioinformatics Resources


Bioinformatics Resources
Heflin Center for Genomic Sciences
Facilities to help
Heflin Center for Genomic Sciences
UAB CCTS
Informatics Institute
Other Genomic Centers
You are here
Department of Pathology Division of Informatics
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
    - (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
- Informatics Institute
  - James Cimino, M.D.
    - ciminoj@uab.edu
    - (205) 996-1958
- **Department of Pathology Division of Informatics**
  - X. Long Zheng, M.D., Ph.D. (Interim)
    - zhengl@uab.edu
    - (205) 975-8161
- **Comprehensive Cancer Center (CCC) Biostatistics and Bioinformatics Shared Facility (BBSF)**
  - Karan Singh, Ph.D.
    - kpsingh@uab.edu
    - (205) 996-6122

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